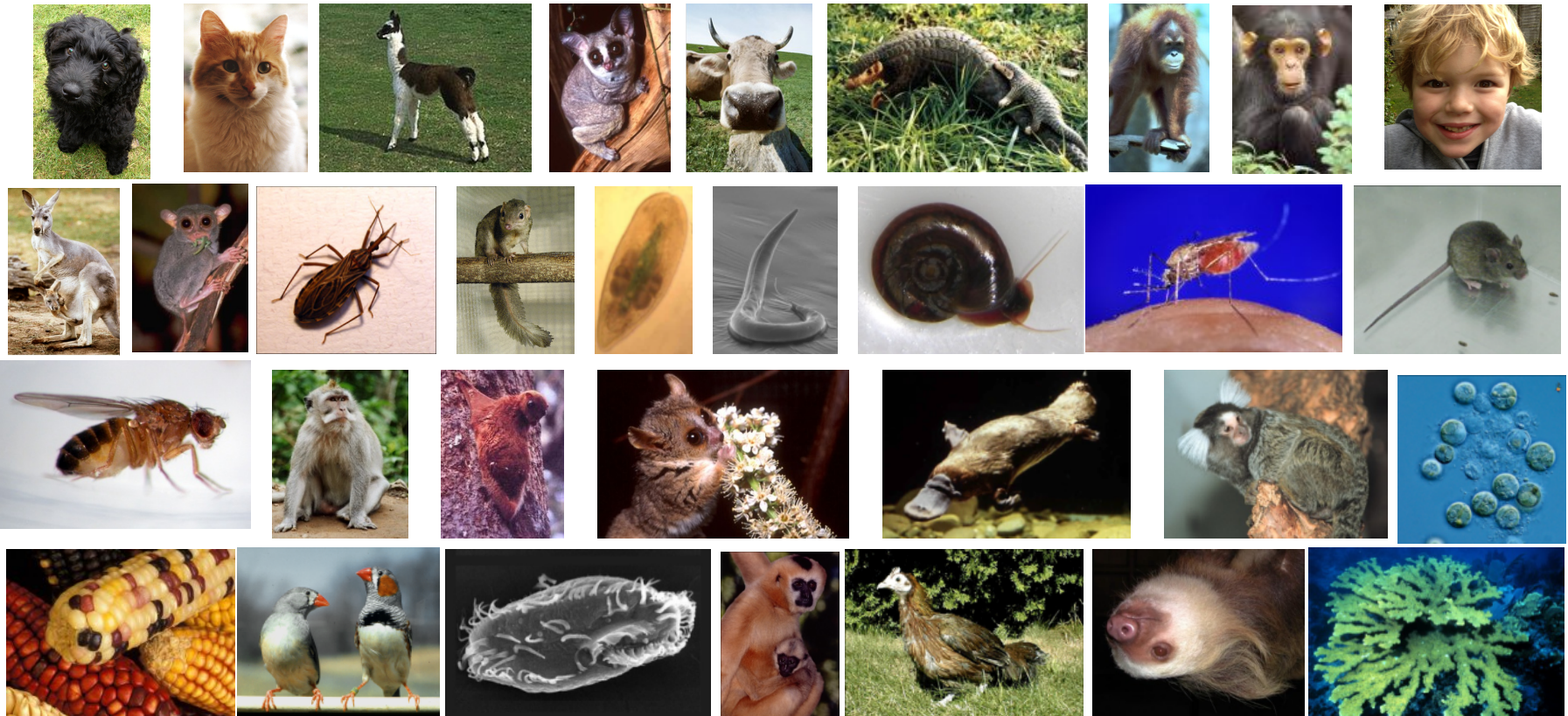


# Module 3

## Comparative Sequence Analysis



Jane Loveland  
Wellcome Trust Sanger Institute

## Overview:

# Introduction to comparative sequence analysis

Two worked examples:

1. The identification and analysis of homologous gene sequences
2. Using of orthologous genome sequences to identify evolutionarily conserved regions

# Comparative Sequence Analysis

Tool for decoding genomic information as it is based upon the tenet that:

Functional sequences evolve more slowly than non-functional sequences, therefore sequences that remain conserved throughout evolution *may* perform a biological function.

# Identify Conserved Regions

## Aligning genome sequences

- Functionally conserved units may be conserved at the sequence level
- Evolutionary Conserved Regions (ECRs)

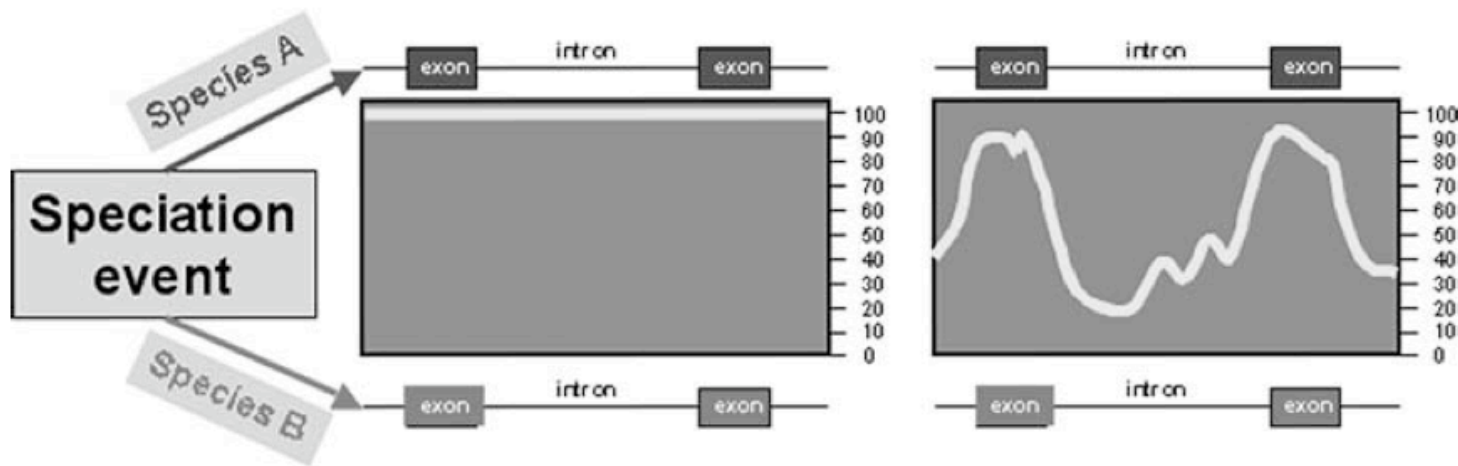
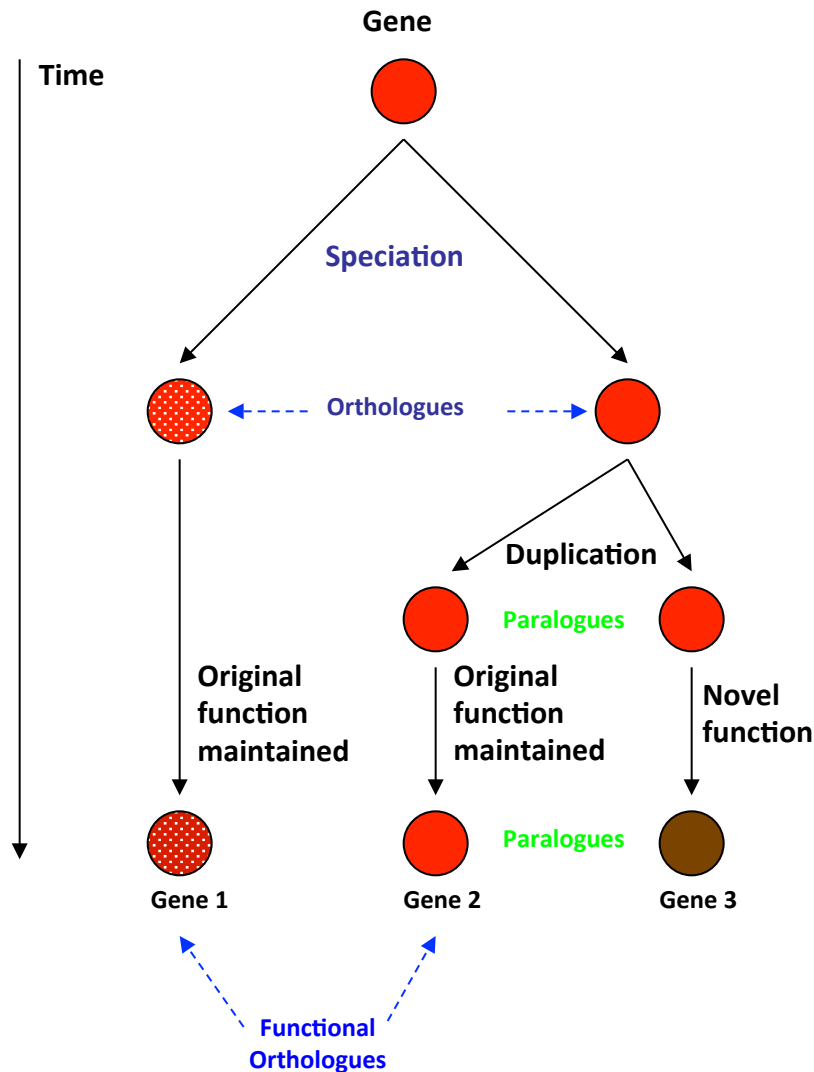


Fig 1. Miller *et al*, 2004. Ann Rev Genomics Hum Gen

# Why Comparative Sequence Analysis?

- allows us to achieve a greater understanding of vertebrate evolution
- tells us what is common and what is unique between different species at the genome level
- the function of human genes and other regions may be revealed by studying their counterparts in lower organisms
- helps identify both coding and non-coding genes and regulatory elements

# Homology, Orthology, Paralogy



**Homologues** - Genes derived from common ancestral gene

**Orthologues** – Genes in different species that are derived from the same gene in last common ancestor

**Paralogues** – Gene families that have diverged within a single species, often by duplication

# Identifying Orthologous Genes

Orthologue Prediction at Ensembl: <http://www.ensembl.org/>

Species	Type	dN/dS	Ensembl identifier & gene name	Compare	Location	Target %id	Query %id
Alpaca ( <i>Vicugna pacos</i> )	1-to-1	n/a	<a href="#">ENSVPAG00000001369</a> MITF microphthalmia-associated transcription factor [Source:HGNC Symbol;Acc:7105]	<ul style="list-style-type: none"> <li>Region Comparison</li> <li>Alignment (protein)</li> <li>Alignment (cDNA)</li> <li>Gene Tree (image)</li> </ul>	<a href="#">GeneScaffold_2364:880354-1101487:1</a>	82	79
Anole lizard ( <i>Anolis carolinensis</i> )	1-to-1	n/a	<a href="#">ENSACAG00000013586</a> MITF Uncharacterized protein [Source: UniProtKB/TrEMBL; acc: G1KP22]	<ul style="list-style-type: none"> <li>Region Comparison</li> <li>Alignment (protein)</li> <li>Alignment (cDNA)</li> <li>Gene Tree (image)</li> </ul>	<a href="#">2:181602837-181633021:-1</a>	86	69
Armadillo ( <i>Dasyus novemcinctus</i> )	1-to-1	n/a	<a href="#">ENSDNOG00000017544</a> MITF microphthalmia-associated transcription factor [Source:HGNC Symbol;Acc:7105]	<ul style="list-style-type: none"> <li>Region Comparison</li> <li>Alignment (protein)</li> <li>Alignment (cDNA)</li> <li>Gene Tree (image)</li> </ul>	<a href="#">GeneScaffold_6638:36209-296923:1</a>	79	79
Bushbaby ( <i>Otolemur garnettii</i> )	1-to-1	0.18451	<a href="#">ENSOGAG00000000043</a> MITF microphthalmia-associated transcription factor [Source:HGNC Symbol;Acc:7105]	<ul style="list-style-type: none"> <li>Region Comparison</li> <li>Alignment (protein)</li> <li>Alignment (cDNA)</li> <li>Gene Tree (image)</li> </ul>	<a href="#">GL873534.1:17118302-17213018:-1</a>	90	87
Caenorhabditis elegans ( <i>Caenorhabditis elegans</i> )	1-to-many	n/a	<a href="#">W02C12.3</a> hlh-30 Protein HLH-30, isoform b [Source:RefSeq]	<ul style="list-style-type: none"> <li>Region Comparison</li> <li>Alignment (protein)</li> <li>Alignment (cDNA)</li> </ul>	<a href="#">IV:4015486-4028129:-1</a>	20	21

Links to the closest putative orthologous genes in other species

Hyperlinks to view alignments & positional information

# Identifying Orthologous Genes

## NCBI Homologene

<http://www.ncbi.nlm.nih.gov/sites/entrez?db=homologene&cmd>

**NCBI HomoloGene**  
Discover Homologs

Search HomoloGene for

Limits Preview/Index History Clipboard Details

HomoloGene is a system for automated detection of homologs among the annotated genes of several completely sequenced eukaryotic genomes.

**HomoloGene Release 62 Statistics**  
Initial numbers of genes from complete genomes, numbers of genes placed in a homology group, and the numbers of groups for each species.

Species	Number of Genes		HomoloGene groups
	Input	Grouped	
Homo sapiens	22,849	19,964	19,351
Pan troglodytes	25,096	17,398	16,913
Canis lupus familiaris	19,766	16,732	16,294
Bos taurus	23,797	18,112	16,639
Mus musculus	25,388	21,538	19,410
Rattus norvegicus	21,991	19,092	17,865
Gallus gallus	17,959	12,988	12,279
Danio rerio	26,288*	17,789	15,288
Drosophila melanogaster	14,085	8,190	7,977
Anopheles gambiae	13,909	8,479	7,921
Caenorhabditis elegans	20,077	5,299	5,070
Schizosaccharomyces pombe	5,043	3,211	3,175
Saccharomyces cerevisiae	5,880	4,744	4,593
Kluyveromyces lactis	5,335	4,458	4,427
Eremothecium gossypii	4,722	3,949	3,940
Magnaporthe grisea	12,832	6,843	6,403
Neurospora crassa	10,079	6,128	6,122
Arabidopsis thaliana	26,981	13,374	13,041
Oryza sativa	26,887	12,973	12,603
Plasmodium falciparum	5,266	990	965

\* indicates organisms where new genome annotation data is used in this build.  
Last updated on: Mon Jul 28 2008

We have recently adopted a new build procedure that makes use of amino acid sequence searching (blastp) to find more distant relationships, but the procedure still refers to the DNA sequence for computation of some of the statistics. The matching strategy is guided by the taxonomic tree such that more closely related organisms are compared first. Moreover, HomoloGene entries now include paralogs in addition to orthologs.

**What's New**

HomoloGene release 62 is now public. It incorporates updated annotation for Danio rerio Zv7 release (NCBI release 3.1, Jun. 12, 2008).

**Tip of The Day**

You can use 'Details' in the tool bar to see how your query was translated and other query details.  
[\[More Tips\]](#)

**Related Resources**

**Entrez Genomes**

A collection of complete genome sequences that includes more than 1000 viruses and over hundred microbes

- Archaea
- Bacteria
- Eukaryota
- Viruses

**COGs**

Phylogenetic classification of proteins encoded in complete genomes.

Contains a wealth of information about homologous genes and links to other resources



# Identifying Orthologous Genes

## BLAST searches

<http://www.ncbi.nlm.nih.gov/BLAST/>

### BLAST Assembled Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

- [Human](#)
- [Mouse](#)
- [Rat](#)
- [Arabidopsis thaliana](#)
- [Oryza sativa](#)
- [Bos taurus](#)
- [Danio rerio](#)
- [Drosophila melanogaster](#)
- [Gallus gallus](#)
- [Pan troglodytes](#)
- [Microbes](#)
- [Apis mellifera](#)

Species specific searches

### Basic BLAST

Choose a BLAST program to run.

- [nucleotide blast](#) Search a **nucleotide** database using a **nucleotide** query  
*Algorithms: blastn, megablast, discontinuous megablast*
- [protein blast](#) Search **protein** database using a **protein** query  
*Algorithms: blastp, psi-blast, phi-blast*
- [blastx](#) Search **protein** database using a **translated nucleotide** query
- [tblastn](#) Search **translated nucleotide** database using a **protein** query
- [tblastx](#) Search **translated nucleotide** database using a **translated nucleotide** query

Nucleotide or protein searches

### Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with [Primer-BLAST](#)
- Search [trace archives](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdart)
- Search sequences that have [gene expression profiles](#) (GEO)
- Search [immunoglobulins](#) (IgBLAST)
- Search for [SNPs](#) (snp)
- Screen sequence for [vector contamination](#) (vecscreen)
- [Align](#) two sequences using BLAST (bl2seq)
- Search [protein](#) or [nucleotide](#) targets in PubChem BioAssay

Trace archives:  
A good place to look  
If you species of interest  
doesn't have a browser

# Paralogues in Ensembl:

The following gene(s) have been identified as putative paralogues (within species):

Taxonomy Level	Gene identifier
Euteleostomi parologue (within species)	<a href="#">ENSG00000068323</a> (TFE3) [ <a href="#">Multi-species comp.</a> ] [ <a href="#">Align</a> ] transcription factor binding to IGHM enhancer 3 [Source:HGNC Symbol;Acc:11752] [Target %id: 34; Query %id: 49]
Euteleostomi parologue (within species)	<a href="#">ENSG00000112561</a> (TFEB) [ <a href="#">Multi-species comp.</a> ] [ <a href="#">Align</a> ] transcription factor EB [Source:HGNC Symbol;Acc:11753] [Target %id: 38; Query %id: 41]
Euteleostomi parologue (within species)	<a href="#">ENSG00000105967</a> (TFEC) [ <a href="#">Multi-species comp.</a> ] [ <a href="#">Align</a> ] transcription factor EC [Source:HGNC Symbol;Acc:11754] [Target %id: 47; Query %id: 31]

[View sequence alignments of all homologues.](#)

[← Paralogues](#)
**Protein families** [help](#)
[Variation Table ▶](#)

Family ID	Consensus annotation	Other Human transcripts in this family	Multiple alignments
ENSFM0025000000692  <a href="#">(3 genes)</a> <a href="#">(all proteins in family)</a>	TRANSCRIPTION FACTOR	<ul style="list-style-type: none"> <li>● <a href="#">ENST00000352241</a> (MITF-001)</li> <li>● <a href="#">ENST00000451708</a> (MITF-003)</li> <li>● <a href="#">ENST00000394351</a> (MITF-004)</li> <li>● <a href="#">ENST00000314557</a> (MITF-005)</li> <li>● <a href="#">ENST00000448226</a> (MITF-007)</li> <li>● <a href="#">ENST00000433517</a> (MITF-009)</li> <li>● <a href="#">ENST00000472437</a> (MITF-014)</li> <li>● <a href="#">ENST00000314589</a> (MITF-015)</li> <li>● <a href="#">ENST00000328528</a> (MITF-201)</li> <li>● <a href="#">ENST00000394355</a> (MITF-202)</li> </ul>	225 Ensembl members of this family <a href="#">JalView</a>  316 members of this family <a href="#">JalView</a>
ENSFM00500000302678  <a href="#">(1 gene)</a> <a href="#">(all proteins in family)</a>	UNKNOWN	<ul style="list-style-type: none"> <li>● <a href="#">ENST00000394348</a> (MITF-006)</li> </ul>	2 Ensembl members of this family <a href="#">JalView</a>  2 members of this family <a href="#">JalView</a>
ENSFM00550000749363  <a href="#">(1 gene)</a> <a href="#">(all proteins in family)</a>	MICROPHthalmia ASSOCIATED TRANSCRIPTION FACTOR FRAGMENT	<ul style="list-style-type: none"> <li>● <a href="#">ENST00000457080</a> (MITF-002)</li> </ul>	1 Ensembl members of this family <a href="#">JalView</a>  2 members of this family <a href="#">JalView</a>
ENSFM00550000751758  <a href="#">(1 gene)</a> <a href="#">(all proteins in family)</a>	UNKNOWN	<ul style="list-style-type: none"> <li>● <a href="#">ENST00000429090</a> (MITF-012)</li> </ul>	1 Ensembl members of this family <a href="#">JalView</a>  1 members of this family <a href="#">JalView</a>

# How best to ensure that you have identified an orthologous gene

- **Percentage identity (protein and nucleotide)**  
(e.g. ClustalOmega, MUSCLE, sometimes Homologene)
- **Compare the size and number of exons in orthologous genes**  
(EST/cDNA to genomes – Splign , Ensembl ExonView)
- **Positional information - neighbouring genes**  
(Ensembl– SyntenyView, UCSC )
- **Confirm that no other paralogous genes are present in your species of interest**  
(BLAST, self-chain @UCSC, paralogues Ensembl)

By comparing two or more genome species we can identify Evolutionary Conserved Regions (ECRs).

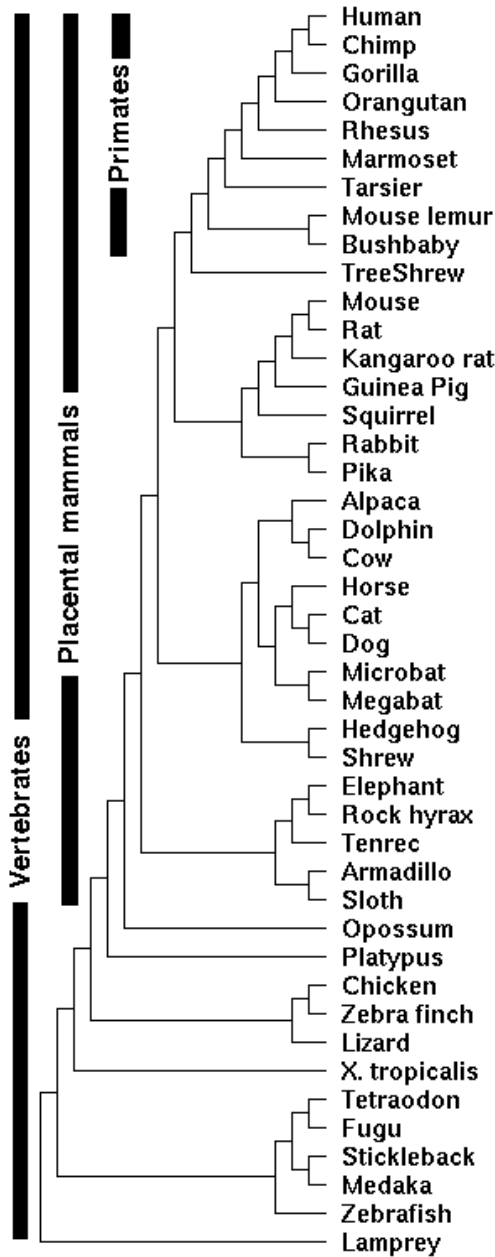
Within genes ECRs tend to be exonic but non-genic ECRs may function as *cis* acting elements such as enhancers













# Comparative Genome Analysis: Where to Start?

To identify conserved regions, you must:

- Decide which species you would like to compare
- Identify and extract the relevant genome sequences
- Annotate genes and other features found in the genome sequences
- Ensure that repetitive sequences are masked

# What vertebrate genomes are available?



- |   |   |   |
|---|---|---|
|  <b>Alpaca</b><br><i>Vicugna pacos</i>               |  <b>Hedgehog</b><br><i>Erinaceus europaeus</i>                         |  <b>Platypus</b><br><i>Ornithorhynchus anatinus</i>      |
|  <b>Anole Lizard</b><br><i>Anolis carolinensis</i>   |  <b>Horse</b><br><i>Equus caballus</i>                                 |  <b>Rabbit</b><br><i>Oryctolagus cuniculus</i>           |
|  <b>Armadillo</b><br><i>Dasyurus novemcinctus</i>    |  <b>Human</b><br><i>Homo sapiens</i>                                   |  <b>Rat</b><br><i>Rattus norvegicus</i>                  |
|  <b>Bushbaby</b><br><i>Otolemur garnettii</i>        |  <b>Hyrax</b><br><i>Procavia capensis</i>                              |  <b>Saccharomyces cerevisiae</b>                         |
|  <b>Caenorhabditis elegans</b>                       |  <b>Kangaroo rat</b><br><i>Dipodomys ordii</i>                         |  <b>Shrew</b><br><i>Sorex araneus</i>                    |
|  <b>Clona intestinalis</b>                           |  <b>Lamprey (preview - assembly only)</b><br><i>Petromyzon marinus</i> |  <b>Sloth</b><br><i>Choloepus hoffmanni</i>              |
|  <b>Clona savignyi</b>                               |  <b>Lesser hedgehog tenrec</b><br><i>Echinops telfairi</i>             |  <b>Squirrel</b><br><i>Spermophilus tridecemlineatus</i> |
|  <b>Cat</b><br><i>Felis catus</i>                    |  <b>Macaque</b><br><i>Macaca mulatta</i>                               |  <b>Stickleback</b><br><i>Gasterosteus aculeatus</i>     |
|  <b>Chicken</b><br><i>Gallus gallus</i>              |  <b>Marmoset</b><br><i>Callithrix jacchus</i>                          |  <b>Tarsier</b><br><i>Tarsius syrichta</i>               |
|  <b>Chimpanzee</b><br><i>Pan troglodytes</i>         |  <b>Medaka</b><br><i>Oryzias latipes</i>                               |  <b>Tetraodon</b><br><i>Tetraodon nigroviridis</i>       |
|  <b>Cow</b><br><i>Bos taurus</i>                    |  <b>Megabat</b><br><i>Pteropus vampyrus</i>                           |  <b>Tree Shrew</b><br><i>Tupaia belangeri</i>           |
|  <b>Dog</b><br><i>Canis familiaris</i>             |  <b>Microbat</b><br><i>Myotis lucifugus</i>                          |  <b>Turkey</b><br><i>Meleagris gallopavo</i>           |
|  <b>Dolphin</b><br><i>Tursiops truncatus</i>       |  <b>Mouse</b><br><i>Mus musculus</i>                                 |  <b>Wallaby</b><br><i>Macropus eugenii</i>             |
|  <b>Elephant</b><br><i>Loxodonta africana</i>      |  <b>Mouse Lemur</b><br><i>Microcebus murinus</i>                     |  <b>Xenopus tropicalis</b>                             |
|  <b>Fruitfly</b><br><i>Drosophila melanogaster</i> |  <b>Opossum</b><br><i>Monodelphis domestica</i>                      |  <b>Zebra Finch</b><br><i>Taeniopygia guttata</i>      |
|  <b>Fugu</b><br><i>Takifugu rubripes</i>           |  <b>Orangutan</b><br><i>Pongo pygmaeus</i>                           |  <b>Zebrafish</b><br><i>Danio rerio</i>                |
|  <b>Gorilla</b><br><i>Gorilla gorilla</i>          |  <b>Pig</b><br><i>Sus scrofa</i>                                     |   |
|  <b>Guinea Pig</b><br><i>Cavia porcellus</i>       |  <b>Pika</b><br><i>Ochotona princeps</i>                             |   |

## Selection of Species for DNA comparisons

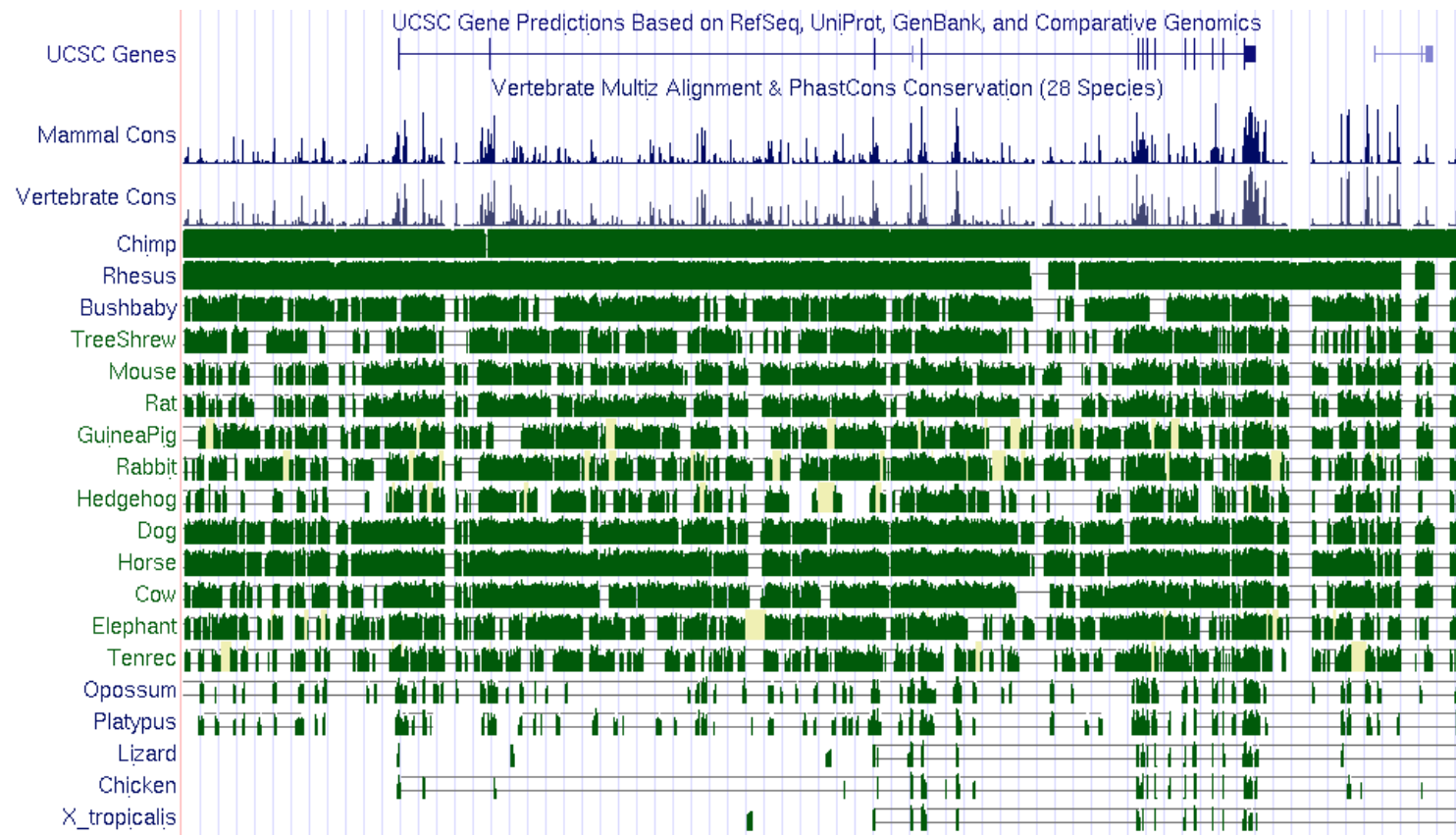


Human vs.	Chimpanzee	Mouse	Opossum	Pufferfish
Size (Gbp)	3.0	2.5	4.2	0.4
Time since divergence	~6 MYA	~ 90 MYA	~150 MYA	~450 MYA
Sequence conservation (in coding regions)	>99%	~80%	~70-75%	~65%
Aids identification of...	Recently changed sequences and genomic rearrangements	Both coding and non-coding sequences	Both coding and non-coding sequences	Primarily coding sequences
Background noise	High	Moderate	Low	Lower

# Aligning genomic sequence

- Pair-wise genome sequence alignments combined with additional phylogenetic information

(eg PhastCons@UCSC, RankVista,)

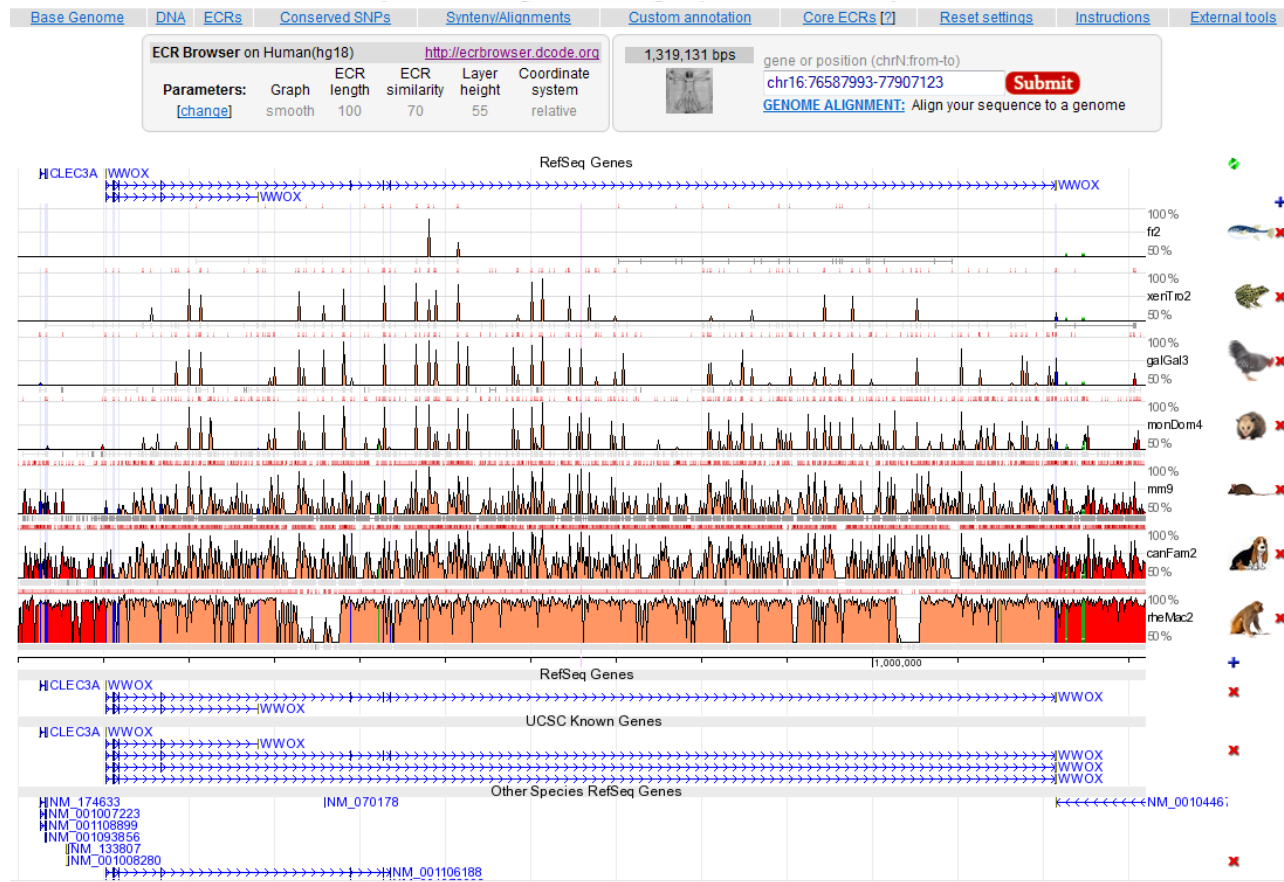




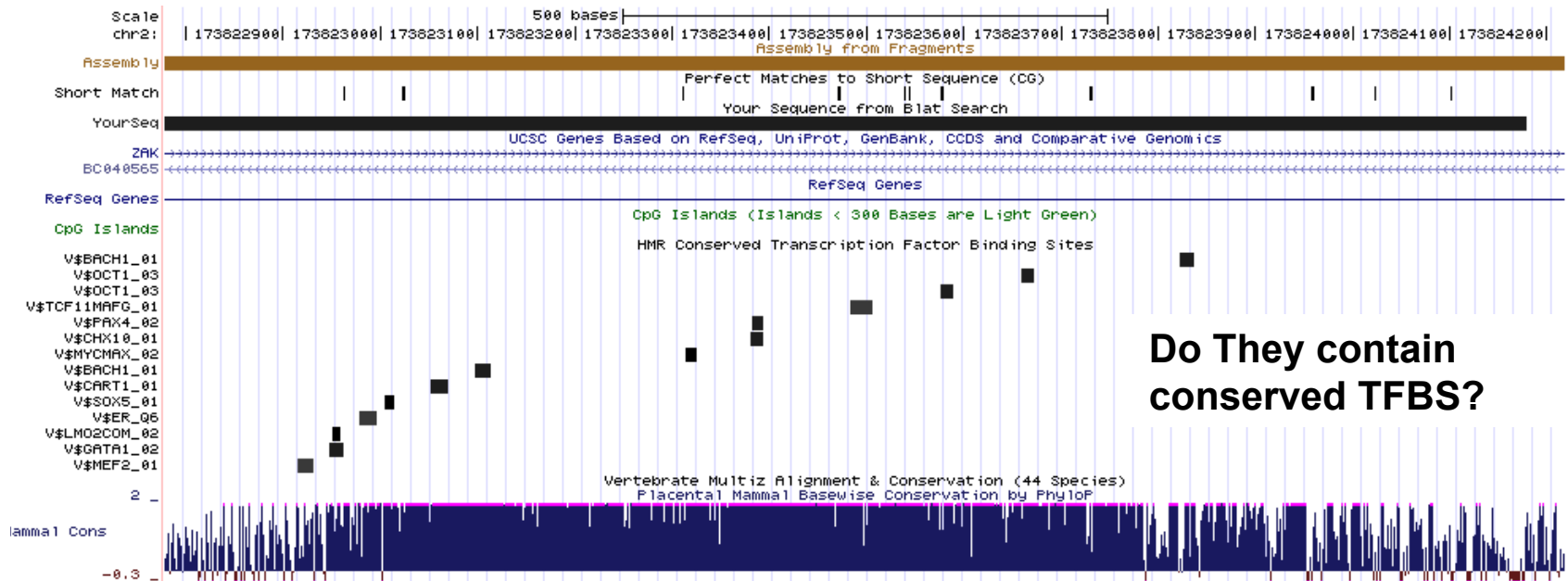
# Aligning genome sequences

- Pair-wise genome sequence alignments

(eg MultiContigView@Ensembl, PipMaker, Vista, ECR viewer, zPicture)

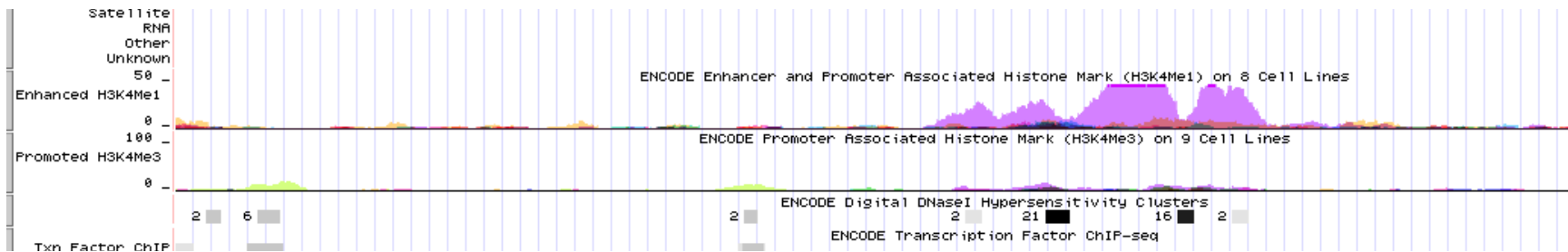


# Are these ECRS functional?



Do They contain conserved TFBS?

## OR Functional Chromatin signatures?



# Do ECRs have in vivo enhancer function?



The **VISTA Enhancer Browser** is a central resource for experimentally validated human noncoding fragments with gene enhancer activity as assessed in transgenic mice. Most of these noncoding elements were selected for testing based on their extreme conservation with other vertebrates. The results of this enhancer screen are provided through this publicly available website.

This program is located at [Lawrence Berkeley National Laboratory](#). See [Handbook](#) for additional details on this work or visit the [Experimental Results](#) to view data.

The browser also features relevant results by external contributors and a large collection of additional [genome-wide conserved noncoding elements](#) which are candidate enhancer sequences. We invite external groups to [submit computational predictions](#) of developmental enhancers.

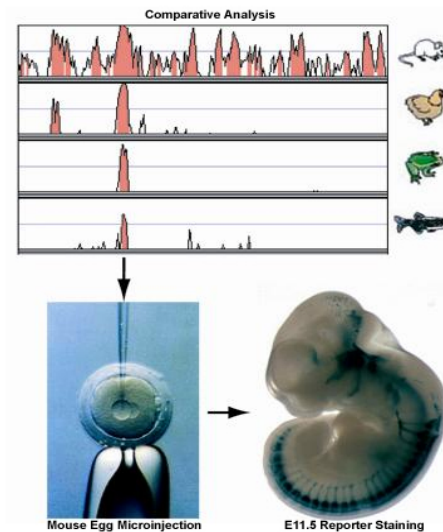
As of **4/14/2010** the database contains information on **1276** in vivo tested elements - **611** elements with enhancer activity.

### Keyword Search

Examples: gene, accession number, locus link, genomic position

[Expression Pattern Search](#)

[Advanced Search](#)



## Summary

In this module we have:

- Identified homologous sequences
- Reviewed the means by which true orthology can be determined
- Viewed evolutionary conserved regions using genome browsers
- Aligned genome sequences and identified both ECRs and conserved transcription factor binding sites
- Identified regions with enhancer associated chromatin signatures

# Evolutionary Conserved Regions

- **Manual**
  - Pipmaker – <http://bio.cse.psu.edu/cgi-bin/pipmaker>  
requires repeatmasked and annotation files  
Local alignment, BLASTZ
  - Vista – <http://www-gsd.lbl.gov/vista>  
requires annotation files, repeat masks for you  
Global alignment, AVID
- **Semi automated (Currently not working)**
  - zPicture - <http://www.dcode.org>  
Local alignment, BLASTZ

## Evolutionary Conserved Regions

- **Automatic**
  - Genome Browsers, e.g UCSC and Ensembl
  - ECRbrowser - <http://www.dcode.org>  
BLAT, BLAST and BLASTZ  
Can link to both zPicture and rVista