
Comparative Genomics

Hendrik-Jan Megens



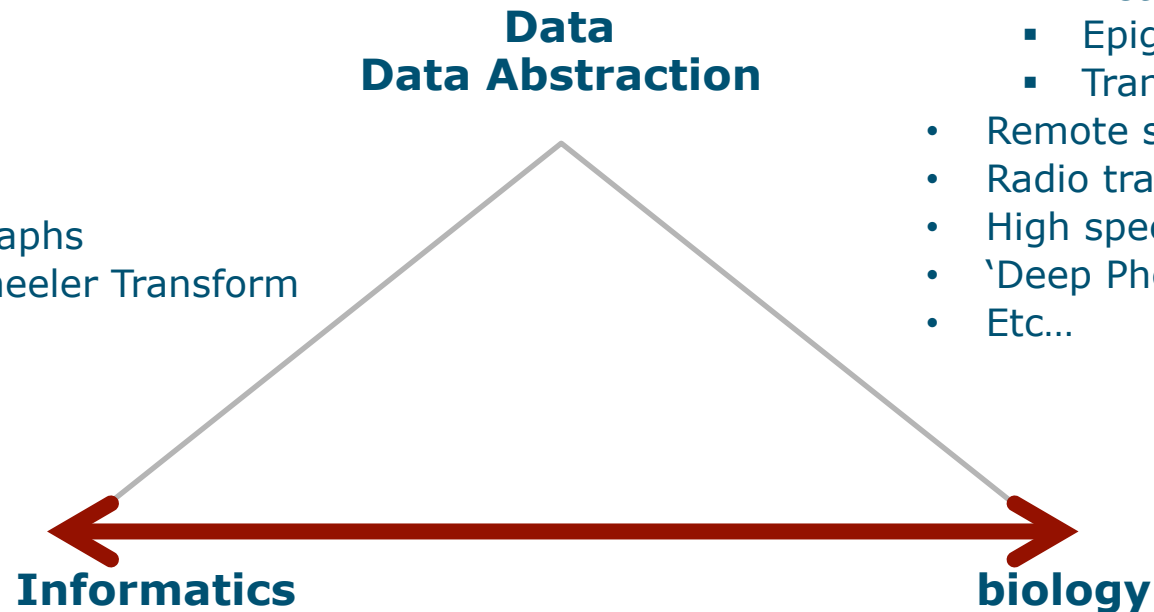
A word on 'bioinformatics'....

Information technology

- Hardware
 - Cpu
 - RAM
 - HD/Storage
- Databases
- Internet
- Cloud
- Algorithms
 - De Bruin Graphs
 - Burrows-Wheeler Transform
 - Etc..
- Etc...

Sensor technology

- Sequencing
 - De novo assembly
 - Variant Calling
 - Metagenomics
 - Epigenetics
 - Transcriptomics
- Remote sensing
- Radio tracking
- High speed cameras
- 'Deep Phenotyping'
- Etc...



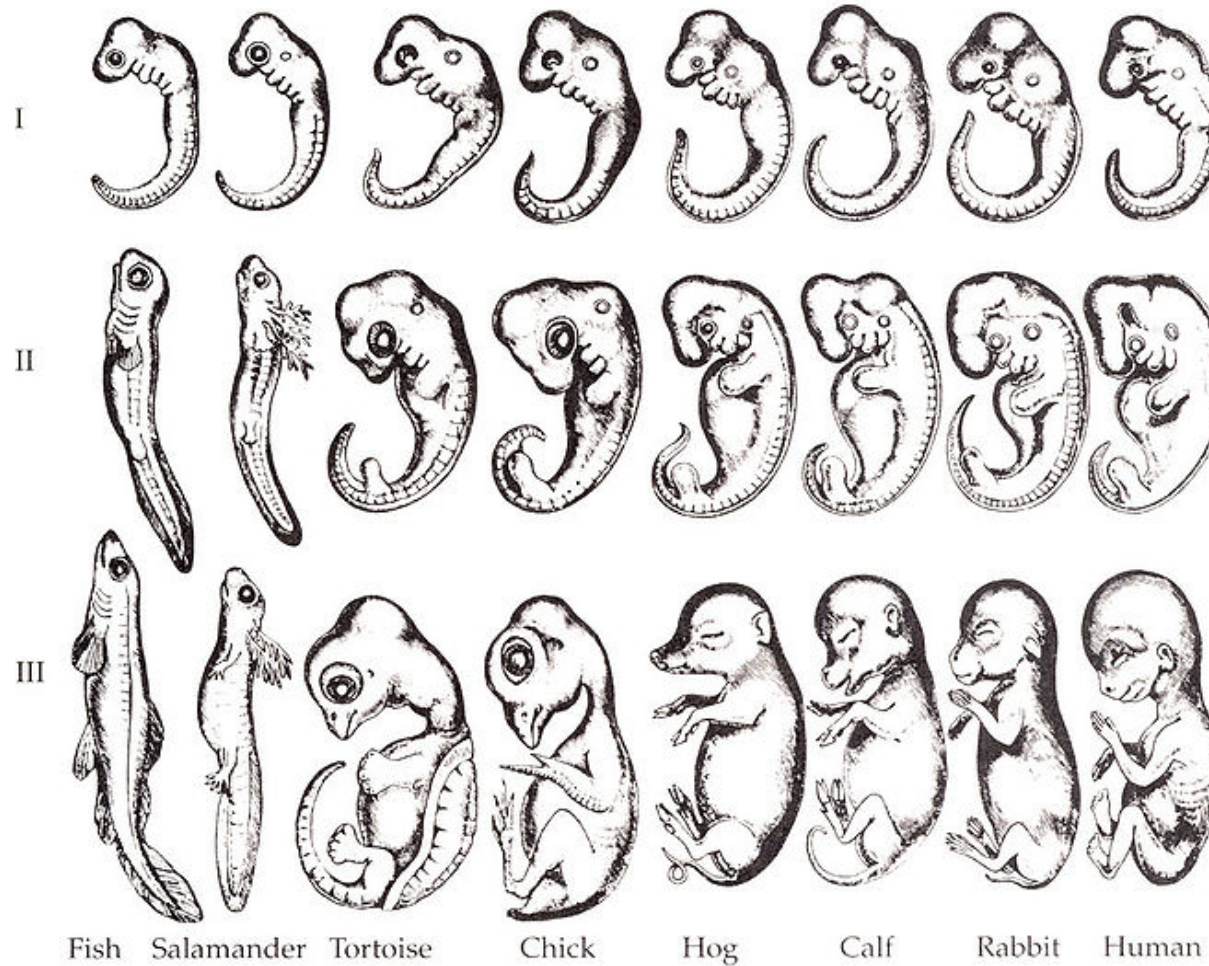
Bioinformatics MacGyverism....



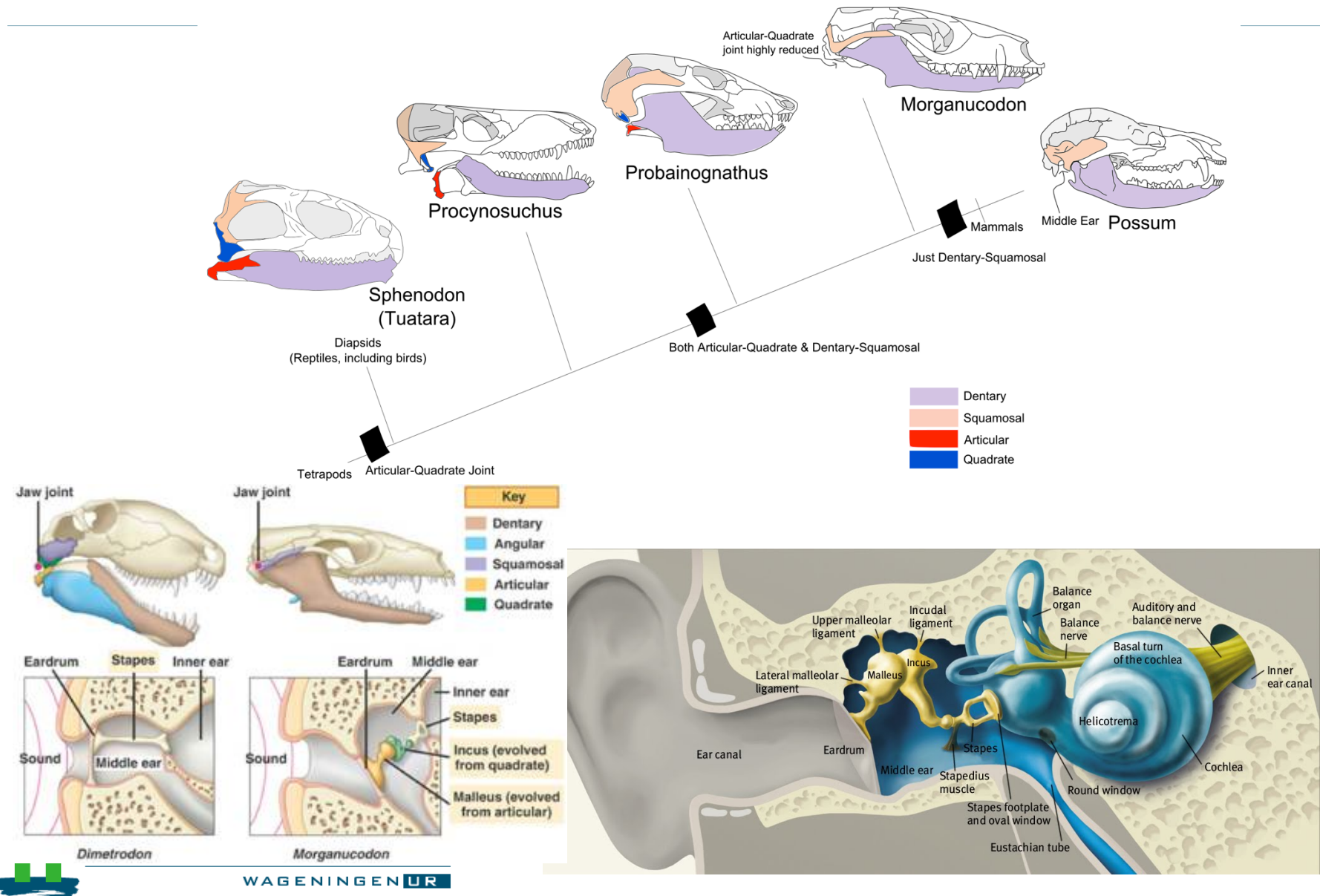
Bioinformatics MacGyverism....



Comparative biology



Comparative biology: an earful of jaw

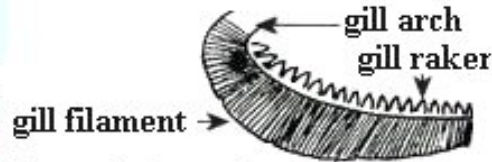
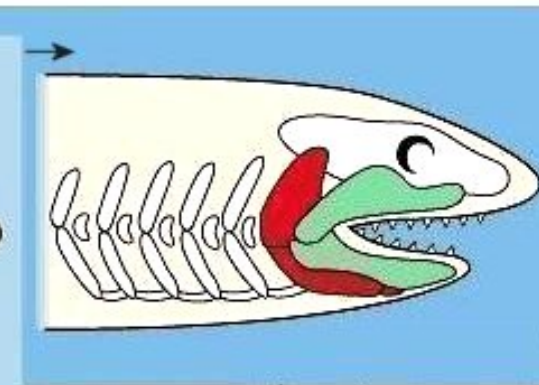
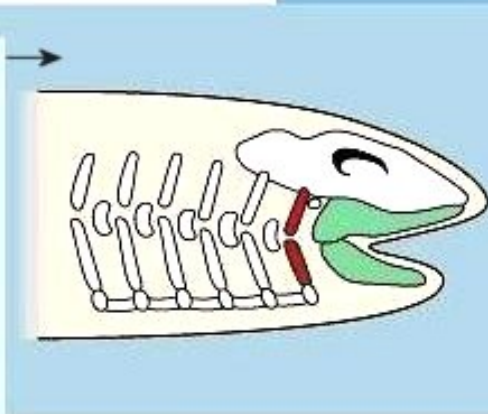
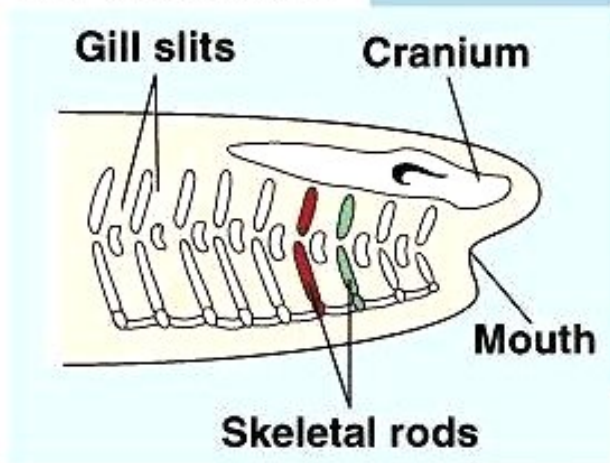


Comparative biology: a jaw full of gill

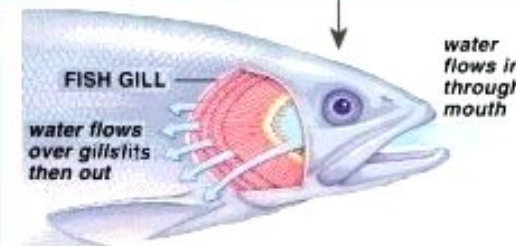
Food filters were modified to form gills

First gill arch became upper and lower jaws

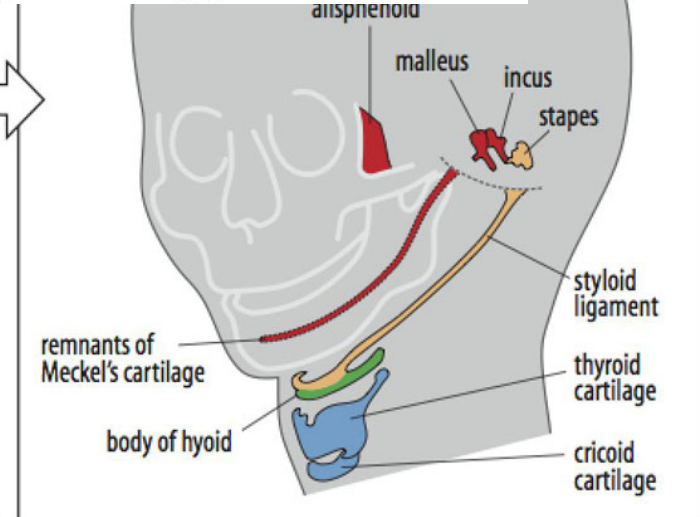
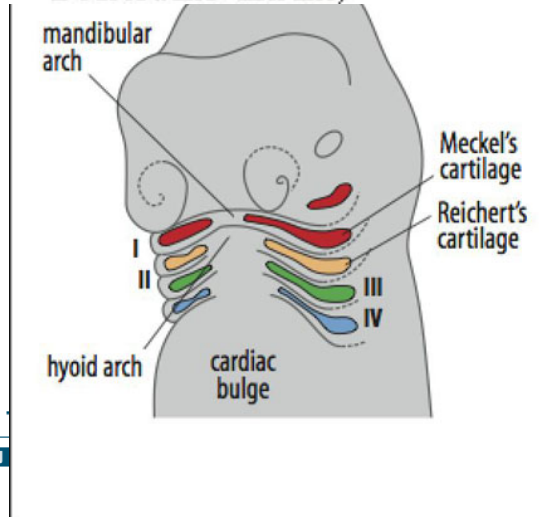
Second gill arch moved forward to brace jaws



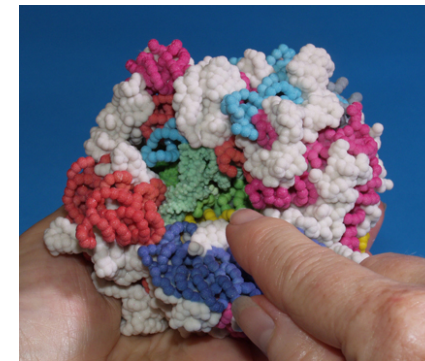
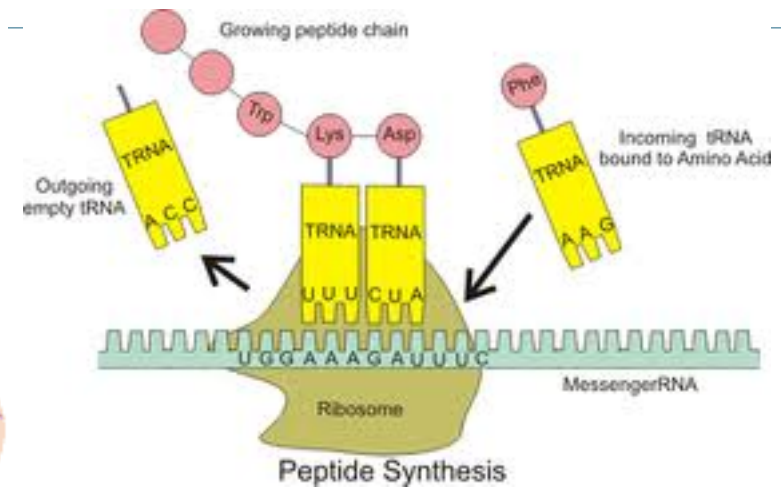
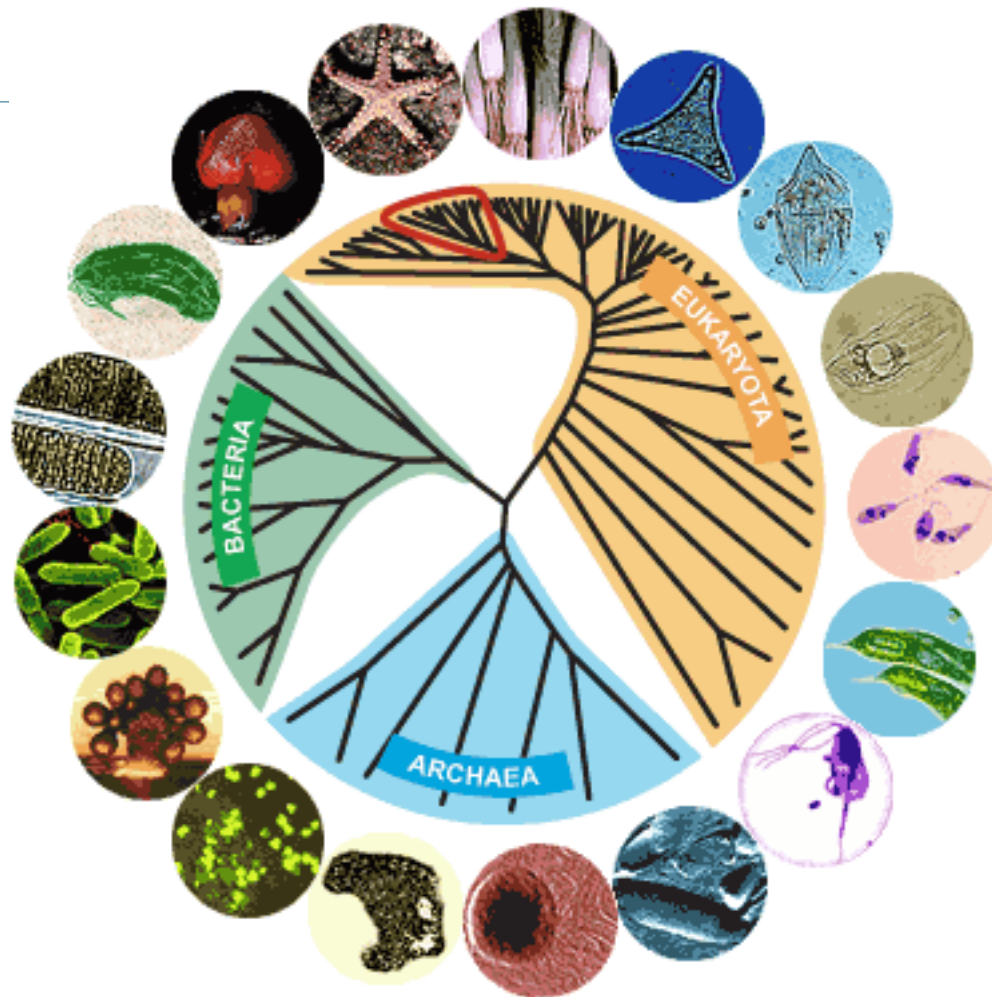
gill filament
(filament collapse when it is out of water - fish dies)



remaining gills in modern fish

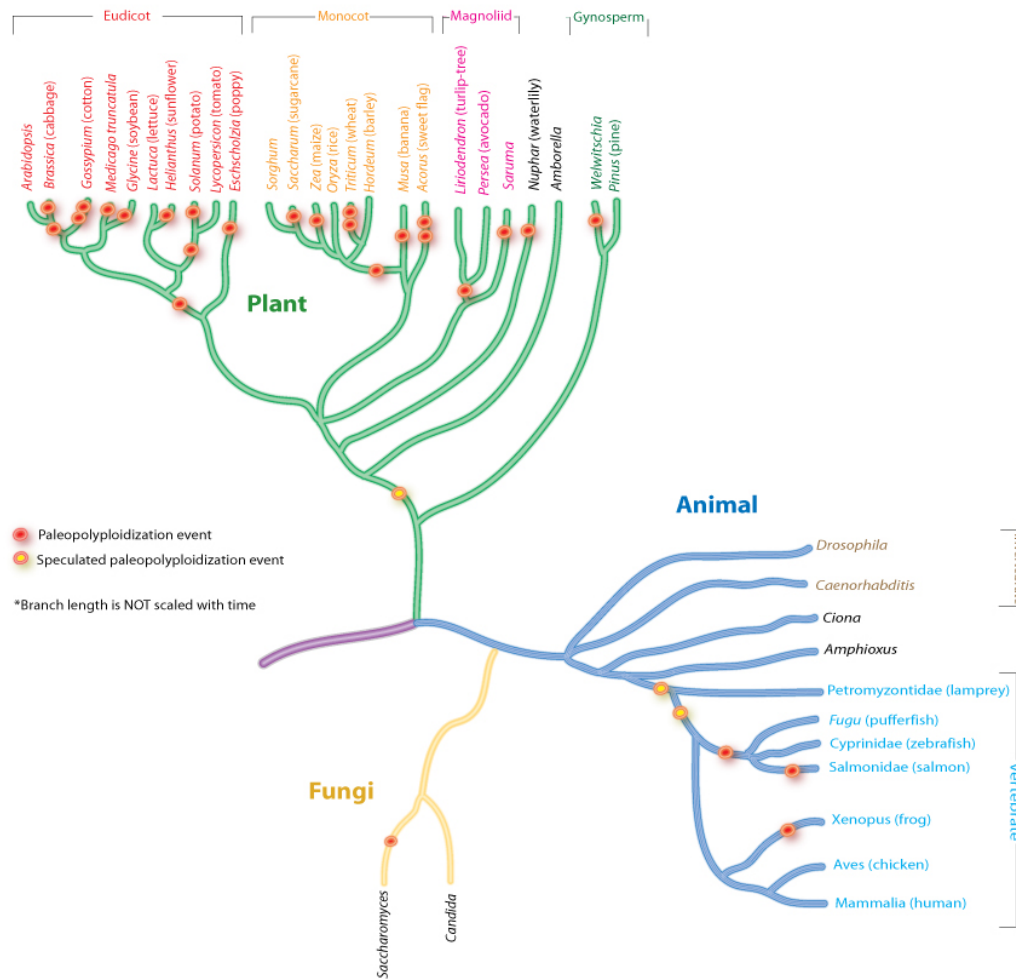


The Tree of Life: comparative biology of rDNA

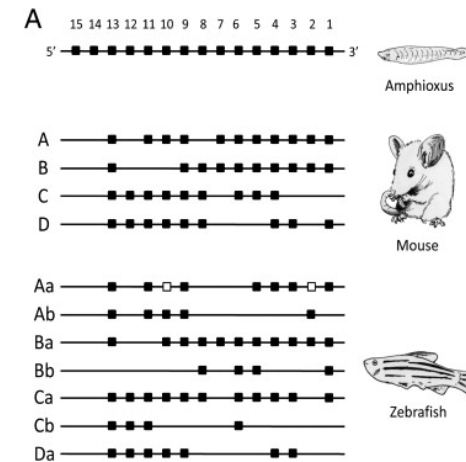
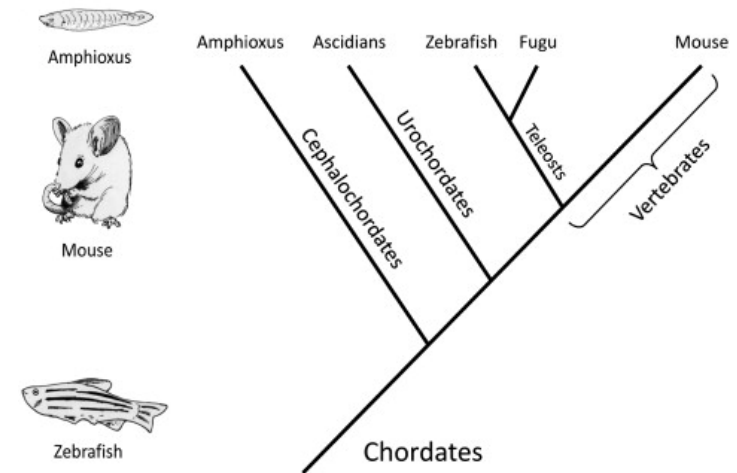


Virtually all Eukaryotes show evidence of paleo-duplication

Known Paleopolyploidy in Eukaryotes

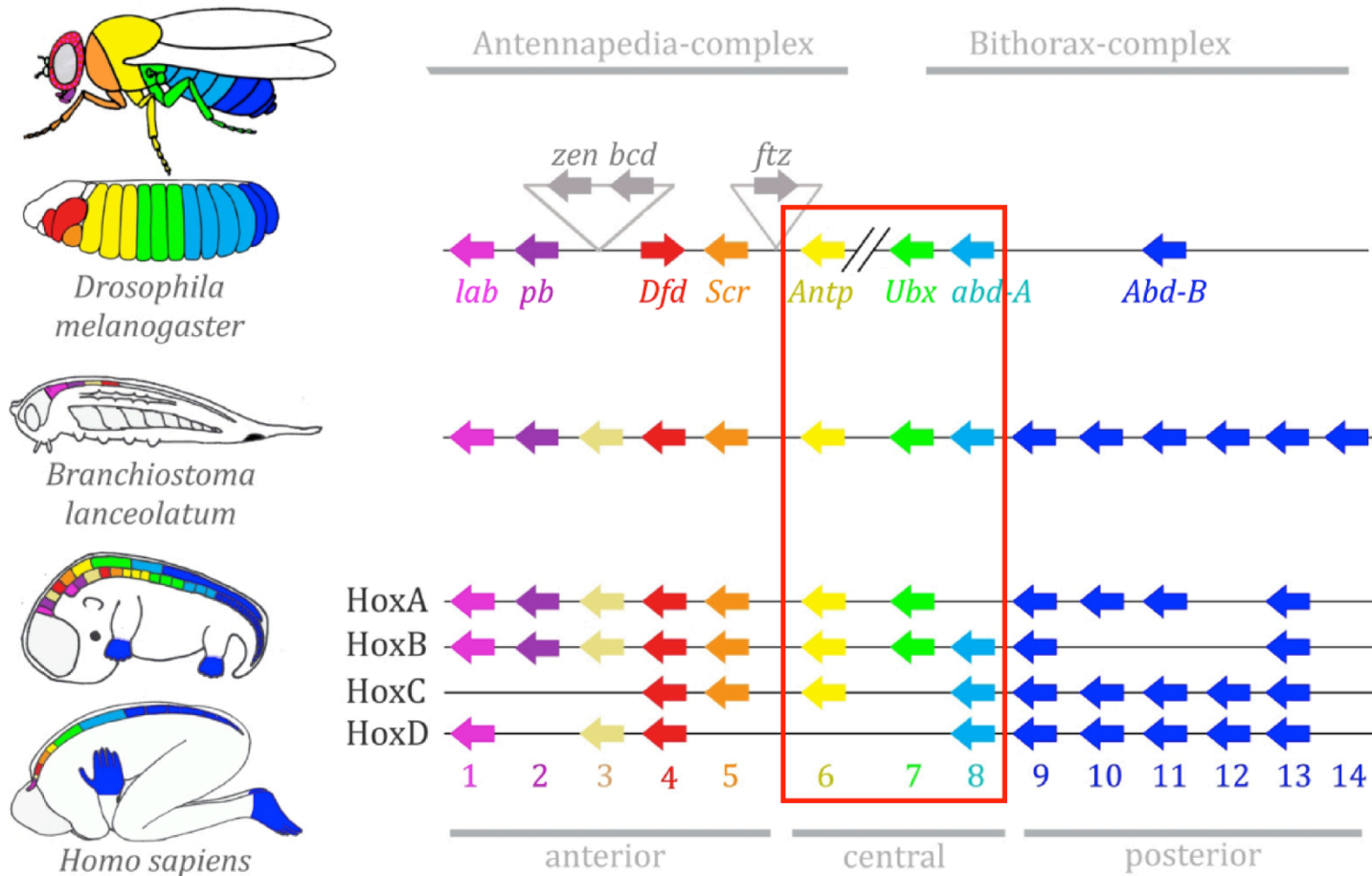


B

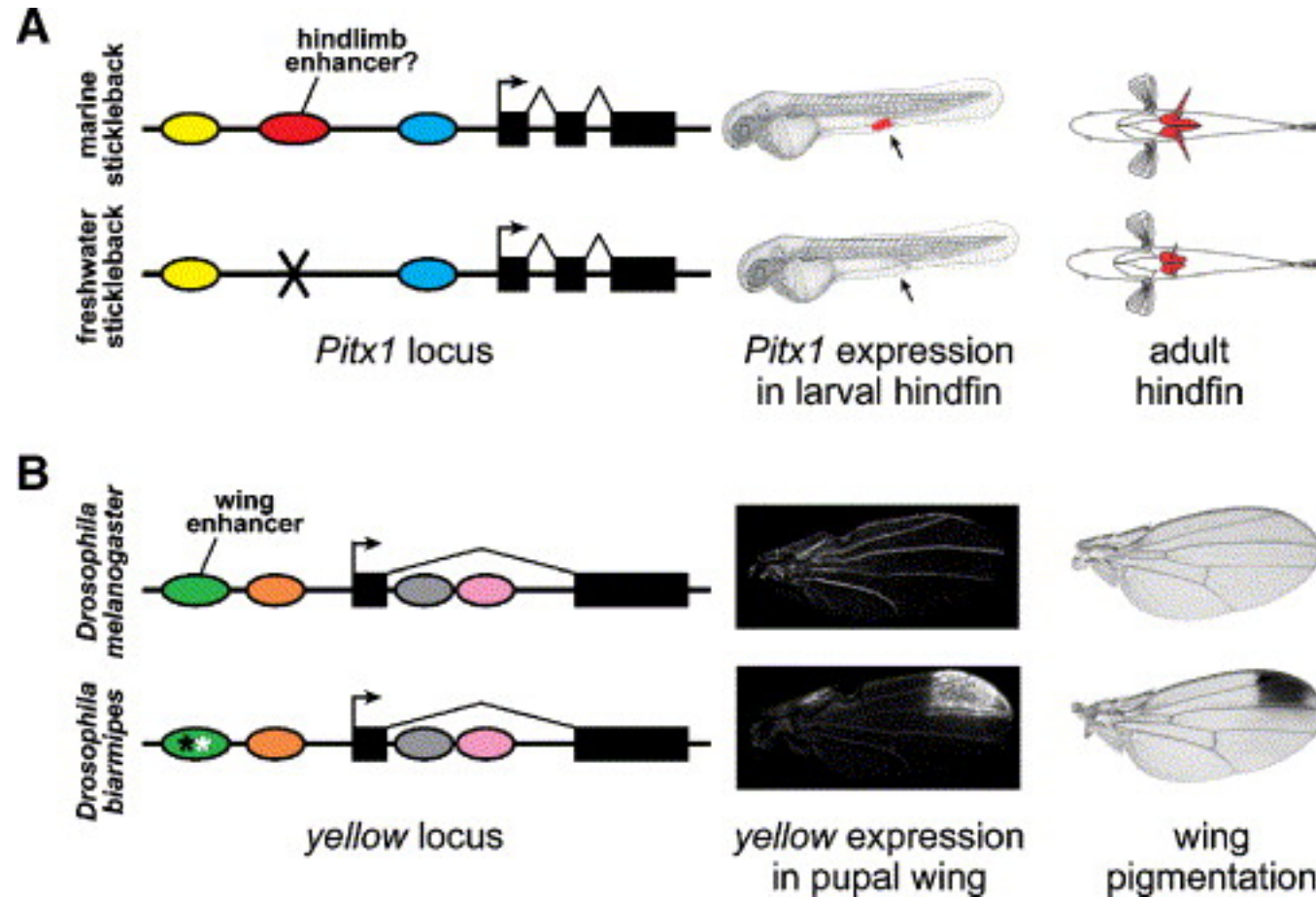


Despite re-diploidization, certain genes may retain duplicated nature (e.g. Hox genes in vertebrates).

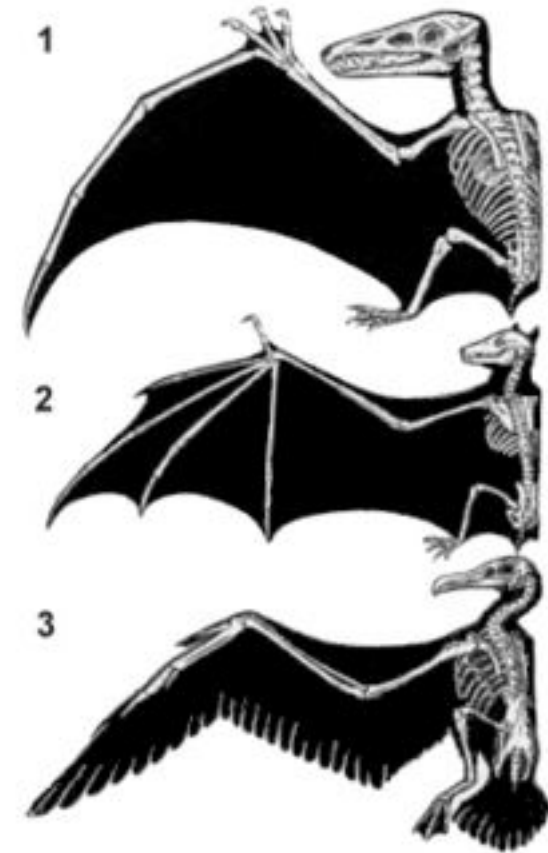
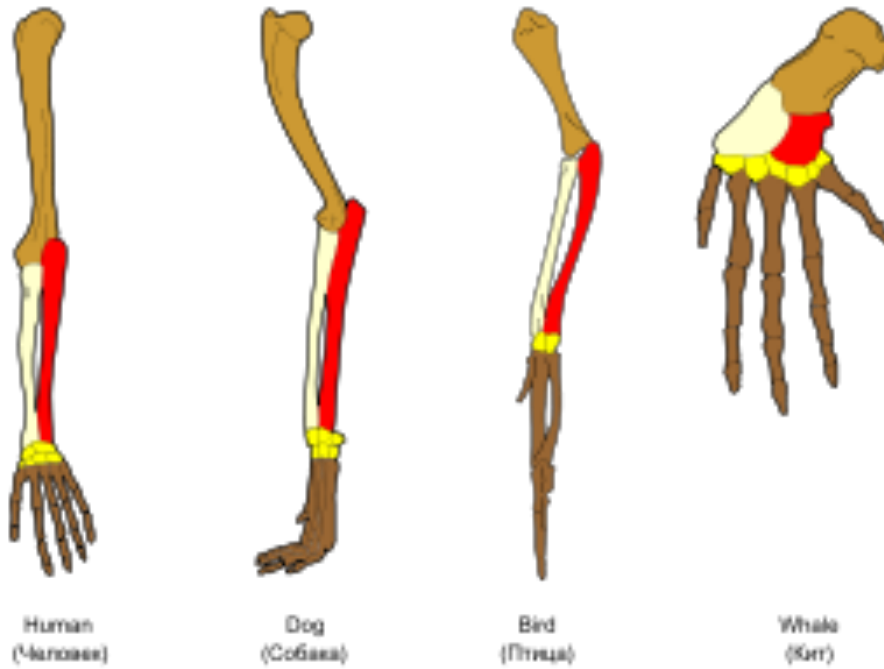
Segmentation is 'homologous' at a molecular level between arthropods and vertebrates



Novel 'switches' → novel functions



Comparative Biology: homology

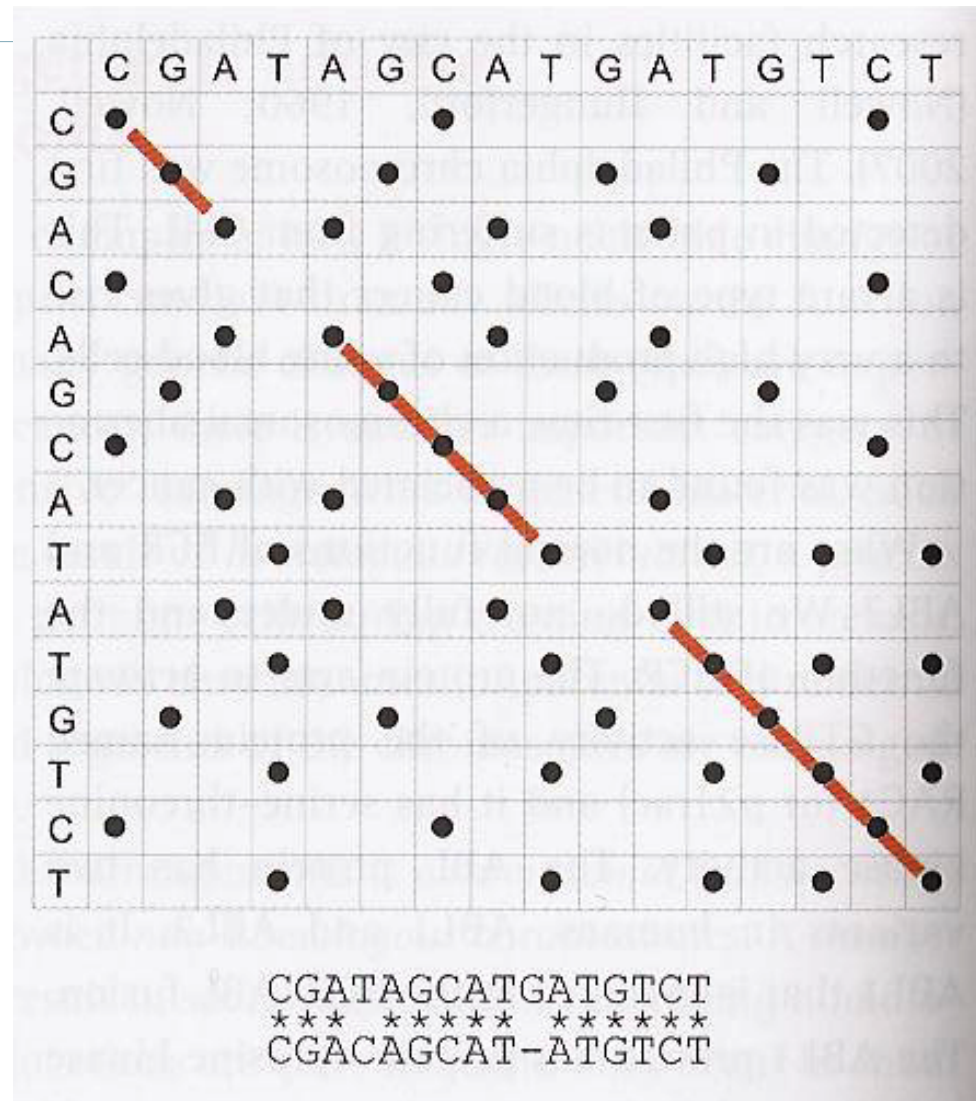


Comparative genomics: sequence alignment

- Sequence homology: entirely defined by sequence similarity.

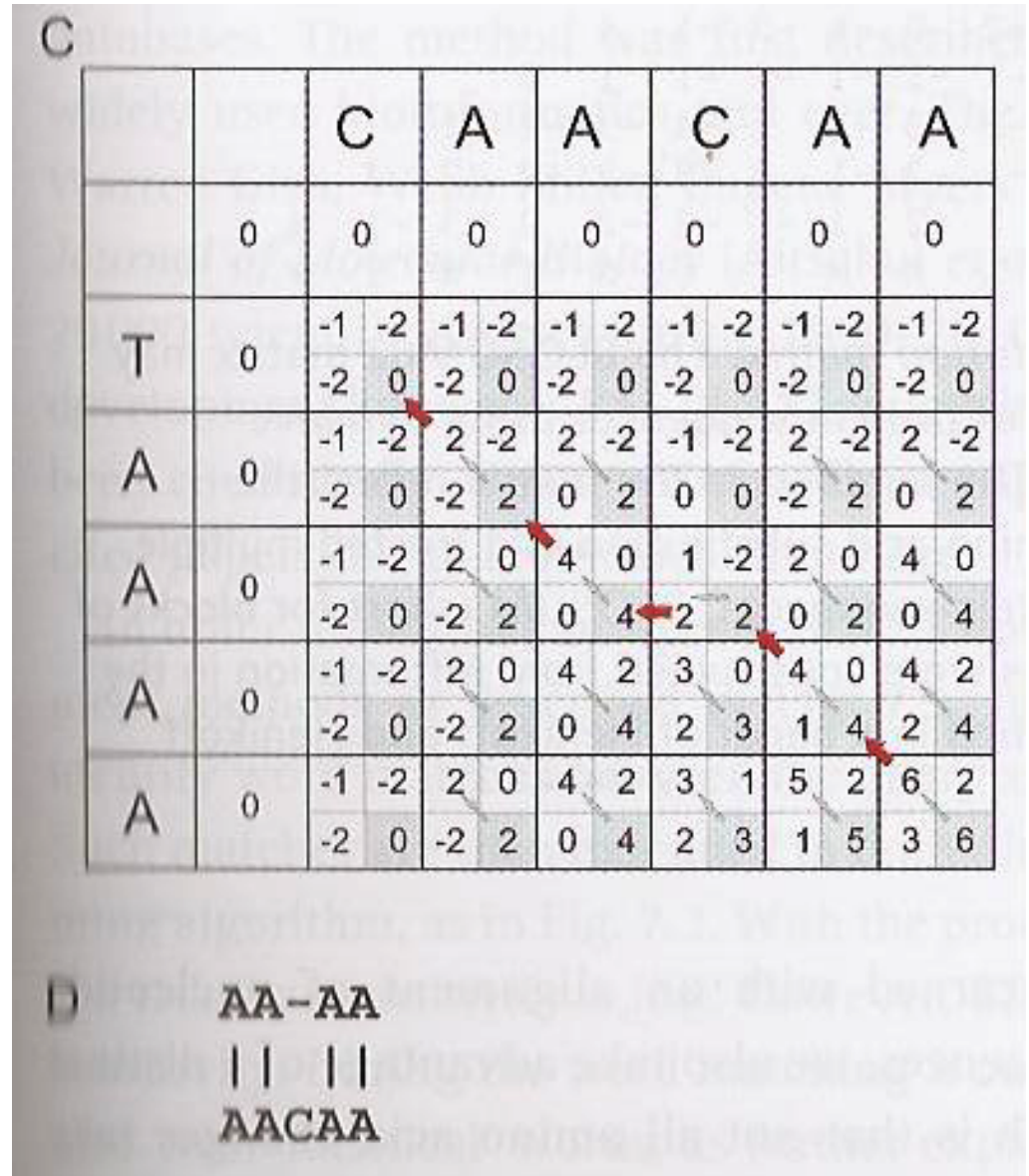
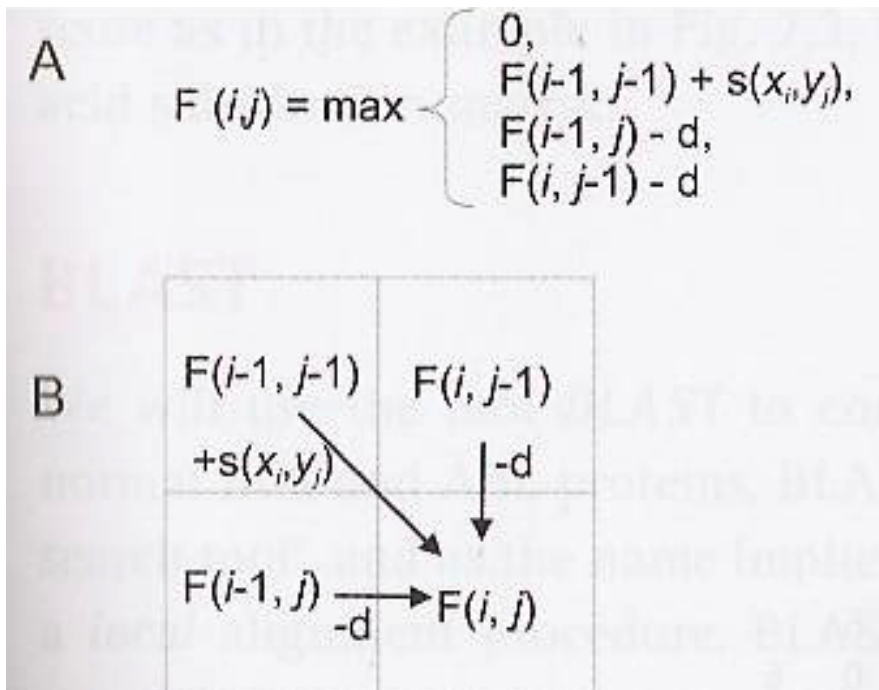
N bases	1/prob occurring	num times by chance in human genome
5	1024	2929687.5
7	16384	183105.4688
9	262144	11444.0918
11	4194304	715.2557373
13	67108864	44.70348358
15	1073741824	2.793967724
17	17179869184	0.174622983
19	2.74878E+11	0.010913936
21	4.39805E+12	0.000682121
23	7.03687E+13	4.26326E-05
25	1.1259E+15	2.66454E-06

DNA alignment as DotPlot



Local alignment using dynamic programming

Gap penalty = $d = 2$
 Mismatch penalty = -1
 Match = 2
 Match or mismatch = s



BLOSUM62: **blocks** of amino acid substitution **matrix**

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	4																	
P	-3	-1	1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	0	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	1



Approximate String Matching using dynamic programming:

example from "Mastering Perl for Bioinformatics"

```
my $pattern = 'EIQADEVRL';  
print "PATTERN:\n$pattern\n";  
  
my $text = 'SVLQDRSMPHQEILAADEVLQESEMRRQDMISHDE';  
print "TEXT:\n$text\n";
```

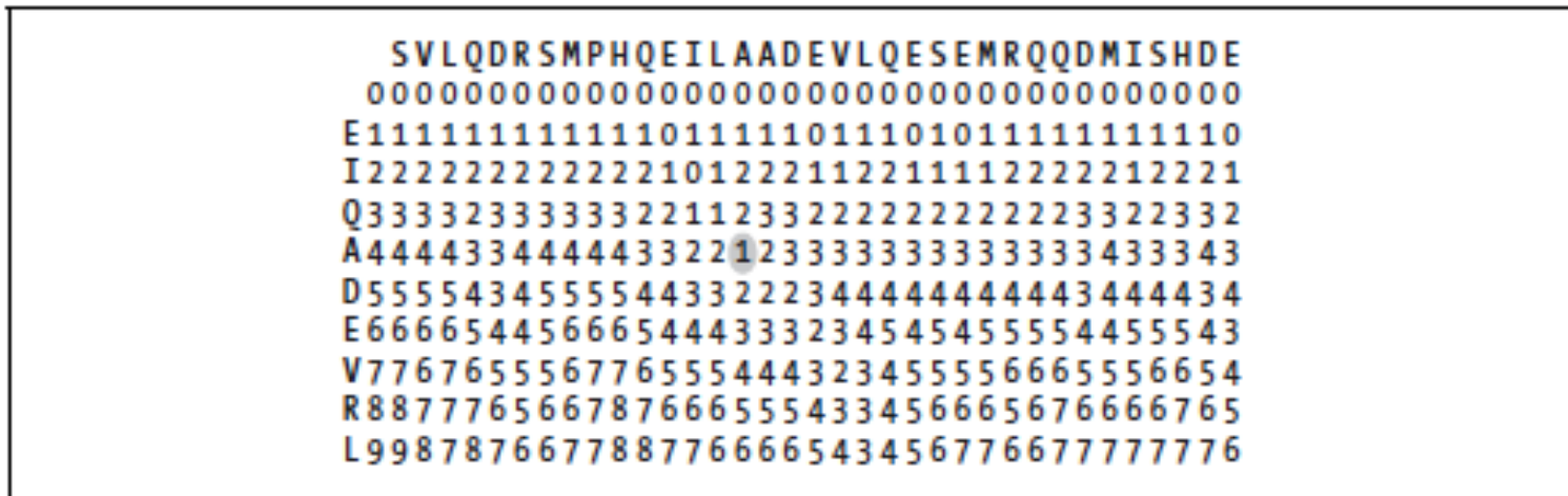


Figure 2-1. Edit distance matrix



Alignment types and alignment software

- Pairwise alignment
 - Database searches
 - BLAST, BLAT
 - Short Read Mappers
 - BWA (Burrows-Wheeler Transform), Mosaik (Smith-Watterman)
 - Genome Aligners
 - MUMmer, (B)LastZ

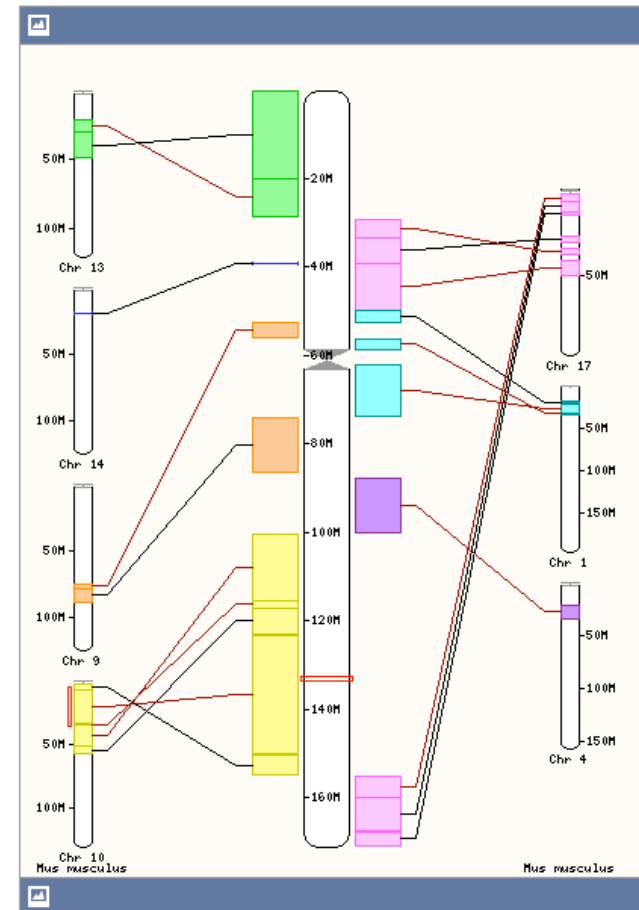
- Multiple sequence alignment
 - Genes, short sequences
 - Clustal(w|x), Muscle, T-Coffee
 - Genome
 - Mauve, Mulan



Synteny: usually defined by gene order

- Configure this page
- Add your data
- Export data
- Bookmark this page
- Share this page

Synteny ⓘ



Change Chromosome:

Change Species:

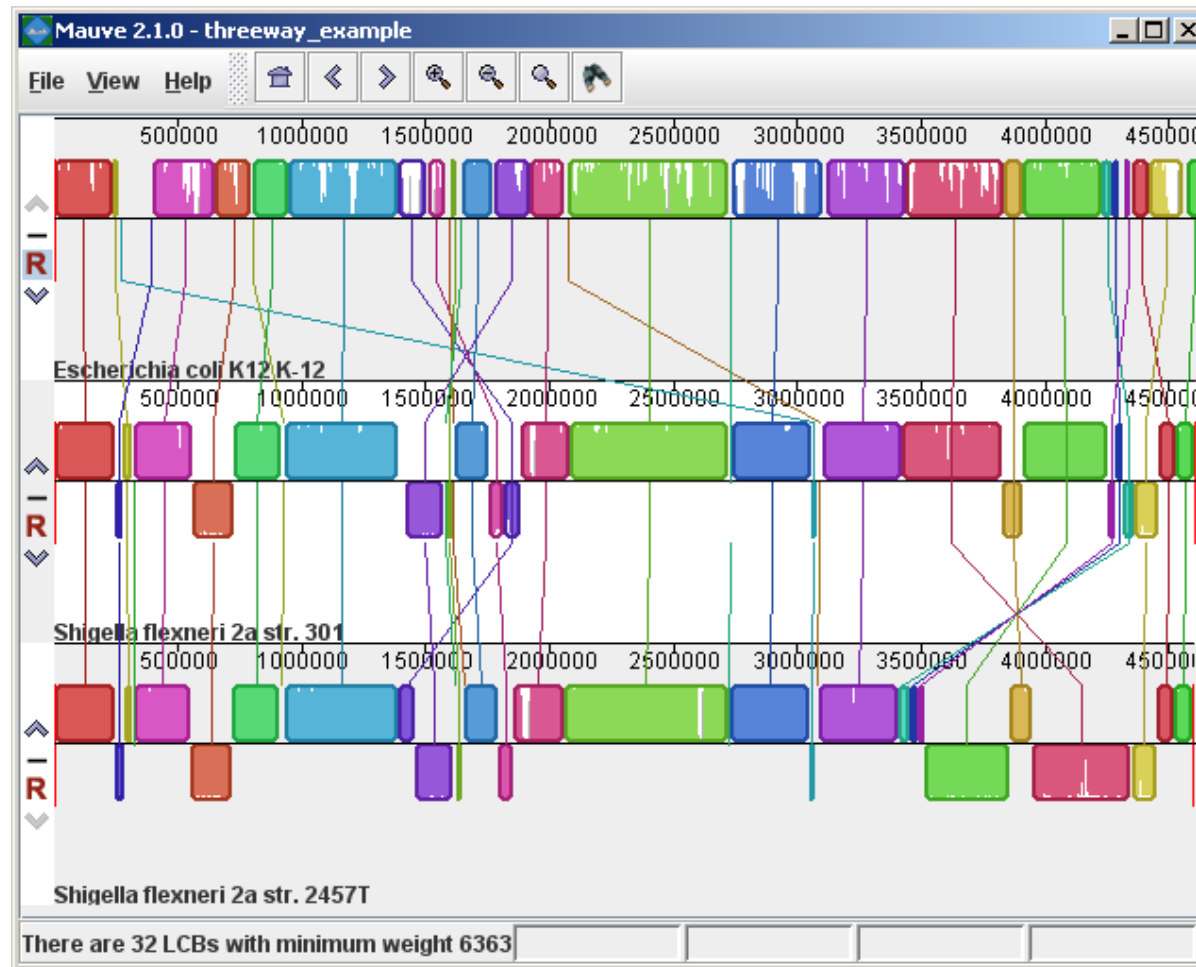
Mouse



◀ 15 upstream genes

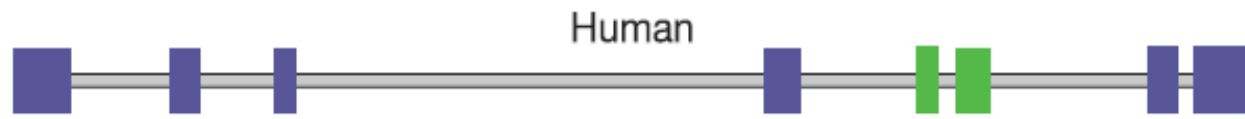
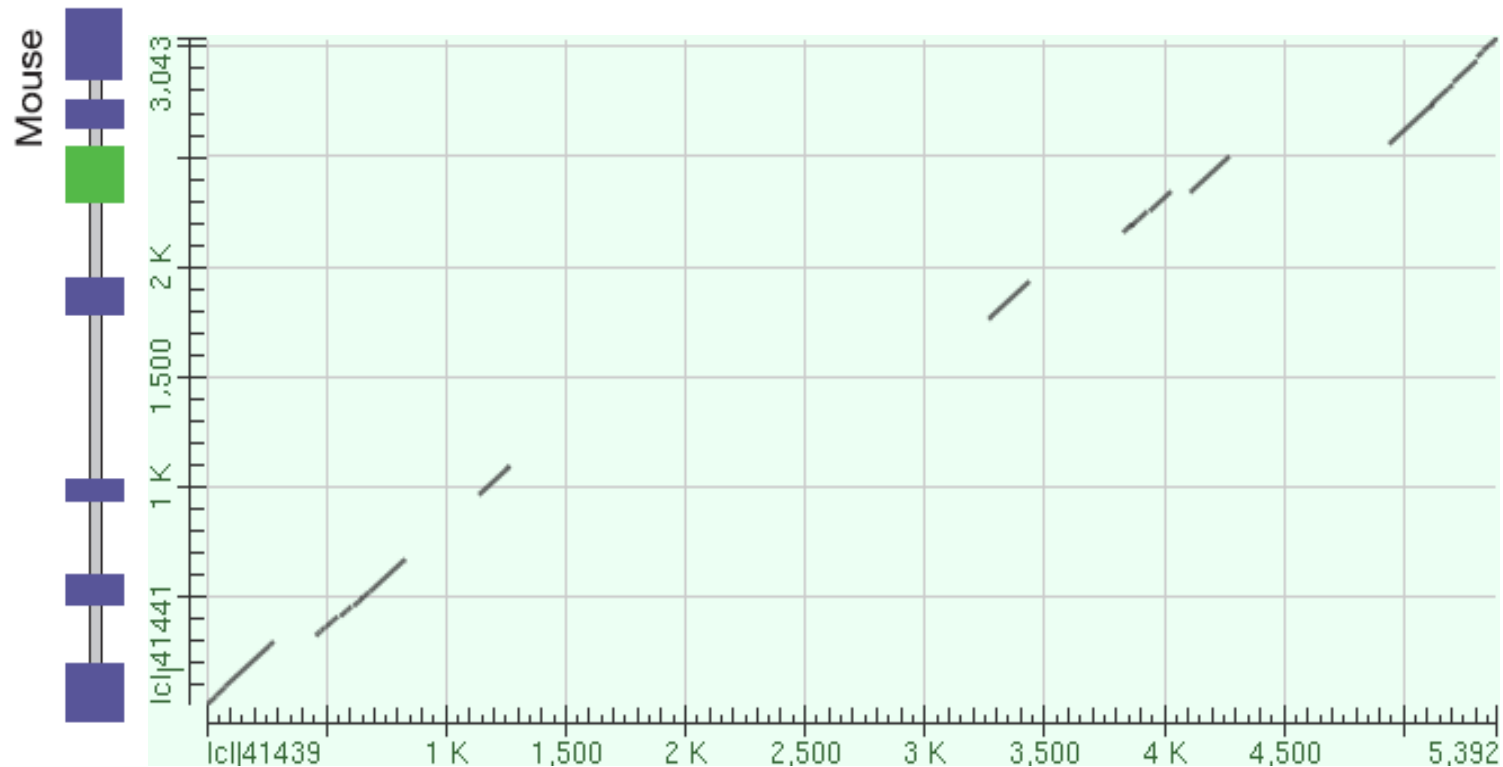
Navigate home

Synteny: whole genome multiple alignment – example using Mauve



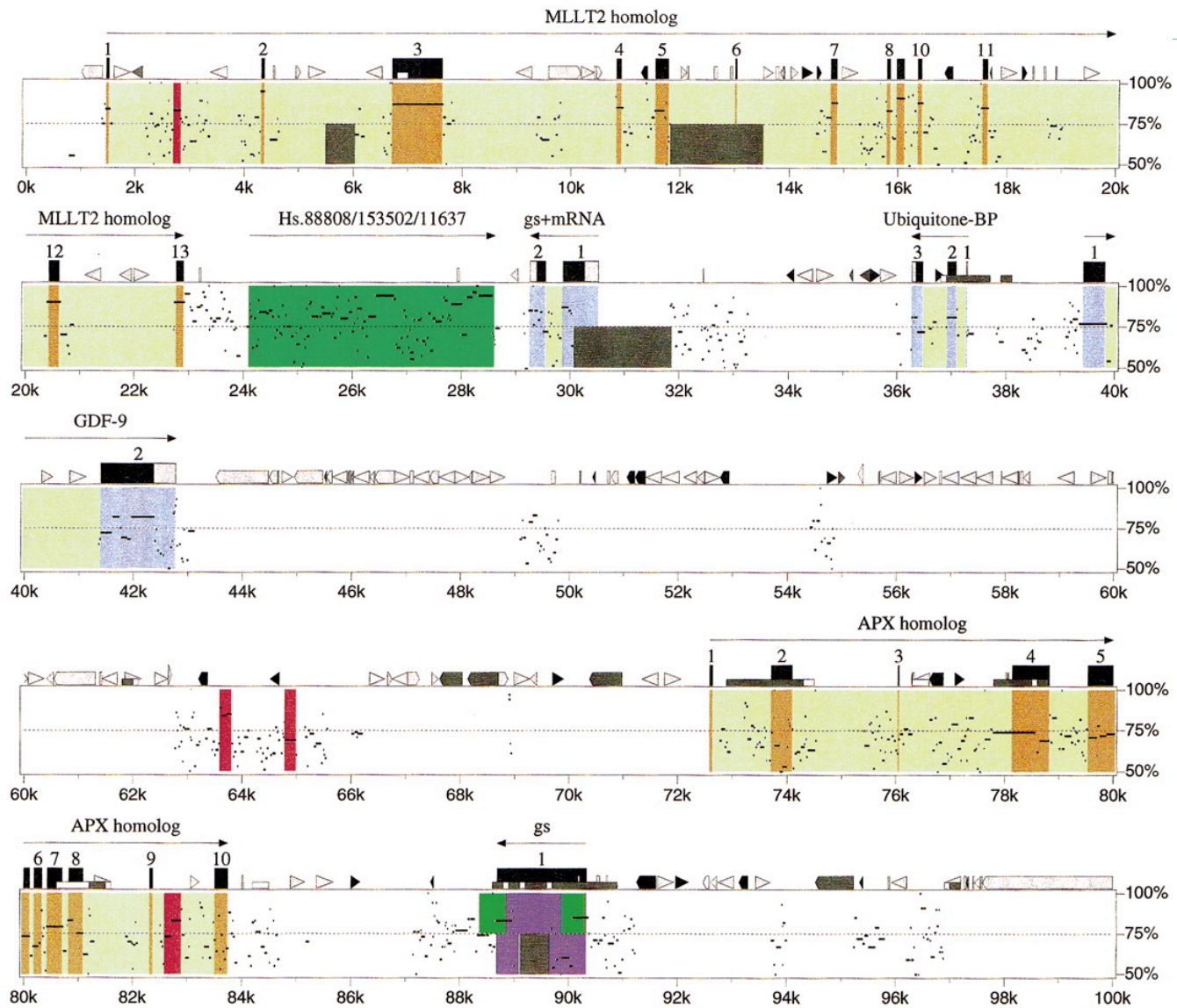
Sequence comparison: Dot plot

Due to purifying selection: High sequence conservation of functional regions



Thu Sep 23 11:01:14 EDT 1999
<http://globin.cse.psu.edu/pipmaker/>

Human clone from Chr. 5q31



WAGENINGEN UNIVERSITY

WAGENINGEN UR

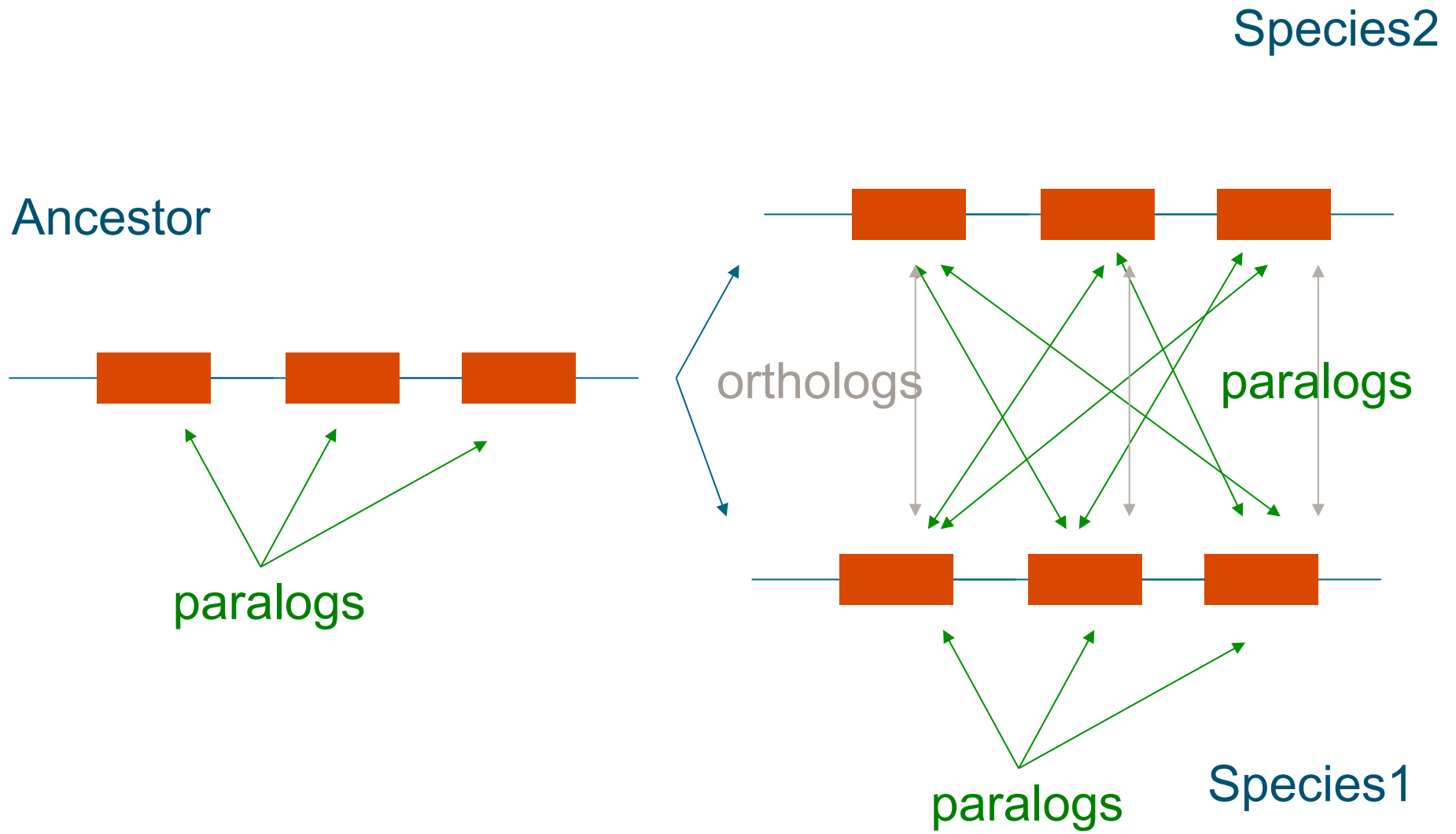
Orthology and Paralogy

Orthology: Sequences are orthologous if they were separated by a **speciation** event

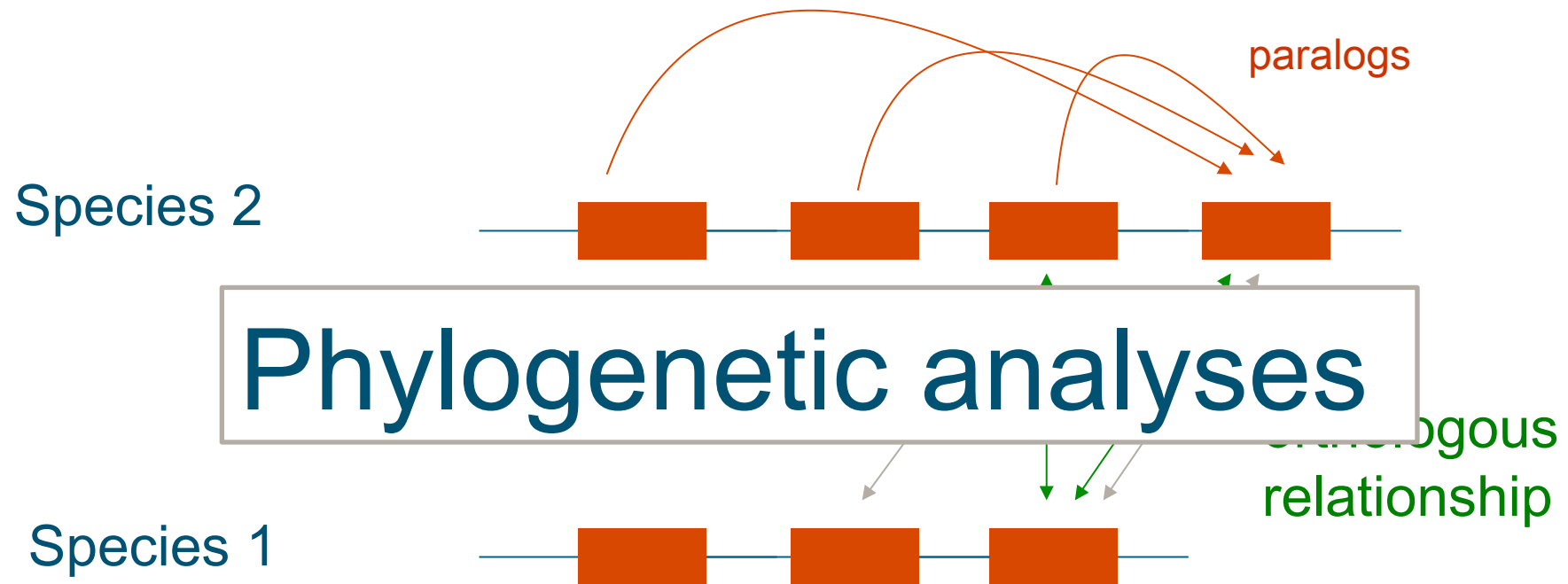
Paralogy: Sequences are paralogous if they were separated by a **gene duplication** event



Orthology vs Paralogy



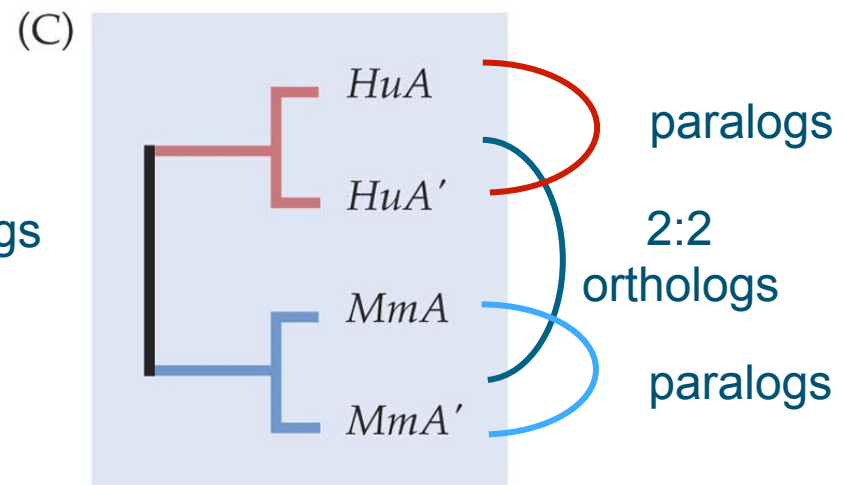
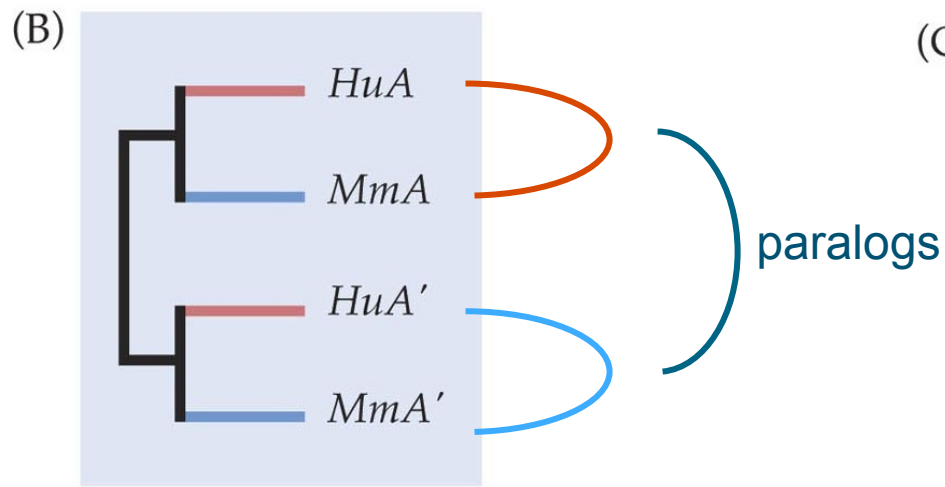
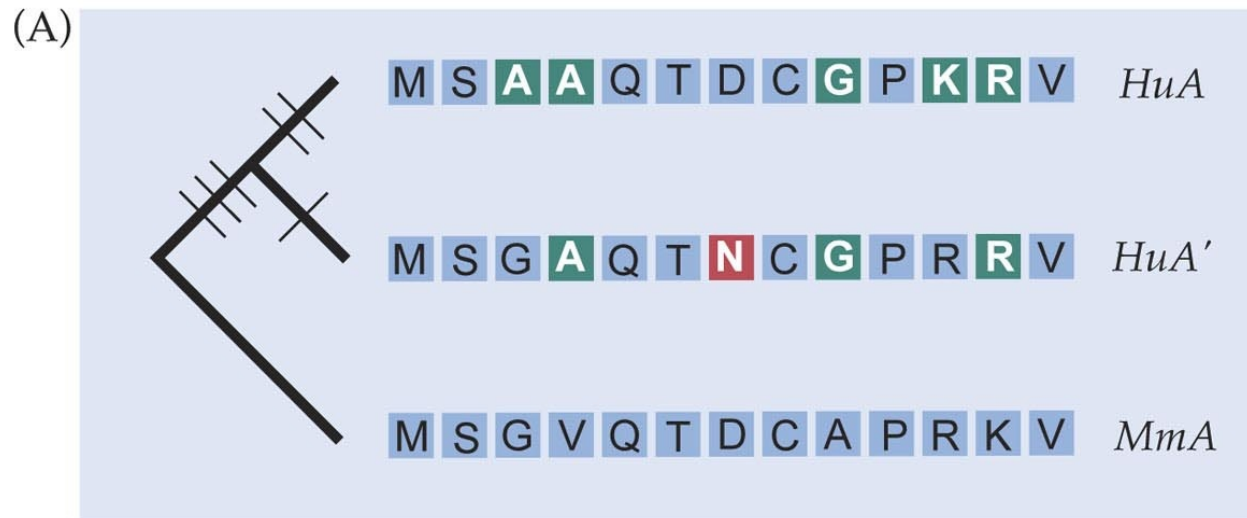
Orthology vs Paralogy complications



It is very likely that you would refer to these two combinations as being the orthologs

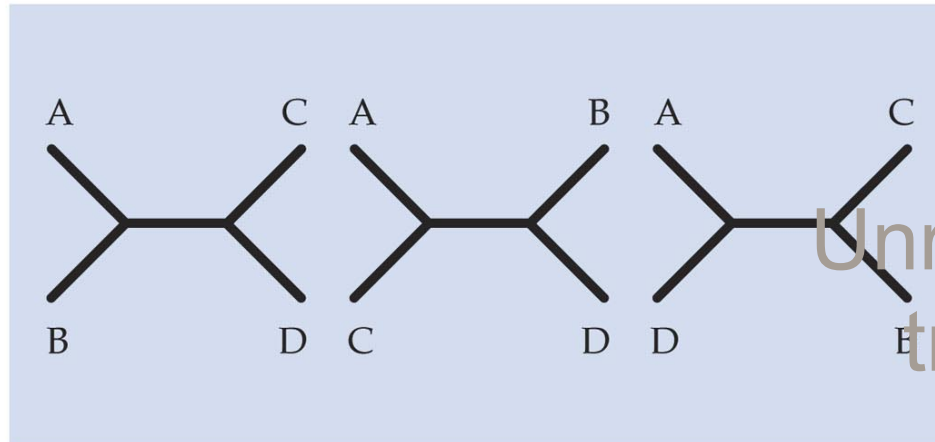
Clustering of sequences and building phylogenetic trees based on genes from species helps to clarify this

Identification of orthologs vs. paralogs

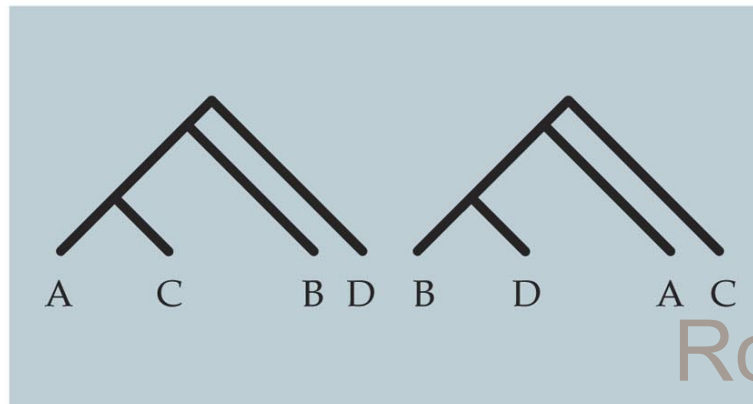


Phylogenetic trees

- Distance Methods
- Parsimony methods
- Maximum Likelihood
- Bayesian



Unrooted trees



Rooted trees

A PRIMER OF GENOME SCIENCE, Second Edition, Box 2.4 (Part 1) © 2005 Sinauer Associates, Inc.

Methods in CB: Phylogeny

evolution.genetics.washington.edu/phylip/software.html

Owing to other pressures on my time, I cannot devote much time to searching for new programs, so their authors are begged to (please!) use the submission form instead.

Methods By computer Cross-referenced Data types Web servers New programs Submitting

Phylogeny Programs

Changes Waiting list Other lists Old programs Not listed News

BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 21 2006, pages 2688–2690
doi:10.1093/bioinformatics/btl446

Phylogenetics

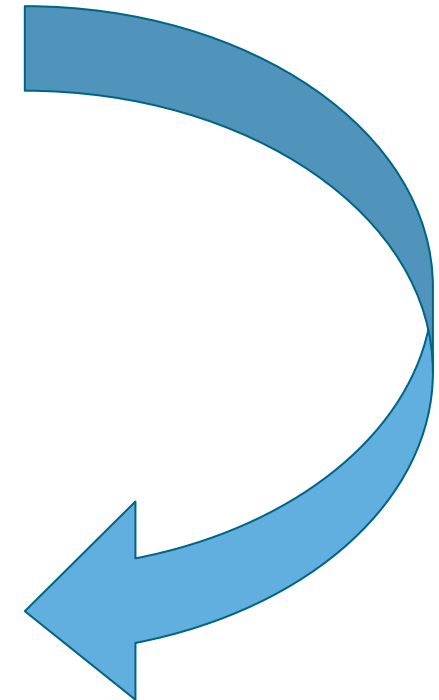
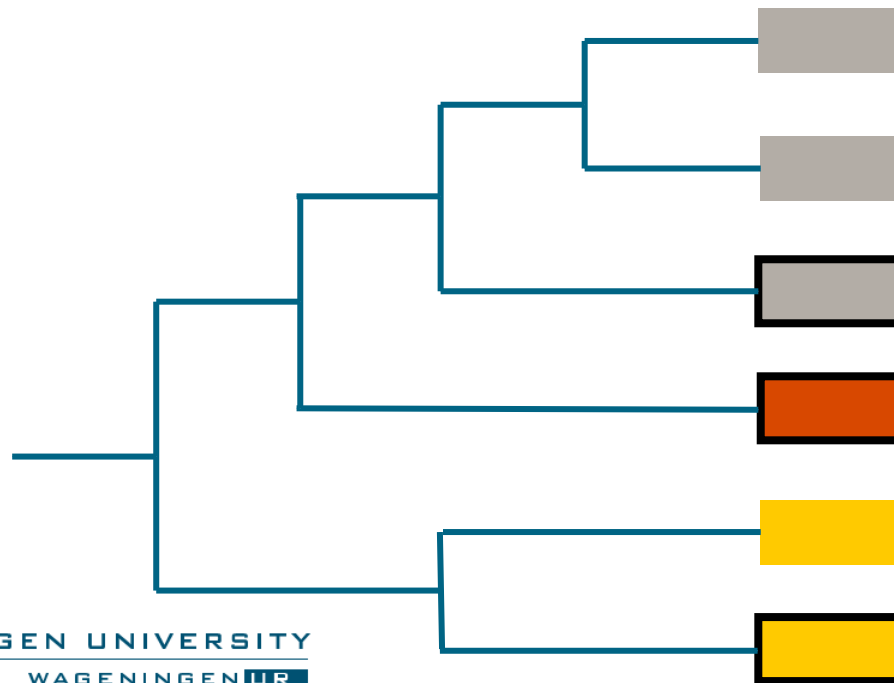
RxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models

Alexandros Stamatakis



WAGENINGEN UNIVERSITY
WAGENINGEN UR

Orthology vs Paralogy complications



Genome Annotation and Phylogenetic Analysis

- Much of the annotation for non-model organisms is based on comparative work with model organisms.
- Transfer/optimization of gene models
- Inference of function (e.g. GO annotation)
- Becoming ever more important with the rapidly increasing number of non-model organisms being sequenced

Gene Naming:

the Hugo Gene Nomenclature Committee is effectively determining gene names across the vertebrates



HGNC
HUGO Gene Nomenclature Committee

Search Genes

Home Search Genes Downloads Gene Families HCOP Useful Links About Contact Us Request Symbol

Comparative Genomics and Orthology Resources

- [Compare](#) Multi Organism Information System
- [EggNog](#) evolutionary genealogy of genes: Non-supervised Orthologous Groups
- [EnsemblCompara](#)
- [EvoLa](#) Evolutionary annotation database
- [Genomicus](#) Genomes in evolution
- [HCOP](#) HGNC Comparison of Orthology Predictions
- [HomoloGene](#)
- [Inparanoid](#) eukaryotic ortholog groups
- [IsoBase](#) A Database of Functionally Related Orthologs
- [Metazome](#) database and graphical user interface enabling comparative genomic studies within the metazoa
- [OMA](#) (Orthologous Matrix project)
- [OPTIC](#) Clade genomics web server
- [OrthoDB](#) The Hierarchical Catalog of Eukaryotic Orthologs
- [OrthoInspector](#)
- [OXGRID](#) the Oxford Grid project
- [P-POD](#) Princeton Protein Orthology Database
- [Panther](#) Classification System
- [PHOG](#) PhyloFacts Orthology Group
- [PhylomeDB](#)
- [TreeFam](#) Tree families database
- [VISTA](#) Tools for Comparative Genomics

EMBL-EBI    

The work of the HGNC is supported by National Human Genome Research Institute (NHGRI) grant P41 HG03345 and Wellcome Trust grant 099129/Z/12/Z

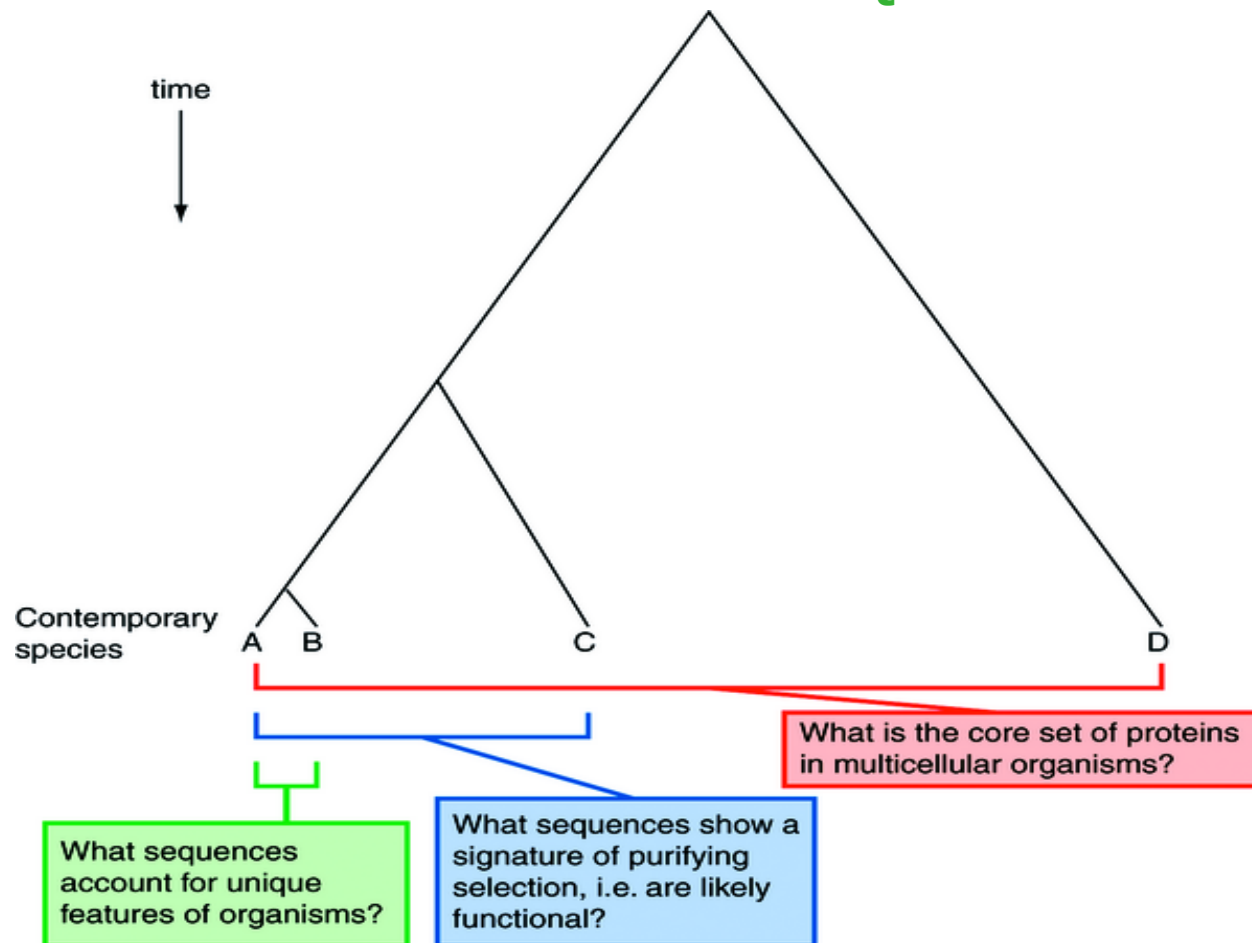
Contact Us: hgnc@genenames.org [Back to top Admin](#)

Site maintained by the External Services team at [EMBL-EBI](#) | [Terms of Use](#) | [Privacy](#) | [Cookies](#)

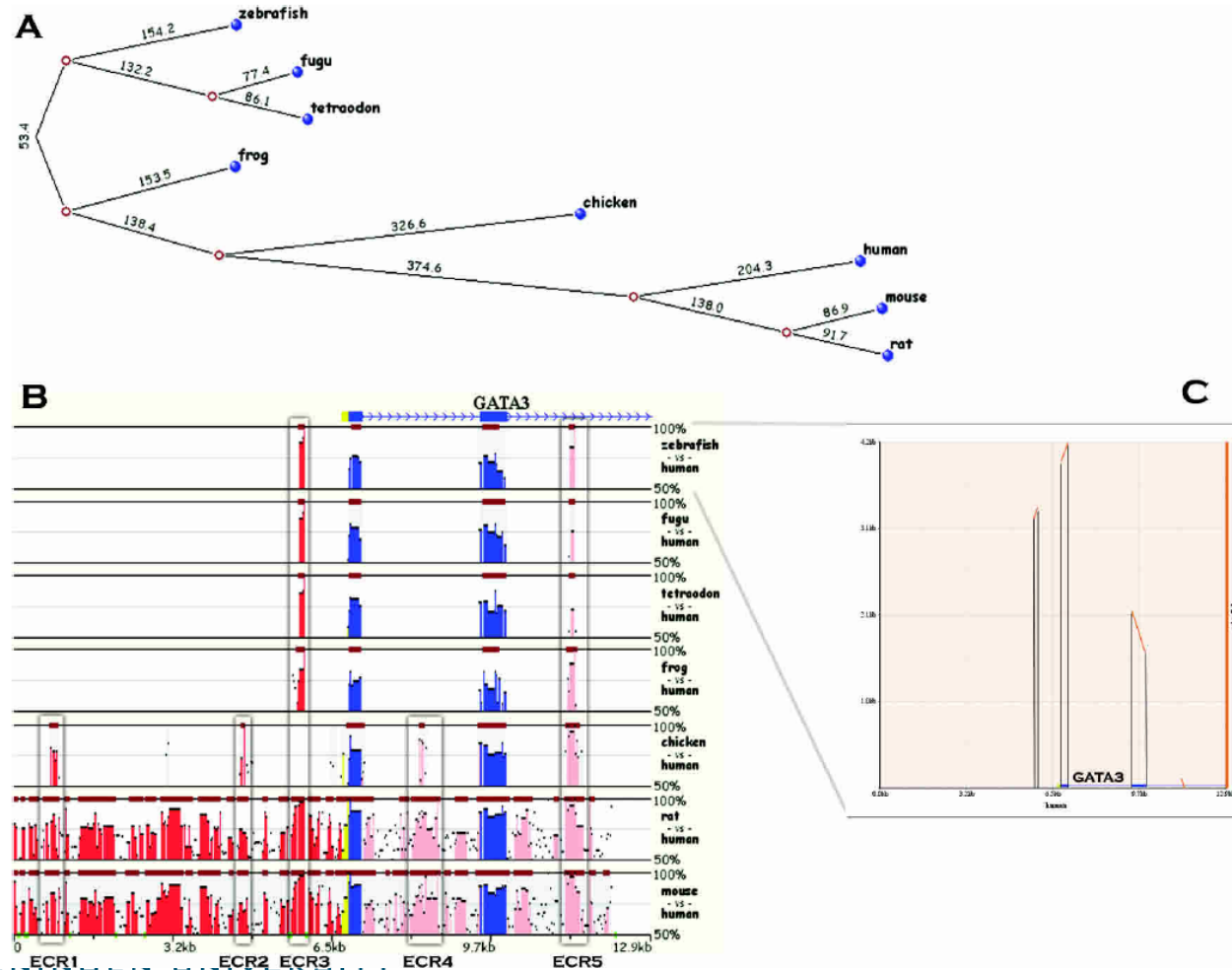


Comparative Biology

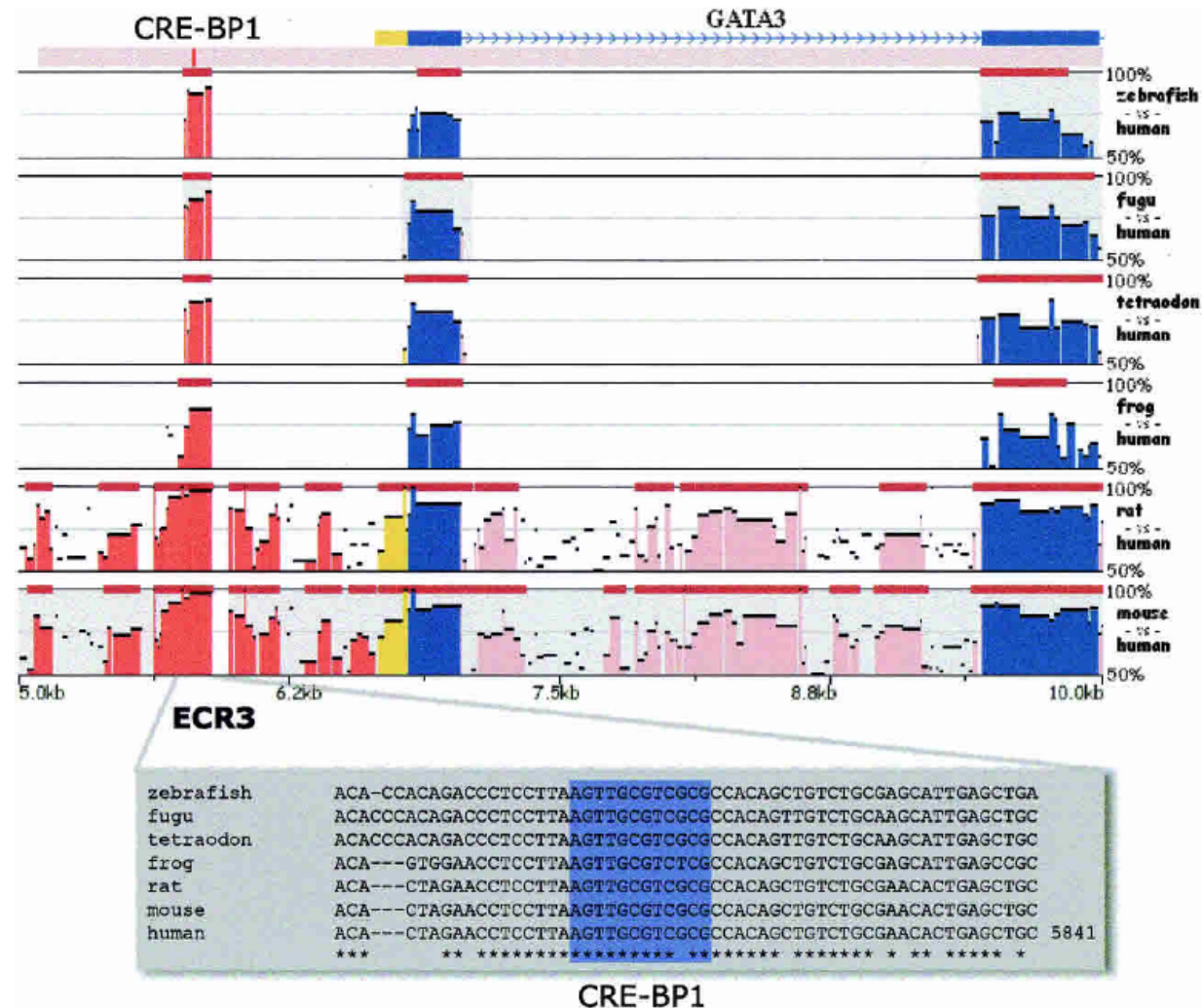
Comparisons of Genomes at Different Phylogenetic Distances Are Appropriate to Address Different Questions



Predicting function: Transcription Factor Binding Site prediction (Mulan)



Predicting function: Transcription Factor Binding Site prediction (Mulan)



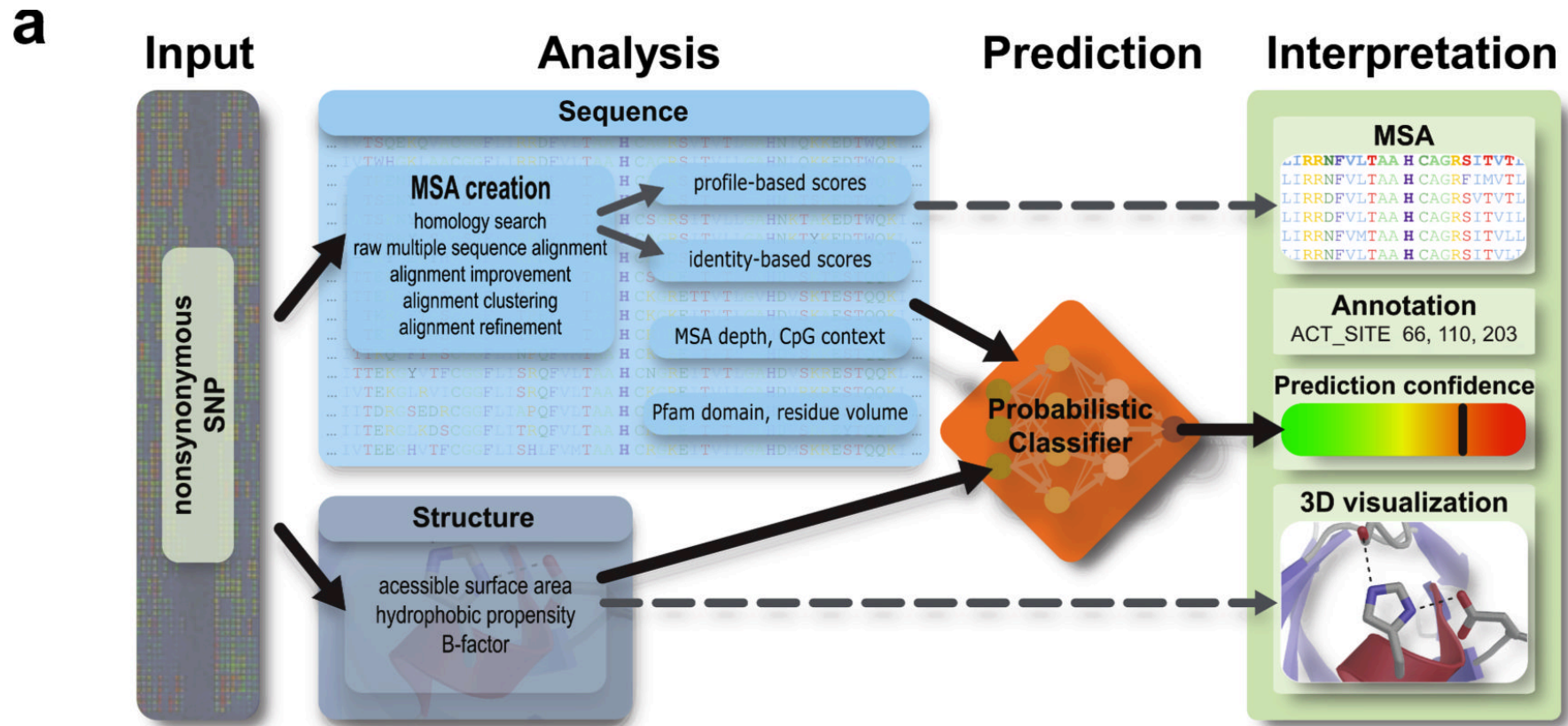
Predicting functional aspect of amino acid substitutions

- An amino-acid substitution that is rare in a certain taxonomic group is more likely to have functional consequences than an amino acid substitution that occurs often.

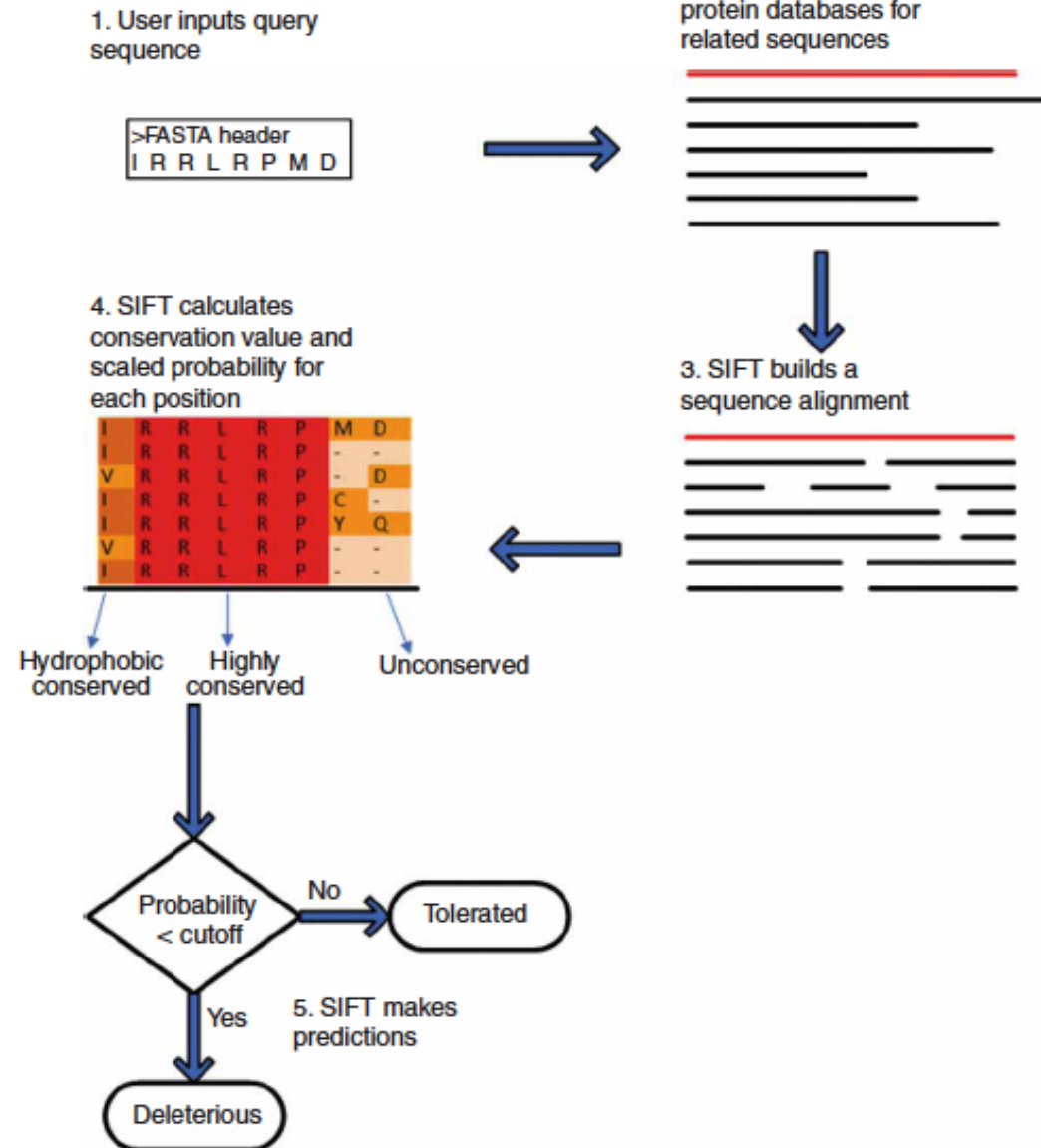
Method	Algorithm
SIFT (http://sift.jcvi.org)	SIFT uses sequence homology; scores are calculated using position-specific scoring matrices with Dirichlet priors
Polyphen ¹² (http://genetics.bwh.harvard.edu/pph/)	Polyphen uses sequence conservation, structure and SWISS-PROT annotation
PMUT ¹³ (http://mmb2.pcb.ub.es:8080/PMut/)	PMUT provides prediction by neural networks, which use internal databases, secondary structure prediction and sequence conservation
SNPs3D ¹⁸ (http://www.snps3d.org/)	SNPs3D is based on support vector machine that uses structural or sequence conservation features
PantherPSEC ¹⁹ (http://www.pantherdb.org/tools/csnpScoreForm.jsp)	Panther PSEC uses sequence homology; scores are calculated using PANTHER Hidden Markov model families
MAPP ¹⁴ (http://mendeL.stanford.edu/SidowLab/downloads/MAPP/index.html)	MAPP considers the physicochemical variation present in a column of a protein sequence alignment to predict the effect of all possible amino acid substitutions on protein function
Align-GVGD ¹⁵ (http://agvgd.iarc.fr/agvgd_input.php)	Align-GVGD combines the biophysical characteristics of amino acids and protein multiple sequence alignments



Predicting function: PolyPhen



Predicting function: SIFT



Predicting function: SIFT

Protein ID	Substitution	dbSNP ID	Prediction	Score	Median Info	Number of Seqs at position
NP_001008498	V352I	rs27180025	TOLERATED	1.00	1.38	99
NP_001008499	Y219C	rs33644739	DAMAGING	0.00	1.90	98
NP_001008501	C192Y	rs32419104	TOLERATED	0.25	1.46	84
82934505	Q23L	rs13484479	DAMAGING *Warning! Low confidence.	0.00	4.32	1
82934725	K52E	rs26925385	TOLERATED	1.00	4.32	2
82934691	P52R	rs26956085	TOLERATED	1.00	4.23	101

* Low confidence means that the protein alignment does not have enough sequence diversity. Because the position artificially appears to be conserved, an amino acid may incorrectly

Click [here](#) to download the following table in tab separated format. You can open it using excel with delimiter set as TAB

If you received a warning that the sequences were not diverse enough, you can have SIFT choose more diverse sequences [here](#).



Predicting function: SIFT

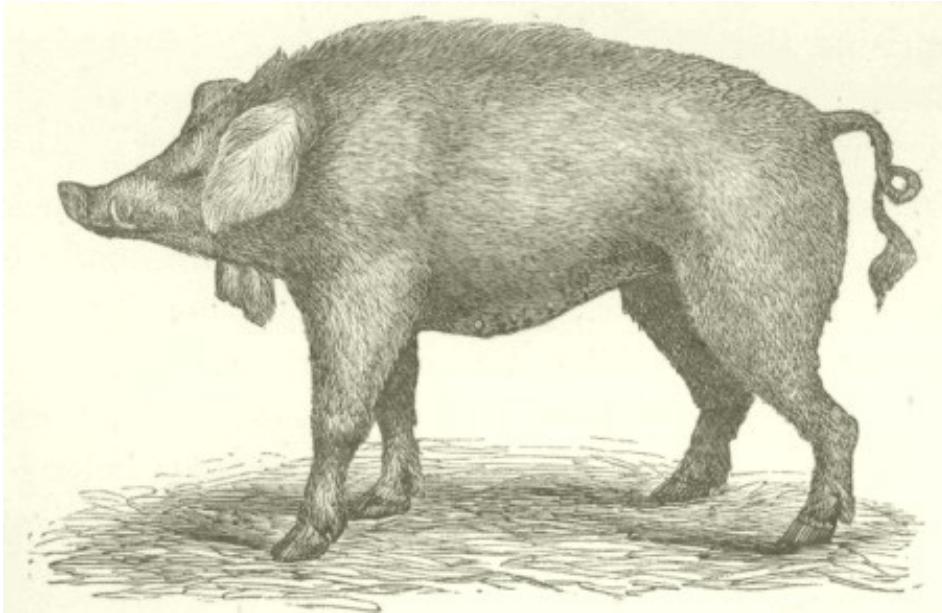
Predict not tolerated	Position	Seq Rep	Predict tolerated
d c g w h n e s p r k q y t	3M	0.60	a f v i M L
w	4A	0.60	c f m y i h v l p r q t n k s e G D A
w m h f	5C	0.80	y i q r e l k d p n v g t a S C
w c m f	6R	0.80	y i h v p l d g n q e t a k S R
w d h q p n c e r g k s	7V	0.80	y t a m F I I V
w d h g	8I	0.80	n c r q p e y k s f m t A v L I
	9N	0.80	w c h p f y M i q r v g e d t a k s l N
c w m f i d	10R	0.80	p v y g s l t n a e q H k R
c w f m y i v d h p g l	11R	0.80	n s t a e k Q R
c w f m y i v d h p g l	12R	0.80	n s t a e k Q R
y w v t s r q p n m l k i g f e d c a	13H	0.80	H

Figure 7 | Output prediction for all substitutions for a single protein. Shown here are amino acid positions 3–13 of a protein. Capital letters indicate amino acids appearing in the protein sequence alignment used for prediction whereas lower case letters indicate amino acids not observed in the protein alignment. The numbers in the 'Seq Rep' column represent the fraction of sequences in the alignment that have an amino acid at the corresponding position.

pos	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1M 0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2E 0.25	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3M 0.50	0.07	0.02	0.02	0.03	0.12	0.02	0.02	0.24	0.03	1.00	0.63	0.02	0.03	0.03	0.03	0.03	0.05	0.17	0.02	0.04
4A 0.50	1.00	0.05	0.13	0.17	0.04	0.68	0.04	0.05	0.16	0.09	0.04	0.12	0.14	0.10	0.10	0.30	0.15	0.11	0.01	0.05
5C 0.75	0.59	1.00	0.17	0.15	0.09	0.33	0.08	0.11	0.17	0.15	0.06	0.19	0.18	0.12	0.12	0.93	0.44	0.21	0.03	0.11
6R 0.75	0.37	0.04	0.24	0.36	0.06	0.23	0.11	0.10	0.58	0.17	0.06	0.26	0.15	0.29	1.00	0.63	0.33	0.15	0.02	0.09
7V 0.75	0.10	0.03	0.02	0.03	0.22	0.04	0.03	0.99	0.04	0.48	0.10	0.03	0.03	0.03	0.04	0.05	0.09	1.00	0.02	0.09
8I 0.75	0.44	0.07	0.07	0.13	0.21	0.07	0.07	0.86	0.13	1.00	0.21	0.08	0.09	0.10	0.10	1.00	0.13	0.24	0.85	0.04
9N 0.75	0.90	0.15	0.73	0.85	0.41	0.60	0.37	0.54	0.94	1.00	0.36	0.95	0.36	0.59	0.66	0.93	0.84	0.64	0.10	0.46
10R 0.75	0.14	0.02	0.08	0.16	0.07	0.10	0.38	0.06	0.46	0.12	0.04	0.13	0.08	0.19	1.00	0.13	0.13	0.09	0.02	0.11
11R 0.75	0.10	0.01	0.06	0.13	0.02	0.06	0.05	0.03	0.33	0.07	0.02	0.08	0.05	0.46	1.00	0.09	0.08	0.05	0.01	0.03
12R 0.75	0.10	0.01	0.06	0.13	0.02	0.06	0.05	0.03	0.33	0.07	0.02	0.08	0.05	0.46	1.00	0.09	0.08	0.05	0.01	0.03
13H 0.75	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Figure 8 | Output scaled probability matrix for a single protein. Shown here are the results for amino acid positions 3–13 of a protein. The threshold score for intolerance of a substitution is 0.05. Rows correspond to positions in the reference protein and columns correspond to all possible 20 amino acids. The header column on the left contains the position, reference amino acid and the fraction of sequences in the alignment that have an amino acid at the corresponding position. Each element of the matrix represents the scaled probability of substitution at each position. Substitutions predicted to be intolerant are highlighted in red.

Transition to modern pig breeds



- Asian 'genes' have been important in shaping modern pig breeds.
- But evidently selection has been the major driver – selection on Asian (e.g. IGF2), and European variation?

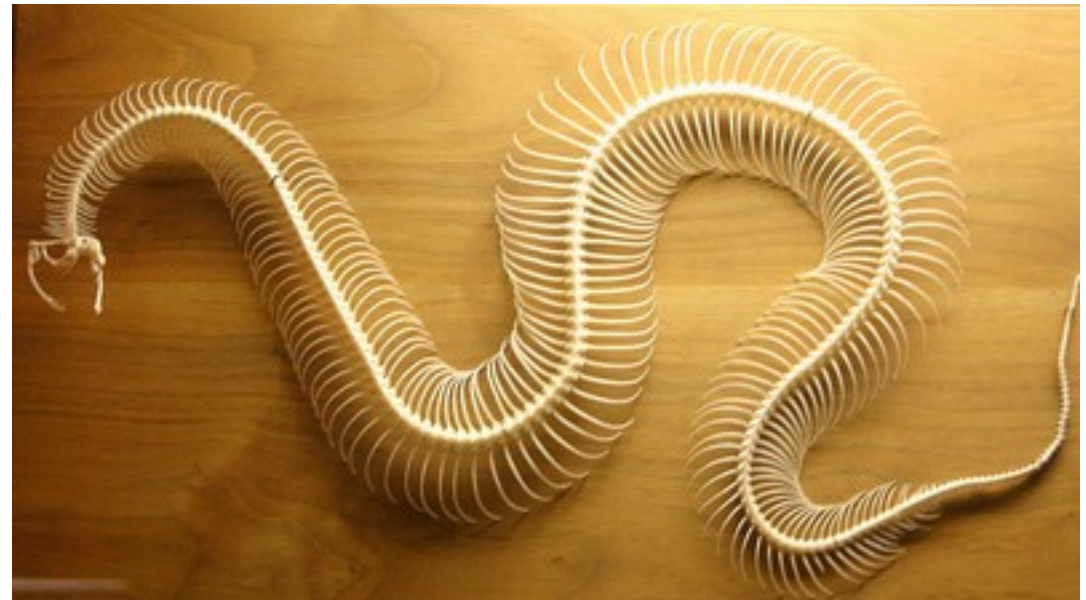
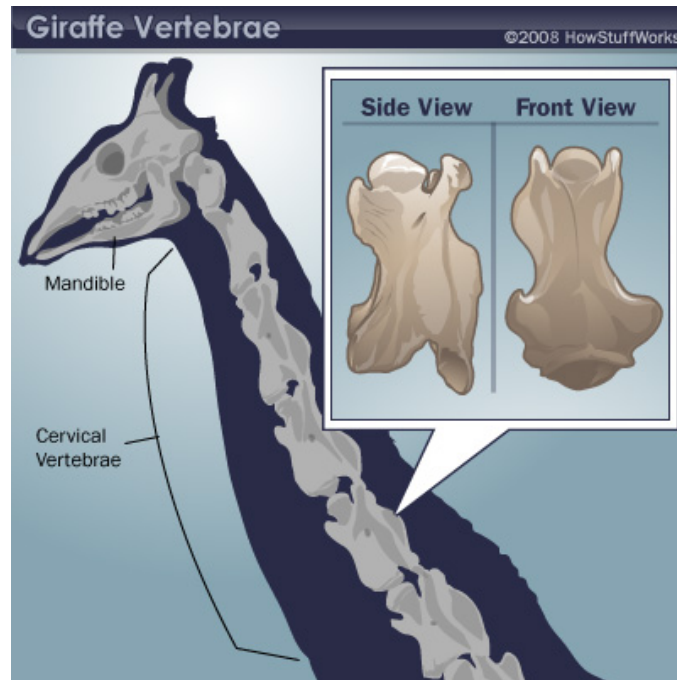


Pigs got longer in the 19th Century

	English Long-legged Male.	African Female.	Chinese Male.	Wild Boar, from Cuvier.	French Domestic Boar, from Cuvier.
Dorsal vertebræ ..	15	13	15	14	14
Lumbar	6	6	4	5	5
Dorsal and lumbar } together }	21	19	19	19	19
Sacral	5	5	4	4	4
Total number of } vertebræ }	26	24	23	23	23

Darwin, C. R. 1868. *The variation of animals and plants under domestication*. London: John Murray. First edition, first issue. Volume 1.

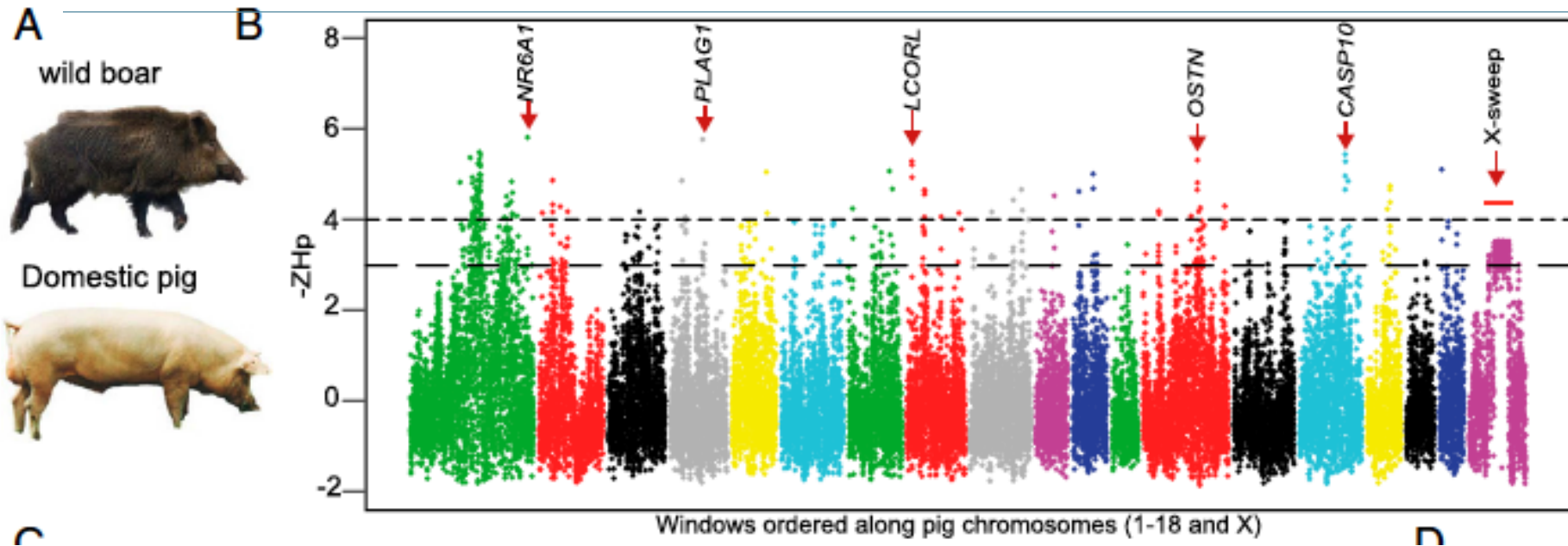
Number of vertebrae are usually very conserved in mammals



Giraffe has 7 neck vertebrae, just like us. Snakes on the other hand vary widely in number of vertebrae

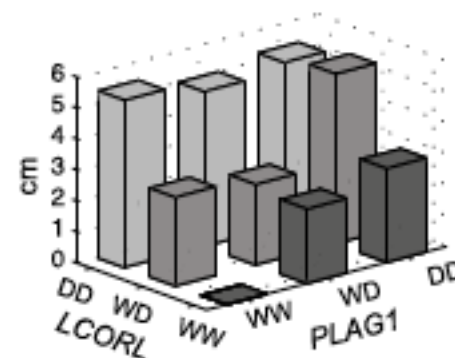


Three of the major signatures of selection in modern pigs are related to body length

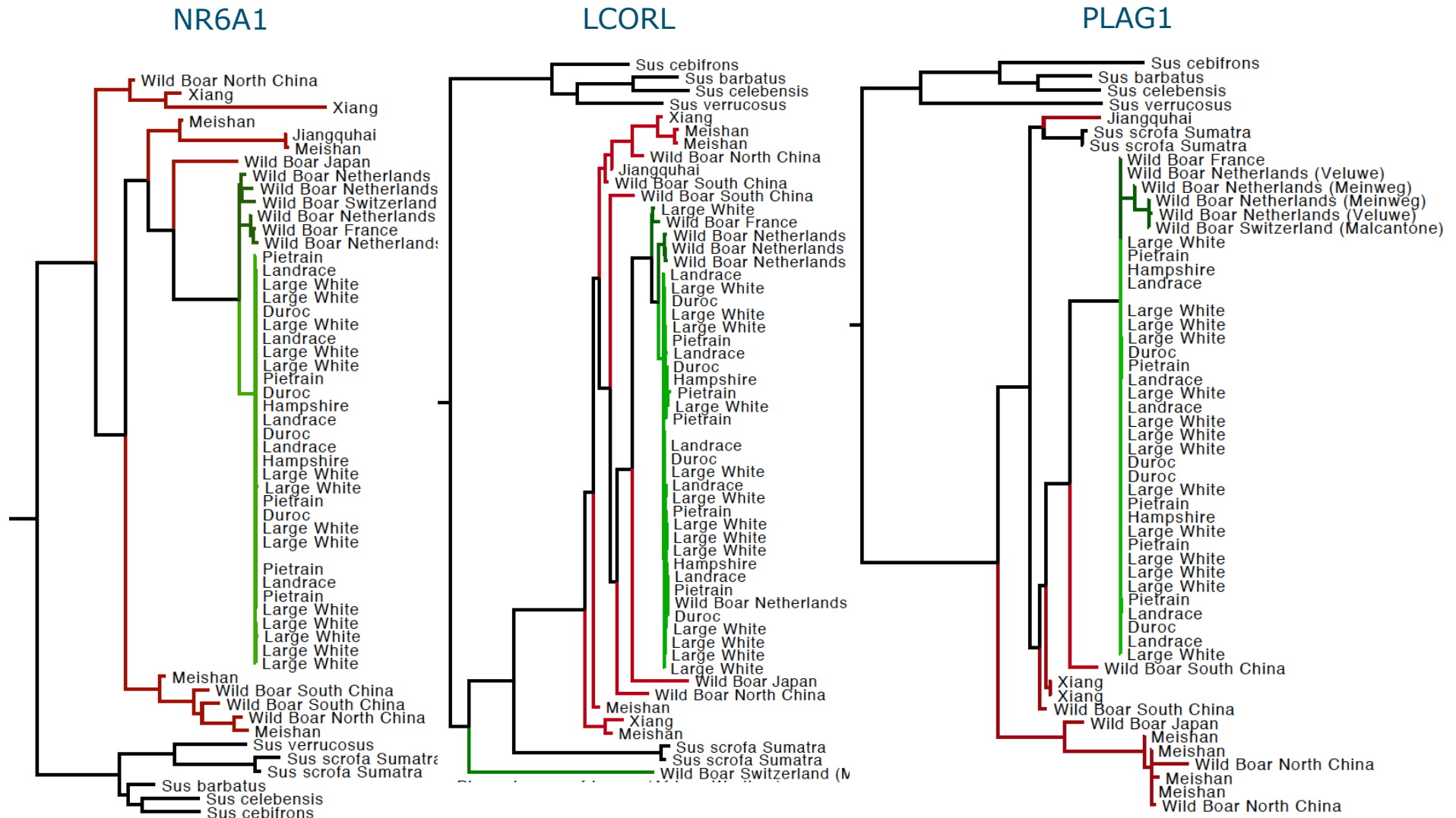


- NR6A1
- PLAG1
- LCORL

Rubin, Megens, et al., 2012, PNAS 109:19529



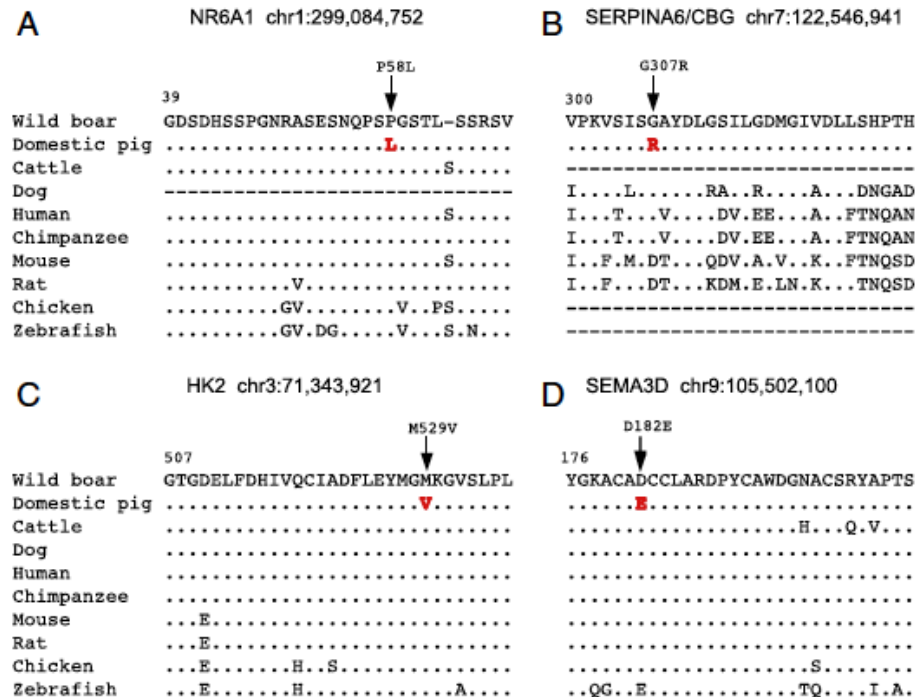
The three 'elongation genes' are not introduced from Asia



Derived nucleotide substitutions showing marked allele frequency differences between wild boar and domestic pigs are enriched in QTn

Substitution	Population	
	Domestic pig	Wild boar
Nonsynonymous	72	6
Synonymous	87	37

Values indicate the number of derived substitutions in which the frequency of the ancestral allele is >0.80 in the indicated population and <0.20 in the other.



Rubin, Megens, et al., 2012, PNAS 109:19529

Comparative Genomics & Genome Databases

Hendrik-Jan Megens

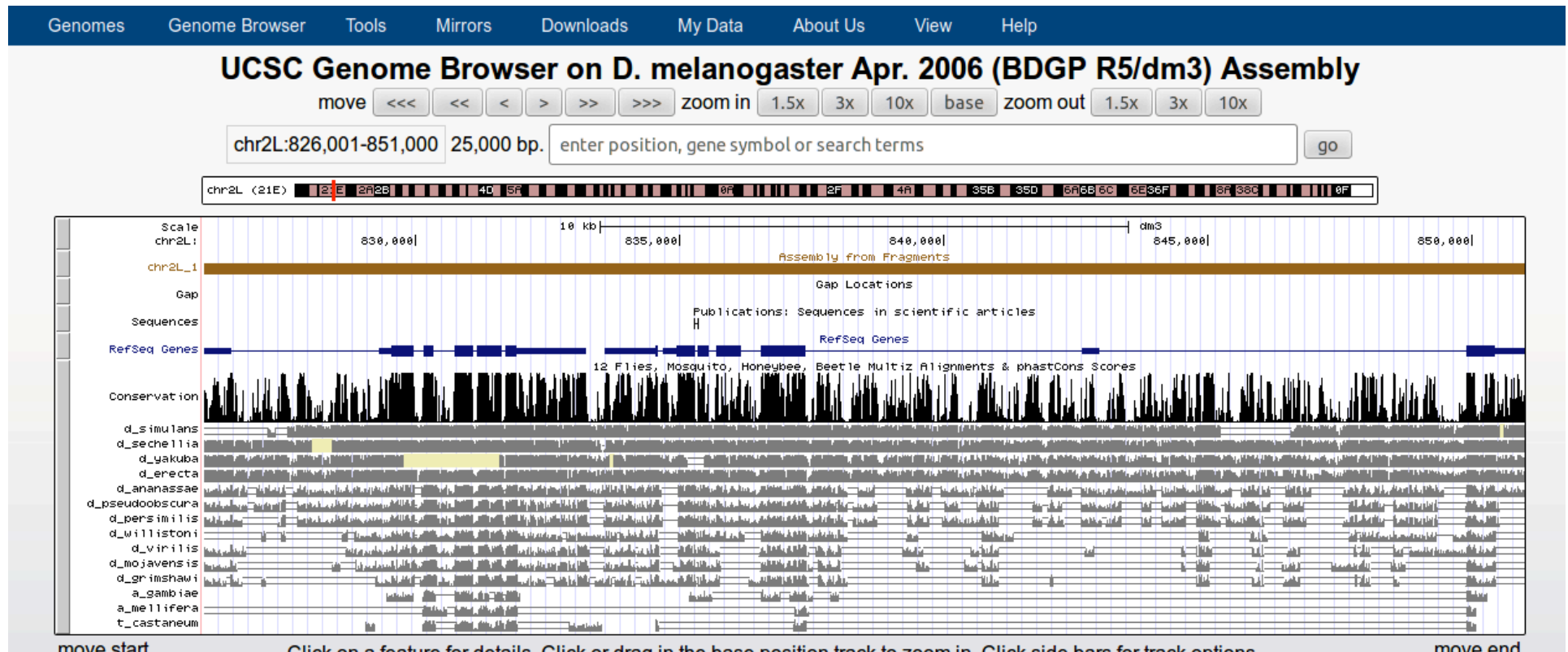


Comparative genomics 'big bang'

- Reference assemblies are currently drafted for many species
- Various -omics data being generated for model, and non model species:
 - Expression/RNAseq
 - Epigenetics/methylomes
 - DNA-protein interactions ('Encode-type data')
 - Variation
- Genome browsers increasingly geared towards comparative tools
 - But buckling under the flood of new species data



Comparing Genomes: UCSC



UCSC and Galaxy

The screenshot shows the Galaxy web interface in a Mozilla Firefox browser. The main content area displays a table of genomic data with columns: #bin, chrom, chromStart, chromEnd, name, and score. The table lists 20 rows of data for chromosome 4 (chr4).

#bin	chrom	chromStart	chromEnd	name	score
73	chr4	786389	786442	lod=72	431
585	chr4	0	15	lod=7	119
585	chr4	199	200	lod=22	272
585	chr4	223	229	lod=13	202
585	chr4	421	437	lod=15	221
585	chr4	972	980	lod=12	191
585	chr4	999	1007	lod=11	179
585	chr4	1110	1118	lod=11	179
585	chr4	1273	1290	lod=12	191
585	chr4	1319	1344	lod=11	179
585	chr4	1678	1695	lod=10	166
585	chr4	3714	3743	lod=18	245
585	chr4	3782	3814	lod=12	191
585	chr4	4244	4268	lod=11	179
585	chr4	4354	4385	lod=17	238
585	chr4	4459	4480	lod=12	191
585	chr4	5160	5197	lod=63	413
585	chr4	5238	5248	lod=19	252
585	chr4	6091	6099	lod=10	166
585	chr4	7184	7207	lod=21	266
585	chr4	7280	7285	lod=13	202
585	chr4	7337	7343	lod=11	179
585	chr4	7517	7521	lod=10	166
585	chr4	7577	7582	lod=18	245
585	chr4	7707	7716	lod=15	221
585	chr4	7823	7828	lod=11	179
585	chr4	7892	7900	lod=16	229
585	chr4	7922	7928	lod=10	166
585	chr4	8123	8128	lod=16	229
585	chr4	8201	8206	lod=10	166
585	chr4	8447	8455	lod=16	229
585	chr4	8477	8482	lod=12	191
585	chr4	8556	8569	lod=24	284
585	chr4	8574	8578	lod=11	179
585	chr4	8579	8584	lod=17	238
585	chr4	8591	8593	lod=10	166
585	chr4	8687	8701	lod=37	342
585	chr4	8741	8746	lod=16	229
585	chr4	8783	8785	lod=10	166
585	chr4	8840	8847	lod=11	179
585	chr4	9010	9015	lod=18	245

The right-hand panel shows the 'History' section with two jobs:

- 4: UCSC Main on D. melanogaster: chainDroYak2 (genome)** - 90.6 MB
- 3: UCSC Main on D. melanogaster: phastConsElements15way (genome)** - ~1,300,000 lines, format: tabular, database: dm3
- 2: UCSC Main on D. melanogaster: phastCons15way (genome)** - 100,004 lines, 9 comments, format: wig, database: dm3

The job details for '2: UCSC Main on D. melanogaster: phastCons15way (genome)' are expanded, showing a track type of 'wiggle_0' and a name of 'Conservation'. The output date is 2013-02-16 19:56. The chromosome is specified as chr4, and the position is 1-1351857. A note indicates the data has been compressed.



UCSC API

- No officially supported API exists
- Third-party API in Ruby

Mishima *et al.* *BMC Bioinformatics* 2012, **13**:240
<http://www.biomedcentral.com/1471-2105/13/240>



SOFTWARE

Open Access

The Ruby UCSC API: accessing the UCSC genome database using Ruby

Hiroyuki Mishima^{1*}, Jan Aerts^{2,3}, Toshiaki Katayama⁴, Raoul J P Bonnal⁵ and Koh-ichiro Yoshiura¹

NCBI: Conserved Domains database

Example: View the phylogenetic sequence tree of the voltage gated chloride channel domain, cd00400: Voltage_gated_CIC

Conserved Domains

cd00400: Voltage_gated_CIC

CLC voltage-gated chloride channel. The CIC chloride channels catalyse the selective flow of Cl⁻ ions across cell membranes, thereby regulating electrical excitation in skeletal muscle and the flow of salt and water across epithelial barriers. This domain is found in the halogen ions (Cl⁻, Br⁻ and I⁻) transport proteins of the CIC family. The CIC channels are found in all three kingdoms of life and perform a variety of functions including cellular excitability regulation, cell volume regulation, membrane potential stabilization, acidification of intracellular organelles, signal transduction, transepithelial transport in animals, and the extreme acid resistance response in eubacteria. They lack any structural or sequence similarity to other known ion channels and exhibit unique properties of ion permeation and gating. Unlike cation-selective ion channels, which form oligomers containing a single pore along the axis of symmetry, the CIC channels form two-pore homodimers with one pore per subunit without axial symmetry. Although lacking the typical voltage-sensor found in cation channels, all studied CIC channels are gated (opened and closed) by transmembrane voltage. The gating is conferred by the permeating ion itself, acting as the gating charge. In addition, eukaryotic and some prokaryotic CIC channels have two additional C-terminal CBS (cystathionine beta synthase) domains of putative regulatory function.

Conserved Features/Sites

Cl⁻ selectivity pore gating Cl⁻ binding dimer interface

Feature 1: Cl⁻ selectivity filter

Evidence:

- Comment: Mutations in these residues affect channel selectivity
- Citation: PMID 11796999

[Download Cn3D for Viewing 3D Structure](#) [Scroll to Sequence Alignment Display](#)

cd00400 is part of a hierarchy of related CD models. Use the graphical representation to navigate this hierarchy. cd00400 is a member of the superfamily c02915.

cd00400 Sequence Cluster **Sub-family Hierarchy**

[Zoom In](#) [Detailed View](#) [Interactive Display with CDTree](#)

cd00400 Voltage_gated_CIC

- cd01031 Eric
- cd01033 CIC_like
- cd01034 Eric_like
- cd01036 CIC_euk
- cd03683 CIC_1_like
- cd03684 CIC_3_like
- cd03685 CIC_6_like
- cd03682 CIC_syc_like

Sequence Alignment

Reformat Format: Hypertext Row Display: up to 10 Color Bits: 2.0 bit Type Selection: the most diverse members

```

1KPL_A 43 GTLTGLVGVAFKAVSNVQNRigalqv---adhafllWFLAFILSALLAMVGYFLvrk---faFEAGSGIPFIEG 114
q1 22295201 19 GLLSGVTITAFYLGKLGKGFALLWgtdpsqvaiatwhehlypYIPLVTAFGSLLVGLSVKfkg---akGLADAIALRHG 94
    
```



NCBI Taxonomy: one of the largest databases with organism names



Taxonomy

The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.

NCBI Taxonomy Browser

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search for as lock

Display 3 levels using filter: none

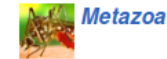
Sus scrofa

Taxonomy ID: 9823
 Genbank common name: **pig**
 Inherited blast name: **even-toed ungulates**
 Rank: species
 Genetic code: [Translation table 1 \(Standard\)](#)
 Mitochondrial genetic code: [Translation table 2 \(Vertebrate Mitochondrial\)](#)
 Other names:
 common name: **wild boar**
 common name: **swine**
 common name: **pigs**
 authority: **Sus scrofa Linnaeus, 1758**

[Lineage\(full \)](#)
[cellular organisms](#); [Eukaryota](#); [Opisthokonta](#); [Metazoa](#); [Eumetazoa](#); [Bilateria](#); [Deuterostomia](#); [Chordata](#); [Cranialia](#); [Vertebrata](#); [Gnathostomata](#); [Teleostomi](#); [Euteleostomi](#); [Sarcopterygii](#); [Tetrapoda](#); [Amniota](#); [Mammalia](#); [Theria](#); [Eutheria](#); [Laurasiatheria](#); [Cetartiodactyla](#); [Suina](#); [Suidae](#); [Sus](#)

Entrez records		
Database name	Subtree links	Direct links
Nucleotide	509,509	508,796
Nucleotide EST	1,669,420	1,669,337
Nucleotide GSS	1,162,428	1,162,428
Protein	42,352	41,853
Structure	606	602
Genome	1	1
Popset	384	376
SNP	566,003	566,003
GEO Datasets	6,020	5,997
UniGene	50,106	50,106
UniSTS	13,987	13,987
PubMed Central	1,250	1,125
Gene	33,533	33,507
OMIA	221	-
SRA Experiments	1,467	1,438
Probe	21,422	21,422
Assembly	6	-
Bio Project	269	266
Bio Sample	1,936	1,905
Bio Systems	1,391	1,391
dbVar	88	88
GEO Profiles	11,692	11,692
Protein Clusters	13	13
Taxonomy	12	1

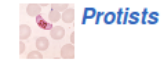
Ensembl: Main databases



Metazoa



Plants



Protists



Fungi



Bacteria

e! or browse vertebrate genomes in Ensembl

■ 'Ensembl' – vertebrate-oriented

- Human, Mouse, Pig, Chicken, Anole lizard, Zebrafish

■ Ensembl genomes

- Metazoans
 - Invertebrates: insects, other arthropods, nematodes, sea urchin, flatworm, etc.
- Plants
 - Arabidopsis, Corn, Rice, Potato, Tomato, etc.
- Fungi
 - Saccharomyces, Aspergillus,
- Protists
 - Plasmodium, Leishmania, Toxoplasma, Giardia, etc.
- Bacteria
 - 6000 bacterial genomes available



Ensembl

- Data internally stored and queried through MySQL databases
- APIs exist to connect to these dbs directly (Perl-API)
- You can download the db, and install locally
- Much of the annotation data (gtf, bed) and sequence data (fasta) can be downloaded
- Has alternative implementation: BioMart
- Various ways of retrieving data programmatically (Perl API, REST, BioMart API for Perl, R)

Biomart Web Portal

The screenshot displays the BioMart web portal interface. At the top, the browser address bar shows the URL `www.biomart.org/biomart/martview/7d13786eee27a042b50d00b8ab14be13`. The BioMart logo is visible in the upper left. Below the logo, there are navigation buttons for 'New', 'Count', and 'Results', along with utility buttons for 'URL', 'XML', 'Perl', and 'Help'. The main content area is titled 'Please restrict your query using criteria below' and contains several query options:

- REGION:**
 - Chromosome: 1
 - Base pair:
 - Gene Start (bp): 1
 - Gene End (bp): 10000000
 - Band:
 - Band Start: p36.33
 - Band End: p36.33
 - Marker:
 - Marker Start: [empty]
 - Marker End: [empty]
 - Encode type: manual_picks
 - Encode region: 5:131256415:132256414
 - Multiple Chromosomal Regions (Chr:Start:End:Strand):
Chromosome Regions (e.g 1:100:10000:-1,1:100000:200000:1): [empty]
- GENE:** [empty]

On the left side, there are sections for 'Dataset' (Homo sapiens genes (GRCh37.p8)), 'Filters' ([None selected]), 'Attributes' (Ensembl Gene ID, Ensembl Transcript ID), and another 'Dataset' section ([None Selected]).



Biomart Perl API

```
# An example script demonstrating the use of BioMart API.
# This perl API representation is only available for configuration versions >= 0.5
use strict;
use BioMart::Initializer;
use BioMart::Query;
use BioMart::QueryRunner;

my $confFile = "PATH TO YOUR REGISTRY FILE UNDER biomart-perl/conf/. For Biomart Central Registry navigate to
                http://www.biomart.org/biomart/martservice?type=registry";

#
# NB: change action to 'clean' if you wish to start a fresh configuration
# and to 'cached' if you want to skip configuration step on subsequent runs from the same registry
#

my $action='cached';
my $initializer = BioMart::Initializer->new('registryFile'=>$confFile, 'action'=>$action);
my $registry = $initializer->getRegistry;

my $query = BioMart::Query->new('registry'=>$registry, 'virtualSchemaName'=>'default');

    $query->setDataset("hsapiens_gene_ensembl");
    $query->addFilter("hgnc_symbol", ["igf1"]);
    $query->addAttribute("ensembl_gene_id");
    $query->addAttribute("ensembl_transcript_id");

$query->formatter("HTML");

my $query_runner = BioMart::QueryRunner->new();
##### GET COUNT #####
# $query->count(1);
# $query_runner->execute($query);
# print $query_runner->getCount();
#####

##### GET RESULTS #####
# to obtain unique rows only
# $query_runner->uniqueRowsOnly(1);

$query_runner->execute($query);
$query_runner->printHeader();
$query_runner->printResults();
$query_runner->printFooter();
#####
```



Ensembl APIs

- Biomart
 - R
 - Perl
 - Java
 - Python
- Perl API
 - Main
 - Compara
 - Variation
 - Functional Genomics
- REST
 - Independent of programming language



Comparing Genomes: Ensembl

■ Gene trees

- are constructed using a representative protein for every gene in Ensembl: proteins are clustered using [hcluster_sg](#) based on WU-BLAST scores, and each cluster of proteins is aligned using M-Coffee. Finally, TreeBeST is used to produce a gene tree from each multiple alignment, reconciling it with the species tree to call duplication events. Homologues are deduced from these trees. We also determine gene gain and loss events using the CAFE software.

■ ncRNA trees

■ Families

- are constructed by MCL clustering of all Ensembl proteins (potentially several per gene). [Metazoan proteins from UniProtKB SwissProt and SPTREMBL are added to extend the protein set.](#)

■ Whole genome alignments

- [are performed either pairwise between two species using BlastZ-net or translated Blat analysis, or using multiple species.](#)

■ Ancestral sequences

- [are calculated from multi-species whole genome alignments.](#)

■ Conservation scores and constrained elements

- [are calculated from the whole genome multiple alignments.](#)

■ Syntenies

- [are calculated from the pairwise alignments.](#)

■ Stable IDs



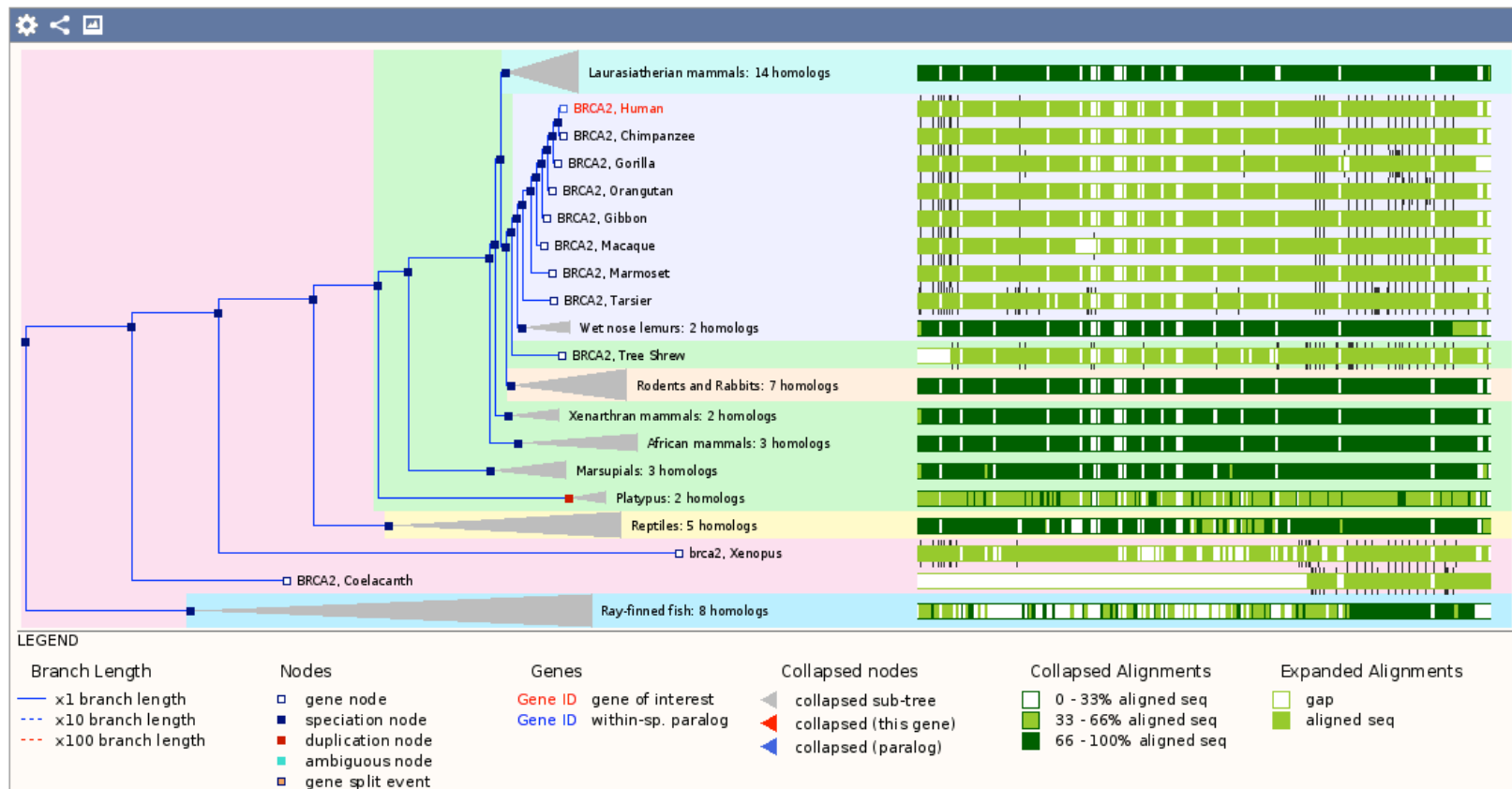
[are provided for Families and Gene Trees and relate exclusively to the content of the Family or the Gene Tree.](#) WELLCOMED GENOME CAMPUS ENGLAND UR

Ensembl: Gene Trees

Gene tree (image)

GeneTree [ENSGT00390000003602](#)

Number of genes	57
Number of speciation nodes	53
Number of duplication	2
Number of ambiguous	1
Number of gene split events	0



biomaRt – R

Load library and create mart object

```
> library(biomaRt)
> ensembl_hs = useMart("ensembl", dataset= "hsapiens_gene_ensembl")
```

Retrieve the Ensembl and Entrezgene IDs based on the Hugo name:

FOXP2

```
> getBM(attributes= c("ensembl_gene_id", "ensembl_gene_id"),
filters= "hgnc_symbol", values=c('FOXP2'), mart=ensembl_hs)
```

Retrieve the protein sequence:

```
> seq = getSequence(id= 'ENSG00000128573', type="ensembl_gene_id",
seqType='peptide', mart=ensembl_hs)
> seq[1,]
```

Retrieve the FOXP2 orthologue from chimpanzee:

```
> getBM(attributes= c("ptroglodytes_homolog_ensembl_gene"), filters=
"ensembl_gene_id", values=c('ENSG00000128573'), mart=ensembl_hs)
```

Ensembl Compara API

Perl API Documentation

Ensembl uses [MySQL](#) relational databases to store its information. A comprehensive set of Application Programme Interfaces (APIs) serve as a middle-layer between underlying database schemes and more specific application programmes. The APIs aim to encapsulate the database layout by providing efficient high-level access to data tables and isolate applications from data layout changes.

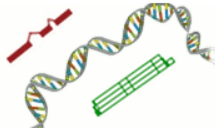


Ensembl's API is written in Perl:

- [Installation instructions](#)
- [full documentation](#) of all modules

Core

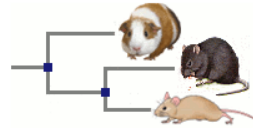
Sequence, genes and other [automated annotation](#)



- [Perl API](#)
- [Database schema](#)
- [Tutorial](#)

Comparative genomics

Homologues, paralogues and protein families



- [Perl API](#)
- [Database schema](#)
- [Tutorial](#)

Variation

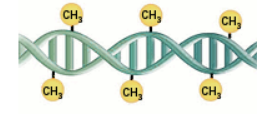
SNPs, somatic mutations and structural variants



- [Perl API](#)
- [Database schema](#)
- [Tutorial](#)

Regulation

Regulatory features, motifs and oligoprobes



- [Perl API](#)
- [Database schema](#)
- [Tutorial](#)

Ensembl Registry

The Registry system allows to tell your programs where to find the Ensembl databases and how to connect to them. It has been implemented for the Ensembl Core and Compara APIs.

[More about the Registry...](#)

Ensembl Software Support

Ensembl is an open project and we would like to encourage correspondence and discussions on any subject on any aspect of Ensembl. Please see the [Ensembl Contacts](#) page for suitable options for getting in touch with us.

If you are interested in undertaking a short-term collaborative project, our "[Geek for a Week](#)" scheme allows developers and researchers to work alongside Ensembl team members.

Ensembl release 70 - January 2013 © [WTSJ](#) / [EBI](#)

[About Ensembl](#) | [Privacy Policy](#) | [Contact Us](#)

[Permanent link](#)



Perl API

```
use Bio::EnsEMBL::Registry;  
Bio::EnsEMBL::Registry->load_registry_from_db(  
    -host => 'ensembl.db.ensembl.org',  
    -user => 'anonymous',  
    -port => 5306);
```

Create adaptor objects

```
my $slice_adaptor =  
    Bio::Ensembl::Registry->get_adaptor($ref_organism, 'core', 'Slice');  
  
my $gene_adaptor =  
    Bio::Ensembl::Registry->get_adaptor($ref_organism, 'core', 'Gene');  
  
my $member_adaptor =  
    Bio::Ensembl::Registry->get_adaptor('Multi', 'compara', 'Member');  
  
my $homology_adaptor =  
    Bio::Ensembl::Registry->get_adaptor('Multi', 'compara', 'Homology');
```



Objects have methods

example of retrieving gene objects by gene symbol

```
my $gene_adaptor =
```

```
    Bio::Ensembl::Registry->get_adaptor($ref_organism, 'core', 'Gene');
```

```
@genes = @{$gene_adaptor->fetch_all_by_external_name('IGF1')};
```



Gene objects have methods too

```
my $gene = $gene_adaptor -> fetch_by_stable_id($stable_id);
```

```
my $start = $gene->start();
```

```
my $end = $gene->end();
```

```
my $chromosome = $gene->seq_region_name();
```

```
my $strand = $gene->strand();
```

Example of retrieving sequence data by coordinates

```
my $slice_adaptor =  
    Bio::Ensembl::Registry->get_adaptor($species, 'core', 'Slice');  
  
my $slice =  
$slice_adaptor->fetch_by_region('chromosome', $chromosome, ($start-1000),  
($end+1000));  
  
my $sequpstream = $slice ->get_repeatmasked_seq()->seq();
```



creating a member object

```
my $member =
```

```
    $member_adaptor -> fetch_by_source_stable_id("ENSEMBLGENE", $gene_id);
```

```
$hum_chrom = $member->chr_name;
```

```
    $hum_chrom_start = $member->chr_start;
```

```
    $hum_chrom_end = $member->chr_end;
```

```
    $descripton = $member->description;
```

```
    my $taxon = $member->taxon;
```



Member objects can be used to retrieve homologies

```
my $homologies = $homology_adaptor >fetch_all_by_Member($member);
```

```
foreach my $homology (@{$homologies}) {...
```

```
}
```



Ensembl – REST API

Ensembl REST API Endpoints

Comparative Genomics

Resource	Description
GET <code>genetree/id/:id</code>	Retrieves Gene Tree dumps for a given Gene Tree stable identifier
GET <code>homology/id/:id</code>	Retrieves homology information by ensembl gene id
GET <code>homology/symbol/:species/:symbol</code>	Retrieves homology information by symbol

Cross References

Resource	Description
GET <code>xrefs/id/:id</code>	Perform lookups of Ensembl Identifiers and retrieve their external cross references in other databases
GET <code>xrefs/name/:species/:name</code>	Performs a lookup based upon the primary accession or display label of an external reference and returning the information we hold about the entry
GET <code>xrefs/symbol/:species/:symbol</code>	Looks up an external symbol and returns all Ensembl objects linked to it. This can be a display name for a gene/transcript/translation, a synonym or an externally linked reference. If a Gene's transcript is linked to the supplied symbol the service will return both Gene and Transcript (it supports transient links).

Features

Resource	Description
GET <code>feature/region/:species/:region</code>	Retrieves multiple types of features for a given region



Example: retrieve the external references of an Ensembl Gene ID

Example Requests

[/xrefs/id/ENSG00000157764?content-type=application/json](#)

Example output

Perl

Python

Ruby

Curl

Wget

```
1. wget -q --header='Content-type:application/json' 'http://beta.rest.ensembl.org/xrefs/id/ENSG00000157764?' -O -
```

Example Requests

[/xrefs/id/ENSG00000157764?content-type=application/json](#)

Example output

Perl

Python

Ruby

Curl


Wget


```
[
  {
    "display_id": "OTTHUMG00000157457",
    "primary_id": "OTTHUMG00000157457",
    "version": "1",
    "description": null,
    "dbname": "OTTG",
    "synonyms": [],
    "info_type": "NONE",
    "info_text": "",
    "db_display_name": "Havana gene"
  },
  {
    "display_id": "ENSG00000157764",
    "primary_id": "ENSG00000157764",
    "version": "0",
    "description": "",

```



NCBI - Genbank


NCBI


Entrez, The Life Sciences Search Engine

HOME SEARCH SITE MAP
PubMed
All Databases
Human Genome
GenBank
Map Viewer
BLAST

Search across databases [Help](#)

- Result counts displayed in gray indicate one or more terms not found

<div style="display: flex; justify-content: space-between; border-bottom: 1px solid #add8e6; padding: 5px;"> <div style="display: flex; align-items: center;"> 273 PubMed: biomedical literature citations and abstracts 🔍 </div> </div> <div style="display: flex; justify-content: space-between; padding: 5px;"> <div style="display: flex; align-items: center;"> 727 PubMed Central: free, full text journal articles 🔍 </div> <div style="display: flex; align-items: center;"> 8 Site Search: NCBI web and FTP sites 🔍 </div> </div>
--

28
Books: online books
🔍

10
OMIM: online Mendelian Inheritance in Man
🔍

NCBI

- Many databases
- Every ID (pubmed, gene, SNP, protein etc) is a number
- Implemented in SQL databases, data can be downloaded as flat text
- Relationship tables between databases (e.g.: genes annotated to pubmed articles)

NCBI – Homology databases

■ Conserved Domain Database (CDD)

- A collection of sequence alignments and profiles representing protein domains conserved in molecular evolution. It also includes alignments of the domains to known 3-dimensional protein structures in the MMDB database.

■ HomoloGene

- A gene homology tool that compares nucleotide sequences between pairs of organisms in order to identify putative orthologs. Curated orthologs are incorporated from a variety of sources via the Gene database.

■ Protein Clusters

- A collection of related protein sequences (clusters), consisting of Reference Sequence proteins encoded by complete prokaryotic and organelle plasmids and genomes. The database provides easy access to annotation information, publications, domains, structures, external links, and analysis tools.



Homologene provides conserved domain perspective

1: HomoloGene:41. Gene conserved in Euteleostomi

Genes

Genes identified as putative homologs of one another during the construction of HomoloGene.

- BRCA2, *H.sapiens*
breast cancer 2, early onset
- BRCA2, *P.troglodytes*
breast cancer 2, early onset
- BRCA2, *M.mulatta*
breast cancer 2, early onset
- BRCA2, *C.lupus*
breast cancer 2, early onset
- BRCA2, *B.taurus*
breast cancer 2, early onset
- Brca2, *M.musculus*
breast cancer 2
- Brca2, *R.norvegicus*
breast cancer 2
- BRCA2, *G.gallus*
breast cancer 2, early onset
- brca2, *D.rerio*
breast cancer 2, early onset

Protein Alignments

Protein multiple alignment, pairwise similarity scores and evolutionary distances.

[Show Multiple Alignment](#)

Proteins

Proteins used in sequence comparisons and their conserved domain architectures.

- NP_000050.2
3418 aa
- XP_509619.2
3418 aa
- XP_001118184.2
3364 aa
- NP_001006654.2
3446 aa
- XP_002691853.1
3427 aa
- NP_033895.2
3329 aa
- NP_113730.1
3343 aa
- NP_989607.2
3397 aa
- NP_001103864.2
2874 aa

Conserved Domains

Conserved Domains from CDD found in protein sequences by rpsblast searching.

BRCA-2_OB1 (pfam09103)

■ BRCA2, oligonucleotide/oligosaccharide-binding, domain



NCBI- Eutils: esearch

```
wget "http://www.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?  
db=pubmed&term=science[journal]+AND+breast+cancer+AND+2008[pdat]&retmax=100"  
-O output.txt
```

Output:

```
<Id>19008416</Id>  
<Id>18927361</Id>  
<Id>18787170</Id>  
<Id>18487186</Id>  
<Id>18239126</Id>  
<Id>18239125</Id>
```



NCBI Eutils: efetch

```
wget 'http://www.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=19008416' -O
pubmed19008416.txt
```

```
},
  cit {
    title {
      name "Genomic loss of microRNA-101 leads to overexpression of histone
methyltransferase EZH2 in cancer."
    },
```

```
.....
  },
```

```
  abstract "Enhancer of zeste homolog 2 (EZH2) is a mammalian histone
methyltransferase that contributes to the epigenetic silencing of target
genes and regulates the survival and metastasis of cancer cells. EZH2 is
overexpressed in aggressive solid tumors by mechanisms that remain unclear.
Here we show that the expression and function of EZH2 in cancer cell lines
are inhibited by microRNA-101 (miR-101). Analysis of human prostate tumors
revealed that miR-101 expression decreases during cancer progression,
paralleling an increase in EZH2 expression. One or both of the two genomic
loci encoding miR-101 were somatically lost in 37.5% of clinically localized
prostate cancer cells (6 of 16) and 66.7% of metastatic disease cells (22 of
33). We propose that the genomic loss of miR-101 in cancer leads to
overexpression of EZH2 and concomitant dysregulation of epigenetic pathways,
resulting in cancer progression.",
```

```
  mesh {
```



Eutils: easy to integrate in any programming language

```
hjm@ubuntu: ~/shared/UbuntuSharedStuff/AdvBioinf/EUtils
#!/usr/bin/perl -w

use strict;
use LWP::Simple;

#search term to find
my $search_term = 'breast cancer';

#maximum number of results to retrieve
my $retmax = 10;

my $utils = 'http://www.ncbi.nlm.nih.gov/entrez/eutils';
my $db_name = 'pubmed';

# Submit the search and retrieve the XML based results
my $search_result = get( $utils . '/esearch.fcgi?db=' . $db_name . '&retmax=' . $retmax . '&term=' . $search_term );

# paper IDs
my @ids = ($search_result =~ m|.*<Id>(.*?)</Id>.*|g);

#loop through all the ids
# get individual papers (if not, then abstracts)
foreach my $id (@ids) {

    #get all details for each paper - full text if available
    my $fetch = $utils . '/efetch.fcgi?db=' . $db_name . '&id=' . $id;

    #prints out to a xml file (file name generated from database name and current paper ID)
    open(OUTFILE, ">$db_name$id.xml");
    print OUTFILE get($fetch);
    close OUTFILE;
}
~
```



NCBI Eutils: esearch

"[http://www.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=gene&term=FOXP2\[sym\]+AND+human\[orgn\]](http://www.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=gene&term=FOXP2[sym]+AND+human[orgn])" -O -

NCBI Eutils: efetch

```
wget "http://www.ncbi.nlm.nih.gov/entrez/eutils/  
efetch.fcgi?db=homologene&id=33482&rettype=FASTA" -  
O -
```

NCBI Eutils: elink

```
wget "http://www.ncbi.nlm.nih.gov/entrez/eutils/  
elink.fcgi?dbfrom=gene&db=homologens&id=93986" -O -
```