

Module 3: Comparative Sequence Analysis and Identification of Regulatory Elements

Aims

- Overview of comparative sequence analysis and sequences available from different organisms for homologous gene identification
- Identify putative paralogous and orthologous genes in Ensembl
- View phylogenetic trees
- Compare genome sequences from different organisms in UCSC
- Identify evolutionary conserved sequences (ECRs) in ECR browser and VISTA Enhancer Browser

Comparative Sequence Analysis

Comparative sequence analysis is a powerful method for aiding human gene identification, inferring function of a gene's product, and identifying novel functional elements such as those involved in transcriptional regulation. This is because biologically important regions of the genome are, generally, under selective constraint. Comparing the genome sequences from a variety of organisms may facilitate the identification of functionally significant units in the human genome.

The information that can be inferred when comparing sequences is dependent on the evolutionary distance between the two organisms. Organisms that are closely related are more likely to share a higher degree of sequence similarity. Organisms more distantly related to human, such as yeast and worm, share less sequence similarity and are likely to show sequence conservation only in coding regions. This may also be true for distantly related vertebrates such as fish. More closely related organisms, such as mouse, are likely to be conserved in coding regions, and other functional elements such as regulatory sequences. However, the closer the evolutionary relationship with human, the more 'sequence noise' is likely to arise where non-functional sequence appears similar because insufficient time has elapsed for the two sequences to diverge.

Evolutionarily Related Gene Sequences

Homologous genes are derived from a common ancestor and may retain a similar sequence or function. In general, homologous genes can be divided into two classes:

1) **Orthologues** are genes that often perform the same function in different organisms. They are defined as being homologous genes in different organisms derived from the same gene during speciation. In general, their sequence similarity reflects the amount of time since they diverged from a common ancestor (i.e. the less time that has elapsed since divergence, the greater the sequence similarity between the two genes).

2) **Paralogues** are gene families that are present within a single species. Often they arise by duplication. These genes are not under the same pressure to maintain their function so that one copy may acquire a novel function.

***note: The terms orthologous and paralogous can apply to sequence without genes as well!**

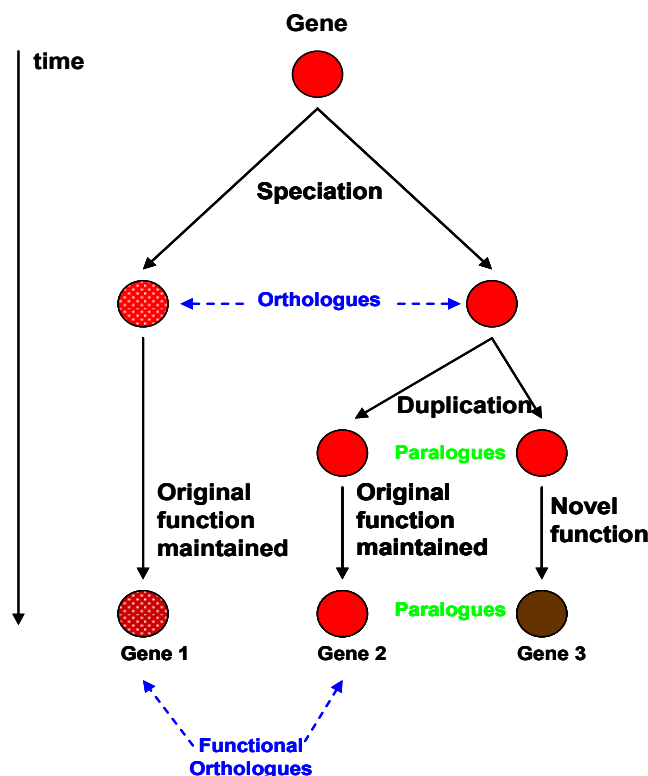


Figure 1 - Homologous Gene Sequences

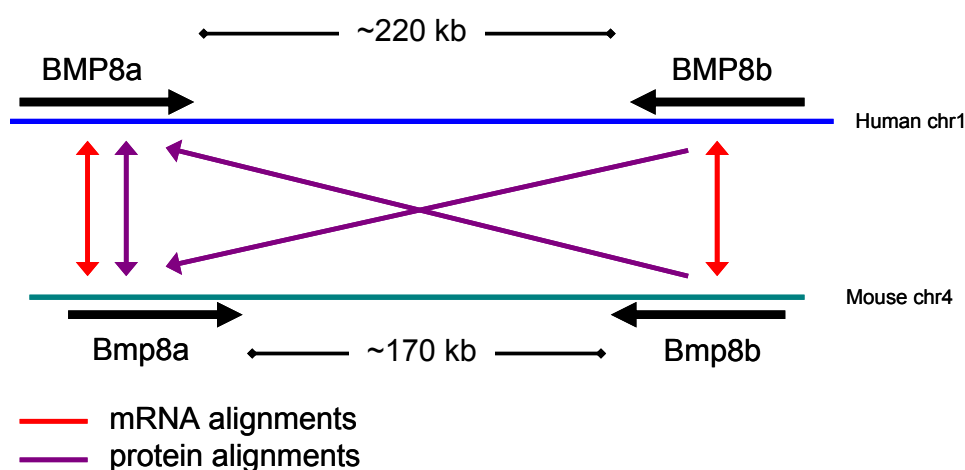
Identifying evolutionarily related gene sequences – where to start...

Most searches for orthologous genes begin with BLAST searches. The type of search that you should perform depends upon what information you have at your disposal. Protein sequence based queries generally find more distantly related matches because of the redundancy in the genetic code (i.e. some amino acids are encoded by more than one codon). Nucleotide searches using the discontinuous megablast parameters are also very useful. We recommend that you survey a number of different databases using a number of different search parameters to obtain the most informative results.

BE CAUTIOUS: uncertain Orthologues

You may encounter gene sequences that appear to be orthologous and may be derived from the same ancestor but no longer perform the same biological function (for example genes 1 and 3 in Figure1). If you choose to analyze such sequences, the sensitivity and specificity of your search will be reduced and it may not yield any informative results.

For example, the gene for bone morphogenetic protein 8 (*BMP8*) was duplicated in a common ancestor of human and mouse, giving rise to *BMP8a* and *BMP8b* (see Figure below). BLAST analysis of these four sequences yields quite confusing results. Human and mouse *BMP8a* are reciprocal best alignments using both nucleotide and protein sequences to search. In contrast, both the nucleotide and protein sequences of mouse *Bmp8b* align best to their human *BMP8a* counterparts. Human *BMP8b* mRNA aligns best to mouse *Bmp8b* mRNA, but human *BMP8b* protein aligns best to mouse *Bmp8a* protein, while mouse protein *Bmp8b* aligns best to human protein *BMP8a* (Nardoneet *al.*, 2004).



Adapted from Nardone *et al.*, 2004

Figure 2 - Uncertain *BMP8a* and *BMP8b* orthologues.

Therefore, we recommend that you perform the following steps to confirm the true functionality and relatedness of your gene sequences. Much of this information can be obtained from genome browsers such as Ensembl, NCBI or the UCSC.

- 1) Identify any other paralogues that may affect your analysis. This can be achieved by performing a BLAST search using your sequence against its source genome, or using the self-chain track at the UCSC genome browser.
- 2) Confirm the percent identity (similarity) at both the nucleotide and protein level between paralogous and orthologous sequences to ensure that you are analyzing the most closely related sequences.
- 3) Perform evolutionary analyses of nucleotide/protein sequence (phylogeny). In contrast to similarity-based methods such as BLAST, phylogenetic methods can better take into account the effects of repeated substitutions at one site and variable rates of evolution among sequences. Multiple genes are placed in an evolutionary tree representing genealogical relationships.
- 4) Compare the exon/intron structure of your orthologous genes. Evolutionary related genes often share a similar gene structure.
- 5) Examine the chromosomal context of the two orthologous genes. Closely related species, such as human and mouse often have large

conserved segments and therefore neighbouring genes are also shared between the two species.

Comparative Genome Analysis

Comparing the DNA sequences of different species is a powerful method for decoding genomic information. This is because functional sequences tend to evolve at a slower rate than non-functional sequences (see Figure 3).

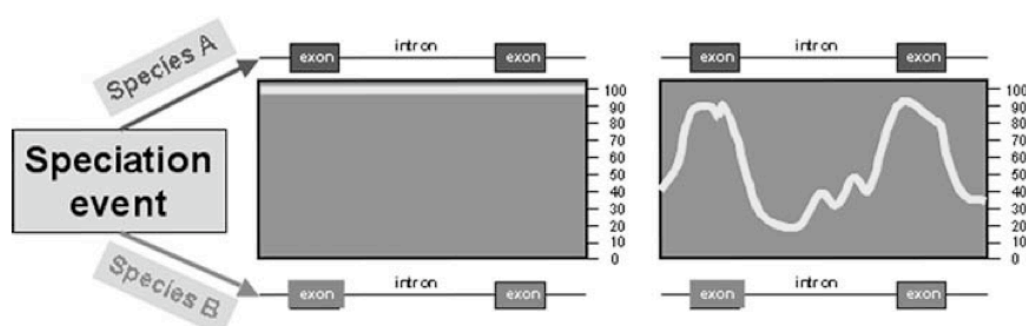


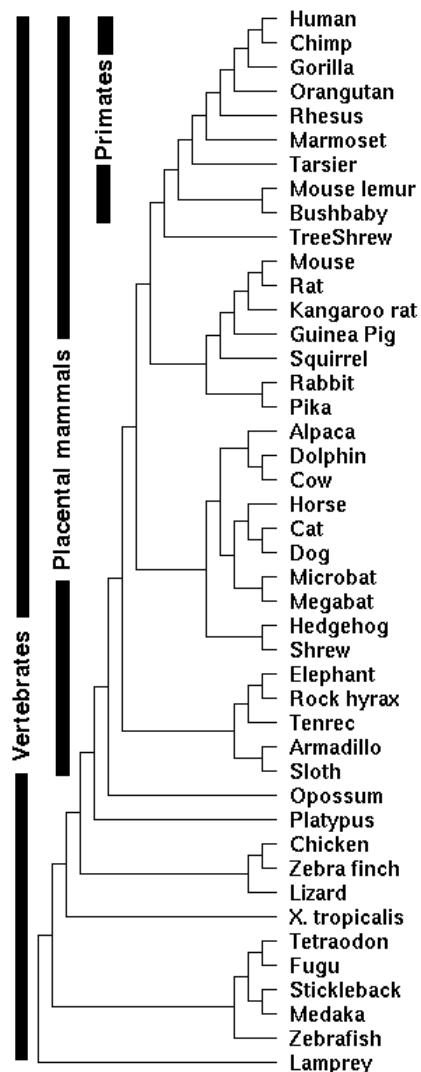
Figure 3 - Diverging sequences. Miller *et al*, 2004. *Annu Rev Genomics Hum Genet*. Immediately after speciation the sequence is 100% identical (see graph on left). Over time, regions under little or no selective pressure, such as introns, are saturated with mutations, whereas regions under negative selection, such as most exons, retain a higher percent identity (see graph on right).

By comparing the genomic sequences of several species at different evolutionary distances it is possible to identify coding sequences, conserved non-coding sequences, and those sequences that are unique to humans or other species.

When two species diverge from a common ancestor, those sequences that maintain their original function are likely to remain conserved in both species throughout their subsequent independent evolution. Therefore, comparing sequences in different species is a powerful tool for increasing the confidence of a predicted functional unit, or identifying novel functional units (e.g. human, mouse and zebrafish).

The sequence of many genomes has been generated using a combination of the clone-by-clone method (adopted for generating the human genome sequence by the public effort) and whole genome shotgun (WGS - used by Celera to generate the

sequence of the fruit fly and their version of the human genome sequence). Finished and unfinished genome assemblies are currently available for many vertebrates.



The key genomes being used to predict conserved elements are shown from the UCSC website.

Figure 4 - The genomes that have been aligned to predict conserved elements at the UCSC genome browser.

In addition to many completely sequenced genomes, low coverage sequence for numerous genomes is currently available and more are in queue to be processed. Many new sequencing platforms offered by Illumina, Roche (454) and Agencourt/ABI (SOLiD) are offering faster ways to obtain many new partial and whole genome sequences for much less cost. The draft sequence of the giant panda genome was the first to use only next generation sequencing (Li *et al* Nature 2010).

In general, greater evolutionary distance between the species is reflected by more divergent sequences and fewer shared functional units (see table below). Comparing sequences that diverged from a common ancestor approximately 450 million years ago (mya) (e.g. human and fish), aids the identification of coding sequences. Conserved non-coding regions are generally not identified. If the evolutionary distance between the two species is reduced to approximately 90 mya (e.g. human and mouse), both non-coding and coding units are commonly conserved. However there can be background noise so comparing a number of different mammalian species (including the more distantly related marsupials) is useful. A large number of features are conserved between recently evolved species such as human and chimpanzee. The inclusion of a closely related species in a comparative analysis makes it possible to identify those genomic sequences that may be responsible for traits that are unique to the reference species.

Table 1 - Selection of Species for DNA Comparisons

Human vs.	Chimpanzee	Mouse	Opossum	Fish
Size (Gbp)	3.0	2.5	3.6	0.4
Time since divergence	~5 MYA	~65MYA	~150 MYA	~450 MYA
Sequence conservation (in coding regions)	>99%	~80%	~75%	~65%
Aids identification of	Recently changed sequences and genomic rearrangements	Both coding and non-coding sequences	Both coding and non-coding sequences	Primary coding sequences
Background noise	HIGH	MODERATE	LOWER	LOW

Gene: MITF ENSG00000187098

Description microphthalmia-associated transcription factor [Source:HGNC Symbol;Acc:HGNC:7105]

Synonyms bHLHe32, MI, WS2, WS2A

Location [Chromosome 3: 69,739,435-69,968,337](#) forward strand.

INSDC coordinates chromosome:GRCh38:CM000665.2:69739435;69968337:1

Transcripts This gene has 17 transcripts (splice variants) [Show transcript table](#)

Summary

Name [MITF](#) (HGNC Symbol)

CCDS This gene is a member of the Human CCDS set: [CCDS2913](#), [CCDS43106](#), [CCDS43107](#), [CCDS46865](#), [CCDS46866](#), [CCDS54607](#), [CCDS74962](#)

UniprotKB This gene has proteins that correspond to the following Uniprot identifiers: [O75030](#)

RefSeq Overlapping RefSeq Gene ID [4286](#) matches and has similar biotype of protein_coding

LRG [LRG 776](#) provides a stable genomic reference framework for describing sequence variations for this gene

Ensembl version ENSG00000187098.12

GRCh37 assembly This gene maps to [69,788,586-70,017,488](#) in GRCh37 coordinates. View this locus in the GRCh37 archive: [ENSG00000187098](#)

Gene type Known protein coding

Prediction Method Annotation for this gene includes both automatic annotation from Ensembl and [Havana](#) manual curation, see [article](#).

Alternative genes This gene corresponds to the following database identifiers:
Havana gene: [OTTHUMG00000149921](#)

Go to Region in Ensembl

Step 3: View protein families

Step 4: Protein families indicates potential paralogues. View location in genome - click where it says 5 genes.

Ensembl protein families

Family ID	Consensus annotation	Other MITF proteins in this family	Multiple alignments
ENSM00250200000692 (5 genes)	TRANSCRIPTION FACTOR	<ul style="list-style-type: none"> ENSP00000295600 (MITF-001) ENSP00000391803 (MITF-007) ENSP00000411389 (MITF-009) ENSP00000418845 (MITF-014) ENSP00000391276 (MITF-002) ENSP00000324443 (MITF-015) ENSP00000398639 (MITF-003) ENSP00000324246 (MITF-005) ENSP00000377880 (MITF-004) ENSP00000433487 (MITF-008) ENSP00000481286 (MITF-006) ENSP00000435909 (MITF-016) ENSP00000327867 (MITF-201) 	398 Ensembl members of this family JaView 662 members of this family JaView
ENSM00550000751758 (1 gene)	UNKNOWN	<ul style="list-style-type: none"> ENSP00000407620 (MITF-012) 	1 Ensembl members of this family JaView 1 members of this family JaView

HUMAN genes in this family [help](#)

[Variation Table](#)

Ensembl genes containing proteins in family ENSFM00250000000692

MITF has 3 potential paralogues. There are 4 entries as one is on a GRC patch.

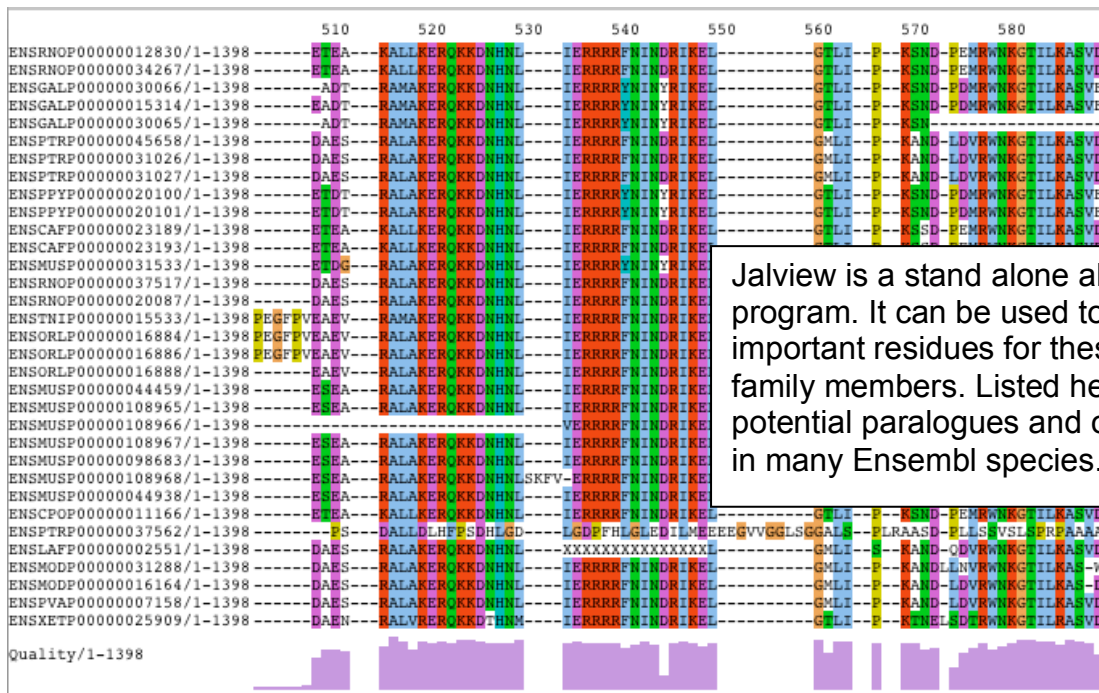
Gene ID and Location	Gene Name	Description (if known)	Protein ID(s)
ENSG00000187098 <small>Chromosome 3: 69.74m</small>	MITF	microphthalmia-associated transcription factor [Source:HGNC Symbol;Acc:HGNC:7105]	
ENSG00000112561 <small>Chromosome 6: 41.68m</small>	TFEB	transcription factor EB [Source:HGNC Symbol;Acc:HGNC:11753]	
ENSG00000105967 <small>Chromosome 7: 115.94m</small>	TFEC	transcription factor EC [Source:HGNC Symbol;Acc:HGNC:11754]	
LRG_776 <small>LRG_776: 202.12k</small>	MITF	microphthalmia-associated transcription factor [Source:HGNC Symbol;Acc:HGNC:7105]	
ENSG00000068323 <small>Chromosome X: 49.03m</small>	TFE3	transcription factor binding to IGHM enhancer 3 [Source:HGNC Symbol;Acc:HGNC:11752]	

Multiple alignments

398 Ensembl members of this family [JalView](#)

662 members of this family [JalView](#)

Step 5: Navigate back to protein families and click JalView under the multiple alignments section. In some browsers, a second JalView button may appear. Click JalView for family members.



- Gene-based displays
 - Summary
 - Splice variants (17)
 - Transcript comparison
 - Supporting evidence
 - Gene alleles
 - [-] Sequence
 - └ Secondary Structure
 - External references
 - Regulation
 - [-] Comparative Genomics
 - Genomic alignments
 - **Gene tree**
 - Gene gain/loss tree
 - Orthologues (78)
 - Paralogues (3)
 - Ensembl protein families (2)
 - Phenotype
 - [-] Genetic Variation
 - Variation table
 - Variation image
 - Structural variation
 - [-] External data
 - Personal annotation
 - [-] ID History
 - Gene history

Step 6: Return to the gene view page for *MITF* and view. This is a different way of getting to the same answer.

Type	Ancestral taxonomy	Ensembl identifier & gene name	Compare	Location	Target %id	Query %id
Paralogues (same species)	Bony vertebrates (Euteleostomi)	ENSG00000068323 TFE3 transcription factor binding to IGHM enhancer 3 [Source:HGNC Symbol;Acc:HGNC:11752]	<ul style="list-style-type: none"> Region Comparison Alignment (protein) Alignment (cDNA) 	X:49028726-49043486:-1	48	53
Paralogues (same species)	Bony vertebrates (Euteleostomi)	ENSG00000105967 TFEC transcription factor EC [Source:HGNC Symbol;Acc:HGNC:11754]	<ul style="list-style-type: none"> Region Comparison Alignment (protein) Alignment (cDNA) 	7:115935148-116159896:-1	47	32
Paralogues (same species)	Vertebrates (Vertebrata)	ENSG00000112561 TFEB transcription factor EB [Source:HGNC Symbol;Acc:HGNC:11753]	<ul style="list-style-type: none"> Region Comparison Alignment (protein) Alignment (cDNA) 	6:41683978-41736259:-1	40	38

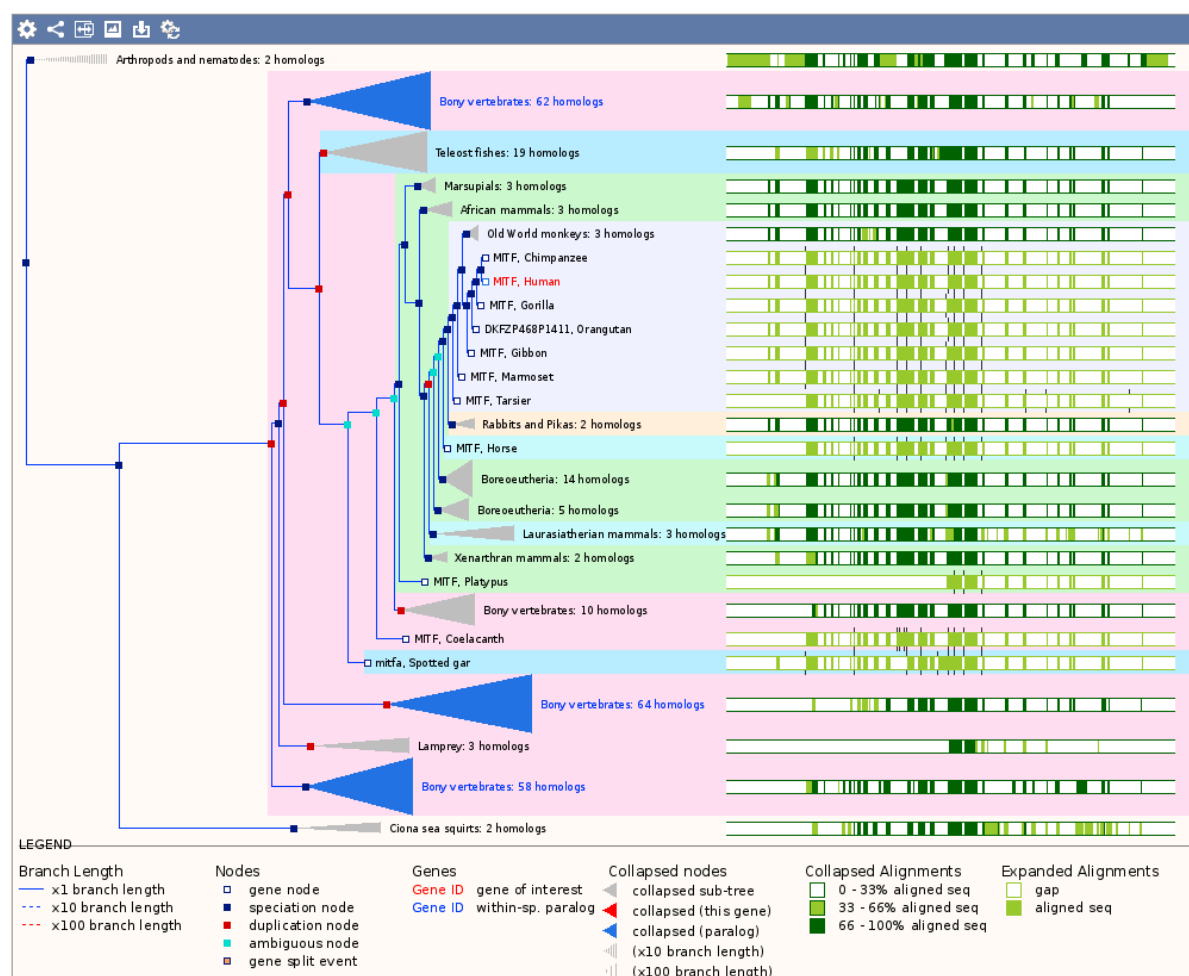
Orthologues

Orthologues are the same gene but in different species. To investigate the evolution of a gene we can investigate the relationships between orthologues and paralogues in all the species for which genomic sequence is available

Gene-based displays

- Summary
- Splice variants (17)
- Transcript comparison
- Supporting evidence
- Gene alleles
- [-] Sequence
 - Secondary Structure
- External references
- Regulation
- [-] Comparative Genomics
 - Genomic alignments
 - Gene tree**
 - Gene gain/loss tree
 - Orthologues (78)
 - Paralogues (3)
 - Ensembl protein families (2)
- Phenotype
- [-] Genetic Variation
 - Variation table
 - Variation image
 - Structural variation
- [-] External data
 - Personal annotation
- [-] ID History
 - Gene history

Step 7: Click on the Gene: *MITF* tab at the top of the page to get back to the gene page and then click Gene Tree (Image)



Maximum likelihood phylogenetic tree drawn using **PHYML**. Red squares represent duplication nodes, while blue squares represent speciation nodes. When the branch length is too long for the display, they are shortened and displayed as dashed lines with color depending on the branch length. The green bars at the right of the tree are a schematic representation of the multiple alignment of the peptides made using **MUSCLE**. Full boxes indicate matches/mismatches, open boxes indicate gaps in the alignment.

Ensembl uses a pipeline to predict orthologues and paralogues in which "maximum likelihood phylogenetic gene trees (generated by TreeBeST) play a central role". However, **BE WARNED** as some of the orthologues predicted in the less characterised genomes are not always well annotated.

Gene-based displays

- Summary
- Splice variants (17)
- Transcript comparison
- Supporting evidence
- Gene alleles
- Sequence
 - Secondary Structure
- External references
- Regulation
- Comparative Genomics
 - Genomic alignments
 - Gene tree**
 - Gene gain/loss trees
 - Orthologues (78)
 - Paralogues (3)
 - Ensembl protein families (2)
- Phenotype
- Genetic Variation
 - Variation table
 - Variation image
 - Structural variation
- External data
 - Personal annotation
- ID History
 - Gene history

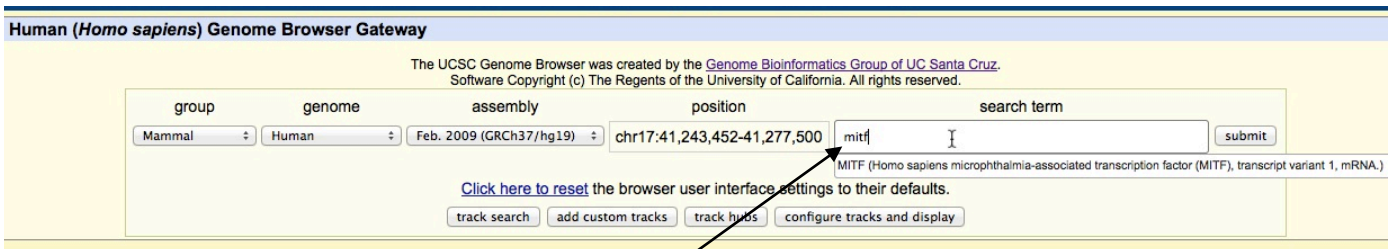
Step 8: Click on orthologues

Make sure you check carefully as not all orthologues are 1-to-1 e.g. zebrafish has 2 copies of *MITF*. Some orthologues may be partial e.g. platypus.

Turkey (<i>Meleagris gallopavo</i>)	1-to-1	n/a	ENSMGAG00000009507 MITF microphthalmia-associated transcription factor [Source:HGNC Symbol;Acc:7105]	<ul style="list-style-type: none"> • Region Comparison • Alignment (protein) • Alignment (cDNA) • Gene Tree (image) 	14:16287512-16329099:1	87	86
Wallaby (<i>Macropus eugenii</i>)	1-to-1	n/a	ENSMEUG0000000099 MITF microphthalmia-associated transcription factor [Source:HGNC Symbol;Acc:7105]	<ul style="list-style-type: none"> • Region Comparison • Alignment (protein) • Alignment (cDNA) • Gene Tree (image) 	GeneScaffold_9177:11014-95772:1	85	86
Xenopus (<i>Xenopus tropicalis</i>)	1-to-1	n/a	ENSXETG00000000154 mitf microphthalmia-associated transcription factor [Source:Jamboree;Acc:XB-GENE-484962]	<ul style="list-style-type: none"> • Region Comparison • Alignment (protein) • Alignment (cDNA) • Gene Tree (image) 	GL173261.1:16072-155698:1	78	73
Zebra Finch (<i>Taeniopygia guttata</i>)	1-to-many	n/a	ENSTGUG00000009847 Novel Ensembl prediction Uncharacterized protein [Source: UniProtKB/TrEMBL; acc: H0ZHV0]	<ul style="list-style-type: none"> • Region Comparison • Alignment (protein) • Alignment (cDNA) • Gene Tree (image) 	12:16777226-16818678:1	92	83
Zebra Finch (<i>Taeniopygia guttata</i>)	1-to-many	n/a	ENSTGUG000000017182 Novel Ensembl prediction Uncharacterized protein [Source: UniProtKB/TrEMBL; acc: H1A3L0]	<ul style="list-style-type: none"> • Region Comparison • Alignment (protein) • Alignment (cDNA) • Gene Tree (image) 	Un:26926871-26938092:-1	81	36
Zebrafish (<i>Danio rerio</i>)	1-to-many	n/a	ENSDARG00000003732 mitfa microphthalmia-associated transcription factor a [Source:ZFIN;Acc:ZDB-GENE-990910-11]	<ul style="list-style-type: none"> • Region Comparison • Alignment (protein) • Alignment (cDNA) • Gene Tree (image) 	6:43321687-43364577:1	62	49
Zebrafish (<i>Danio rerio</i>)	1-to-many	n/a	ENSDARG000000037833 mitfb microphthalmia-associated transcription factor b [Source:ZFIN;Acc:ZDB-GENE-010919-1]	<ul style="list-style-type: none"> • Region Comparison • Alignment (protein) • Alignment (cDNA) • Gene Tree (image) 	23:730122-781915:-1	66	63

2. Comparing genome sequences in the UCSC.

We will next use the UCSC genome browser to identify evolutionarily conserved regions and then investigate them to see if they have any epigenetic signatures that suggest they have regulatory roles in the genome.

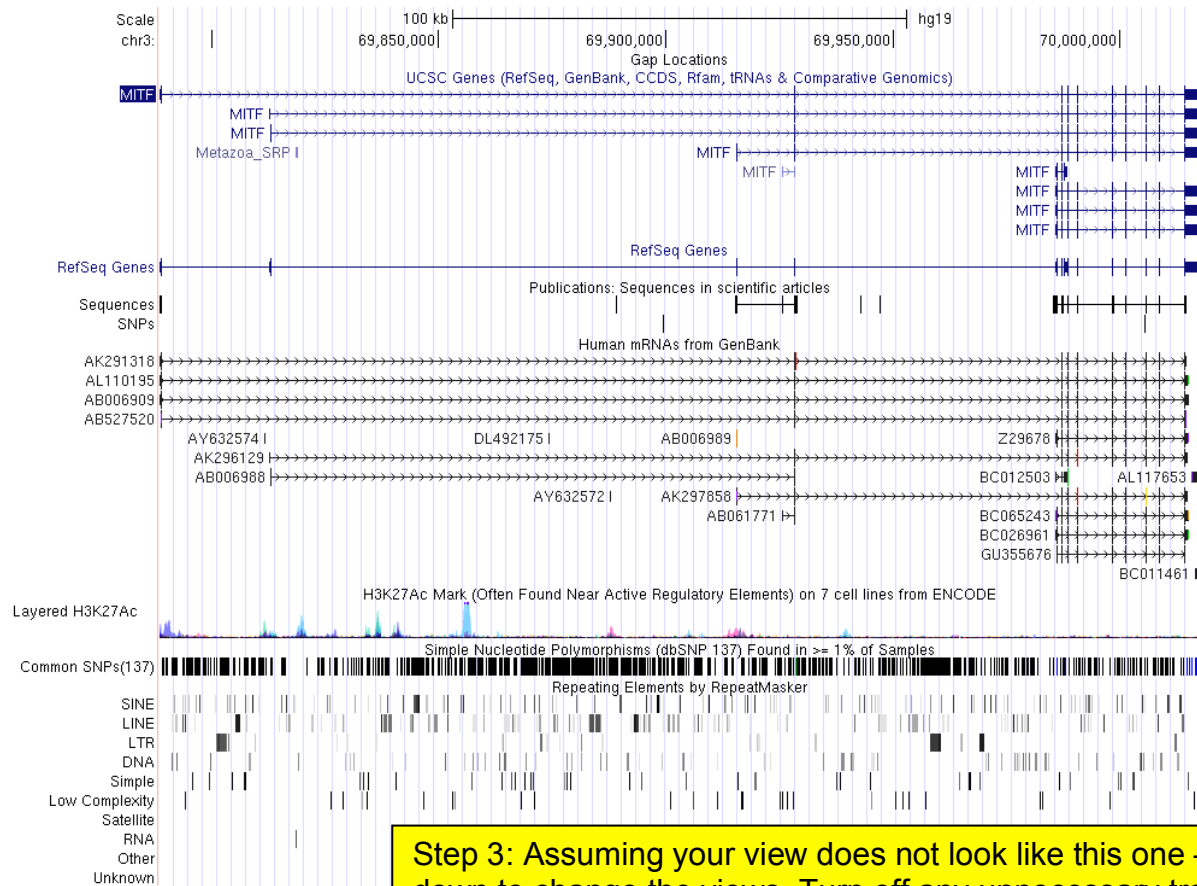


Step 1: Search for the *MITF* gene. The search box will autocomplete if the term is a known gene name.

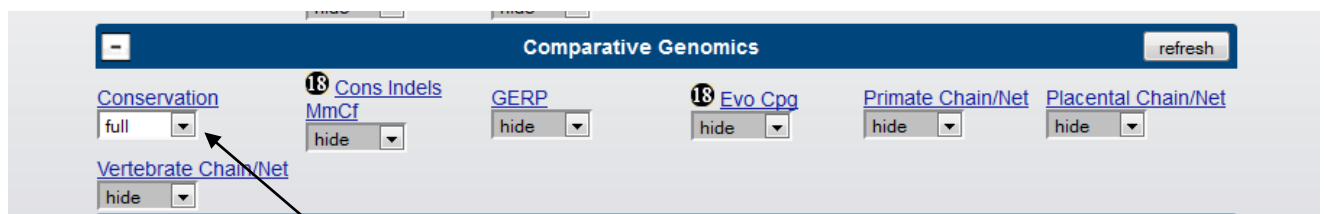
UCSC Genes

MITF (uc021xam.1) at chr3:69985751-70017488	- Homo sapiens microphthalmia-associated transcription factor (<i>MITF</i>), transcript variant 6, mRNA.
MITF (uc021xal.1) at chr3:69985751-69988216	- Homo sapiens microphthalmia-associated transcription factor (<i>MITF</i>), transcript variant 8, mRNA.
MITF (uc011bgb.2) at chr3:69812707-70017488	- Homo sapiens microphthalmia-associated transcription factor (<i>MITF</i>), transcript variant 7, mRNA.
MITF (uc003dof.3) at chr3:69985751-70017488	- Homo sapiens microphthalmia-associated transcription factor (<i>MITF</i>), transcript variant 4, mRNA.
MITF (uc003doe.3) at chr3:69985751-70017488	- Homo sapiens microphthalmia-associated transcription factor (<i>MITF</i>), transcript variant 5, mRNA.
MITF (uc003doc.1) at chr3:69925491-69928534	- Homo sapiens microphthalmia-associated transcription factor (<i>MITF</i>), transcript variant 2, mRNA.
MITF (uc003dob.3) at chr3:69915375-70017488	- Homo sapiens microphthalmia-associated transcription factor (<i>MITF</i>), transcript variant 2, mRNA.
MITF (uc003doa.3) at chr3:69812962-70017488	- Homo sapiens microphthalmia-associated transcription factor (<i>MITF</i>), transcript variant 3, mRNA.
MITF (uc003danz.3) at chr3:69788586-70017488	- Homo sapiens microphthalmia-associated transcription factor (<i>MITF</i>), transcript variant 1, mRNA.
TFE3 (uc004dmb.3) at chrX:48886242-48900990	- Homo sapiens transcription factor binding to IGHM enhancer 3 (<i>TFE3</i>), mRNA.
USP13 (uc003fkh.3) at chr3:179370933-179507189	- Homo sapiens ubiquitin specific peptidase 13 (isopeptidase T-3) (<i>USP13</i>), mRNA.
IKZF4 (uc001sjd.1) at chr12:56414689-56432219	- Homo sapiens IKAROS family zinc finger 4 (<i>Eos</i>) (<i>IKZF4</i>), mRNA.
TFE3 (uc004dmc.3) at chrX:48886242-48900990	- Homo sapiens transcription factor binding to IGHM enhancer 3 (<i>TFE3</i>), mRNA.
TFEB (uc003oqs.1) at chr6:41651716-41702798	- Homo sapiens transcription factor EB (<i>TFEB</i>), transcript variant 1, mRNA.
TFEB (uc031sop.1) at chr6:41651716-41702139	- Homo sapiens transcription factor EB (<i>TFEB</i>), transcript variant 4, mRNA.
TFEB (uc010jxo.1) at chr6:41655083-41701591	- Homo sapiens transcription factor EB (<i>TFEB</i>), transcript variant 2, mRNA.
TFEB (uc003oqv.1) at chr6:41654267-41701591	- Homo sapiens transcription factor EB (<i>TFEB</i>), transcript variant 3, mRNA.
TFEB (uc003oqt.2) at chr6:41651716-41703265	- Homo sapiens transcription factor EB (<i>TFEB</i>), transcript variant 5, mRNA.
IKZF4 (uc001sjc.1) at chr12:56414689-56432219	- Homo sapiens IKAROS family zinc finger 4 (<i>Eos</i>) (<i>IKZF4</i>), mRNA.
IKZF4 (uc001sjb.1) at chr12:56413104-56432219	- Homo sapiens IKAROS family zinc finger 4 (<i>Eos</i>) (<i>IKZF4</i>), mRNA.
CADM1 (uc031gei.1) at chr11:115044345-115375241	- Homo sapiens cell adhesion molecule 1 (<i>CADM1</i>), transcript variant 1, mRNA.
CADM1 (uc001ppi.4) at chr11:115044345-115375241	- Homo sapiens cell adhesion molecule 1 (<i>CADM1</i>), transcript variant 1, mRNA.
CADM1 (uc031geh.1) at chr11:115044345-115285467	- Homo sapiens cell adhesion molecule 1 (<i>CADM1</i>), transcript variant 1, mRNA.
KARS (uc002feq.3) at chr16:75661622-75681585	- Homo sapiens lysyl-tRNA synthetase (<i>KARS</i>), transcript variant 2, mRNA.
TFEC (uc003vhj.2) at chr7:115575202-115670867	- Homo sapiens transcription factor EC (<i>TFEC</i>), transcript variant 1, mRNA.
PIAS3 (uc001eoc.1) at chr11:145575988-145586546	- Homo sapiens protein inhibitor of activated STAT, 3 (<i>PIAS3</i>), mRNA.
TYR (uc001pcs.3) at chr11:88911040-89028927	- Homo sapiens tyrosinase (<i>TYR</i>), mRNA.

Step 2: Click on the longest transcript.



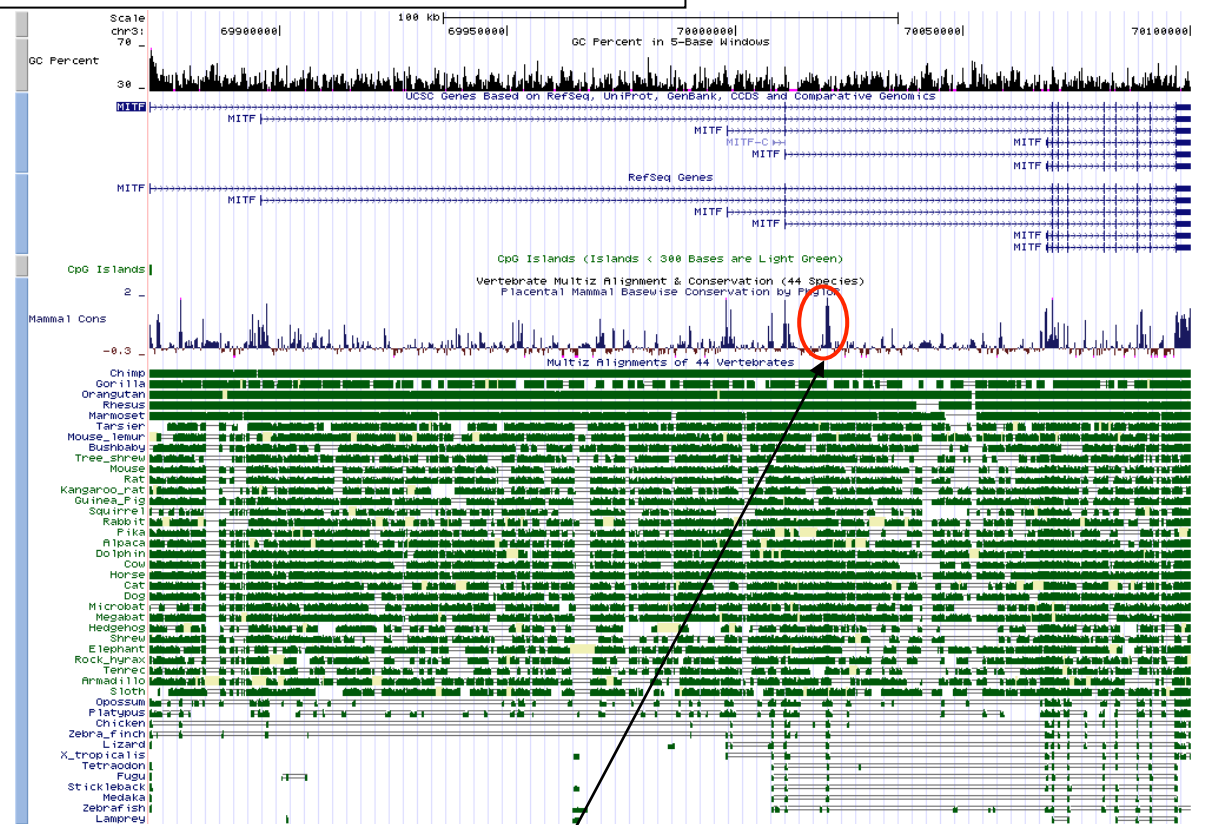
Step 3: Assuming your view does not look like this one – scroll down to change the views. Turn off any unnecessary tracks and then look at the Comparative Genomics section.



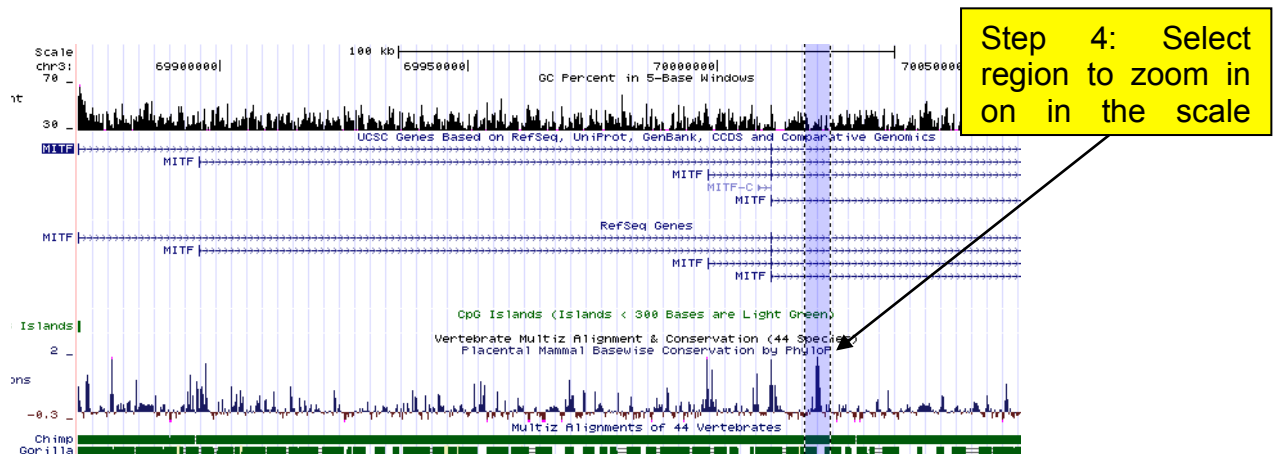
Click here to set adjust the settings of the conservation

Select all species and Multiz Alignments to full. Press submit.

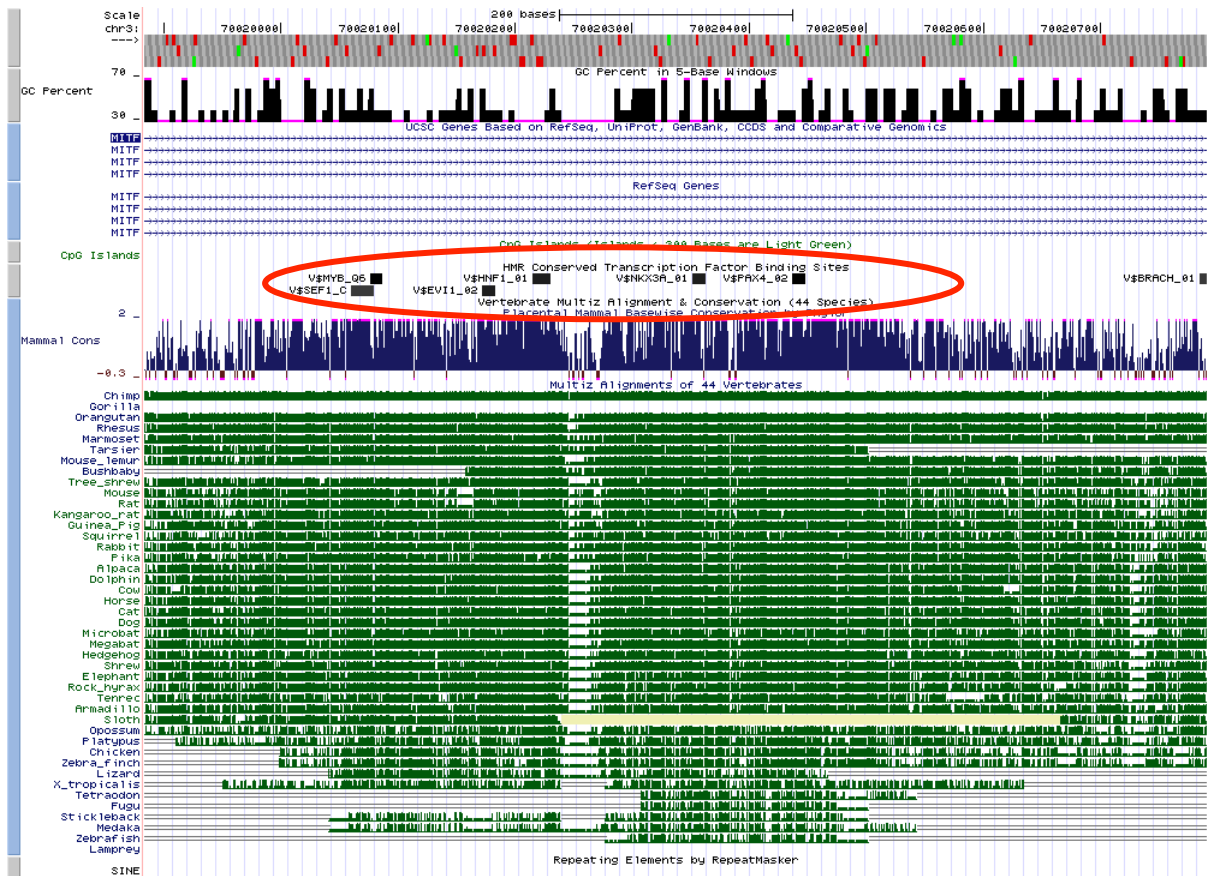
Are there any conserved non-coding regions?



Yes, there is a non-coding Evolutionary Conserved Region (ECR) in intron 2. This is conserved back to fish suggesting it is an important regulatory element.



Zoom in on the ECR. Is it associated with any transcription factor binding sites? Switch on TFBS conserved Track to “pack” (Regulation section).



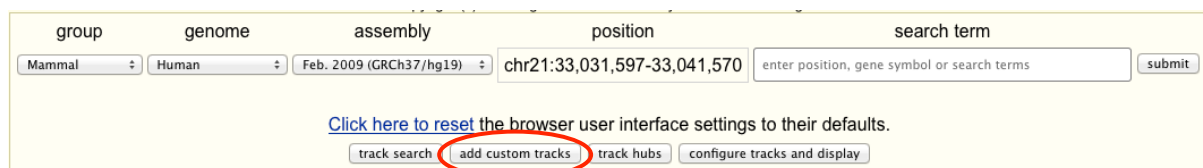
Yes, the ECR has a number of conserved transcription factor binding sites. This track shows TFBS that are conserved between human, mouse and rat.

Adding your own track to UCSC.

It is really easy to make and upload data files for your own tracks to UCSC. We are going to add a very short bed file containing putative enhancers identified next to the H19 non-coding RNA gene. The file also contains the known control region for H19 and IGF2 genes. The file is called UCSC-enhancers_track.bed.

```
browser position chr11:2000000-2030000
track name=PutativeEnhancers description="Putative_Enhancers" color=250,0,0
chr11 2013859 2014040 CS1
chr11 2012311 2012615 CS2
chr11 2011146 2011294 CS3
chr11 2009454 2009727 CS4
chr11 2001080 2001174 CS6
chr11 2021124 2023944 H19-DMR
```

This is a very basic track with only 5 regions. The format has a description line including track name, description and in this case colour. This also includes the position I am interested in in the genome. Below you need at least 3 columns chromosome, start and end. I have also included a name column. Details of all of the file types you can upload can be found on the Add Custom Tracks page



The screenshot shows the UCSC Genome Browser search interface. It includes dropdown menus for 'group' (Mammal), 'genome' (Human), and 'assembly' (Feb. 2009 (GRCh37/hg19)). The 'position' field contains 'chr21:33,031,597-33,041,570'. A search term input field is present with a 'submit' button. Below the search area, there is a link: 'Click here to reset the browser user interface settings to their defaults.' At the bottom, there are four buttons: 'track search', 'add custom tracks' (circled in red), 'track hubs', and 'configure tracks and display'.

Add Custom Tracks

clade Mammal genome Human assembly Feb. 2009 (GRCh37/hg19)

Display your own data as custom annotation tracks in the browser. Data must be formatted in [BED](#), [bigBed](#), [bedGraph](#), [GFF](#), [GTF](#), [WIG](#), [bigWig](#), [MAF](#), [BAM](#), [BED detail](#), [Personal Genome SNP](#), [VCF](#), [broadPeak](#), [narrowPeak](#), or [PSL](#) formats. To configure the display, set [track](#) and [browser](#) line attributes as described in the [User's Guide](#). Data in the bigBed, bigWig, BAM and VCF formats can be provided via only a URL or embedded in a track line in the box below. Publicly available custom tracks are listed [here](#). Examples are [here](#).

Paste URLs or data: Or upload: No file selected.

```
color=250,0,0
chr11 2013859 2014040 CS1
chr11 2012311 2012615 CS2
chr11 2011146 2011294 CS3
chr11 2009454 2009727 CS4
chr11 2001080 2001174 CS6
chr11 2021124 2023944 H19-DMR
```

Optional track documentation: Or upload: No file selected.

Click [here](#) for an HTML document template that may be used for Genome Browser track descriptions.

This is a short file so it can just be pasted in.

Manage Custom Tracks

genome: Human assembly: Feb. 2009 (GRCh37/hg19) [hg19]

Name	Description	Type	Doc	Items	Pos	delete
PutativeEnhancers	Putative_Enhancers	bed		6	chr11	<input type="checkbox"/>

Click on "chr11" to return to browser

Make sure the display is set to "full" for your custom track, otherwise it won't display the feature names.

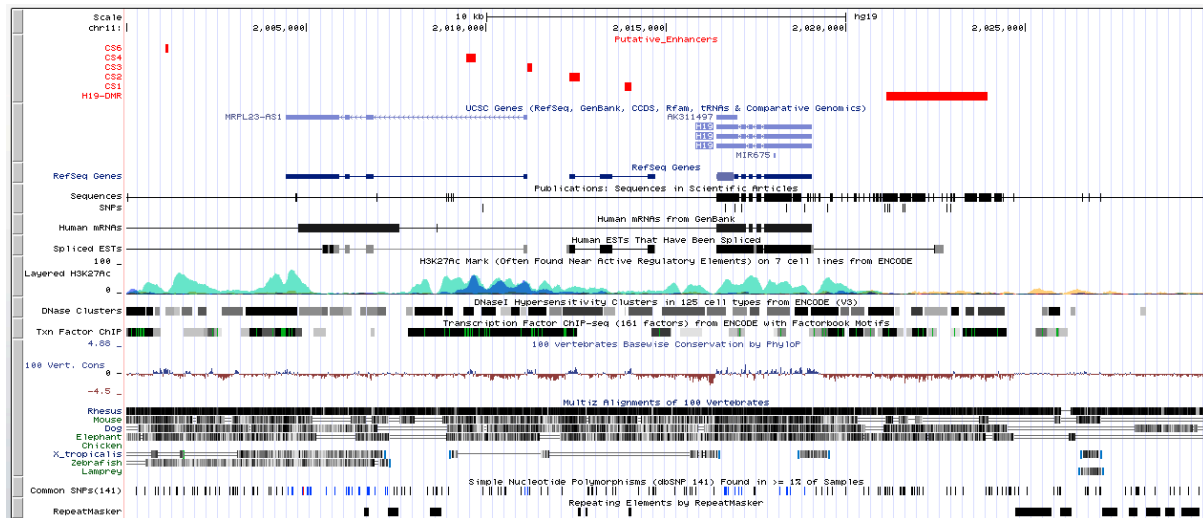
collapse all Use drop-down controls below and press refresh to alter tracks displayed. expand all
 Tracks with lots of items will automatically be displayed in more compact modes.

Custom Tracks

[PutativeEnhancers](#)
 full

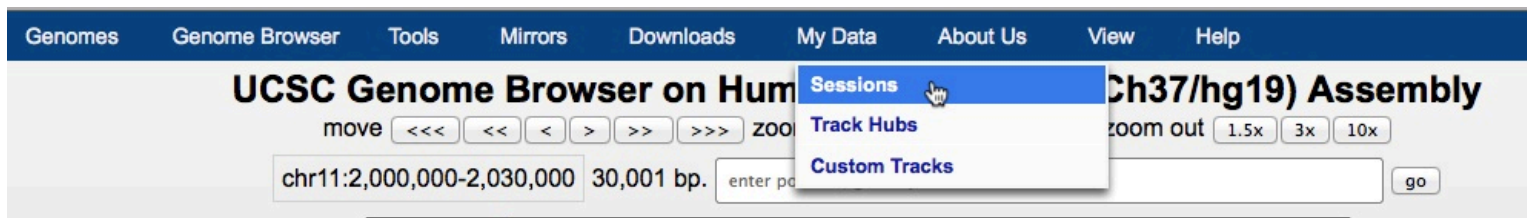
Mapping and Sequencing

Genes and Gene Predictions



These regions, which are viewed here as red bars, can be explored to see if they lie in areas of the genome that act as regulatory elements.

It is possible to save your session in UCSC so you can come back to it at a later date. Go to the Sessions page, which is under my data. On your first visit you can request an account.



Sign in to UCSC Genome Bioinformatics

[Login](#)

[Create an account](#)

Signing in enables you to save current settings into a named session, and then restore settings from the session later. If you wish, you can share named sessions with other users.

Session Management

See the [Sessions User's Guide](#) for more information about this tool.

[Click here to reset](#) the browser user interface settings to their defaults.

If you [sign in](#), you will also have the option to save named sessions.

Save Settings

Save current settings to a local file:

file: file type returned:

(leave file blank to get output in browser window)

Restore Settings

Use settings from another user's saved session:

user: session name:

Use settings from a local file:

Use settings from a URL (http://..., ftp://...):

You can then save and share your sessions.

3. Comparing genome sequences in the ECR browser and identifying potential transcription factor binding sites.

The Comparative Genomics Developments website is a very useful resource comparing genomes to decipher the code of gene regulation. It encompasses much of the historical databases for genome alignments such as VISTA, rVISTA and zPicture. In this exercise we will explore the *WWOX* gene for conserved elements that may have regulatory function. But remember, conserved elements could be anything.

Step 1: Choose the ECR Browser

Step 2: Search for the WWOX gene in human

Features

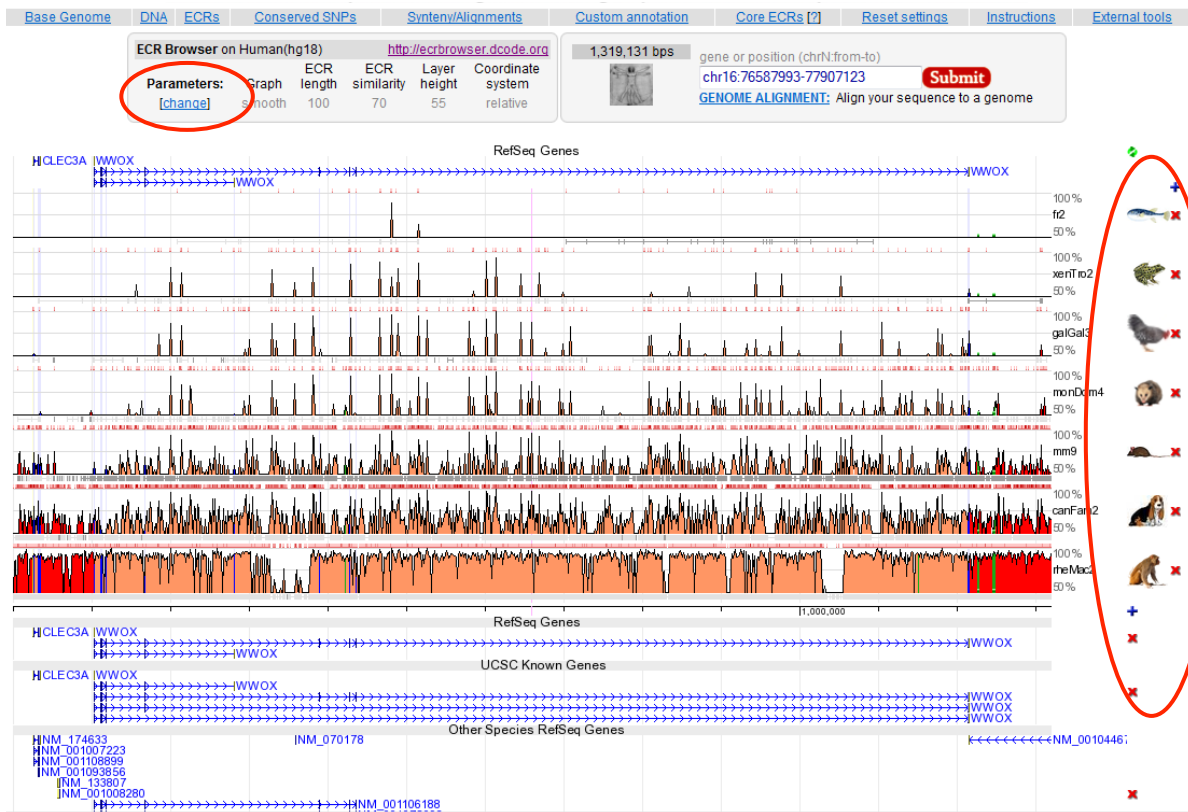
- multiple gene annotation tracks
- possibility to submit your own custom genome annotation
- share your custom annotation by submitting it to the [ECR Browser User Annotation Database](#)
- enhanced search through RefSeq, UCSC, Ensembl, mRNA, STS, SNPs, etc.
- zoom in/out using mouse wheel (similar to Google Maps)
- drag & drop rearrangement of gene annotation tracks
- drag & drop recentering of the conservation plot
- gene annotation drop on the conservation plot changes the reference annotation
- new alignments with repetitive elements included (colored in green)
- synteny annotation under each conservation track
- on-the-fly ECR and gene annotation
- single mouse-click *Grab* ECR function
- keyboard shortcuts:
 - o - zoom out 3x, i - zoom in 3x, > - shift to the right, < - shift to the left, l - flip the plot, g - genome selection window, p - parameters window, c - highlight coreECRs, r - reset parameters to defaults, f - refresh the page, a - additional alignments, m - main gene annotation, z - blast-based genome alignment

Note: ECR Browser was tested on Internet Explorer 7 and Mozilla Firefox 2. We were also told it works on both Safari and Opera. Please update your Internet browser, if you experience unexpected behavior of ECR Browser.

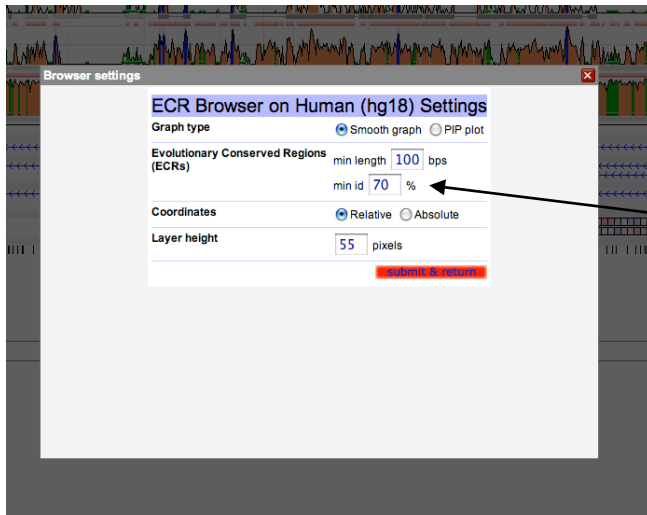
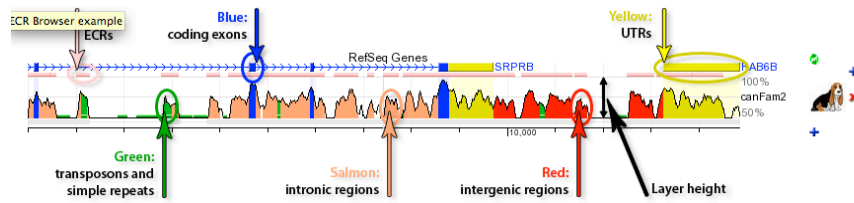
Citing ECR Browser
 I. Ovcharenko, M.A. Nobrega, G.G. Loots, and L. Stubbs, *Nucleic Acids Research*, 32, W280-W286 (2004) [\[PDF\]](#)

Step 3: Select the longest Refseq transcript for *WWOX*

RefSeq genes...	WWOX	WW domain-containing oxidoreductase isoform 3
chr16:78133327-78134096	WWOX	WW domain-containing oxidoreductase isoform 2
chr16:78133327-78132593	WWOX	WW domain-containing oxidoreductase isoform 1
UCSC genes...		
chr16:78133327-79246564	WWOX	Homo sapiens WW domain containing oxidoreductase (WWOX), transcript variant 1, mRNA
chr16:78133327-78312593	WWOX	Homo sapiens WW domain containing oxidoreductase (WWOX), transcript variant 2, mRNA
chr16:78133327-78134096	WWOX	Homo sapiens WW domain containing oxidoreductase (WWOX), transcript variant 3, mRNA
chr16:78133327-79246564	WWOX	Homo sapiens WW domain containing oxidoreductase (WWOX), transcript variant 1, mRNA
chr16:78133327-79246564	WWOX	Homo sapiens WW domain containing oxidoreductase (WWOX), transcript variant 2, mRNA
chr16:78133327-79246564	WWOX	Homo sapiens WW domain containing oxidoreductase (WWOX), transcript variant 1, mRNA
MGC genes...		
chr16:78133597-79246564	BC003184	WWOX



Step 4: The database is based on original VISTA plots and contains gene information imported from the UCSC. Be careful because it may not be reflecting the same genome build. On the right hand side, you can remove or add organisms (add cow and zebrafish). Click instructions at the top to get the helpful key shown below. Change the parameters using the [change] function.



Step 5: Change the minimum identity (min id) to 85% to filter low identity ECRs.



Step 6: There is one very highly conserved ECR in intron 8 (the large final intron). It's easier to see if you zoom out first! Zoom in a bit and you may have to re-centre. Then mouse over the ECR on the fugu track. This will bring up an alignment between the two species.

If you find it tricky to get to this ECR, simply open ECR browser from the following co-ordinates:
chr16:78464965-78550000

ECR :: Evolutionary Conserved Region

location: chr16:78511090-78511441
length: 352 bps
identity: 88.1%

Alignment

overlapping alignment block:
chr16:78511090-78511465 -vs- fr3:chr9:2566448-2566823

[conserved transcription factor binding sites \(TFBS\)](#)

2566823	2566803	2566783
fr3	GCTGATGGACTATTTTCAAATTGATTTCAAATAATGACATAAGCAGCCGATCCOCTAGC	
hg19	GTGATGGACTATTTTCAAATTGATTTCAAATAATGATTTAAGGGGATGACTTCTAGT	
78511090	78511110	78511130
2566763	2566743	2566723
fr3	CCAGATTACCTATTGATTTTAAATAGAAAAGCTCATTATATAAGCAGGAACGGCAACAC	
hg19	CTAGATTACCTATTGATTTTAAATAGAAAAGCTCATTATATAAGCAGTAACCCG-ATAT	
78511150	78511170	78511190
2566703	2566683	2566663
fr3	AAAAGCTAGCCCAACCTTTGCATAAATCCITTAATGAATTTCCAGAGCCCGTGGTCTTA	
hg19	AAAACCTAGCAACCTTTGCATAAATCCITTAATGAATTTCCAGAGCCTGTGGTCTTA	
78511209	78511229	78511249
2566643	2566625	2566605
fr3	C--ATTTTTTAATTAATCCATTTCTTTTTTAAGGGTTGCTGTGTAATTTGCAITGCTGT	
hg19	CTTTTTTTTAATTAATCTATTTCTTTTTTAAGTGTACTGTGTAATTTGCAITGCTGT	
78511269	78511289	78511309
2566585	2566565	2566545
fr3	GAAGTGGGTGCTGTCCAGATAAAGTGCCATTGATCCTTATTAGGCTCACCTCTGGGCT	

Step 7. From here, you can go directly to rVISTA to assess known transcription factor binding sites.

RVISTA DOWNLOADING

Blastz alignment... ok
Fetching annotation files... ok

Defining transcription factor binding sites

TRANSFAC professional V10.2 library

Biological species

- vertebrates
- plants
- nematodes
- insects
- fungi
- bacteria

Matrix similarity

- Optimized for function
- Predefined as

Matrix selection

- use only high-specificity matrices

User-defined consensus sequences

SUBMIT

Step 8. You are now in rVISTA, but importantly, all the parameters have been filled in for you, and all you need to do is select the TFBS of interest.

REGULATORY VISTA

RVISTA | VISUALIZATION & CLUSTERING

Picture
 Bases per layer: 0.5kb
 Picture width (in pixels): 800
 Smooth plot

Clustering
 Individual clustering
 Combinatorial clustering
 1 site(s) per 100 bps

flip
SUBMIT

Show
 conserved
 aligned
 all

Select TF Subset Summary page List clustered TFBS

Step 11: Redraw to view 0.5 kb per line.



Can we find any other functional information about this ECR?

ECR :: Evolutionary Conserved Region

ECR [Evolutionary Conserved Region]
 Location: chr16:78511090-78511441
 length: 352 bps
 identity: 88.1%

Close

Step 15: Go to the Vista Enhancer browser page and paste in the coordinates from the UCSC website.

VISTA Enhancer Browser
whole genome enhancer browser

Home | Browser Handbook and Methods | Experimental Data | Advanced Search | Gallery | Contact

The **VISTA Enhancer Browser** is a central resource for experimentally validated human and mouse 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 in other vertebrates or epigenomic evidence (ChIP-Seq) of putative enhancer marks. The results of this *in vivo* 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. We invite external groups to [submit requests](#) for candidate enhancers to be tested at this single developmental time-point.

As of **9/5/2013** the database contains information on **1951** *in vivo* tested elements - **1056** elements with enhancer activity. Lawrence Berkeley National Laboratory and OpenHelix announce [comprehensive training programs for VISTA](#).

Keyword Search

Both (Human & mouse enhancers)
 Human only (hg19 coordinates)
 Mouse only (mm9 coordinates)

chr16:78511090-78511441

Examples: gene, accession number, locus link, genomic position (e.g. chr1:3000000-5000000)

[Advanced Search](#)

Enhancer Discovery Strategies

ChIP-seq from tissues | Comparative Analysis

Mouse Egg Microinjection | E11.5 Reporter Staining

VISTA Enhancer Browser
whole genome enhancer browser

Home | Browser Handbook and Methods | Experimental Data | Advanced Search | Gallery | Contact

(i) Hyperlinks indicate coordinates in species of DNA origin. Non-hyperlinked coordinates indicate orthologous region in respective other species.

[Download Data](#) 1 element(s). Elements per page: 50 100 500 All

ID	Human (hg19)	Mouse (mm9)	Expression	Section
	Coordinates	Coordinates		
	Bracketing Genes	Bracketing Genes		
hs12	chr16:78,510,608-78,511,944 WVOX-MAF	chr8:117,268,335-117,269,838 Wwox(intragenic)		

Home | Browser Handbook and Methods | Experimental Data | Advanced Search | Gallery | Contact

Step 16: Click on location. The blue coloured mouse indicates that *in vivo* analysis has been performed on this ECR

Human element [hs12]

Position: chr16:78,510,608-78,511,944 ([UCSC browser](#))

Source: Lawrence Berkeley National Laboratory

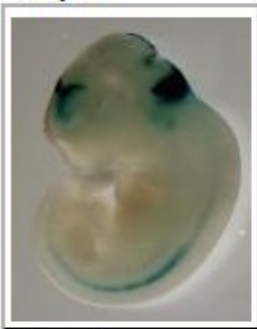
Flanking genes: [WWOX](#) - [MAF](#)

Expression Pattern

forebrain (9 out of 11 embryos)

hindbrain (rhombencephalon) (9 out of 11 embryos)

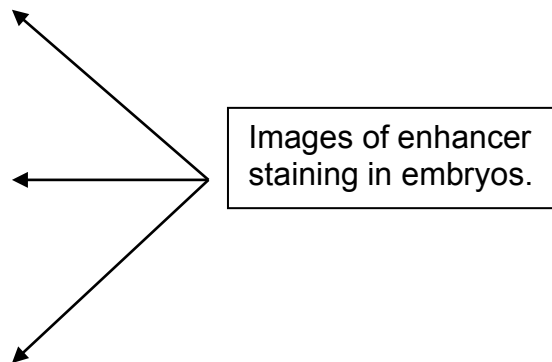
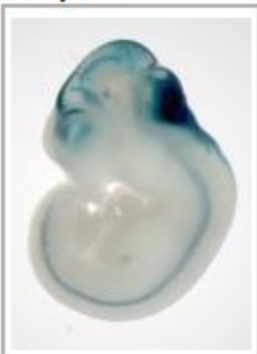
Embryo 1



Embryo 3



Embryo 7



Tasks:

1. Ensembl:

- a. Identify the *Bax* gene in Zebrafish – hint: you may not be able find it by searching in the zebrafish Ensembl.
- b. Can you identify any predicted paralogues
- c. View the orthologues for the zebrafish *Bax* gene
- d. View the Gene tree view for *Bax*

2. UCSC genome browser:

- a. Find the *ARID1A* gene in the genome.
- b. Are there any non-coding ECRs within or close to this gene?
- c. A recent paper has identified putative heart enhancers in developing mouse heart via p300 ChIP (Blow *et al.* Nat Genet. 2010 September; 42(9): 806–810). On your computer is a file of the coordinates of these putative elements (p300_Heart). Read this file in to UCSC and see if any of the peaks coincide with the regions identified above.

3. ECR browser. Using the Human *ZAK* gene:

- a. View the gene in the ECR browser
- b. Can you find non-coding ECRs?
- c. Set up rVISTA analysis between human and mouse on the most conserved ECR. Also try looking at the conserved TFBS for this ECR at UCSC
- d. Is there any evidence for this being a regulatory element in the Vista Enhancer browser?

Answers:

1. Ensembl:

- a. Identify the *Bax* gene in Zebrafish – hint: you may not be able find it by searching in the zebrafish Ensembl.

Search for the human *BAX* gene and view the orthologues. There will be a zebrafish orthologue called *baxa*. Interestingly there are two copies of *BAX* in the zebrafish genome - it appears to have been tandemly duplicated *baxa* and *baxb*.

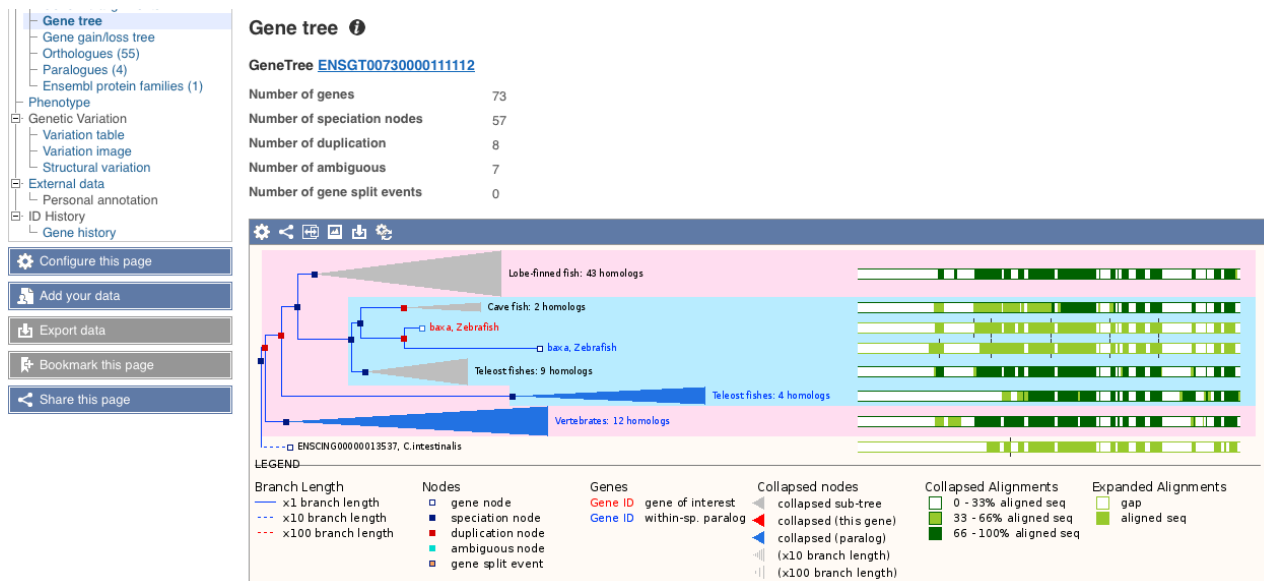
- b. Can you identify any predicted paralogues for *baxa*?

Type	Ancestral taxonomy	Ensembl identifier & gene name	Compare	Location	Target %id	Query %id
Paralogue (within species)	Danio rerio	ENSDARG00000089129 <i>baxa</i> bcl2-associated X protein, a [Source:ZFIN;Acc:ZDB-GENE-000511-6]	<ul style="list-style-type: none"> Region Comparison Alignment (protein) Alignment (cDNA) 	3:37761255-37768582:-1	50	52
Paralogue (within species)	Euteleostomi	ENSDARG00000030881 <i>baxb</i> bcl2-associated X protein, b [Source:ZFIN;Acc:ZDB-GENE-050227-21]	<ul style="list-style-type: none"> Region Comparison Alignment (protein) Alignment (cDNA) 	3:32211782-32216892:-1	20	22
Paralogue (within species)	Euteleostomi	ENSDARG00000089995 BX511080.2 Uncharacterized protein [Source:UniProtKB/TrEMBL (E7F560)]	<ul style="list-style-type: none"> Region Comparison Alignment (protein) Alignment (cDNA) 	3:32204378-32208411:-1	19	16
Paralogue (within species)	Euteleostomi	ENSDARG00000068102 zgc:153993 zgc:153993 [Source:ZFIN;Acc:ZDB-GENE-060929-176]	<ul style="list-style-type: none"> Region Comparison Alignment (protein) Alignment (cDNA) 	7:54755948-54779600:-1	32	34

- c. View the orthologues for the zebrafish *Bax* gene

There are 55 orthologues

d. View the Gene tree image for *Bax*

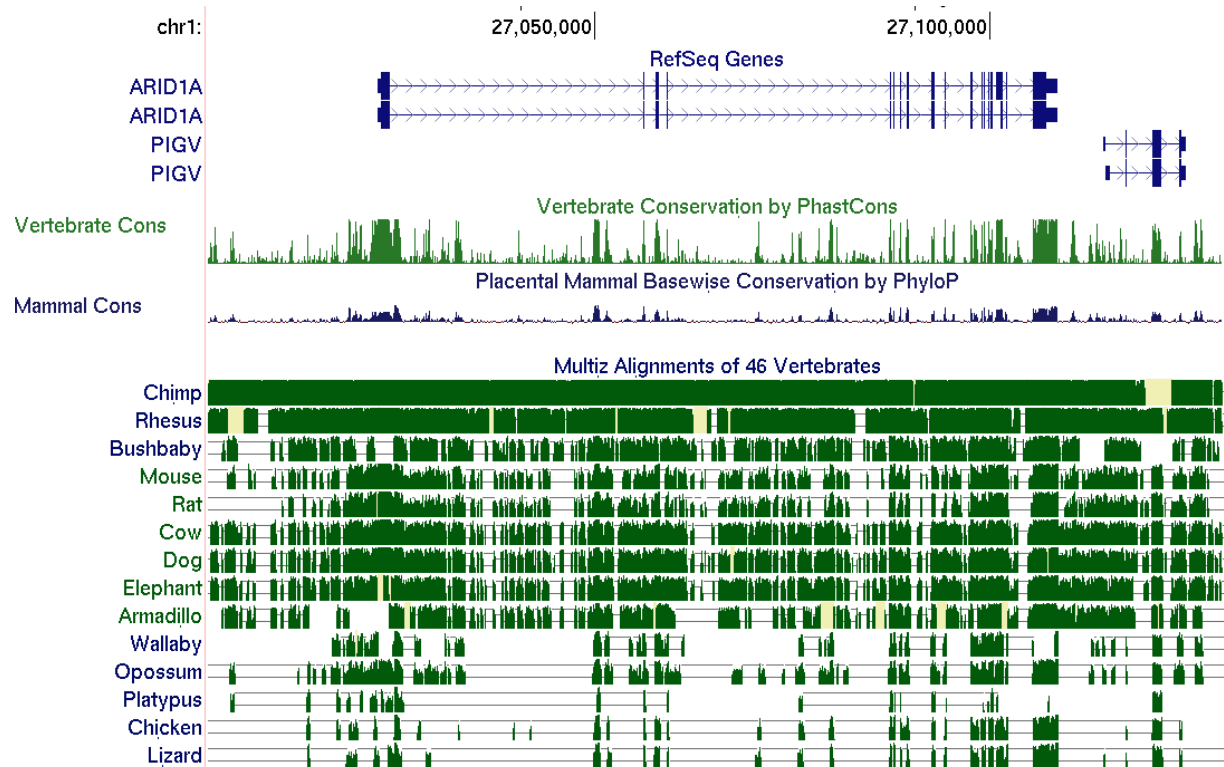


You can expand the image to view the full tree, and alter the display parameters under configure this page.

2. UCSC genome browser

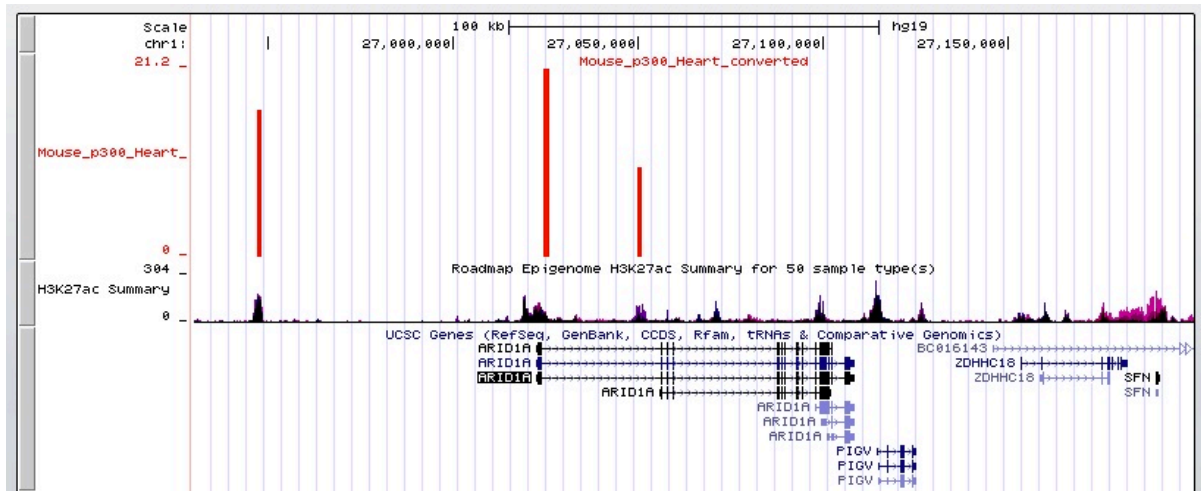
a. Find the *ARID1A* gene in the human genome.*ARID1A* is found at chr1:27022522-27108601

b. Are there any non-coding ECRs within or close to this gene?



Yes there are a number of non exonic ECRs

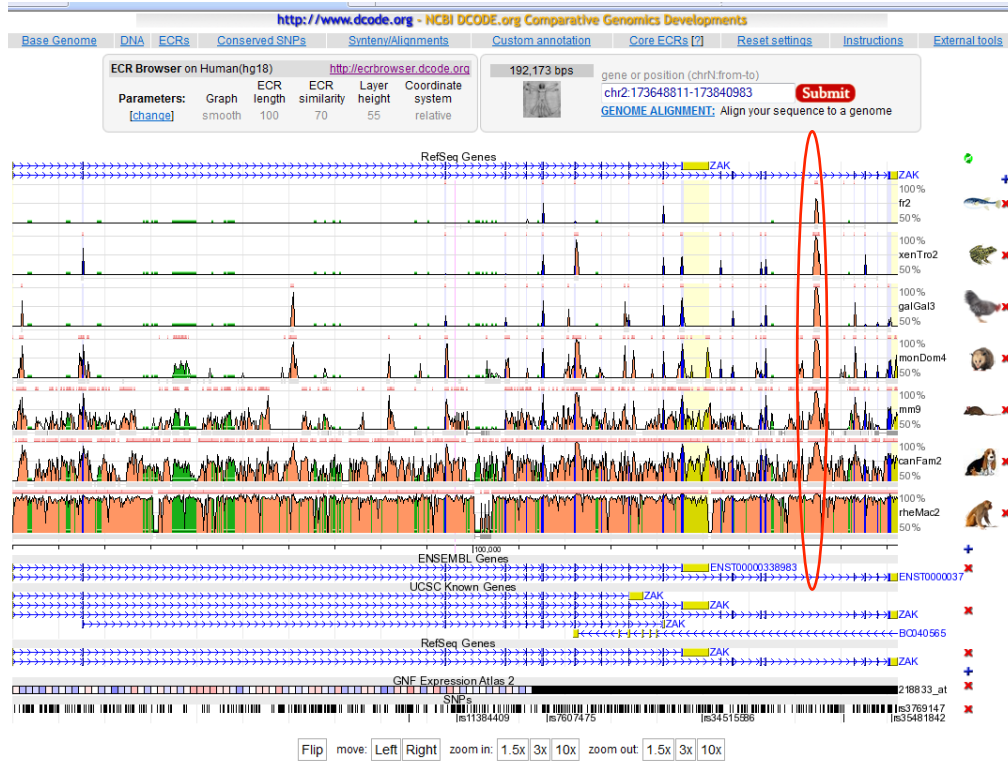
c. A recent paper has identified putative heart enhancers in developing mouse heart via p300 ChIP (Blow et al Nat Genet. 2010 September; 42(9): 806–810). On your computer is a file of the coordinates of these putative elements (p300_Heart). Read this file in to UCSC and see if any of the peaks coincide with the regions identified above.



The file contained 3 peaks all of which correspond to heart H3K27ac peaks in heart derived tissues. Together these data suggest that these regions act as enhancers in the mammalian heart.

3. ECR browser. Using either the Human ZAK gene:

a. View the gene in the ECR browser



b. Can you find non-coding ECRs?

There are a couple of non-coding ECRs that may be real. There is a really well conserved ECR in intron 14 of ZAK.

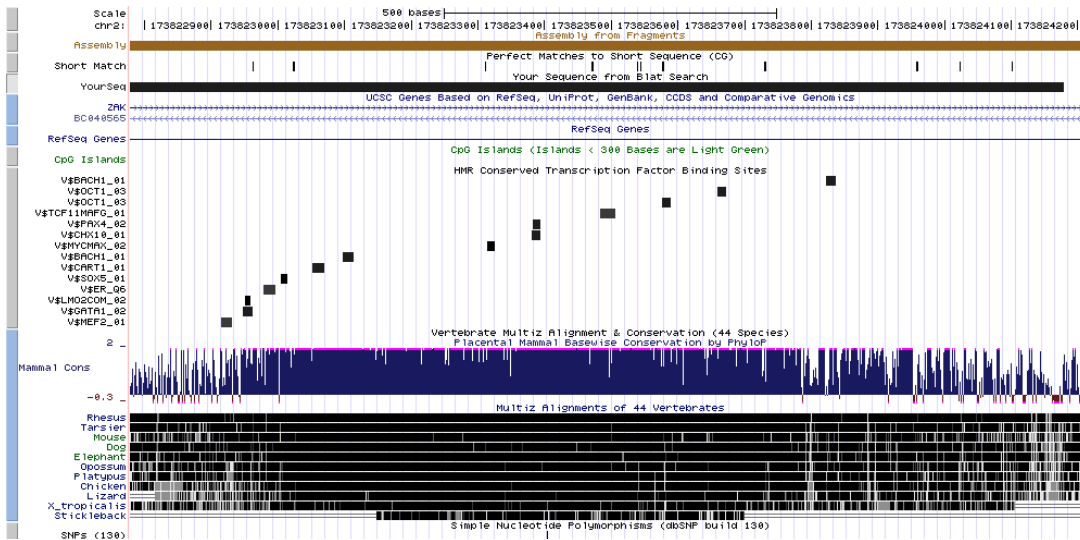
c. Set up rVISTA analysis between human and mouse



244 conserved transcription factor binding sites were identified between human and mouse.

This is quite meaningless; it is more informative to see which TFBS are conserved between more than 2 species.

You can also look at the region in UCSC and see which conserved transcription factor binding sites were identified between human, mouse and rat



- d. Is there any evidence for this being a regulatory element in the Vista Enhancer browser?

Yes - there is evidence for enhancer activity in the forebrain, midbrain and neural tube.

