

Gene Family Evolution

Toni Gabaldón
January, 2024

Comparative genomics of unicellular eukaryotes:

Interactions and symbioses

30 Sept – 5 October 2024 | Sant Feliu de Guixols, Spain

Organizers

Alexandra Z Worden

Marine Biological Laboratory, USA
University of Chicago, USA

Co-organizers

Toni Gabaldón

Institute for Research in Biomedicine, ES

Patrick Keeling

University of British Columbia, CA

Julius Lukeš

Institute of Parasitology, Biology Centre, CZ

Gwenaél Piganeau

Observatoire Océanologique de Banyuls, FR

Courtney Stairs

Lund University, SE

All Inclusive Meeting Fee & Key Dates
(includes accommodations, meals, airport bus)

Abstract & Applic. Deadline

Opens 22 Jan., Closes 9 Feb. 2024

Registration deadline

8 March 2024

Student/Postdocs 785 EUR

Academic 990 EUR

Industry 1450 EUR

Note a 21% Spanish VAT (tax) must be collected on top of the above fees

Confirmed speakers

Manny Ares, Jr.

University of California Santa Cruz, US

David Booth

University of California San Francisco, US

Fabien Burki

Uppsala University, SE

ThankGod Ebenezer

University of Cambridge, UK

Matthias Fischer

Max Planck Institute for Medical Research, DE
Royal Netherlands Institute for Sea Res, NL

Isabelle Florent

Muséum National d'Histoire Naturelle, FR

Rachel Foster

Stockholm University, SE

Lillian Fritz-Laylin

University of Mass. Amherst, USA

Filip Husnik

Okinawa Institute of Science & Tech, JP

Anna Karkowska

University of Warsaw, PL

Patrick Keeling

University of British Columbia, CA

Puri Lopez-Garcia

CNRS & Université Paris-Saclay, FR

Varsha Mathur

Oxford University, UK

Kika Pašuthová

Charles University, CZ

Anja Spang

Max Planck Institute for Medical Research, DE
Royal Netherlands Institute for Sea Res, NL

Flora Vincent

European Molecular Biology Laboratory, FR

Iñaki Ruiz-Trillo

CSIC-Universitat Pompeu Fabra, ES

Ross Waller

University of Cambridge, UK

Kenneth Wolfe

University College Dublin, IE

Norico Yamada

University of Konstanz, DE

Contact

Alexandra Worden

cgue@mbi.edu

Early Career Scientist Events

Daily ECS 'Meet the Speakers'
Coffee Breaks

Day 2 Special ECS Gathering &
Select. of Round Table Topics

Day 4 ECS RT Discussions &
Cross Disciplinary Career Talk

ECS Mentors

Wideman (USA), Stairs (SE),
del Campo (ES), Eme (FR)

Meeting Website

<https://go.mbl.edu/cgue>

#CGUE2024

Sponsors



International Society



**Pre-Register Now
and Get 5%
Discount over the
Early Bird Rate!**



**Follow us & Spread
the Word!**



[ESEB2025](#)



[@eseb2025](#)



ESEB2025

BARCELONA • 17-22 August 2025

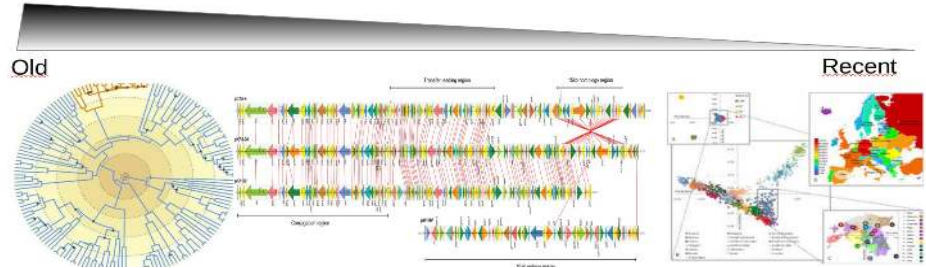
**CONGRESS OF THE EUROPEAN SOCIETY FOR
EVOLUTIONARY BIOLOGY**





Phylogenomics can be regarded as the intersect of the fields of evolution and genomics

Eisen J. (2005)



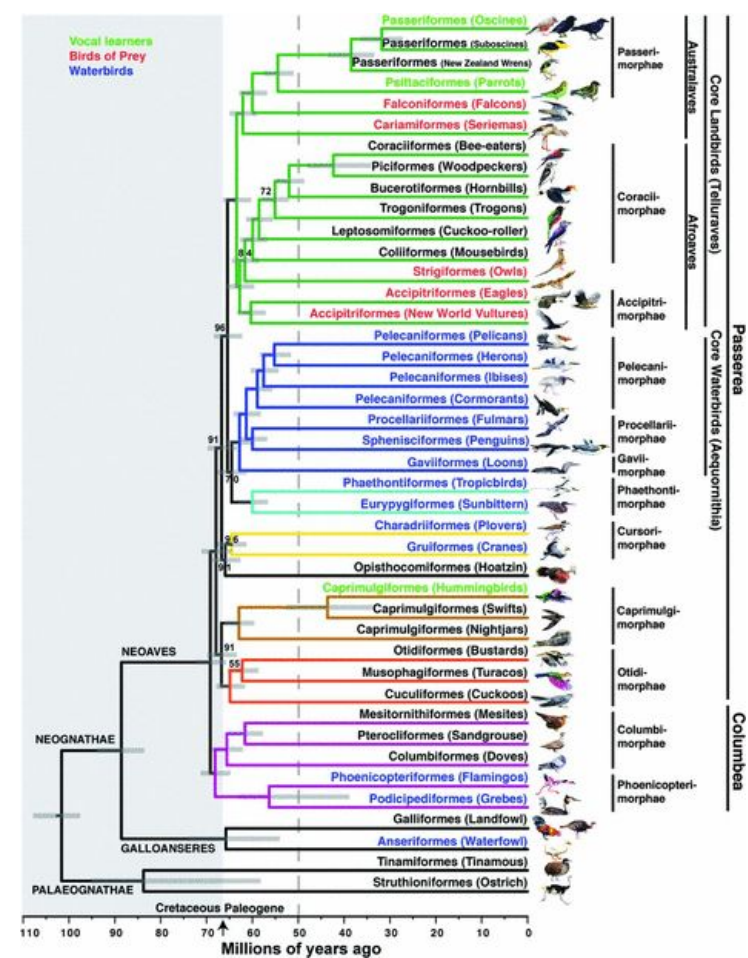
- Origin and early evolution of eukaryotes
- Reticulated evolution in eukaryotes
- Genomic and phenotypic evolution in yeast pathogens (Candida)
- Host-microbiome interactions
- Phylogenomics applied to study of biodiversity and phenotypic transitions

Gene Family Evolution

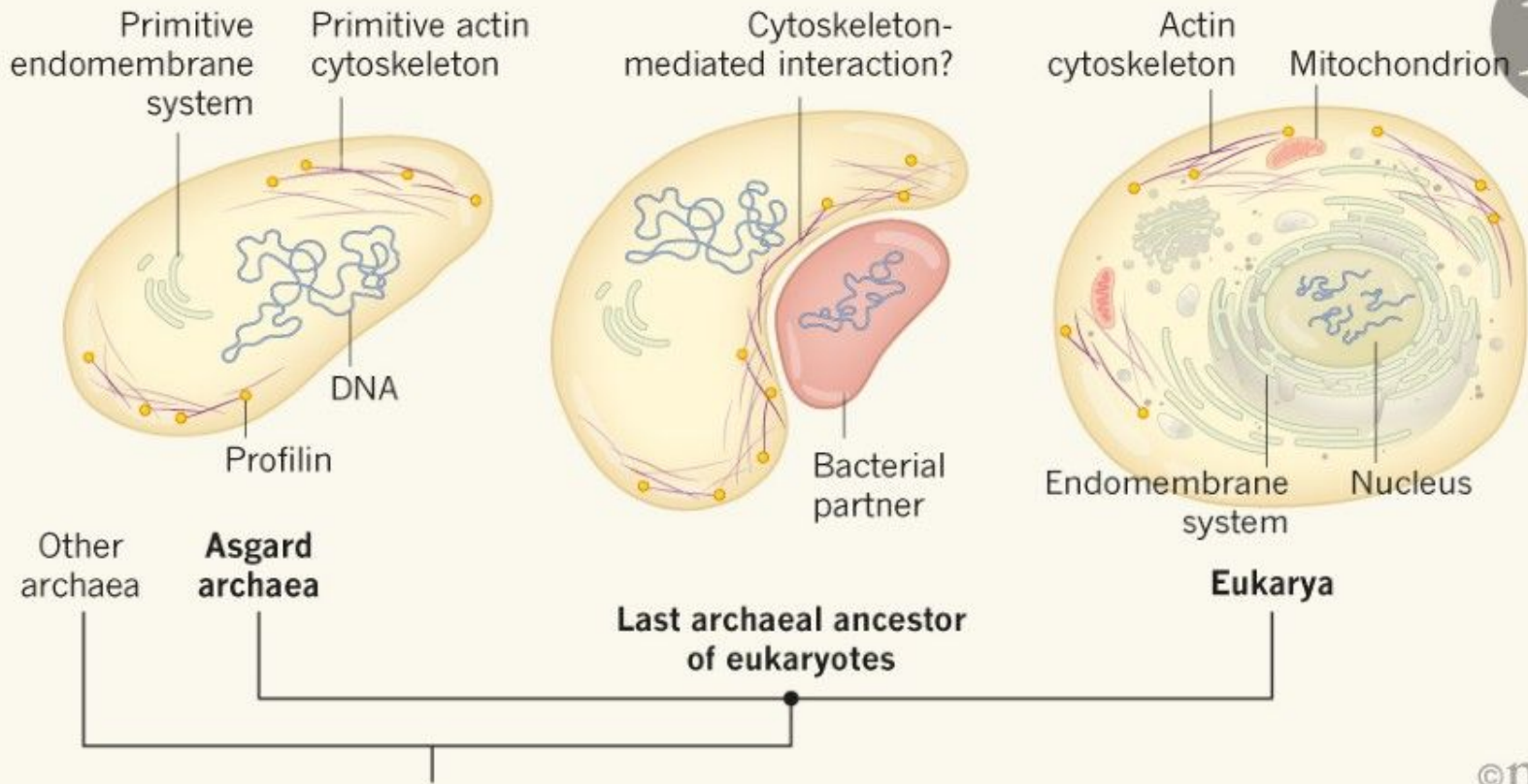
Toni Gabaldón
January, 2024



DO NOT MISS THE FOREST FOR THE TREE



363 bird genomes



Why care about gene family evolution?

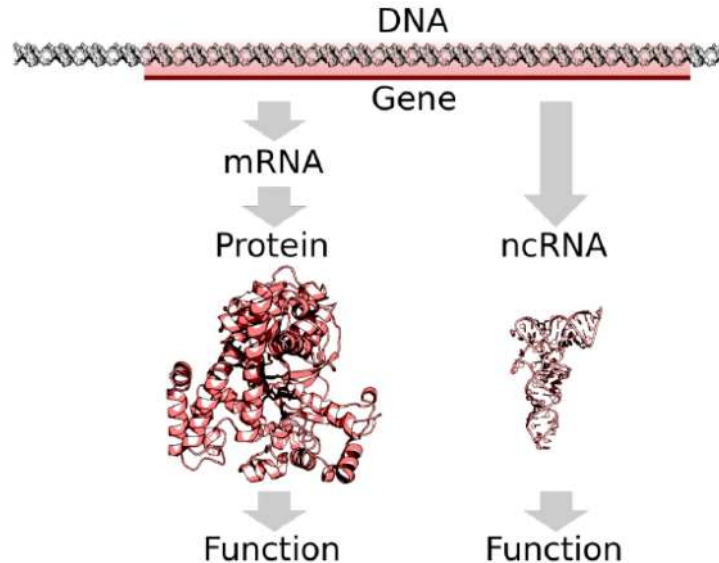
- Gene repertoires encode the phenotypic potentials of a given organism
- Changes in gene content or gene functions underlie phenotypic evolution
- Gene family evolution can reveal how the current diversity of molecular and biological functions has evolved
- Genes can be regarded as evolutionary units that evolve (in part) independently from the species tree
- Genes retain footprints of past evolutionary events
- Functional annotation of genes requires an evolutionary insight
- Co-evolution of gene families reveal functional interactions

But.....what is a gene?



A modern definition:

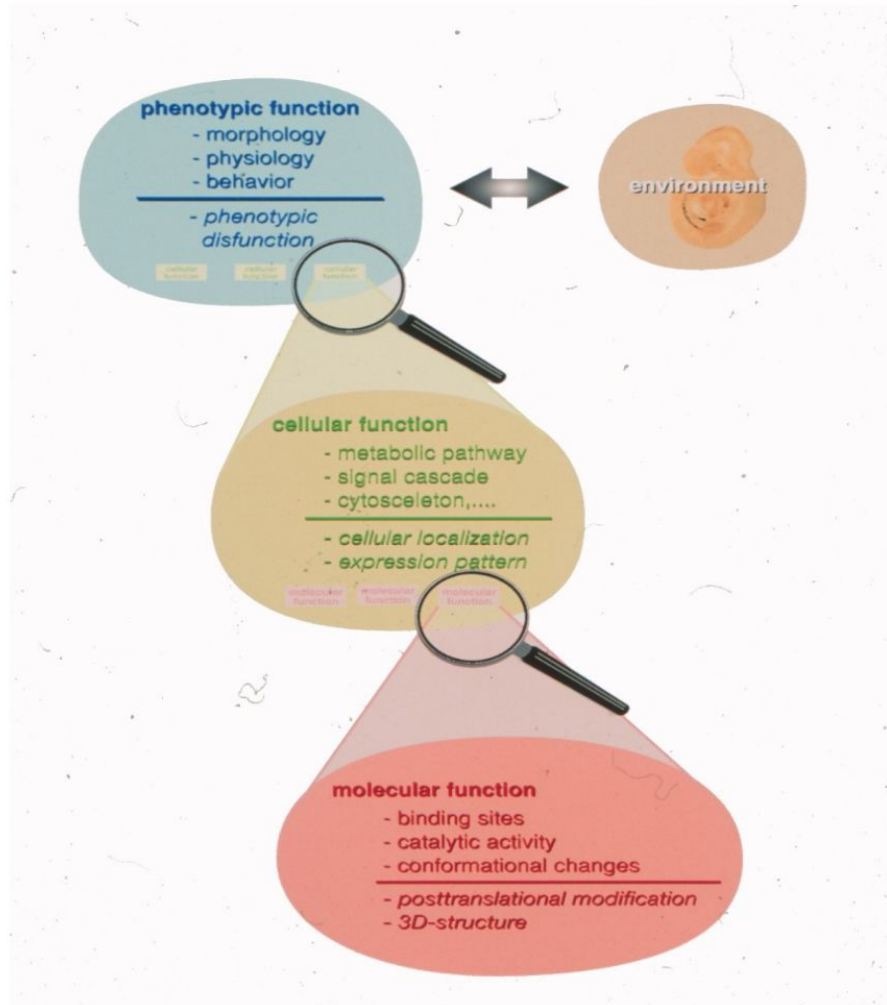
A piece of DNA or RNA which codes for a molecule that has a function



But.....what is a gene function?



Functional roles of genes.



Is difficult to formalize
functional annotations.
Attempts include E.C.
numbers, GO terms, etc

Most annotations are **indirect**

**They are far from optimal,
but better than nothing**



The Gene Ontology

A GO annotation is ...

...a statement that a gene product;

1. has a particular molecular function
or is involved in a particular biological process
or is located within a certain cellular component
2. as described in a particular reference
3. as determined by a particular method

Accession	Name	GO ID	GO term name	Reference	Evidence code
P00505	GOT2	GO:0004069	aspartate transaminase activity	PMID:2731362	IDA



From genome to gene content: gene prediction

- De novo
- Homology-based
- RNAseq based

From genome to gene content: gene prediction

- De novo
- Homology-based
- RNAseq based

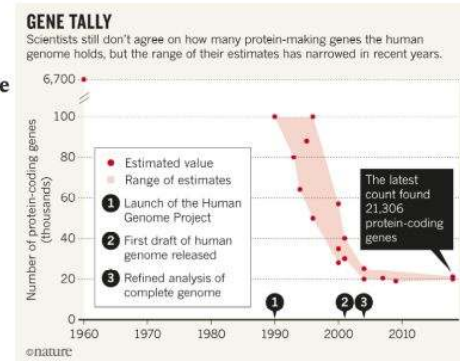
[nature](#) > [news](#) > [article](#)

NEWS | 19 June 2018

Still an issue!

New human gene tally reignites debate

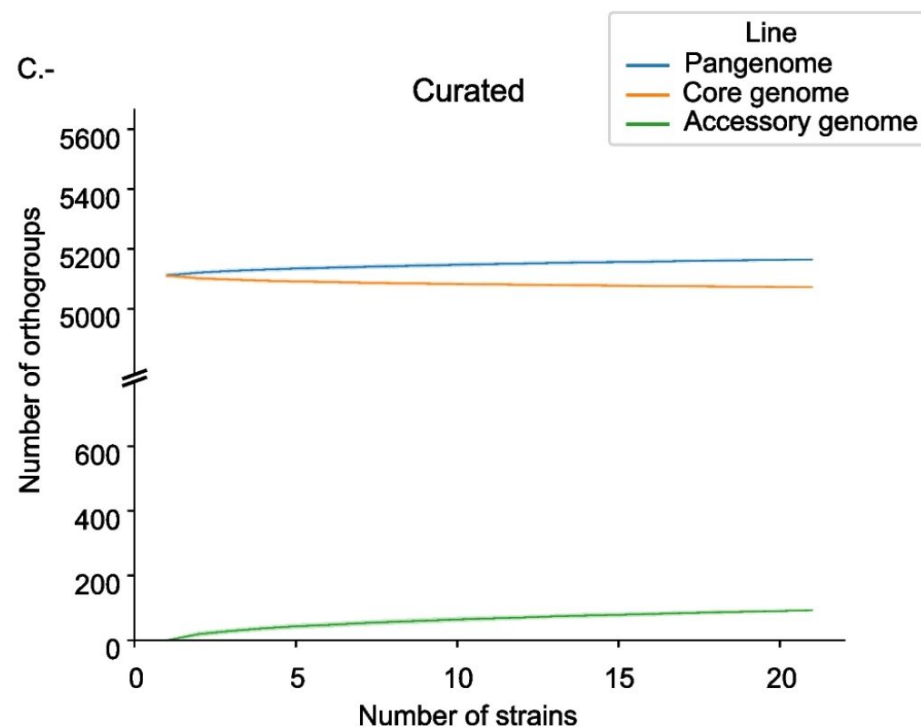
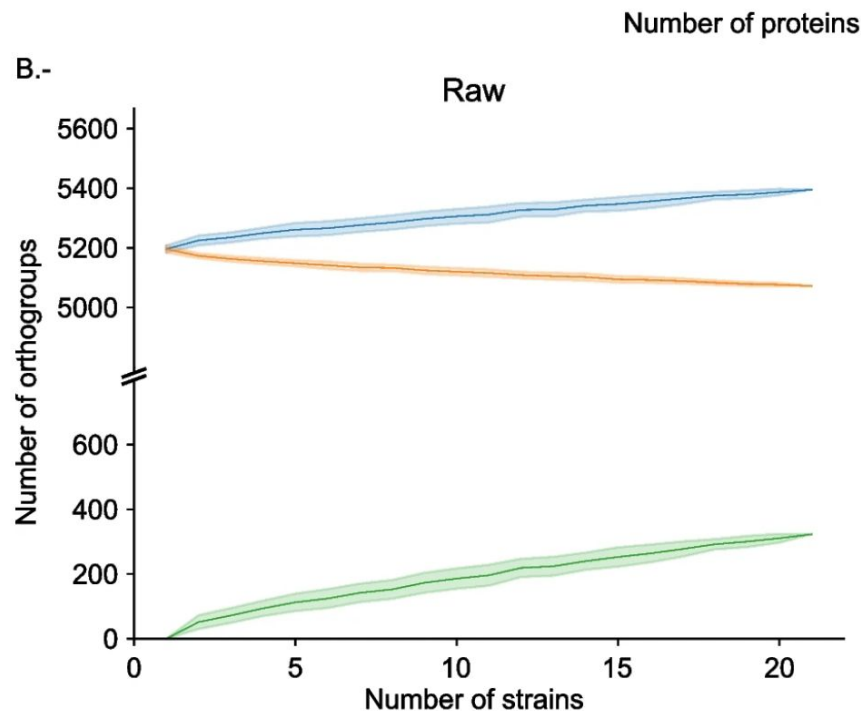
Some fifteen years after the human genome was sequenced, researchers still can't agree on how many genes it contains.



Chromosome-level assemblies from diverse clades reveal limited structural and gene content variation in the genome of *Candida glabrata*

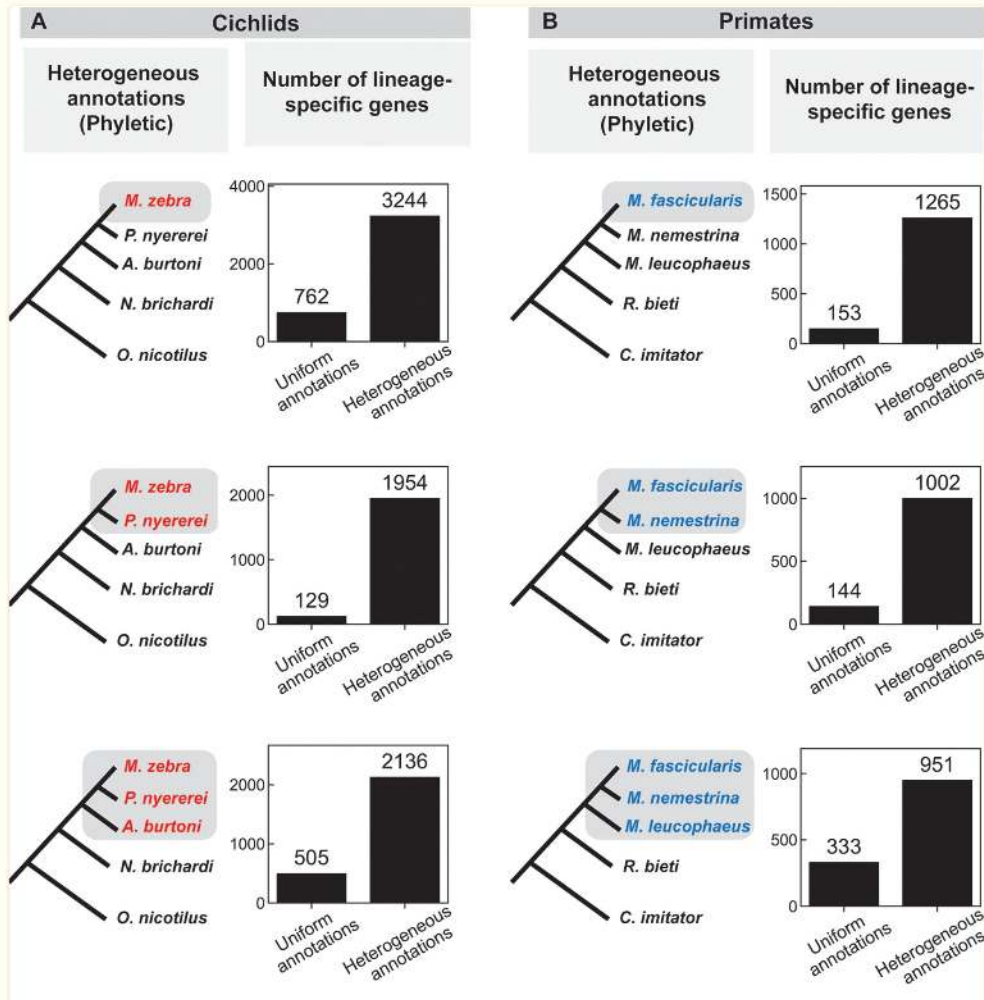
[Marina Marcet-Houben](#), [María Alvarado](#), [Ewa Ksiezopolska](#), [Ester Saus](#), [Piet W. J. de Groot](#) & [Toni Gabaldón](#) 

[BMC Biology](#) 20, Article number: 226 (2022) | [Cite this article](#)

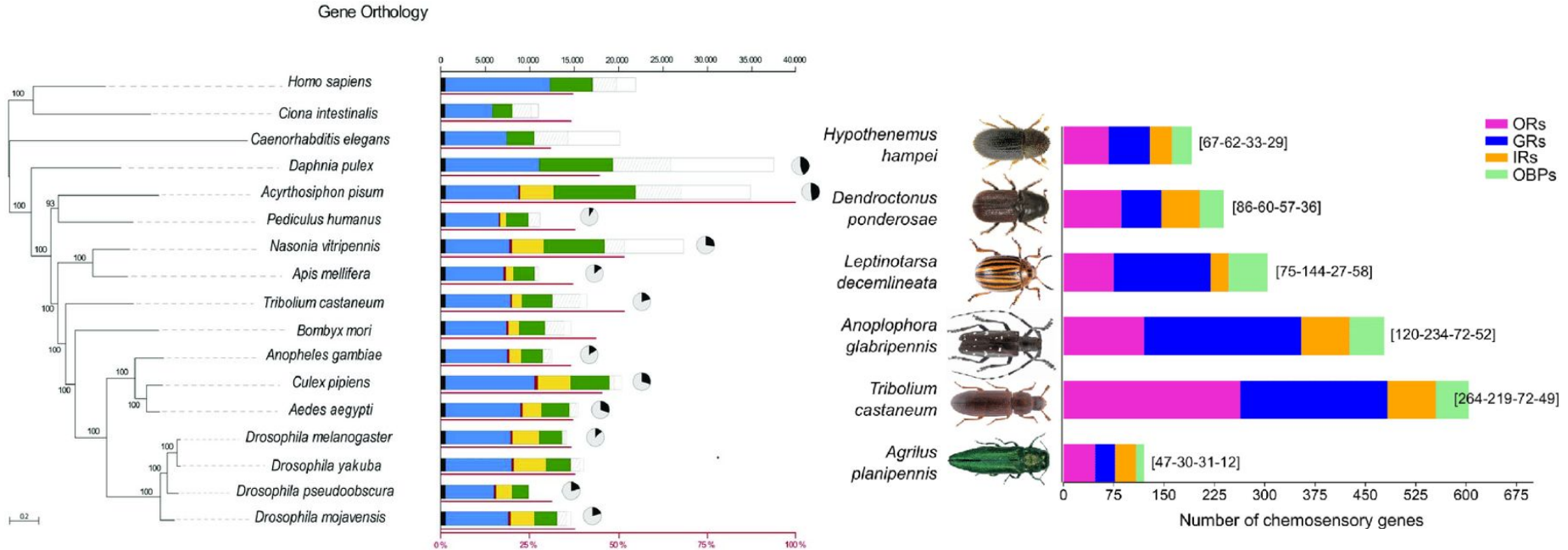


Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes

Caroline M. Weisman ^{5, 6} • Andrew W. Murray • Sean R. Eddy • Show footnotes



Variation of gene content across species



A gene family:

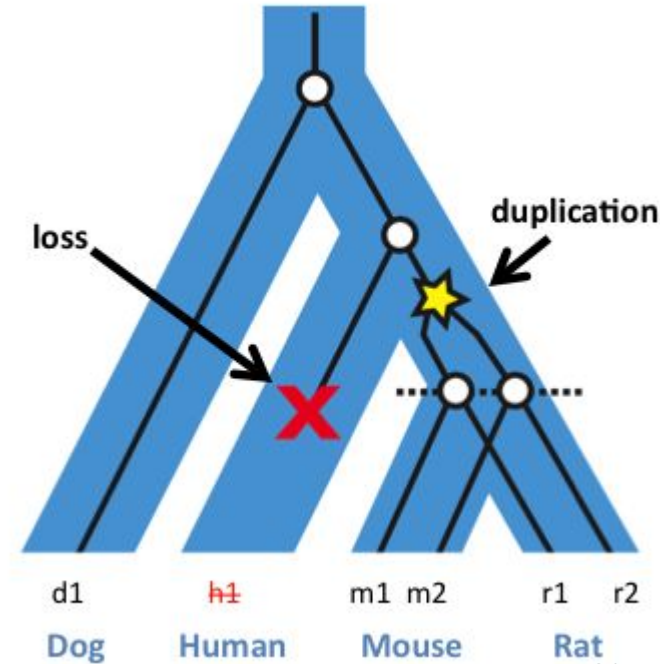
A set of genes with shared ancestry (homologs)

Gene families have hierarchical evolutionary relationship (**best represented by a tree**)

Members of a gene family can be orthologs or paralogs between them

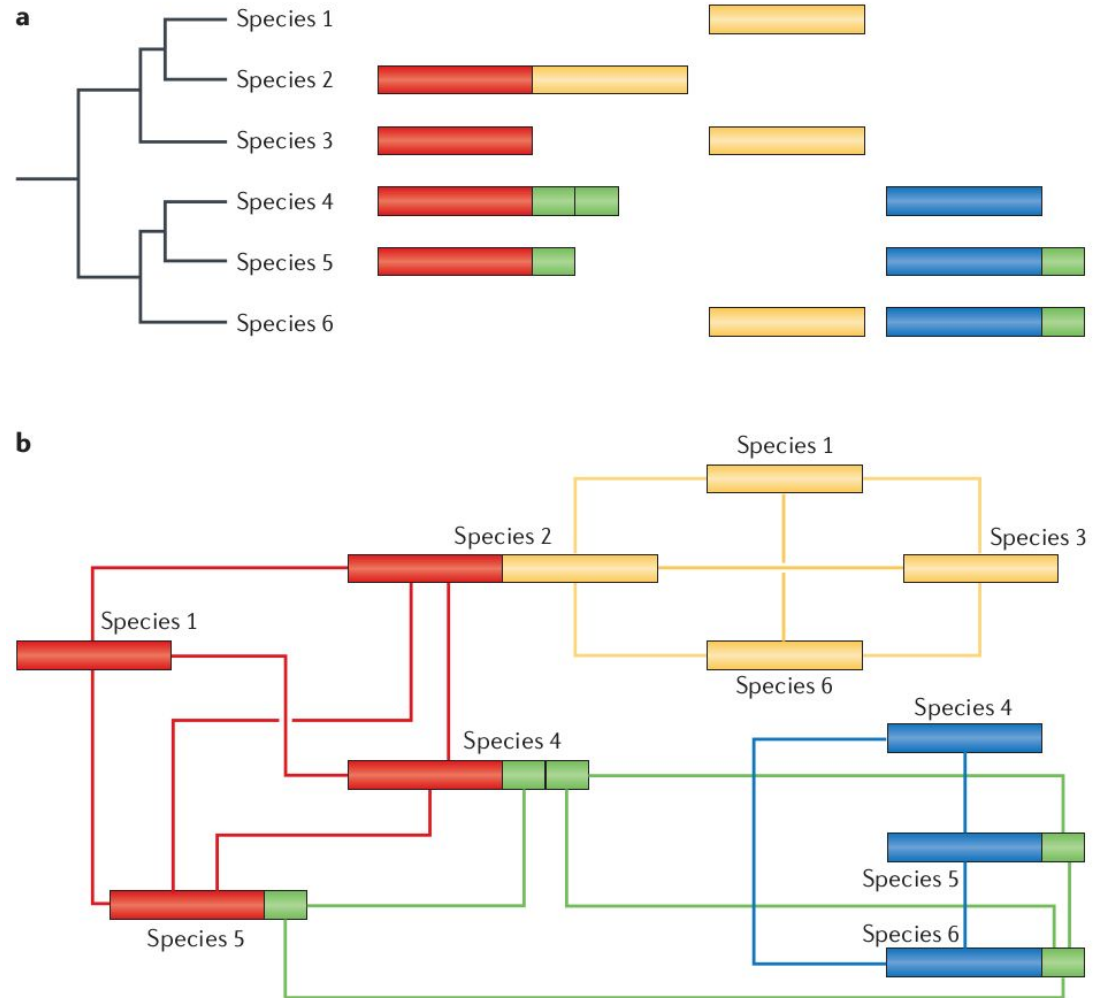
An orthologous group is a (or part of) a gene family

Gene families evolve by duplication and loss (birth and death)



But

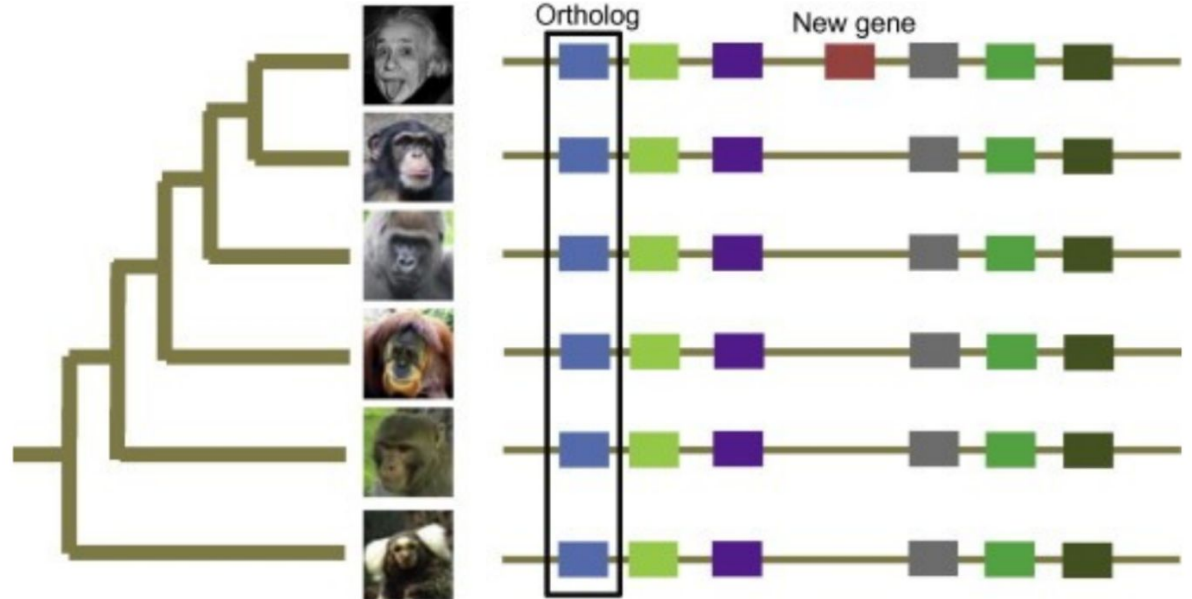
- Genes also evolve by reticulate evolution (HGT and Hybridization)
- Genes also evolve by fusion and fission



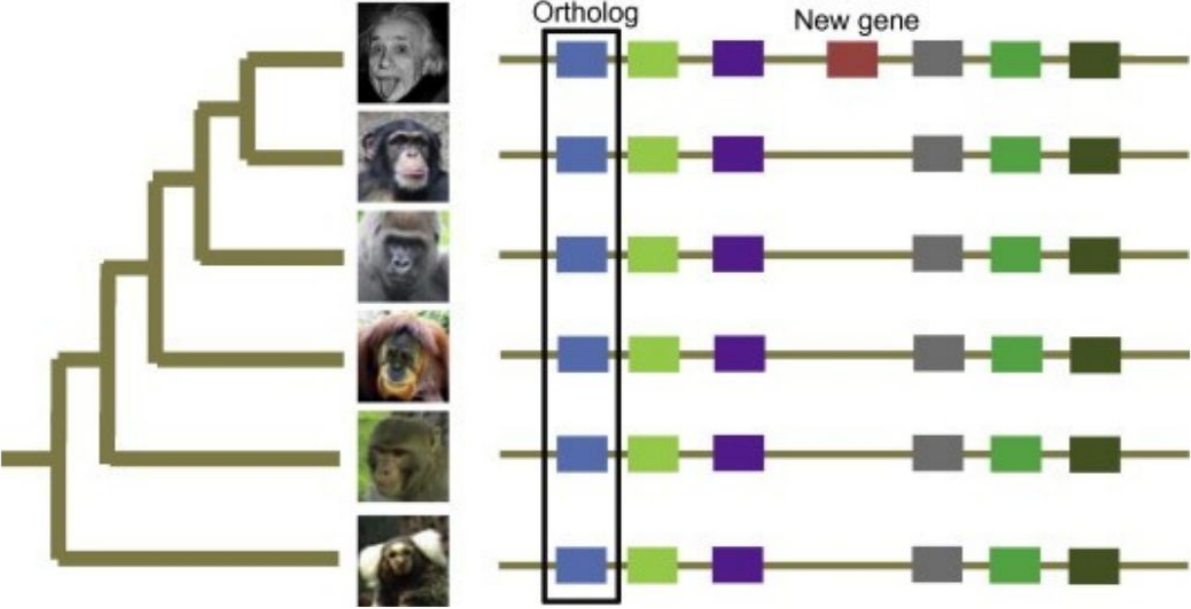
But how they originate in the first place?

Every newly sequenced genomes has predicted “orphans” for which no homolog can be found:

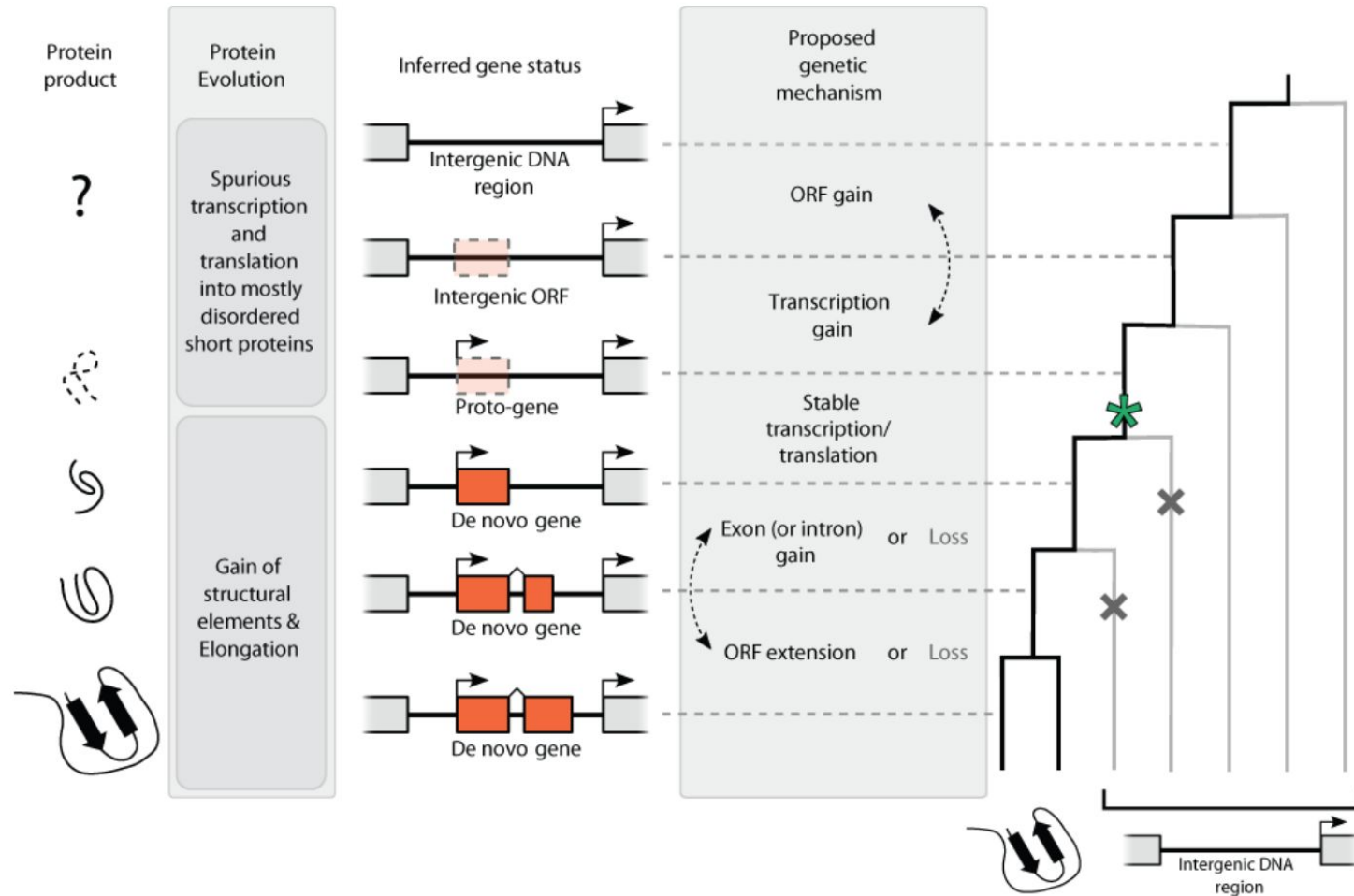
- Spurious predictions?
- Undetected homology?
- Newly emerged gene?



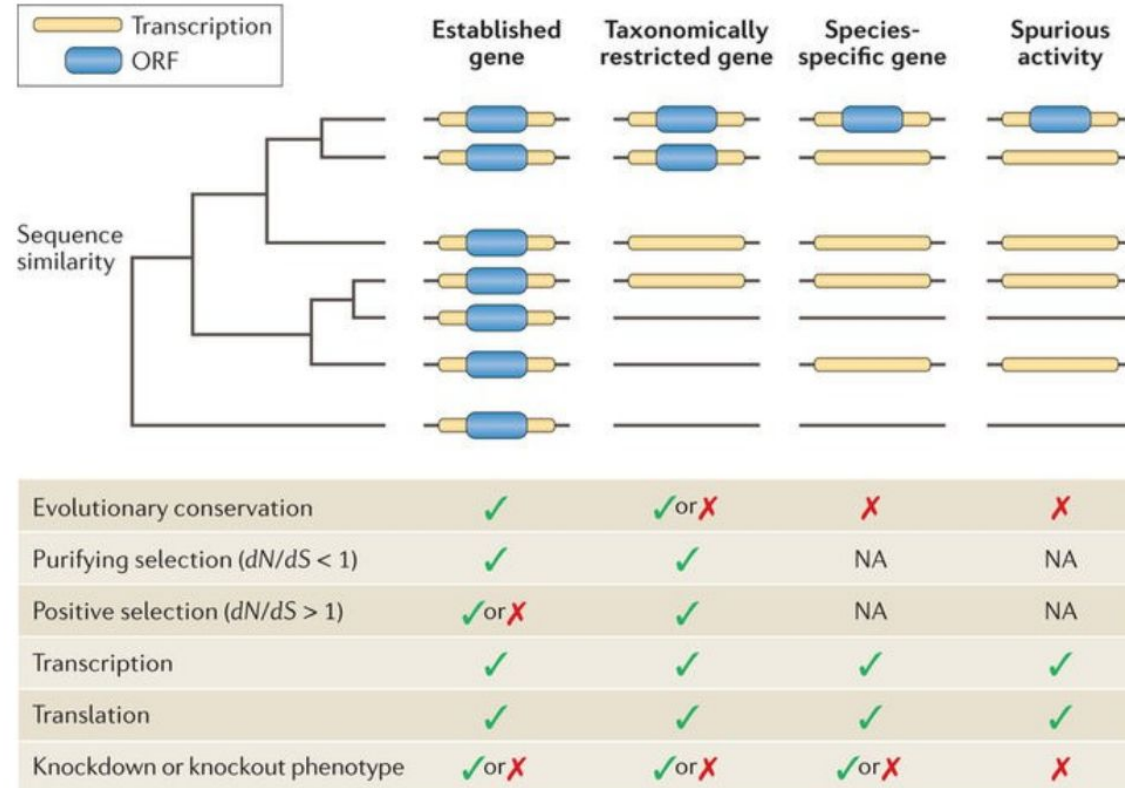
De novo origin of genes.

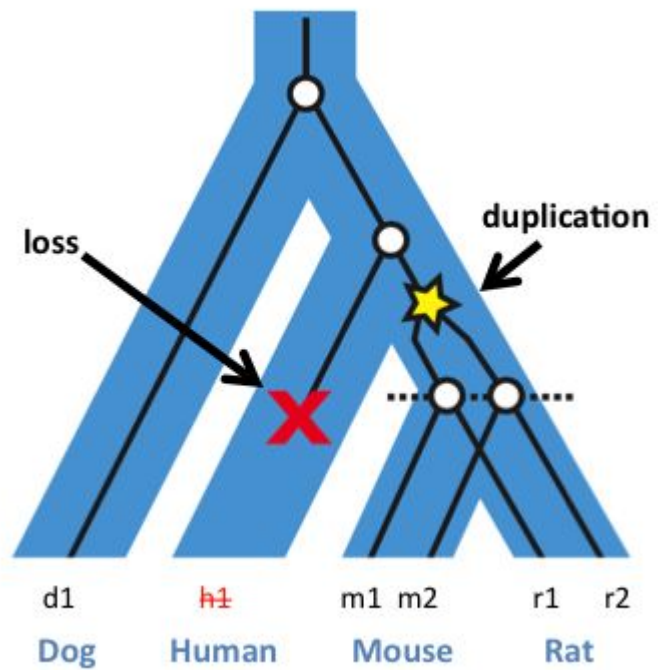


De novo origin of genes.



De novo origin of genes.



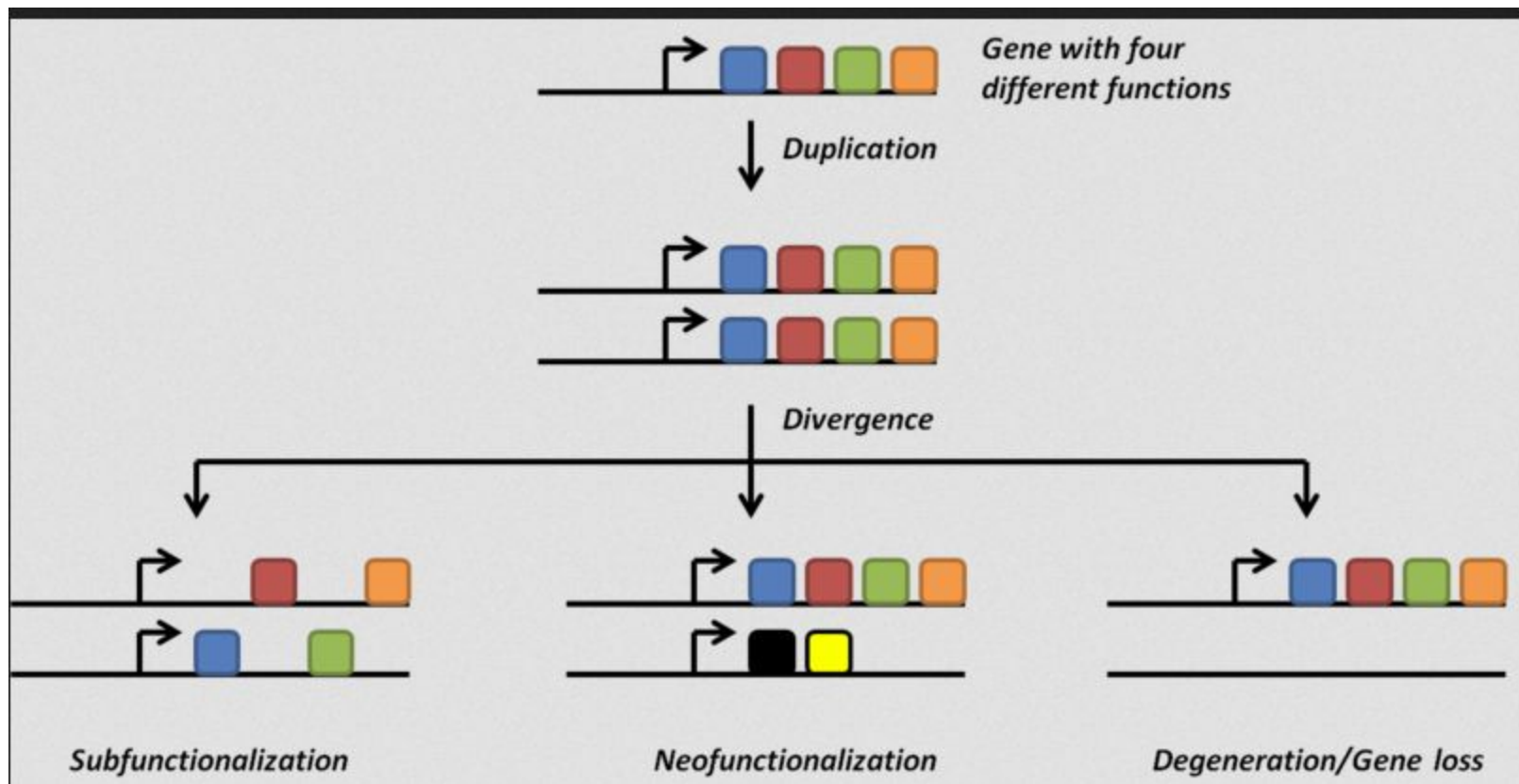


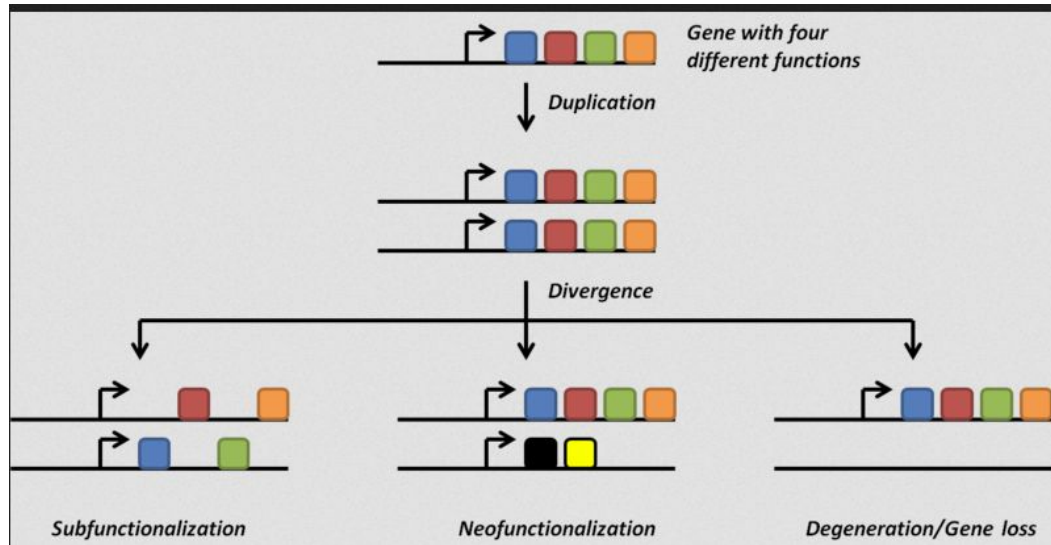
Why genes duplicate?

Spontaneous duplications are common due to:

- DNA breaks and repair: unequal crossing over, replication slippage, ectopic recombination
- Retrotranscription
- Mobile elements
- Aneuploidies, Polyploidies

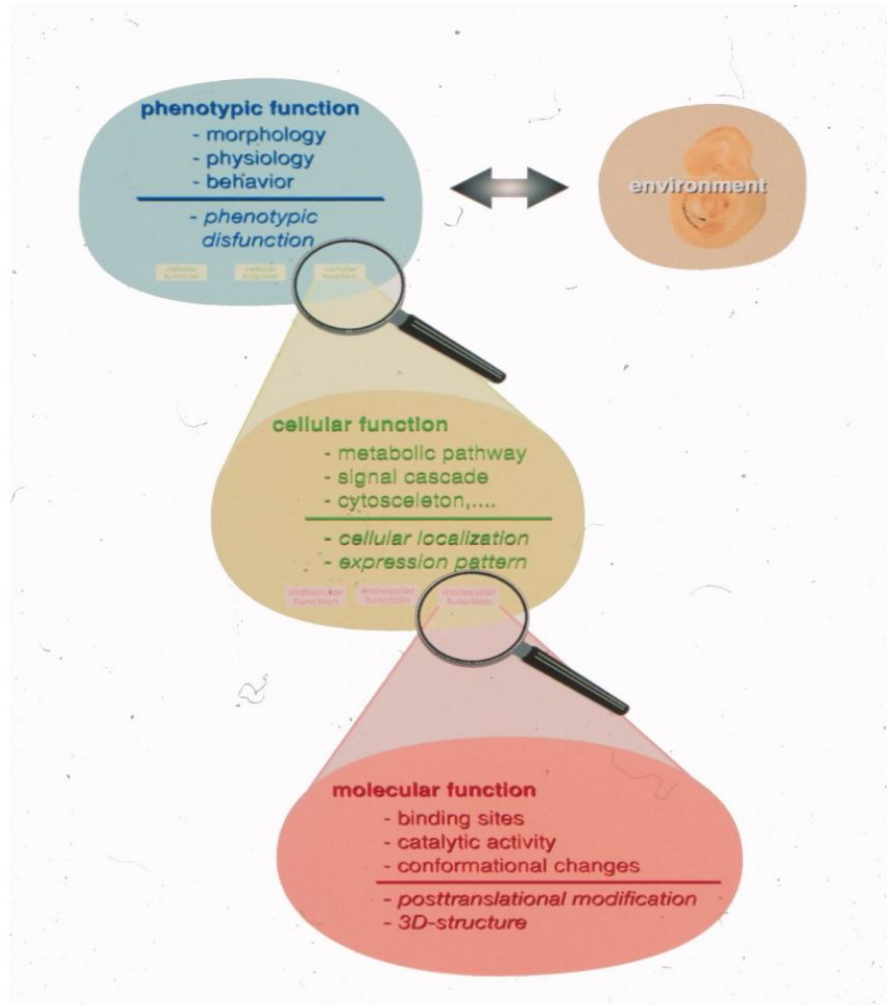
They are common, but the most common outcome of duplication is degeneration (loss) of one of the duplicates

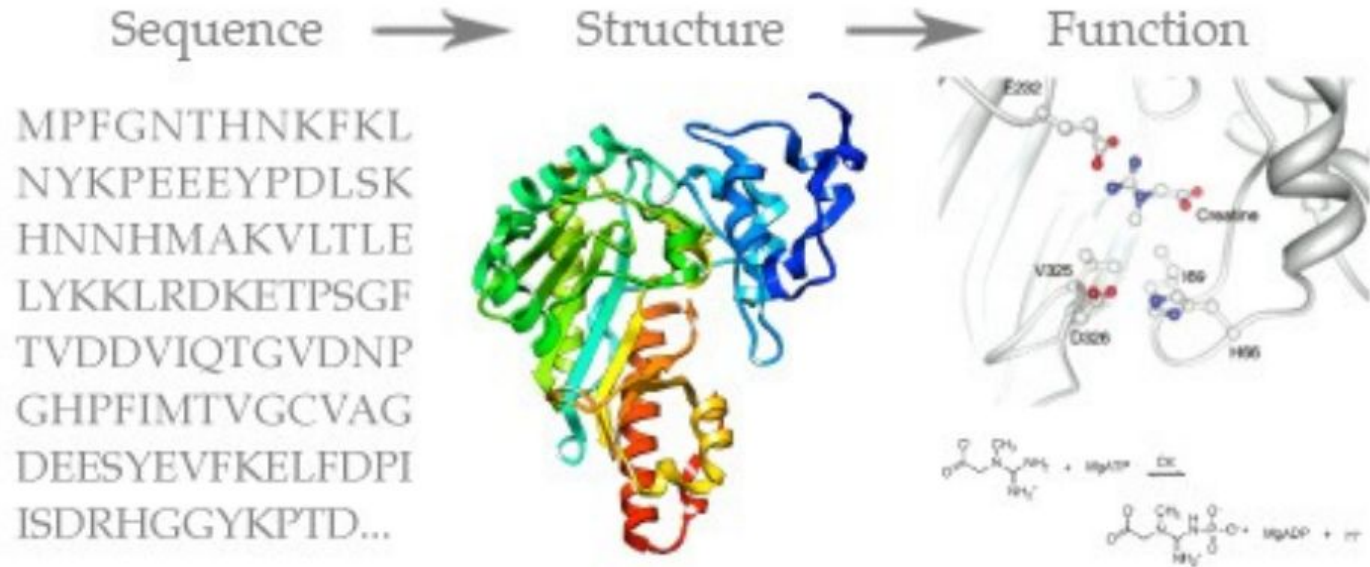




According to this model, gene duplicate retention is associated to functional change

Functional roles of genes.

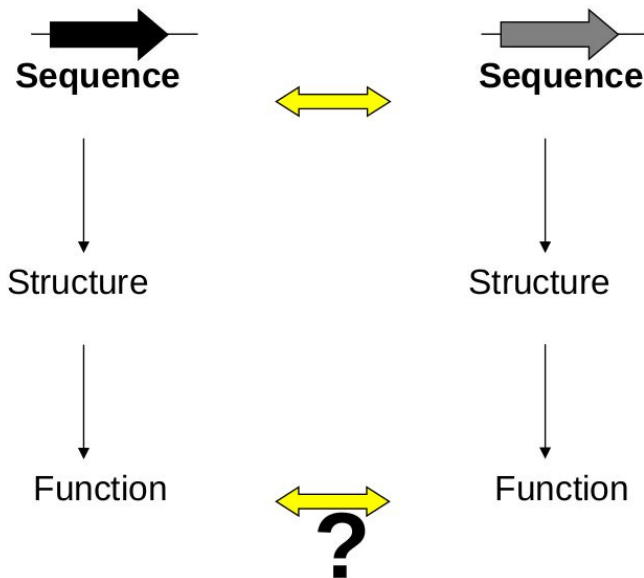




Gene sequence encode protein (or RNA) structure, and its (dynamic) physico-chemical properties, which in turn perform some activity (in a given context)

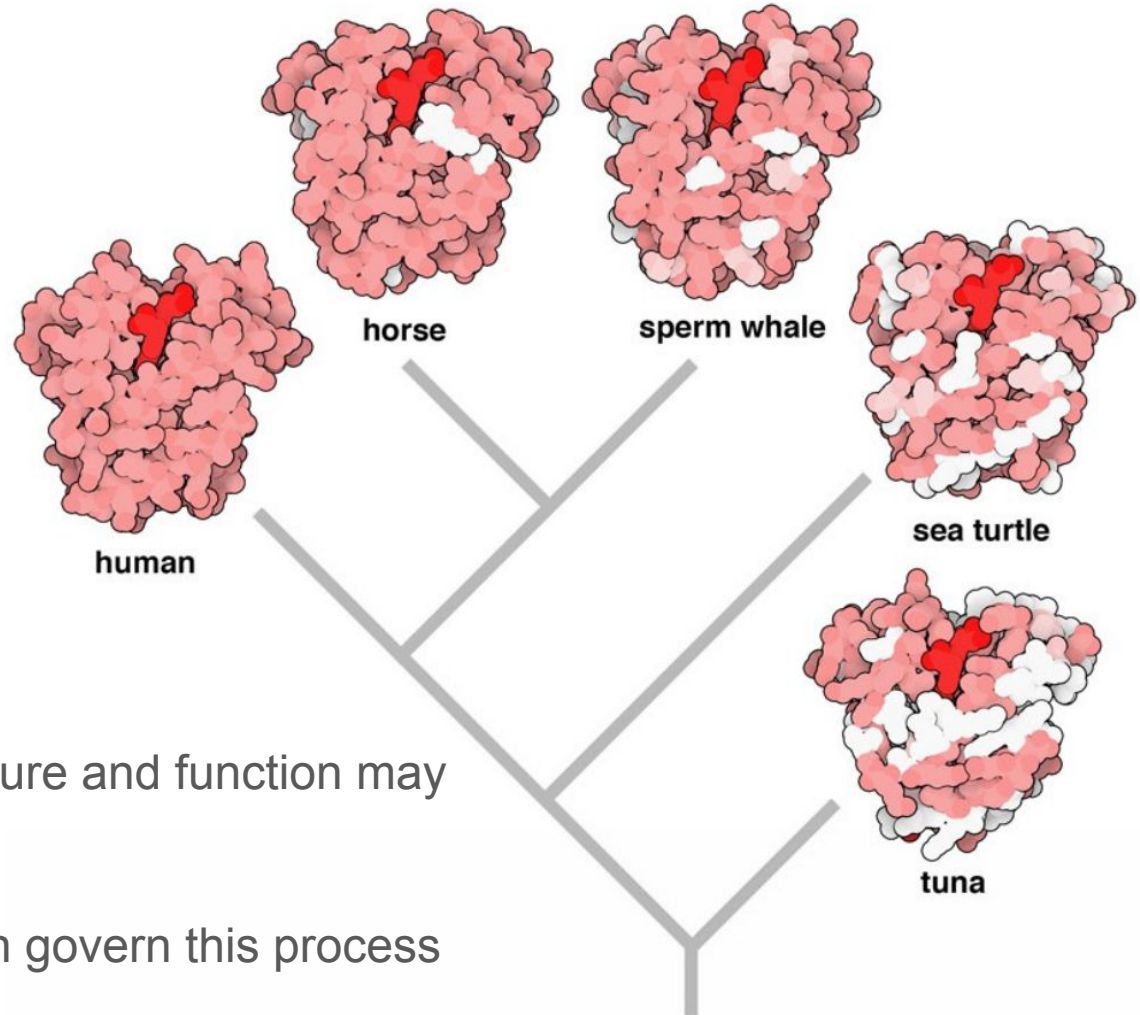
Homology based functional inference.

If sequence determines structure, which determines function, can we predict function from Sequence?



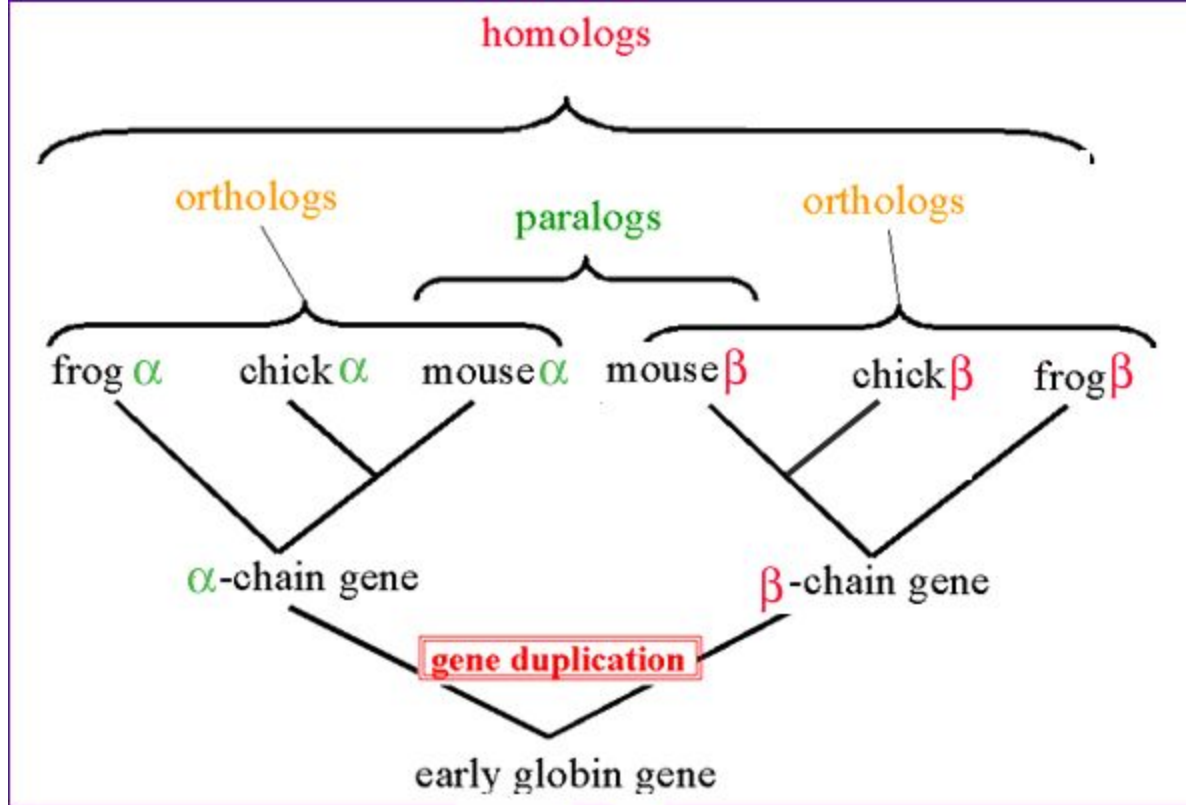
The overwhelming majority of functional annotations are based on this concept

One family one function?



If sequence changes, structure and function may or may not change.

Mutation, drift, and selection govern this process



If duplications promote functional genes, and paralogs are the result of duplications, we expect them to diverge in function.

The orthology conjecture: orthologs, as compared to paralogs, are more likely to share function

Questioning the orthology conjecture

Opinion

Cell
PRESS

How confident can we be that orthologs are similar, but paralogs differ?

Romain A. Studer and Marc Robinson-Rechavi

Department of Ecology and Evolution, Biophore, Lausanne University, CH-1015 Lausanne, Switzerland and Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland

OPEN ACCESS Freely available online

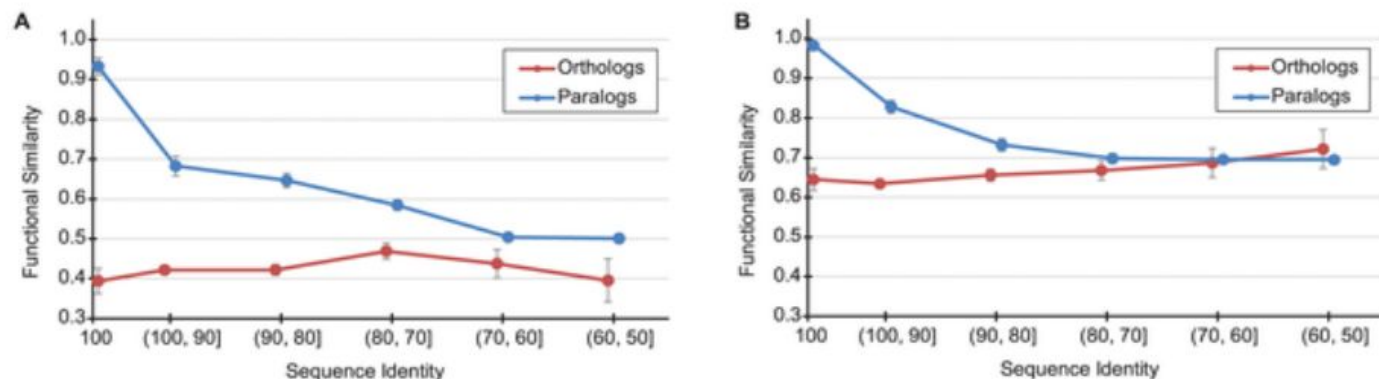
PLoS COMPUTATIONAL BIOLOGY

Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals

Nathan L. Nehrt^{1,3}, Wyatt T. Clark^{1,3}, Predrag Radivojac^{1*}, Matthew W. Hahn^{1,2*}

¹ School of Informatics and Computing, Indiana University, Bloomington, Indiana, United States of America, ² Department of Biology, Indiana University, Bloomington, Indiana, United States of America

Figure 1. The relationship between functional similarity and sequence identity for human-mouse orthologs (red) and all paralogs (blue).



Nehrt NL, Clark WT, Radivojac P, Hahn MW (2011) Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. *PLoS Comput Biol* 7(6): e1002073. doi:10.1371/journal.pcbi.1002073
<http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1002073>

On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report

Paul D. Thomas^{1*}, Valerie Wood², Christopher J. Mungall³, Suzanna E. Lewis³, Judith A. Blake⁴ on behalf of the Gene Ontology Consortium

1 Division of Bioinformatics, Department of Preventive Medicine, University of Southern California, Los Angeles, California, United States of America, **2** Cambridge Systems Biology Centre and Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom, **3** Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, **4** Bioinformatics and Computational Biology, The Jackson Laboratory, Bar Harbor, Maine, United States of America

OPEN ACCESS Freely available online

Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs

Adrian M. Altenhoff^{1,2}, Romain A. Studer^{2,3,4}, Marc Robinson-Rechavi^{2,3}, Christophe Dessimoz^{1,2,5*}

1 ETH Zurich, Department of Computer Science, Zurich, Switzerland, **2** Swiss Institute of Bioinformatics, Lausanne, Switzerland, **3** Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland, **4** Institute of Structural and Molecular Biology, Division of Biosciences, University College London, London, United Kingdom, **5** EMBL-European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom

Nature Reviews Genetics | AOP, published online 4 April 2013; doi:10.1038/nrg3456

PERSPECTIVES

BRIEFINGS IN BIOINFORMATICS, VOL. 12, NO. 5, 442–448
Advance Access published on 22 April 2011

doi:10.1093/bib/bbr022

OPINION

Functional and evolutionary implications of gene orthology

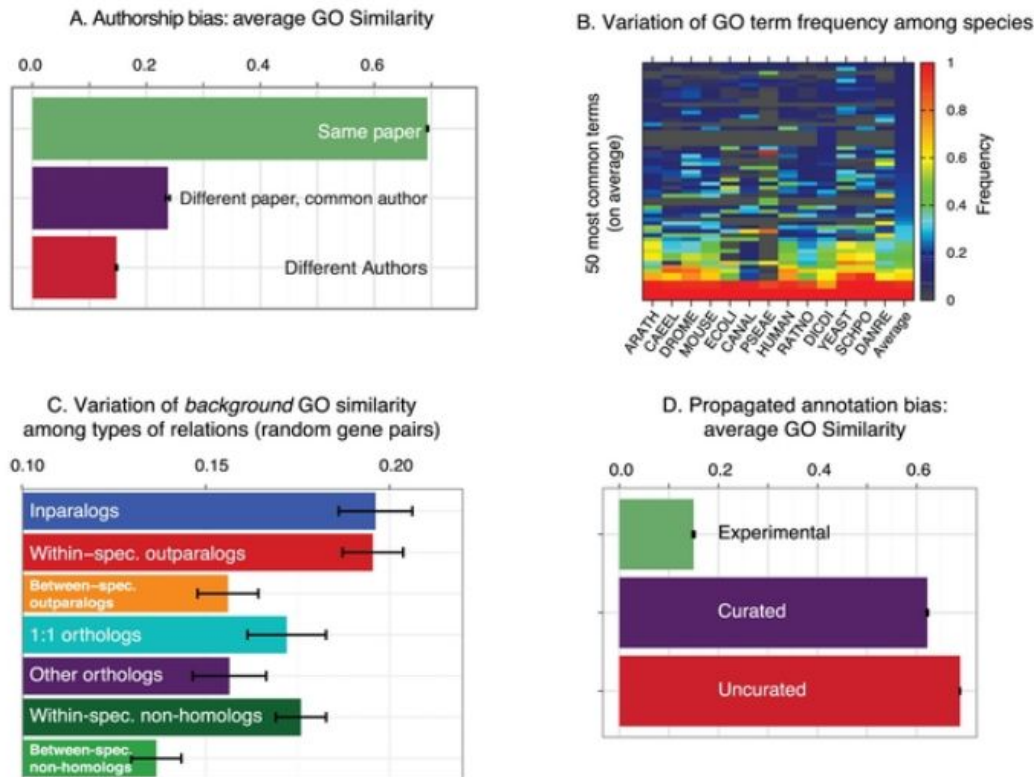
Toni Gabaldón and Eugene V. Koonin

Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication

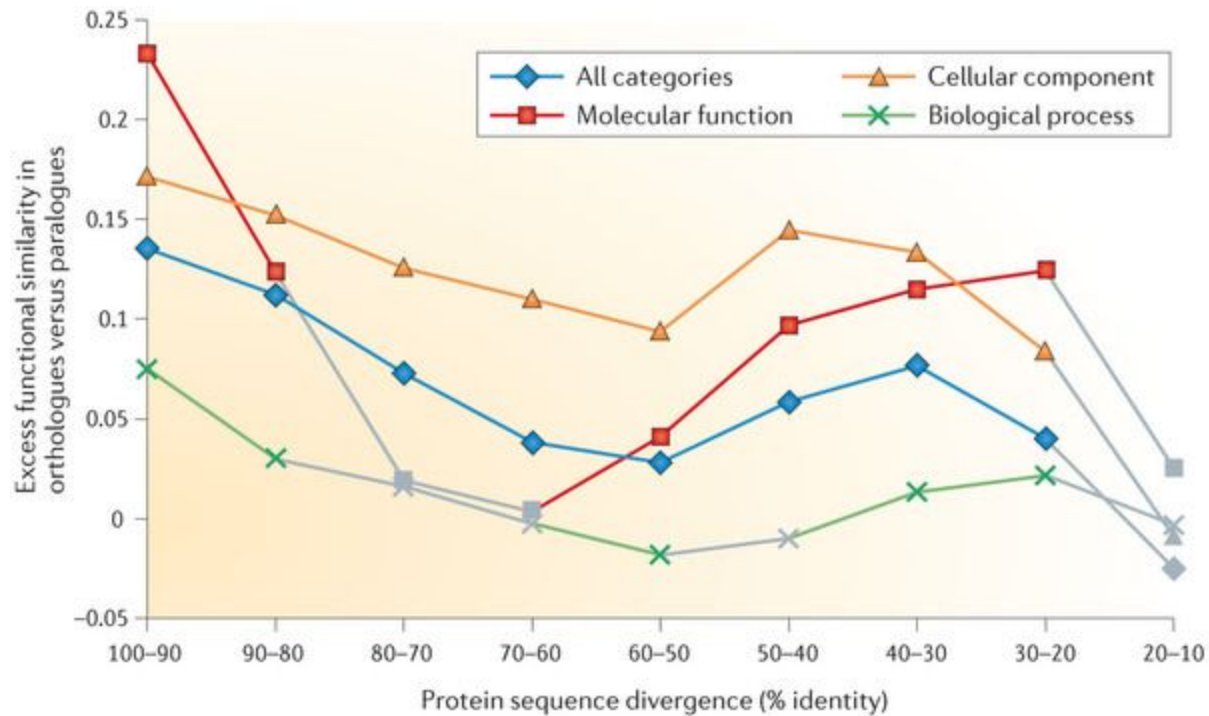
Jaime Huerta-Cepas, Joaquín Dopazo, Martijn A. Huynen and Toni Gabaldón

Submitted: 19th January 2011; Received (in revised form): 22nd March 2011

Figure 1. Potential confounding factors in GO analyses.



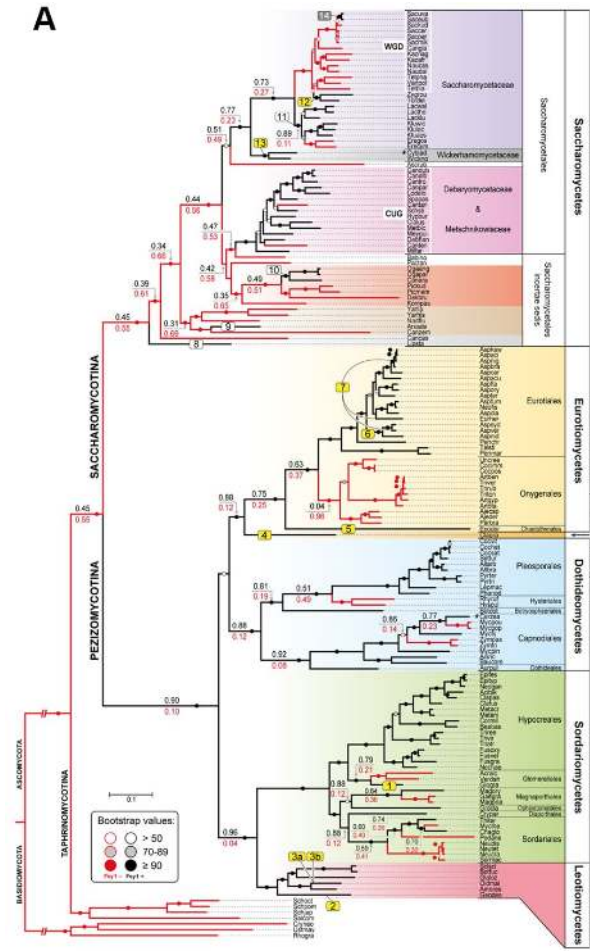
Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C (2012) Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. *PLoS Comput Biol* 8(5): e1002514. doi:10.1371/journal.pcbi.1002514 <http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1002514>



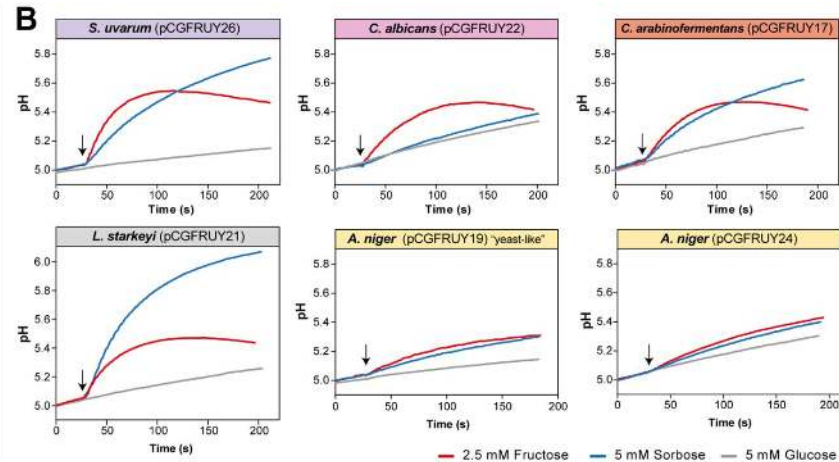
Nature Reviews | Genetics

Functional divergence through speciation

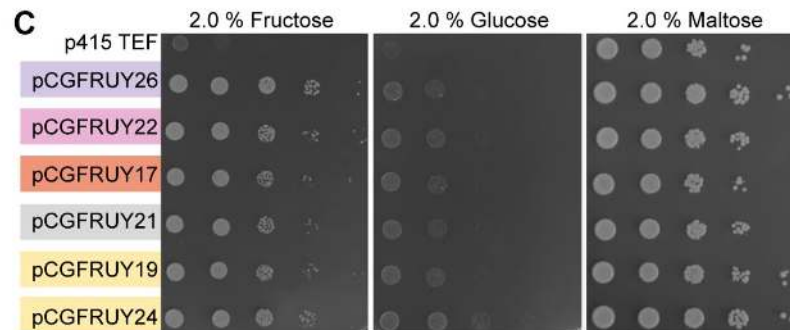
A



B



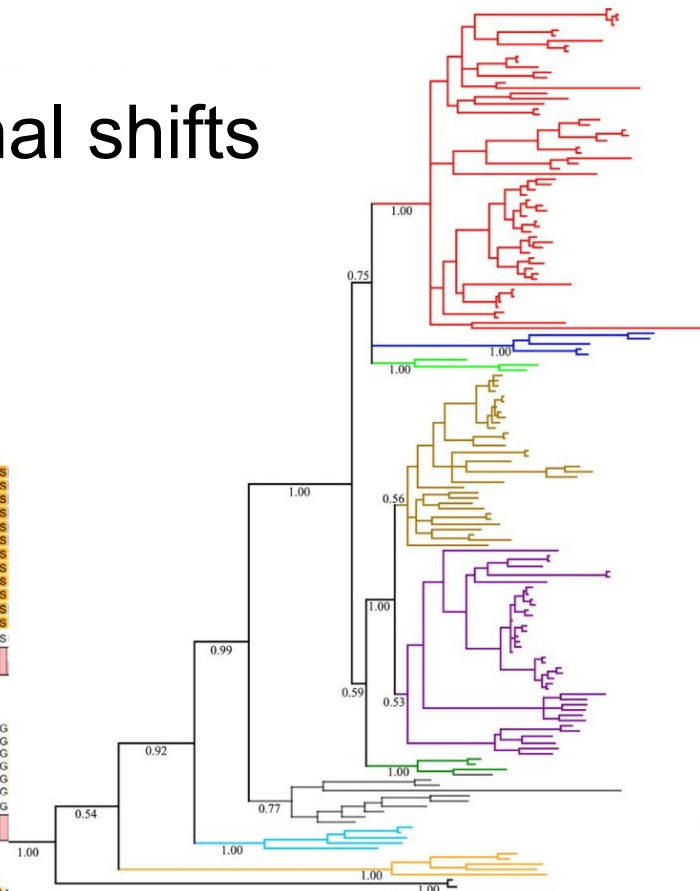
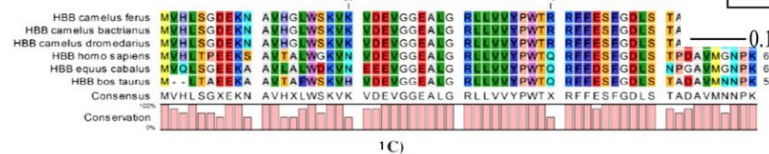
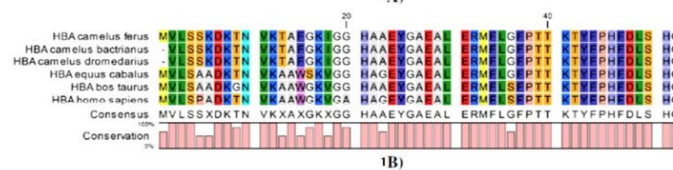
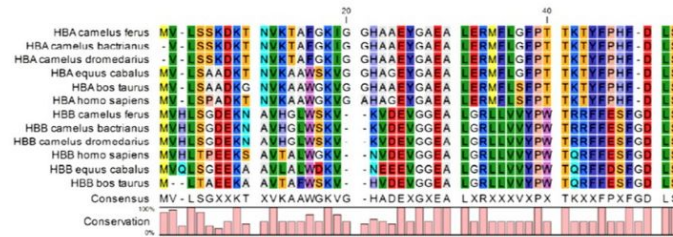
C



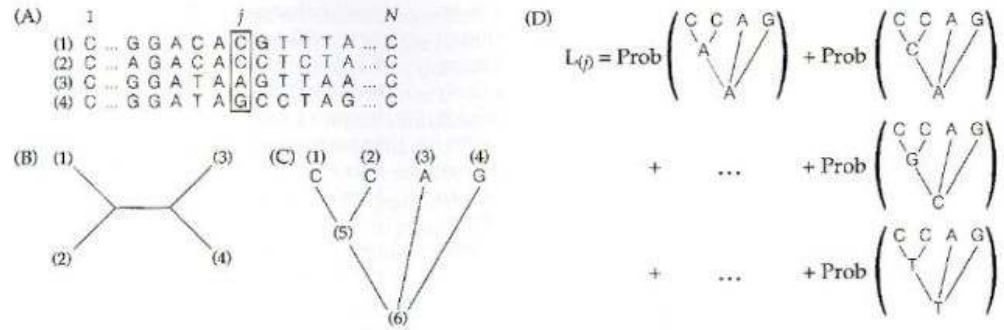
Conclusions

- Orthologs (slightly?) more likely than paralogs to share function
- One function per gene family?: not totally, variation over a common theme (e.g. transporter with different substrate affinities)
- Broadly defined functions probably conserved, specific functions more variable.

Gene trees can inform on functional shifts



Maximum likelihood methods provide not only a topology and branch length, but also a hypothesis of sequence evolution along the tree



- Tree after rooting in an arbitrary node (reversible model).
- The likelihood for a particular site is the sum of the probabilities of every possible reconstruction of ancestral states given some model of base substitution.
- The likelihood of the tree is the product of the likelihood at each site.

$$L = L_{(1)} \cdot L_{(2)} \cdot \dots \cdot L_{(N)} = \prod_{j=1}^N L_{(j)}$$

- The likelihood is reported as the sum of the log likelihood of the full tree.

$$\ln L = \ln L_{(1)} + \ln L_{(2)} + \dots + \ln L_{(N)} = \sum_{j=1}^N \ln L_{(j)}$$

Nonsynonymous and synonymous substitutions are expected to be subject to selection to different degrees

A Nonsynonymous / Synonymous substitution

<u>TCC</u>	GAT	<u>ATA</u>	TGG	<u>CAA</u>	CCC	<u>GAC</u>	AAA
S	D	I	W	Q	P	D	K

<u>TCA</u>	GAT	<u>CTA</u>	TGG	<u>CAG</u>	CCC	<u>CAC</u>	AAA
S	D	L	W	Q	P	R	K

B Radical / Conservative substitution

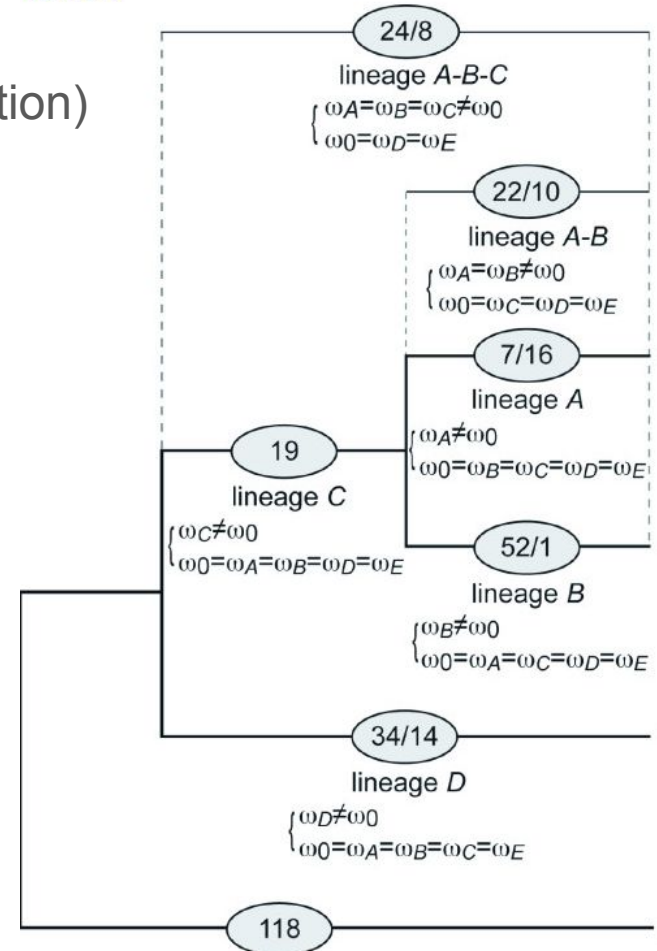
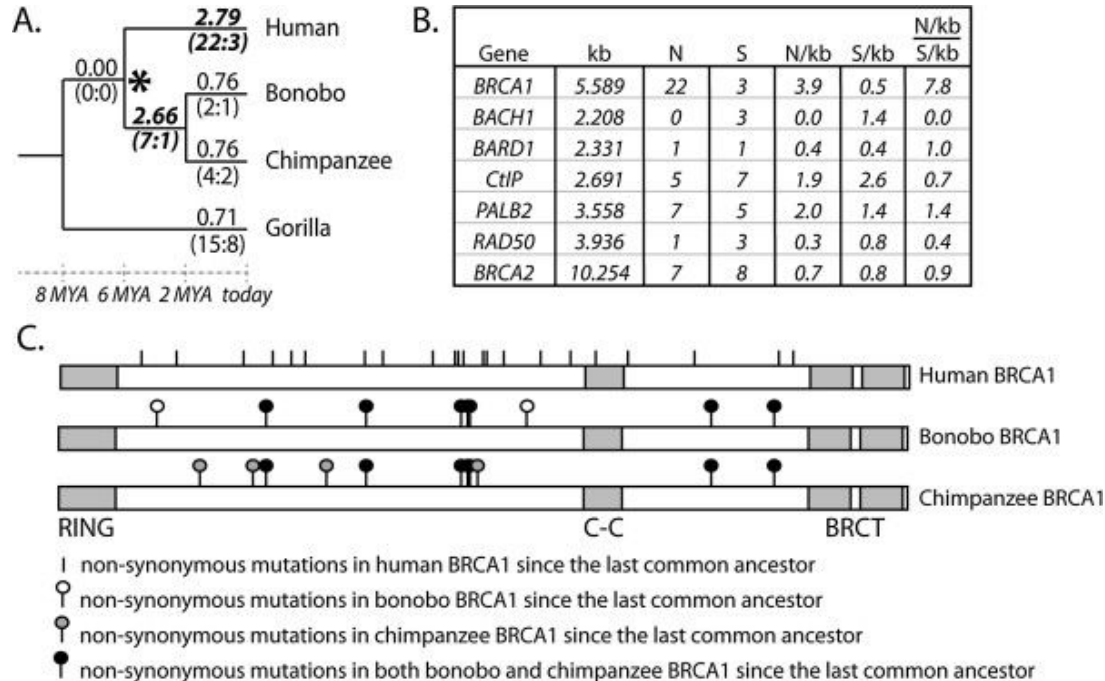
ATT	<u>GAC</u>	TATTCC	<u>TGT</u>	TGGTTT	<u>GAA</u>	CCAGGC	<u>AGA</u>			
I	<u>D⁻</u>	Y	S	<u>C^N</u>	W	F	<u>E⁻</u>	P	G	<u>R⁺</u>

ATT	<u>CAC</u>	TACTCC	<u>GGT</u>	TGGTTC	<u>GCA</u>	CCAGGA	<u>AAA</u>			
I	<u>R⁺</u>	Y	S	<u>G^N</u>	W	F	<u>A^N</u>	P	G	<u>K⁺</u>

+ positive
 - negative
 N neutral

Figure 2.

We can use branch-site models to compute rates for each branch (i.e. to detect lineage specific selection) (e.g. PAML)



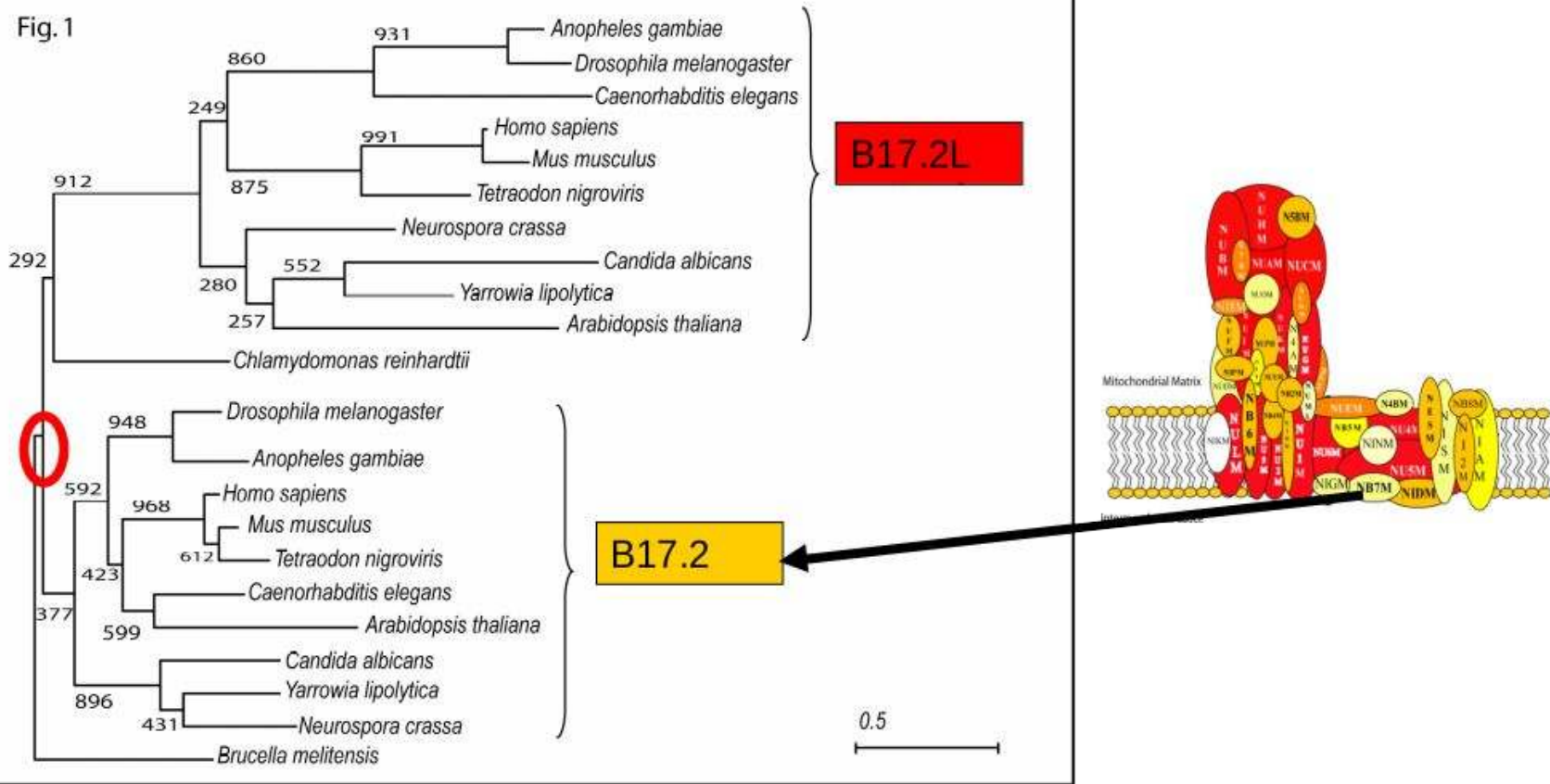
Can we predict change of function?

DIVERGE2= compare sub-alignments of different clades that differ radically in specific domains



Probably group 2a and group 2b, perform different functions

Fig. 1



Gabaldón and Huynen 2007

Prediction: B17.2L has a function that is linked to Complex I (co-evolution) but likely Very different from what B17.2 (never identified as a subunit, large sequence divergence different constraint) ⁵⁰

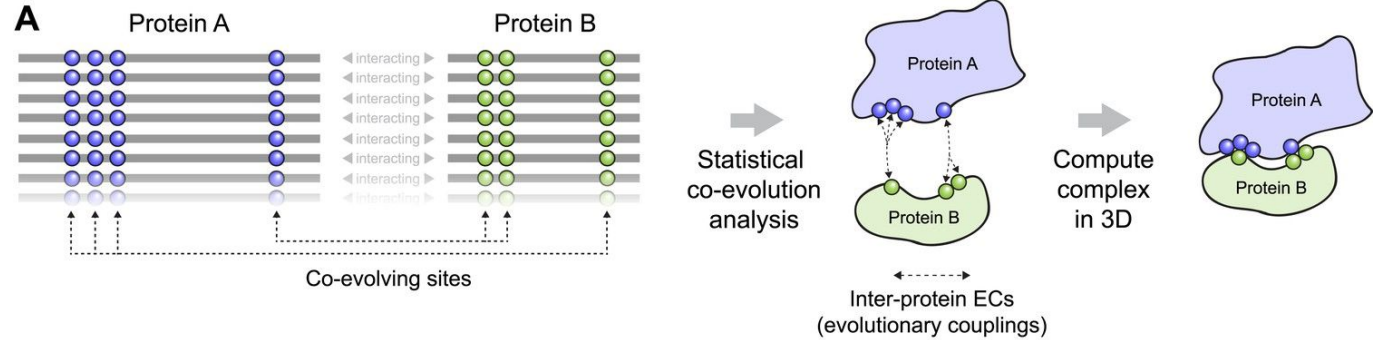


A molecular chaperone for mitochondrial complex I assembly is mutated in a progressive encephalopathy

Isla Ogilvie,¹ Nancy G. Kennaway,² and Eric A. Shoubridge^{1,3}

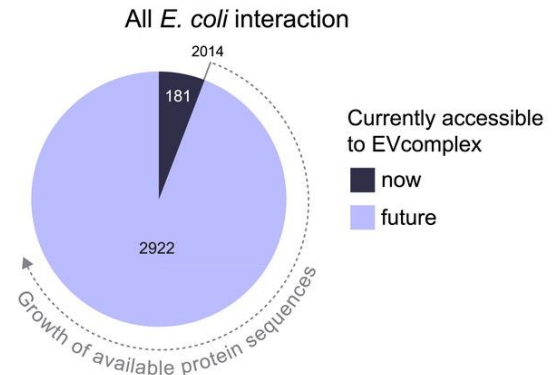
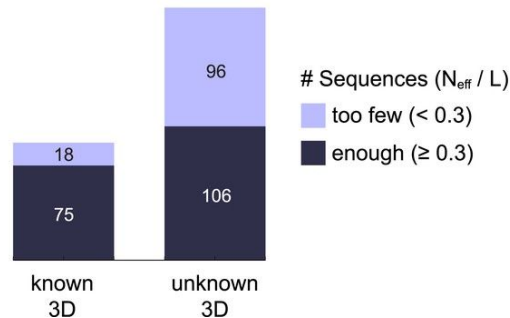
¹Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada. ²Department of Molecular and Medical Genetics, Oregon Health & Science University, Portland, Oregon, USA. ³Department of Human Genetics, McGill University, Montreal, Quebec, Canada.

You can also model co-evolution between sequences



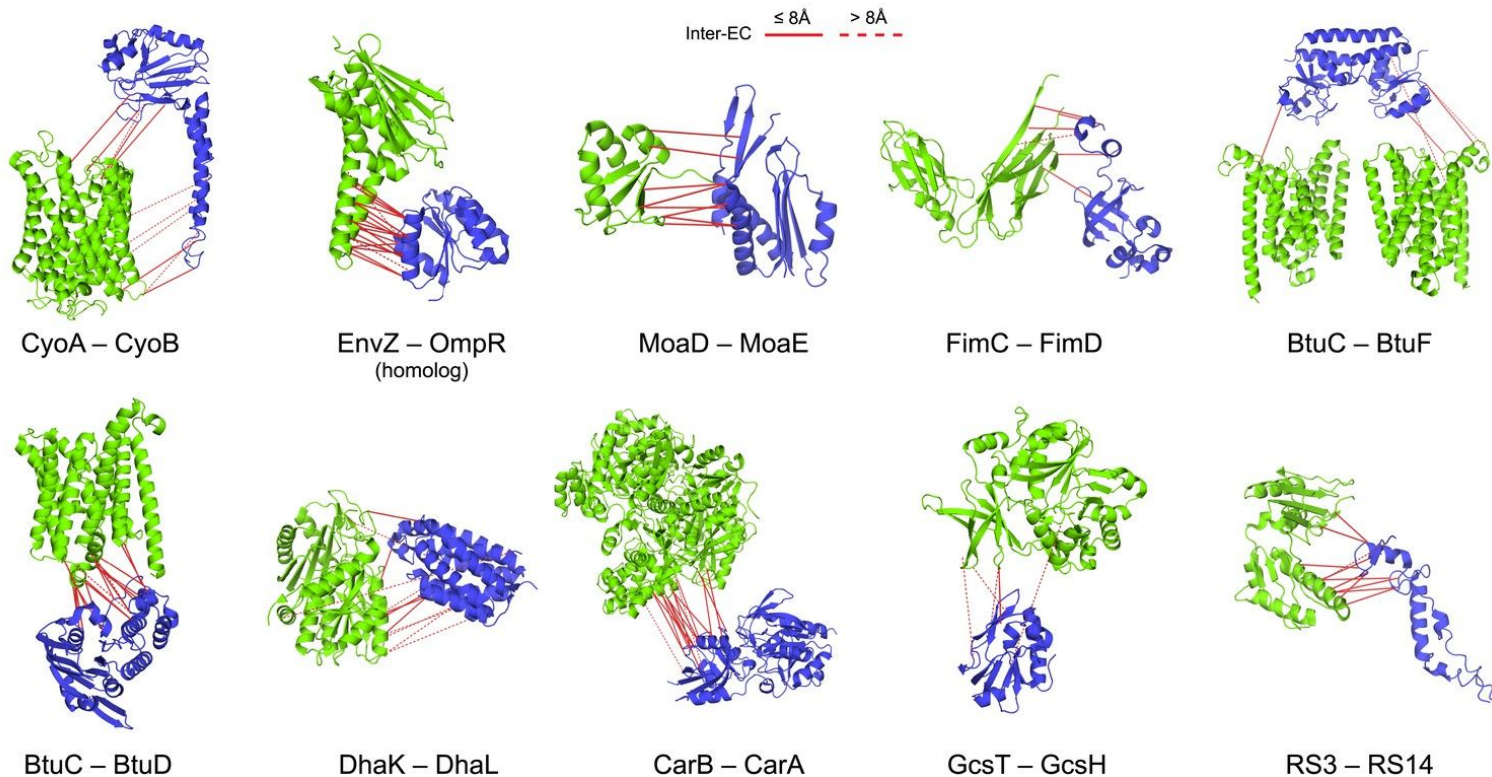
B

Complexes with subunits close on *E. coli* genome

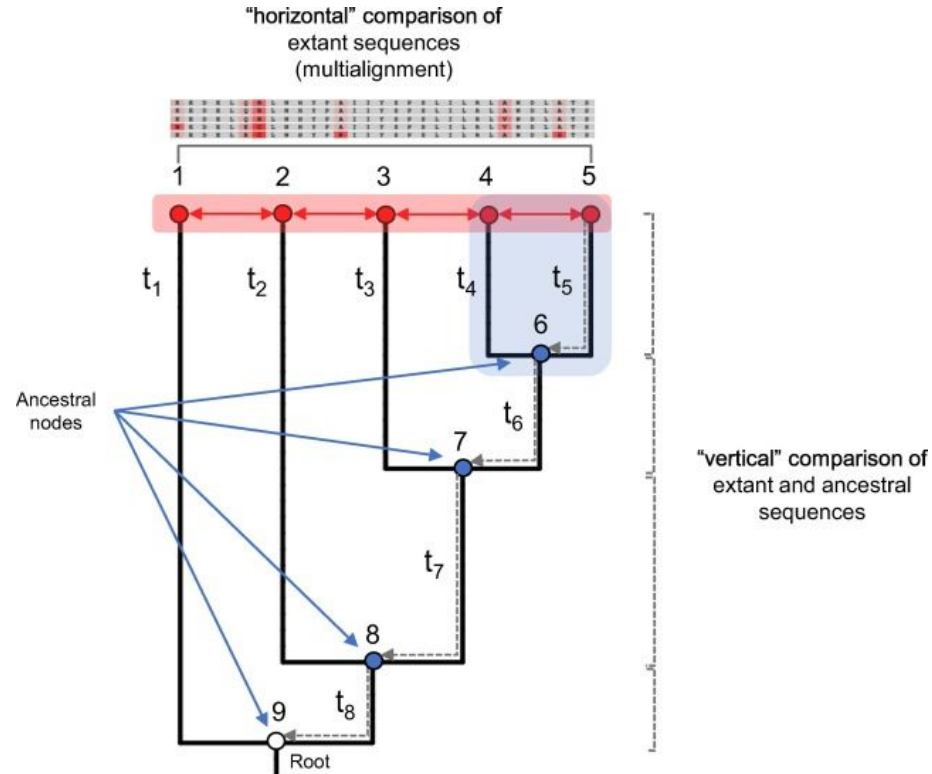


You can also model co-evolution between sequences

Blinded prediction of inter-protein contacts in complexes with known 3D structure



You can even reconstruct ancestral sequences



News

Triassic reptile saw red

Resurrected protein suggests that crocodiles' ancestors roamed at night.

Helen Pearson

A reptile from the Triassic period may have done its stalking at night. So suggest scientists who have resurrected a 240-million-year-old eye protein that sees dim light¹.

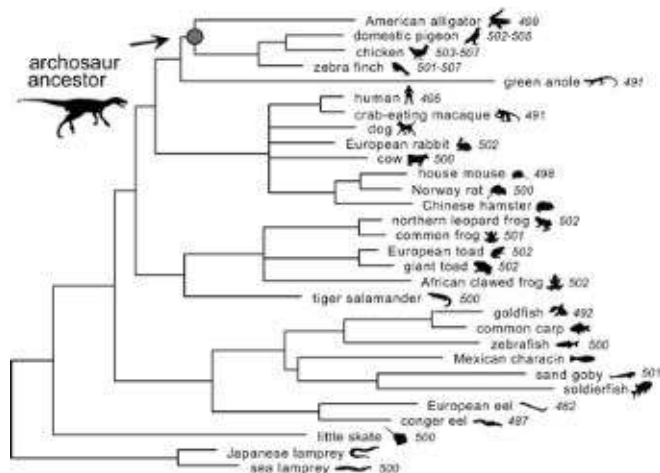
Such a molecule may have been found in the eyes of the earliest archosaurs, which were predecessors of the dinosaurs. Similar proteins, called rhodopsins, perceive low levels of light in humans and other animals.

Thomas Sakmar of Rockefeller University in New York and his colleagues used a computer program to extrapolate the DNA sequence of the ancient rhodopsin from known sequences in alligator, birds, frogs and fish.



Gene reconstruction gives researchers a dim view of the distant past.

© SPL

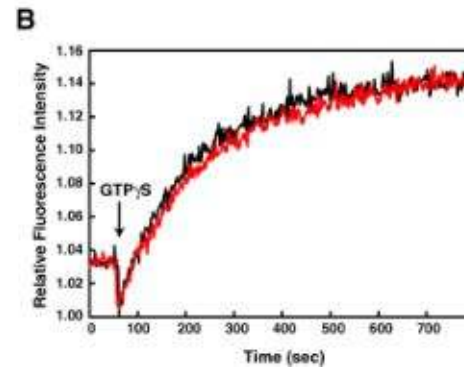
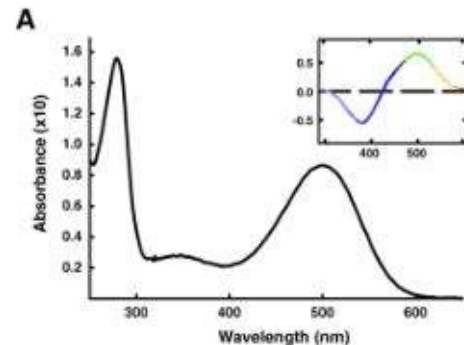


Recreating a Functional Ancestral Archosaur Visual Pigment

Belinda S. W. Chang, Karolina Jönsson, Manija A. Kazmi, Michael J. Donoghue, Thomas P. Sakmar

Molecular Biology and Evolution, Volume 19, Issue 9, 1 September 2002, Pages 1483–1489,
<https://doi.org/10.1093/oxfordjournals.molbev.a004211>

Published: 01 September 2002 **Article history** ▼



News

Triassic reptile saw red

Resurrected protein suggests that crocodiles' ancestors roamed at night.

Helen Pearson

A reptile from the Triassic period may have done its stalking at night. So suggest scientists who have resurrected a 240-million-year-old eye protein that sees dim light¹.



Such a molecule may have been found

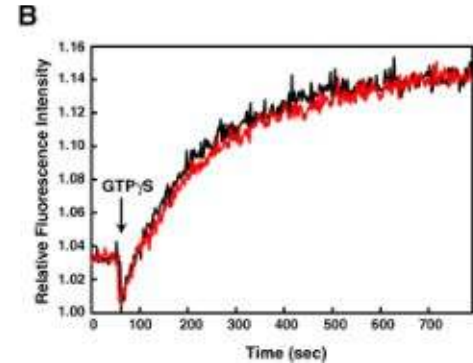
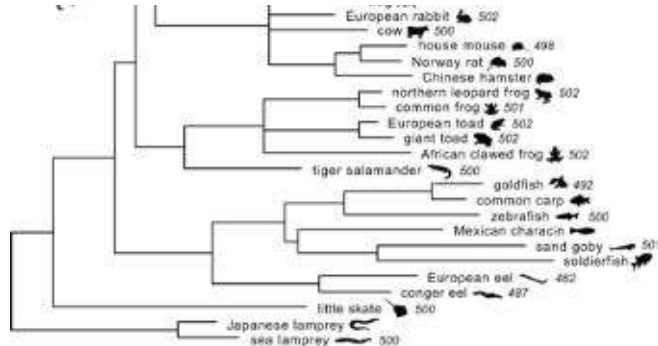
Recreating a Functional Ancestral Archosaur Visual Pigment

Belinda S. W. Chang, Karolina Jönsson, Manija A. Kazmi, Michael J. Donoghue, Thomas P. Sakmar

Molecular Biology and Evolution, Volume 19, Issue 9, 1 September 2002, Pages 1483–1489,
<https://doi.org/10.1093/oxfordjournals.molbev.a004211>

Published: 01 September 2002 Article history ▾

displayed similar functional characteristics. This indicates that archosaurs may have had a class of visual pigments that would support dim-light vision, which is consistent with the intriguing possibility that nocturnal, not diurnal, life histories may have been the ancestral state in amniotes (Gauthier 1994), though further studies will be needed to clarify this issue.



Conclusions

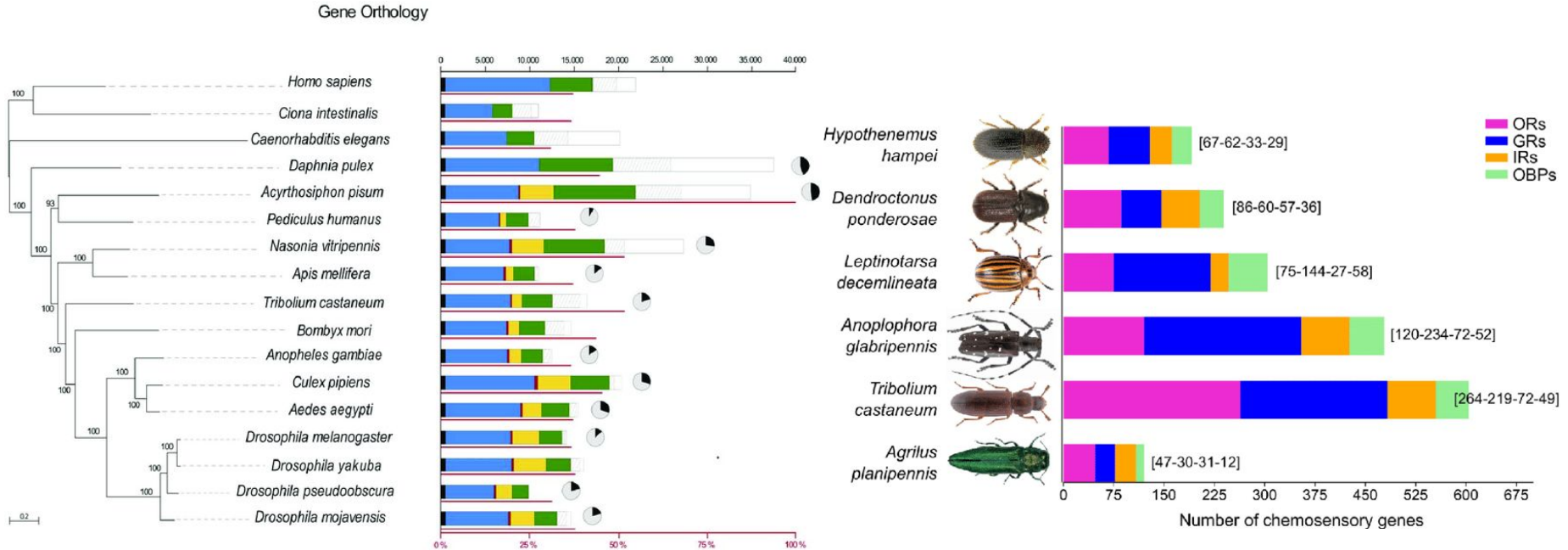
- Gene trees and their underlying alignments provide a plethora of information that can be exploited for different purposes.
- Most such analysis have been used in particular case-studies
- But large computing capacities, automated pipelines and more efficient algorithms enable to scale up such analyses .

Break?

How to study gene family evolution at genomic scales?

- 1) **Model gene family content across a species tree**
- 2) Reconstruct gene (family) phylogenies and compare them with the species tree

Variation of gene content across species



A gene family:

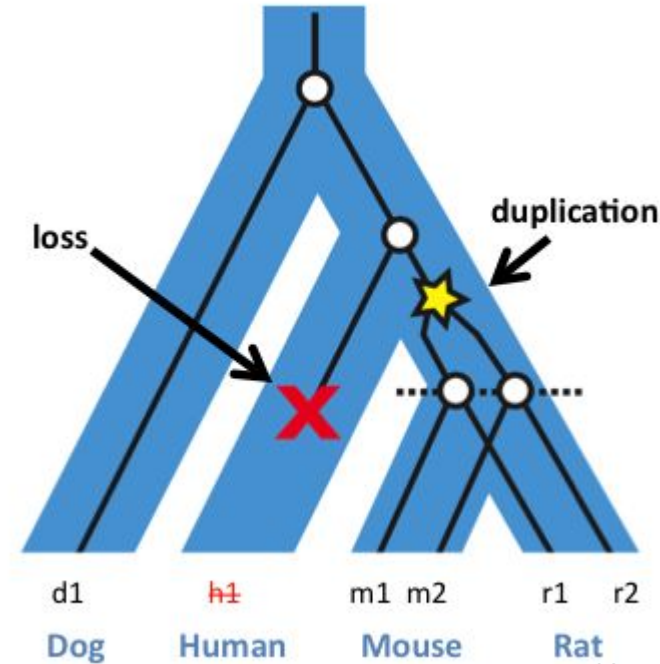
A set of genes with shared ancestry (homologs)

Gene families have hierarchical evolutionary relationship (**best represented by a tree**)

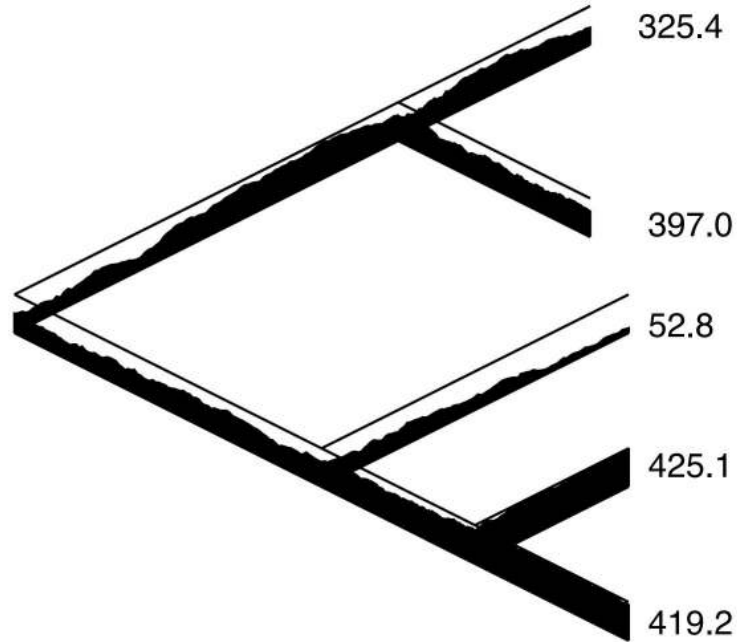
Members of a gene family can be orthologs or paralogs between them

An orthologous group is a (or part of) a gene family

Gene families evolve by duplication and loss (birth and death)

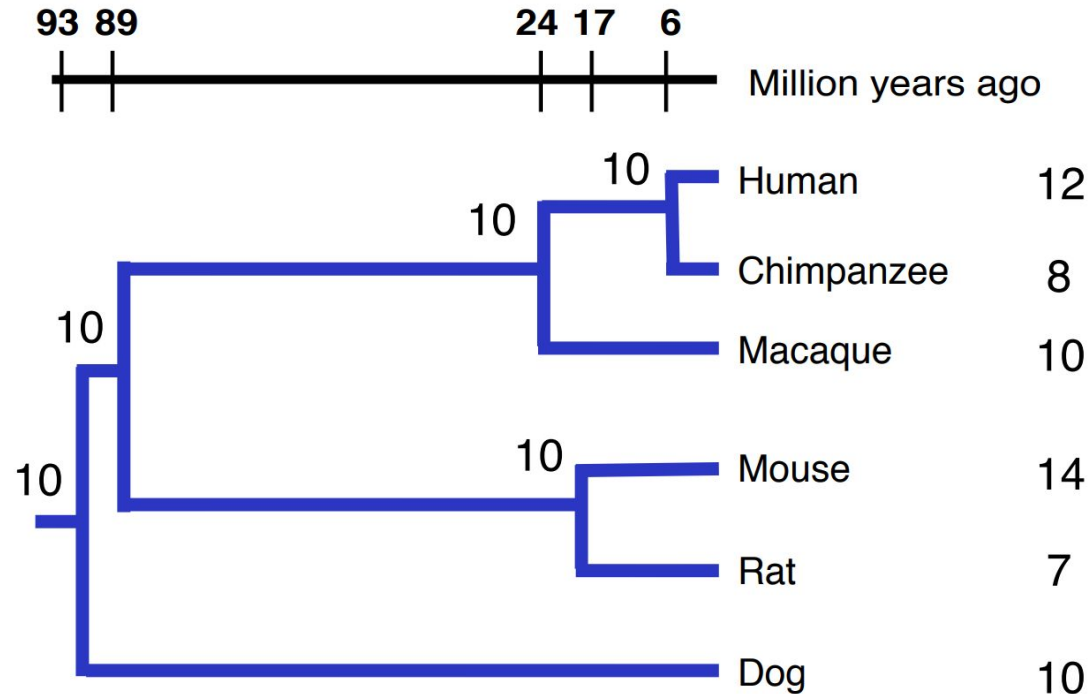


Models for gene family evolution: Model family gene numbers as quantitative traits



Felsenstein (2005)

Models for gene family evolution



$$\lambda=0.002$$

(assuming birth=death)

Models for gene family evolution

JOURNAL ARTICLE

CAFE 5 models variation in evolutionary rates among gene families

Fábio K Mendes, Dan Vanderpool , Ben Fulton, Matthew W Hahn [Author Notes](#)

Bioinformatics, Volume 36, Issue 22-23, December 2020, Pages 5516–5518,

<https://doi.org/10.1093/bioinformatics/btaa1022>

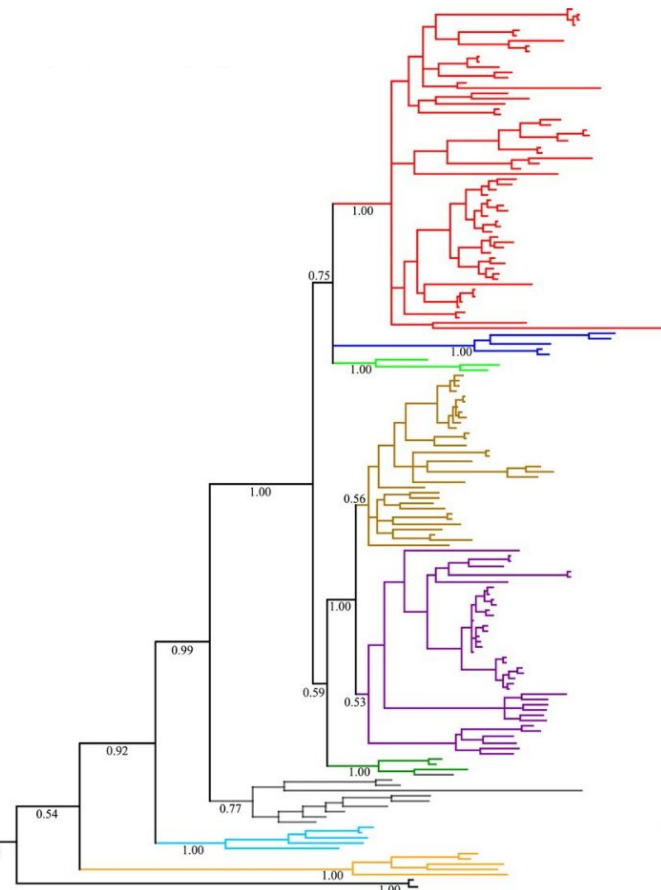
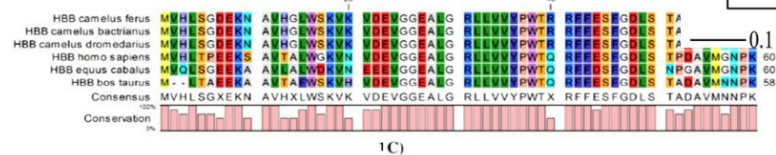
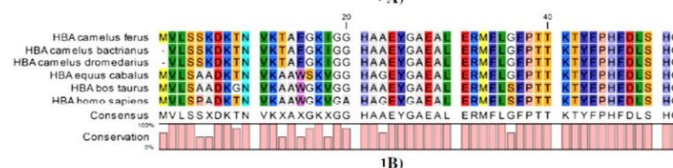
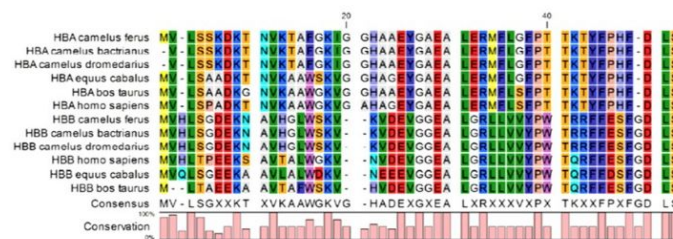
Published: 16 December 2020 **Article history** ▼

- Allows different rates in different branches and across families
- Models gene annotation errors

How to study gene family evolution?

- 1) Model gene family content across a species tree
- 2) Reconstruct gene (family) phylogenies and compare them with the species tree**

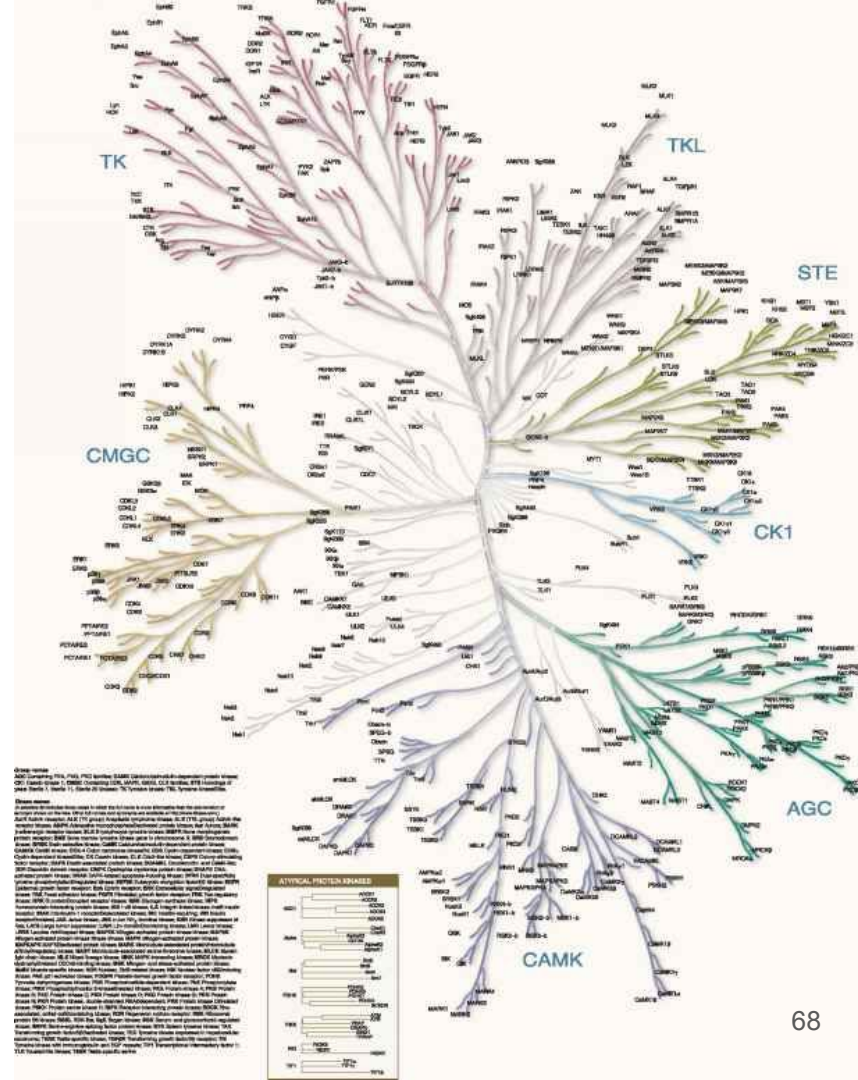
Gene trees

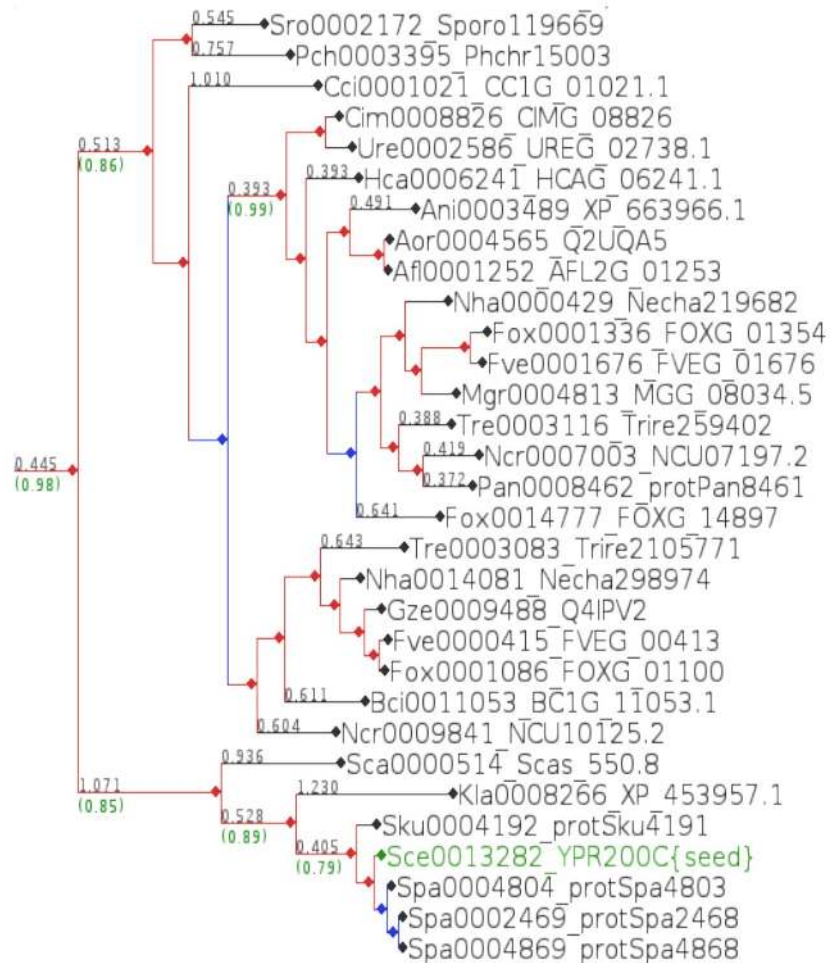


The Protein Kinase Complement of the Human Genome

G. Manning,^{1*} D. B. Whyte,¹ R. Martinez,¹ T. Hunter,²
S. Sudarsanam^{1,3}

We have catalogued the protein kinase complement of the human genome (the "kinome") using public and proprietary genomic, complementary DNA, and expressed sequence tag (EST) sequences. This provides a starting point for comprehensive analysis of protein phosphorylation in normal and disease states, as well as a detailed view of the current state of human genome analysis through a focus on one large gene family. We identify 518 putative protein kinase genes, of which 71 have not previously been reported or described as kinases, and we extend or correct the protein sequences of 56 more kinases. New genes include members of well-studied families as well as previously unidentified families, some of which are conserved in model organisms. Classification and comparison with model organism kinomes identified orthologous groups and highlighted expansions specific to human and other lineages. We also identified 106 protein kinase pseudogenes. Chromosomal mapping revealed several small clusters of kinase genes and revealed that 244 kinases map to disease loci or cancer amplicons.





Tree collections can be interrogated to:

- Find families that show a particular topology
- Detect and date duplication events
- Genes that have accelerated evolutionary rates at a particular lineage (positive/relaxed selection)
- Detect families expanded at particular lineages
- Detect footprints of horizontal gene transfer, lineage sorting, gene conversion and other evolutionary processes
- Search for co-evolving genes
- Predict functional properties
- Across-species prediction of orthology and paralogy

Approaches

Interrogate gene trees independent of species tree

Compare gene trees and species tree: reconciliation, species-overlap

Co-estimate gene trees and species trees: GeneRax, ALE

Approaches

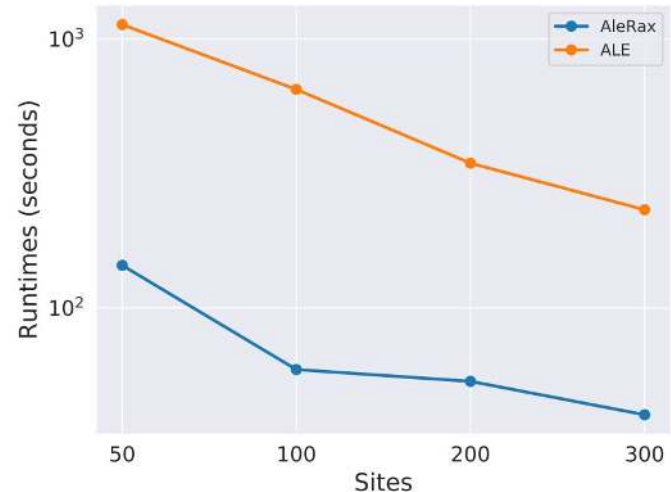
Interrogate gene trees independent of species tree

Compare gene trees and species tree: reconciliation, species-overlap

Co-estimate gene trees and species trees: GeneRax, ALE

**AleRax: A tool for gene and species tree co-estimation and reconciliation
a probabilistic model of gene duplication, transfer, and loss**

 Benoit Morel, Tom A. Williams, Alexandros Stamatakis, Gergely J. Szöllősi
doi: <https://doi.org/10.1101/2023.10.06.561091>



Approaches

A) Family centric approach (Most used)

Build gene families by a blast-based clustering approach (e.g. Orthofinder)

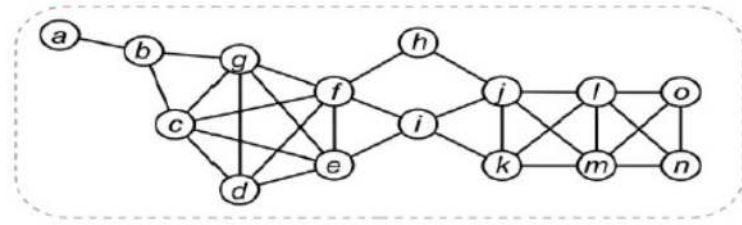
Then make a gene tree per family

B) Gene centric approach (PhylomeDB)

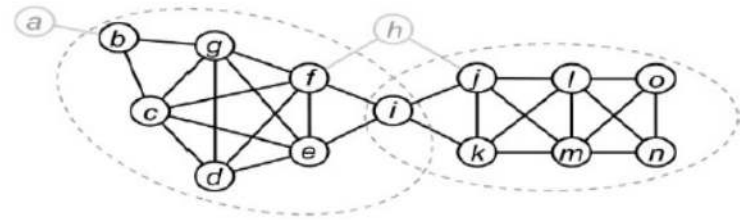
Take a seed genome, for every gene find homologs with blast, reconstruct a gene tree per gene (multiple gene trees per family are possible)

Finding optimal granularity
might be tricky

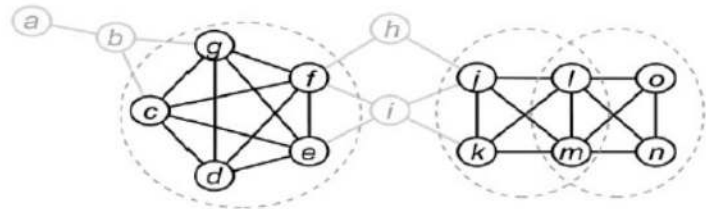
2



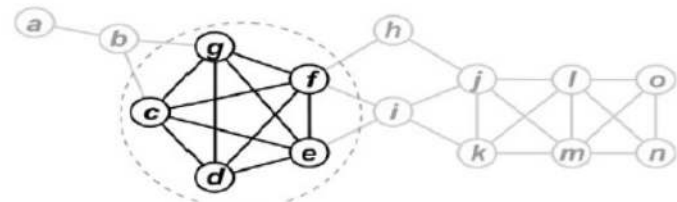
3

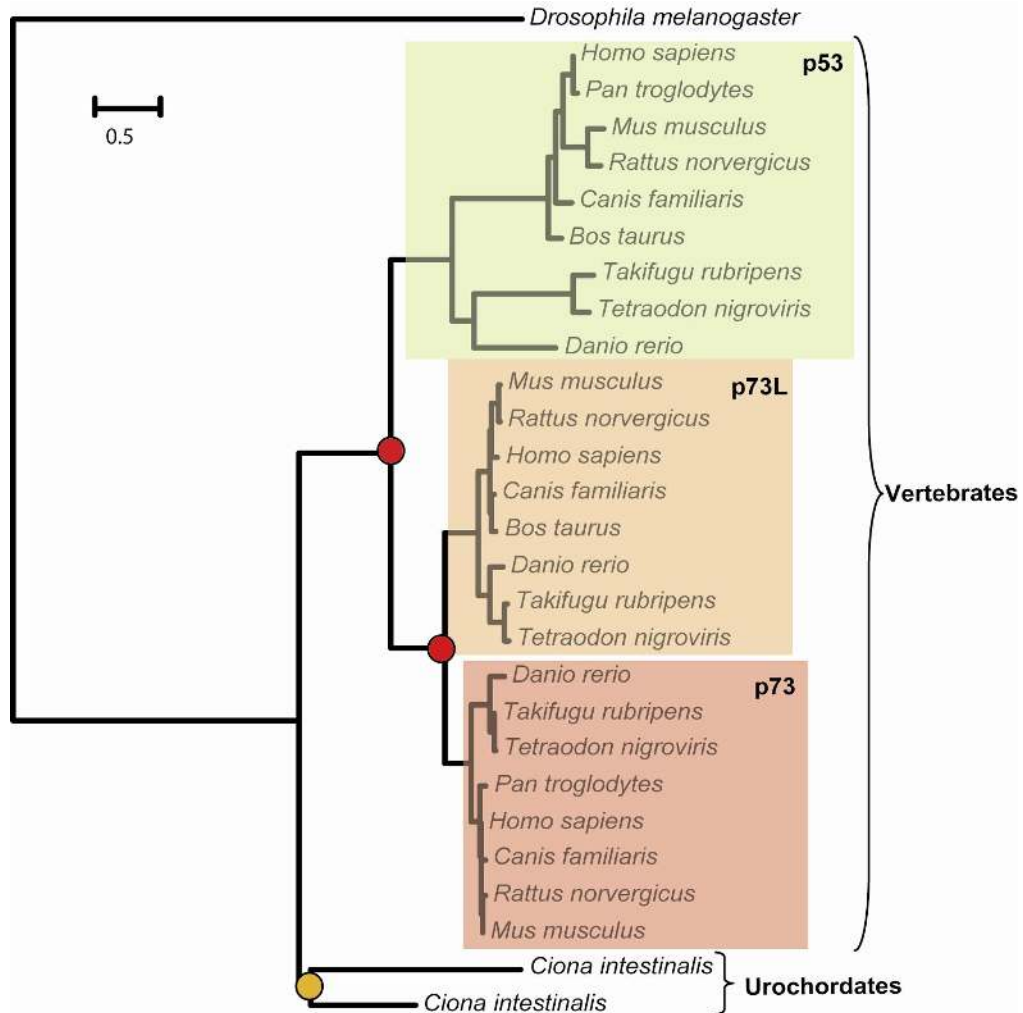


4



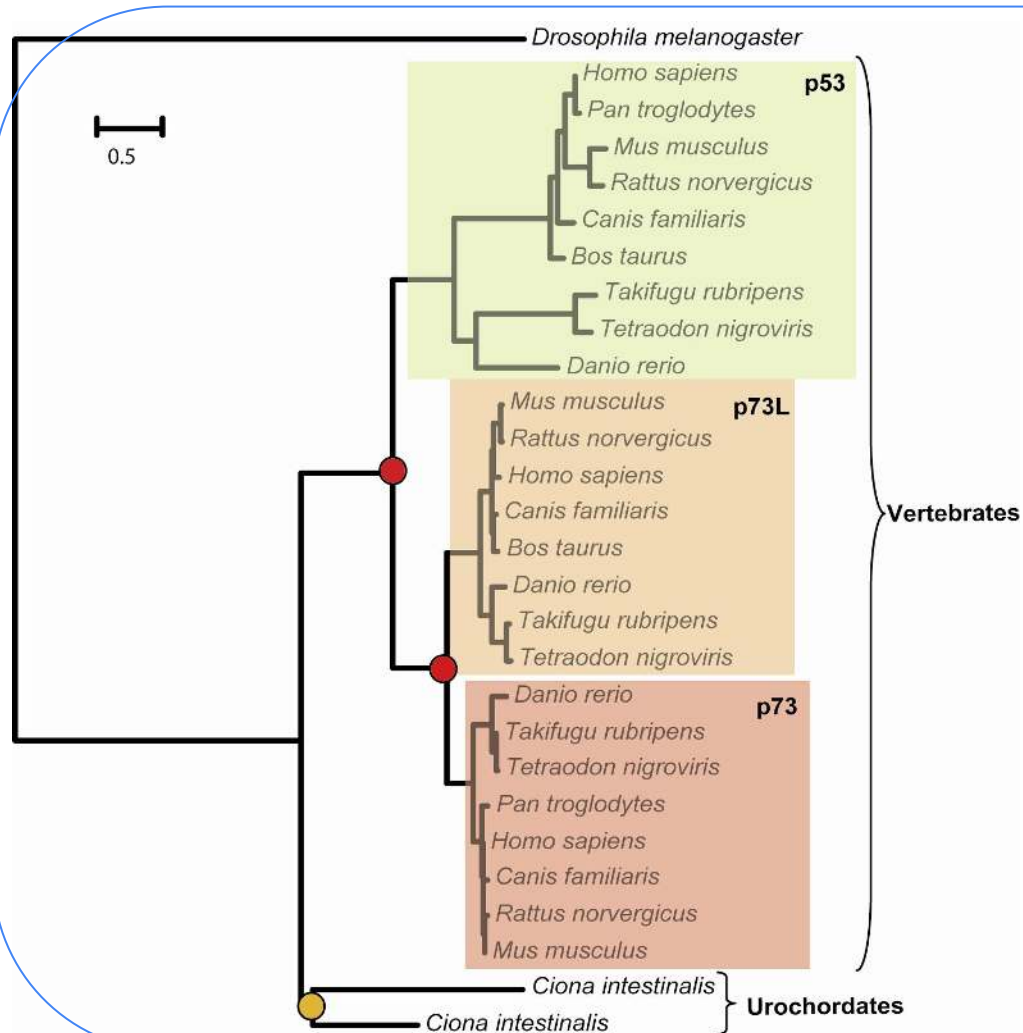
5



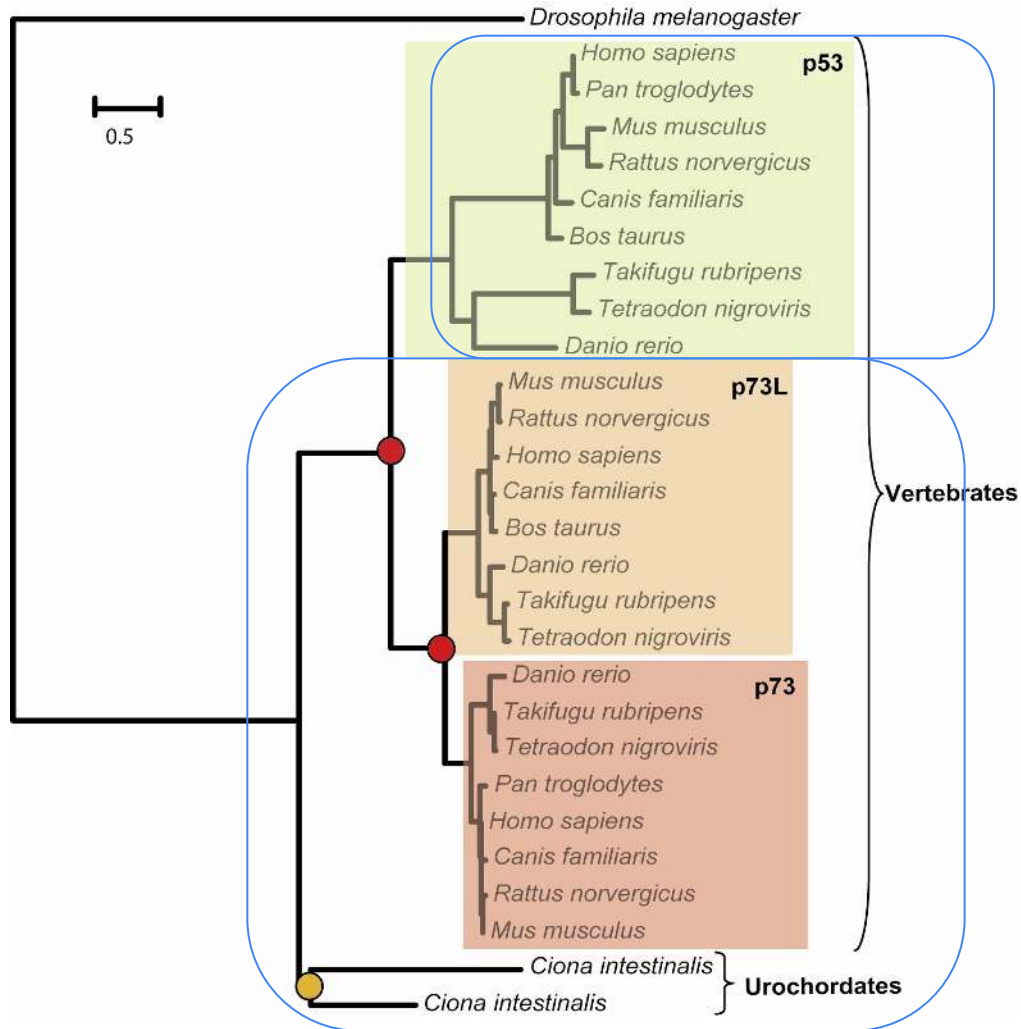


How many families?

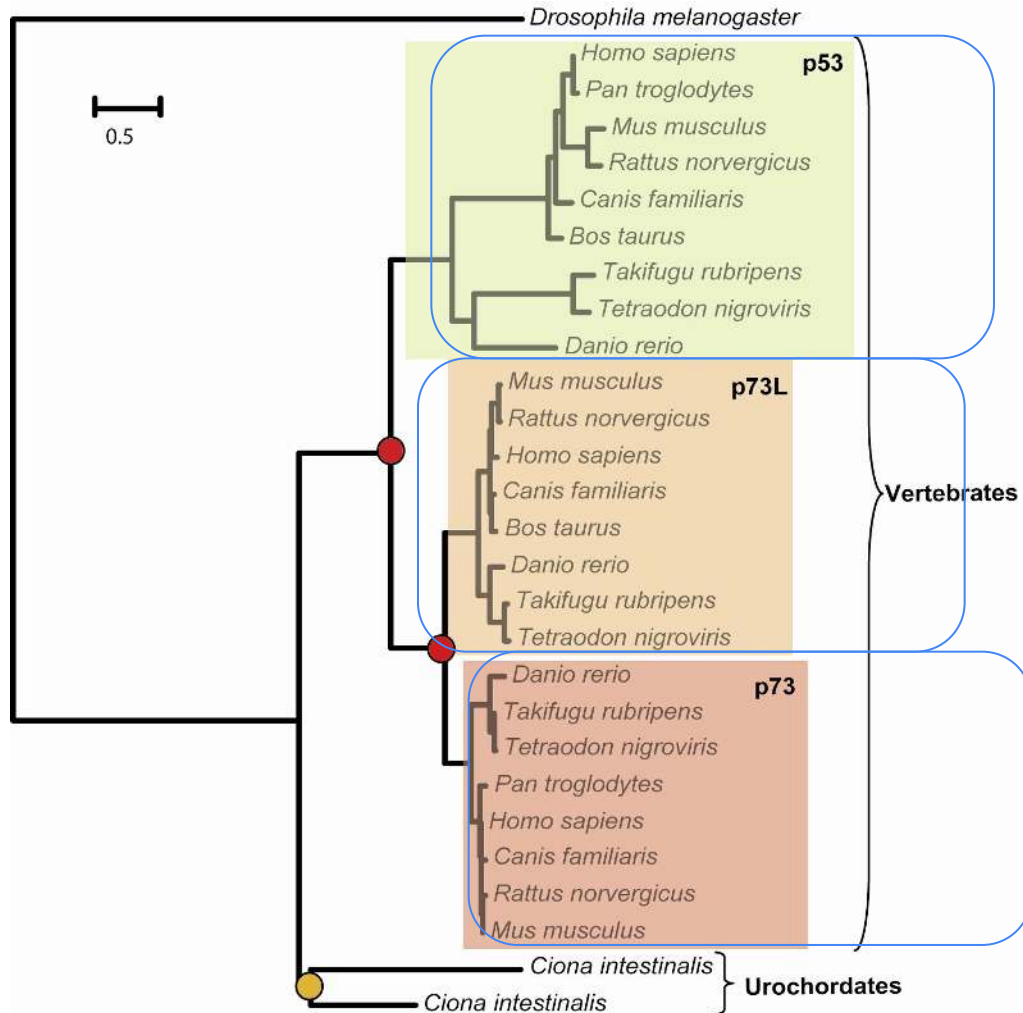
How many families?

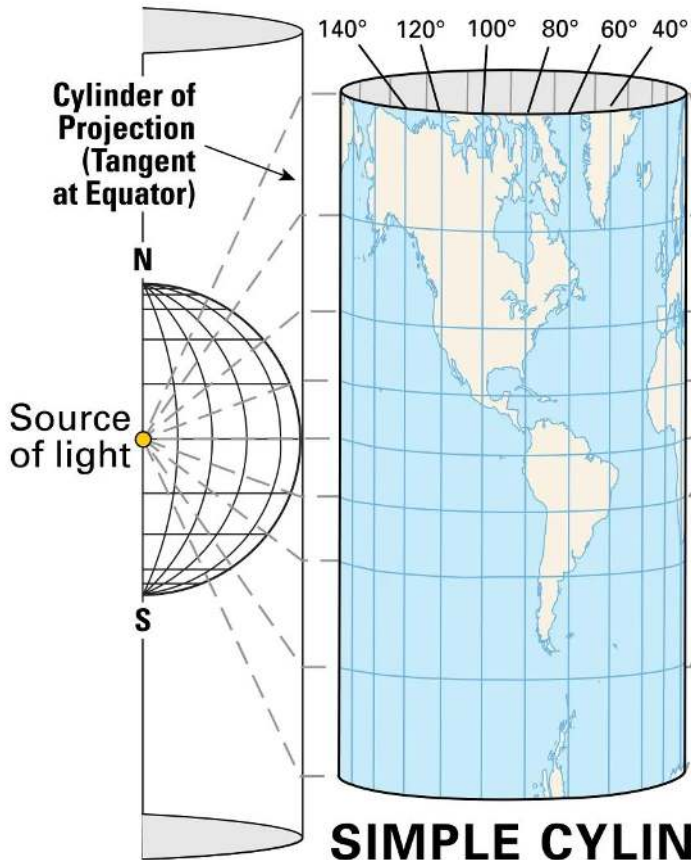


How many families?



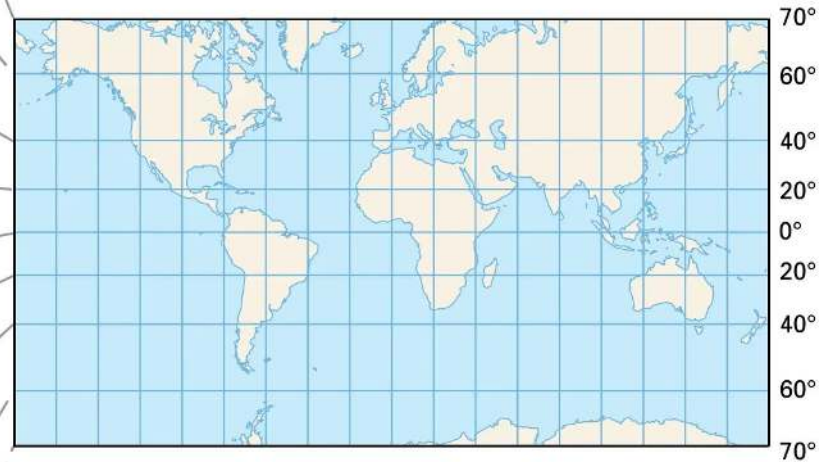
How many families?





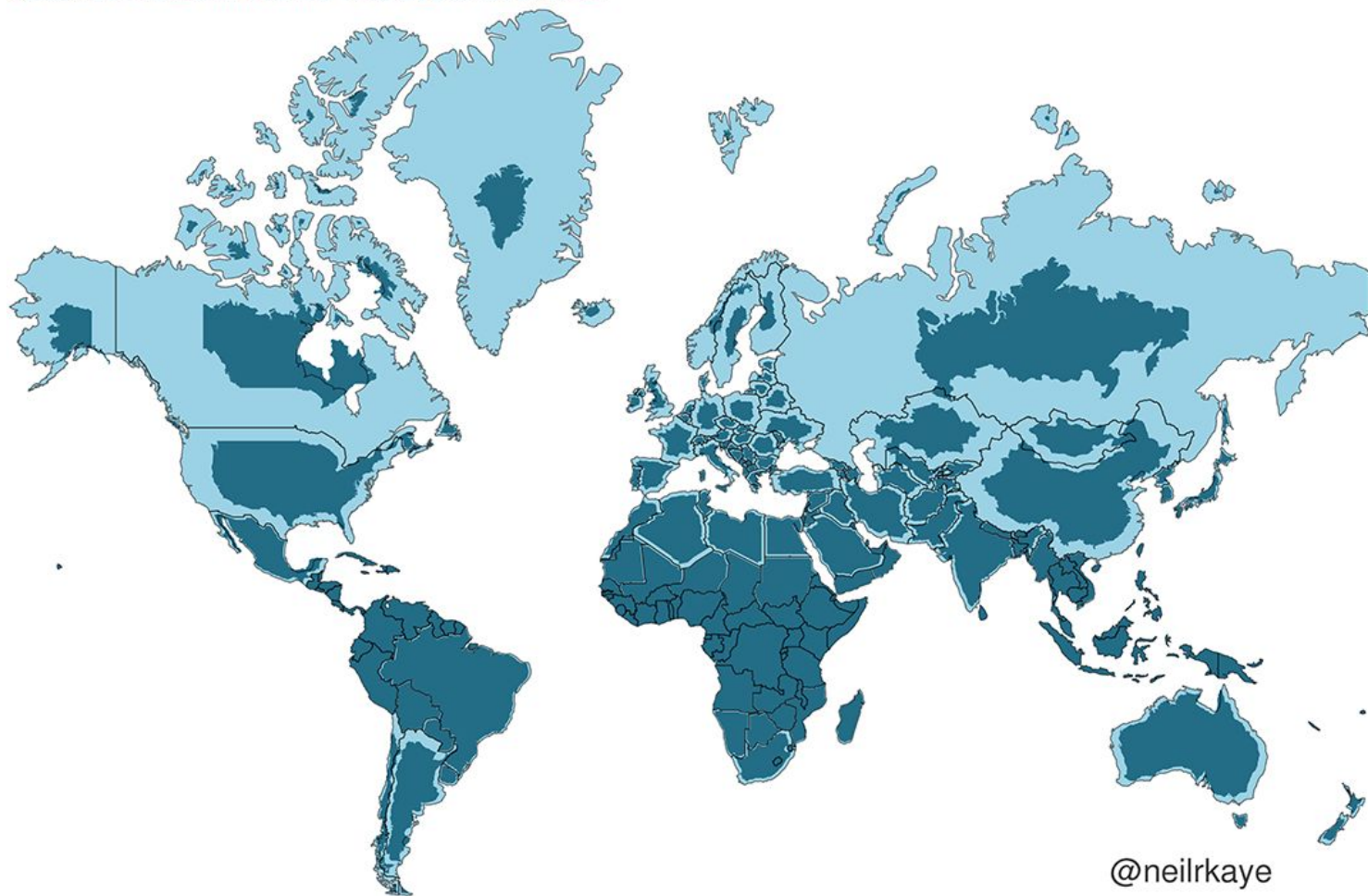
**SIMPLE CYLINDRICAL
PROJECTION**

MERCATOR PROJECTION



© Encyclopædia Britannica, Inc.

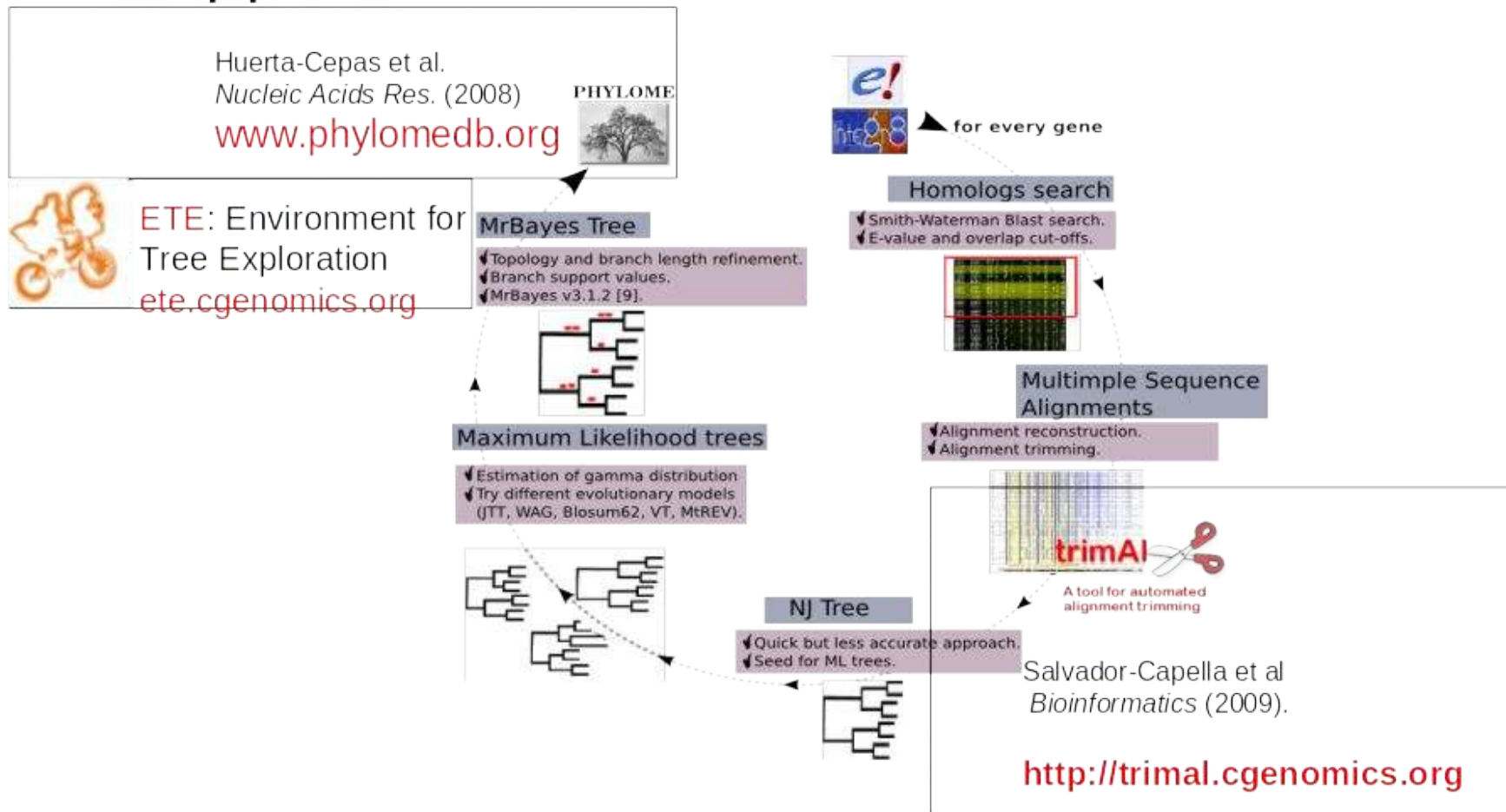
MERCATOR PROJECTION VS THE TRUE SIZE OF COUNTRIES



Orthogroups are useful but

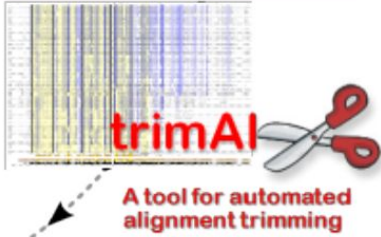
- Bad name choice (= gene family, and it contains paralogs)
- A 1-dimensional projection of a hierarchical relationship based on indirect measure of that hierarchy (blast-based distance)
- Nested relationships, must be defined at each taxonomic level
- ...

Our pipeline:

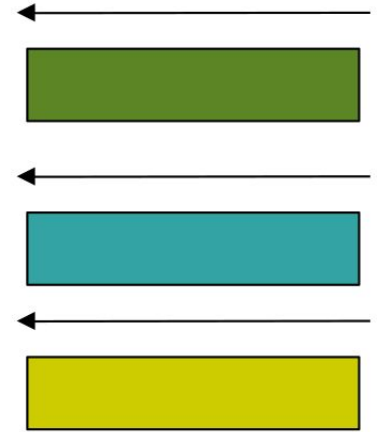
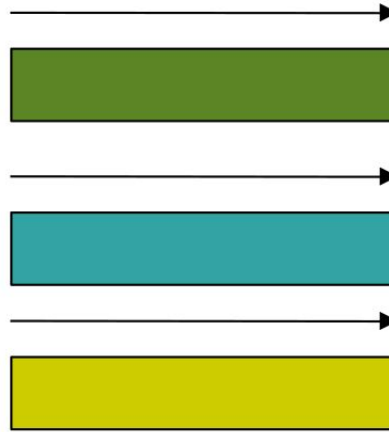


Multiple Sequence Alignments

- Alignment reconstruction.
- Alignment trimming.



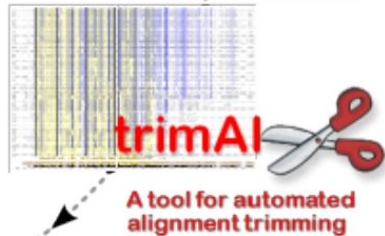
te approach.



Homologous sequences aligned in forward and reversed (head or tail approach), and each of them with three different algorithms: $2 \times 3 = 6$ different alignments

Multiple Sequence Alignments

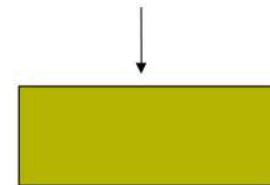
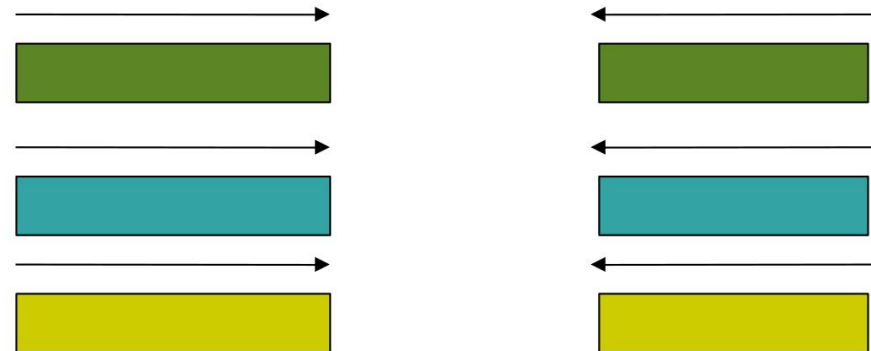
- Alignment reconstruction.
- Alignment trimming.



te approach.

A consensus is built from the 6 different alignments (M-Cofee)

TrimAl trims based on a consistency score



```
sw_DSBA_PSESM/1
sw_DSBA_SALTY/1
sw_DSBA_ENTAM/3
sw_DSBA_LEGPN/1
```

```
---MRNLIISAALVAASLFGMSAQAAEPIESGRQYV-ELTSAVPV
---MKKIWLA---LAGMVLAFSASAAQISDGRQYI-TLDKP--V
AKWINSIFKSVVLTAAALALPFTAS--AFTE-GTDYM-VLEKP---
-----LMPMTALATQFIEGRKYQTVASAQ-LS
```

cons



```
sw_DSBA_PSESM/1
sw_DSBA_SALTY/1
sw_DSBA_ENTAM/3
sw_DSBA_LEGPN/1
```

```
AVPGK-IEVIELFWYGCPCYAFEPET---NPWVEKLPSDVNFVR
--AGE-PQVLEFFSFYCPHCYQFEEVLHVSNDVKKKLPEGTQKTR
-IPDAADKTLIKVFSYACPFCKYKRAVT--GPVADKVVADLVTFVP
TNKDKTPLITEFFSYGCPWCYKIDAPLN--D-WATRMGRGAHLER
```

cons

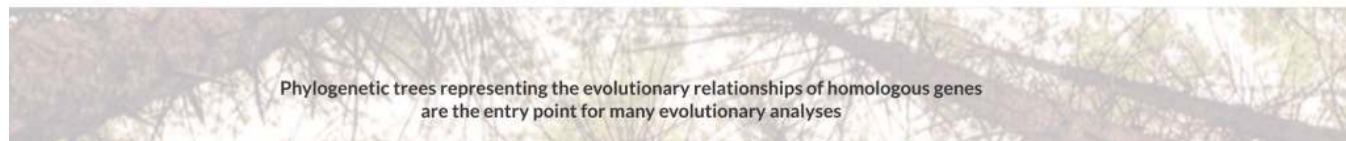




WHAT IS PHYLOMEDB?

PhylomeDB is a public database for complete **catalogs of gene phylogenies** (phylogenies). It allows users to interactively explore the evolutionary history of genes through the visualization of phylogenetic trees and multiple sequence alignments. Moreover, phylomeDB provides genome-wide orthology and paralogy predictions which are based on the analysis of the phylogenetic trees. The automated pipeline used to reconstruct trees aims at providing a high-quality phylogenetic analysis of different genomes, including Maximum Likelihood tree inference, alignment trimming and evolutionary model testing.

PhylomeDB includes also a public download section with the complete set of trees, alignments and orthology predictions. Finally, phylomeDB provides an advanced tree visualization interface based on the ETE toolkit, which integrates tree topologies, taxonomic information, domain mapping and alignment visualization in a single and interactive tree image.



Search in PhylomeDB

(i.e. ENSG00000139618, YBL058W,
TP53)

[RandomTree!](#)
[BLAST search](#)

Latest Phylomes

Clogmia albipunctata	2013
Penicillium digitatum	2012
Schistosoma mansoni	2012
Cucumis melo	2012

[see all phylomes](#)

PhylomeDB uses

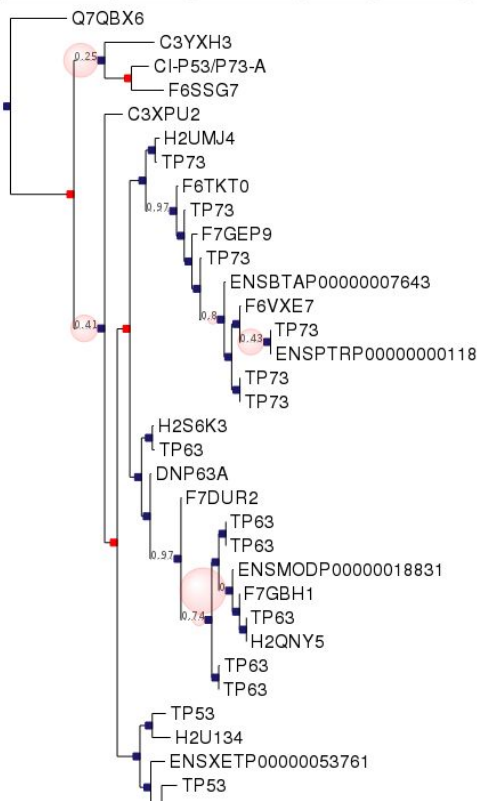


TP53 tree in phylome 218

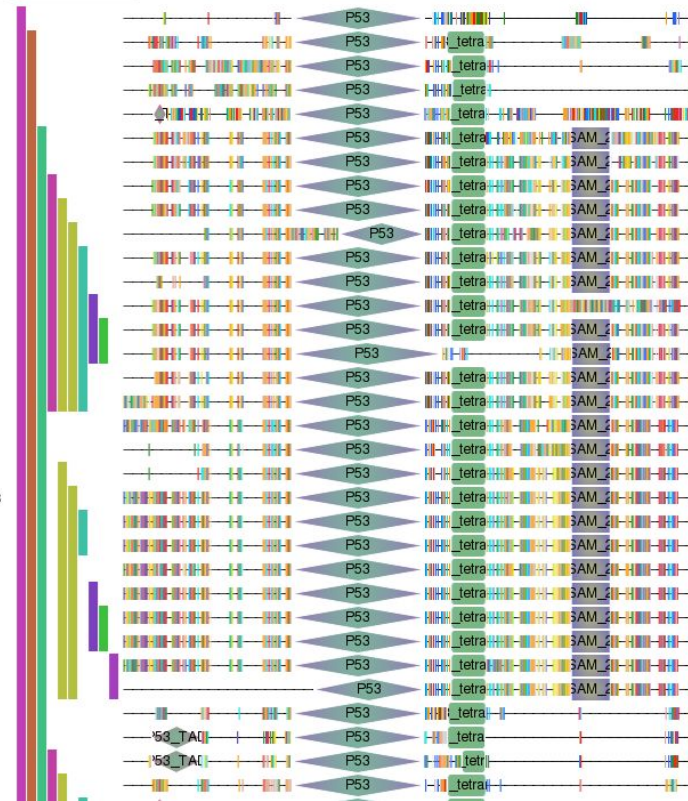
AS seed in Rat phylome

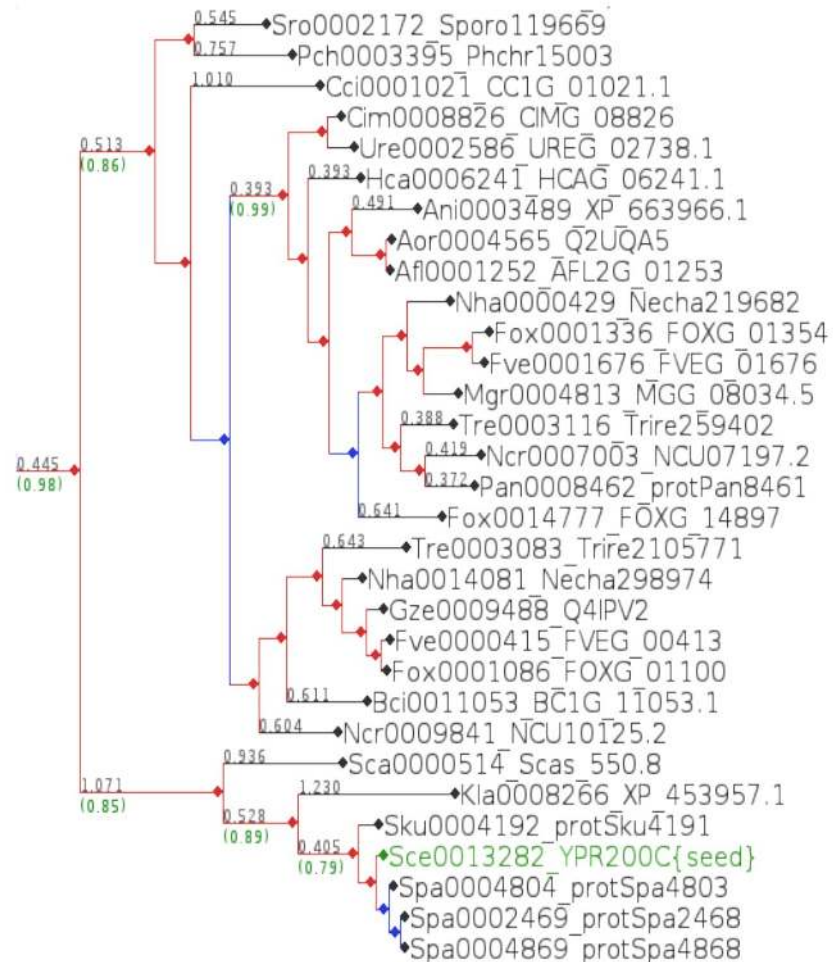
JTT (k:-18130.4)

-- in collateral trees --



Anopheles gambiae
Branchiostoma floridae
Ciona intestinalis
Ciona intestinalis
Branchiostoma floridae
Takifugu rubripes
Danio rerio
Xenopus tropicalis
Gallus gallus
Monodelphis domestica
Canis familiaris
Bos taurus
Macaca mulatta
Homo sapiens
Pan troglodytes
Mus musculus
Rattus norvegicus
Takifugu rubripes
Danio rerio
Gallus gallus
Ornithorhynchus anatinus
Rattus norvegicus
Mus musculus
Monodelphis domestica
Macaca mulatta
Homo sapiens
Pan troglodytes
Canis familiaris
Bos taurus
Danio rerio
Takifugu rubripes
Xenopus tropicalis
Gallus gallus





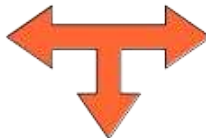
Tree collections can be interrogated to:

- Find families that show a particular topology
- Detect and date duplication events
- Genes that have accelerated evolutionary rates at a particular lineage (positive/relaxed selection)
- Detect families expanded at particular lineages
- Detect footprints of horizontal gene transfer, lineage sorting, gene conversion and other evolutionary processes
- Search for co-evolving genes
- Predict functional properties
- Across-species prediction of orthology and paralogy

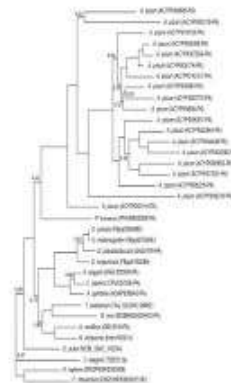
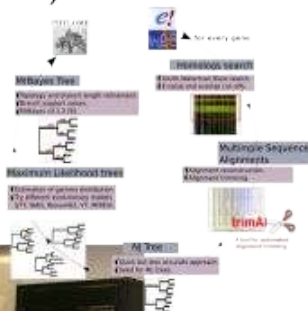




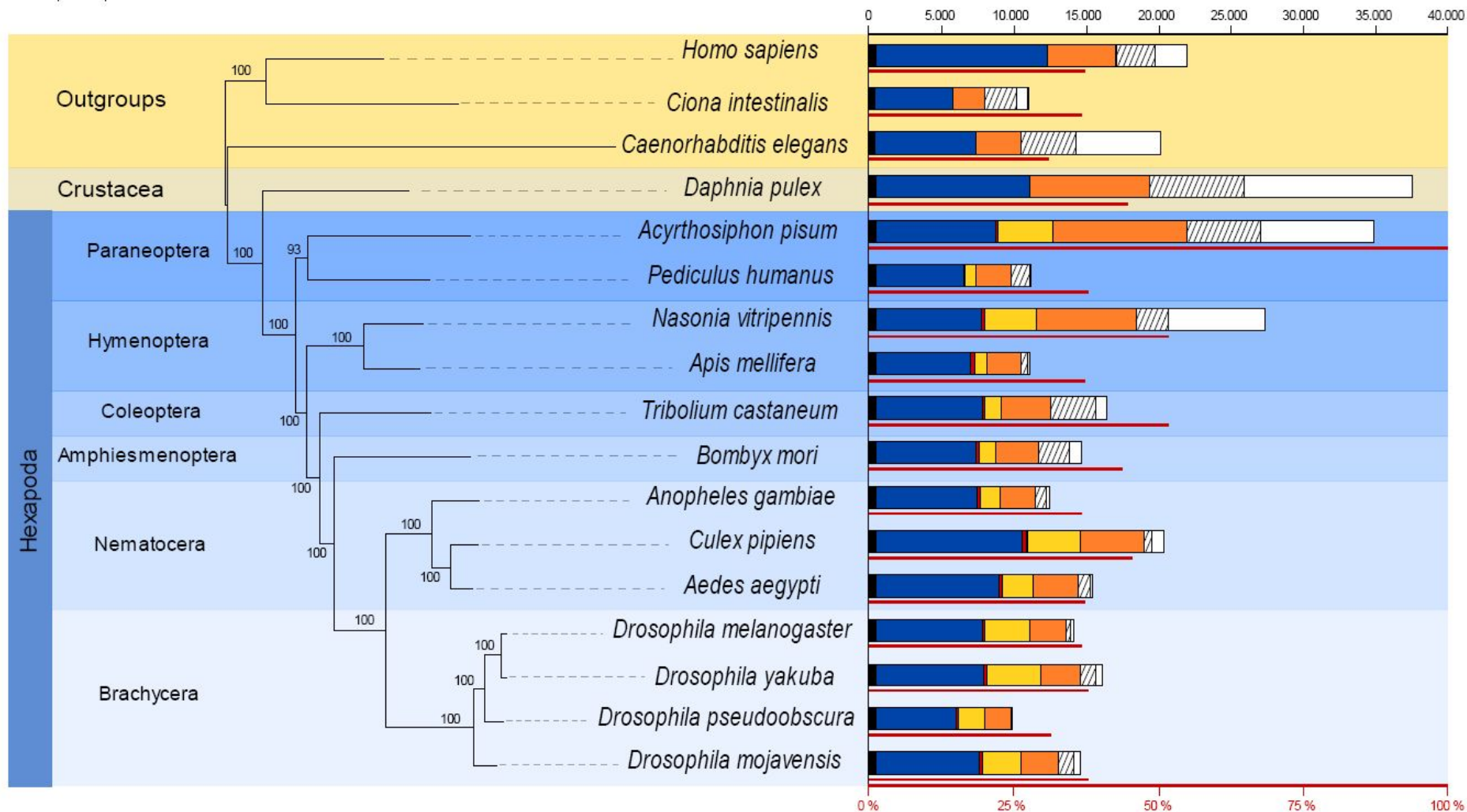
Acyr_1.0 (34,600 genes)

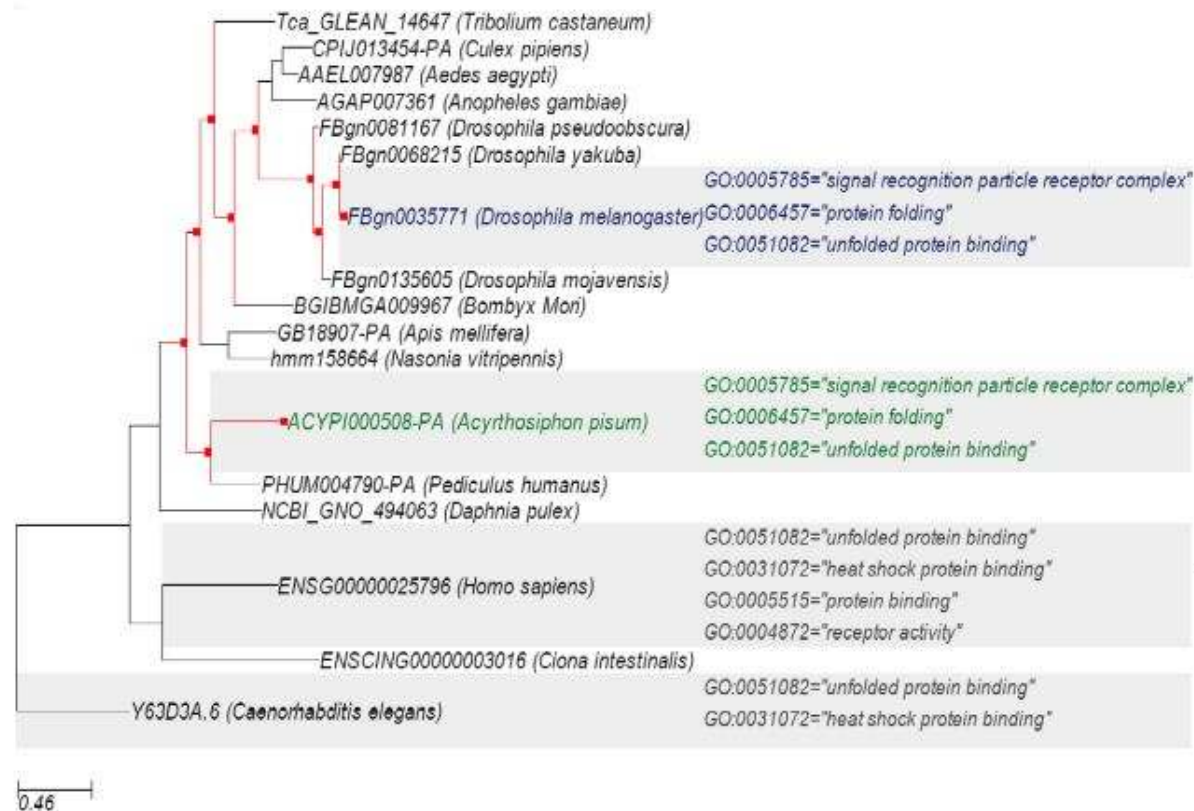


13 other sequenced arthropods and 3 out-groups



0.2



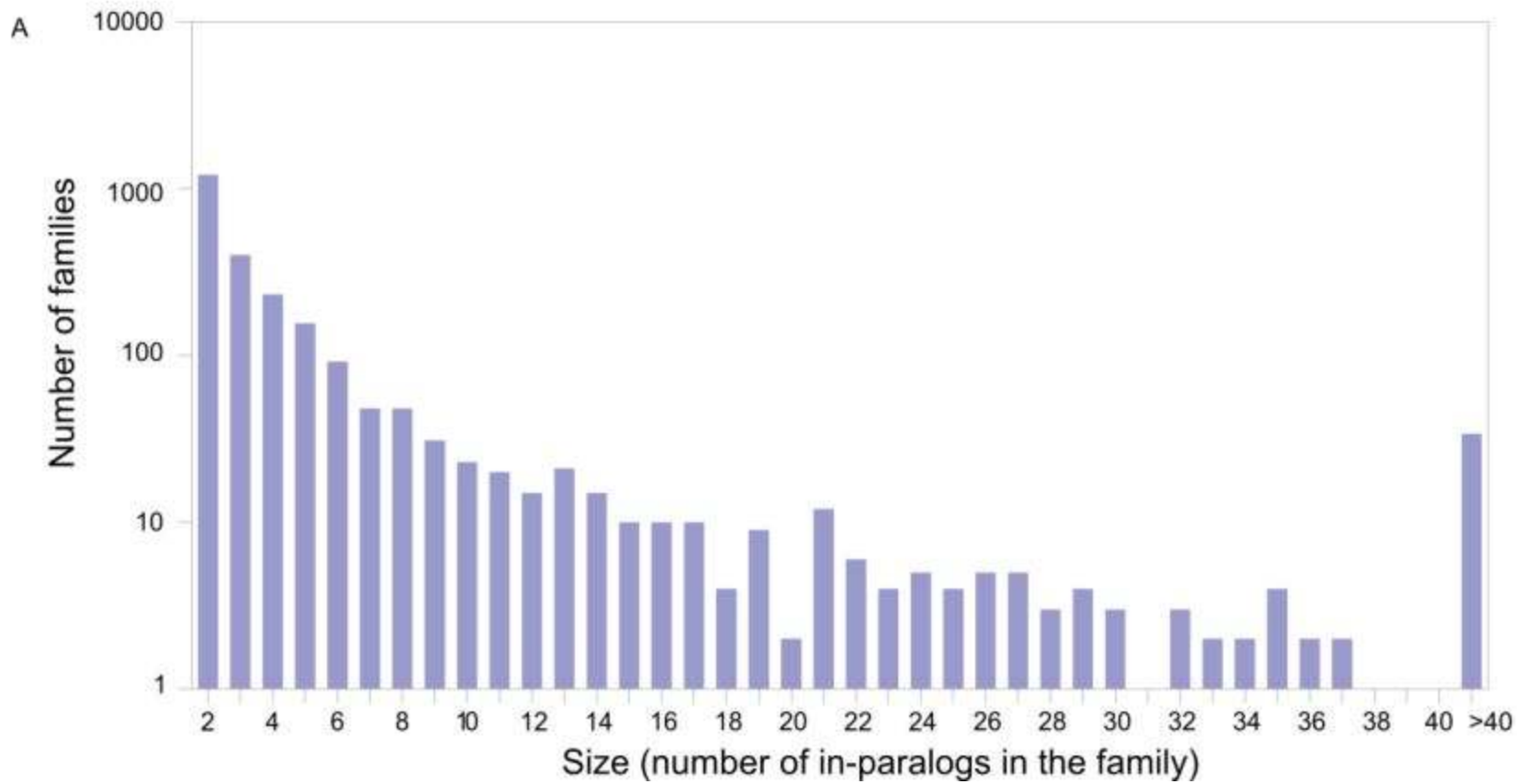


Phylogeny-based
one-to-one orthology
functional annotation

Orthologies with annotated *Drosophila melanogaster* genes:

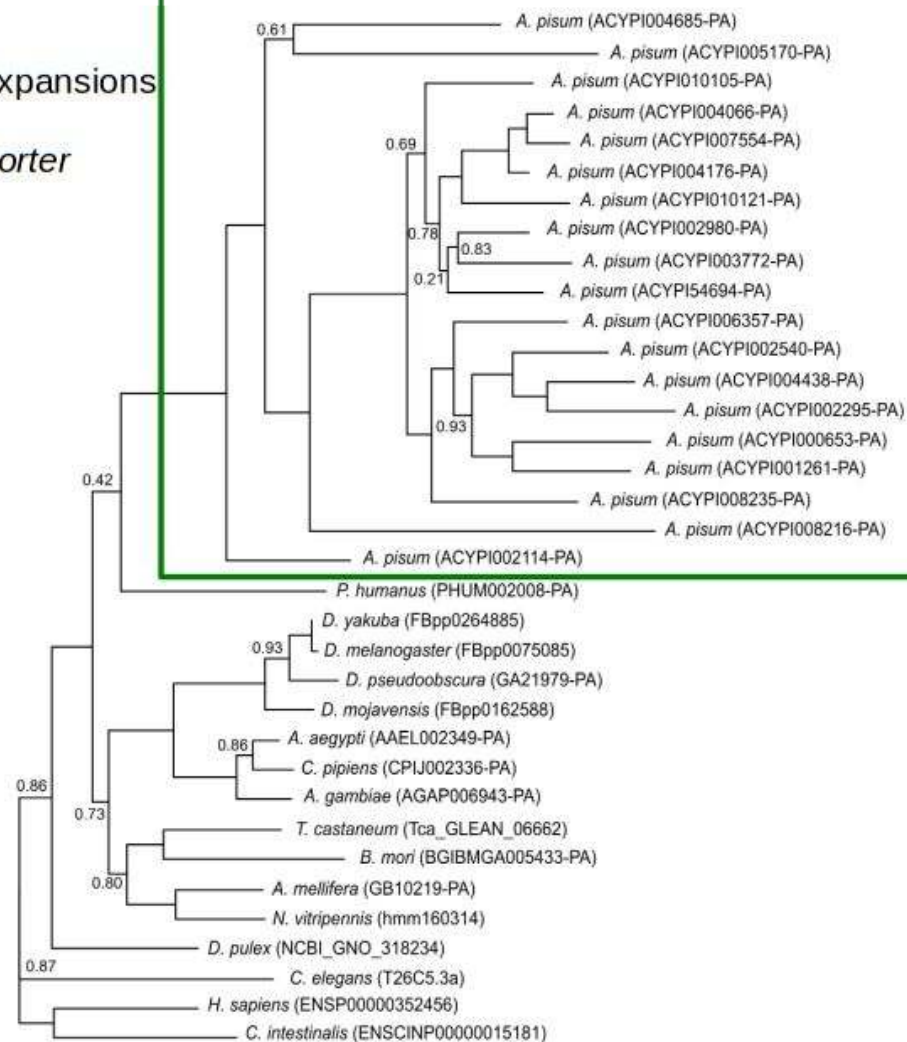
4,059 (one-to-one), **2,282** (one-to-many, many-to-many or many-to-one)

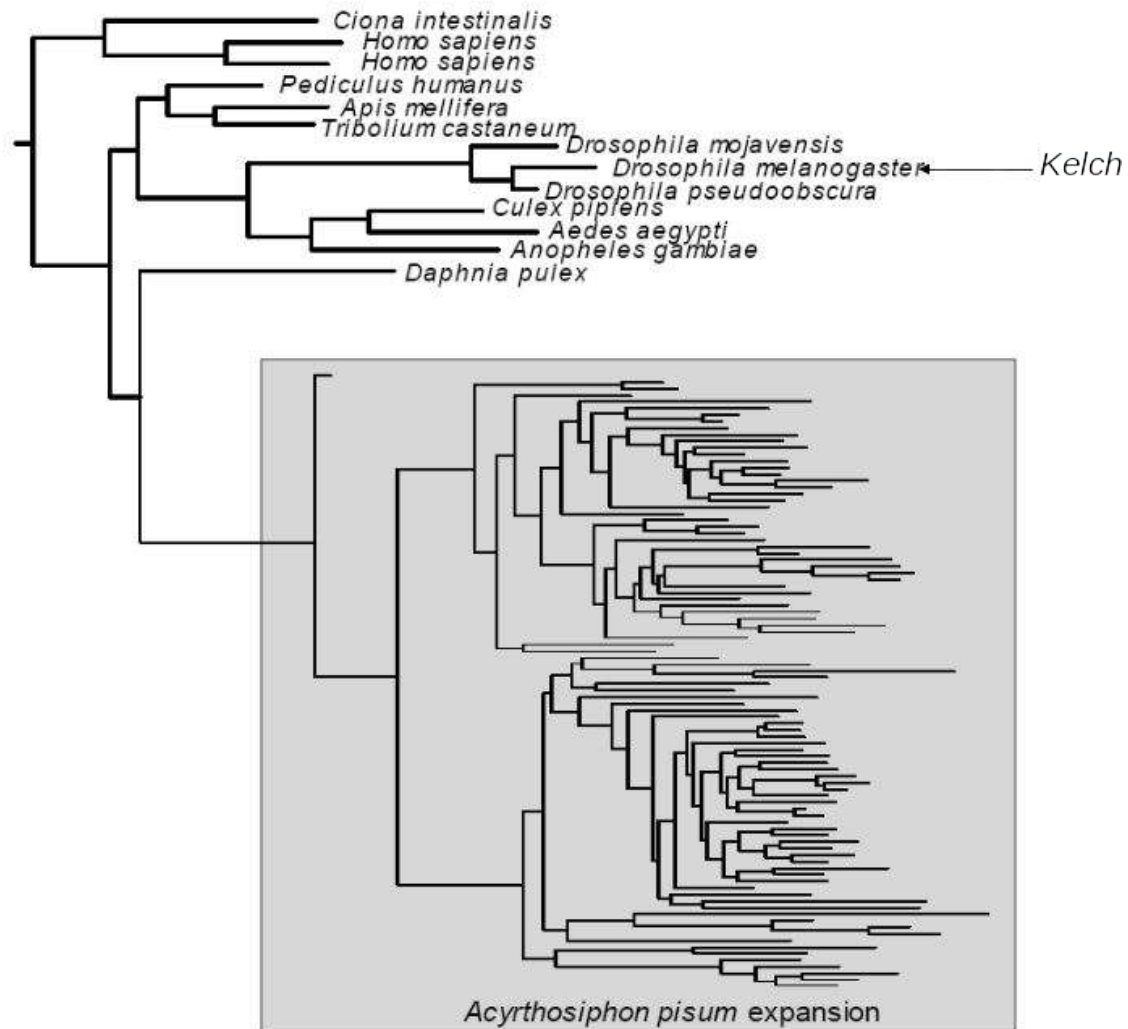
A wave of lineage-specific expansions in the pea aphid



B

Lineage-specific expansions

Acetyl-CoA transporter



In *Drosophila*, kelch protein is involved in the organization and morphology of the ovarian ring channel.

A particularity of pea aphids is a complex life cycle with reproductive polyphenism and extensive differences in ovarian morphology between the different female morphs.

Is the kelch family expansion in aphids related to such diversity?

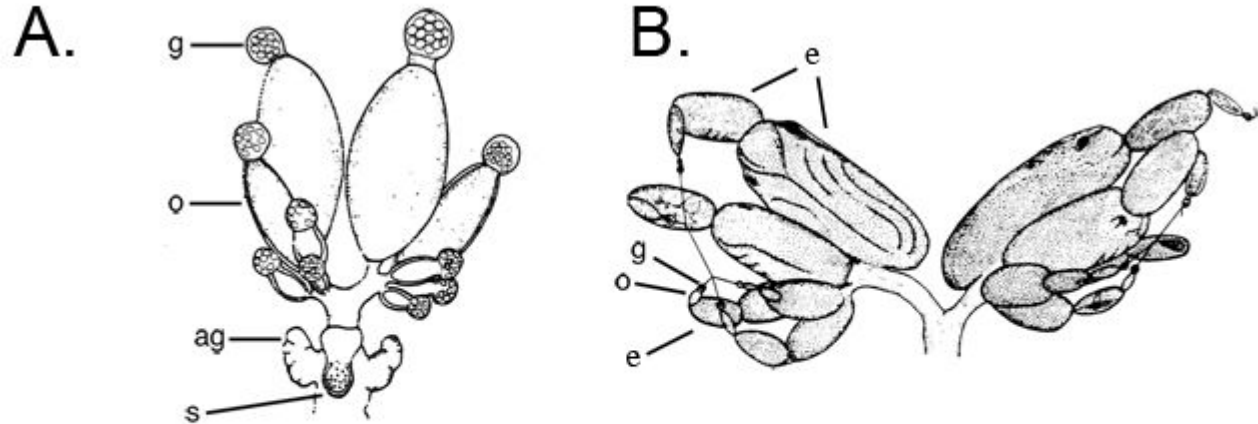


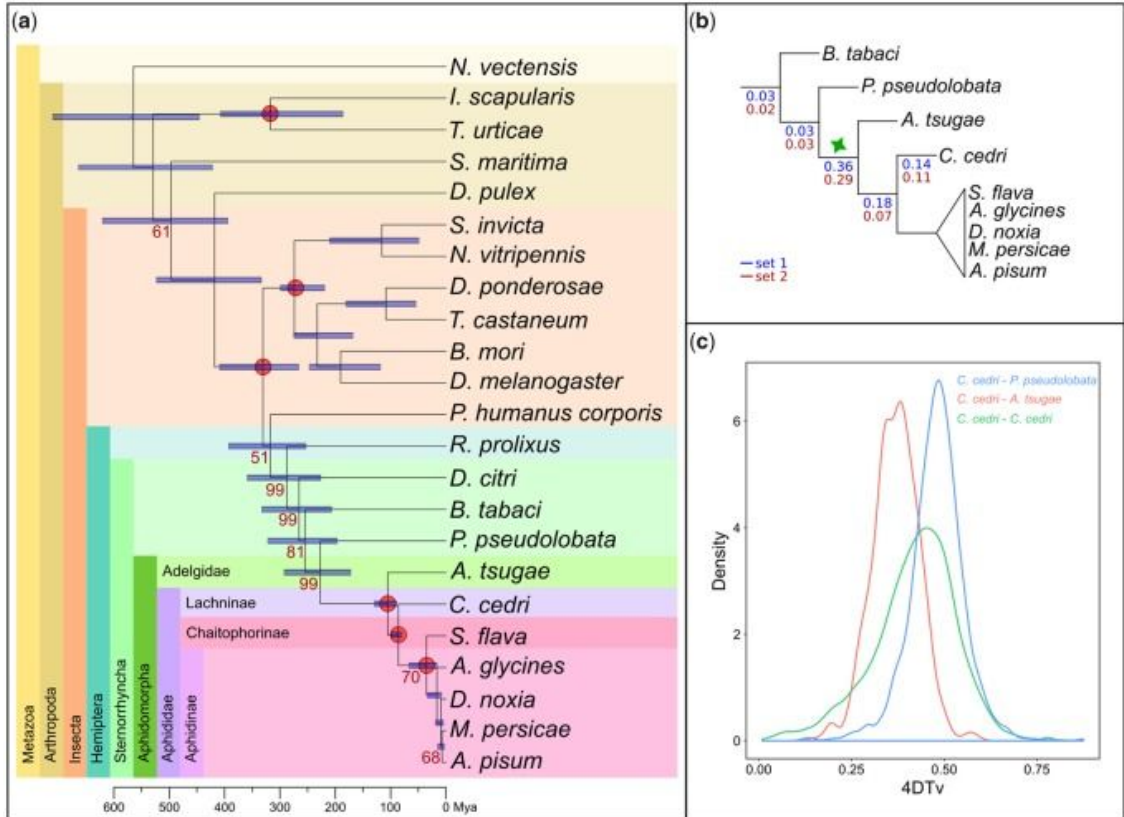
Figure 2. Viviparous and oviparous development. Oviparous (A) and viviparous (B) ovaries differ not only as to whether they possess embryos, accessory glands and spermathecae, but also in the relative size of germaria and oocytes. Abbreviations: g is germarium, o is oocyte, e is viviparous embryo, ag is accessory gland, s is spermatheca. Images are modified from Blackman, 1987.

Probable ancestral WGD(s) in the ancestor of aphids

> Mol Biol Evol. 2020 Mar 1;37(3):730-756. doi: 10.1093/molbev/msz261.

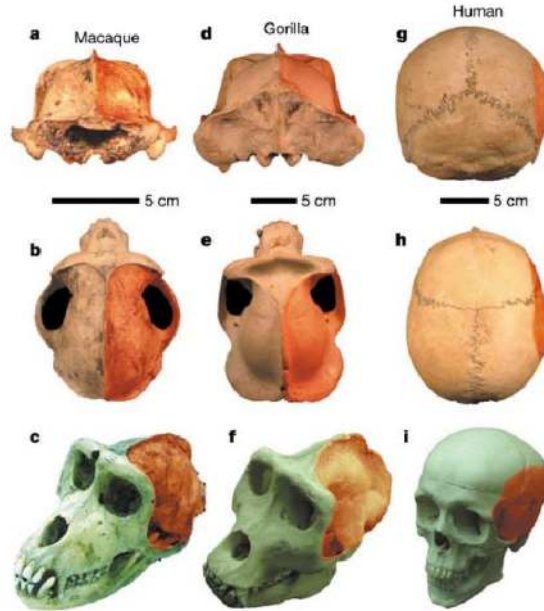
Phylogenomics Identifies an Ancestral Burst of Gene Duplications Predating the Diversification of Aphidomorpha

Irene Julca ¹, Marina Marcet-Houben ¹, Fernando Cruz ², Carlos Vargas-Chavez ³, John Spencer Johnston ⁴, Jèssica Gómez-Garrido ², Leonor Frías ², André Corvelo ^{2 5}, Damian Loska ¹, Francisco Cámara ¹, Marta Gut ^{2 6}, Tyler Alioto ^{2 6}, Amparo Latorre ^{3 7}, Toni Gabaldón ^{1 6 8}



Gene loss also drives adaptation

Loss of *Myh16* associated with cranial enlargement

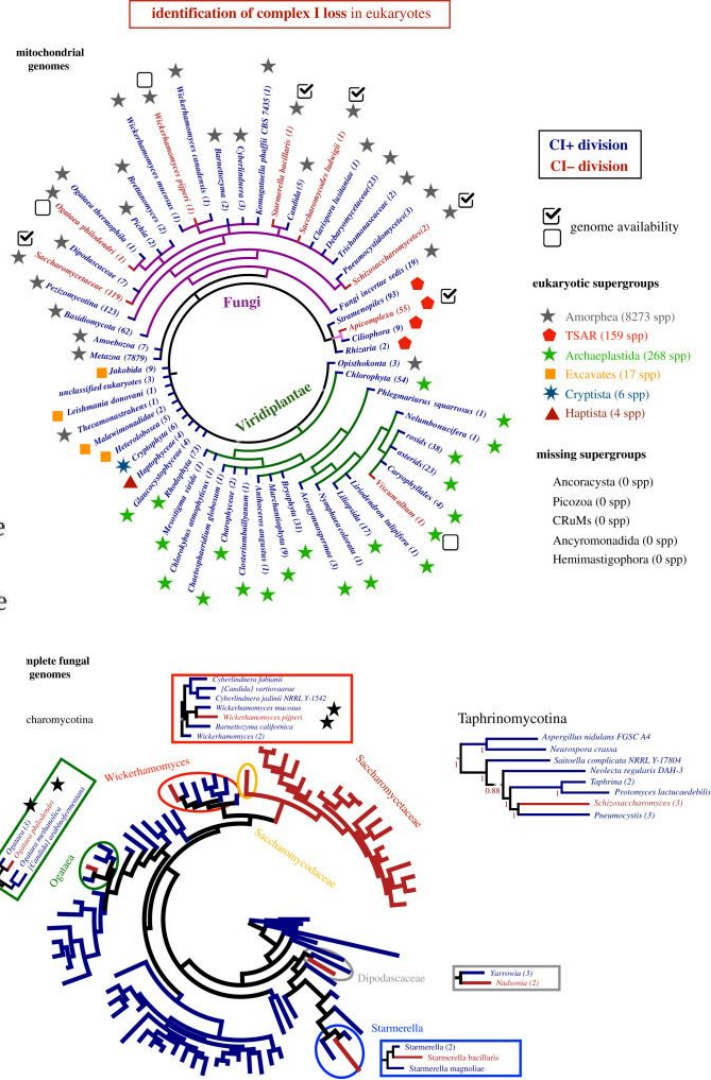


Stedman et al. (2004)

[Miguel Àngel Schikora-Tamarit](#),^{1, 2} [Marina Marcet-Houben](#),^{1, 2} [Jozef Nosek](#),³ and [Toni Gabaldón](#)^{1, 2, 4}

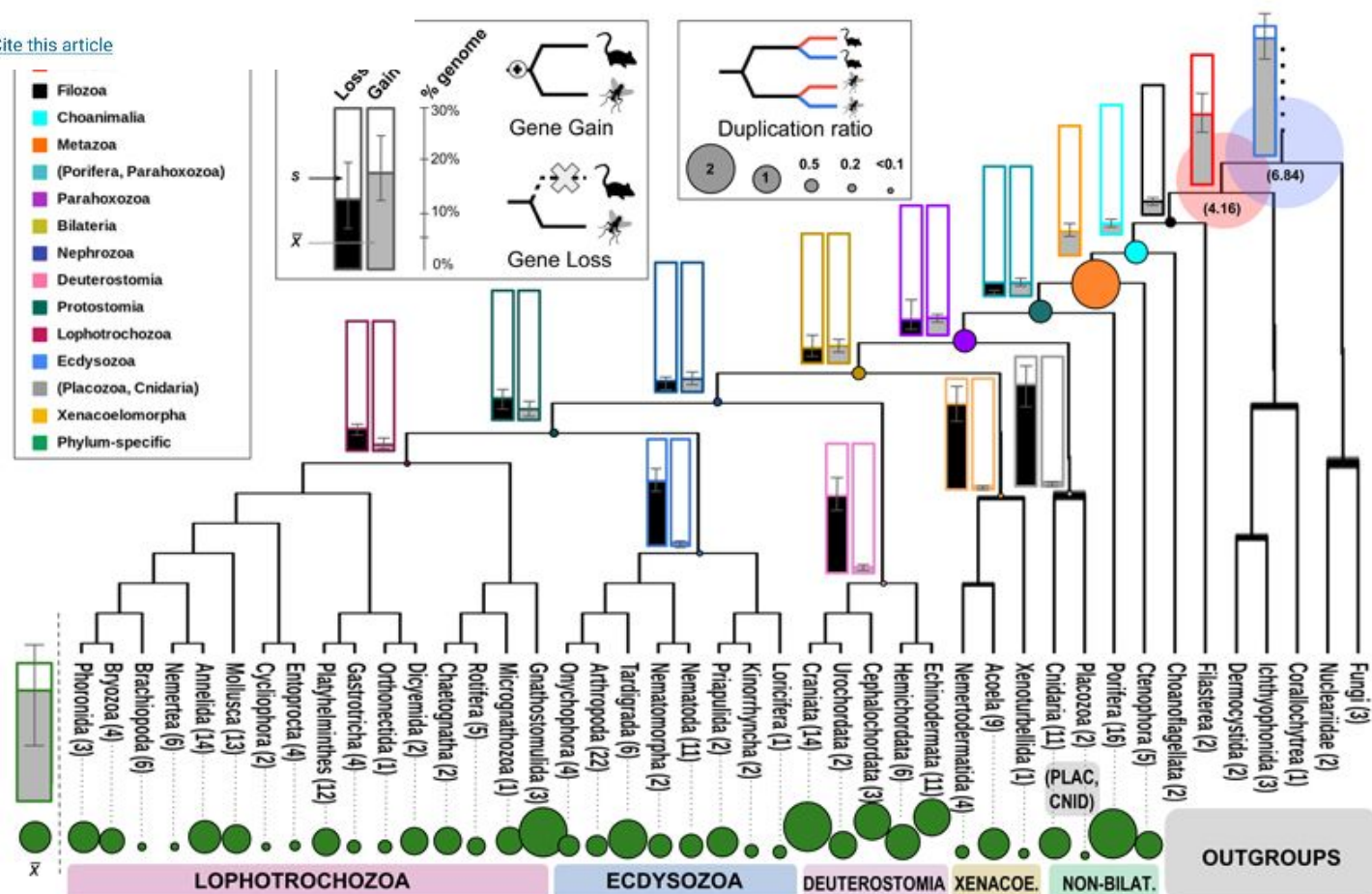
- Complex I was lost 8 independent times in fungi
- Other genomic changes correlate with CI loss

inferred genomic changes convergently associated with complex I loss. Based on these results, we predict novel complex I functional partners and relate the loss of complex I with the presence of increased mitochondrial antioxidants, higher fermentative capabilities, duplications of alternative dehydrogenases, loss of alternative oxidases and adaptation to antifungal compounds. To explain these findings, we hypothesize that a combination of previously acquired compensatory mechanisms and exposure to environmental triggers of oxidative stress (such as hypoxia and/or toxic chemicals) induced complex I loss in fungi.



Gene gain and loss across the metazoan tree of life

Rosa Fernández & Toni Gabaldón ✉

[Nature Ecology & Evolution](#) 4, 524–533 (2020) | [Cite this article](#)


Beyond duplication and loss

- Selection and recombination can explain anomalous gene trees

Convergent evolution

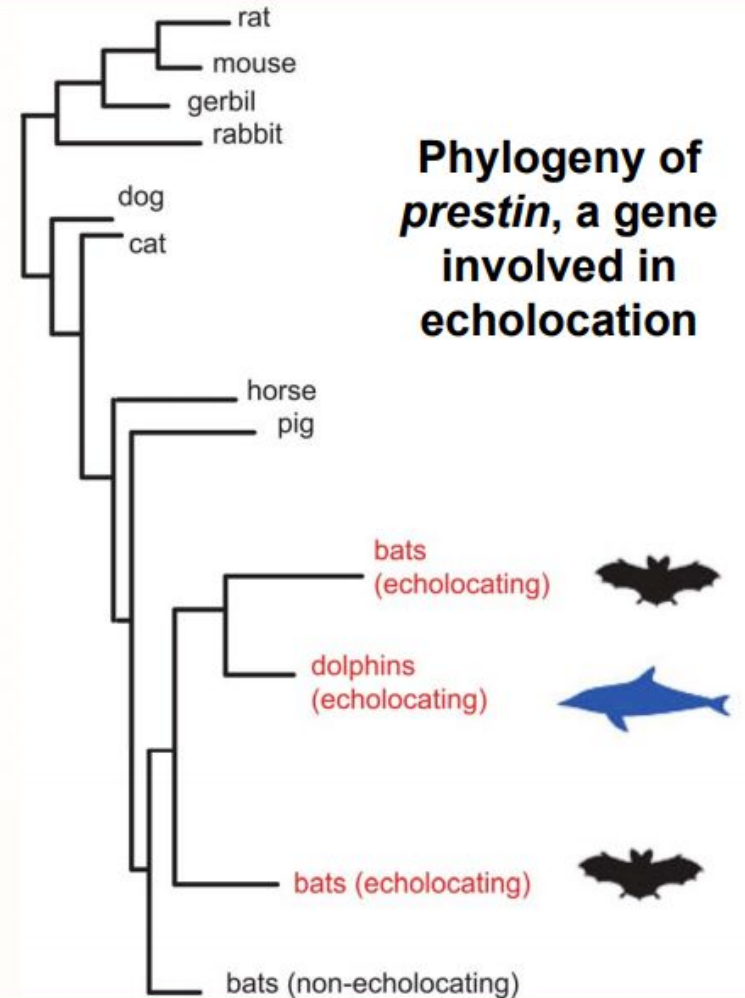
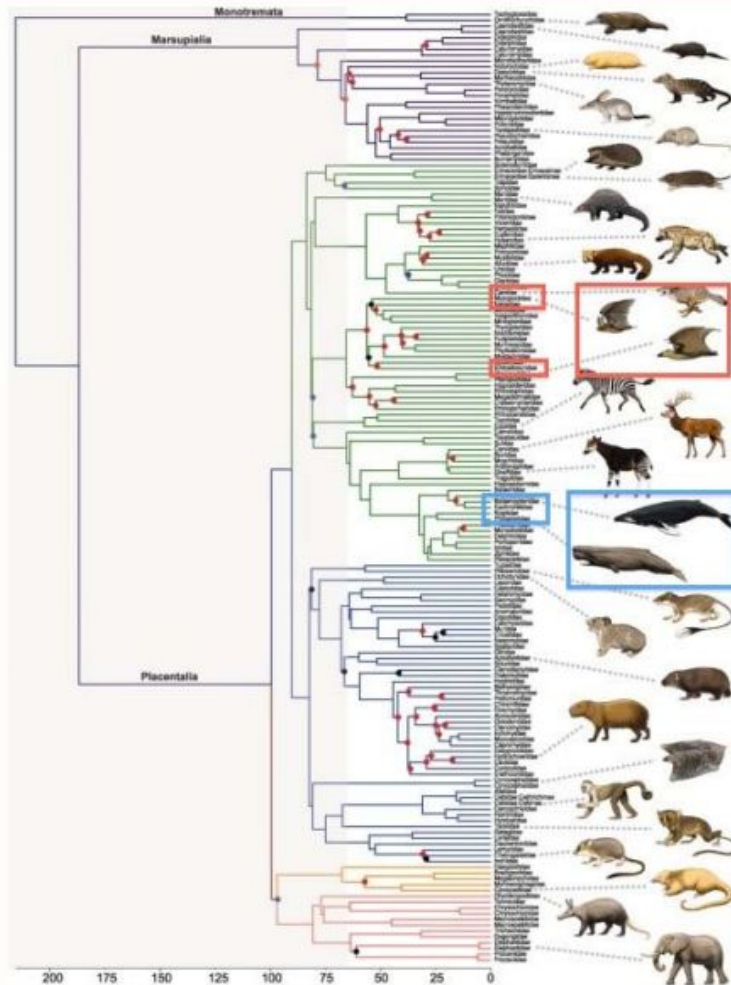


Convergent evolution

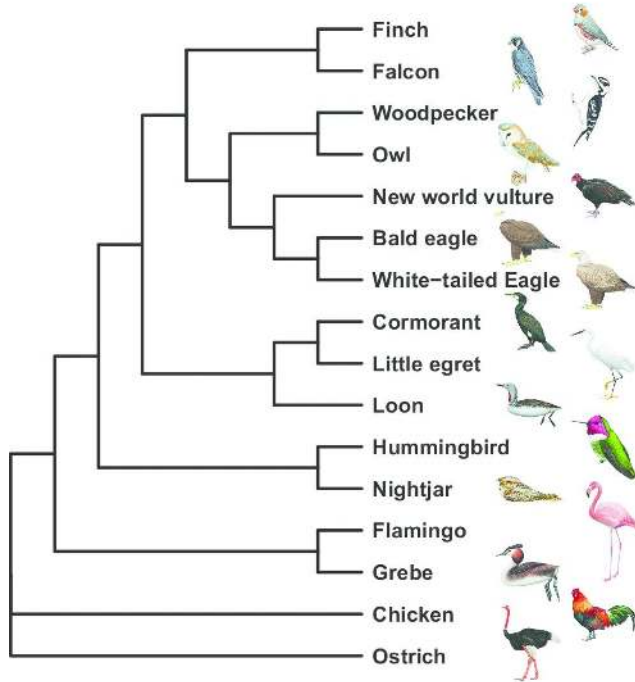


Cactus

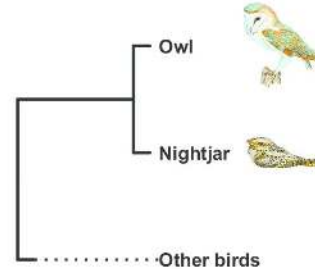
Euphorbia



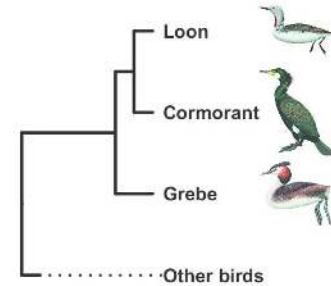
H_0 : species phylogeny



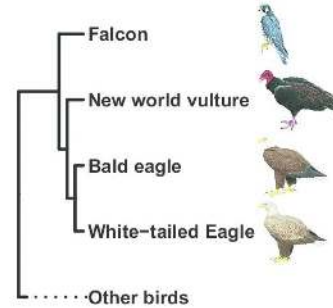
H_{noc} : nocturnal
convergence



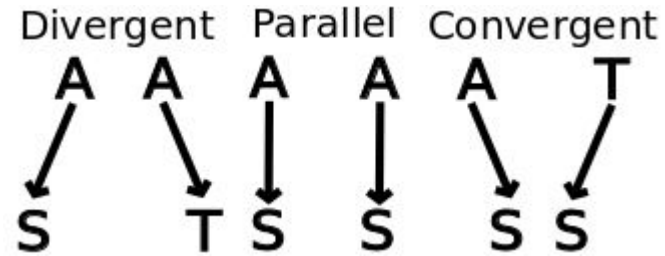
H_{foot} : foot-propelled diving
convergence



H_{rap} : diurnal raptorial
convergence

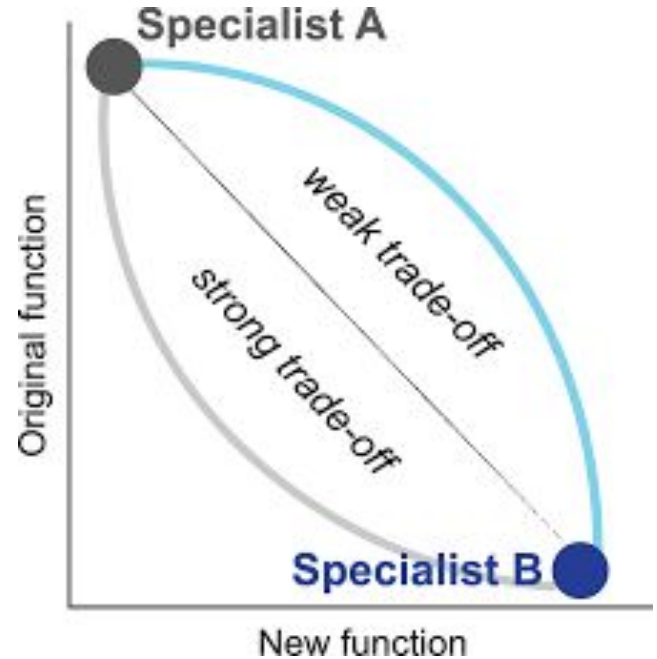


Parallel evolution



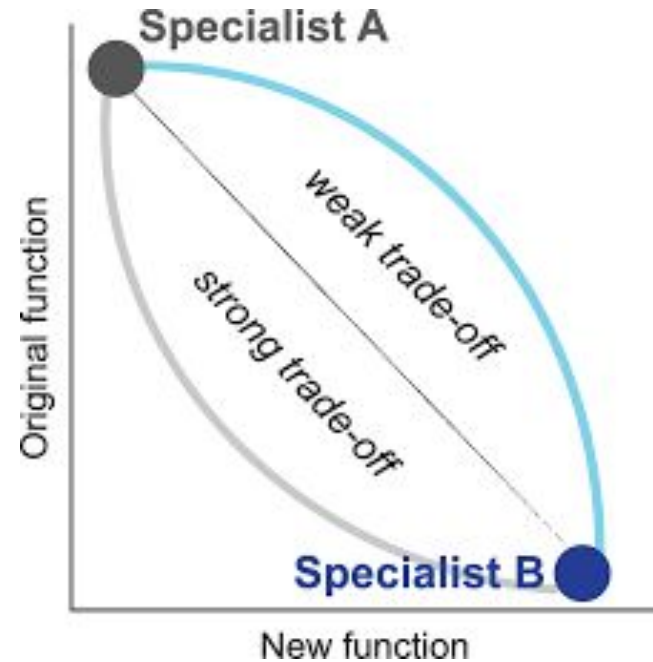
Escape from adaptive conflict

Gene with two functions in conflict



Escape from adaptive conflict

Gene with two functions in conflict



Escape from adaptive conflict

Gene with two functions in conflict

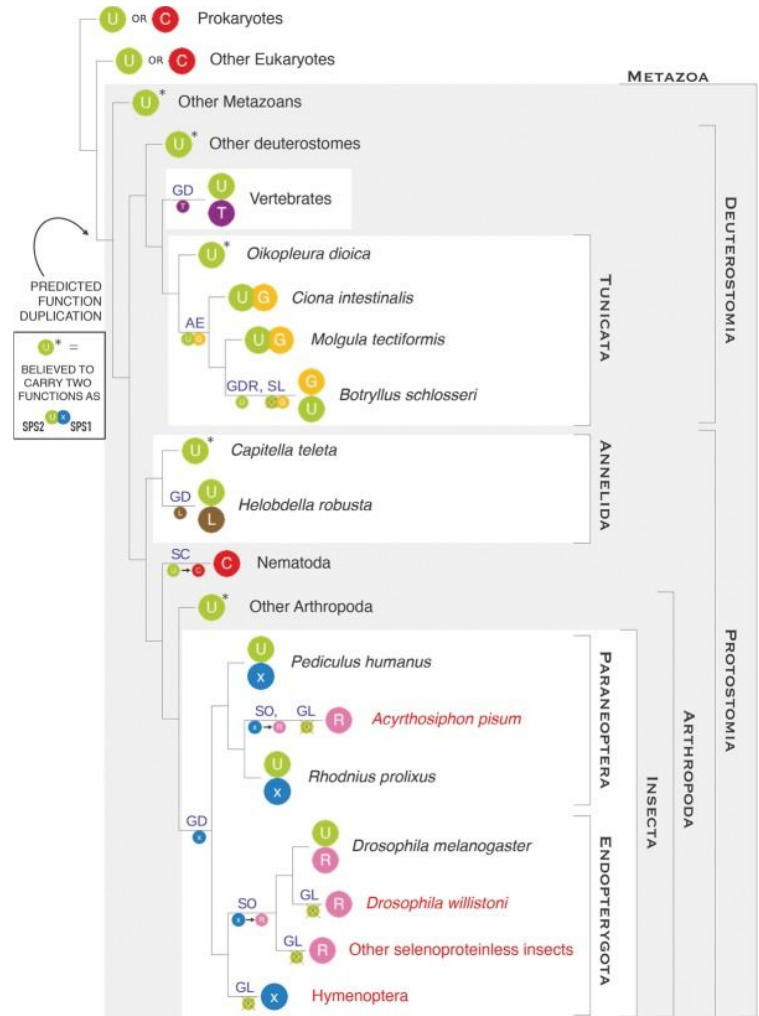


> [Genome Res.](#) 2015 Sep;25(9):1256-67. doi: 10.1101/gr.190538.115. Epub 2015 Jul 20.

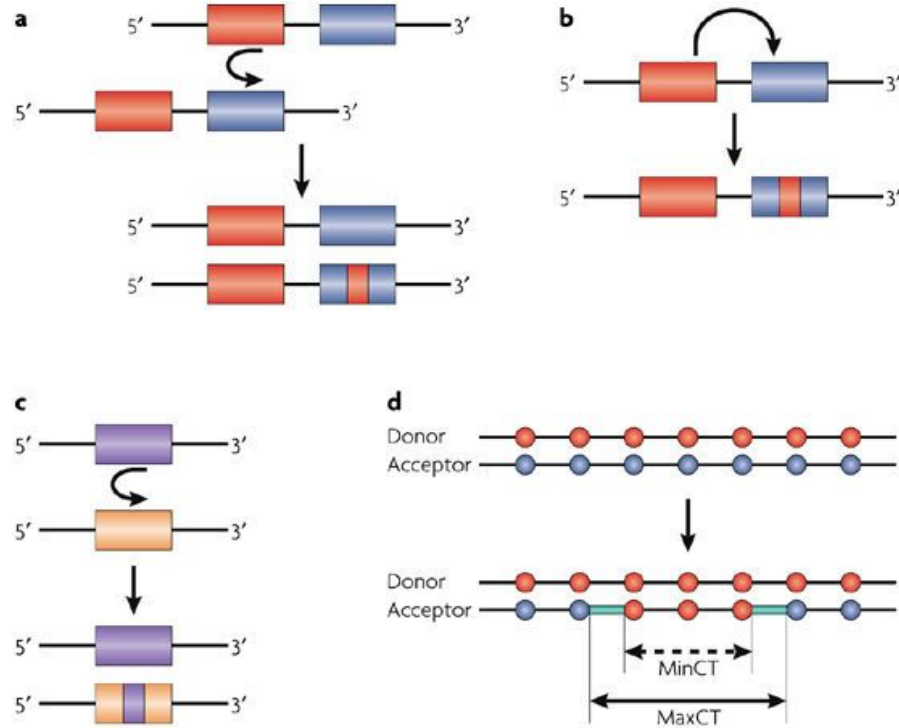
Evolution of selenophosphate synthetases: emergence and relocation of function through independent duplications and recurrent subfunctionalization

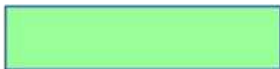
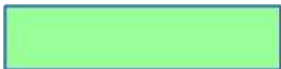
Marco Mariotti ¹, Didac Santesmasses ², Salvador Capella-Gutierrez ³, Andrea Mateo ⁴, Carme Arnan ², Rory Johnson ², Salvatore D'Aniello ⁵, Sun Hee Yim ⁶, Vadim N Gladyshev ⁶, Florenci Serras ⁴, Montserrat Corominas ⁴, Toni Gabaldón ⁷, Roderic Guigó ²

Escape from adaptive conflict



Concerted evolution (gene conversion)

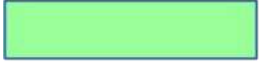




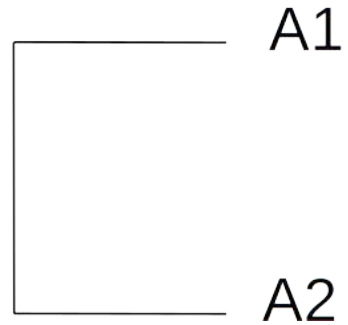
Duplication

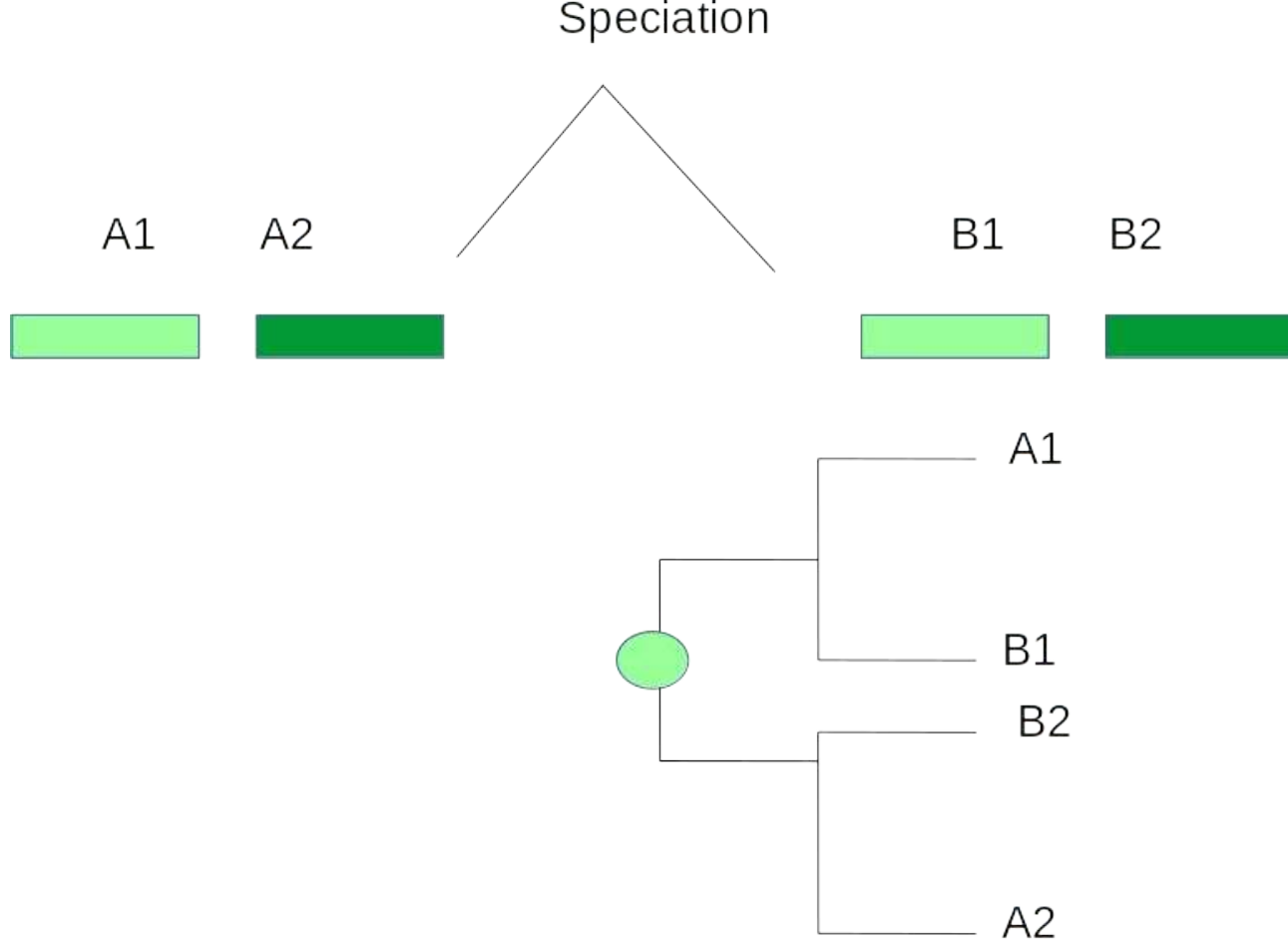
A1

A2

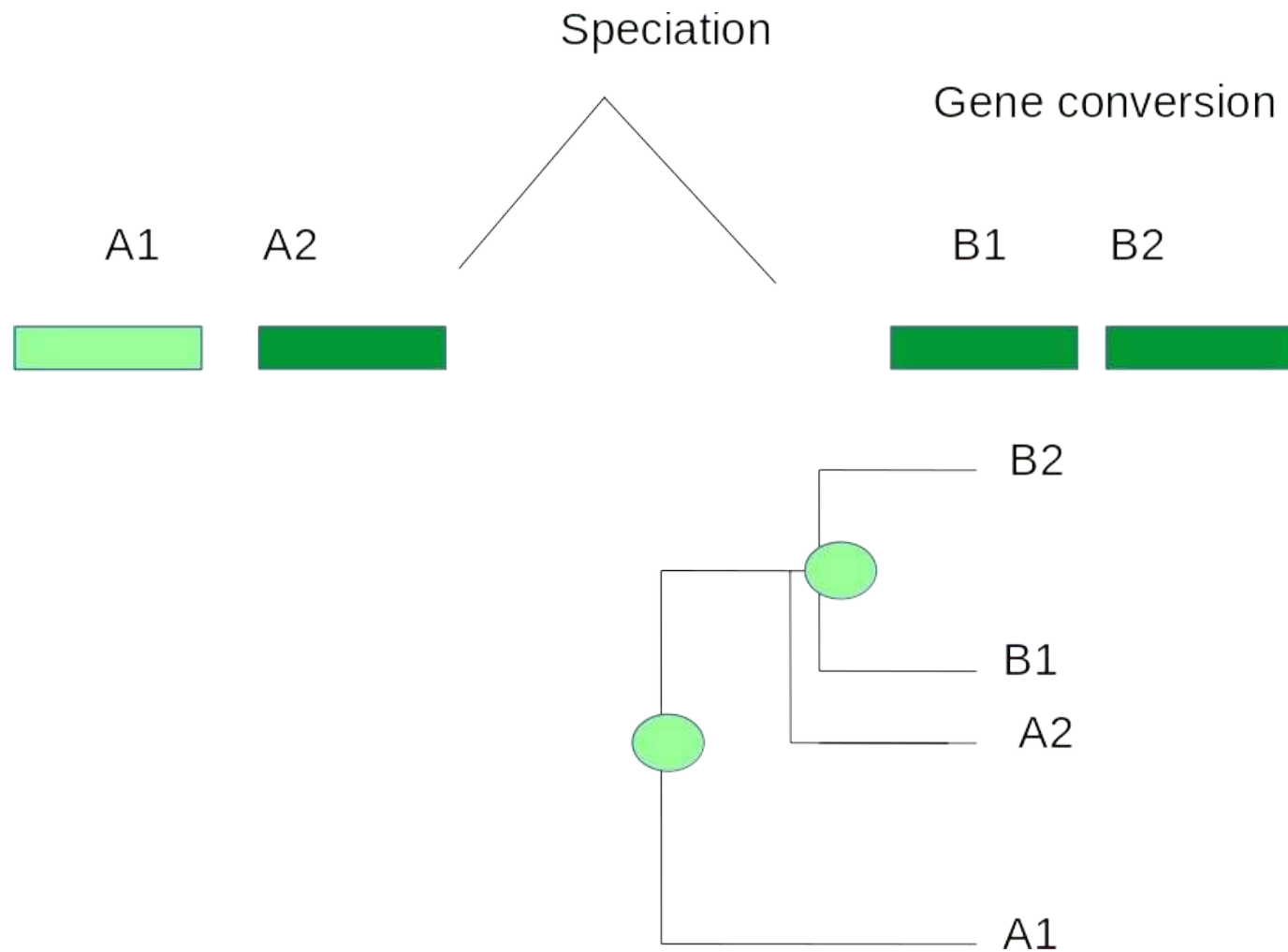


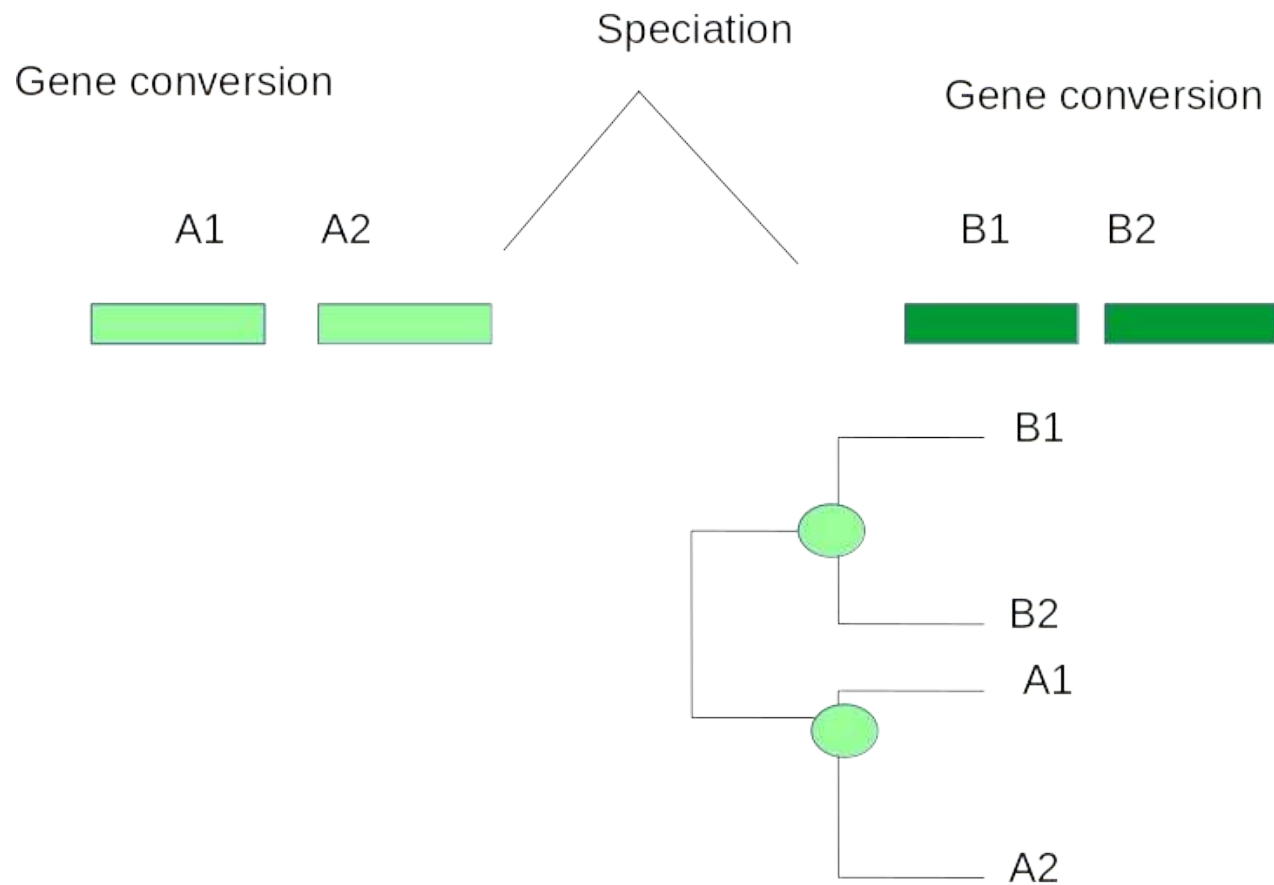
Divergence





Orthologs are closer than ancient paralogs





Paralogs are closer than orthologs, apparent parallel duplication

Phylogenomic analysis demonstrates a pattern of rare and long-lasting concerted evolution in prokaryotes

Sishuo Wang✉ & Youhua Chen

Communications Biology **1**, Article number: 12 (2018)

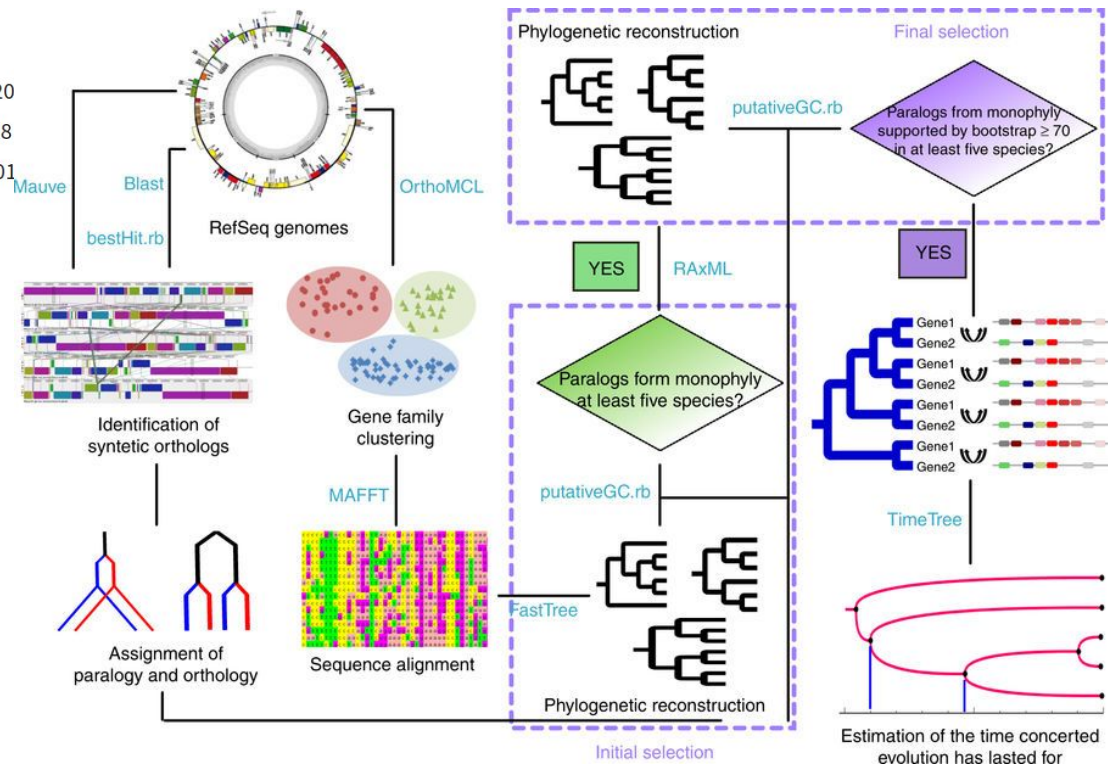
doi:10.1038/s42003-018-0014-x

Download Citation

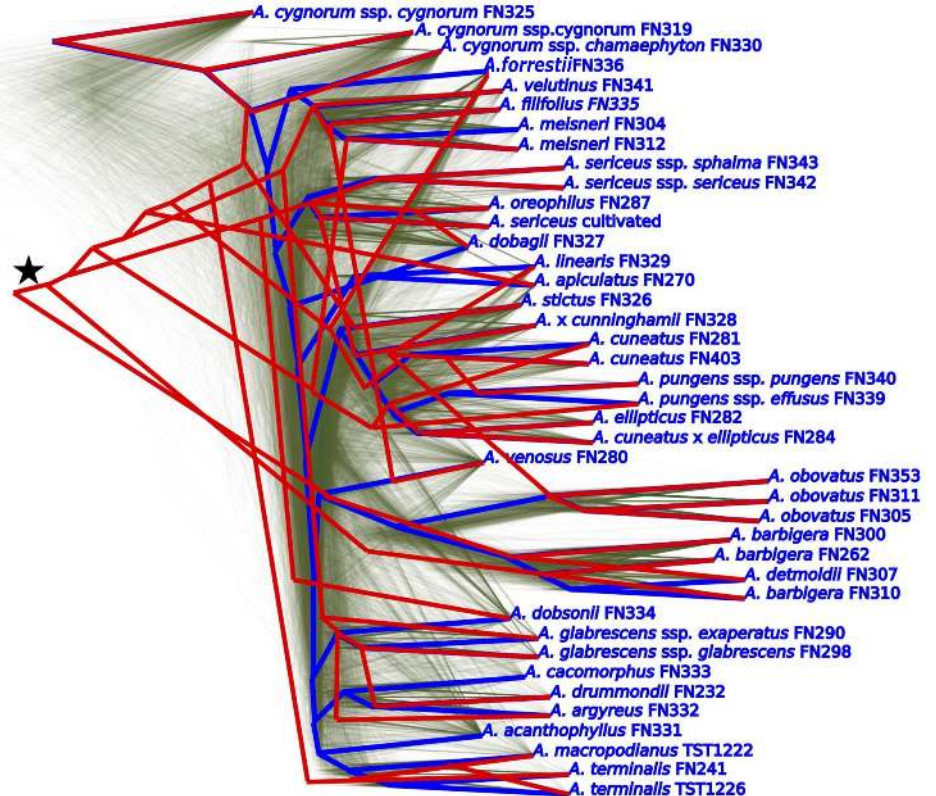
Received: 20 September 20

Accepted: 11 January 2018

Published: 08 February 2018



Reticulate (non-vertical) gene evolution



Conclusions:

- Genome-wide analyses of gene trees provide useful information to trace the evolution of genes, species, and traits
- Gene trees and species trees provide distinct information
- Now is computationally feasible to massively look at gene evolution: more powerful computers, new algorithms, data is there

Challenges:

- Gene family definition in the context of domain shuffling, and alternative splicing is unresolved
- Scalability is compromised, well-thought designs in taxonomic focus and genome choice are more important as data accumulates
- Genome annotation and the lack of common ground is a growing problem
- Functional interpretation is limited due to poor and non-specific annotations
- Green computing considerations: shall we recompute all once a new genome is added (e.g. ensembl, OMA)

THANKS