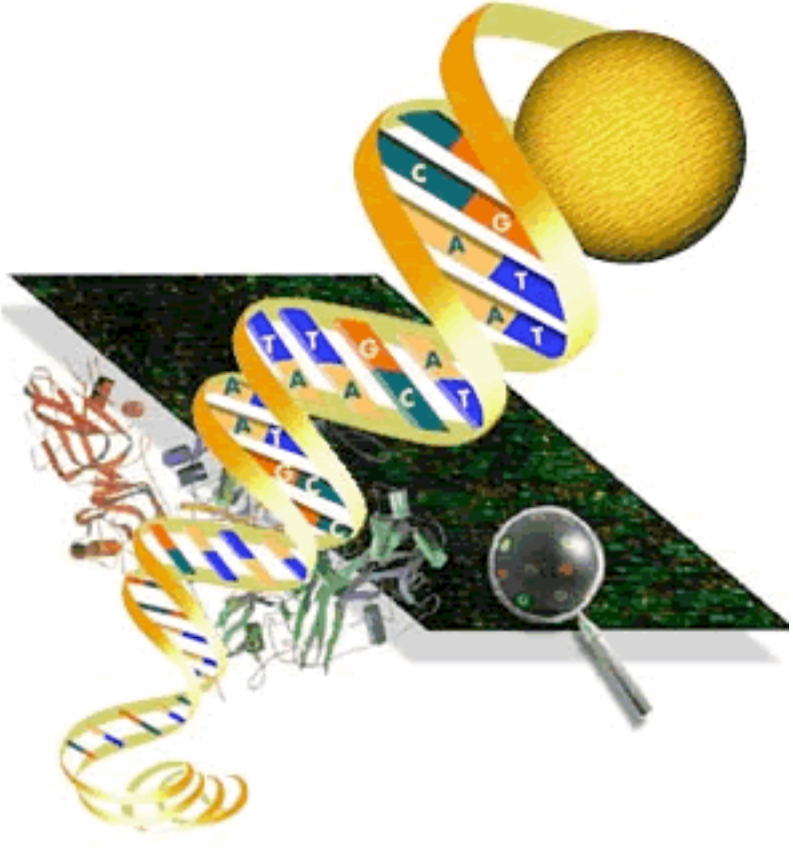


http://www.plantgdb.org/tutorial/annotatemodule/

Annotating Genomes:

where are all the genes?

Jane Loveland

Havana group

Introduction to Making Sense of Genomes
2nd October 2012

Overview

- A bit of background
- What is manual genome annotation?
- Why do we need it?
- How do we do it?

havana

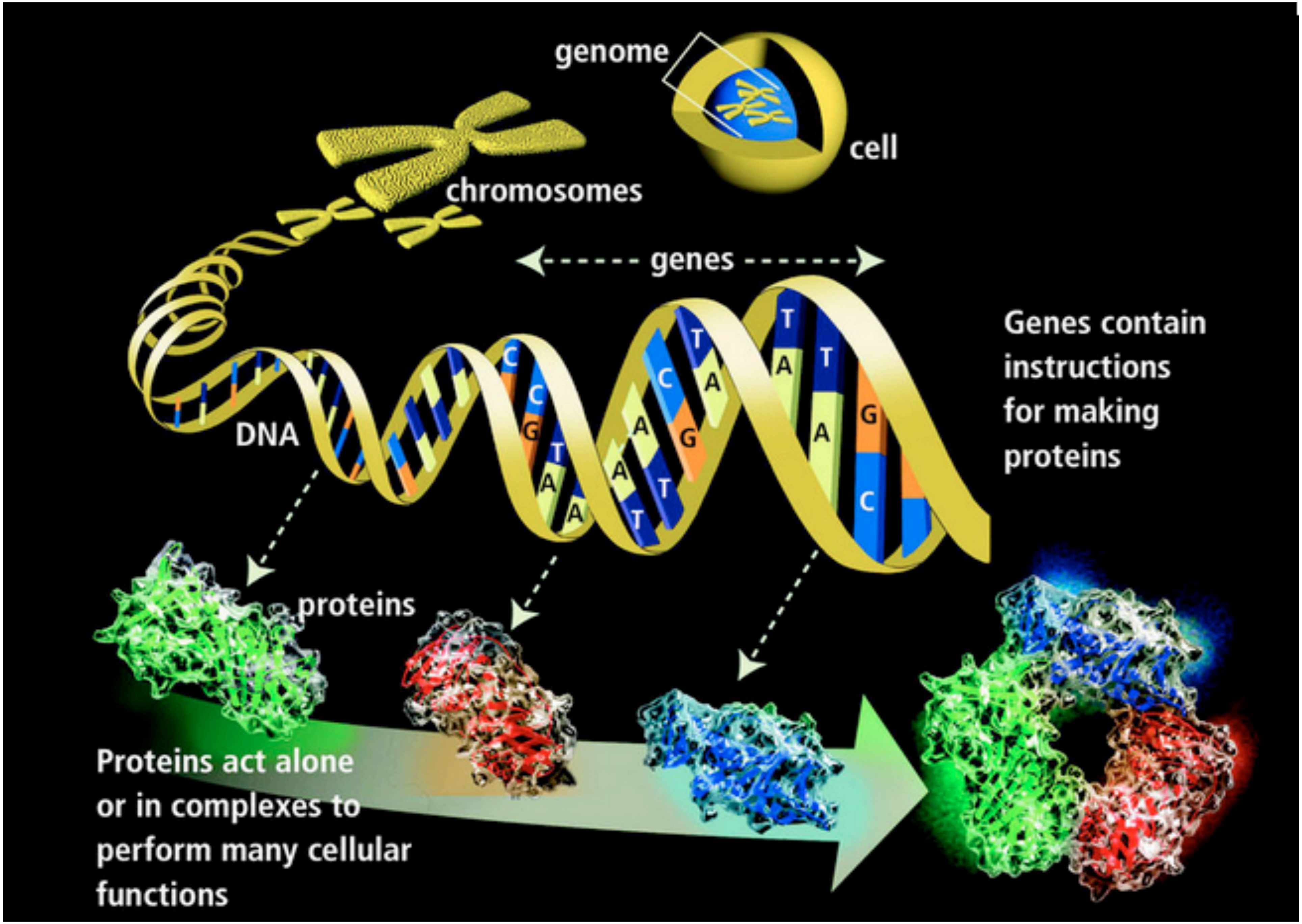
human and vertebrate analysis and annotation

Havana: Human And Vertebrate Analysis aNd Annotation

Manual annotation of human, mouse and zebrafish whole chromosomes or genomes

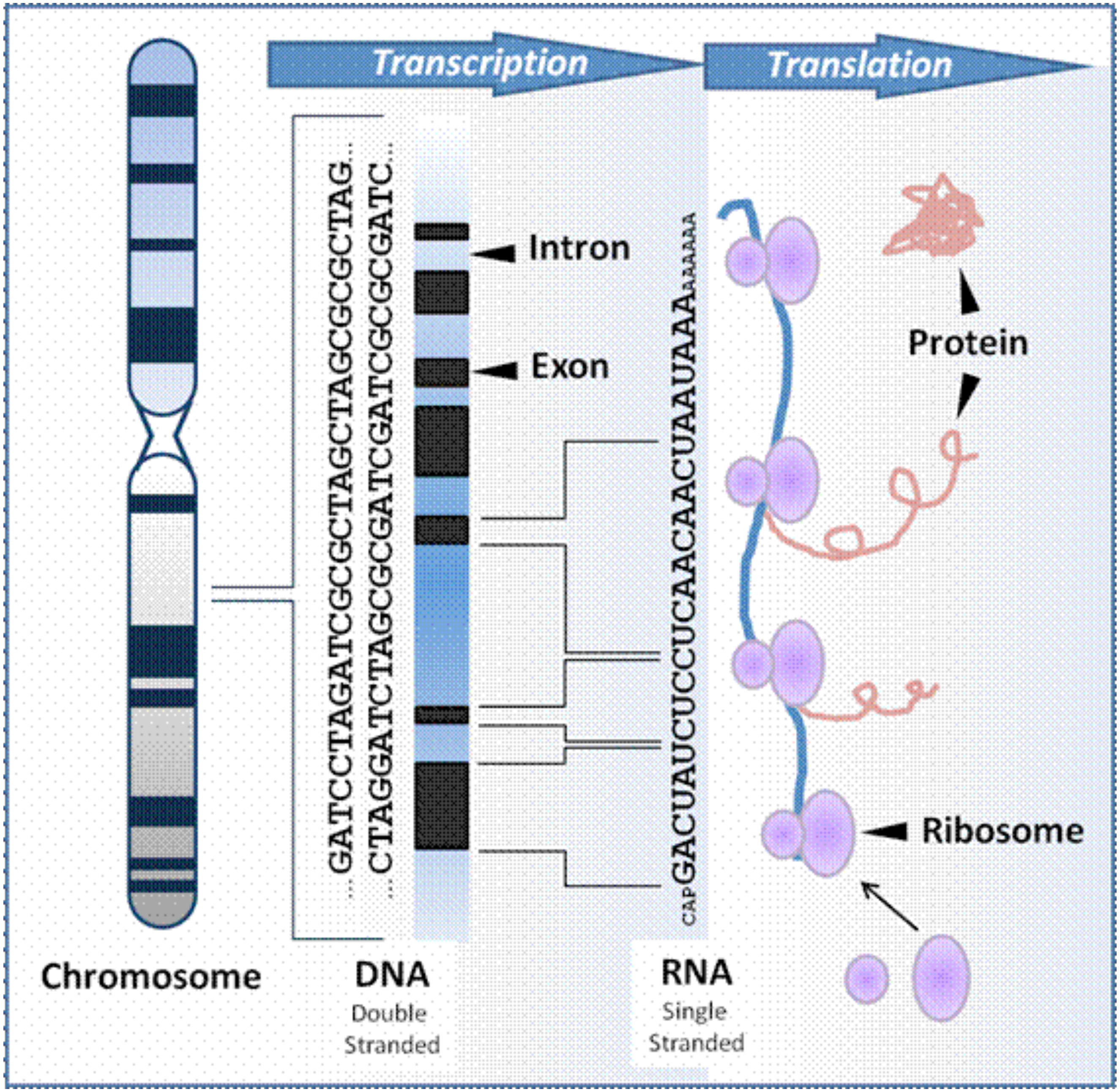
Vega

Vega: VErtebrate Genome Annotation database


genome
cell
chromosomes
genes
DNA
Genes contain instructions for making proteins
proteins
Proteins act alone or in complexes to perform many cellular functions
C
C
G
T
A
A
A
A
T
C
G
T
T
A
T
A
T
A
G
C

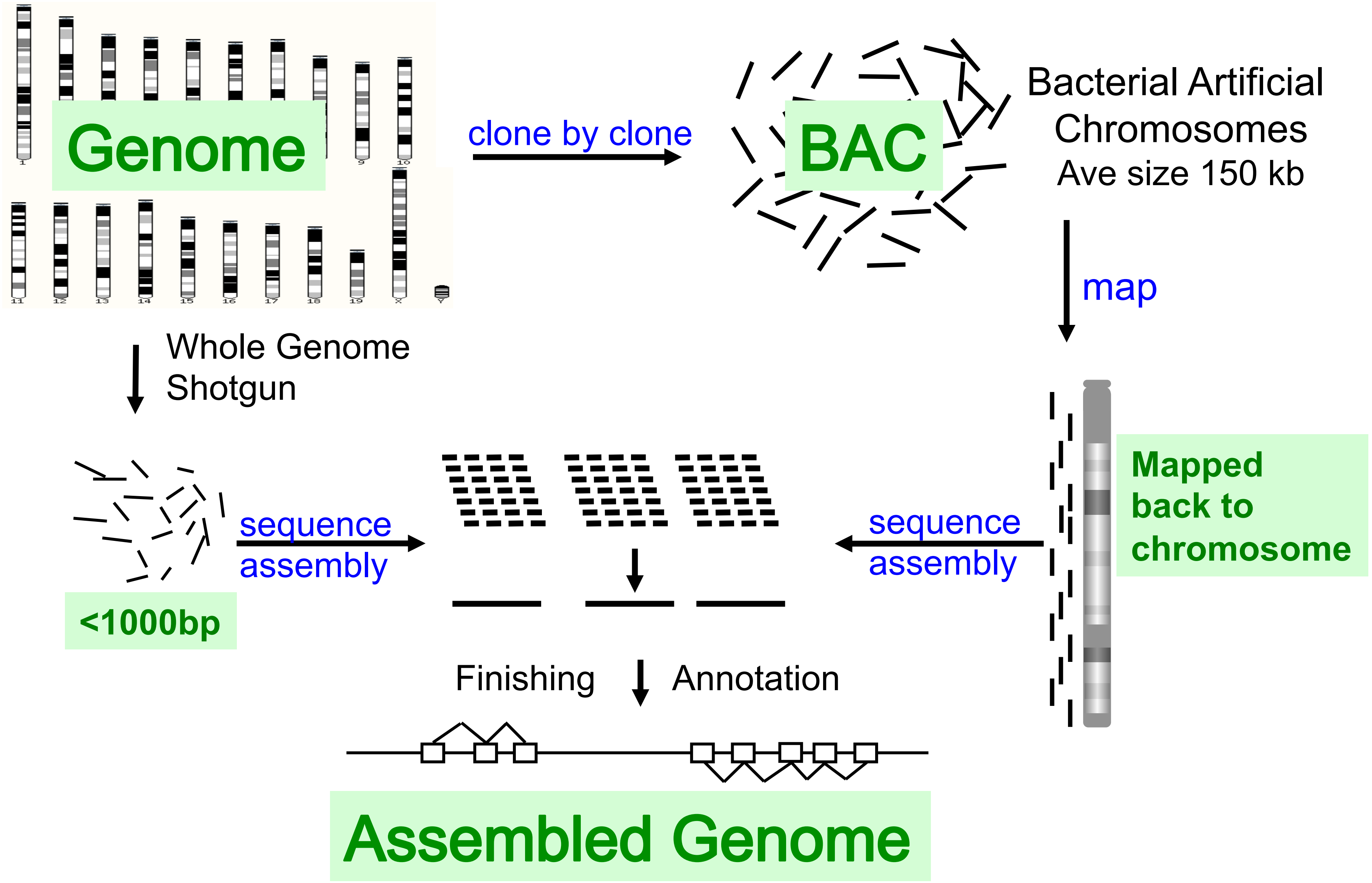
http://chslearns.vail.k12.az.us/pluginfile.php/1738/course/section/616/central%20dogma.png

Central Dogma Of Molecular Biology


Transcription
Translation
...GATCCTAGATCGCGCTAGCTAGCGCGCTAG...
...CTAGGATCTAGCGCGATCGATCGCGCGATC...
Intron
Exon
CAPGACUAUCUCCCUCAACAACUAAAUAAAAAAAAA
Protein
Ribosome
Chromosome
DNA
Double
Stranded
RNA
Single
Stranded

http://www4.ncsu.edu/~bjreadin/

Hybrid Sequencing Strategy



Genome

clone by clone

BAC

Bacterial Artificial
Chromosomes
Ave size 150 kb

map

Whole Genome
Shotgun

Mapped
back to
chromosome

sequence
assembly

<1000bp

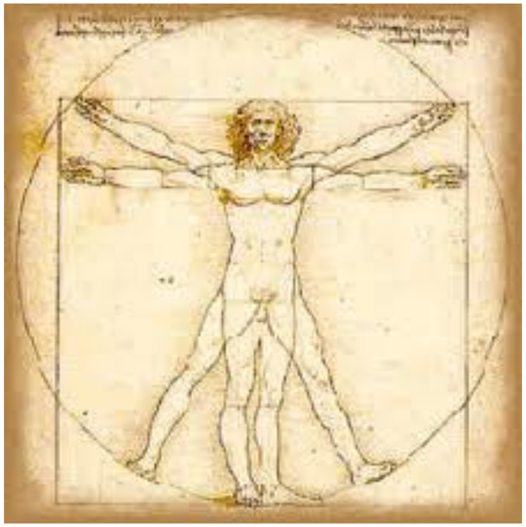
sequence
assembly

Finishing

Annotation

Assembled Genome

Reference genomes:



Human ~3Gb:
22 chromosomes + sex chromosomes

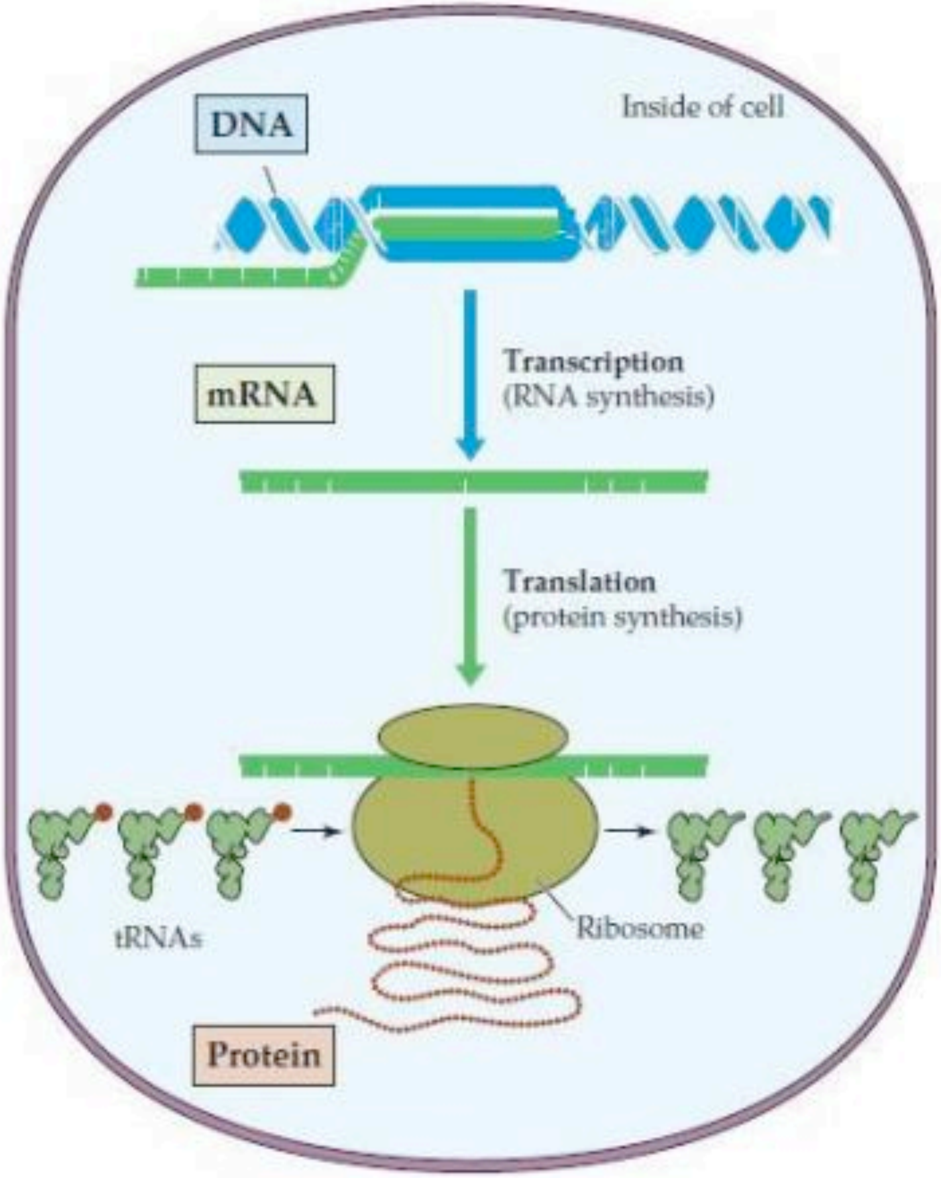
Mouse ~3 Gb:
19 chromosomes + sex chromosomes

SPL

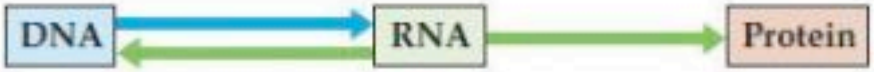
Zebrafish ~1.4 Gb:
25 chromosomes, no specific sex chromosomes

ENCODE

RNA polymerase



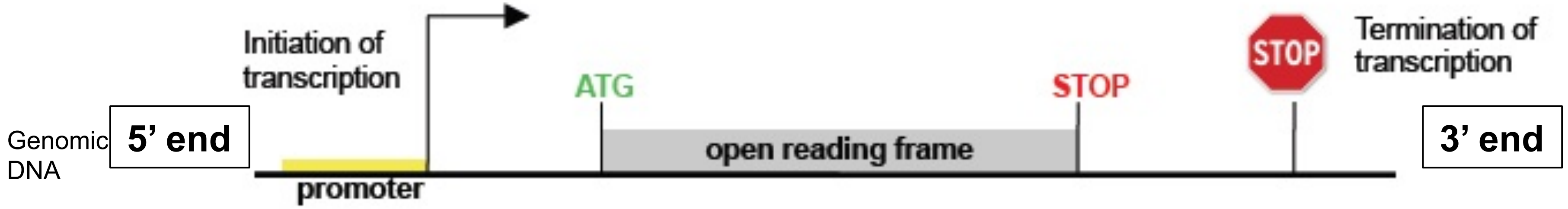
Ribosome binds to RNA generates amino acids



http://www.tutorvista.com/content/biology/biology-iii/gene-expression/molecular-central-dogma.php#

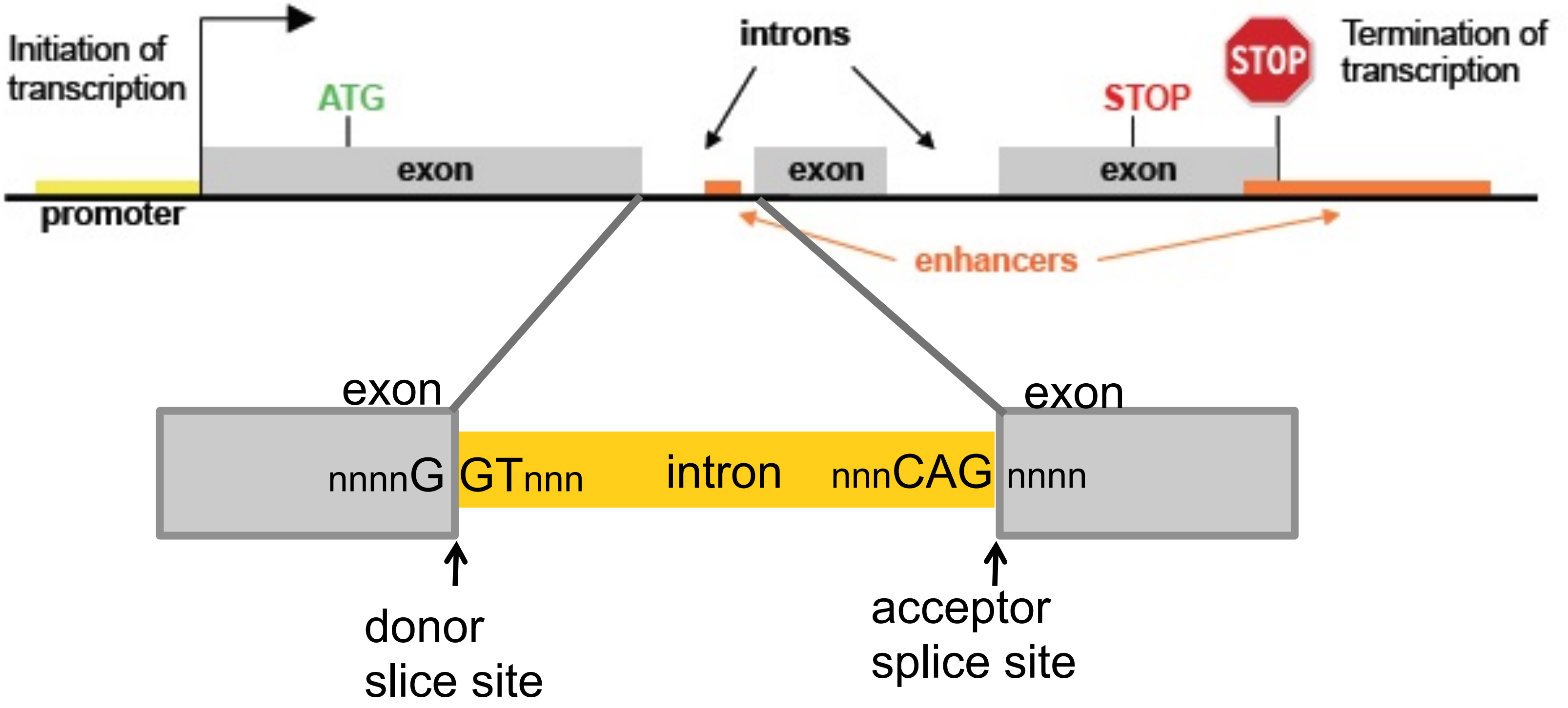
Prokaryotes:

Simple protein-coding gene



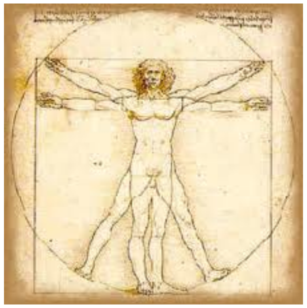
Eukaryotes:

More complex: Introns and Exons



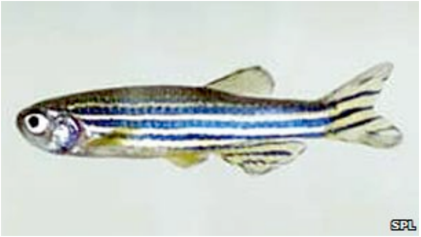
Do we know how many genes there are?

Protein coding genes



1980’s	100, 000
2000	40, 000
Today	~ 21, 000

21,638



SPL

25,120

Evidence for genes: DNA (*in vitro)*

5’ end

mRNA extracted from tissue sample

3’ end

mRNA AUGCCTG~~~~~~~~~~~~~~~~~~~~~~~~~~~~AATAAA~~~~~~~~AAAAAAAA

Reverse transcriptase transcribes the mRNA into a complimentary strand of DNA (cDNA)

cDNA

TACGGAC~~~~~~~~~~~~~~~~~~~~~~~~~~~~TTATTT~~~~~~~~TTTTTTTTT

ATGCCTG~~~~~~~~~~~~~~~~~~~~~~~~~~~~AATAAA~~~~~~~AAAAAAAA

mRNA degraded and a second strand of cDNA made

EST

ATGCCTG~~~~~~~~~~~~~

~TTATTT~~~~~~~~TTTTTTTTT

5‘ –> 3’

Expressed Sequence Tags (ESTs)
shorter stretches of cDNA (`800bp) sequences
from both ends

3‘ <– 5’

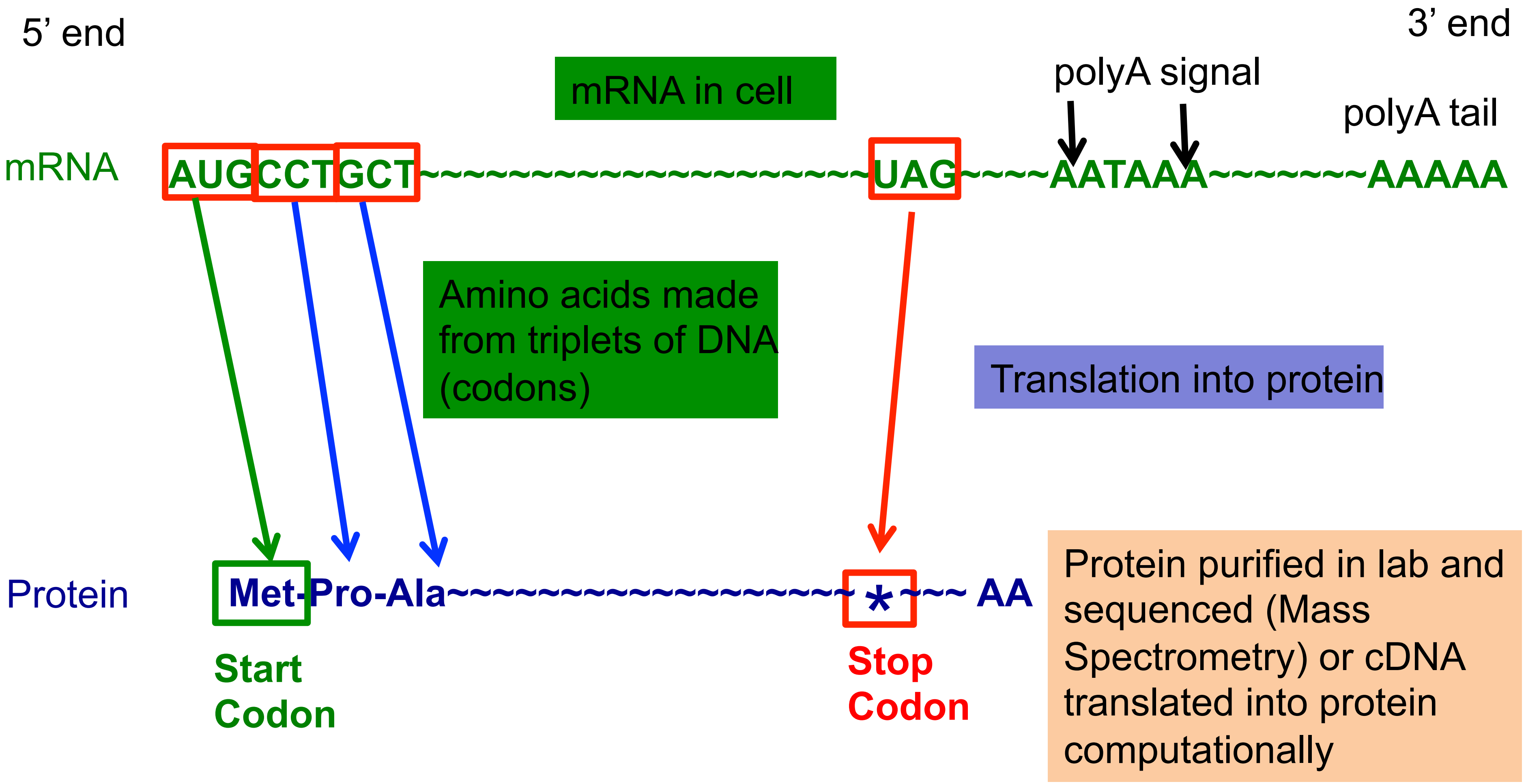
Sequences submitted to databases

ENA European Nucleotide Archive

NCBI GenBank

DDBJ DNA Data Bank of Japan

Evidence for genes: Protein



5’ end

3’ end

mRNA in cell

polyA signal

polyA tail

mRNA AUGCCTGCT~~~~~~~~~~~~~~~~~~~~~~UAG~~~~AATAAA~~~~~~~~AAAAA

Amino acids made from triplets of DNA (codons)

Translation into protein

Protein Met-Pro-Ala~~~~~~~~~~~~~~~~~~~~ * ~~~~ AA

Start Codon

Stop Codon

Protein purified in lab and sequenced (Mass Spectrometry) or cDNA translated into protein computationally

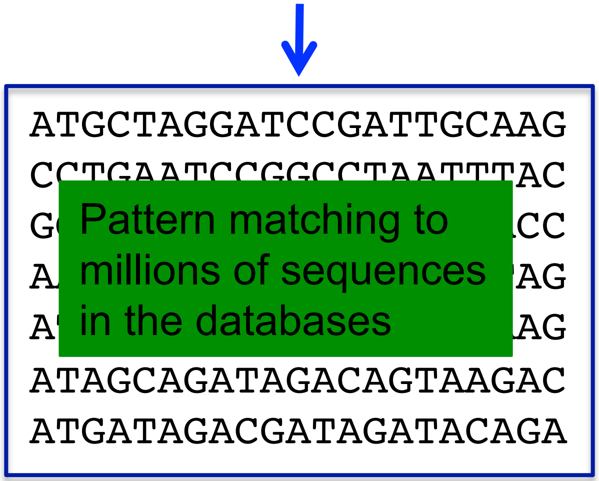
Sequences submitted to database

UniProt

Searching the databases: How to find gene location

BLAST

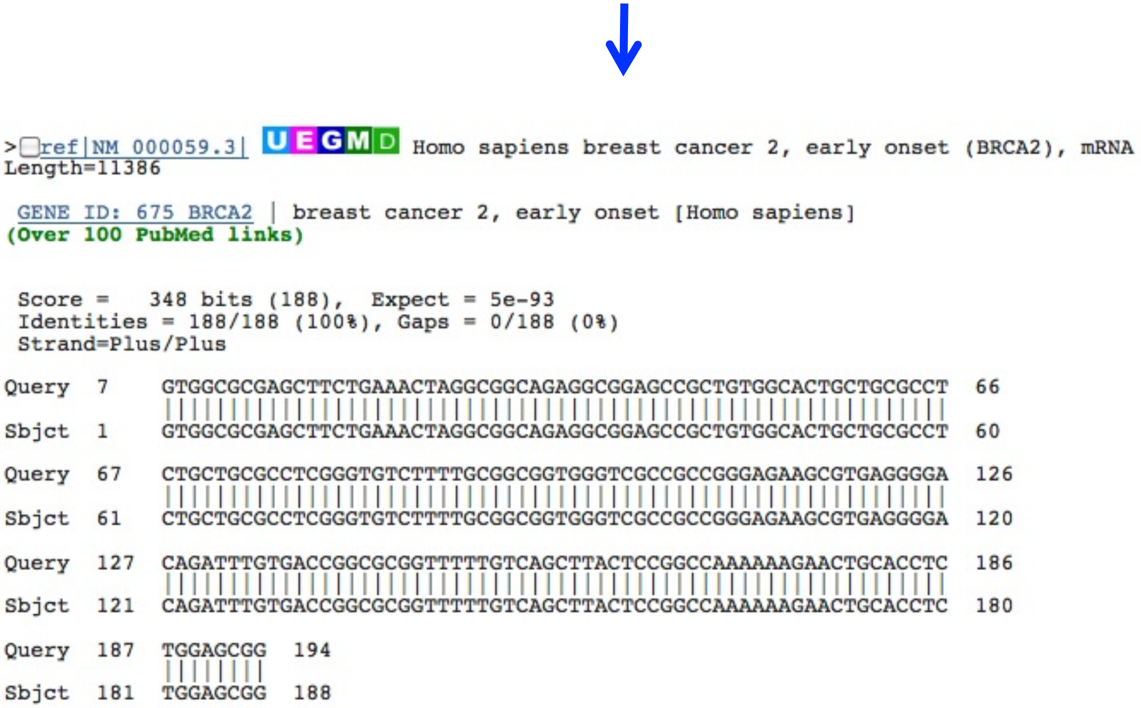
(Basic Local Alignment Search Tool)



ATGCTAGGATCCGATTGCAAG
CCTGAATCCGGCCTAATTTAC
G CC
A AG
A AG
ATAGCAGATAGACAGTAAGAC
ATGATAGACGATAGATACAGA

Pattern matching to millions of sequences in the databases

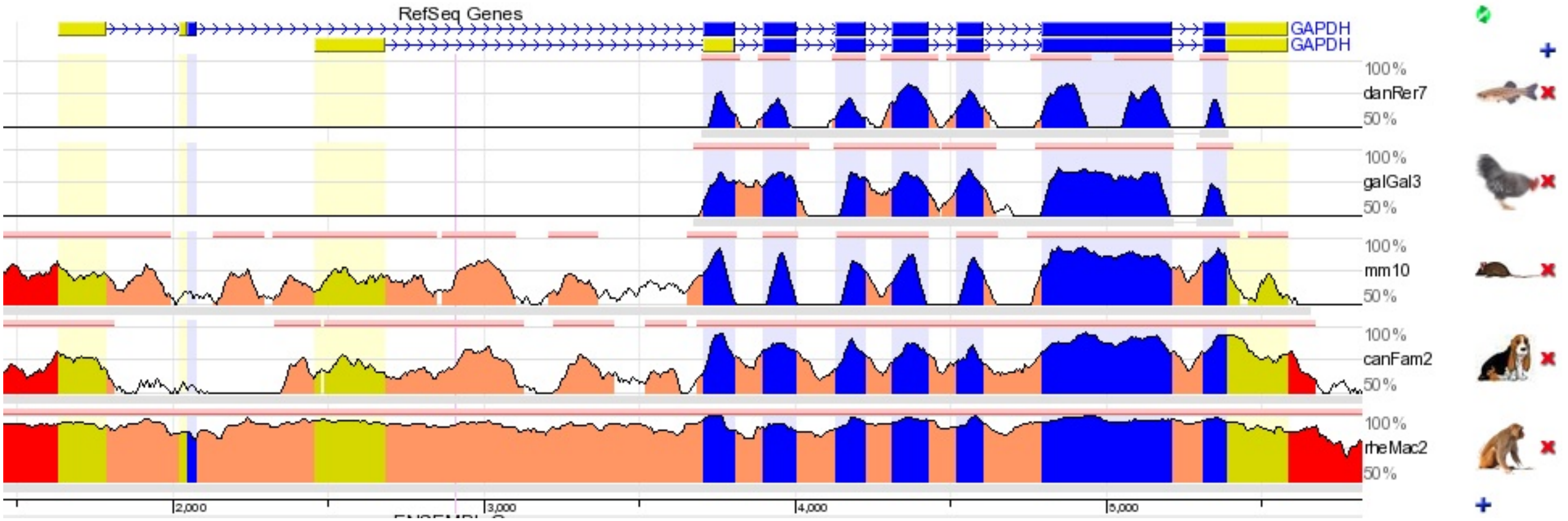
Genomic DNA as query against the databases



```
>□ref|NM 000059.3| UEGMD Homo sapiens breast cancer 2, early onset (BRCA2), mRNA
Length=11386

 GENE ID: 675 BRCA2 | breast cancer 2, early onset [Homo sapiens]
(Over 100 PubMed links)

 Score =   348 bits (188),  Expect = 5e-93
 Identities = 188/188 (100%), Gaps = 0/188 (0%)
 Strand=Plus/Plus

Query  7    GTGGCGCGAGCTTCTGAAACTAGGCGGCAGAGGCGGAGCCGCTGTGGCACTGCTGCGCCT  66
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    GTGGCGCGAGCTTCTGAAACTAGGCGGCAGAGGCGGAGCCGCTGTGGCACTGCTGCGCCT  60

Query  67   CTGCTGCGCCTCGGGTGTCTTTTGCGGCGGTGGGTCGCCGCCGGGAGAAGCGTGAGGGGA  126
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   CTGCTGCGCCTCGGGTGTCTTTTGCGGCGGTGGGTCGCCGCCGGGAGAAGCGTGAGGGGA  120

Query  127  CAGATTTGTGACCGGCGCGGTTTTTGTCAGCTTACTCCGGCCAAAAAAGAACTGCACCTC  186
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  CAGATTTGTGACCGGCGCGGTTTTTGTCAGCTTACTCCGGCCAAAAAAGAACTGCACCTC  180

Query  187  TGGAGCGG  194
            ||||||||
Sbjct  181  TGGAGCGG  188
```

BLAST results and alignment

Conservation:
GAPDH gene coding exons



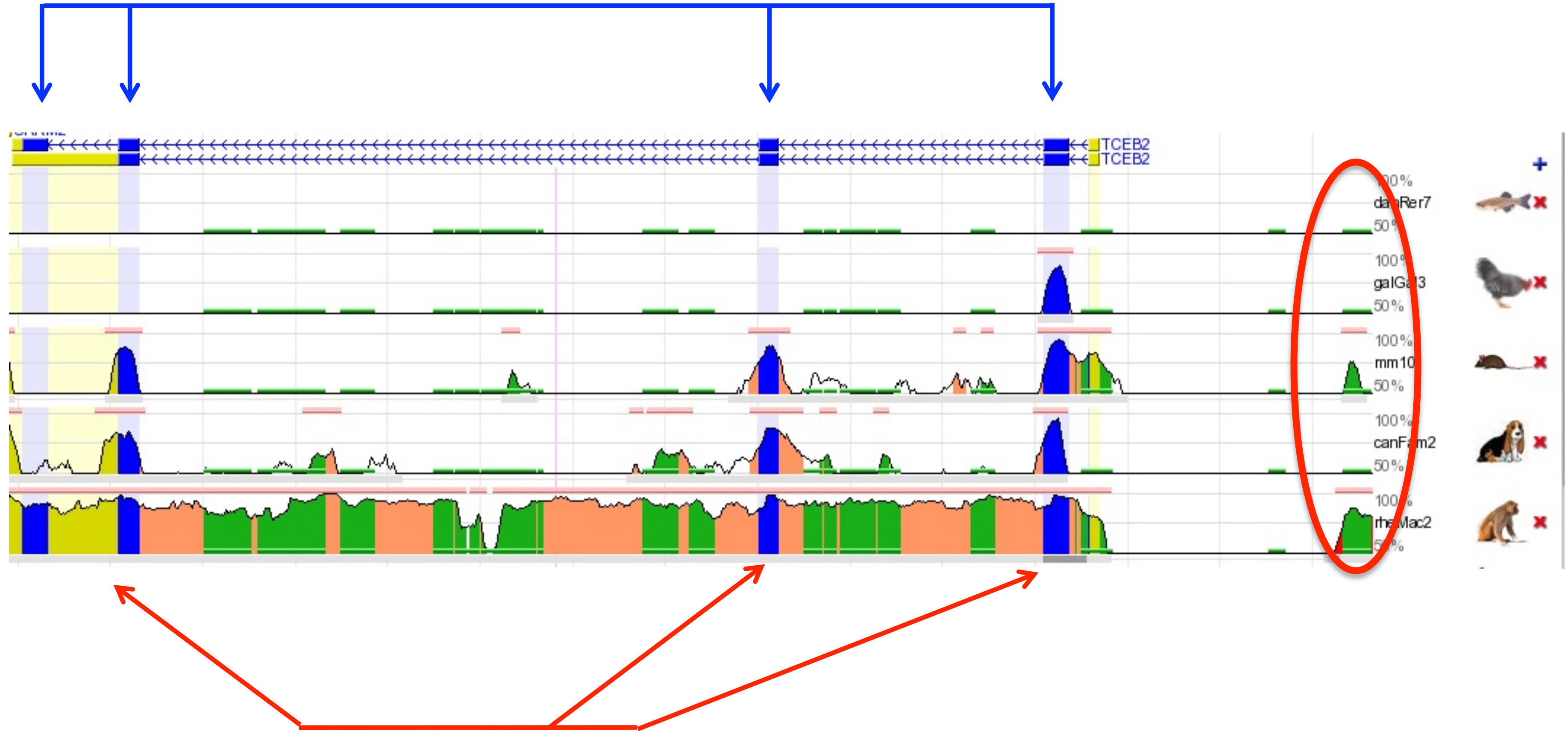
Conservation:

TCEB2 gene

3' end

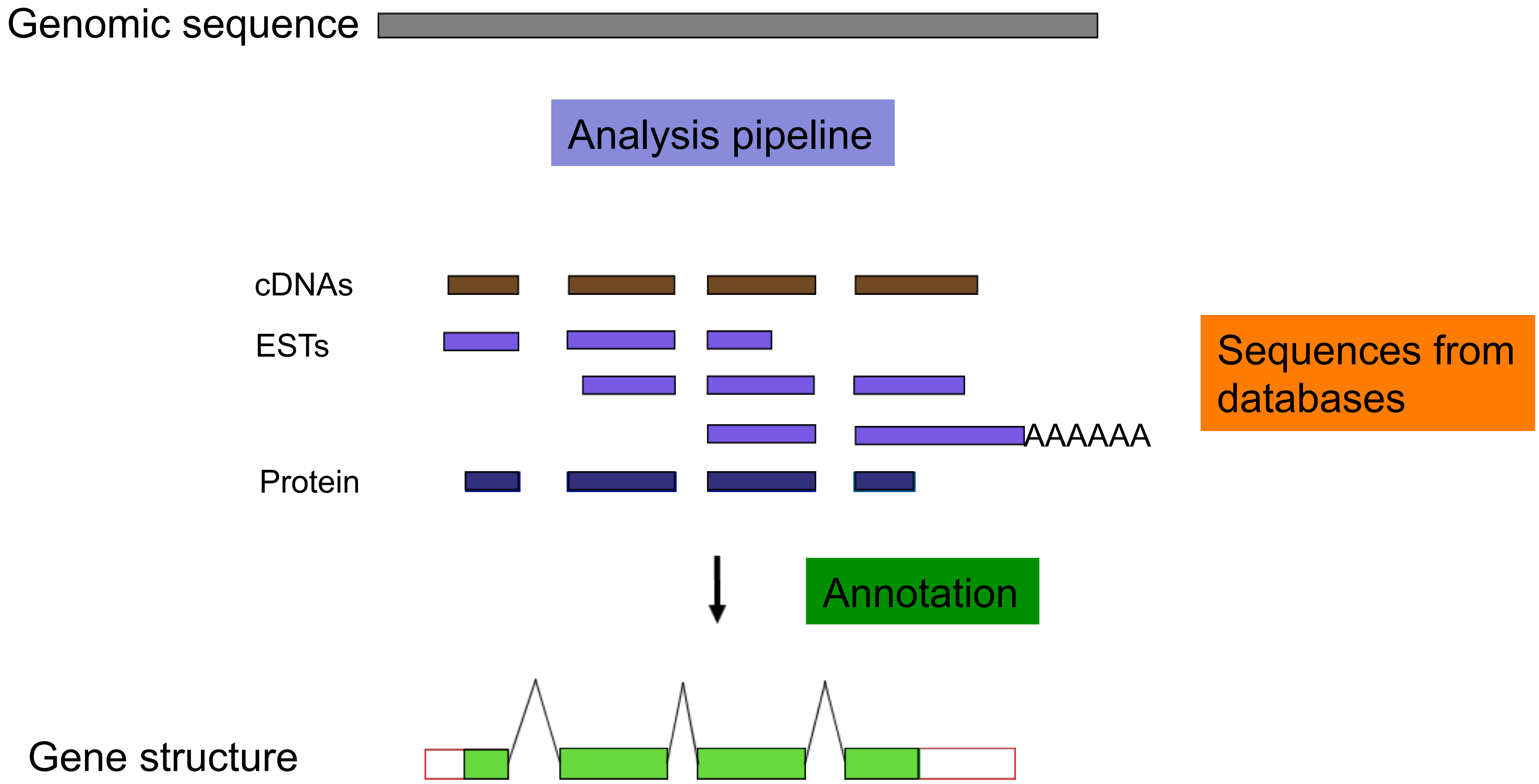
Coding exons

5' end



Conserved regions

Making the transcript from evidence:



Genomic sequence

Analysis pipeline

cDNAs

ESTs

AAAAAA

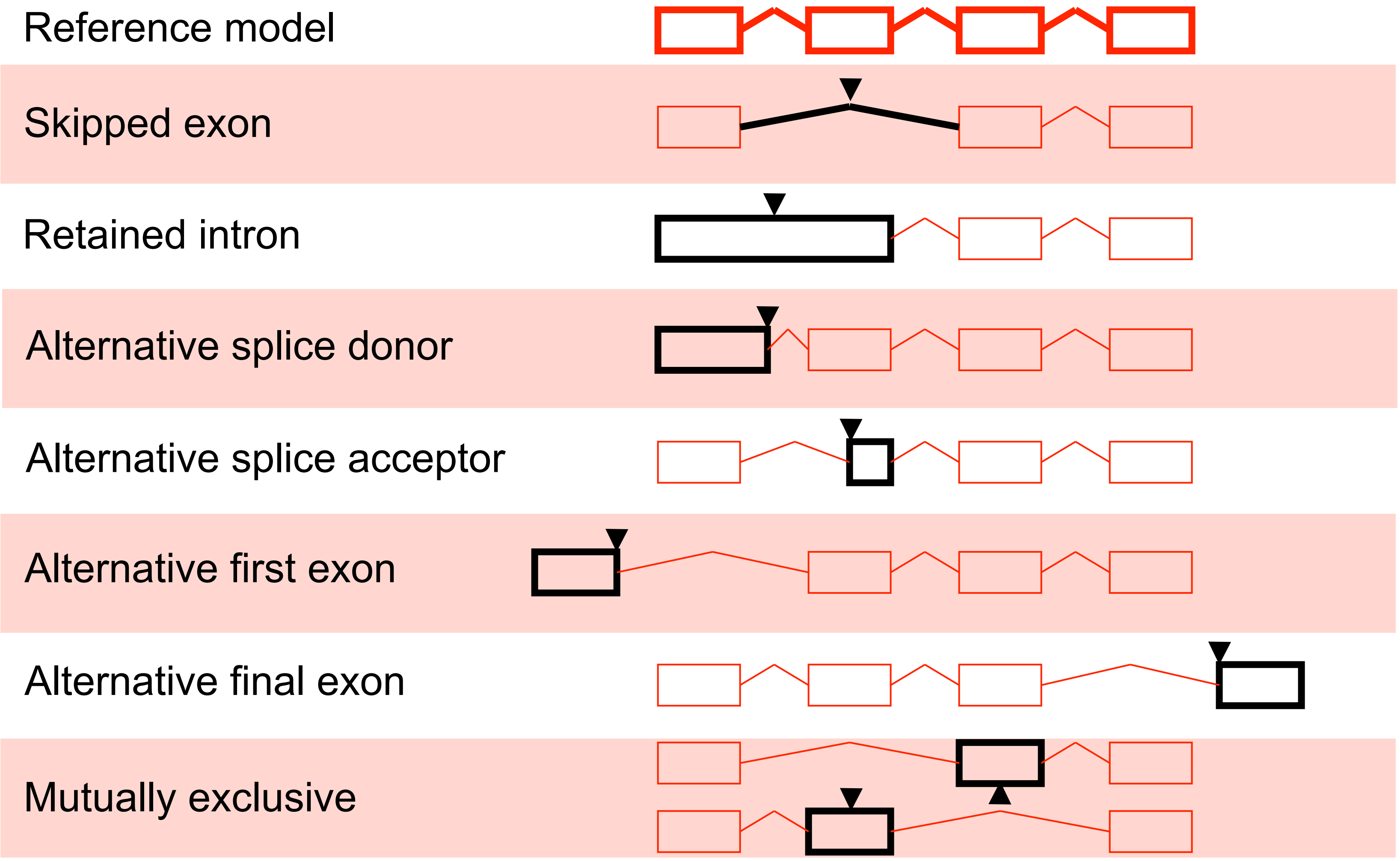
Protein

Sequences from databases

Annotation

Gene structure

Alternative Splicing



Reference model

Skipped exon

Retained intron

Alternative splice donor

Alternative splice acceptor

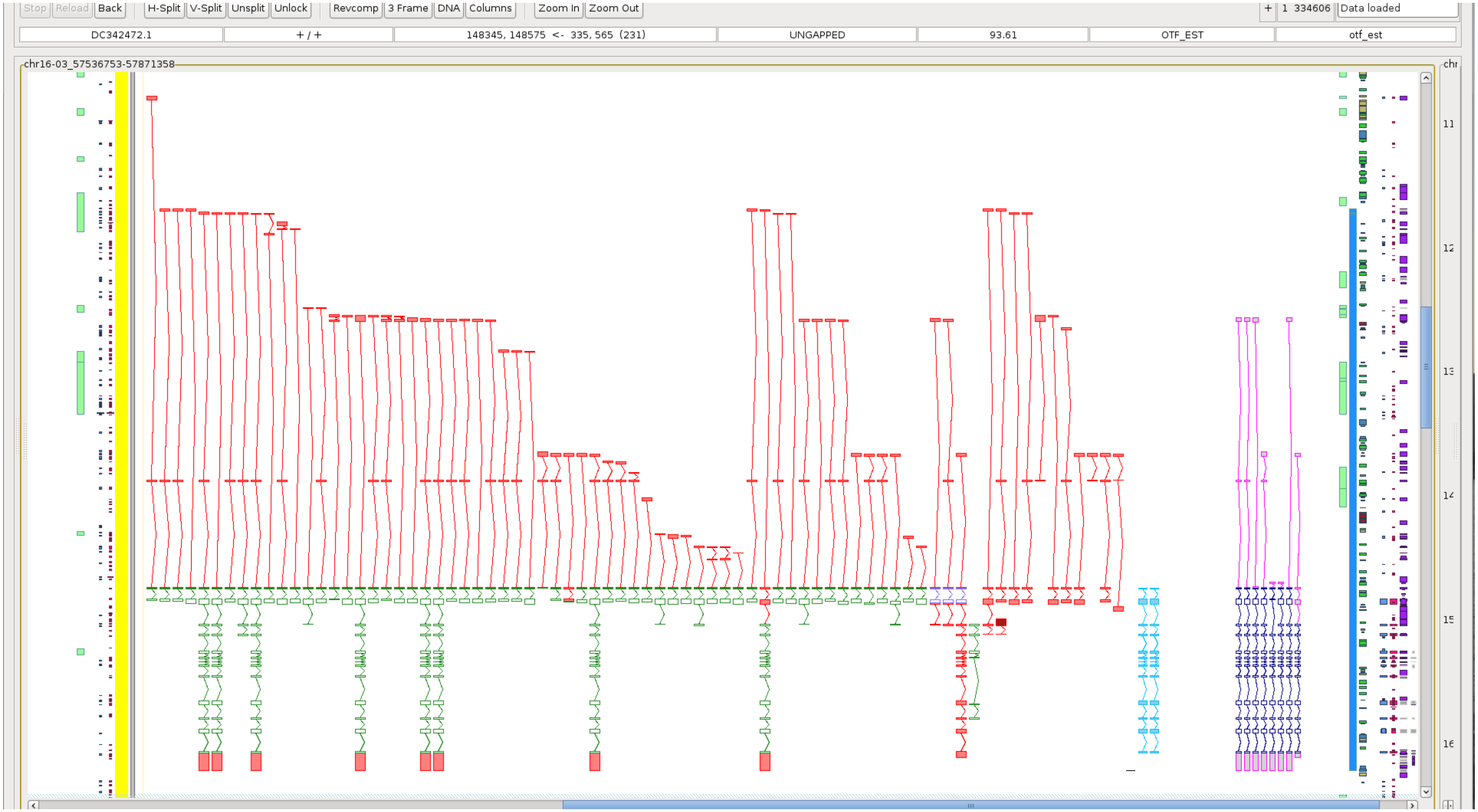
Alternative first exon

Alternative final exon

Mutually exclusive

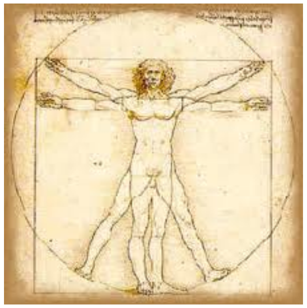
GPR56:

Human G protein-coupled receptor 56 gene



DC342472.1 | + / + | 148345, 148575 <- 335, 565 (231) | UNGAPPED | 93.61 | OTF_EST | otf_est

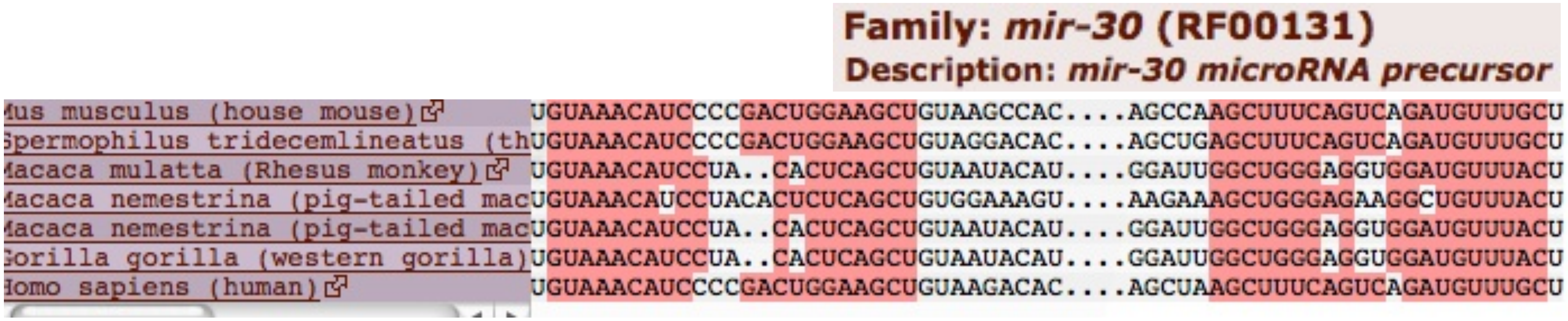
Do we know how many genes there are?

	Coding	Non-coding	Pseudogene
	21,224	15,962	14,427
	21,638	6,875	5,510
	25,120	4,556	224

Non-coding genes:

microRNAs:

Under 200 residues
Highly conserved signature
Imported from Rfam database

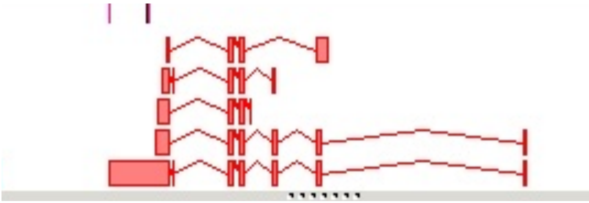


Family: ***mir-30*** **(RF00131)**
Description: ***mir-30 microRNA precursor***

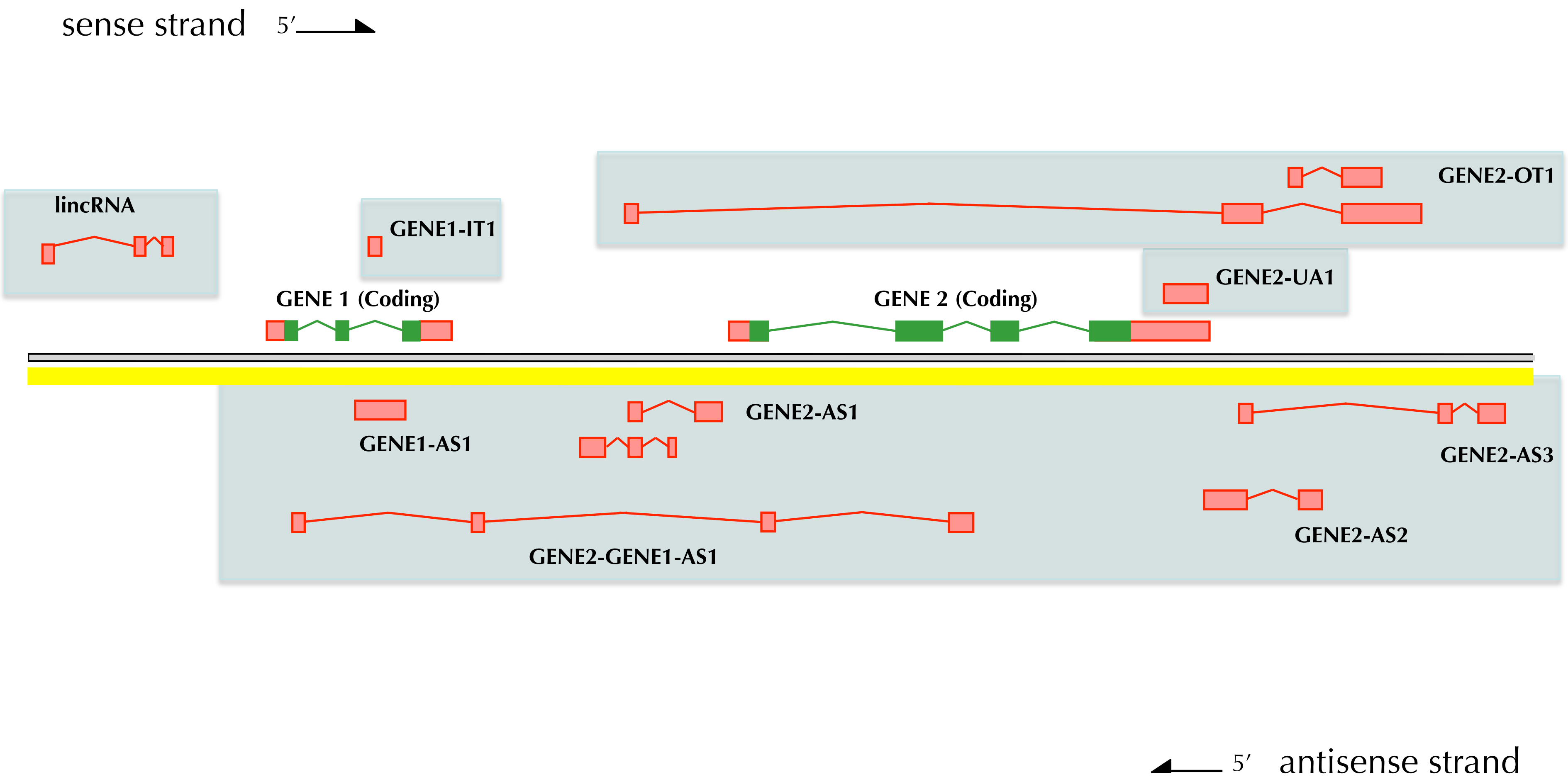
```
Mus musculus (house mouse)                 UGUAAACAUCCCCGACUGGAAGCUGUAAGCCAC....AGCCAAGCUUUCAGUCAGAUGUUUGCU
Spermophilus tridecemlineatus (th          UGUAAACAUCCCCGACUGGAAGCUGUAGGACAC....AGCUGAGCUUUCAGUCAGAUGUUUGCU
Macaca mulatta (Rhesus monkey)             UGUAAACAUCCUA..CACUCAGCUGUAAUACAU....GGAUUGGCUGGGAGGUGGAUGUUUACU
Macaca nemestrina (pig-tailed mac          UGUAAACAUCCUACACUCUCAGCUGUGGAAAGU....AAGAAAGCUGGGAGAAGGCUGUUUACU
Macaca nemestrina (pig-tailed mac          UGUAAACAUCCUA..CACUCAGCUGUAAUACAU....GGAUUGGCUGGGAGGUGGAUGUUUACU
Gorilla gorilla (western gorilla)          UGUAAACAUCCUA..CACUCAGCUGUAAUACAU....GGAUUGGCUGGGAGGUGGAUGUUUACU
Homo sapiens (human)                       UGUAAACAUCCCCGACUGGAAGCUGUAAGACAC....AGCUAAGCUUUCAGUCAGAUGUUUGCU
```

Long non-coding RNAs:
(lncRNAs)

Over ~ 200 residues
Not highly conserved between species
Manually annotated
Some very well-known e.g. Hotair (HOX antisense intergenic RNA)
Many others not yet characterised



lncRNAs

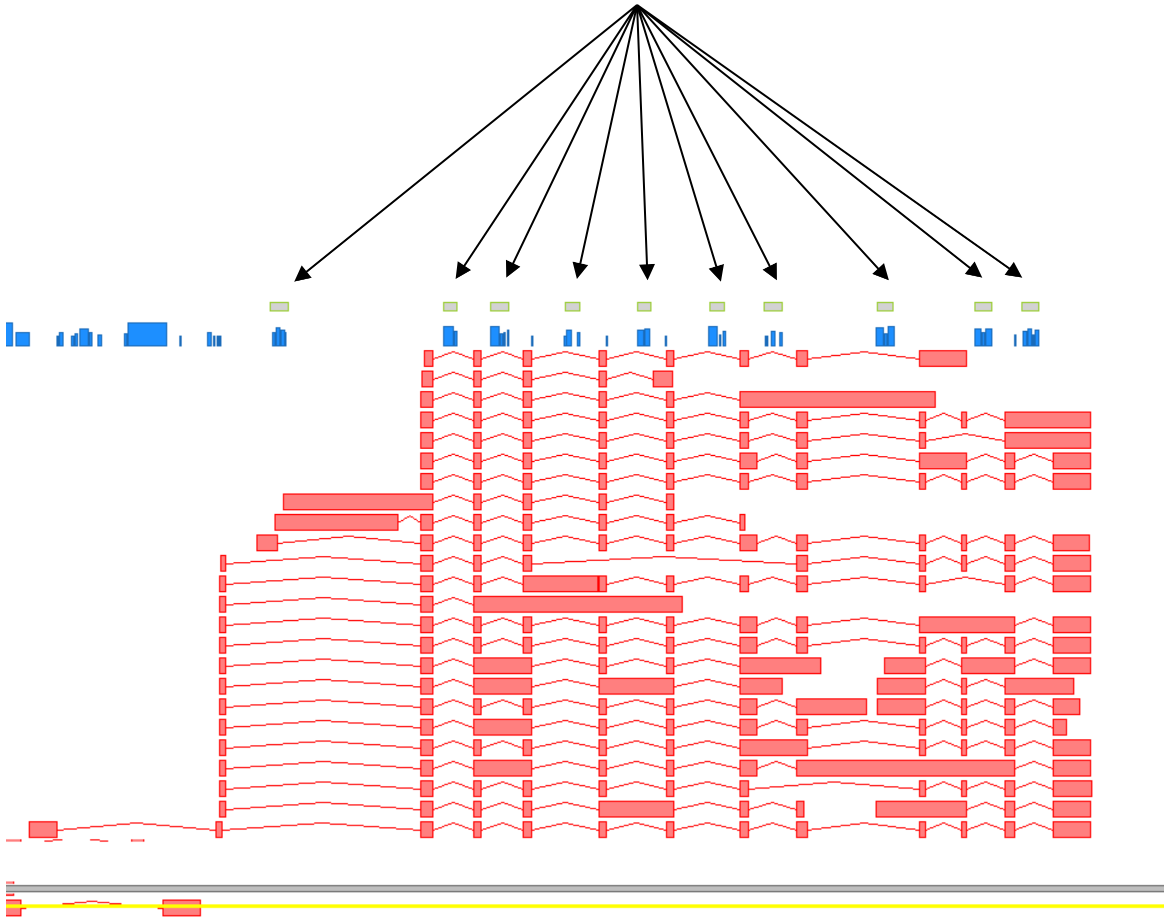


sense strand 5′

5′ antisense strand

GAS5: growth arrest-specific 5 (non-protein coding)

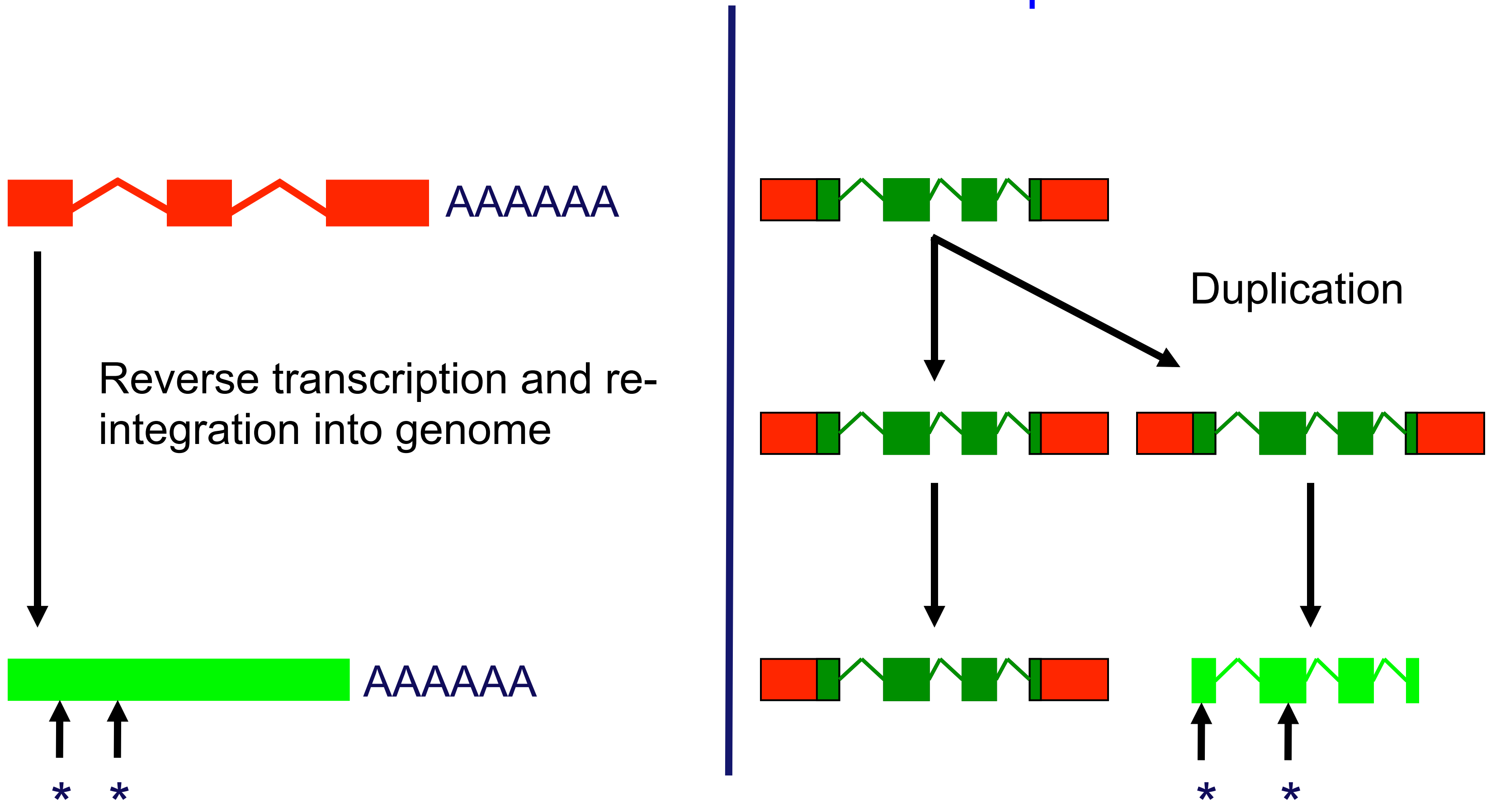
intronic snoRNAs



Havana pseudogenes

Processed

Unprocessed



AAAAAA

Reverse transcription and re-integration into genome

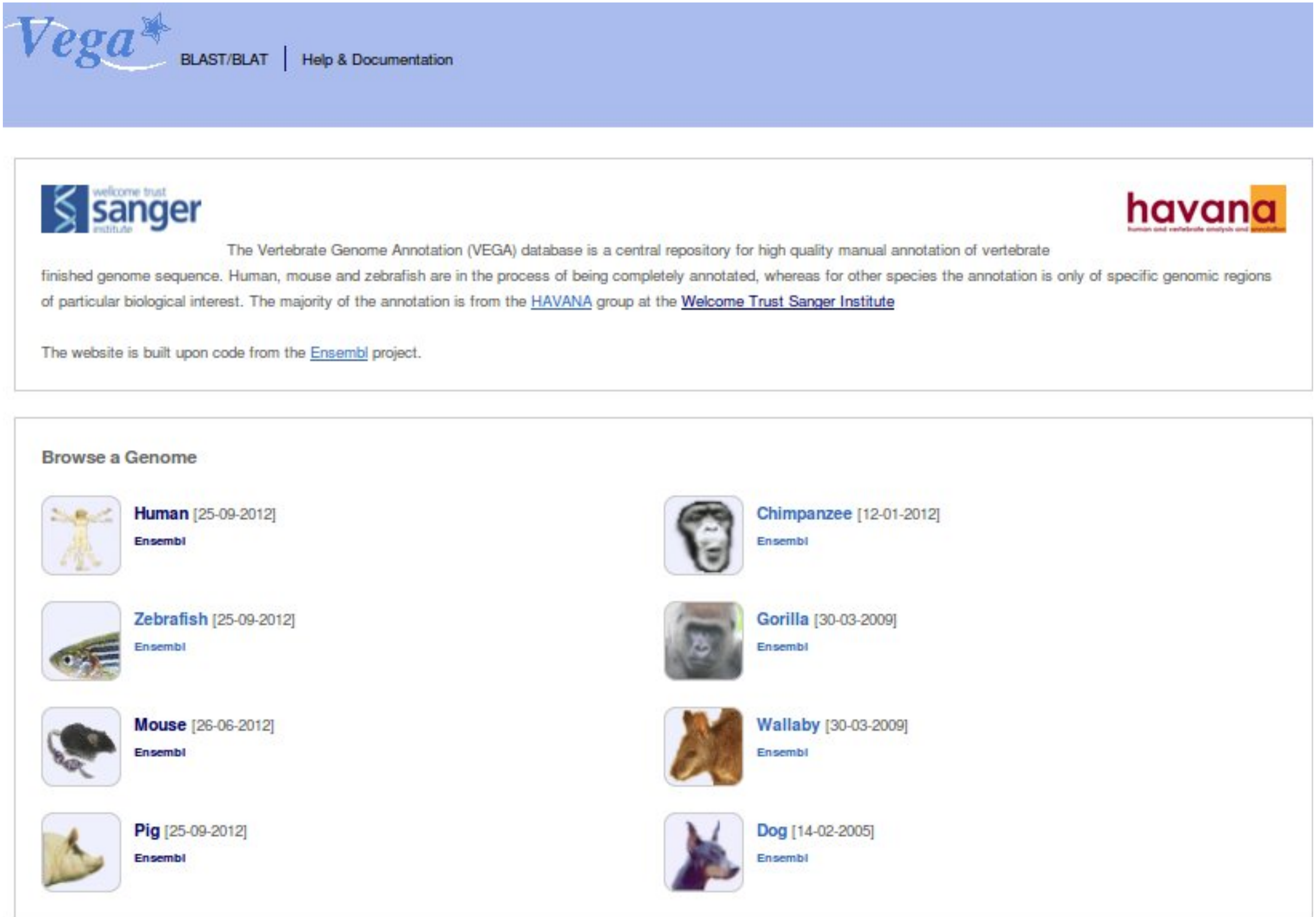
AAAAAA

* *

Duplication

* *

Disruption to coding sequence and stop codons



Vega
BLAST/BLAT
Help & Documentation
wellcome trust sanger institute
havana
human and vertebrate analysis and annotation

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.

Browse a Genome

Human [25-09-2012]
Ensembl

Chimpanzee [12-01-2012]
Ensembl

Zebrafish [25-09-2012]
Ensembl

Gorilla [30-03-2009]
Ensembl

Mouse [26-06-2012]
Ensembl

Wallaby [30-03-2009]
Ensembl

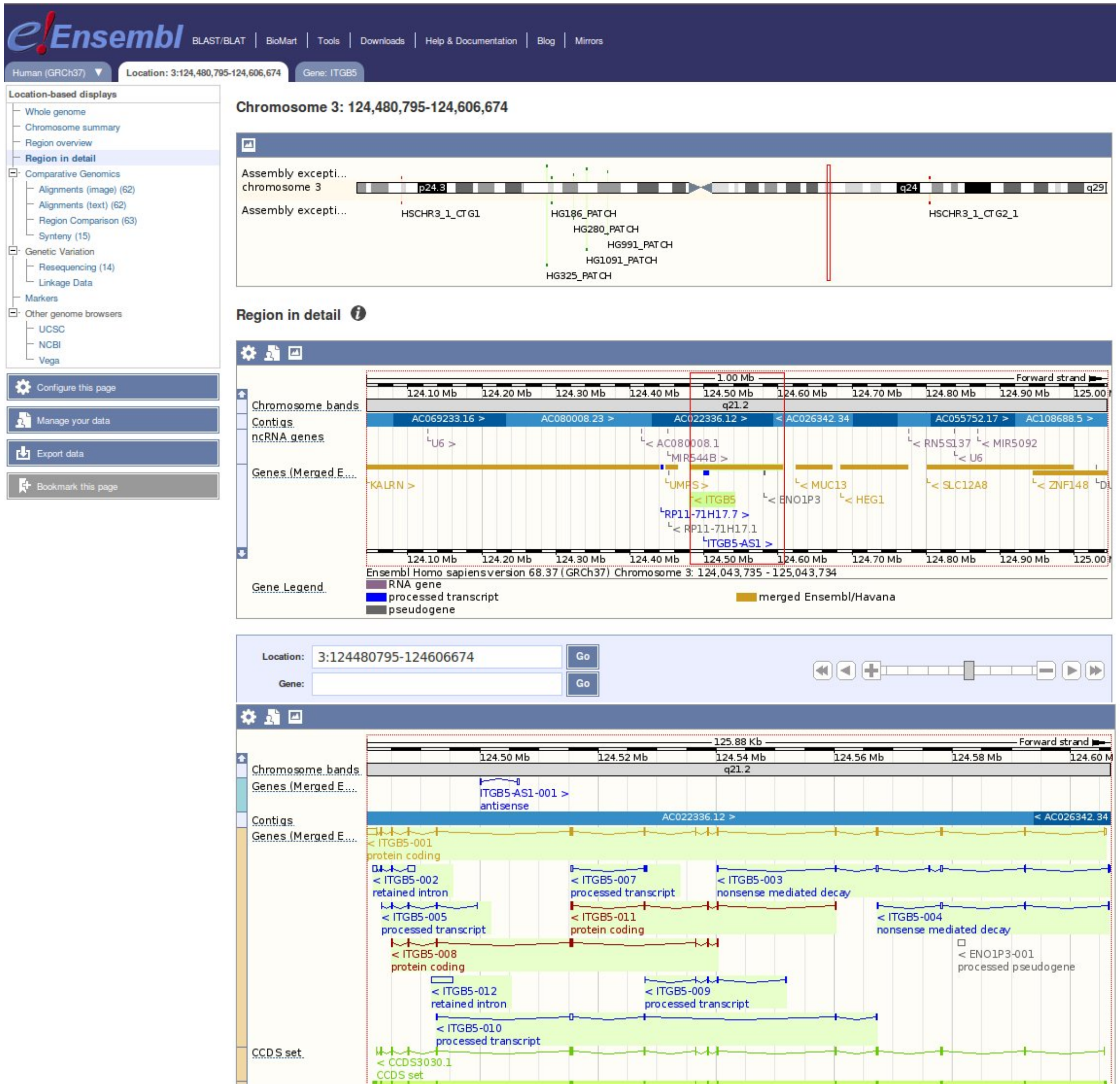
Pig [25-09-2012]
Ensembl

Dog [14-02-2005]
Ensembl

Genome browsers include Vega genes:

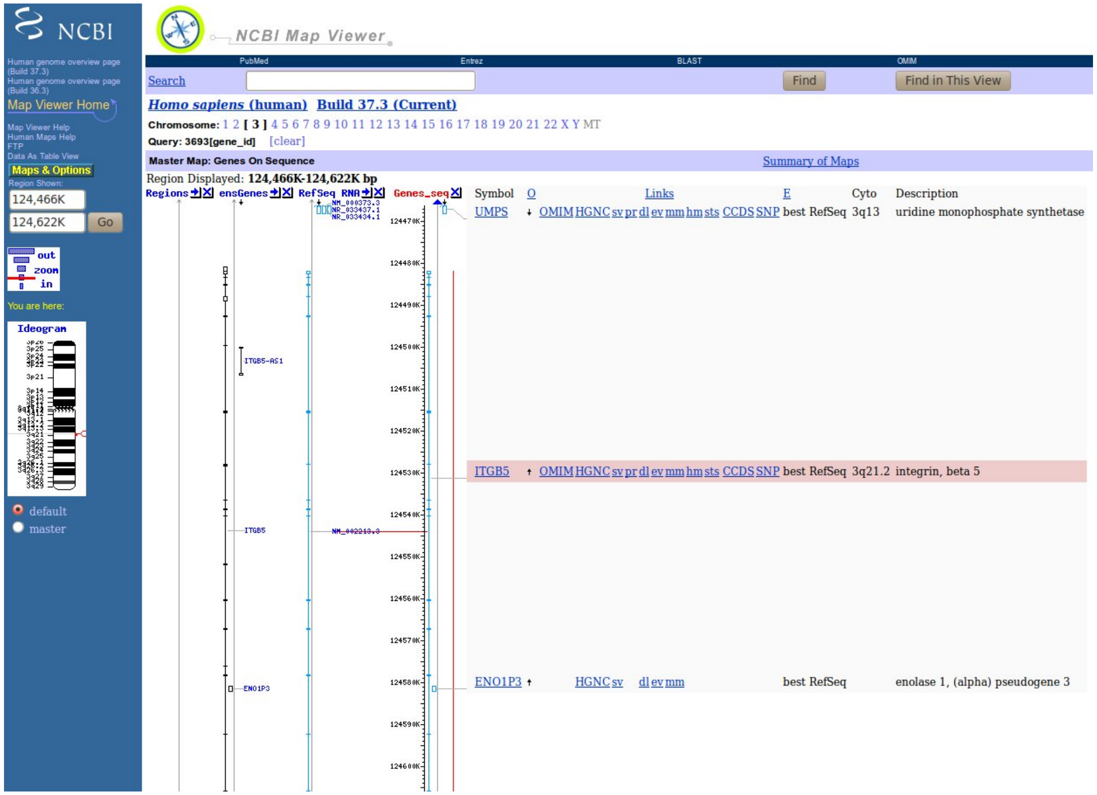
GENCODE geneset:

ENCODE



Genome browsers include Vega genes:

GENCODE geneset:



***Homo sapiens* (human)** **Build 37.3 (Current)**

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: 3693[gene_id] [clear]

Master Map: Genes On Sequence Summary of Maps

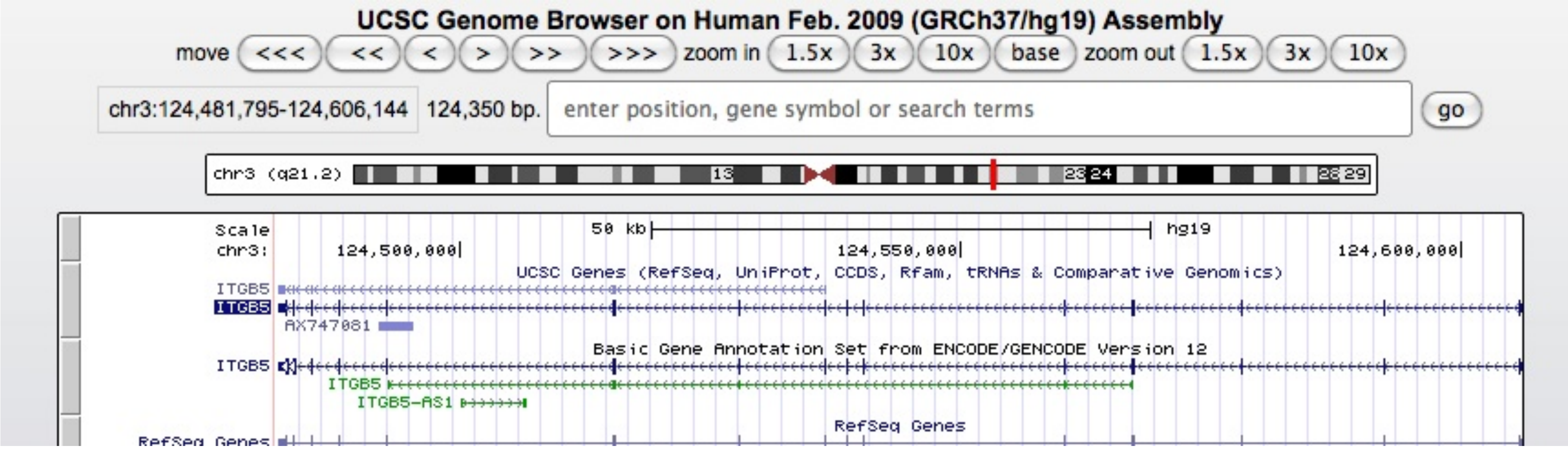
Region Displayed: **124,466K-124,622K bp**

Symbol	O	Links	E	Cyto	Description
UMPS	↓	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	3q13	uridine monophosphate synthetase
ITGB5	↑	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	3q21.2	integrin, beta 5
ENO1P3	↑	HGNC sv dl ev mm	best RefSeq		enolase 1, (alpha) pseudogene 3

Genome browsers include Vega genes:

GENCODE geneset:

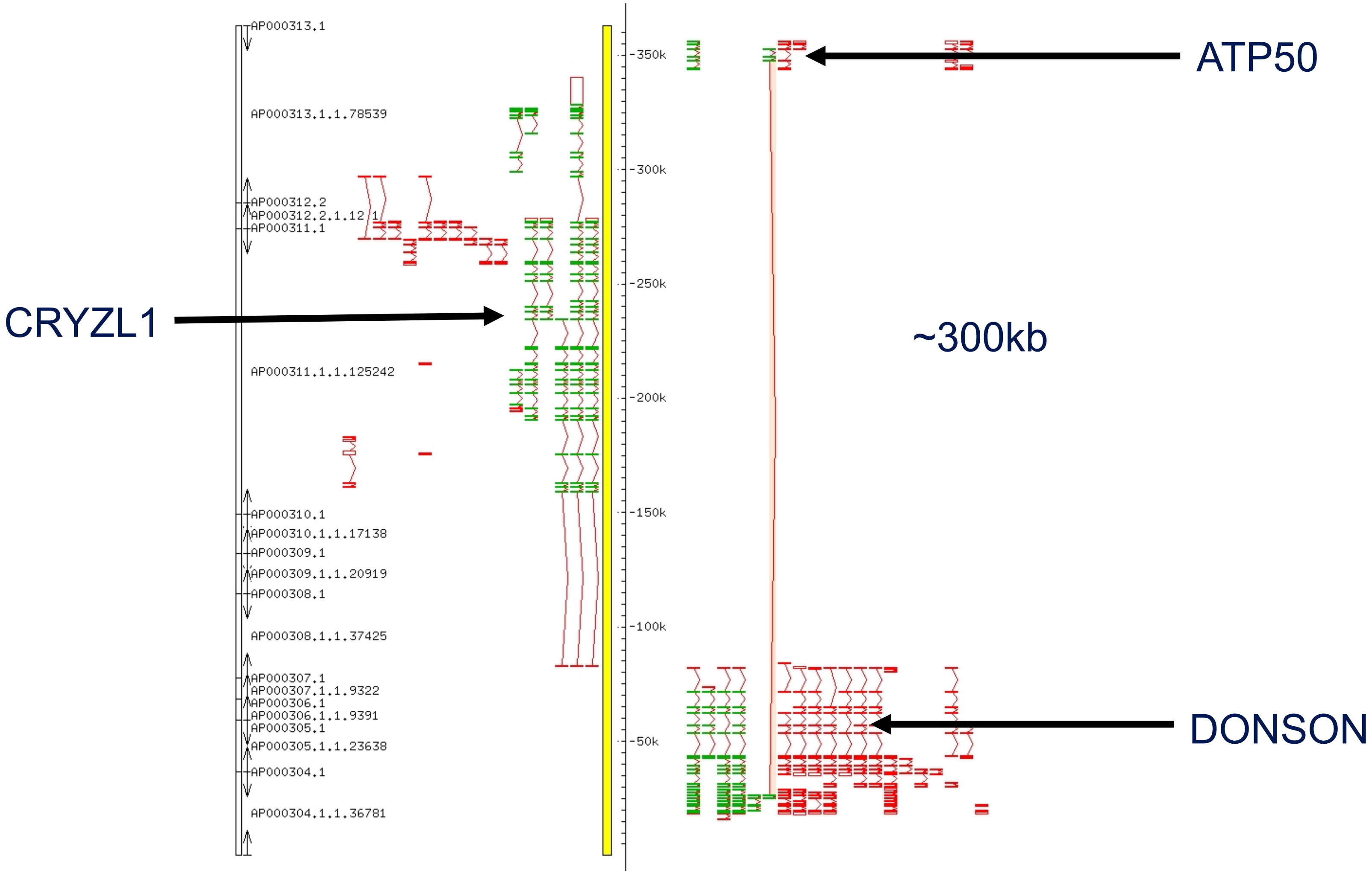
ENCODE



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr3:124,481,795-124,606,144 124,350 bp.

Linked loci



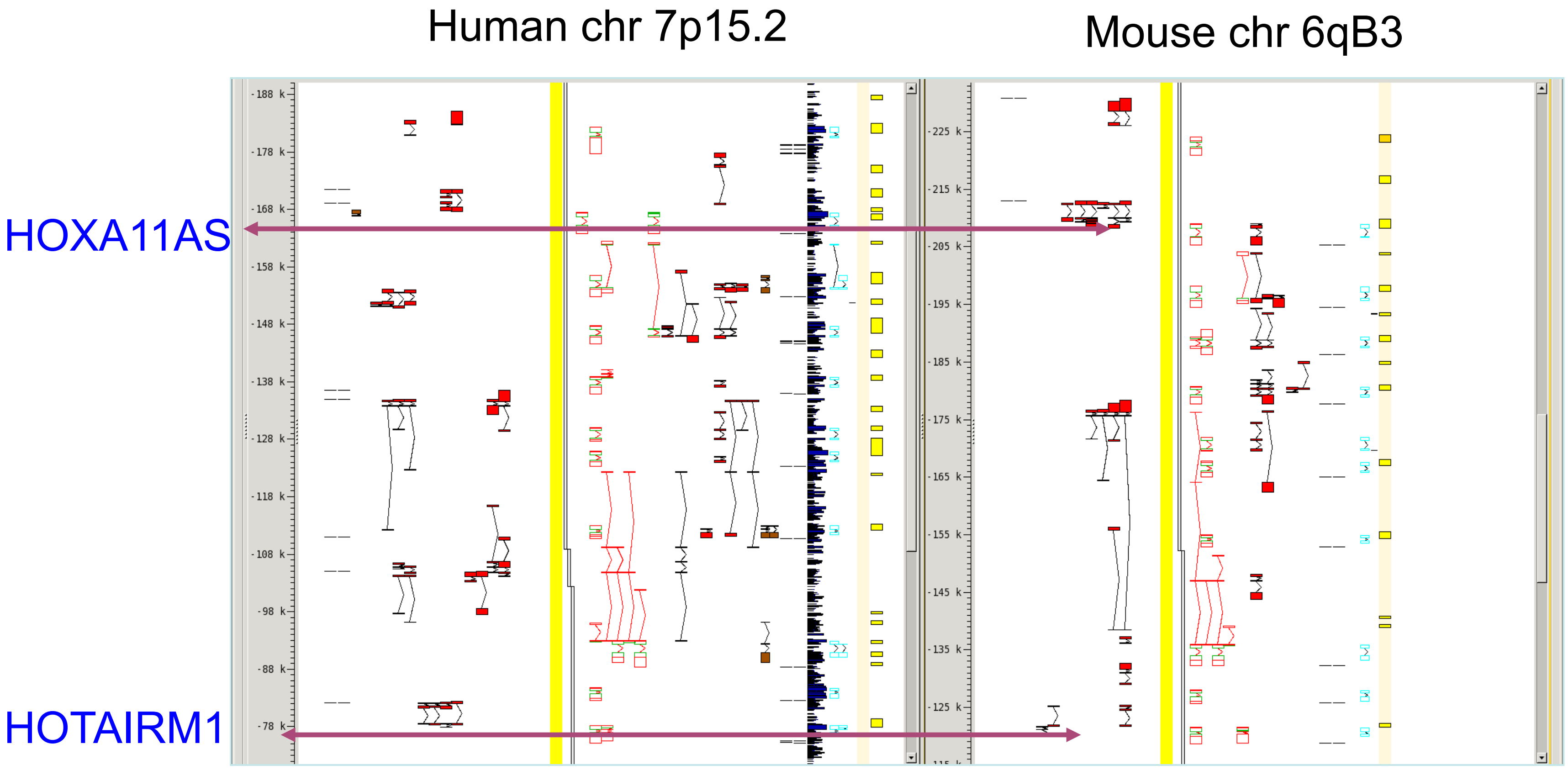
ATP50

CRYZL1

~300kb

DONSON

HOXA gene cluster



Long non-coding transcripts are conserved across species and regulate expression of HOX genes

RIVIERA

Acknowledgements

Havana:
Jen Harrow
If Barnes
Ruth Bennett
Alex Bignell
Veronika Boychenko
Gloria Despacio-Reyes
Sarah Donaldson
Adam Frankish
Matt Hardy
Toby Hunt
Mike Kay
Gavin Laird
David Lloyd
Jane Loveland
Deepa Manthravadi
Gaurab Mukherjee
Jonathan Mudge
Jeena Rajan
Gary Saunders
Catherine Snow
Charles Steward
Marie-Marthe Suner
Mark Thomas
Laurens Wilming

Encode:
Jose Gonzalez
Sarah Grubb
Electra Tapanari

Anacode:
James Gilbert
Matthew Astley
Michael Gray
Jeremy Henty

Vega:
Stephen Trevanion
Dan Sheppard

Zmap:
Ed Griffiths
Gemma Barson
Malcolm Hinsley

Project Coordinator:
Tim Hubbard

http://vega.sanger.ac.uk