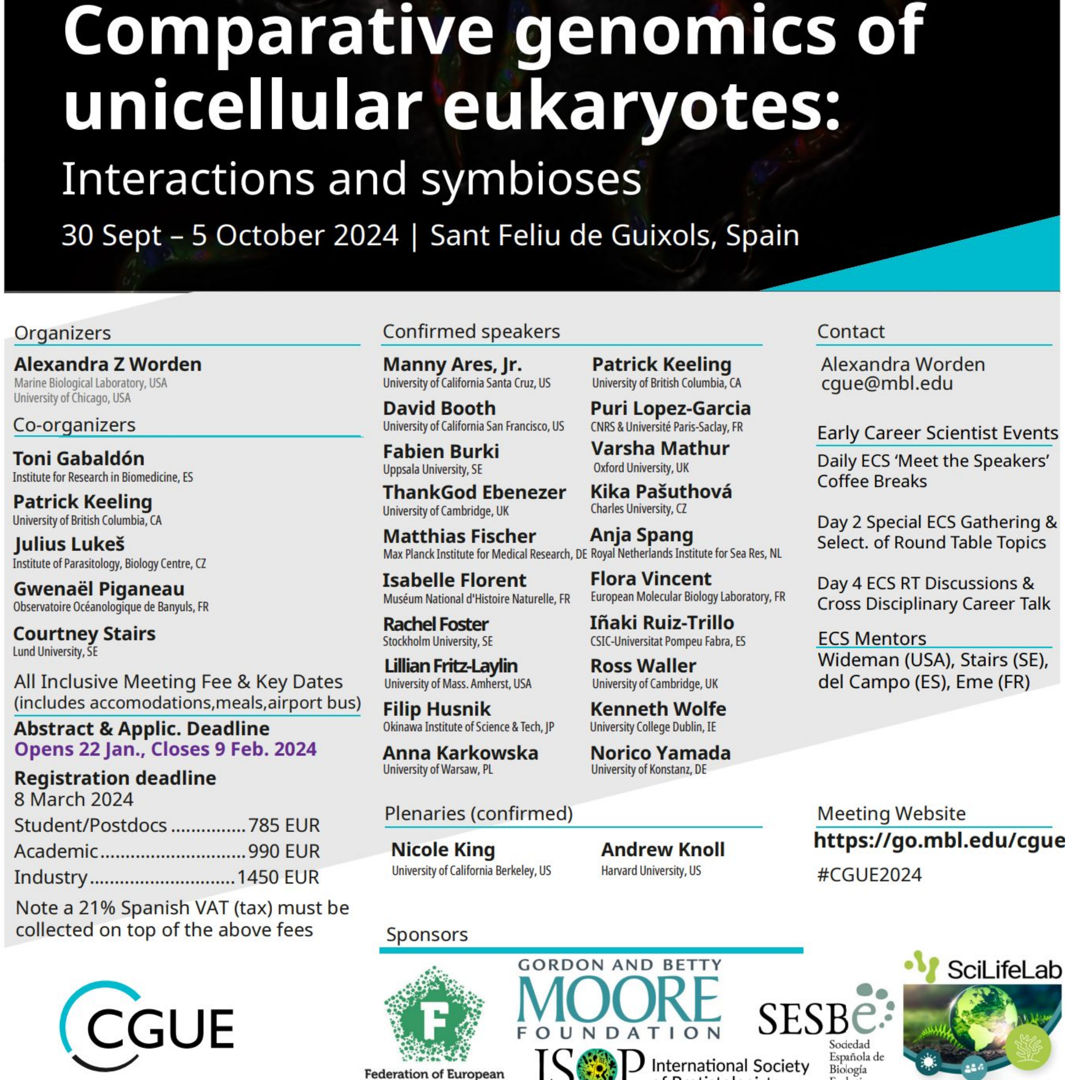
Gene Family Evolution

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

Gwenaël Piganeau
Observatoire Océanologique de Banyuls, FR

Courtney Stairs
Lund University, SE

All Inclusive Meeting Fee & Key Dates (includes accomodations,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

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
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

Plenaries (confirmed)

Nicole King
University of California Berkeley, US

Andrew Knoll
Harvard University, US

Sponsors

Federation of European

GORDON AND BETTY MOORE FOUNDATION

ISOP International Society of Protistologists

SESBE Sociedad Española de Biología Evolutiva

SciLifeLab

Contact

Alexandra Worden
cgue@mbl.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

CGUE

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

Follow us & Spread the Word!

ESEB2025

@eseb2025

ESEB2025

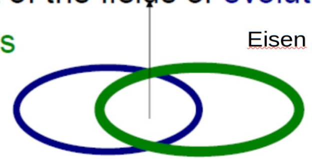
BARCELONA • 17-22 August 2025

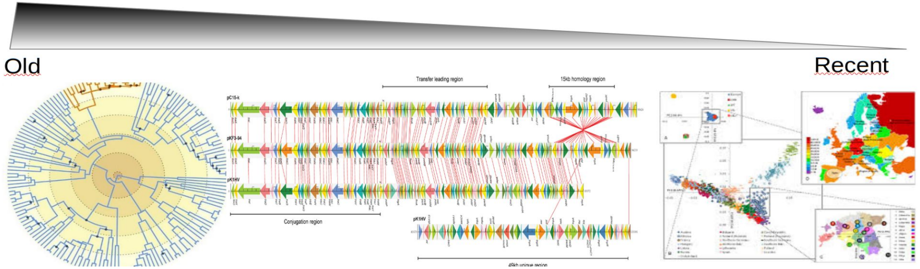
CONGRESS OF THE EUROPEAN SOCIETY FOR EVOLUTIONARY BIOLOGY

BSC
Barcelona
Supercomputing
Center

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

Eisen J. (2005)





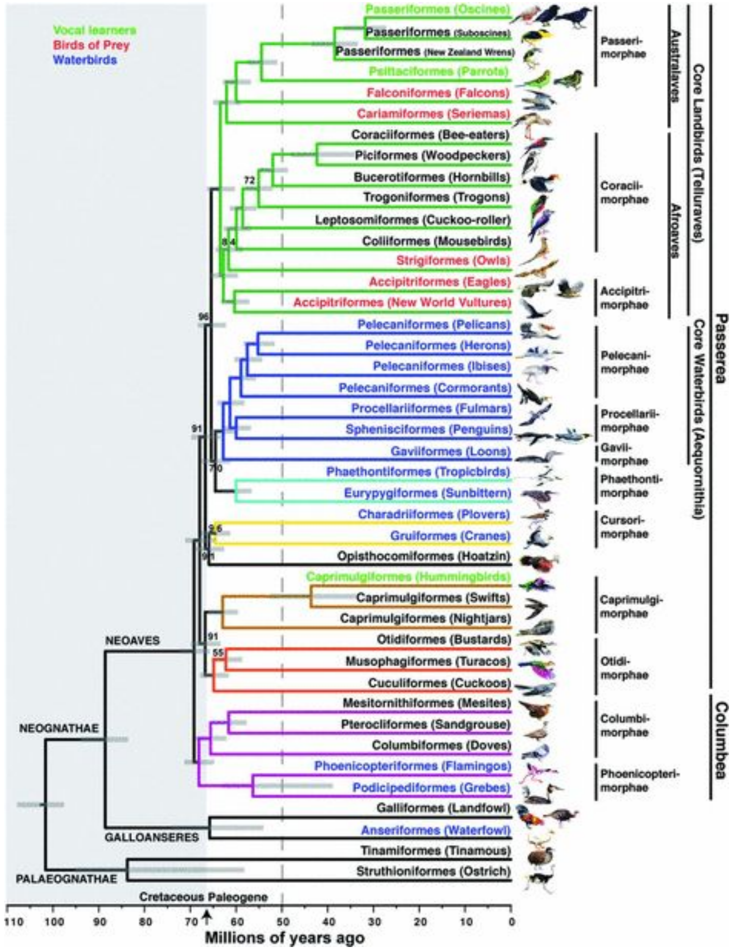
- Origin and early evolution of eukaryotes
- Reticulated evolution in eukaryotes
- Genomic and phenotypic evolution in yeast pathogens (Candida)
- Host-microbiome interactions
- Phylogenomics applied to study of biodiversity and phenotypic transitions

Gene Family Evolution

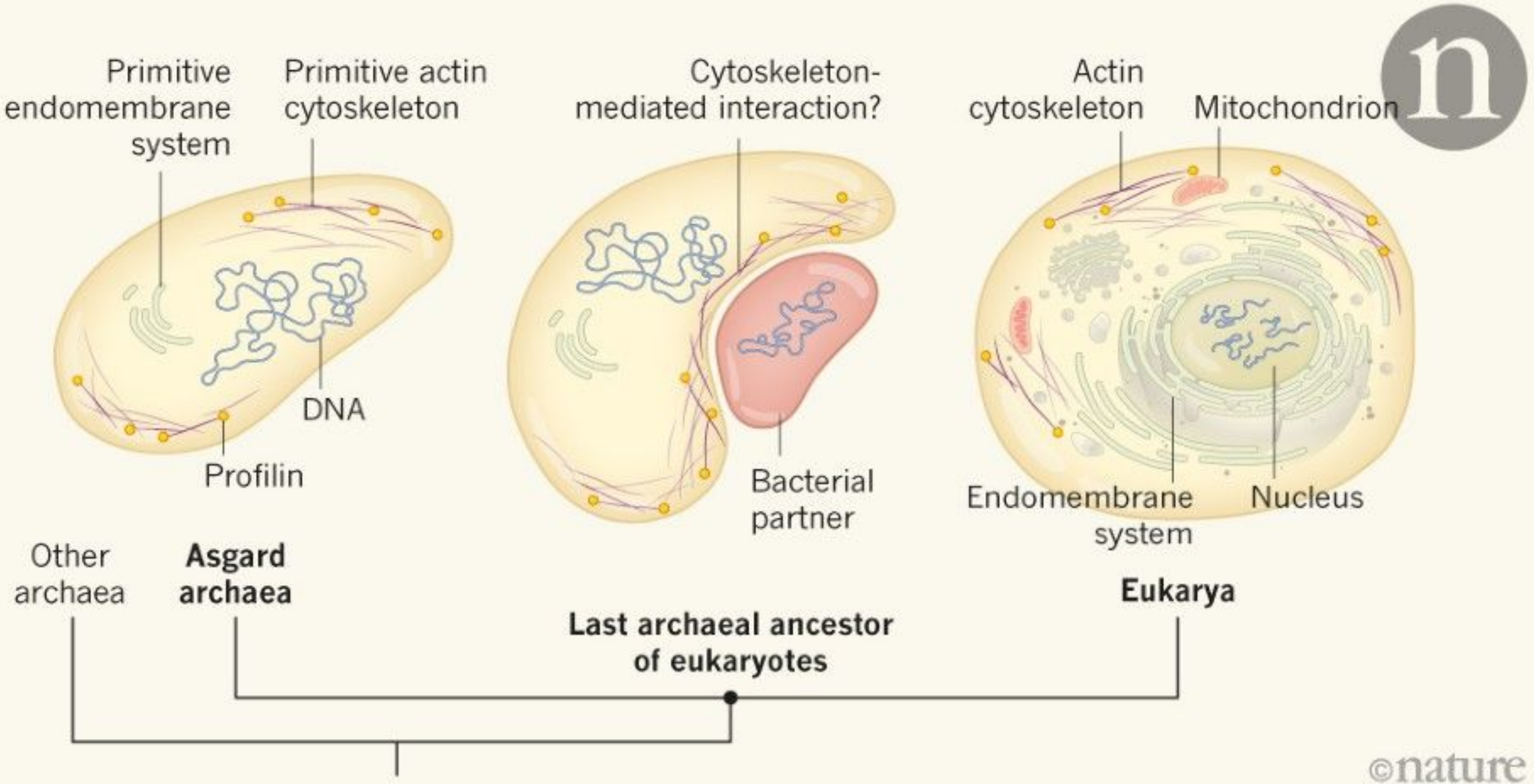
Toni Gabaldón
January, 2024

DO NOT MISS THE FOREST FOR THE TREE

Does DNA act as a telephone line? p. 1284
Mutations enhancing leukemia development pp. 1291 & 1373
A stable gold support gives sharper resolution p. 1377
Science
$10
12 DECEMBER 2014
sciencemag.org
AAAS
SPECIAL ISSUE
Avian genomes
Sequencing across the bird species tree
p. 1308



363 bird genomes


Primitive endomembrane system
Primitive actin cytoskeleton
DNA
Profilin
Cytoskeleton-mediated interaction?
Bacterial partner
Actin cytoskeleton
Mitochondrion
Endomembrane system
Nucleus
Other archaea
Asgard archaea
Last archaeal ancestor of eukaryotes
Eukarya
©nature

Eme and Ettema (2018)

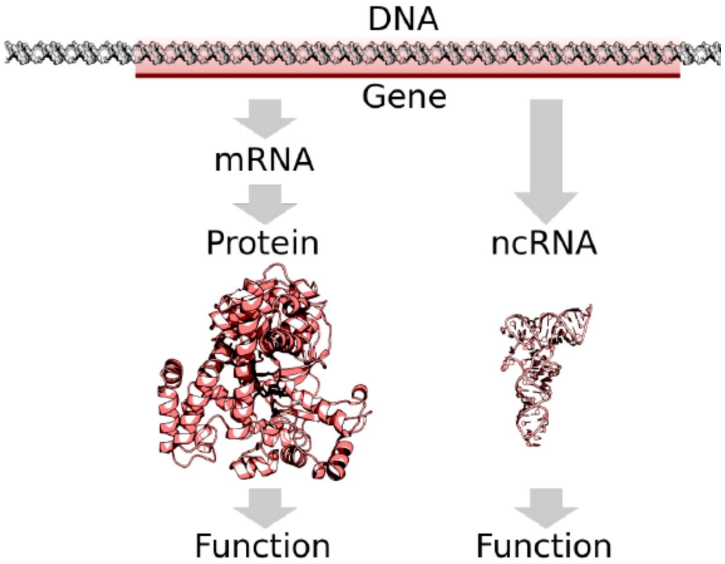
Why care about gene family evolution?

- 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?

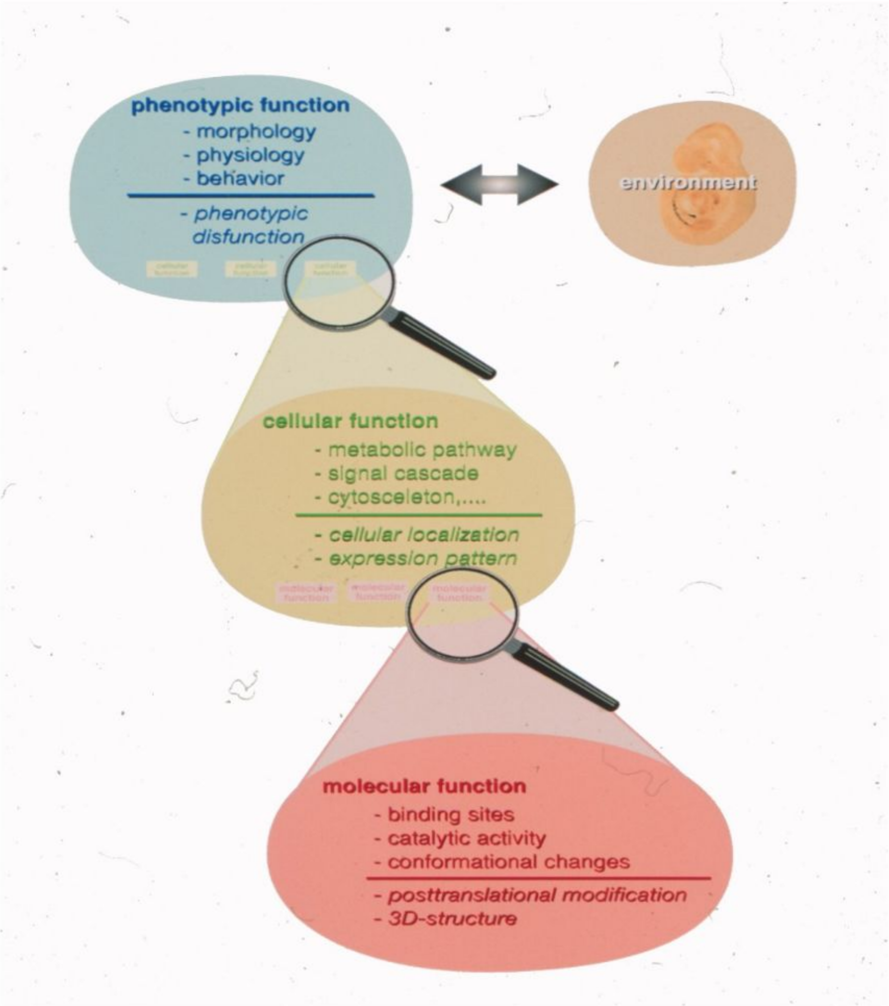
A modern definition:

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.



phenotypic function

- morphology
- physiology
- behavior

- *phenotypic disfunction*

cellular function

- metabolic pathway
- signal cascade
- cytosceleton,....

- *cellular localization*
- *expression pattern*

molecular function

- binding sites
- catalytic activity
- conformational changes

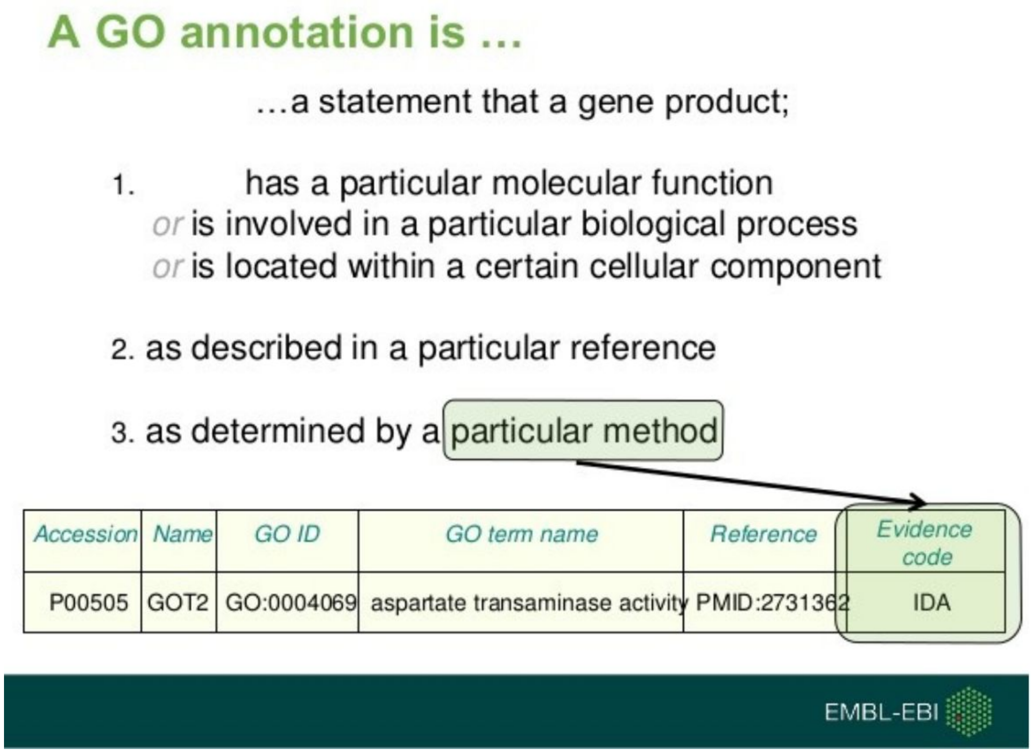
- *posttranslational modification*
- *3D-structure*

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

EMBL-EBI

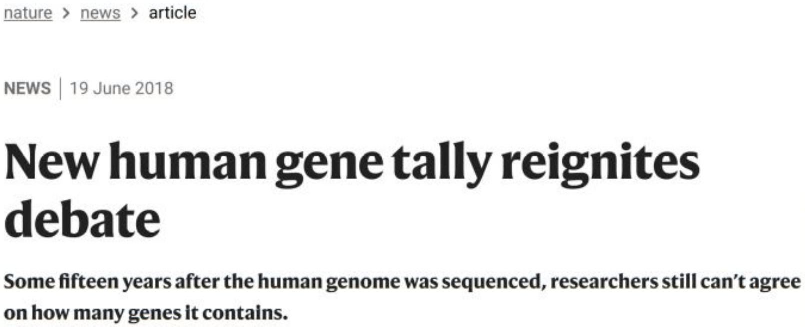
From genome to gene content: gene prediction

- De novo
- Homology-based
- RNAseq based

From genome to gene content: gene prediction

- De novo
- Homology-based
- RNAseq based

Still an issue!

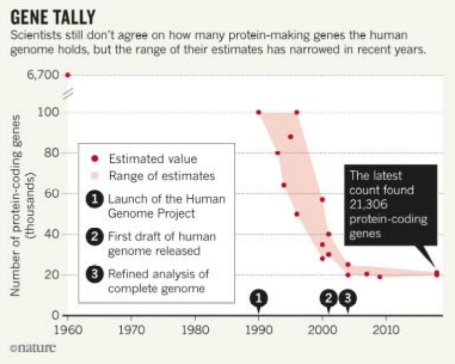


nature > news > article

NEWS | 19 June 2018

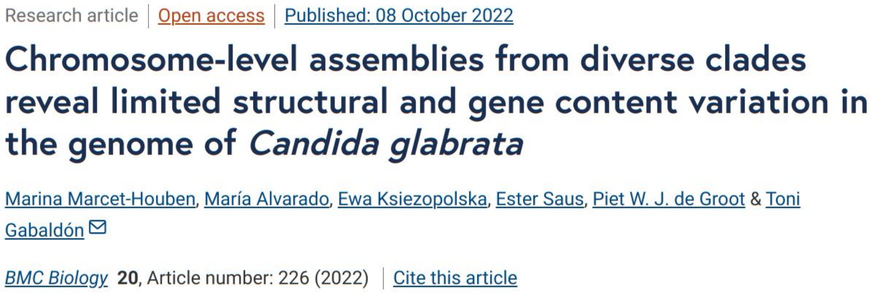
New human gene tally reignites debate

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



GENE TALLY

Scientists still don't agree on how many protein-making genes the human genome holds, but the range of their estimates has narrowed in recent years.

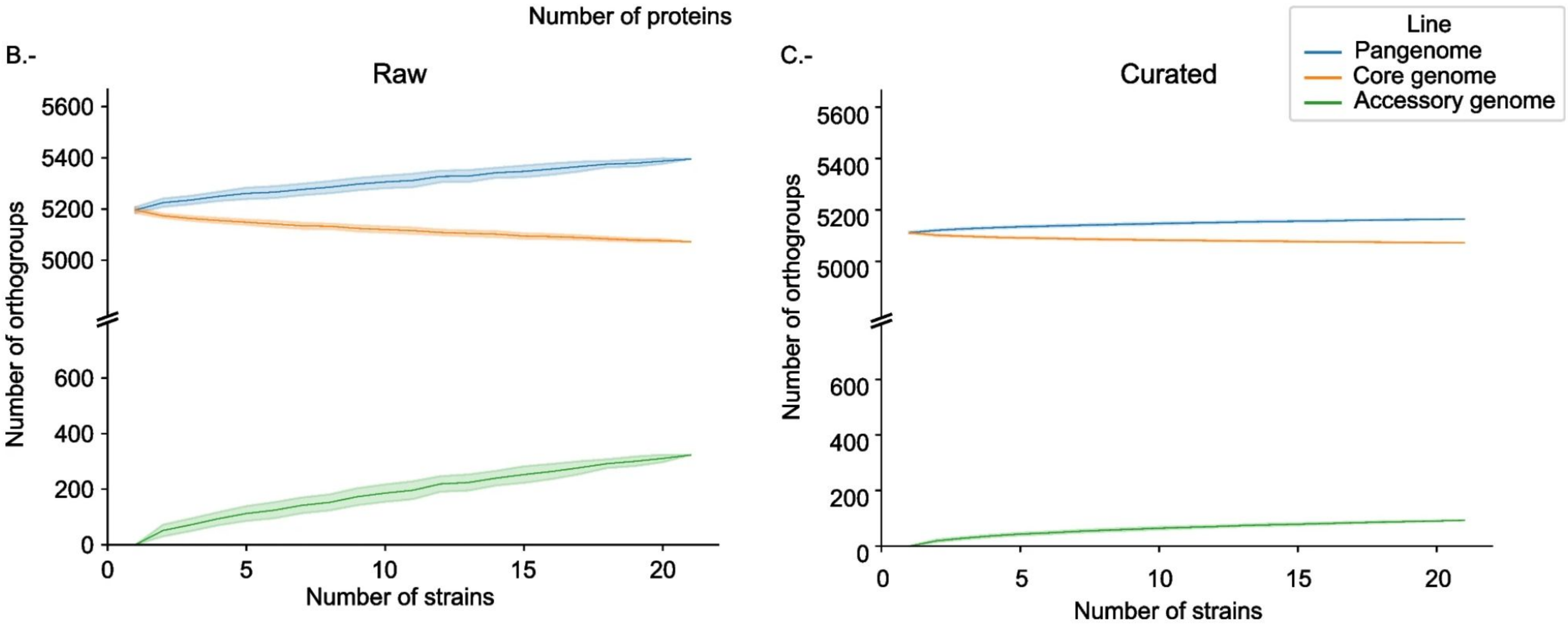


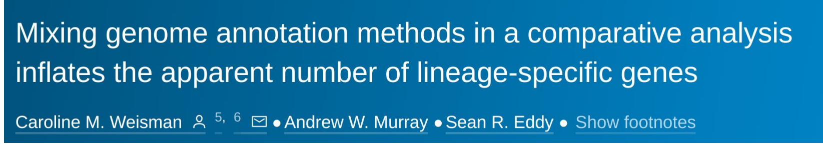
Research article | Open access | Published: 08 October 2022

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

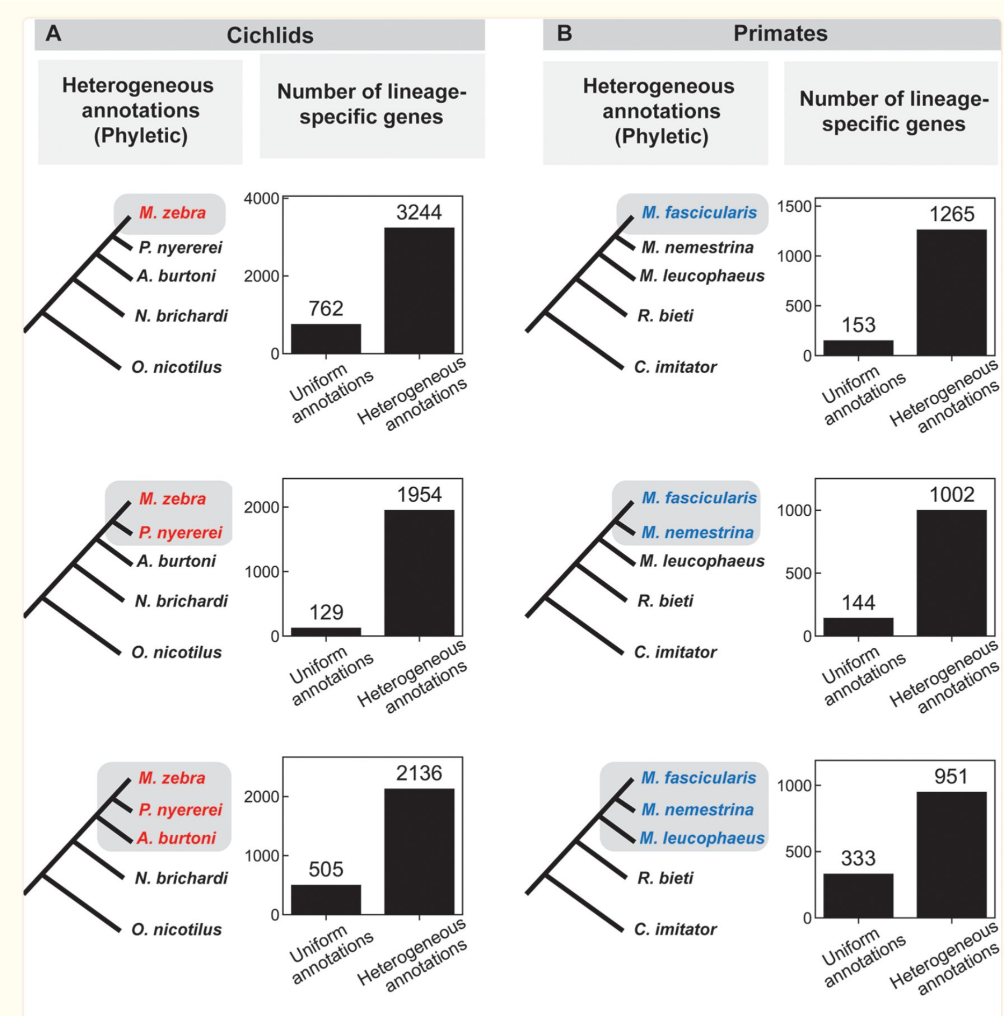
BMC Biology **20**, Article number: 226 (2022) | Cite this article



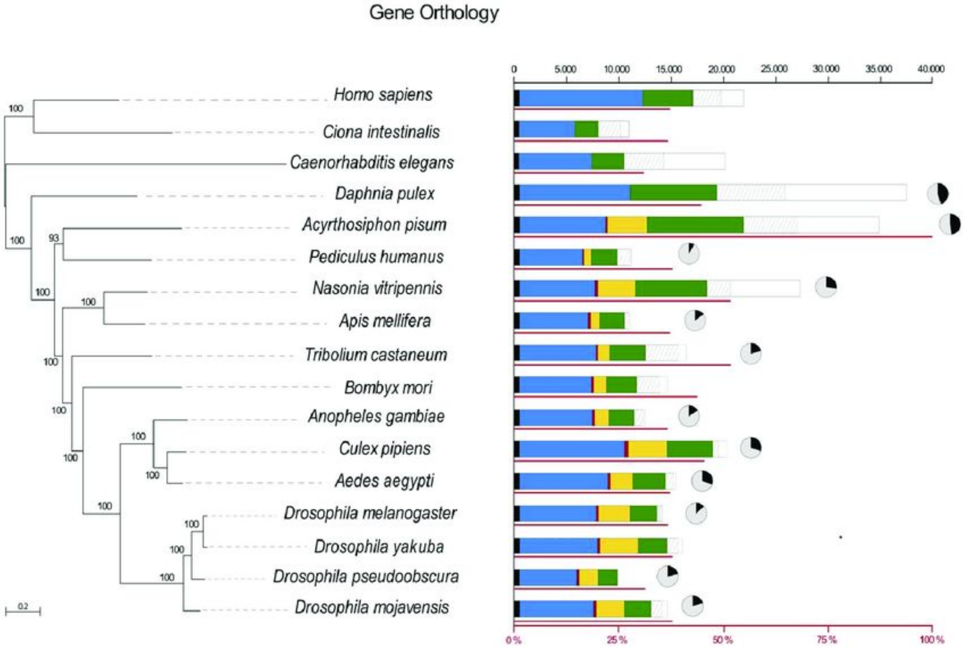


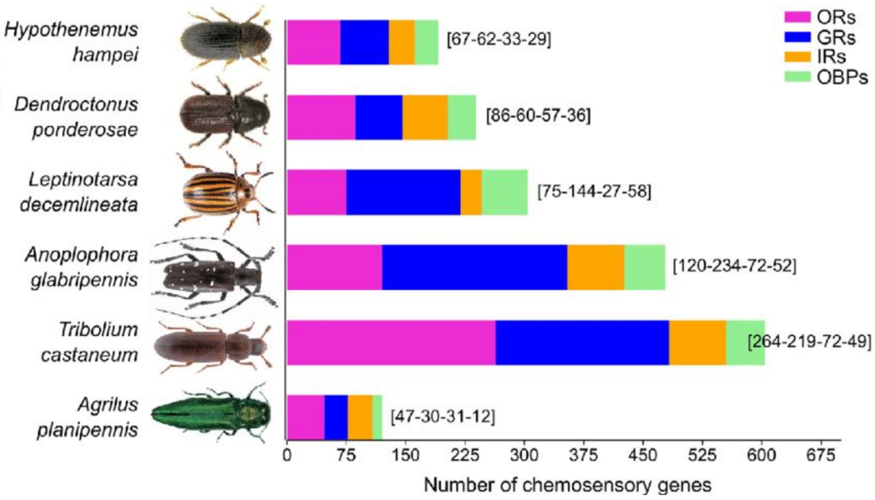
Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes

Caroline M. Weisman [5, 6] • Andrew W. Murray • Sean R. Eddy • Show footnotes



Variation of gene content across species





A gene family:

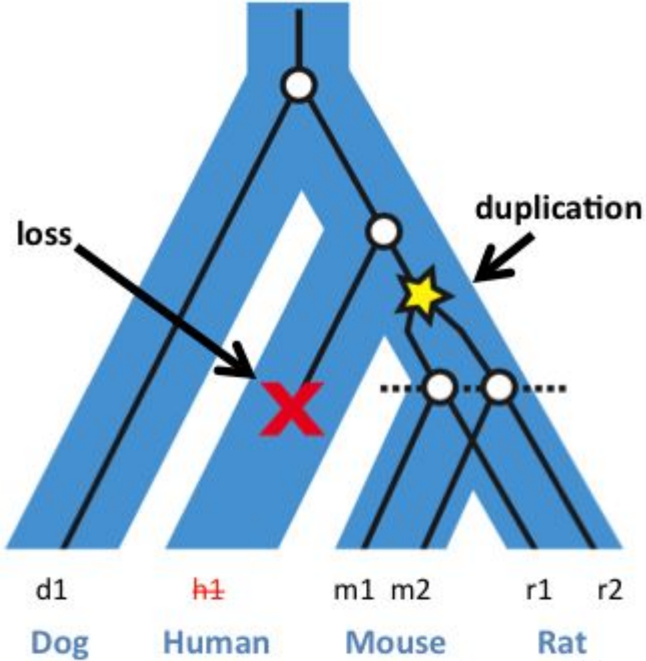
A set of genes with shared ancestry (homologs)

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

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

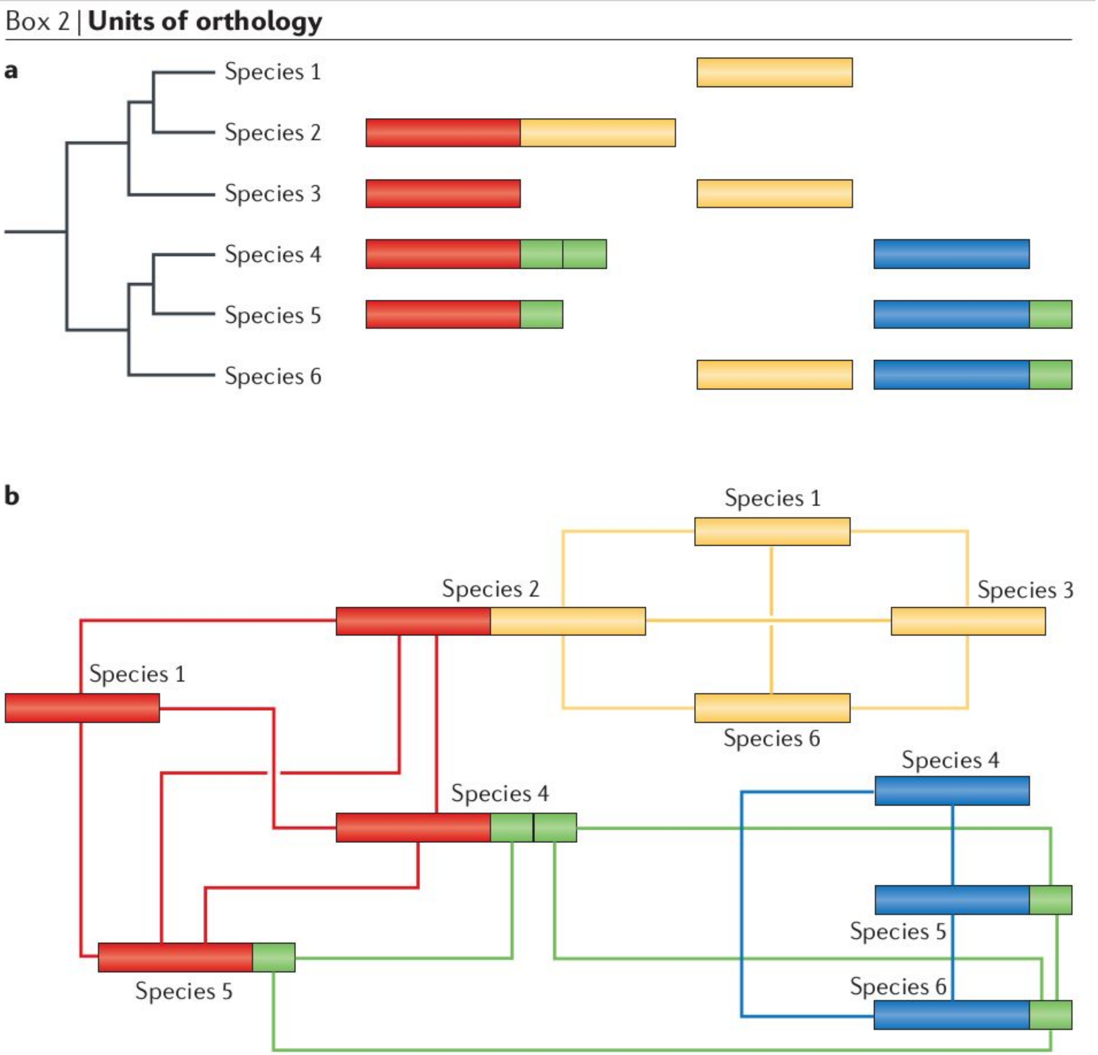
An orthologous group is a (or part of) a gene family

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



But

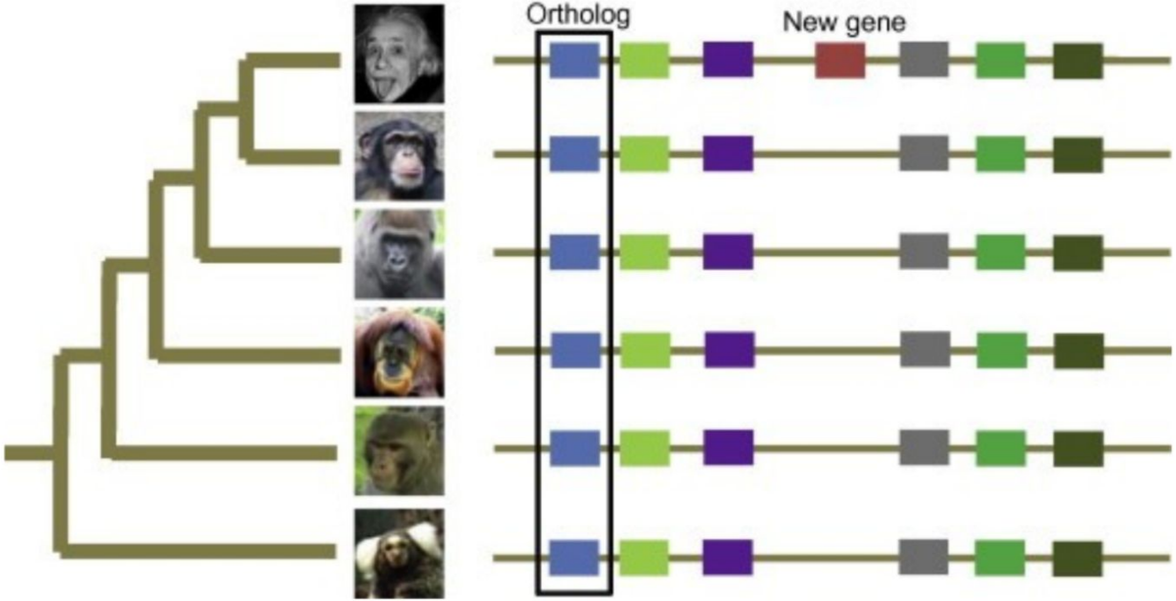
- Genes also evolve by reticulate evolution (HGT and Hybridization)
- Genes also evolve by fusion and fission



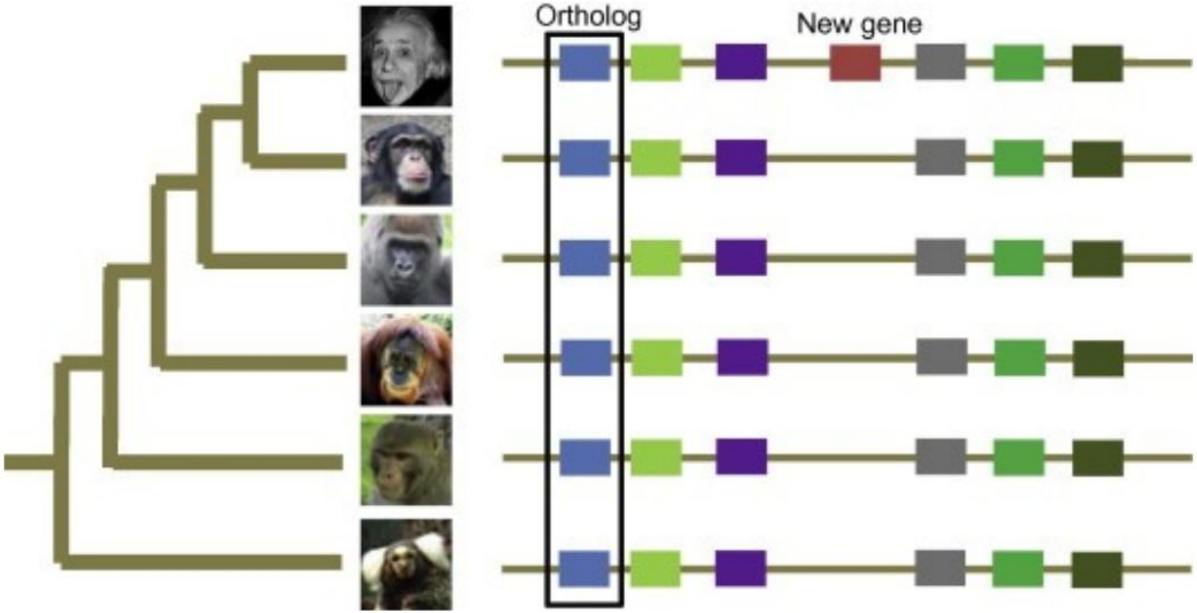
But how they originate in the first place?

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

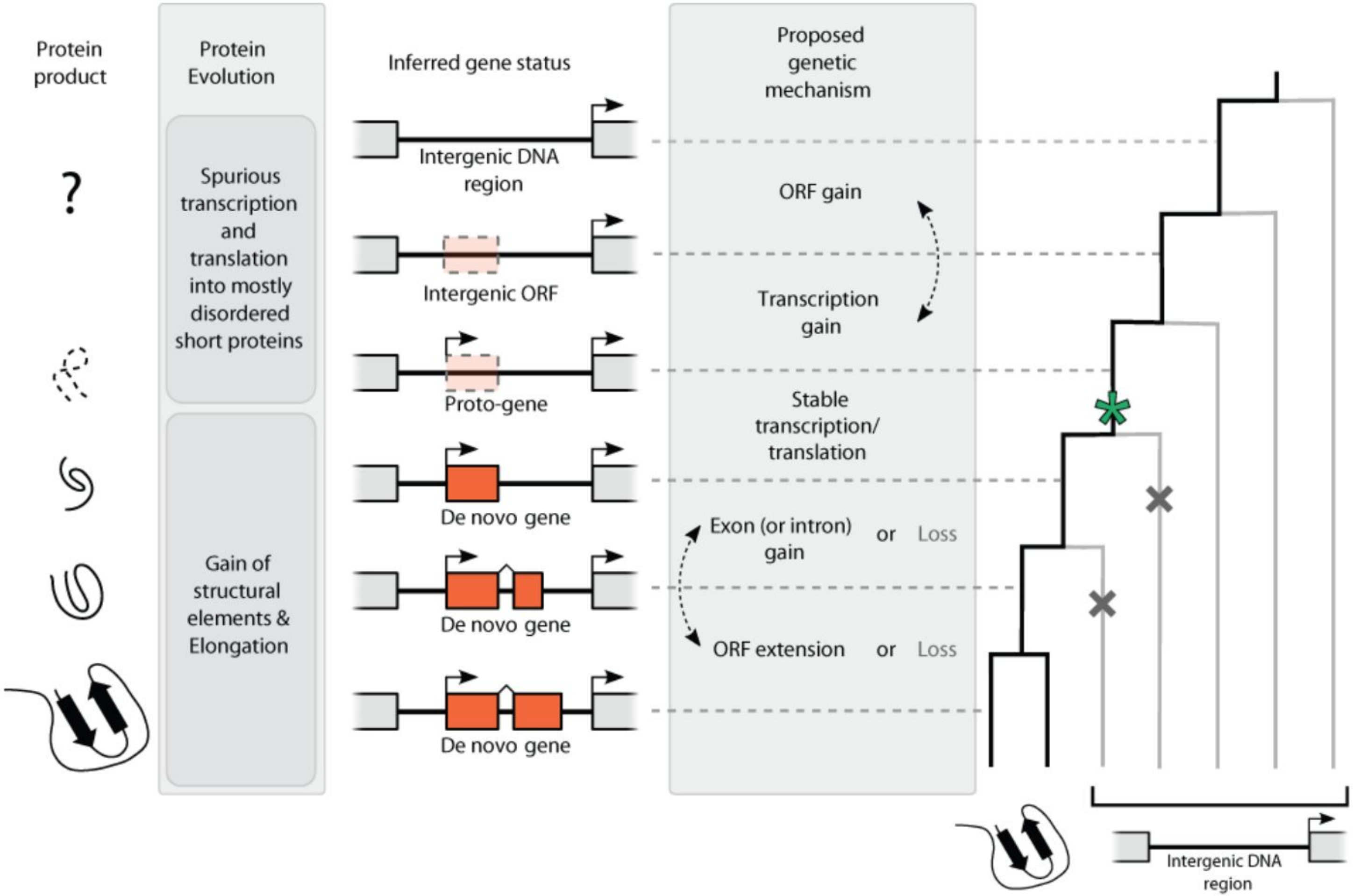
- Spurious predictions?
- Undetected homology?
- Newly emerged gene?



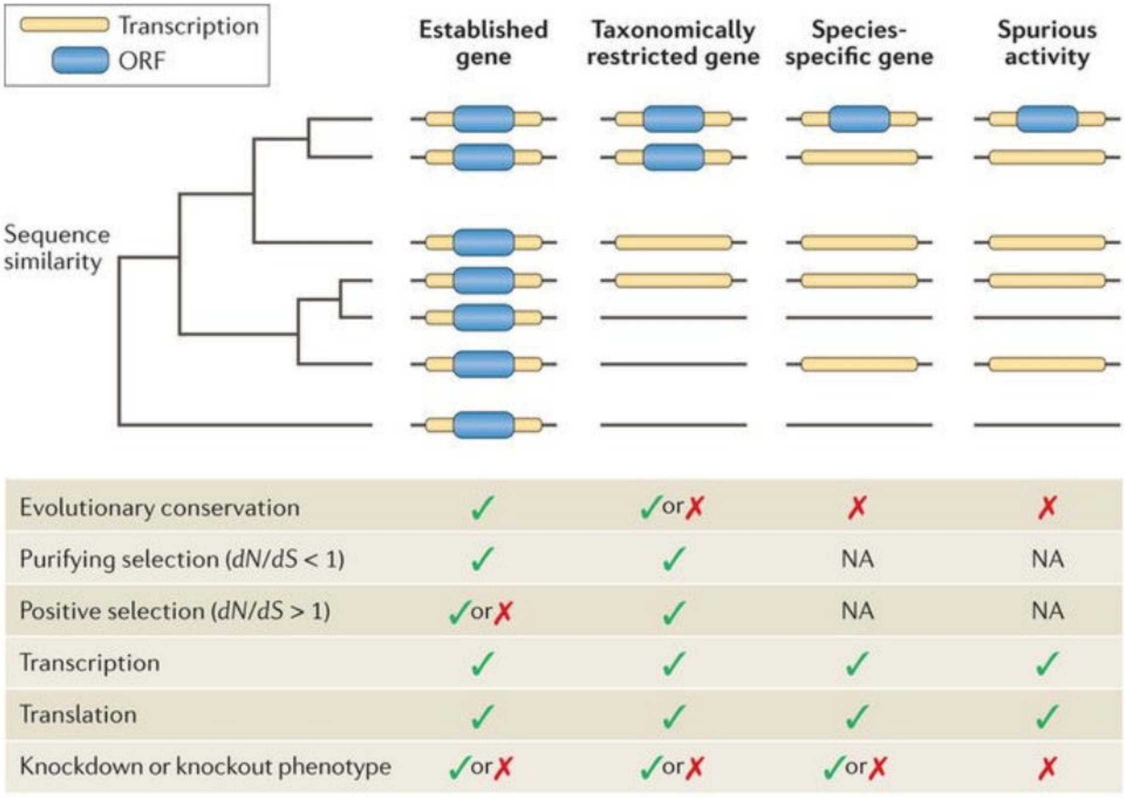
De novo origin of genes.



De novo origin of genes.

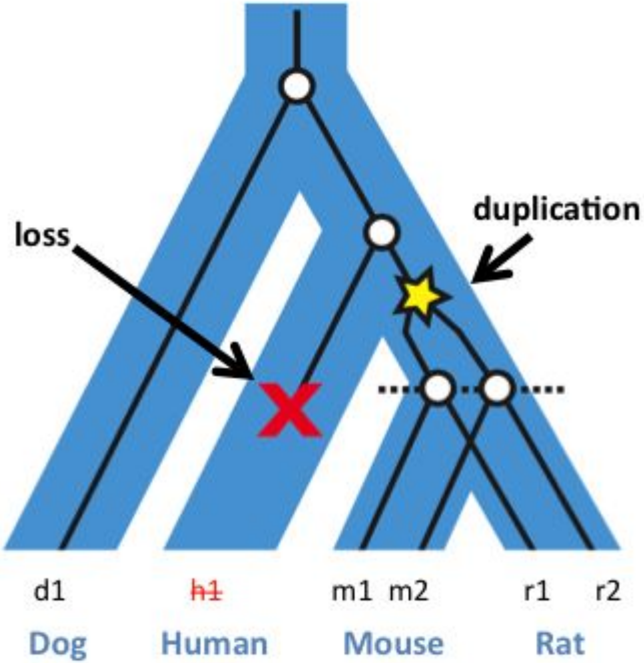


De novo origin of genes.



	Established gene	Taxonomically restricted gene	Species-specific gene	Spurious activity
Evolutionary conservation	✓	✓ or ✗	✗	✗
Purifying selection (*dN*/*dS* < 1)	✓	✓	NA	NA
Positive selection (*dN*/*dS* > 1)	✓ or ✗	✓	NA	NA
Transcription	✓	✓	✓	✓
Translation	✓	✓	✓	✓
Knockdown or knockout phenotype	✓ or ✗	✓ or ✗	✓ or ✗	✗

Nature Reviews | Genetics

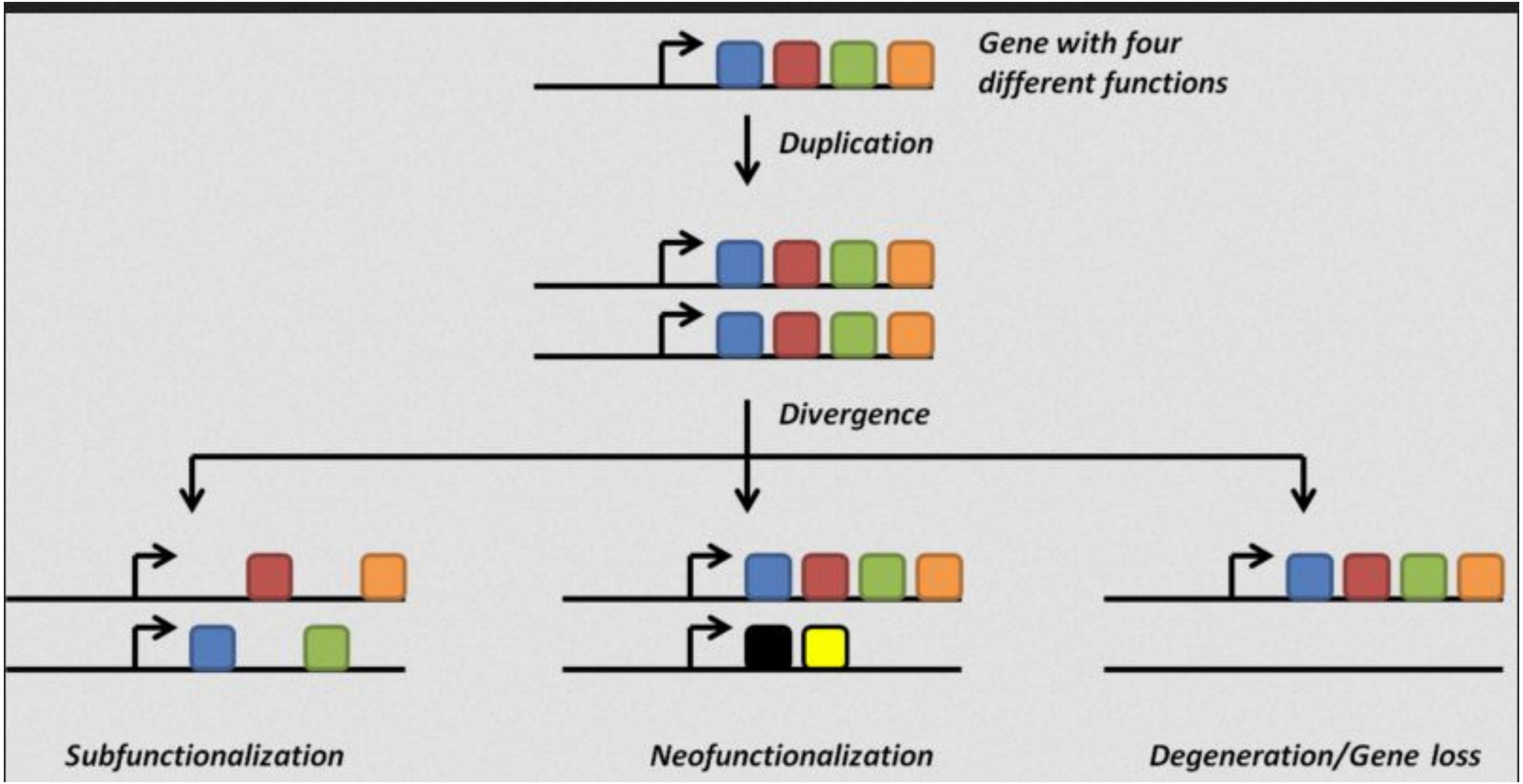

loss
duplication
d1
~~h1~~
m1
m2
r1
r2
Dog
Human
Mouse
Rat

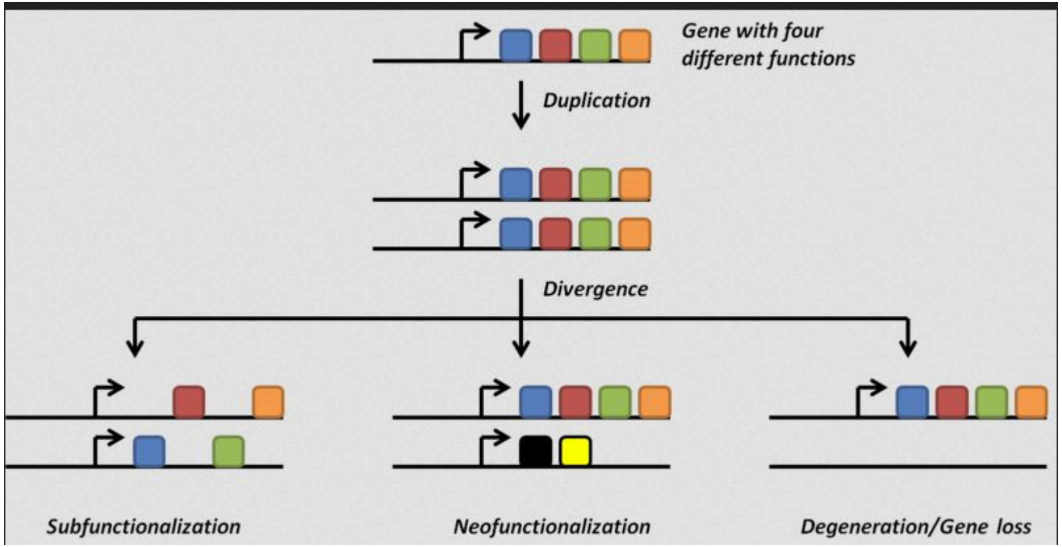
Why genes duplicate?

Spontaneous duplications are common due to:

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

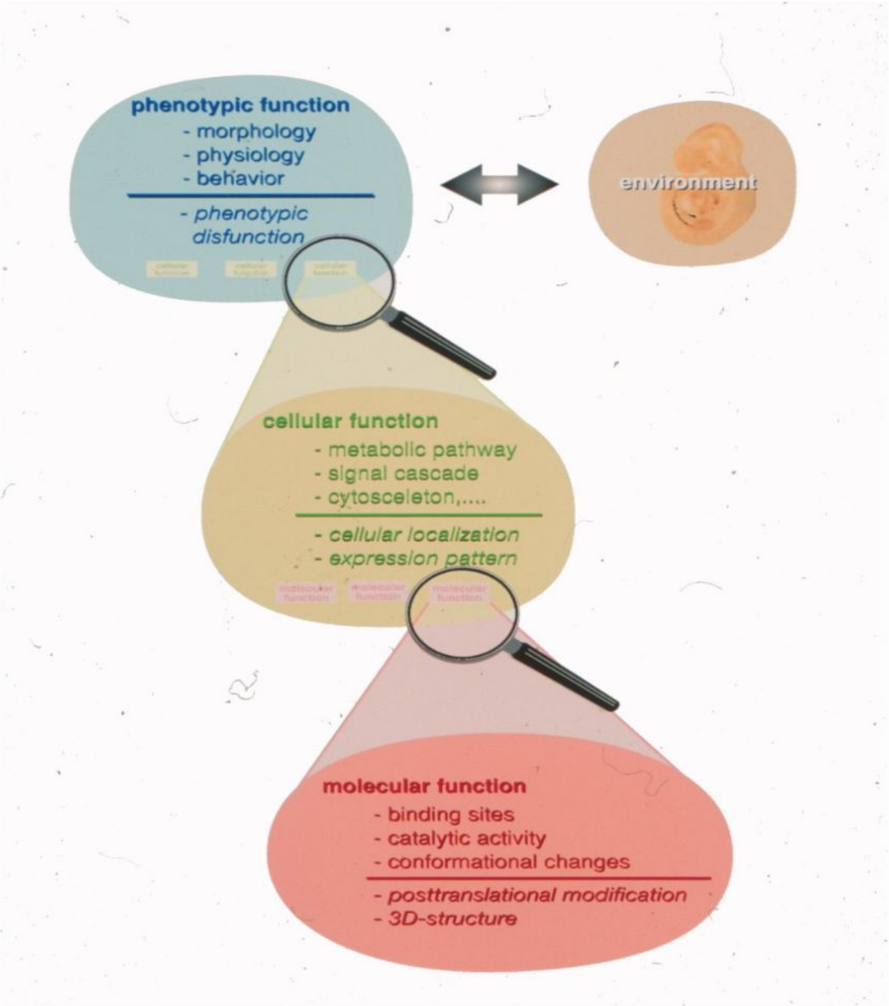
They are common, but the most common outcome of duplication is degeneration (loss) of one of the duplicates


Gene with four different functions
Duplication
Divergence
Subfunctionalization
Neofunctionalization
Degeneration/Gene loss



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

Functional roles of genes.



phenotypic function

- morphology
- physiology
- behavior

- phenotypic disfunction

cellular function

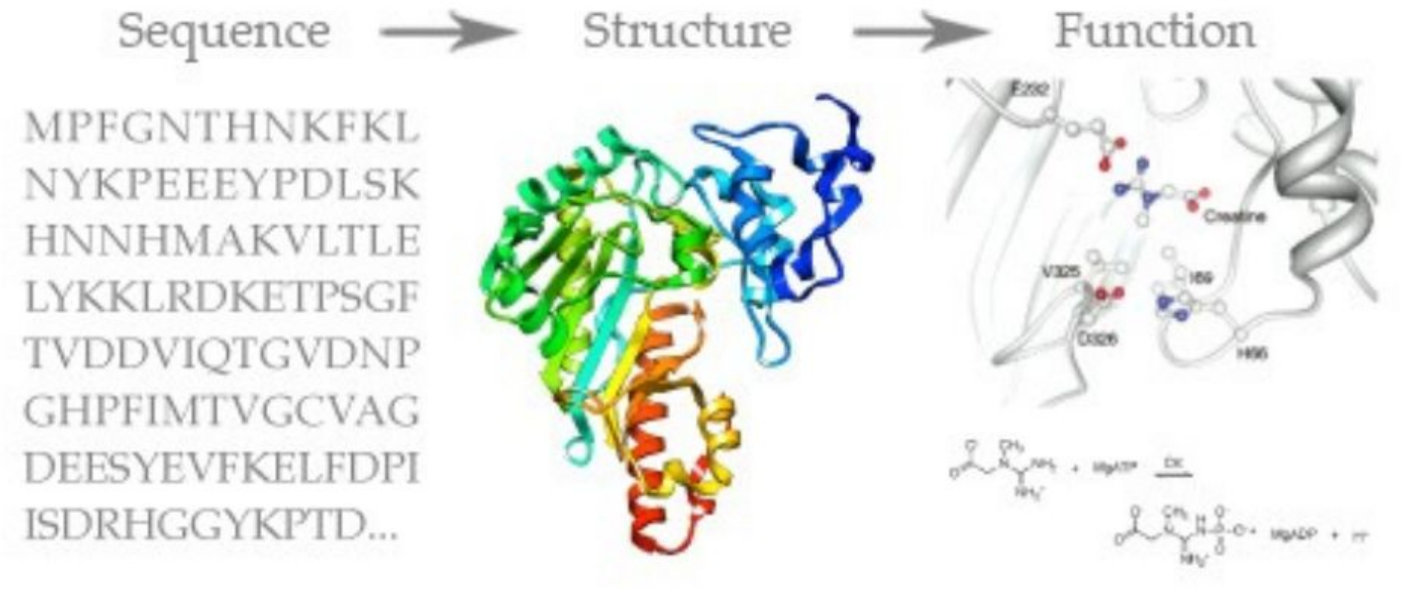
- metabolic pathway
- signal cascade
- cytosceleton,....

- cellular localization
- expression pattern

molecular function

- binding sites
- catalytic activity
- conformational changes

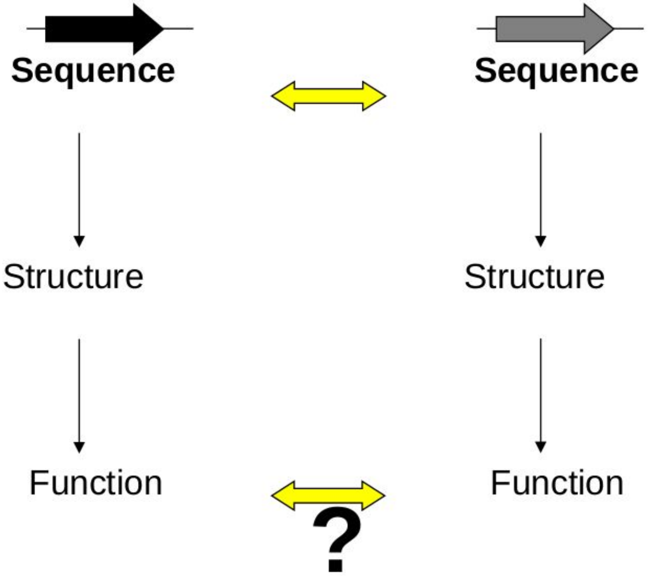
- posttranslational modification
- 3D-structure



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

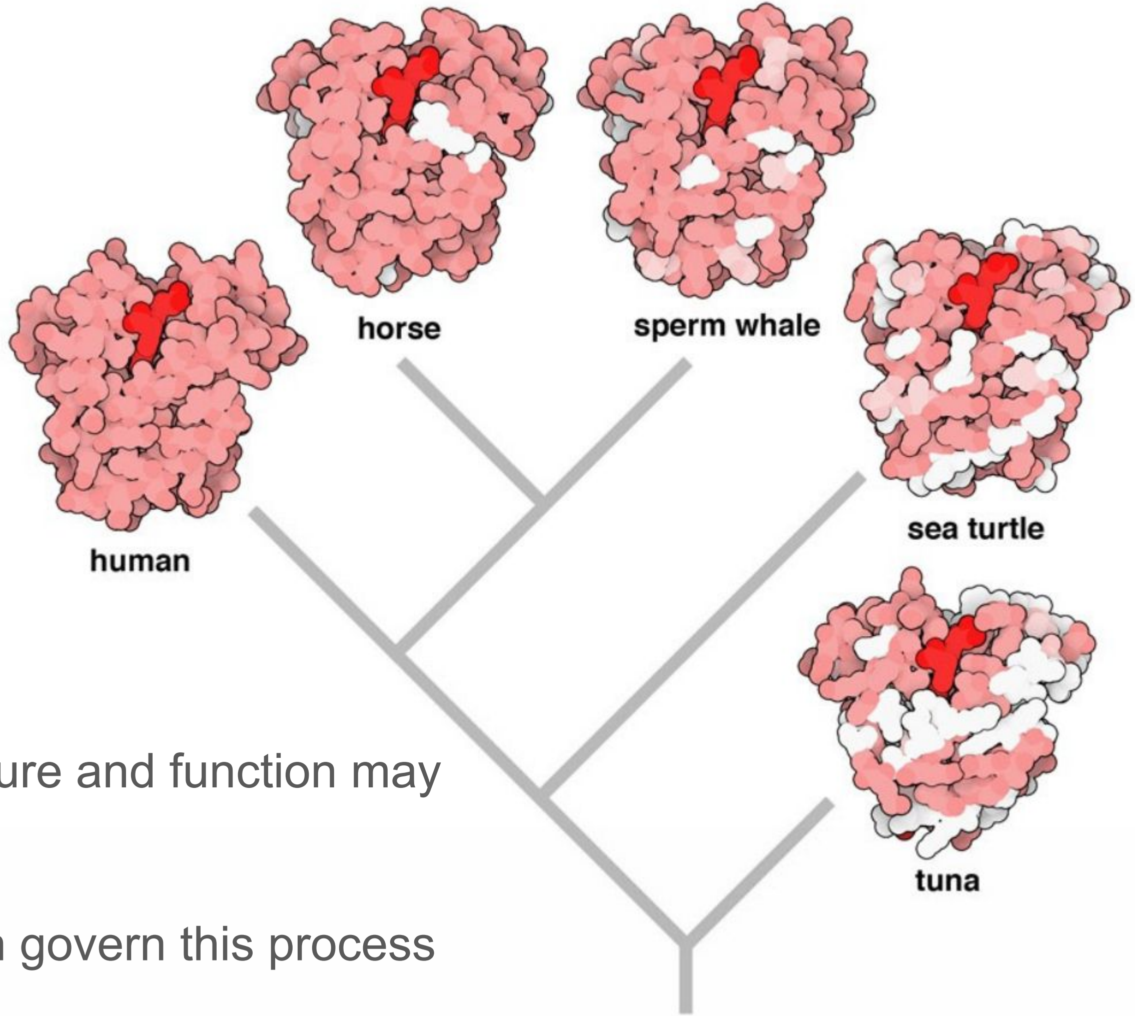
Homology based functional inference.

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



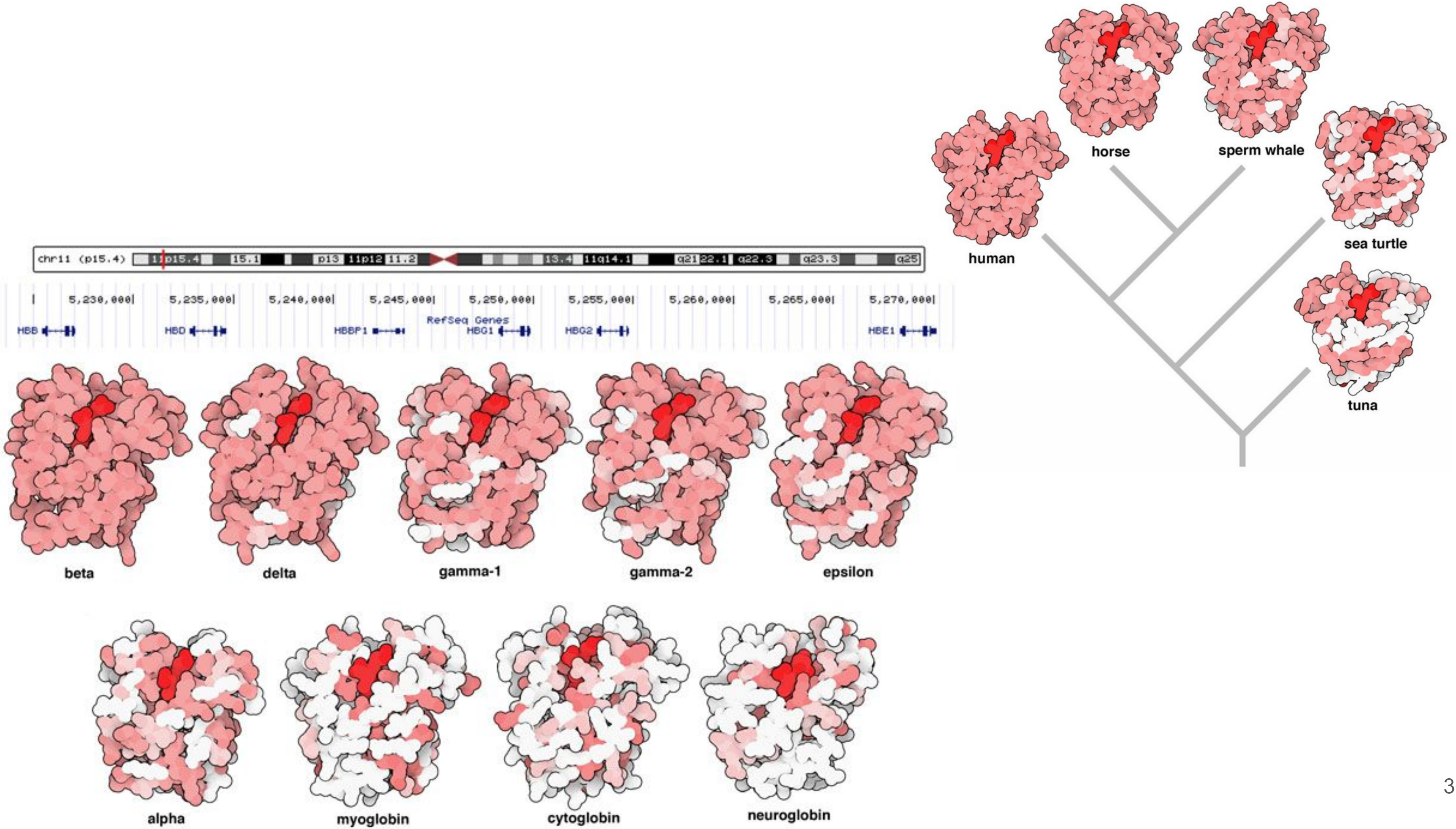
The overwhelming majority of functional annotations are based on this concept

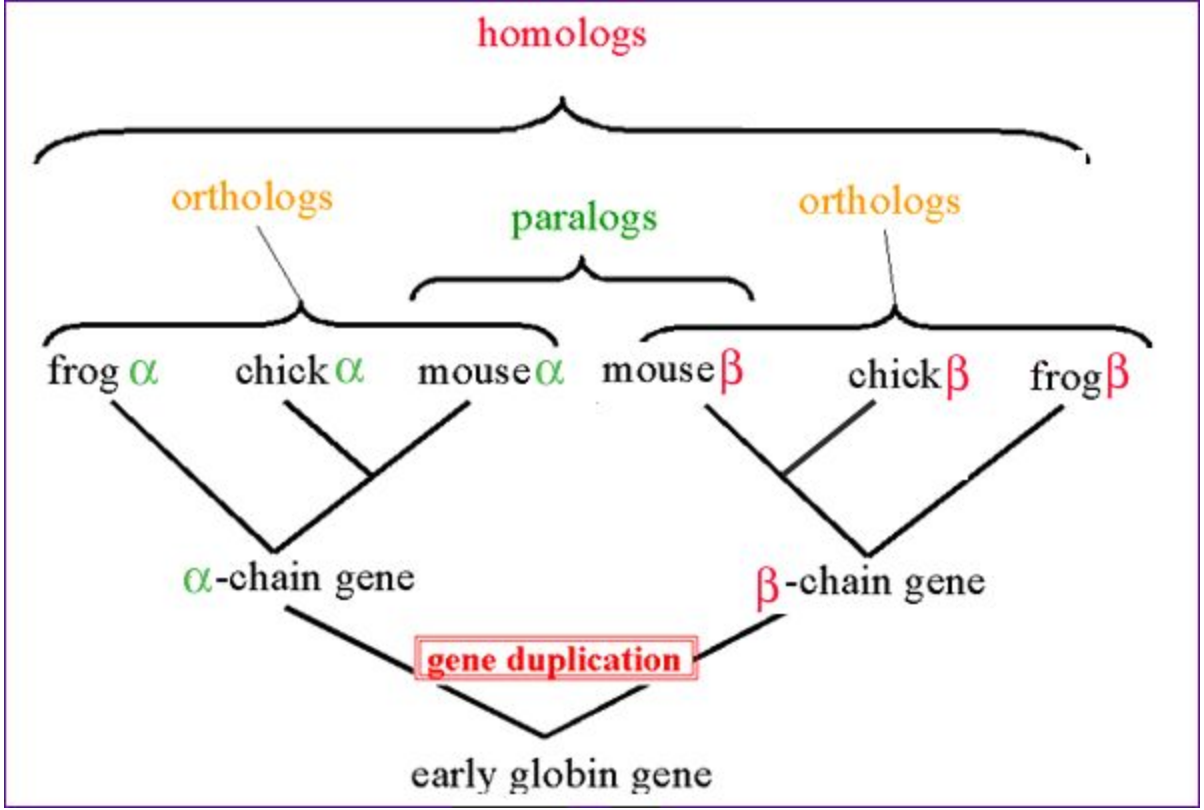
One family one function?



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

Mutation, drift, and selection govern this process

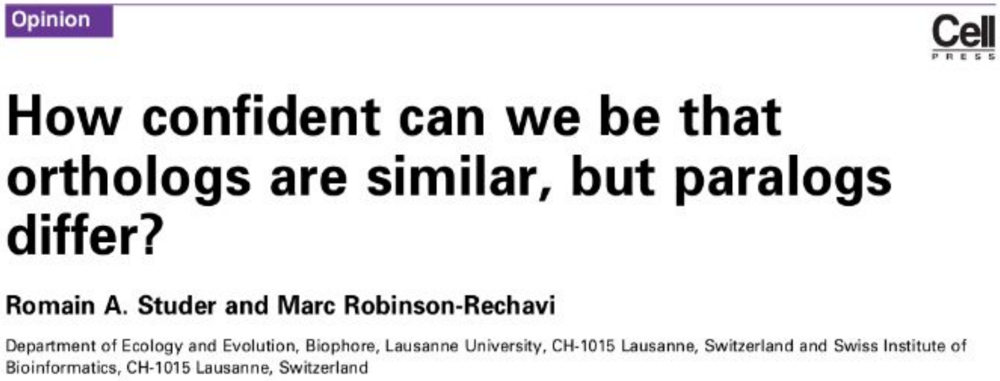

human
horse
sperm whale
sea turtle
tuna
chr11 (p15.4)
11p15.4
15.1
p13
11p12
11.2
13.4
11q14.1
q21
22.1
q22.3
q23.3
q25
5,230,000|
5,235,000|
5,240,000|
5,245,000|
5,250,000|
5,255,000|
5,260,000|
5,265,000|
5,270,000|
RefSeq Genes
HBB
HBD
HBBP1
HBG1
HBG2
HBE1
beta
delta
gamma-1
gamma-2
epsilon
alpha
myoglobin
cytoglobin
neuroglobin



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



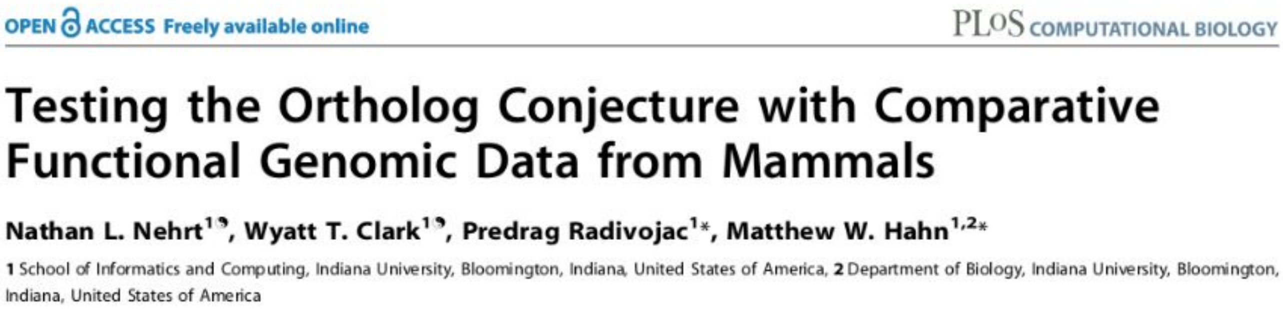
Opinion

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

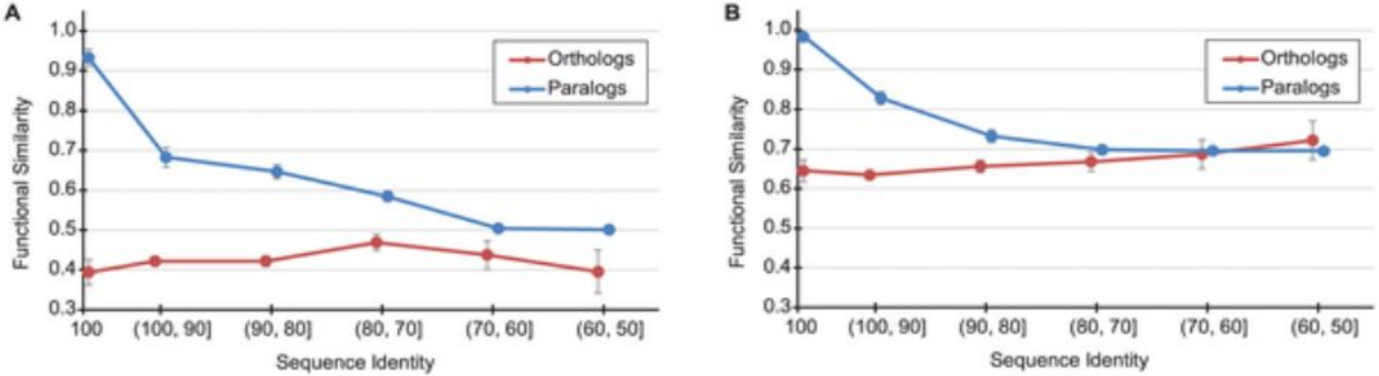
PLoS COMPUTATIONAL BIOLOGY

Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals

Nathan L. Nehrt[1¶], Wyatt T. Clark[1¶], Predrag Radivojac[1*], Matthew W. Hahn[1,2*]

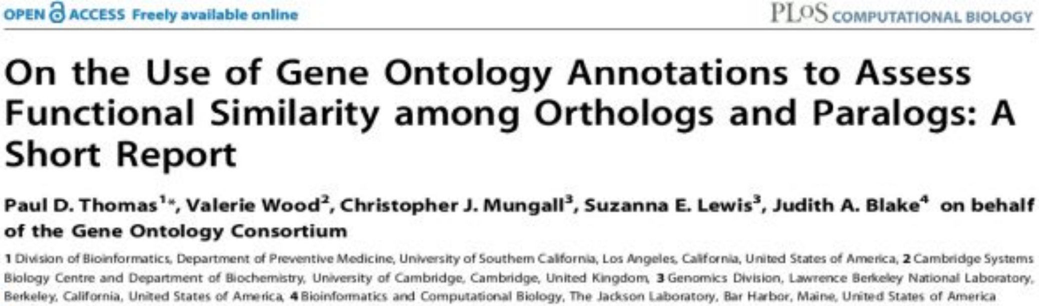
[1] School of Informatics and Computing, Indiana University, Bloomington, Indiana, United States of America, [2] Department of Biology, Indiana University, Bloomington, Indiana, United States of America

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

PLOS | COMPUTATIONAL BIOLOGY



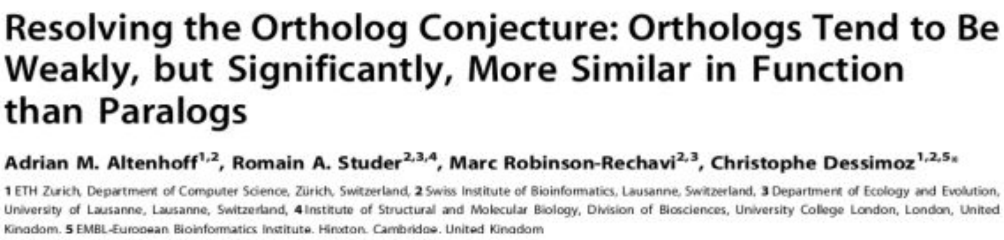
OPEN ACCESS Freely available online PLoS COMPUTATIONAL BIOLOGY

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

Paul D. Thomas[1]*, Valerie Wood[2], Christopher J. Mungall[3], Suzanna E. Lewis[3], Judith A. Blake[4] 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 onlin



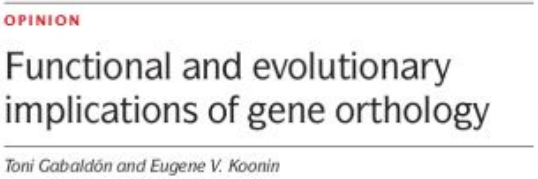
Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs

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, Zürich, 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

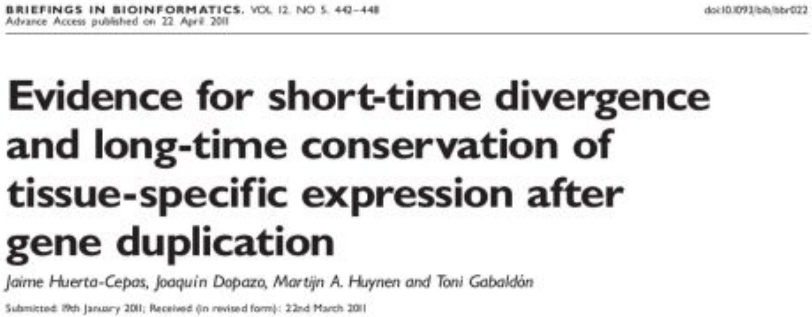
PERSPECTIVES



OPINION

Functional and evolutionary implications of gene orthology

Toni Gabaldón and Eugene V. Koonin



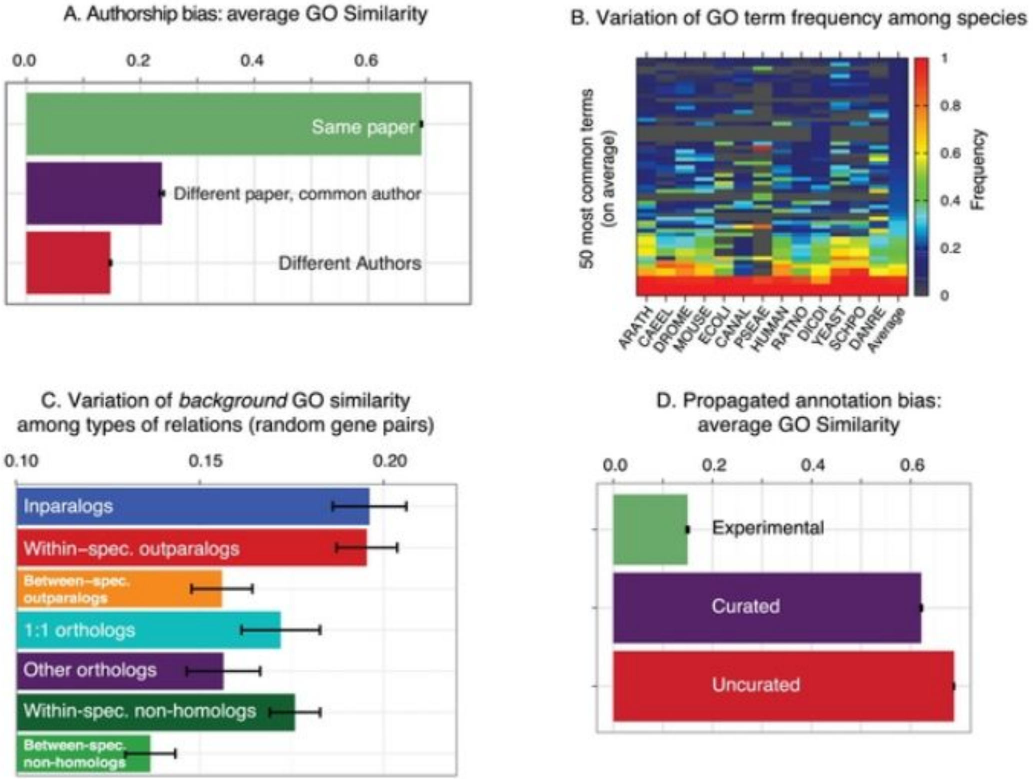
BRIEFINGS IN BIOINFORMATICS. VOL 12. NO 5. 442–448 doi:10.1093/bib/bbr022
Advance Access published on 22 April 2011

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

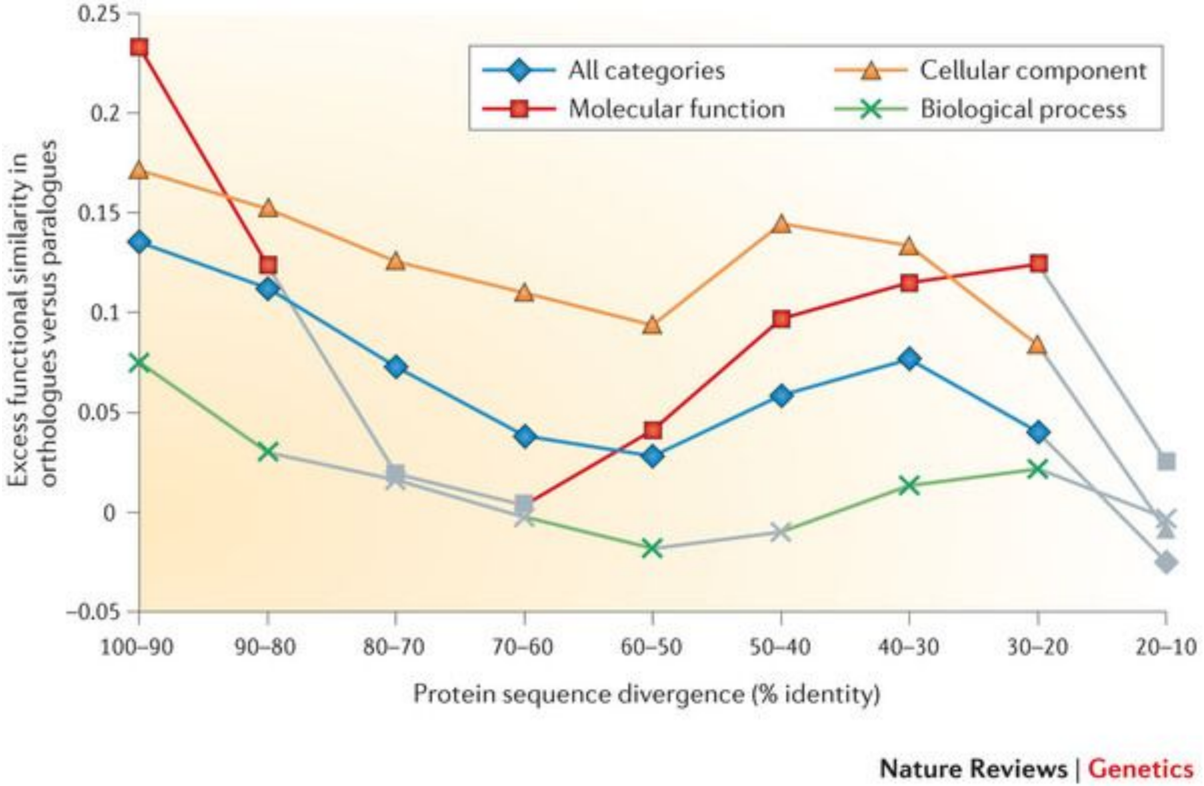
Jaime Huerta-Cepas, Joaquín Dopazo, Martijn A. Huynen and Toni Gabaldón

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

Figure 1. Potential confounding factors in GO analyses.

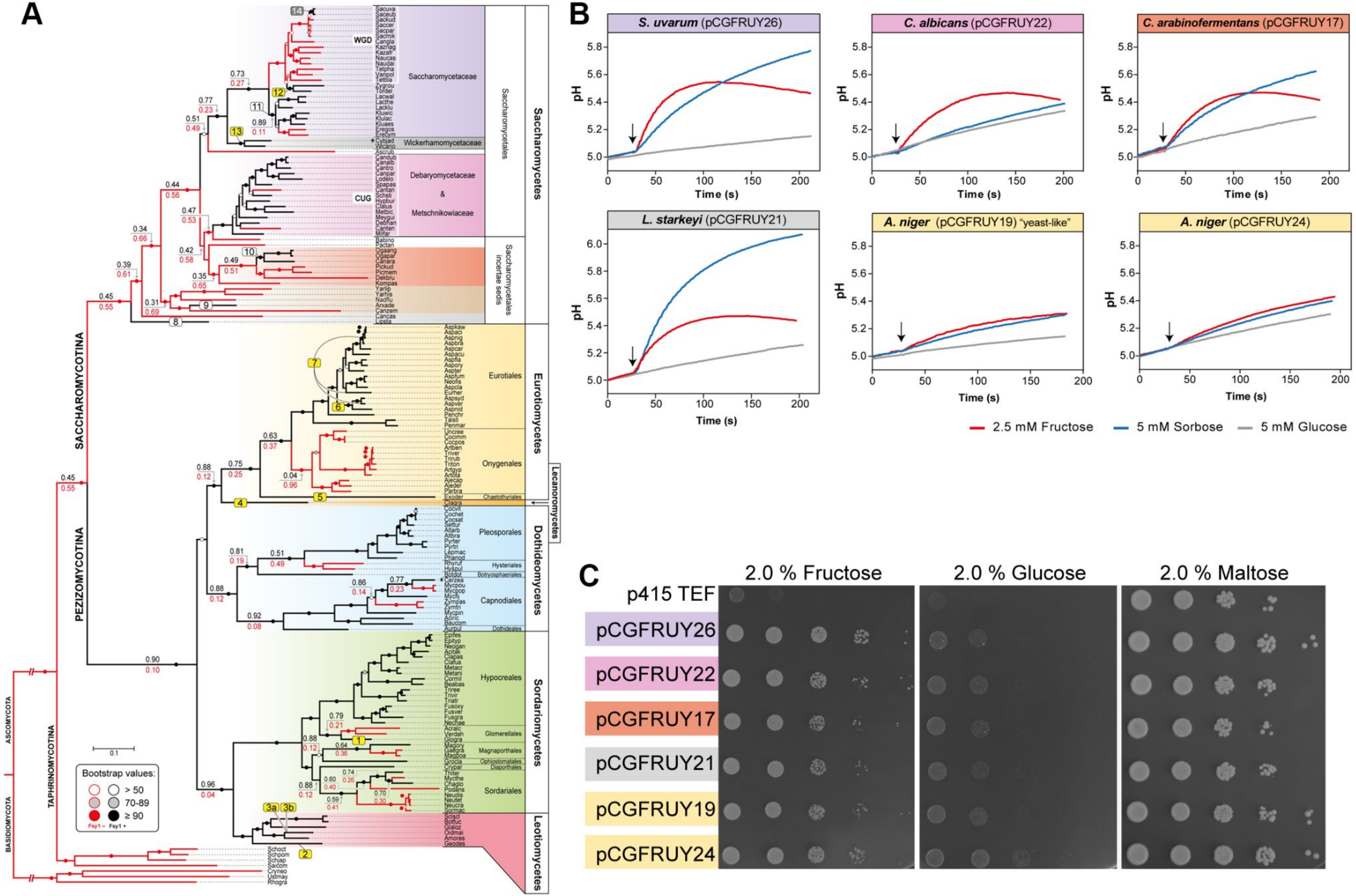


Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C (2012) Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. PLoS Comput Biol 8(5): e1002514. doi:10.1371/journal.pcbi.1002514
http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1002514



Gabaldón, Koonin (2013)

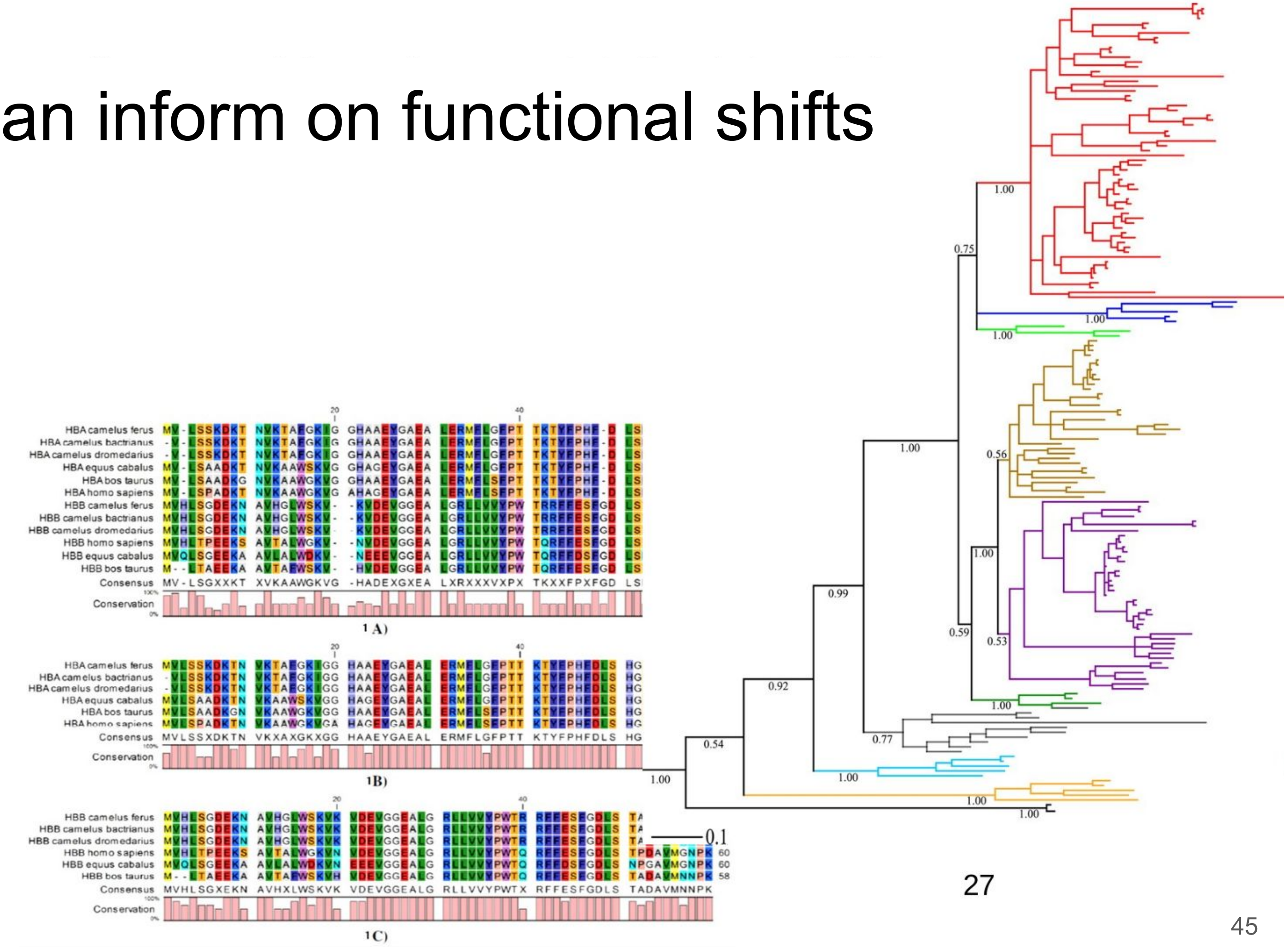
Functional divergence through speciation



Conclusions

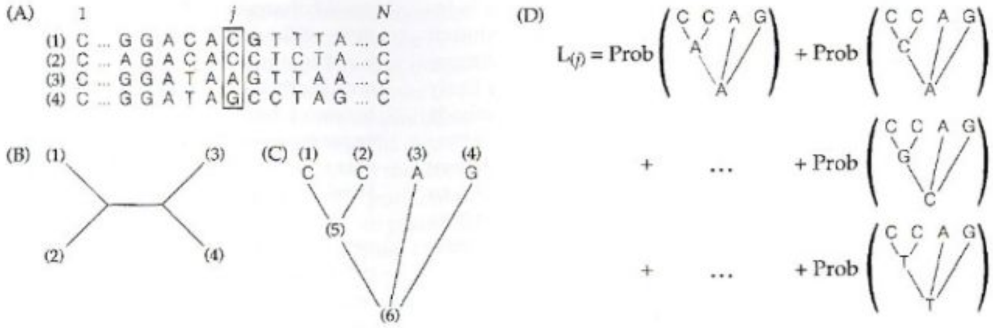
- Orthologs (slightly?) more likely that paralogs to share function
- One function per gene family?: not totally, variation over a common theme (e.g. transporter with different substrate affinities)
- Broadly defined functions probably conserved, specific functions more variable.

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

computation in a real problem


(A)
1
j
N
(1) C ... G G A C A C G T T T A ... C
(2) C ... A G A C A C C T C T A ... C
(3) C ... G G A T A A G T T A A ... C
(4) C ... G G A T A G C C T A G ... C
(B)
(1)
(2)
(3)
(4)
(C)
(1)
(2)
(3)
(4)
C
C
A
G
(5)
(6)
(D)
L(j) = Prob
C C A G
A
A
+ Prob
C C A G
C
A
+
...
+ Prob
C C A G
G
C
+
...
+ Prob
C C A G
T
T

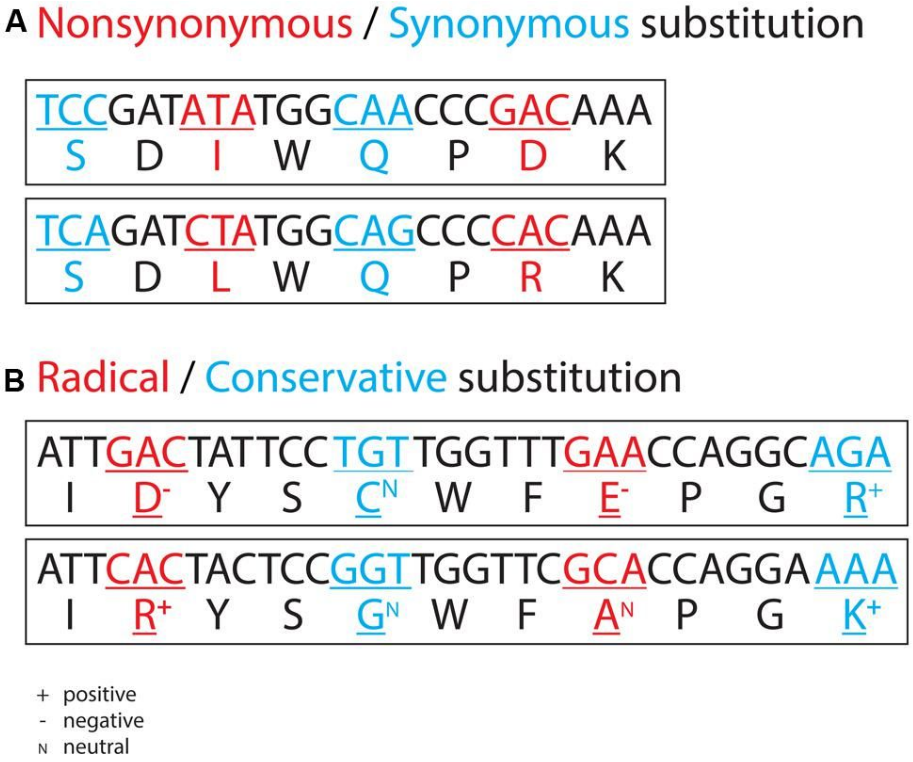
- Tree after rooting in an arbitrary node (reversible model).
- The likelihood for a particular site is the sum of the probabilities of every possible reconstruction of ancestral states given some model of base substitution.
- The likelihood of the tree is the product of the likelihood at each site.

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

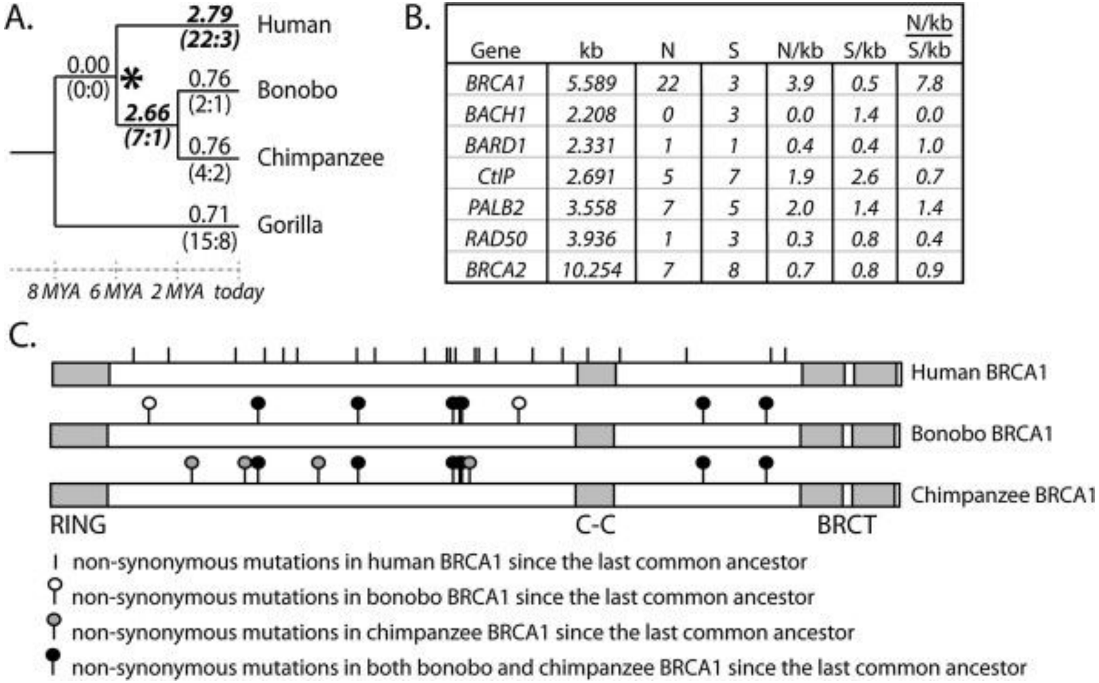
- The likelihood is reported as the sum of the log likelihhod of the full tree.

$$lnL = lnL_{(1)} + lnL_{(2)} + ... + lnL_{(N)} = \sum_{j=1}^{N} lnL_{(j)}$$

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

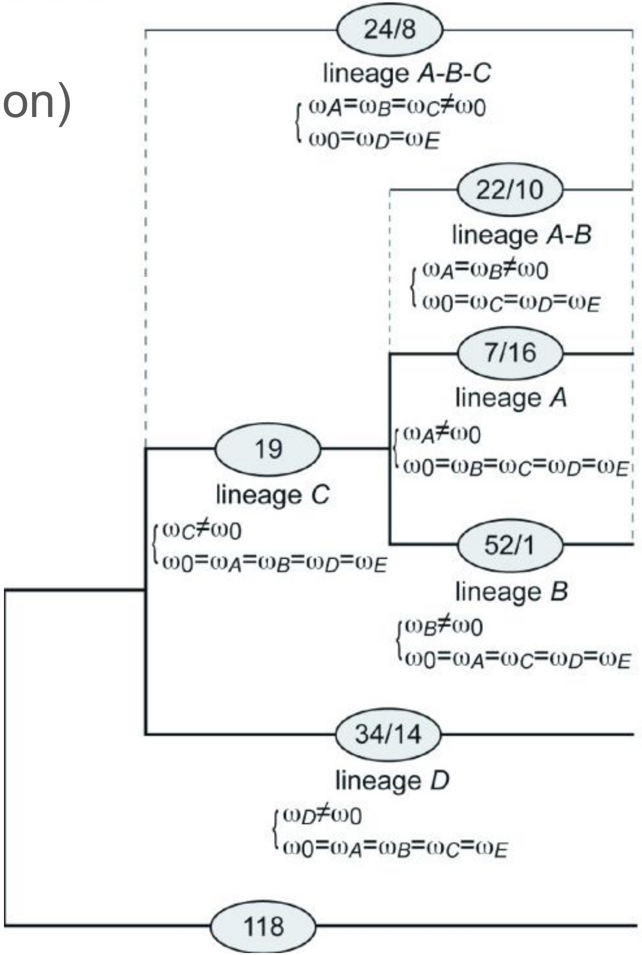

A Nonsynonymous / Synonymous substitution
TCCGATATATGGCAACCCGACAAA
S D I W Q P D K
TCAGATCTATGGCAGCCCCACAAA
S D L W Q P R K
B Radical / Conservative substitution
ATTGACTATTCCTGTTGGTTTGAACCAGGCAGA
I D- Y S CN W F E- P G R+
ATTCACTACTCCGGTTGGTTCGCACCAGGAAAA
I R+ Y S GN W F AN P G K+
+ positive
- negative
N neutral

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


A.
2.79
(22:3)
Human
0.00
(0:0)
*
0.76
(2:1)
Bonobo
2.66
(7:1)
0.76
(4:2)
Chimpanzee
0.71
(15:8)
Gorilla
8 MYA 6 MYA 2 MYA today
B.
C.
Human BRCA1
Bonobo BRCA1
Chimpanzee BRCA1
RING
C-C
BRCT
non-synonymous mutations in human BRCA1 since the last common ancestor
non-synonymous mutations in bonobo BRCA1 since the last common ancestor
non-synonymous mutations in chimpanzee BRCA1 since the last common ancestor
non-synonymous mutations in both bonobo and chimpanzee BRCA1 since the last common ancestor

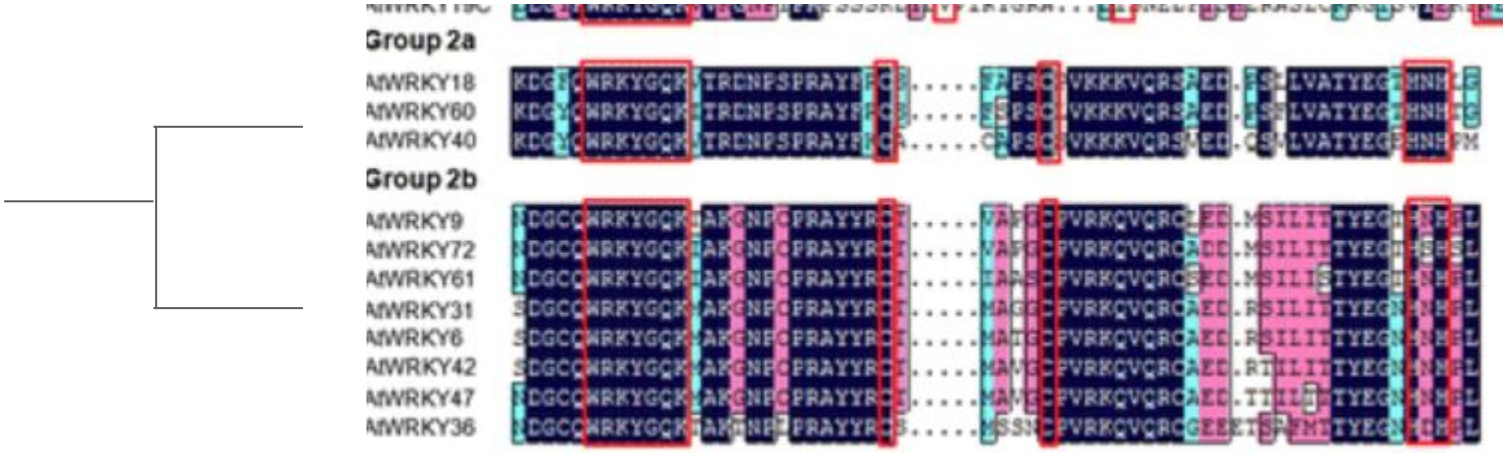
Gene	kb	N	S	N/kb	S/kb	N/kb / S/kb
BRCA1	5.589	22	3	3.9	0.5	7.8
BACH1	2.208	0	3	0.0	1.4	0.0
BARD1	2.331	1	1	0.4	0.4	1.0
CtIP	2.691	5	7	1.9	2.6	0.7
PALB2	3.558	7	5	2.0	1.4	1.4
RAD50	3.936	1	3	0.3	0.8	0.4
BRCA2	10.254	7	8	0.7	0.8	0.9

Figure 2.



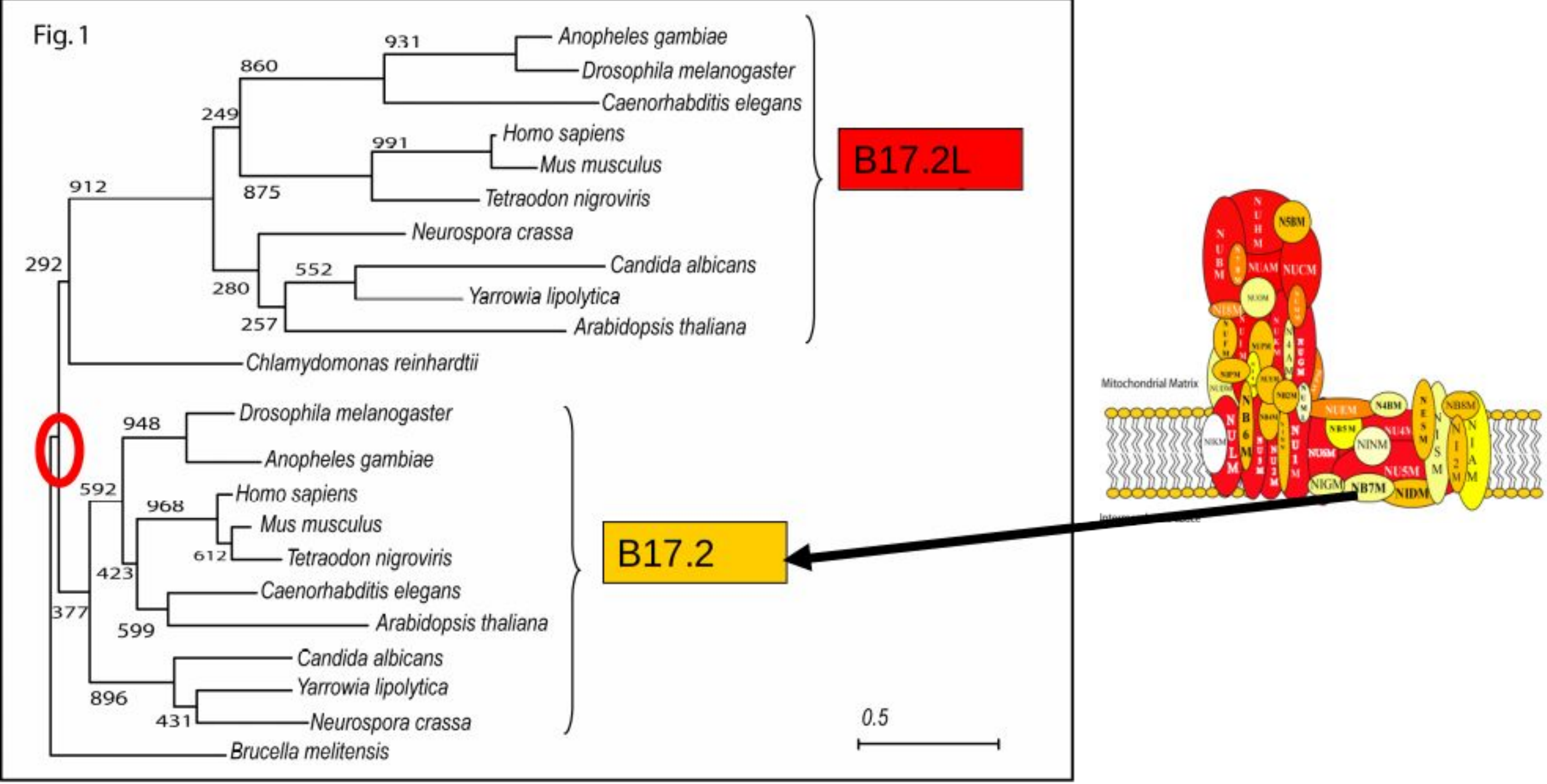
Can we predict change of function?

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

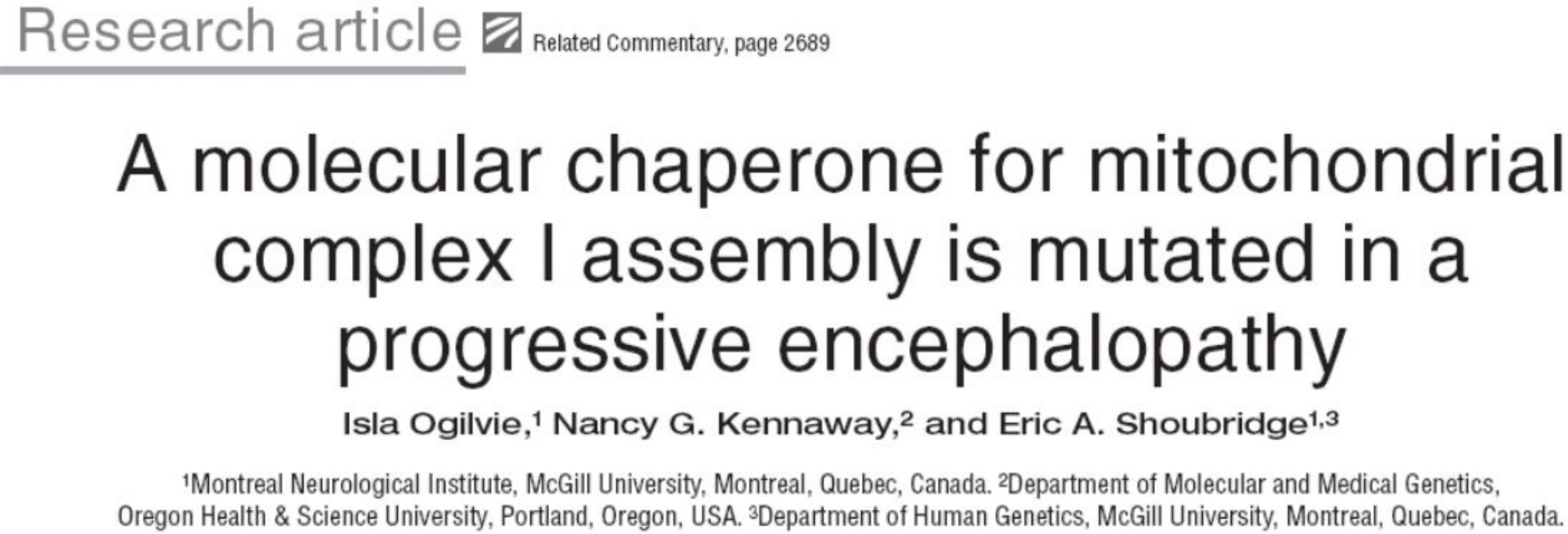


Probably group 2a and group 2b, perform different functions

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)



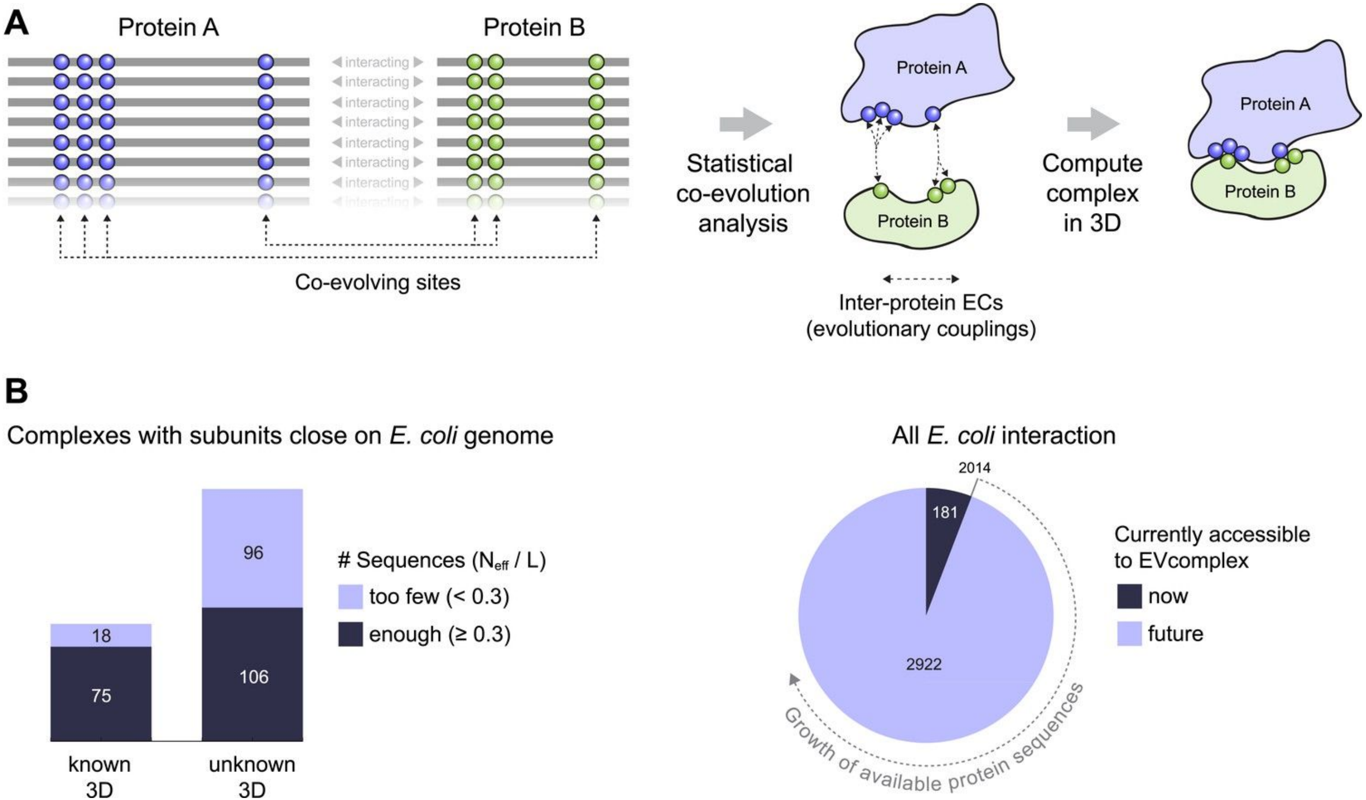
Research article Related Commentary, page 2689

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

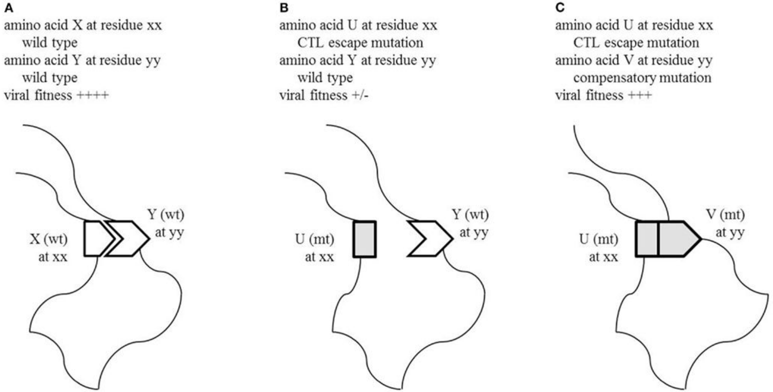
Isla Ogilvie,[1] Nancy G. Kennaway,[2] and Eric A. Shoubridge[1,3]

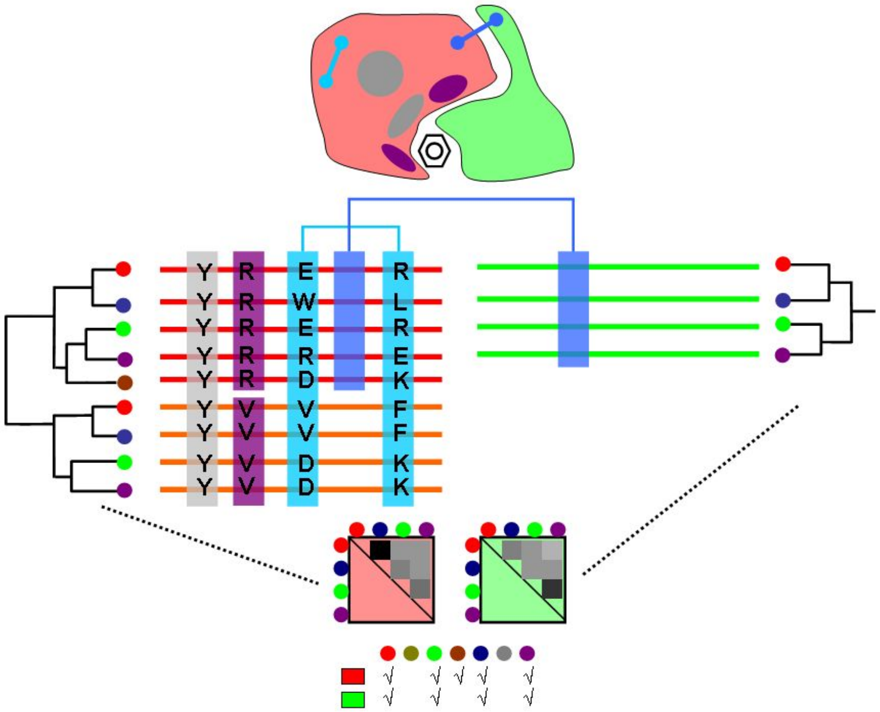
[1]Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada. [2]Department of Molecular and Medical Genetics, Oregon Health & Science University, Portland, Oregon, USA. [3]Department of Human Genetics, McGill University, Montreal, Quebec, Canada.

You can also model co-evolution between sequences

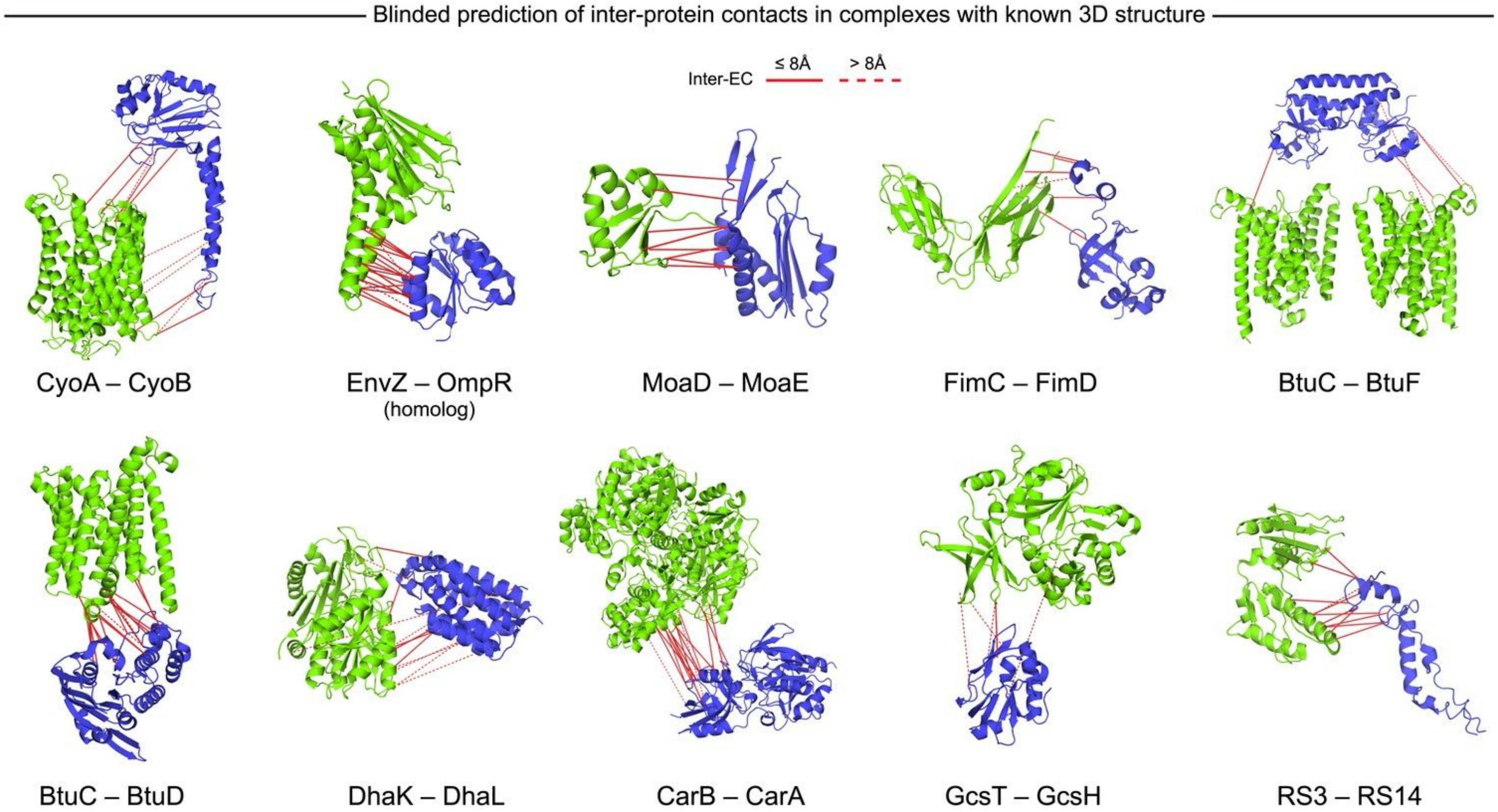


You can also model co-evolution between sequences

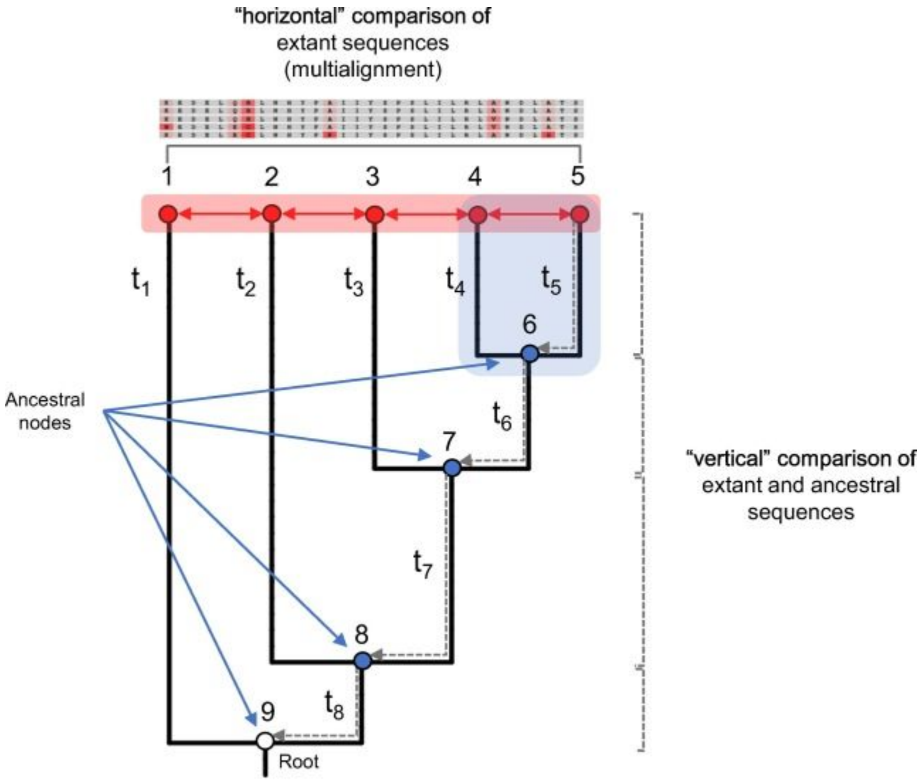


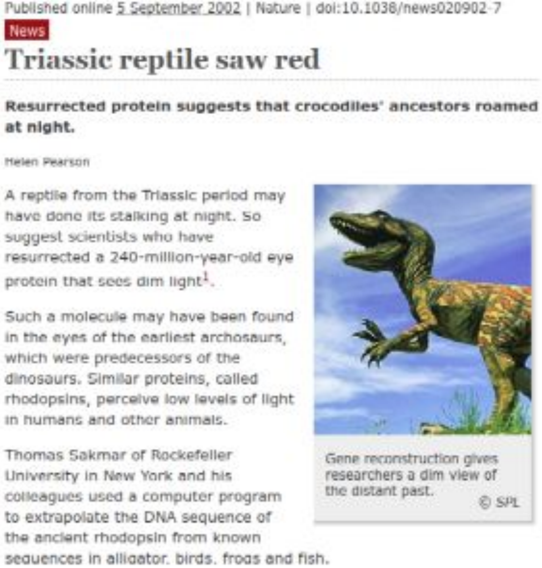


You can also model co-evolution between sequences



You can even reconstruct ancestral sequences





Published online 5 September 2002 | Nature | doi:10.1038/news020902-7

News

Triassic reptile saw red

Resurrected protein suggests that crocodiles' ancestors roamed at night.

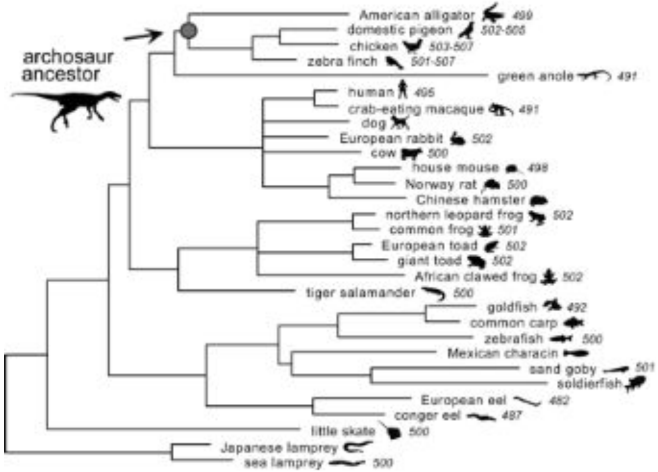
Helen Pearson

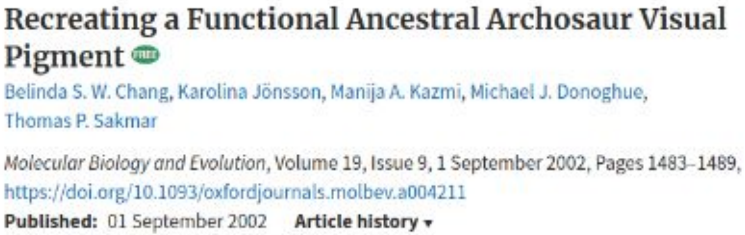
A reptile from the Triassic period may have done its stalking at night. So suggest scientists who have resurrected a 240-million-year-old eye protein that sees dim light[1].

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



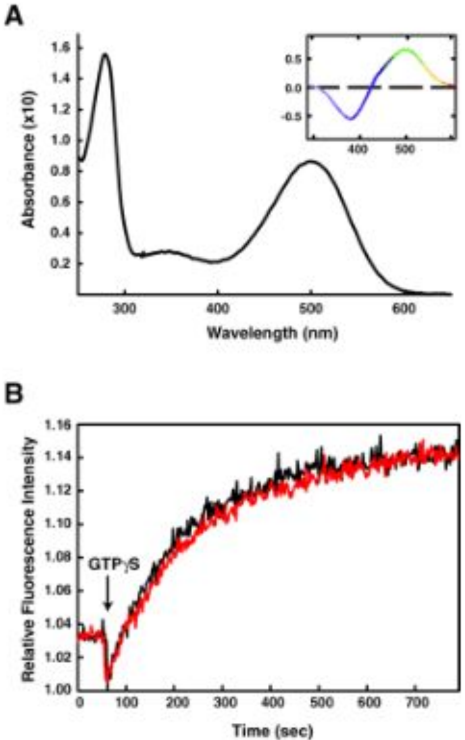


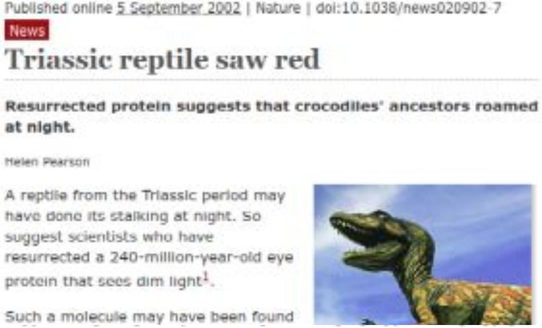
Recreating a Functional Ancestral Archosaur Visual Pigment

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 ▾**





Published online 5 September 2002 | Nature | doi:10.1038/news020902-7

News

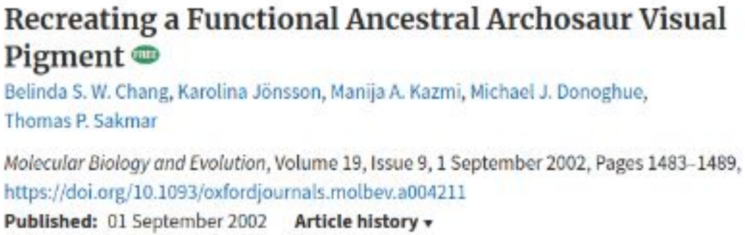
Triassic reptile saw red

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 light[1].

Such a molecule may have been found

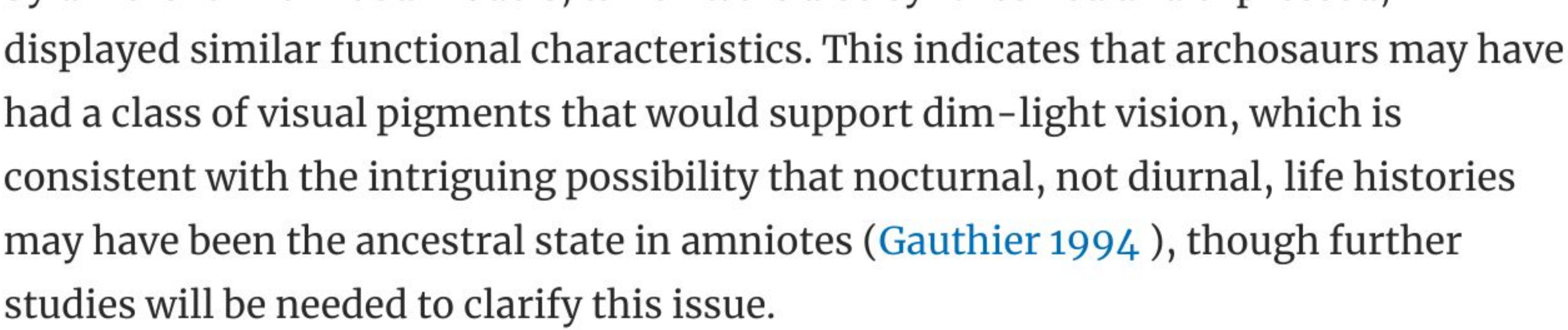


Recreating a Functional Ancestral Archosaur Visual Pigment

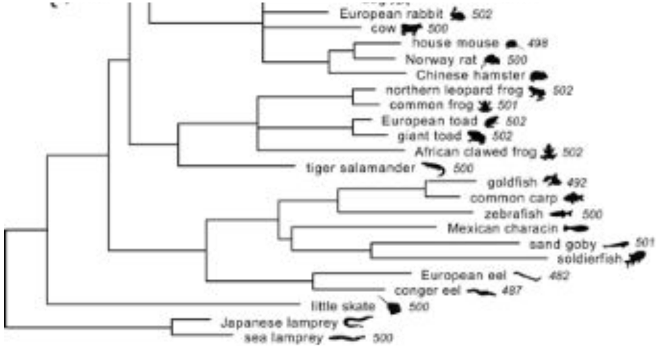
Belinda S. W. Chang, Karolina Jönsson, Manija A. Kazmi, Michael J. Donoghue, Thomas P. Sakmar

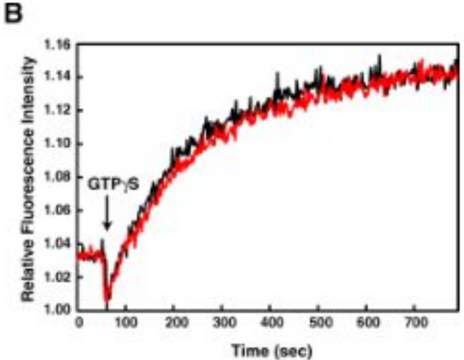
Molecular Biology and Evolution, Volume 19, Issue 9, 1 September 2002, Pages 1483–1489, https://doi.org/10.1093/oxfordjournals.molbev.a004211

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 .

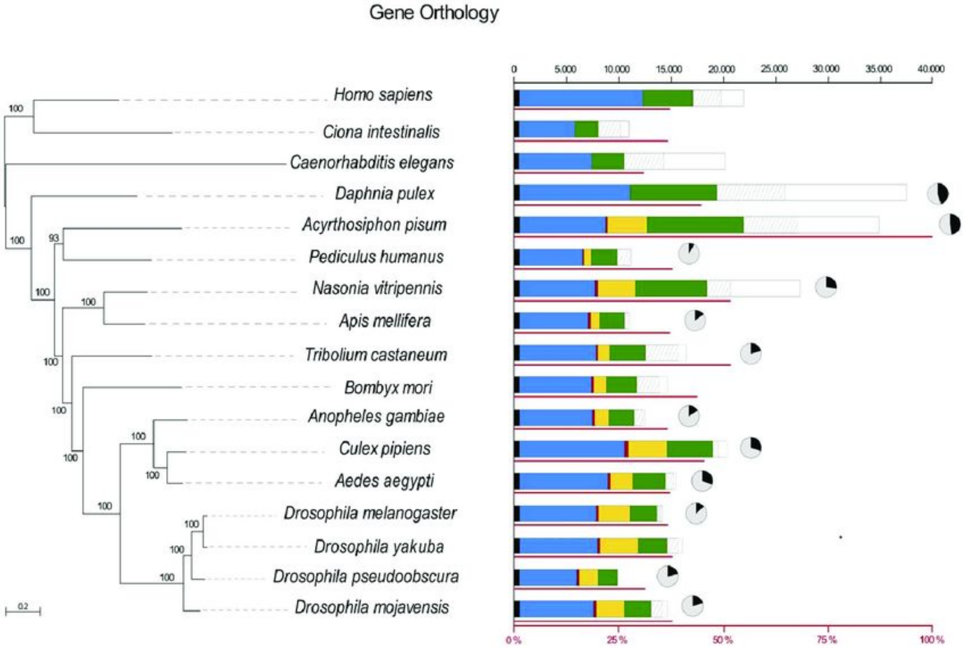
Break?

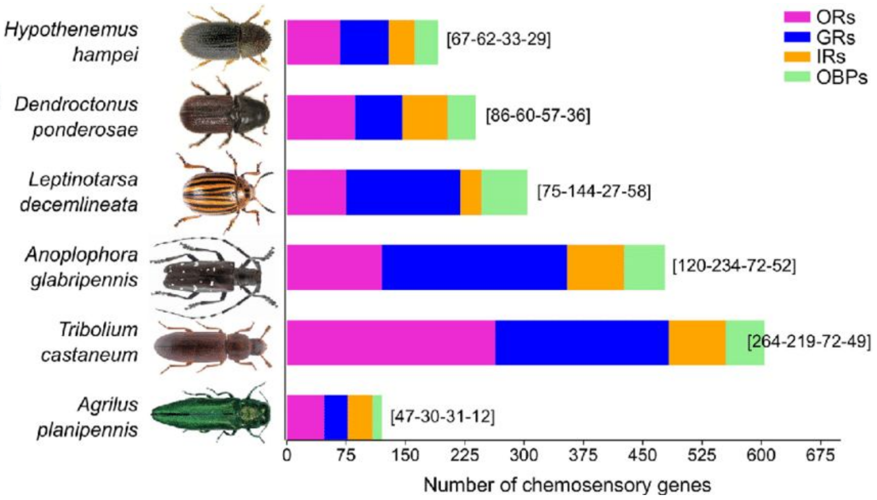
How to study gene family evolution at genomic scales?

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:

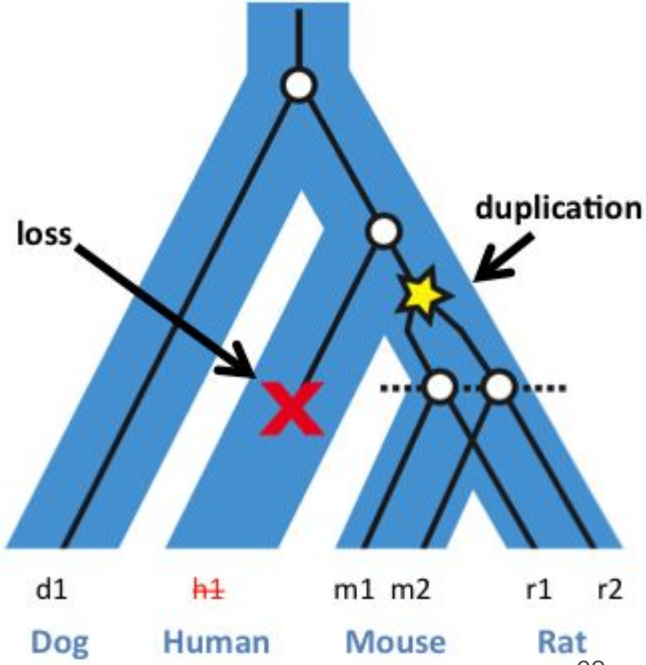
A set of genes with shared ancestry (homologs)

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

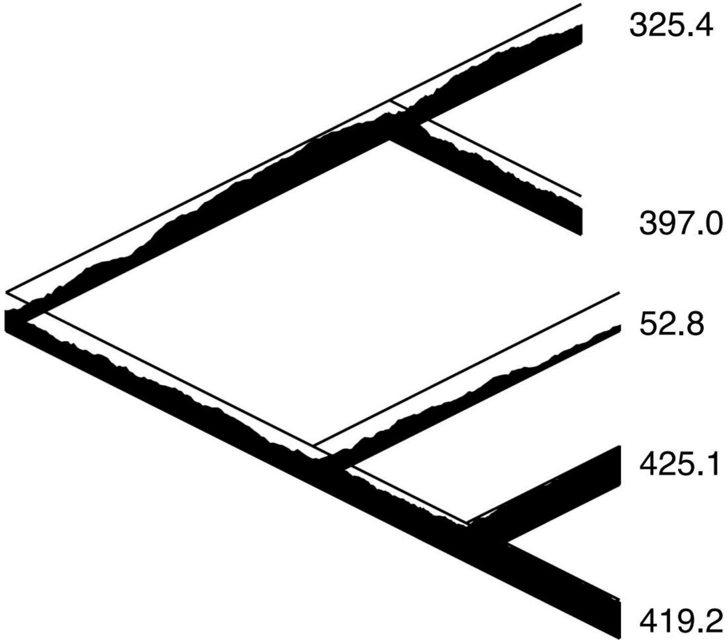
Members of a gene family can be orthologs or paralogs between them

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

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

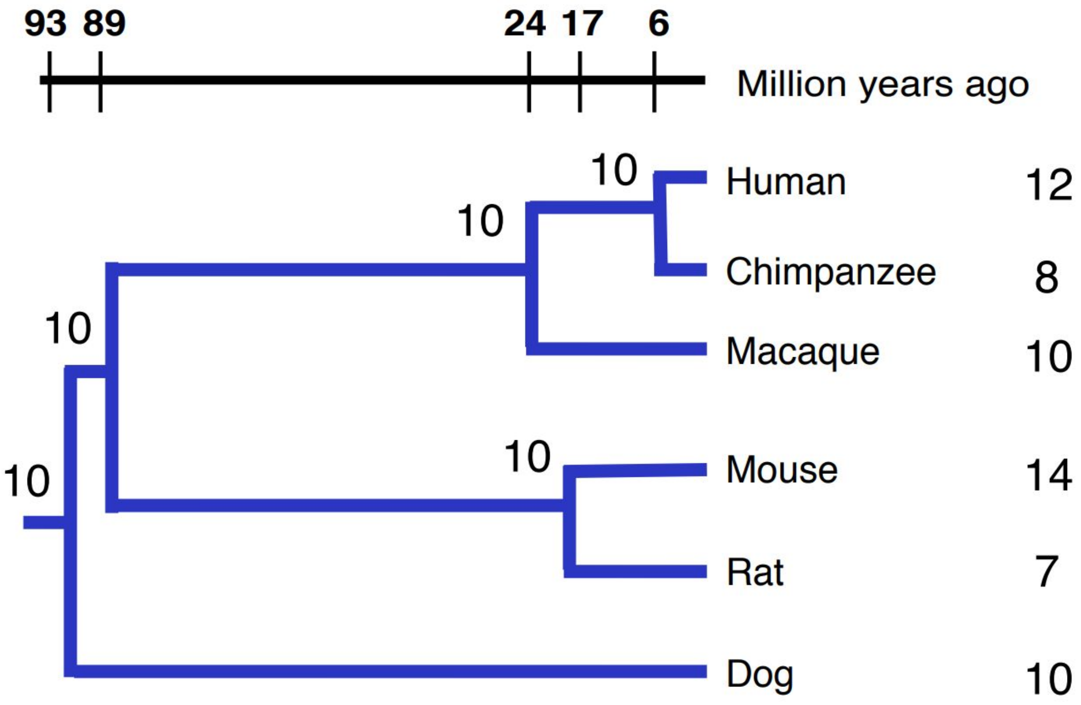


Models for gene family evolution:
Model family gene numbers as quantitative traits



Felsenstein (2005)

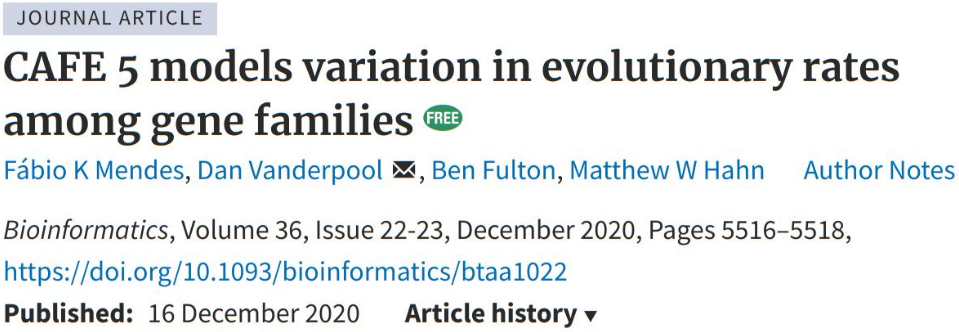
Models for gene family evolution



λ=0.002

(assuming birth=death)

Models for gene family evolution



JOURNAL ARTICLE

CAFE 5 models variation in evolutionary rates among gene families FREE

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** ▾

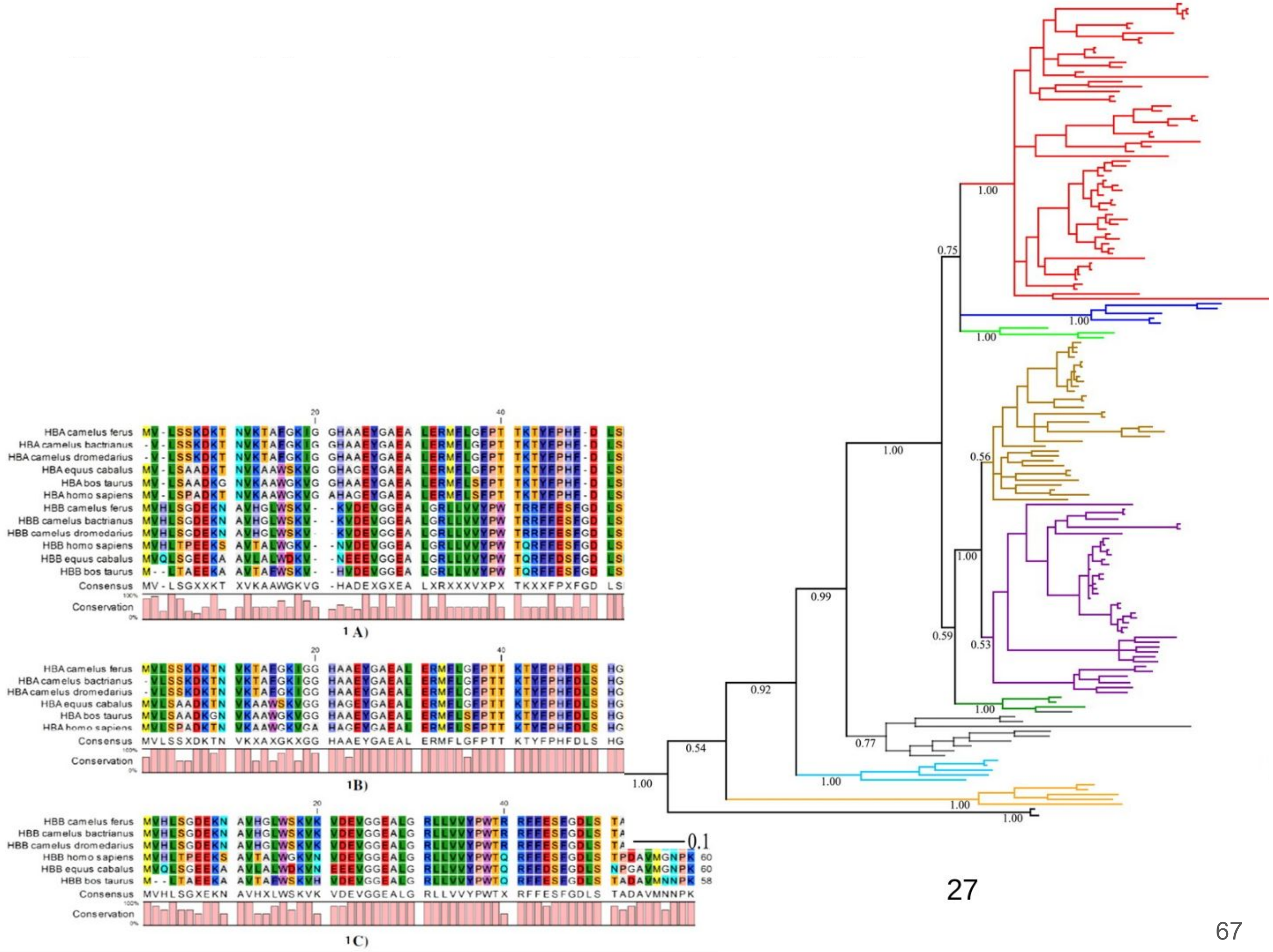
- Allows different rates in different branches and across families
- Models gene annotation errors

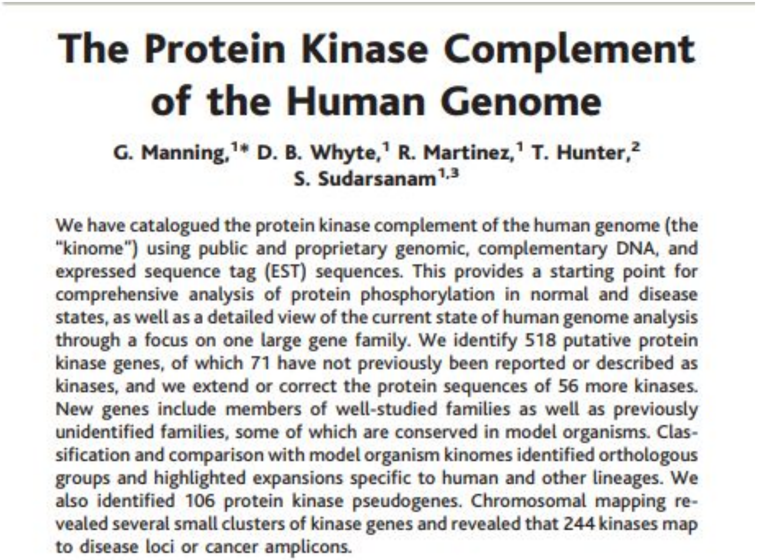
How to study gene family evolution?

1) Model gene family content across a species tree

2) **Reconstruct gene (family) phylogenies and compare them with the species tree**

Gene trees

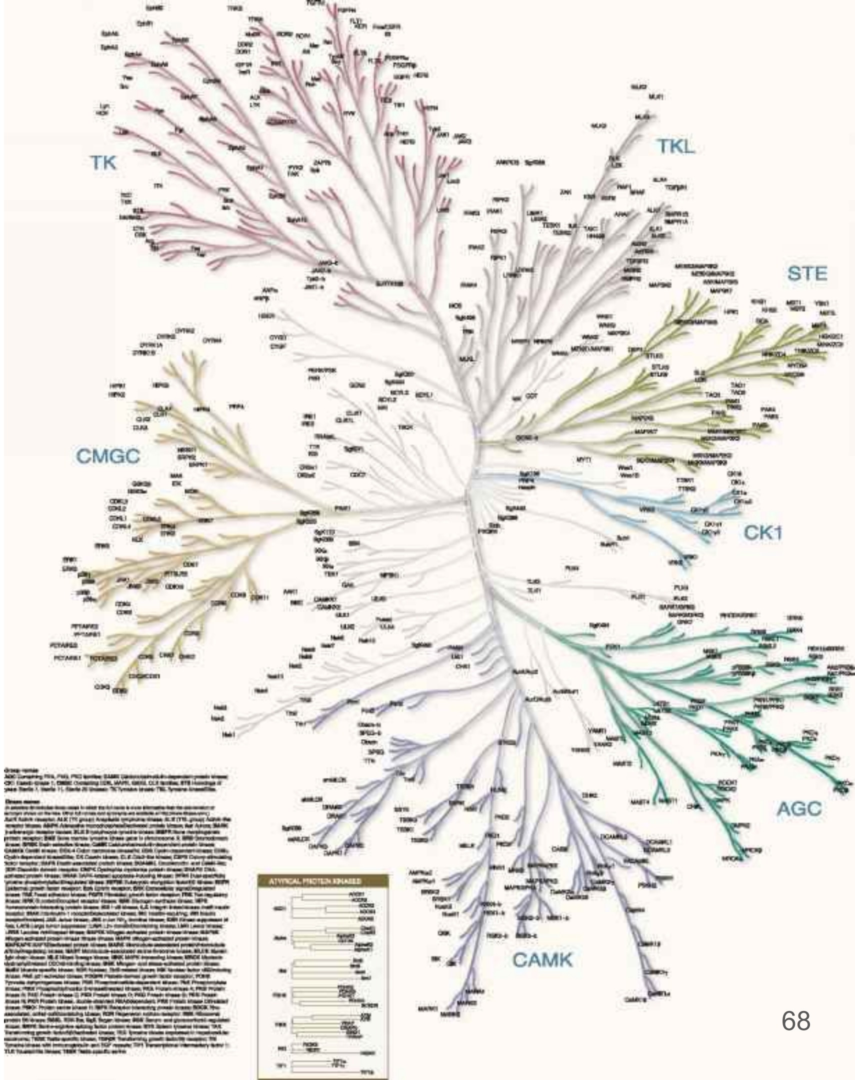

20
40
HBA camelus ferus MV-LSSKDKT NVKTAFGKIG GHAAEYGAEA LERMFLGFPT TKTYFPHF-D LS
HBA camelus bactrianus -V-LSSKDKT NVKTAFGKIG GHAAEYGAEA LERMFLGFPT TKTYFPHF-D LS
HBA camelus dromedarius -V-LSSKDKT NVKTAFGKIG GHAAEYGAEA LERMFLGFPT TKTYFPHF-D LS
HBA equus cabalus MV-LSAADKT NVKAAWSKVG GHAGEYGAEA LERMFLGFPT TKTYFPHF-D LS
HBA bos taurus MV-LSAADKG NVKAAWGKVG GHAAEYGAEA LERMFLSFPT TKTYFPHF-D LS
HBA homo sapiens MV-LSPADKT NVKAAWGKVG AHAGEYGAEA LERMFLSFPT TKTYFPHF-D LS
HBB camelus ferus MVHLSGDEKN AVHGLWSKV- -KVDEVGGEA LGRLLVVYPW TRRFFESFGD LS
HBB camelus bactrianus MVHLSGDEKN AVHGLWSKV- -KVDEVGGEA LGRLLVVYPW TRRFFESFGD LS
HBB camelus dromedarius MVHLSGDEKN AVHGLWSKV- -KVDEVGGEA LGRLLVVYPW TRRFFESFGD LS
HBB homo sapiens MVHLTPEEKS AVTALWGKV- -NVDEVGGEA LGRLLVVYPW TQRFFESFGD LS
HBB equus cabalus MVQLSGEEKA AVLALWDKV- -NEEEVGGEA LGRLLVVYPW TQRFFDSFGD LS
HBB bos taurus M--LTAEEKA AVTAFWSKV- -HVDEVGGEA LGRLLVVYPW TQRFFESFGD LS
Consensus MV-LSGXXKT XVKAAWGKVG -HADEXGXEA LXRXXXVXPX TKXXFPXFGD LS
100%
Conservation
0%
1 A)
20
40
HBA camelus ferus MVLSSKDKTN VKTAFGKIGG HAAEYGAEAL ERMFLGFPTT KTYFPHFDLS HG
HBA camelus bactrianus -VLSSKDKTN VKTAFGKIGG HAAEYGAEAL ERMFLGFPTT KTYFPHFDLS HG
HBA camelus dromedarius -VLSSKDKTN VKTAFGKIGG HAAEYGAEAL ERMFLGFPTT KTYFPHFDLS HG
HBA equus cabalus MVLSAADKTN VKAAWSKVGG HAGEYGAEAL ERMFLGFPTT KTYFPHFDLS HG
HBA bos taurus MVLSAADKGN VKAAWGKVGG HAAEYGAEAL ERMFLSFPTT KTYFPHFDLS HG
HBA homo sapiens MVLSPADKTN VKAAWGKVGA HAGEYGAEAL ERMFLSFPTT KTYFPHFDLS HG
Consensus MVLSSXDKTN VKXAXGKXGG HAAEYGAEAL ERMFLGFPTT KTYFPHFDLS HG
100%
Conservation
0%
1B)
20
40
HBB camelus ferus MVHLSGDEKN AVHGLWSKVK VDEVGGEALG RLLVVYPWTR RFFESFGDLS TA
HBB camelus bactrianus MVHLSGDEKN AVHGLWSKVK VDEVGGEALG RLLVVYPWTR RFFESFGDLS TA
HBB camelus dromedarius MVHLSGDEKN AVHGLWSKVK VDEVGGEALG RLLVVYPWTR RFFESFGDLS TA
HBB homo sapiens MVHLTPEEKS AVTALWGKVN VDEVGGEALG RLLVVYPWTQ RFFESFGDLS TPDAVMGNPK 60
HBB equus cabalus MVQLSGEEKA AVLALWDKVN EEEVGGEALG RLLVVYPWTQ RFFDSFGDLS NPGAVMGNPK 60
HBB bos taurus M--LTAEEKA AVTAFWSKVH VDEVGGEALG RLLVVYPWTQ RFFESFGDLS TADAVMNNPK 58
Consensus MVHLSGXEKN AVHXLWSKVK VDEVGGEALG RLLVVYPWTX RFFESFGDLS TADAVMNNPK
100%
Conservation
0%
1C)
1.00
0.75
1.00
1.00
1.00
0.56
1.00
1.00
0.53
0.59
0.99
1.00
0.77
0.92
0.54
1.00
1.00
1.00
1.00
1.00
0.1
27

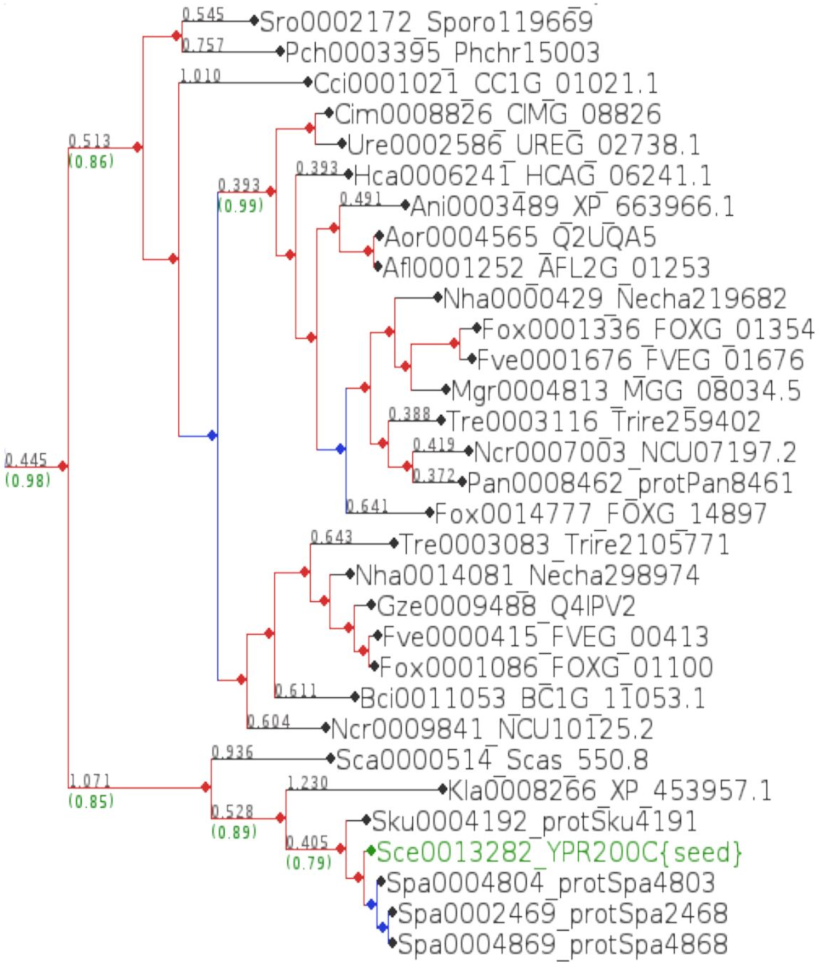


The Protein Kinase Complement of the Human Genome

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

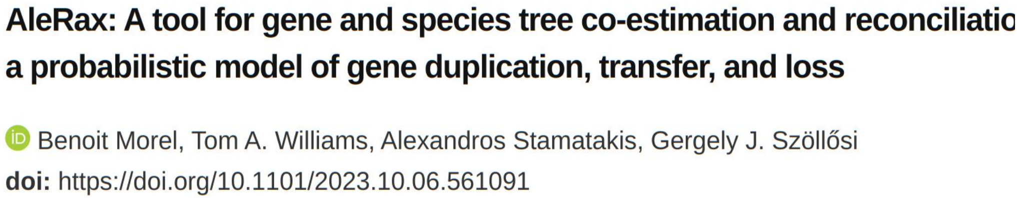
Co-estimate gene trees and species trees: GeneRax, ALE

Approaches

Interrogate gene trees independent of species tree

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

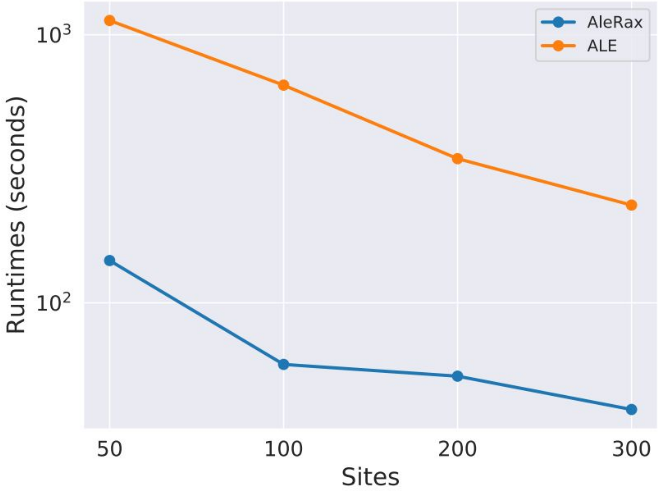
Co-estimate gene trees and species trees: GeneRax, ALE



AleRax: A tool for gene and species tree co-estimation and reconciliatio 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)

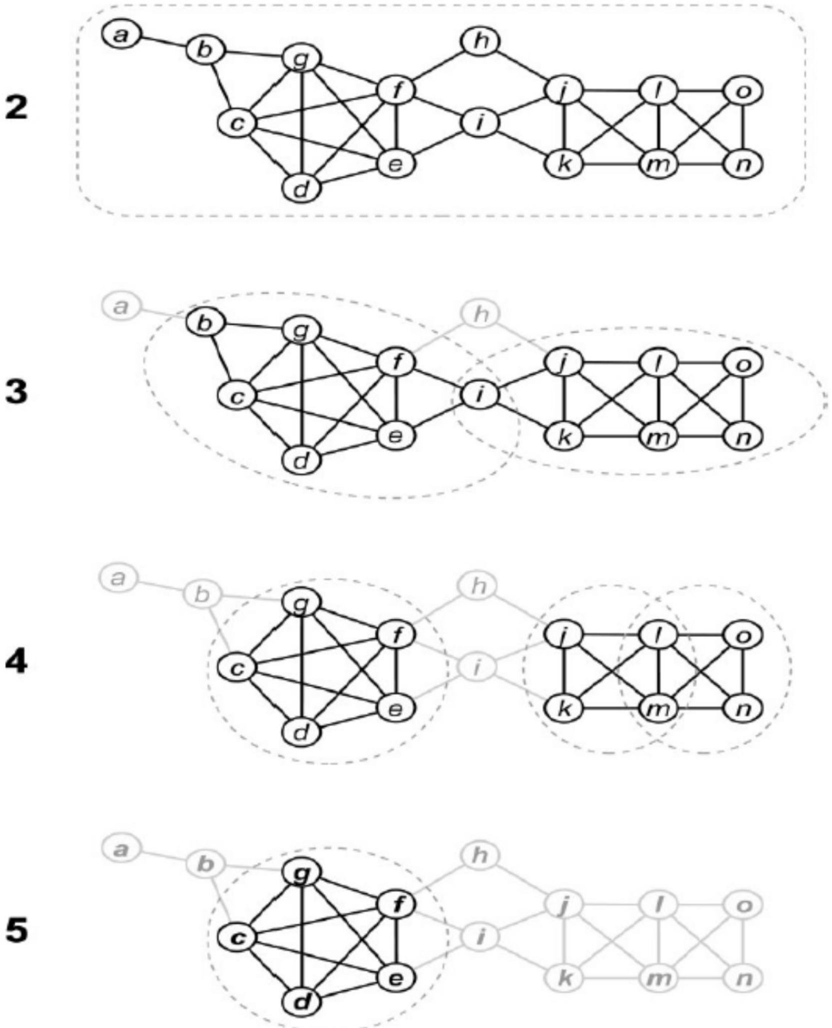
Build gene families by a blast-based clustering approach (e.g. Orthofinder)

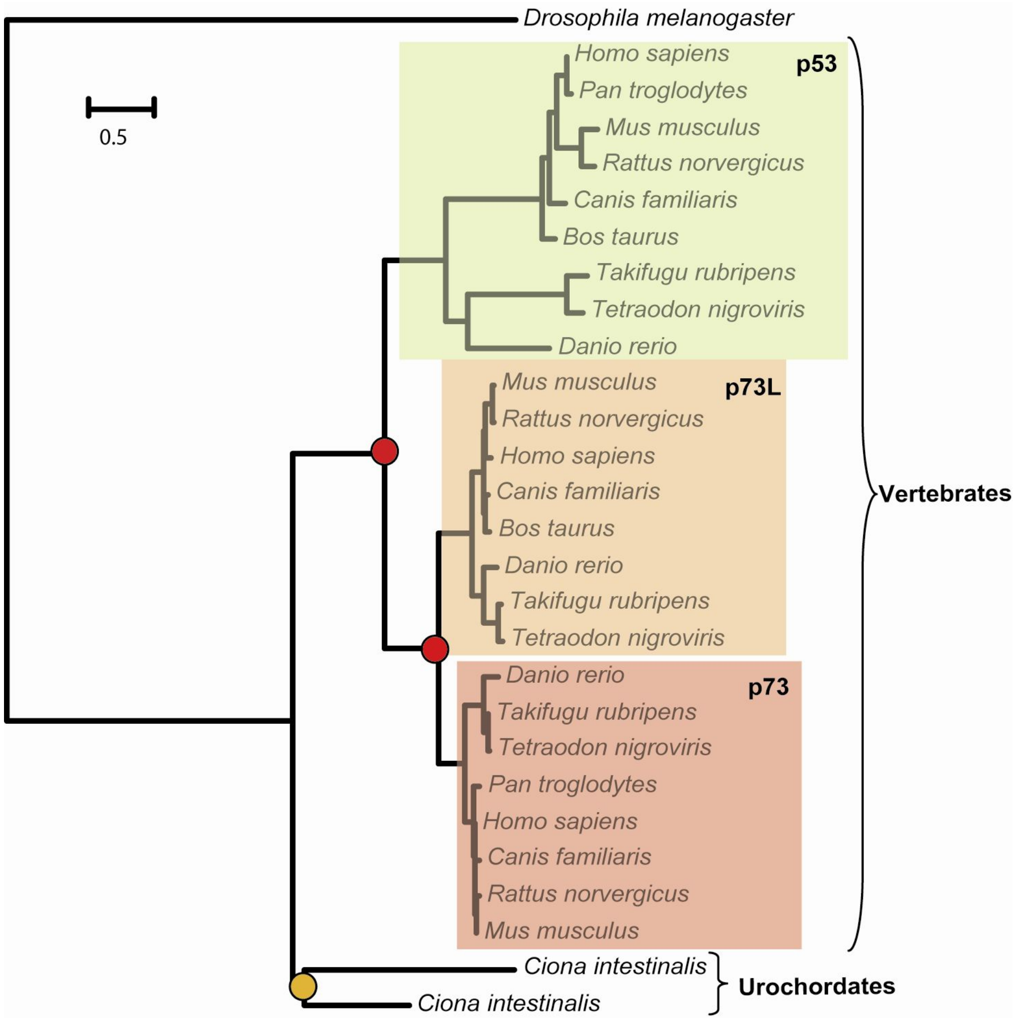
Then make a gene tree per family

B) Gene centric approach (PhylomeDB)

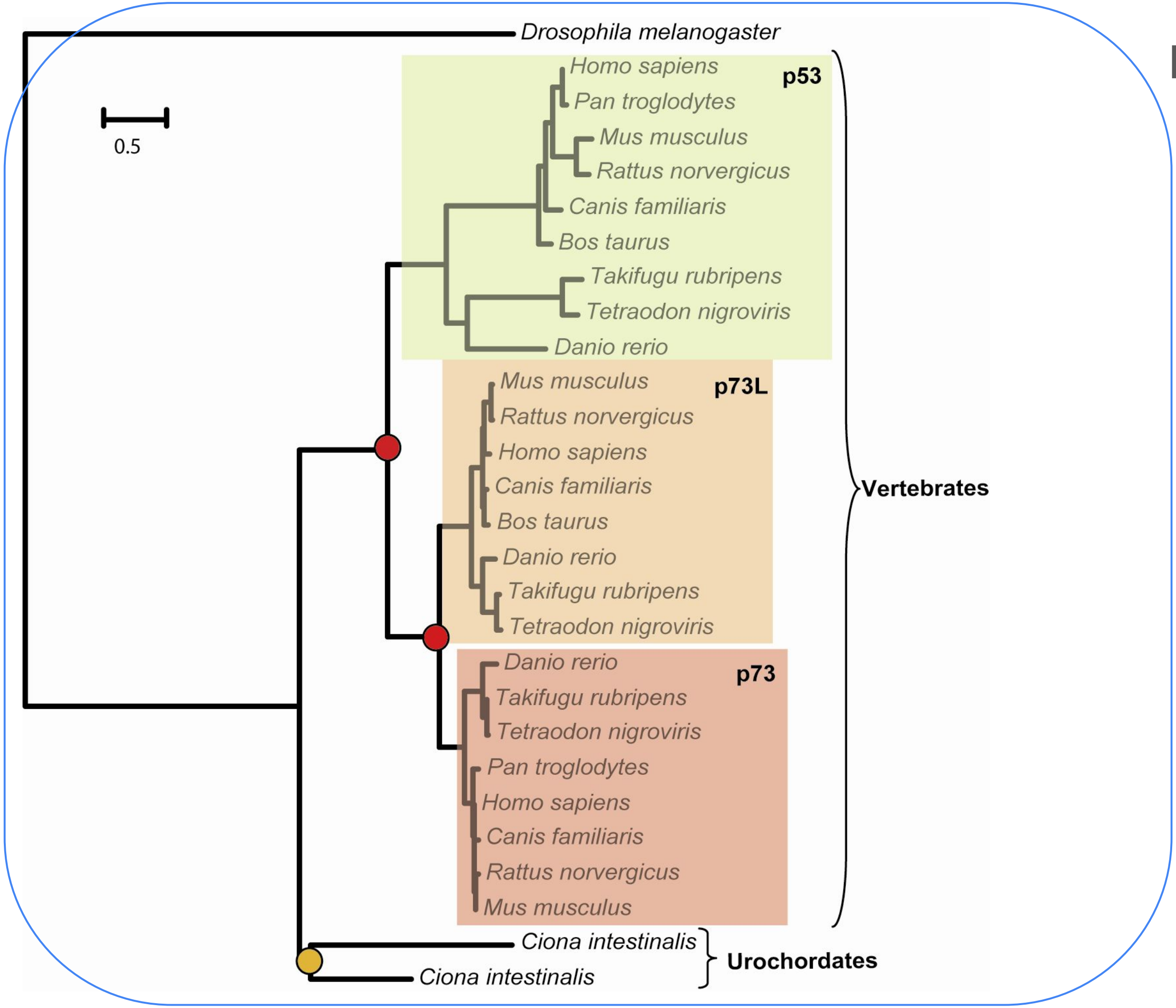
Take a seed genome, for every gene find homologs with blast, reconstruct a gene tree per gene (multiple gene trees per family are possible)

Finding optimal granularity might be tricky


2
a b g h f c i j l o k m n e d
3
a b g h f c i j l o k m n e d
4
a b g h f c i j l o k m n e d
5
a b g h f c i j l o k m n e d

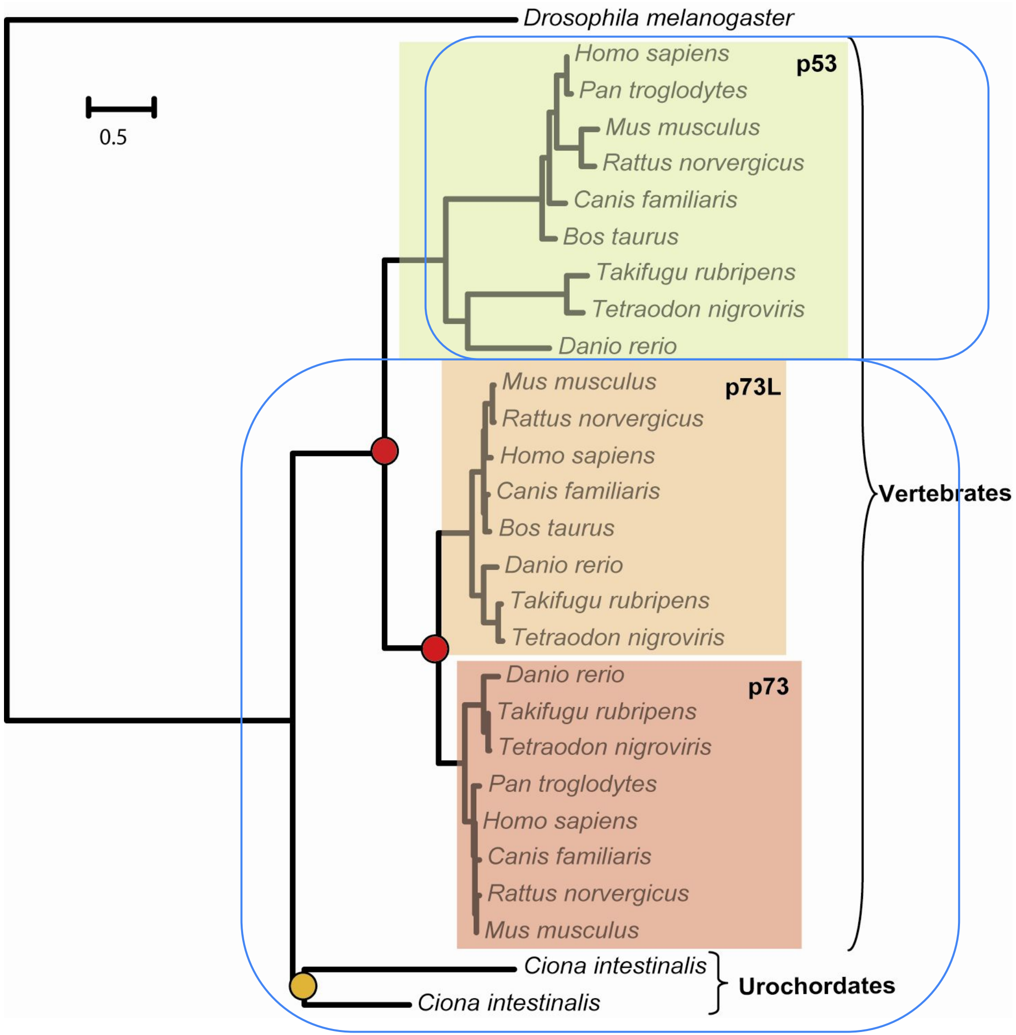


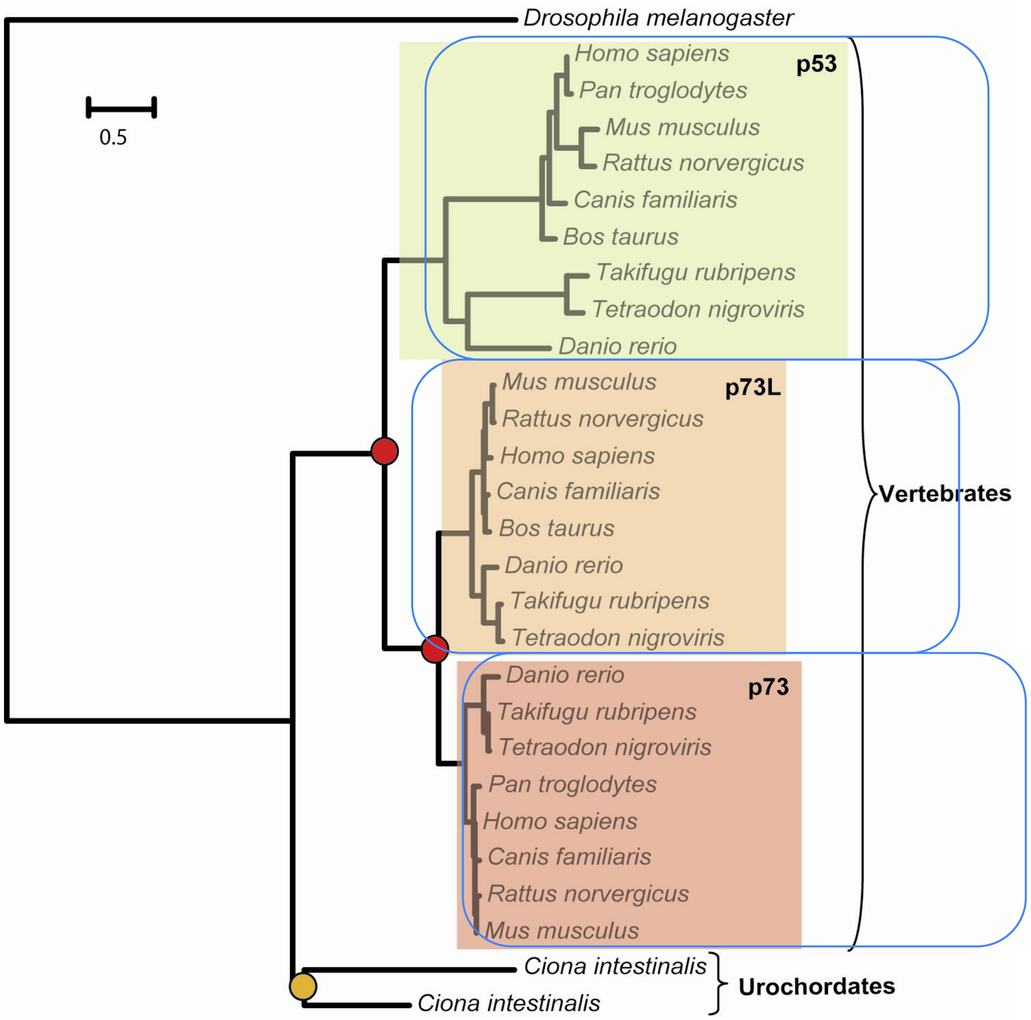
How many families?



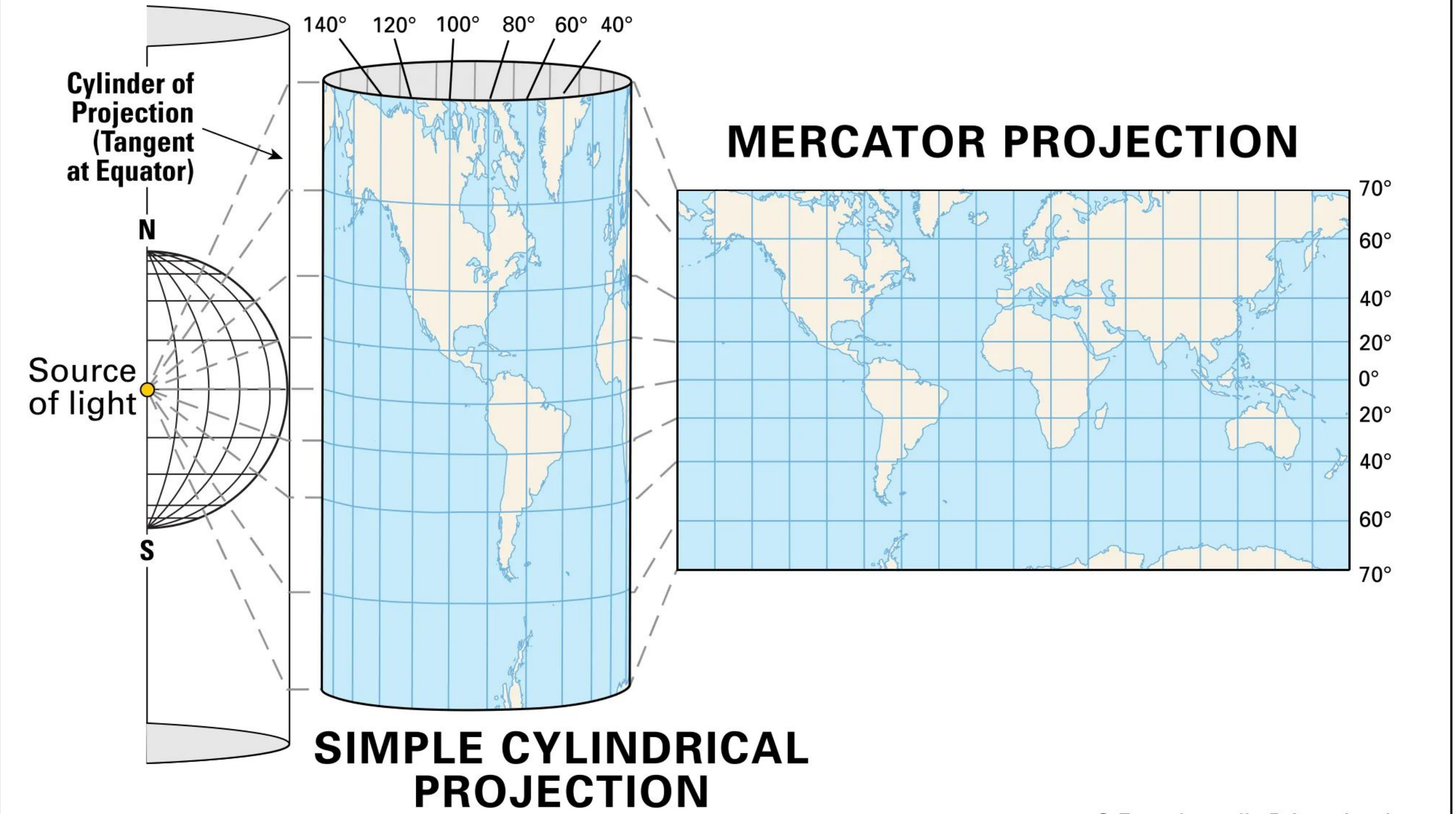
How many families?

How many families?

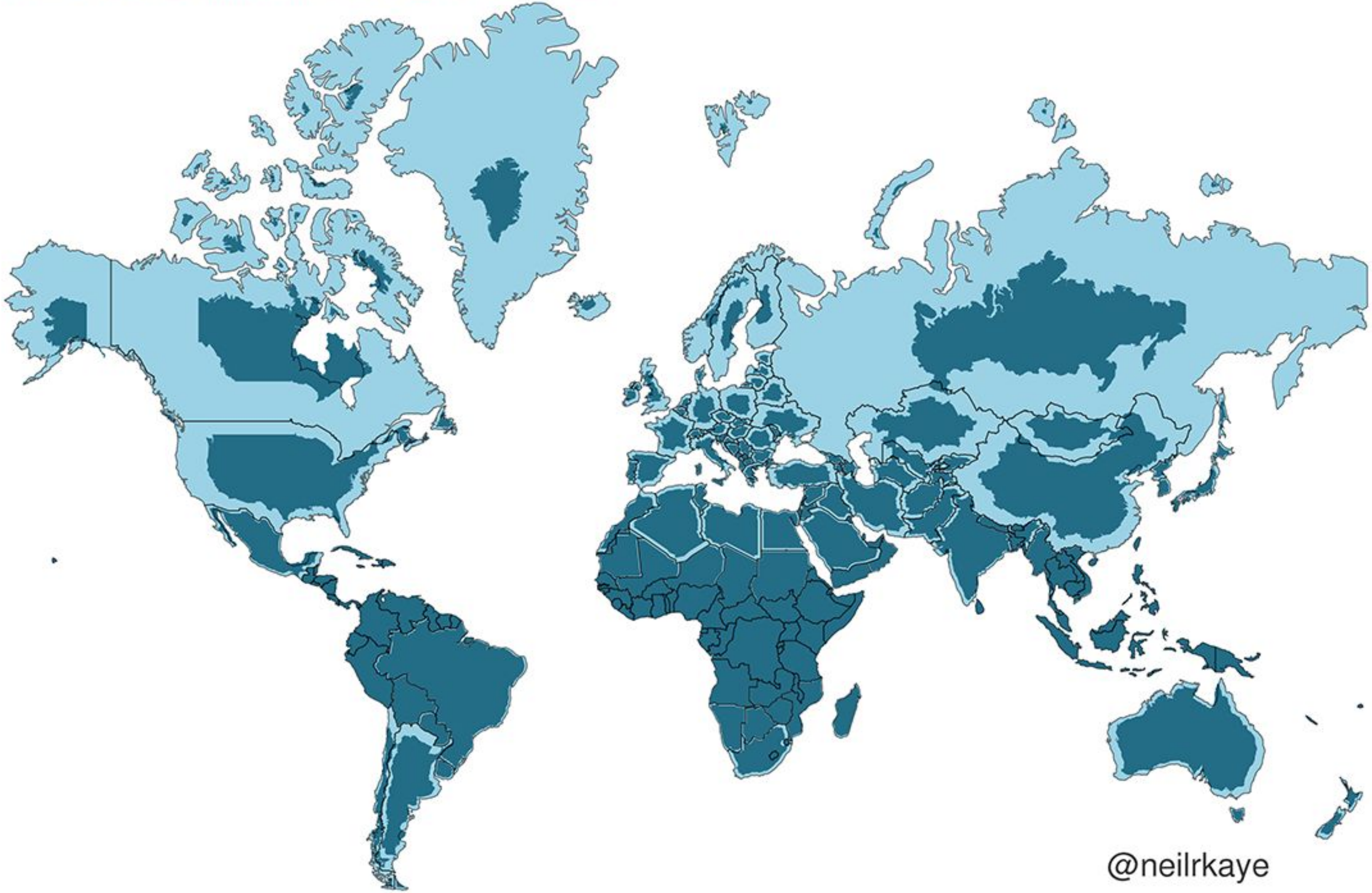




How many families?


140°
120°
100°
80°
60°
40°
Cylinder of Projection (Tangent at Equator)
N
Source of light
S
MERCATOR PROJECTION
70°
60°
40°
20°
0°
20°
40°
60°
70°
SIMPLE CYLINDRICAL PROJECTION

MERCATOR PROJECTION VS THE TRUE SIZE OF COUNTRIES

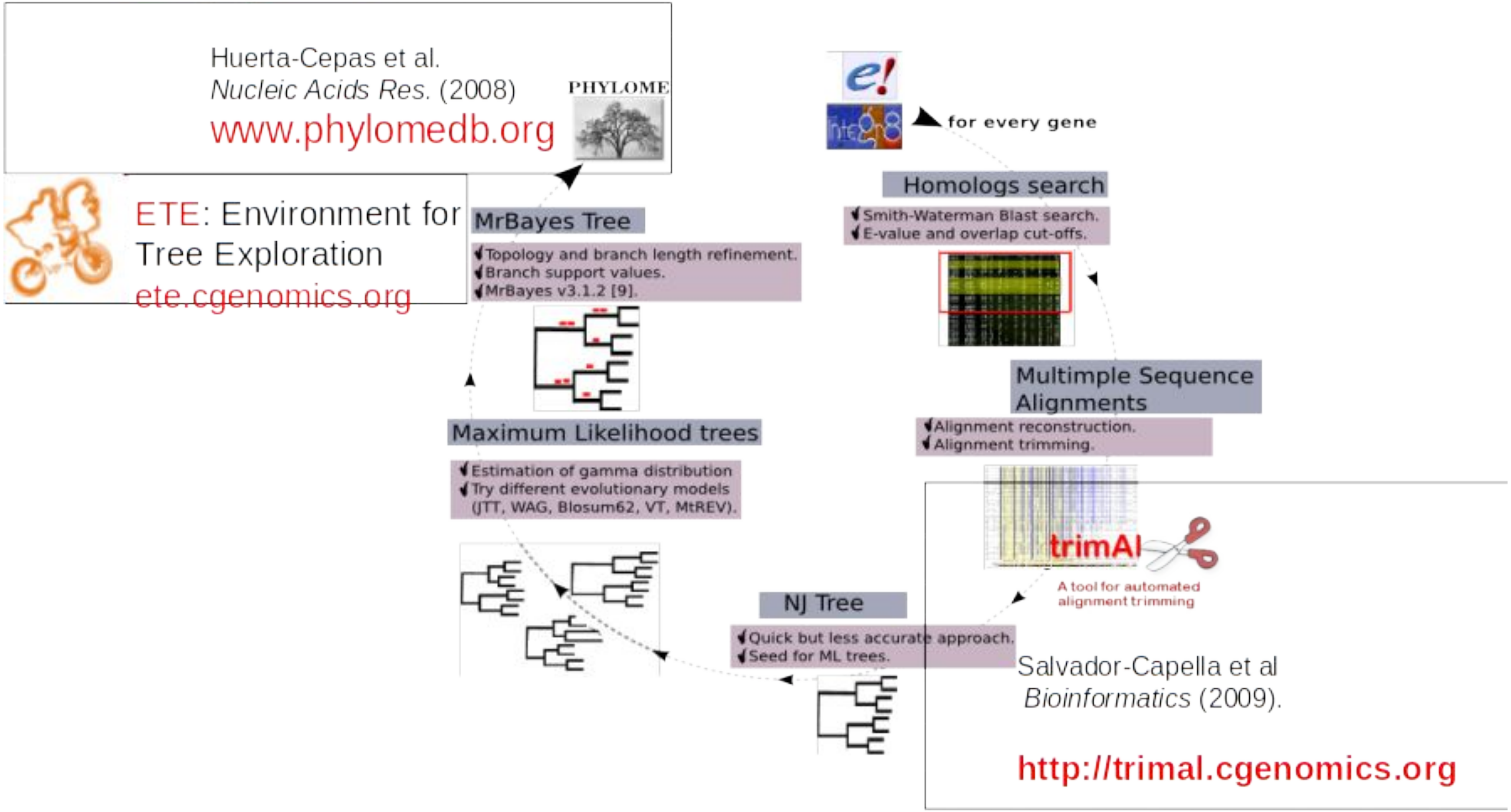


@neilrkaye

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:



Huerta-Cepas et al.
Nucleic Acids Res. (2008)
www.phylomedb.org

PHYLOME

ETE: Environment for Tree Exploration
ete.cgenomics.org

for every gene

Homologs search

✓ Smith-Waterman Blast search.
✓ E-value and overlap cut-offs.

Multimple Sequence Alignments

✓ Alignment reconstruction.
✓ Alignment trimming.

trimAl
A tool for automated alignment trimming

NJ Tree

✓ Quick but less accurate approach.
✓ Seed for ML trees.

Maximum Likelihood trees

✓ Estimation of gamma distribution
✓ Try different evolutionary models (JTT, WAG, Blosum62, VT, MtREV).

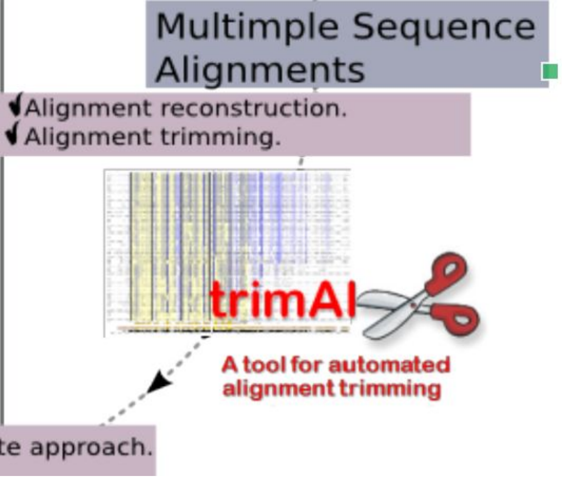
MrBayes Tree

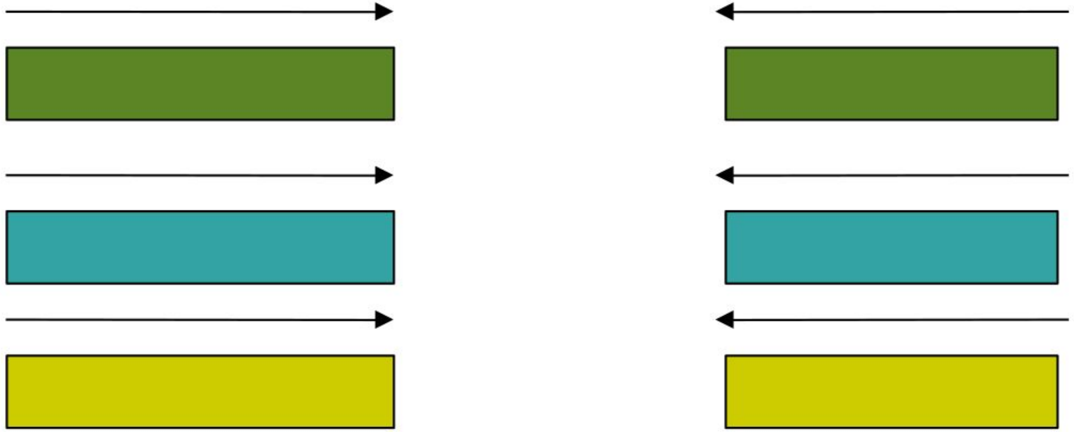
✓ Topology and branch length refinement.
✓ Branch support values.
✓ MrBayes v3.1.2 [9].

Salvador-Capella et al
Bioinformatics (2009).

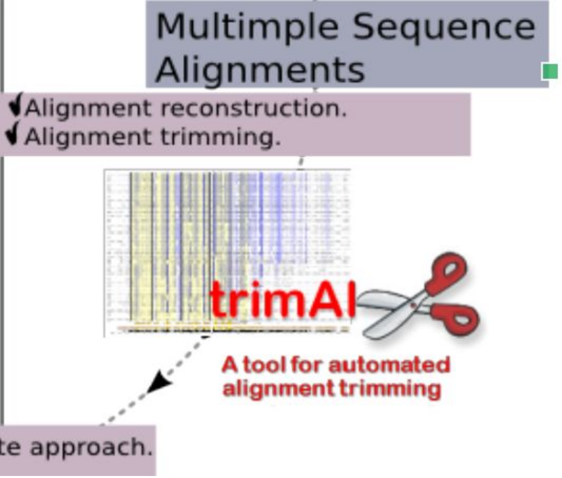
http://trimal.cgenomics.org

Last pipeline described in Fuentes et al NAR (2020)


Multimple Sequence Alignments
✓Alignment reconstruction.
✓Alignment trimming.
trimAl
A tool for automated alignment trimming
te approach.



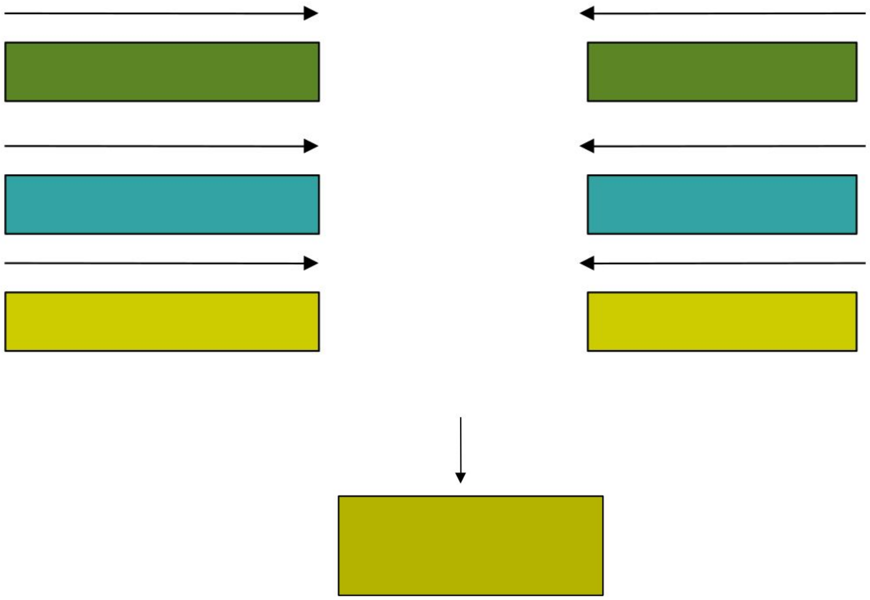
Homologous sequences aligned in forward and reversed (head or tail approach), and each of them with three different algorithms: 2 x 3 = 6 different alignments



Multimple Sequence Alignments

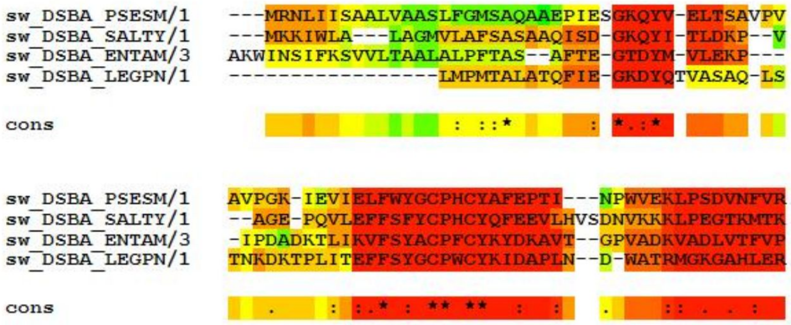
✓Alignment reconstruction.
✓Alignment trimming.

te approach.



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

TrimAl trims based on a consistency score

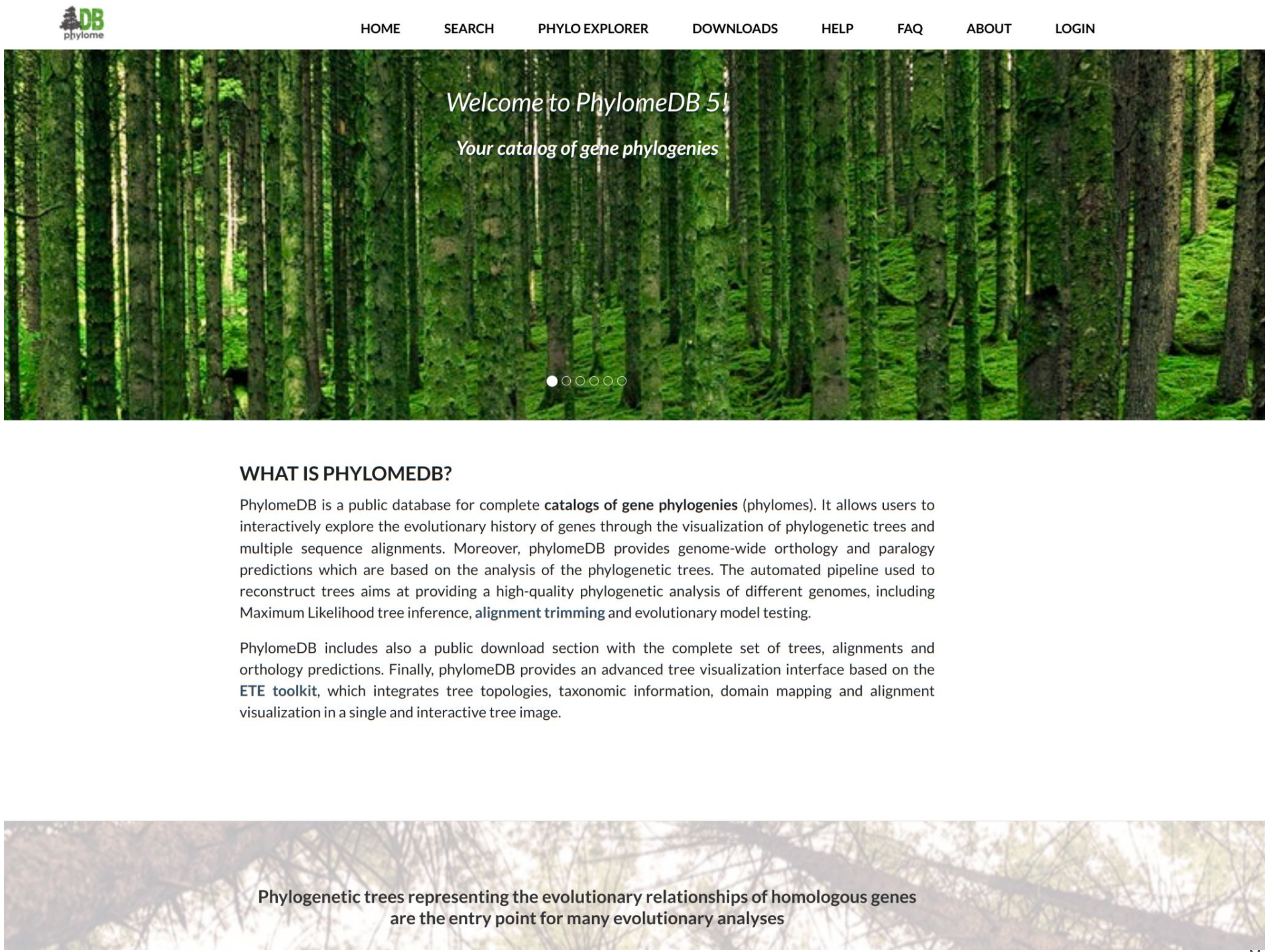


```
sw_DSBA_PSESM/1   ---MRNLIISAALVAASLFGMSAQAAEPIESGKQYV-ELTSAVPV
sw_DSBA_SALTY/1   ---MKKIWLA---LAGMVLAFSASAAQISD-GKQYI-TLDKP--V
sw_DSBA_ENTAM/3   AKWINSIFKSVVLTAALALPFTAS--AFTE-GTDYM-VLEKP---
sw_DSBA_LEGPN/1   ----------------LMPMTALATQFIE-GKDYQTVASAQ-LS

cons                                : ::*      : *.:*

sw_DSBA_PSESM/1   AVPGK-IEVIELFWYGCPHCYAFEPTI---NPWVEKLPSDVNFVR
sw_DSBA_SALTY/1   --AGE-PQVLEFFSFYCPHCYQFEEVLHVSDNVKKKLPEGTKMTK
sw_DSBA_ENTAM/3   -IPDADKTLIKVFSYACPFCYKYDKAVT--GPVADKVADLVTFVP
sw_DSBA_LEGPN/1   TNKDKTPLITEFFSYGCPWCYKIDAPLN--D-WATRMGKGAHLER

cons                 .    : :.* : ** **  :  :   .    :: . . :
```



Welcome to PhylomeDB 5!

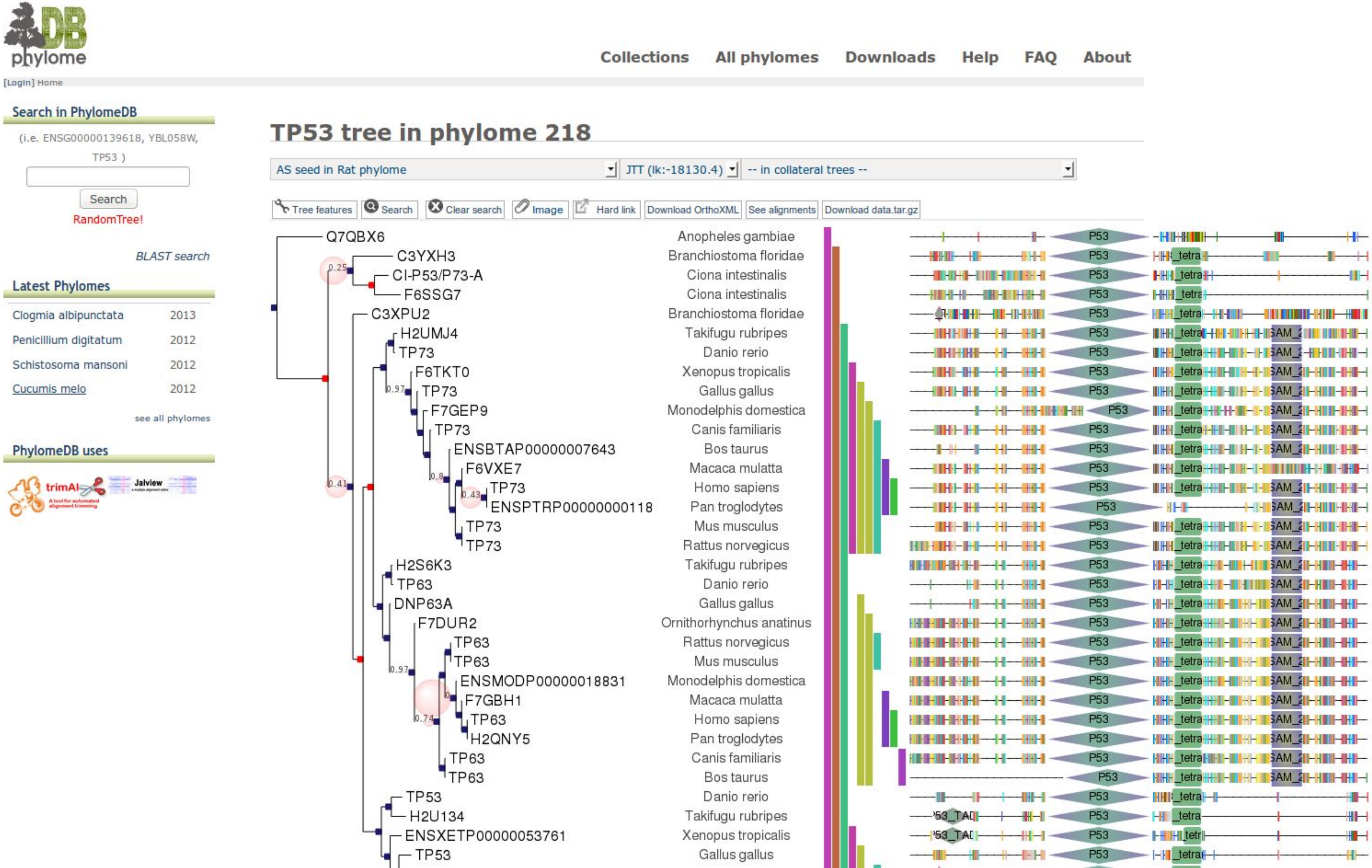
Your catalog of gene phylogenies

WHAT IS PHYLOMEDB?

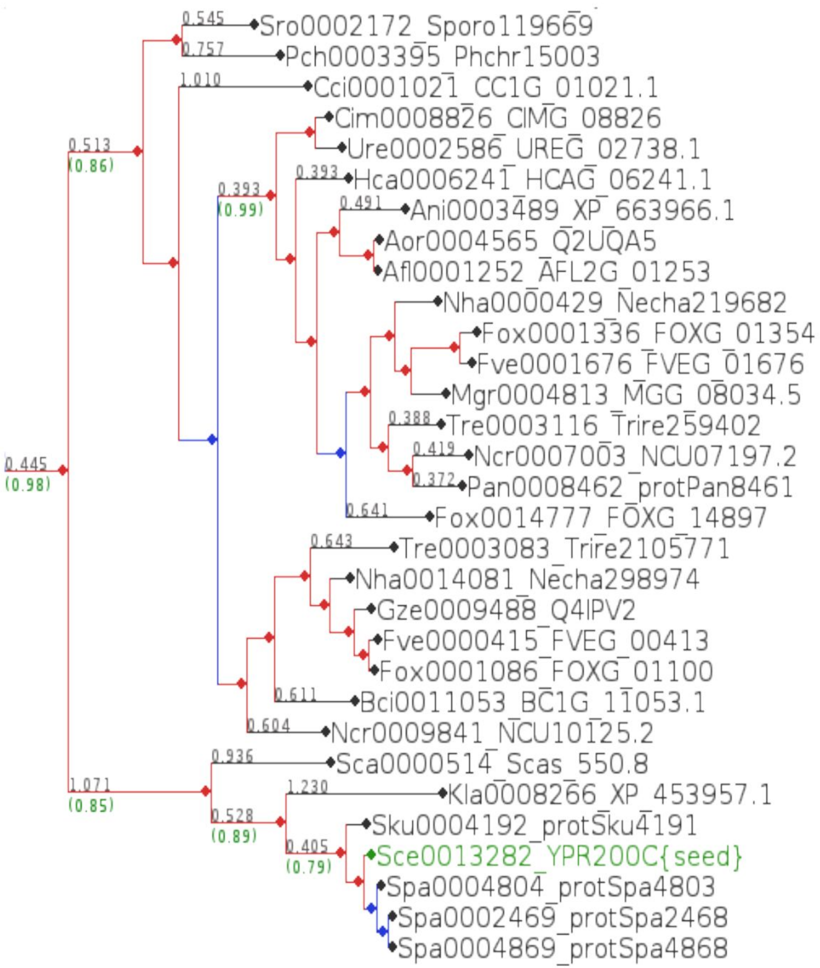
PhylomeDB is a public database for complete **catalogs of gene phylogenies** (phylomes). It allows users to interactively explore the evolutionary history of genes through the visualization of phylogenetic trees and multiple sequence alignments. Moreover, phylomeDB provides genome-wide orthology and 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.

Phylogenetic trees representing the evolutionary relationships of homologous genes are the entry point for many evolutionary analyses

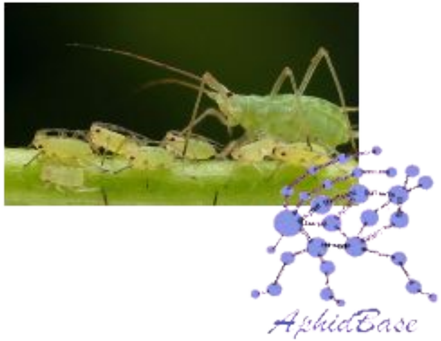


TP53 tree in phylome 218



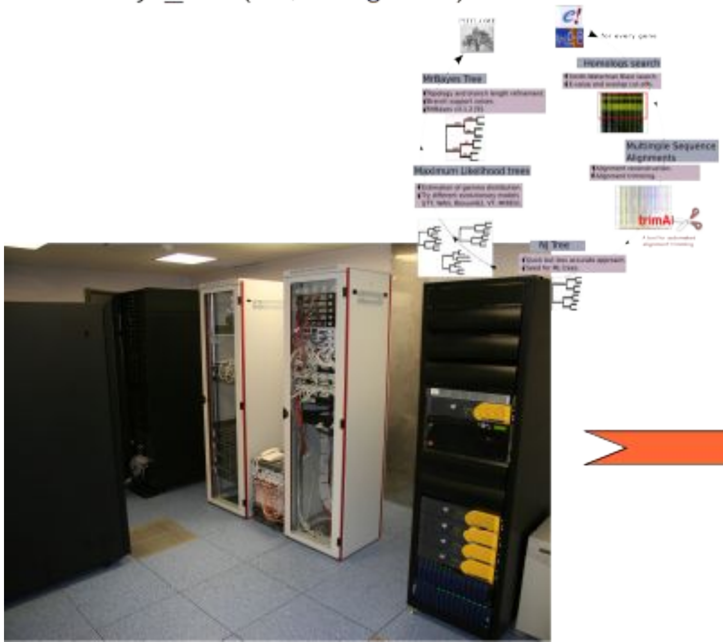
Tree collections can be interrogated to:

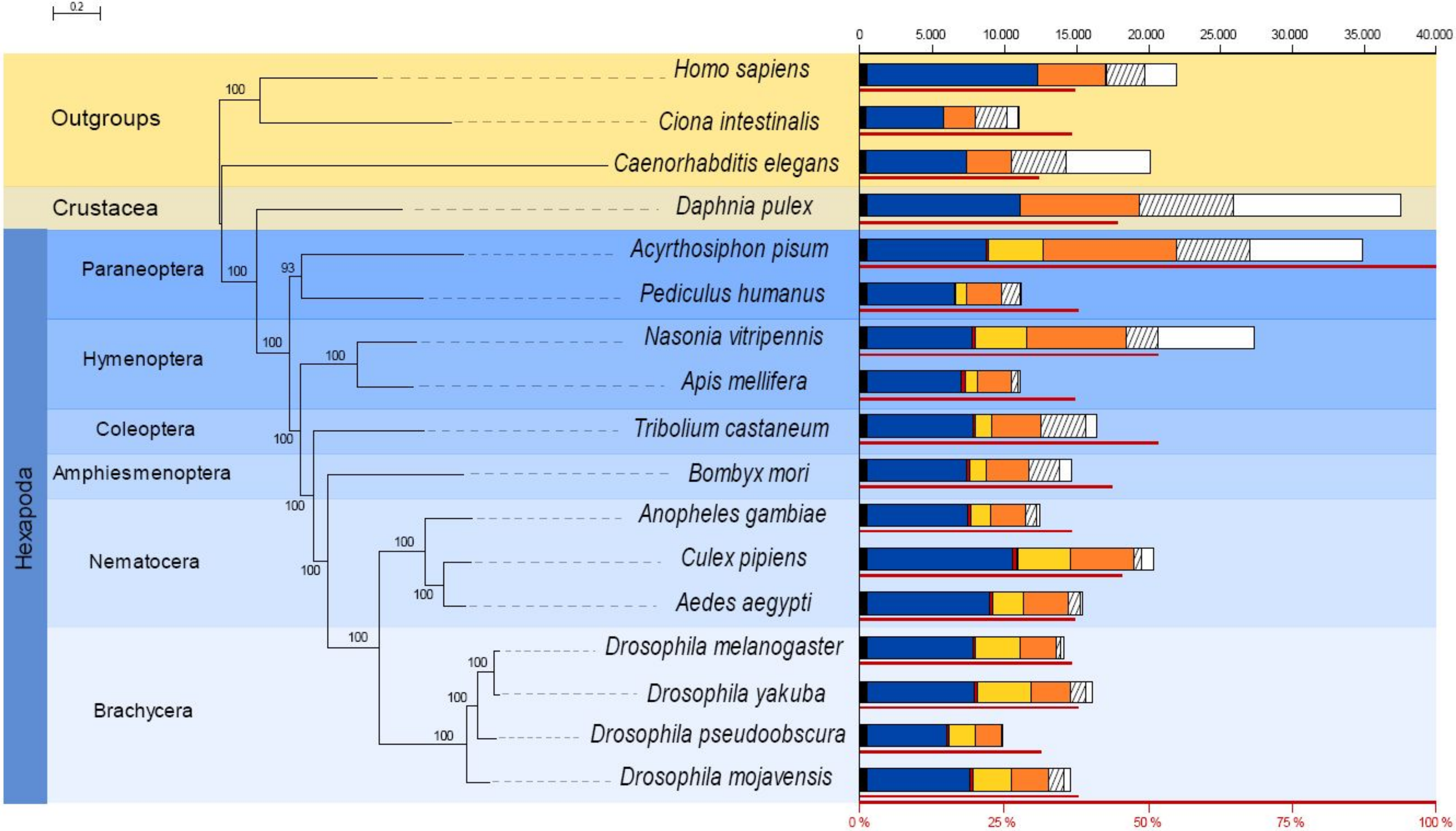
- Find families that show a particular topology
- Detect and date duplication events
- Genes that have accelerated evolutionary rates at a particular lineage (positive/relaxed selection)
- Detect families expanded at particular lineages
- Detect footprints of horizontal gene transfer, lineage sorting,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



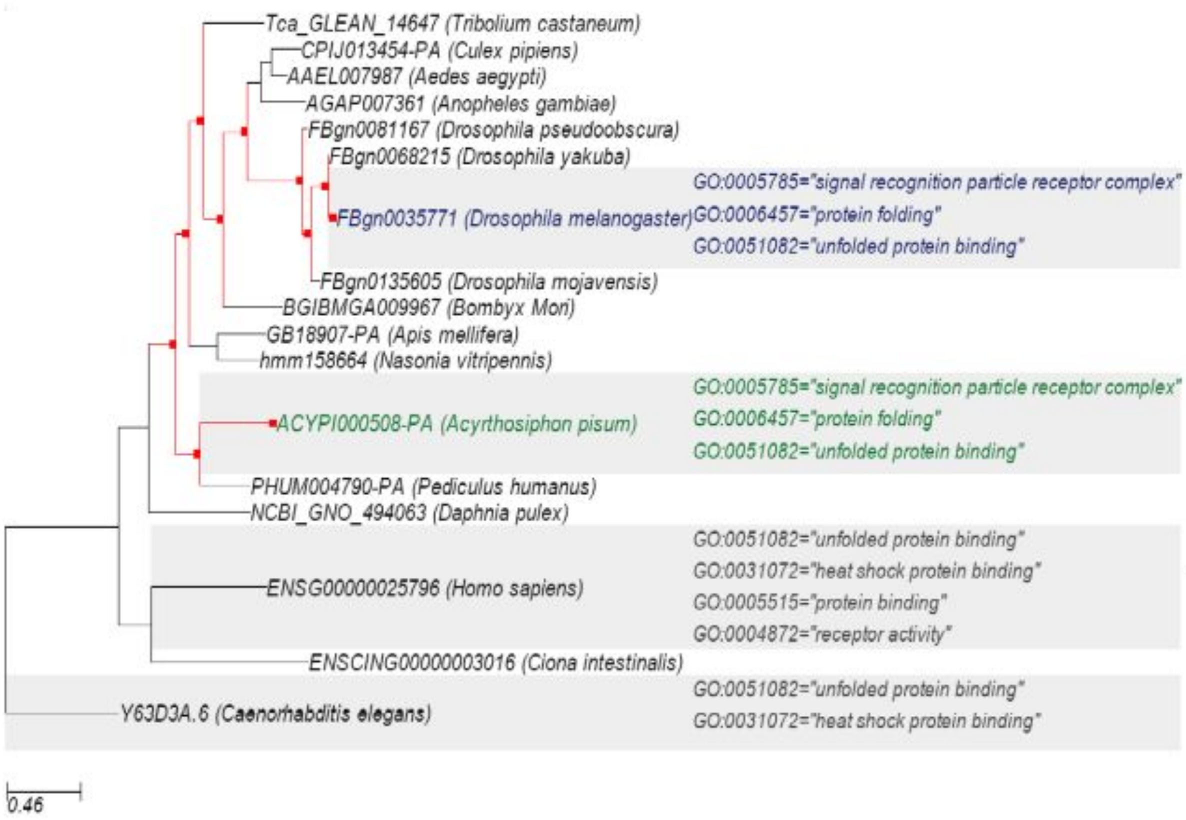
Acyr_1.0 (34,600 genes)

13 other sequenced arthropods and 3 out-groups


for every gene
Homologs search
Smith-Waterman blast search
E-value and overlap cut-offs
MrBayes Tree
Topology and branch length refinement
Branch support values
MrBayes v3.1.2 [3]
Maximum Likelihood trees
Estimation of gamma distribution
Try different evolutionary models (JTT, WAG, Blosum62, VT, MtREV)
Multiple Sequence Alignments
Alignment reconstruction
Alignment trimming
trimAl
A tool for automated alignment trimming
NJ Tree
Quick but less accurate approach
Seed for ML trees



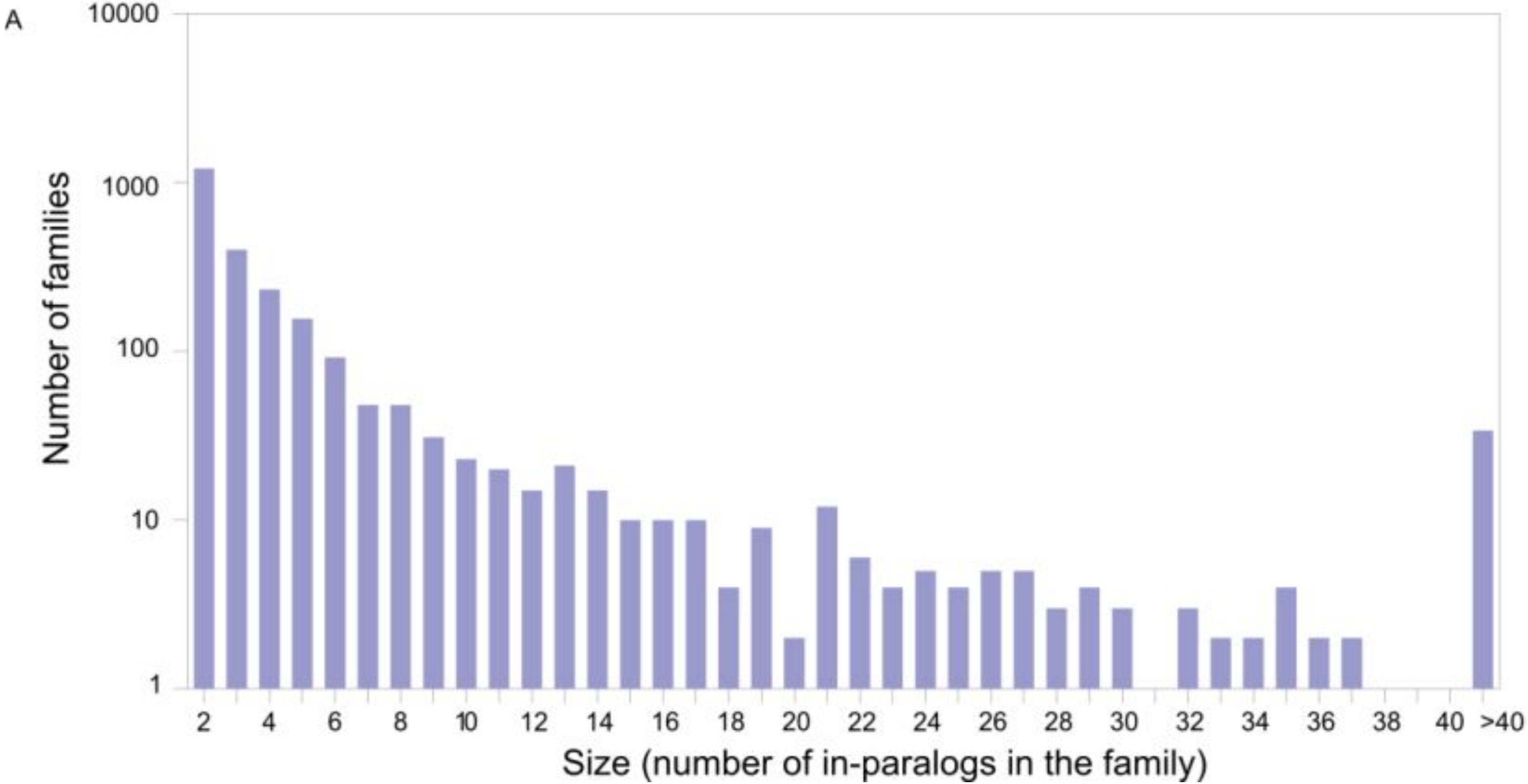
0.2
0
5.000
10.000
15.000
20.000
25.000
30.000
35.000
40.000
Outgroups
Crustacea
Hexapoda
Paraneoptera
Hymenoptera
Coleoptera
Amphiesmenoptera
Nematocera
Brachycera
Homo sapiens
Ciona intestinalis
Caenorhabditis elegans
Daphnia pulex
Acyrthosiphon pisum
Pediculus humanus
Nasonia vitripennis
Apis mellifera
Tribolium castaneum
Bombyx mori
Anopheles gambiae
Culex pipiens
Aedes aegypti
Drosophila melanogaster
Drosophila yakuba
Drosophila pseudoobscura
Drosophila mojavensis
100
100
93
100
100
100
100
100
100
100
100
100
100
100
0 %
25 %
50 %
75 %
100 %


Tca_GLEAN_14647 (Tribolium castaneum)
CPIJ013454-PA (Culex pipiens)
AAEL007987 (Aedes aegypti)
AGAP007361 (Anopheles gambiae)
FBgn0081167 (Drosophila pseudoobscura)
FBgn0068215 (Drosophila yakuba)
GO:0005785="signal recognition particle receptor complex"
FBgn0035771 (Drosophila melanogaster)
GO:0006457="protein folding"
GO:0051082="unfolded protein binding"
FBgn0135605 (Drosophila mojavensis)
BGIBMGA009967 (Bombyx Mori)
GB18907-PA (Apis mellifera)
hmm158664 (Nasonia vitripennis)
GO:0005785="signal recognition particle receptor complex"
ACYPI000508-PA (Acyrthosiphon pisum)
GO:0006457="protein folding"
GO:0051082="unfolded protein binding"
PHUM004790-PA (Pediculus humanus)
NCBI_GNO_494063 (Daphnia pulex)
GO:0051082="unfolded protein binding"
GO:0031072="heat shock protein binding"
ENSG00000025796 (Homo sapiens)
GO:0005515="protein binding"
GO:0004872="receptor activity"
ENSCING00000003016 (Ciona intestinalis)
GO:0051082="unfolded protein binding"
Y63D3A.6 (Caenorhabditis elegans)
GO:0031072="heat shock protein binding"
0.46

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)

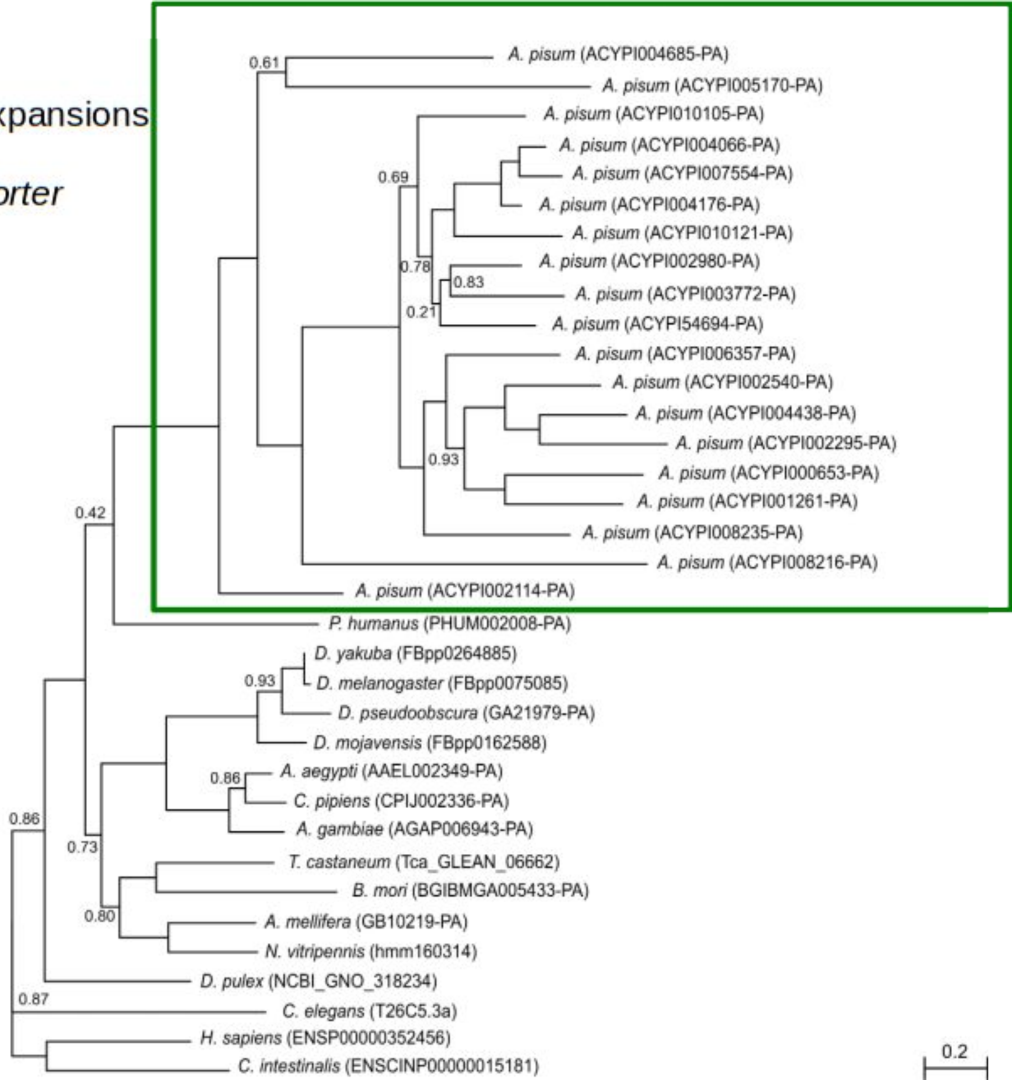
A wave of lineage-specific expansions in the pea aphid



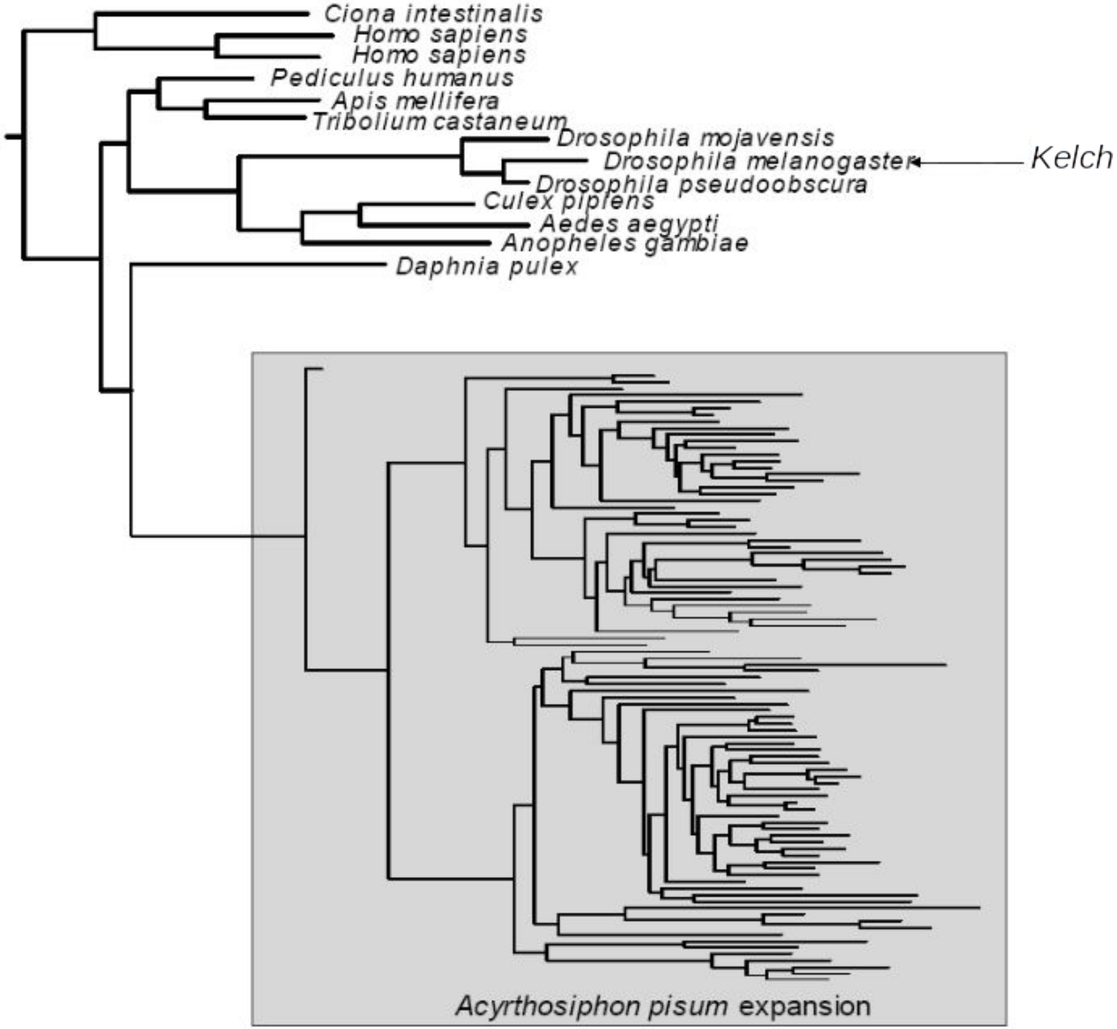
B

Lineage-specific expansions

Acetyl-CoA transporter



A. pisum (ACYPI004685-PA)
A. pisum (ACYPI005170-PA)
0.61
A. pisum (ACYPI010105-PA)
A. pisum (ACYPI004066-PA)
A. pisum (ACYPI007554-PA)
0.69
A. pisum (ACYPI004176-PA)
A. pisum (ACYPI010121-PA)
A. pisum (ACYPI002980-PA)
0.78
0.83
A. pisum (ACYPI003772-PA)
0.21
A. pisum (ACYPI54694-PA)
A. pisum (ACYPI006357-PA)
A. pisum (ACYPI002540-PA)
A. pisum (ACYPI004438-PA)
A. pisum (ACYPI002295-PA)
0.93
A. pisum (ACYPI000653-PA)
A. pisum (ACYPI001261-PA)
0.42
A. pisum (ACYPI008235-PA)
A. pisum (ACYPI008216-PA)
A. pisum (ACYPI002114-PA)
P. humanus (PHUM002008-PA)
D. yakuba (FBpp0264885)
0.93
D. melanogaster (FBpp0075085)
D. pseudoobscura (GA21979-PA)
D. mojavensis (FBpp0162588)
0.86
A. aegypti (AAEL002349-PA)
C. pipiens (CPIJ002336-PA)
0.86
A. gambiae (AGAP006943-PA)
0.73
T. castaneum (Tca_GLEAN_06662)
B. mori (BGIBMGA005433-PA)
0.80
A. mellifera (GB10219-PA)
N. vitripennis (hmm160314)
D. pulex (NCBI_GNO_318234)
0.87
C. elegans (T26C5.3a)
H. sapiens (ENSP00000352456)
C. intestinalis (ENSCINP00000015181)
0.2


Ciona intestinalis
Homo sapiens
Homo sapiens
Pediculus humanus
Apis mellifera
Tribolium castaneum
Drosophila mojavensis
Drosophila melanogaster
Kelch
Drosophila pseudoobscura
Culex pipiens
Aedes aegypti
Anopheles gambiae
Daphnia pulex
Acyrthosiphon pisum expansion

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?

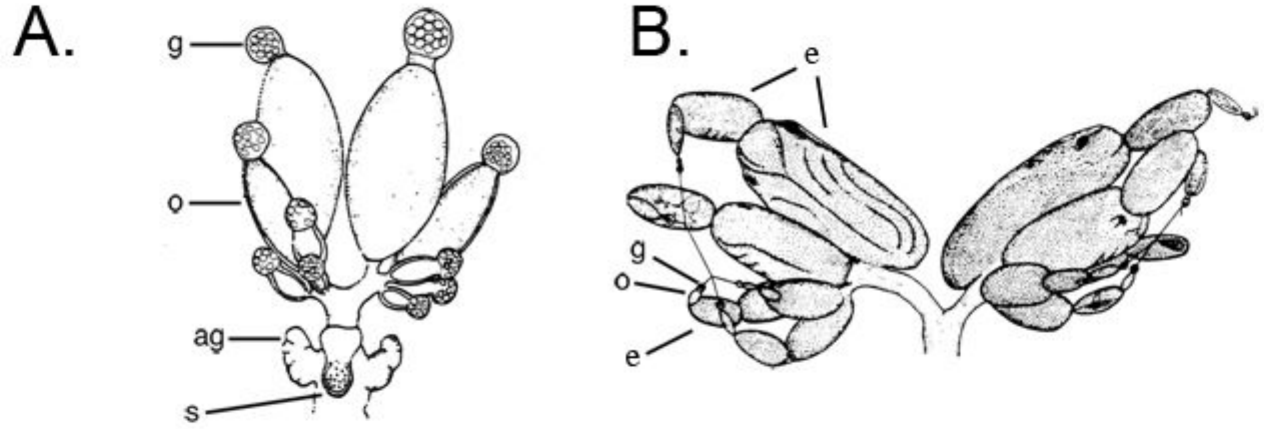
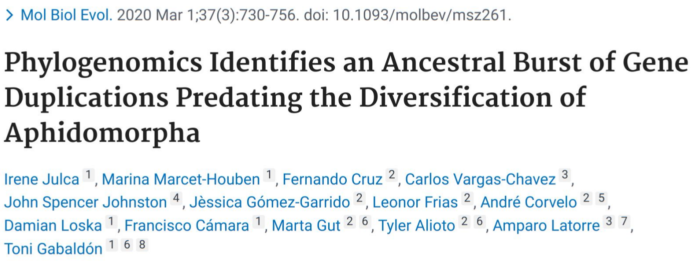


Figure 2. Viviparous and oviparous development. Oviparous (A) and viviparous (B) ovaries differ not only as to whether they possess embryos, accessory glands and spermathecae, but also in the relative size of germaria and oocytes. Abbreviations: g is germarium, o is oocyte, e is viviparous embryo, ag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

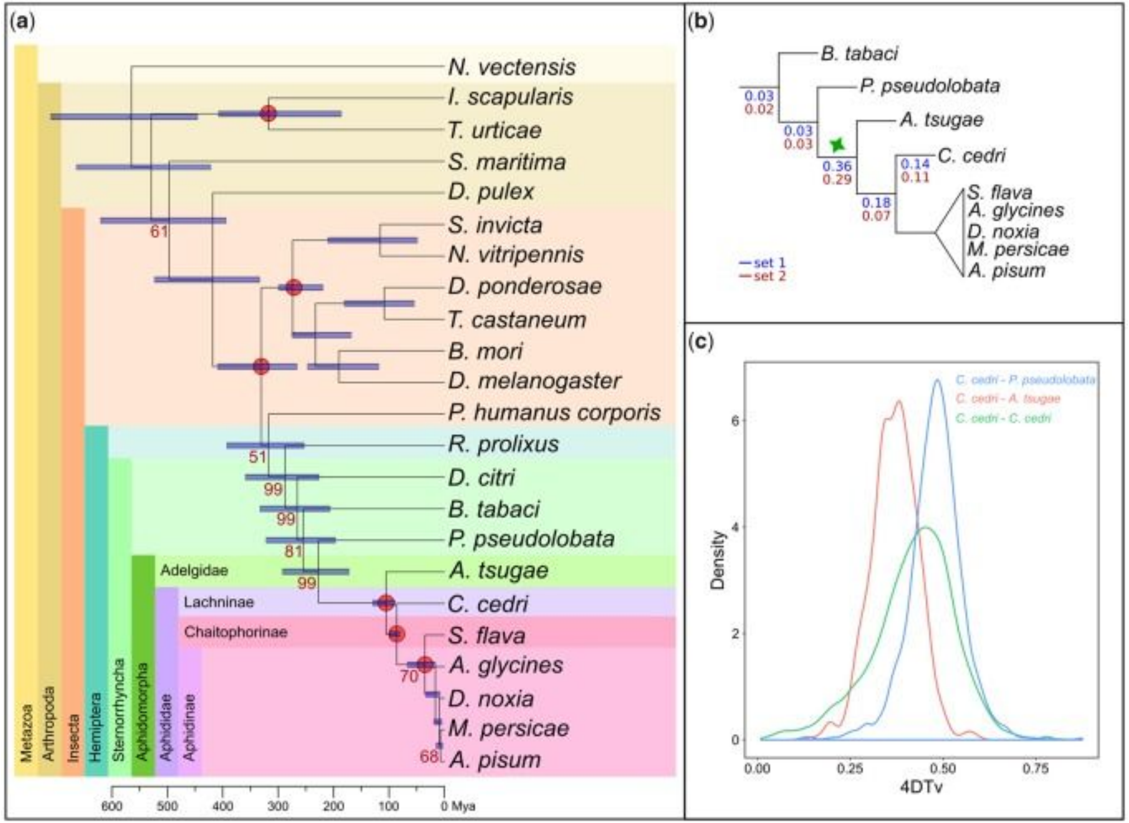
Probable ancestral WGD(s) in the ancestor of aphids



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

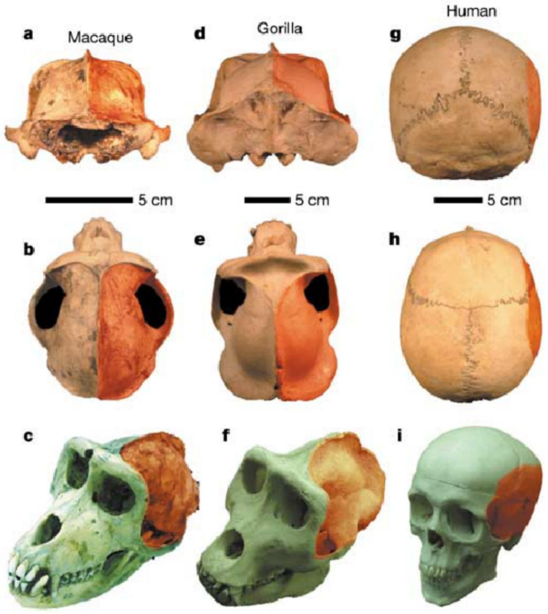
Phylogenomics Identifies an Ancestral Burst of Gene Duplications Predating the Diversification of Aphidomorpha

Irene Julca [1], Marina Marcet-Houben [1], Fernando Cruz [2], Carlos Vargas-Chavez [3], John Spencer Johnston [4], Jèssica Gómez-Garrido [2], Leonor Frias [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]

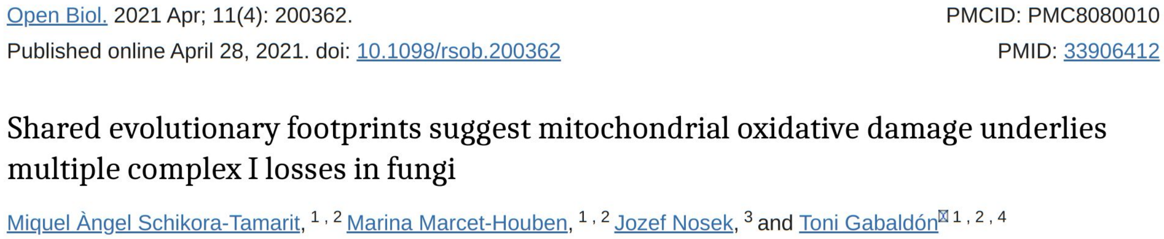


Gene loss also drives adaptation

Loss of *Myh16* associated with cranial enlargement



Stedman et al. (2004)

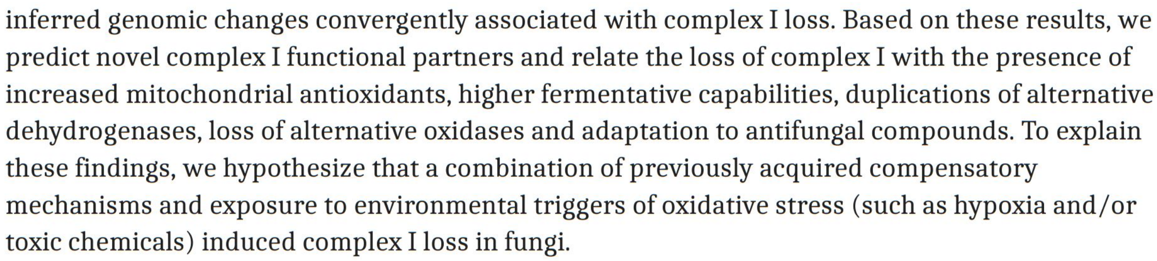


Open Biol. 2021 Apr; 11(4): 200362. PMCID: PMC8080010
Published online April 28, 2021. doi: 10.1098/rsob.200362 PMID: 33906412

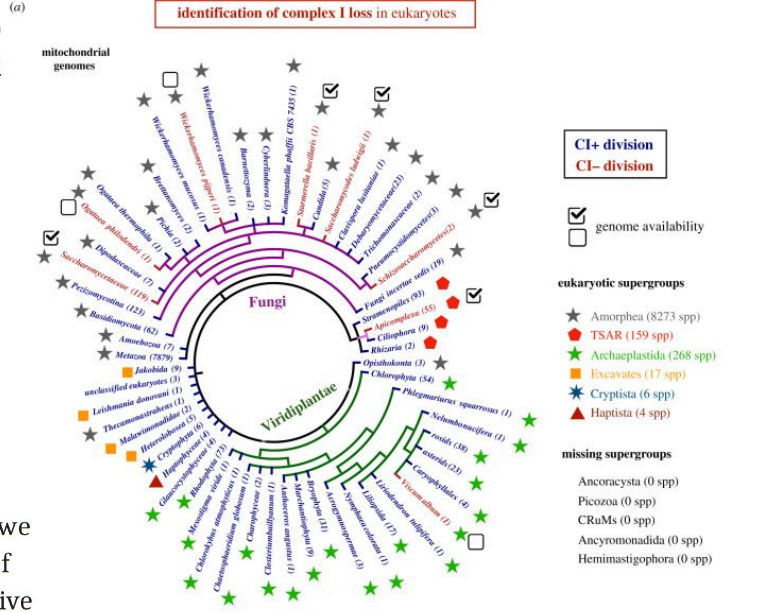
Shared evolutionary footprints suggest mitochondrial oxidative damage underlies multiple complex I losses in fungi

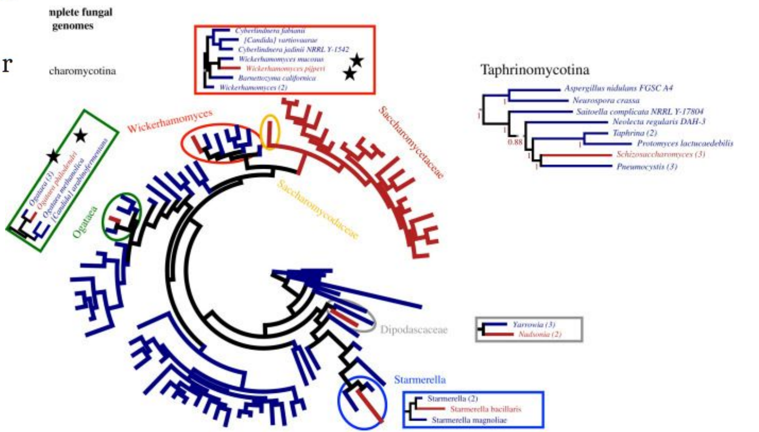
Miquel Àngel Schikora-Tamarit,[1,2] Marina Marcet-Houben,[1,2] Jozef Nosek,[3] and Toni Gabaldón[1,2,4]

- 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.



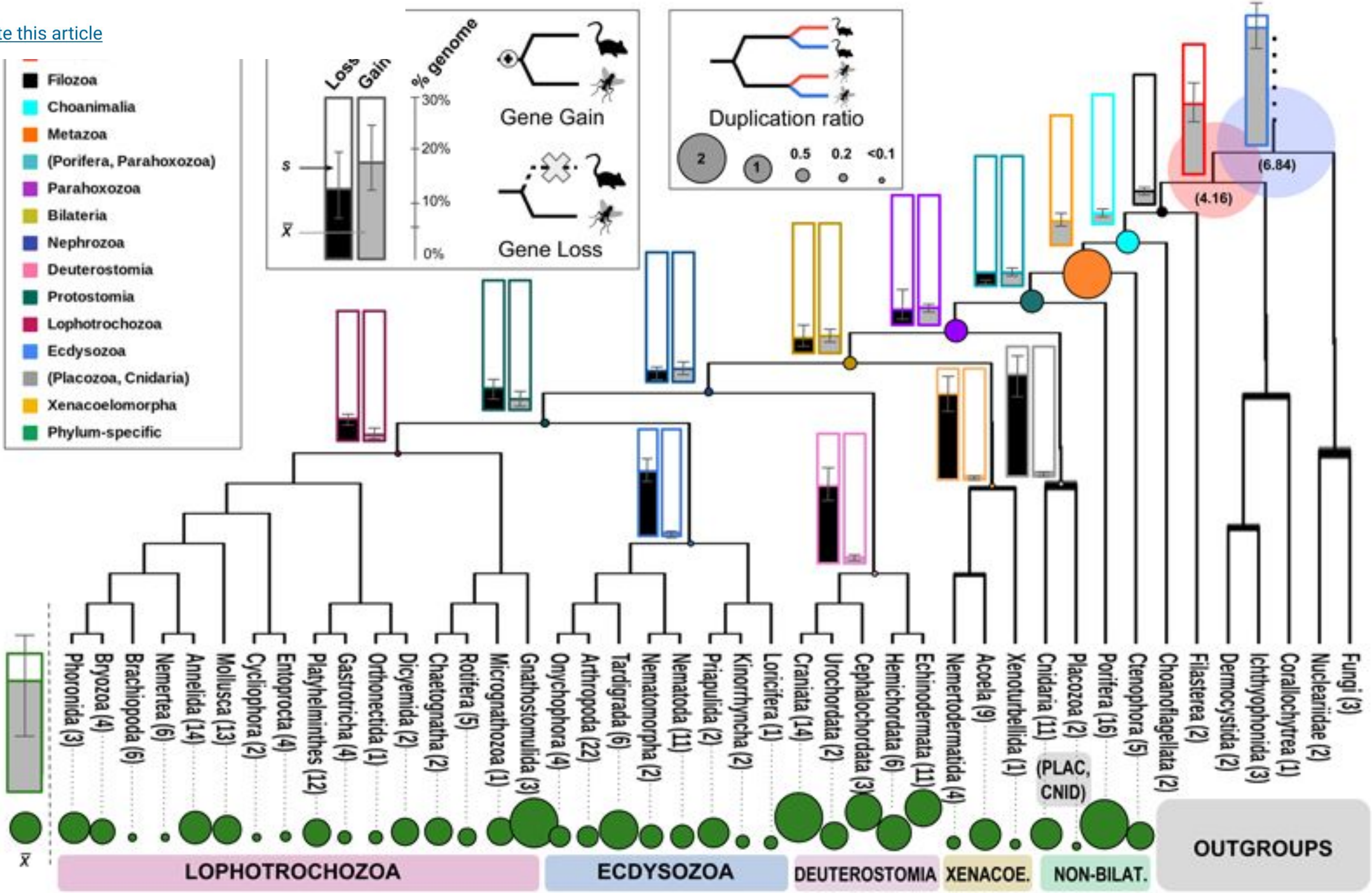


Article | Published: 28 January 2020

Gene gain and loss across the metazoan tree of life

Rosa Fernández & Toni Gabaldón

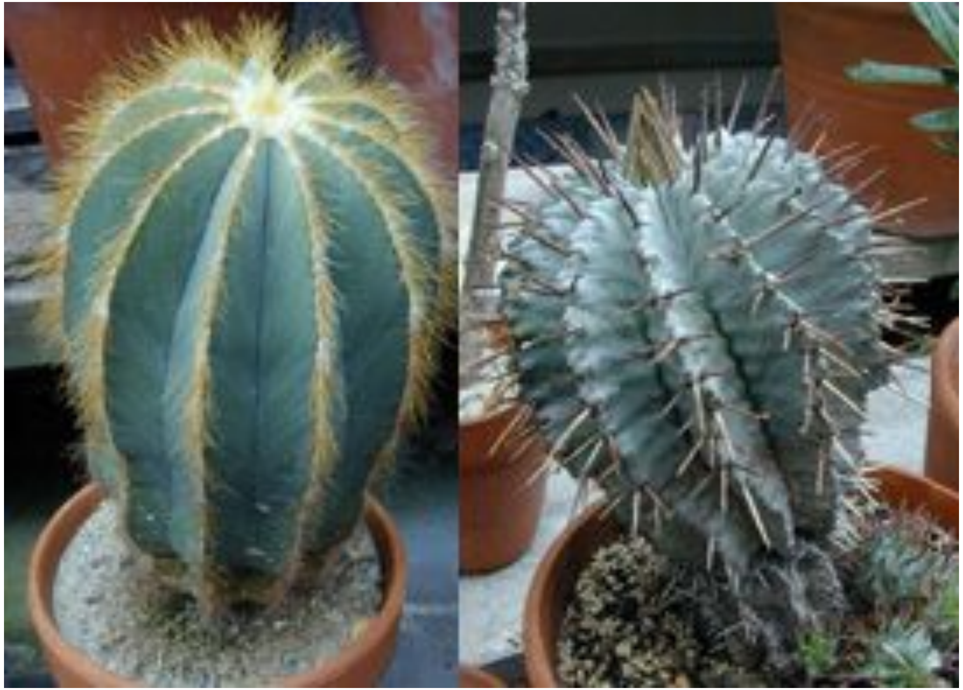
Nature Ecology & Evolution **4**, 524–533 (2020) | Cite this article



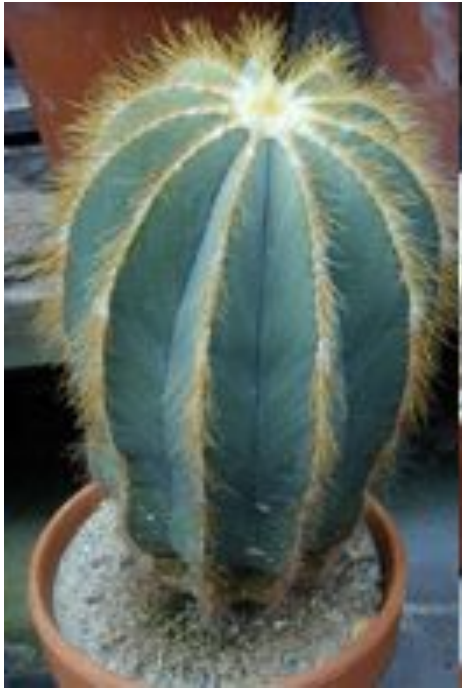
Beyond duplication and loss

- Selection and recombination can explain anomalous gene trees

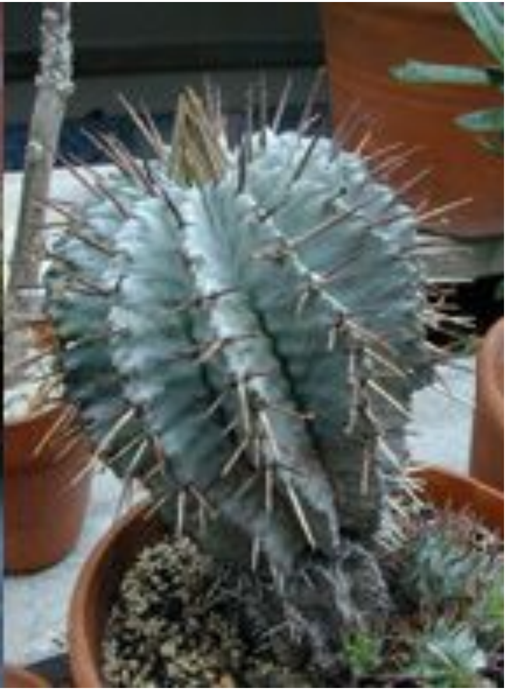
Convergent evolution



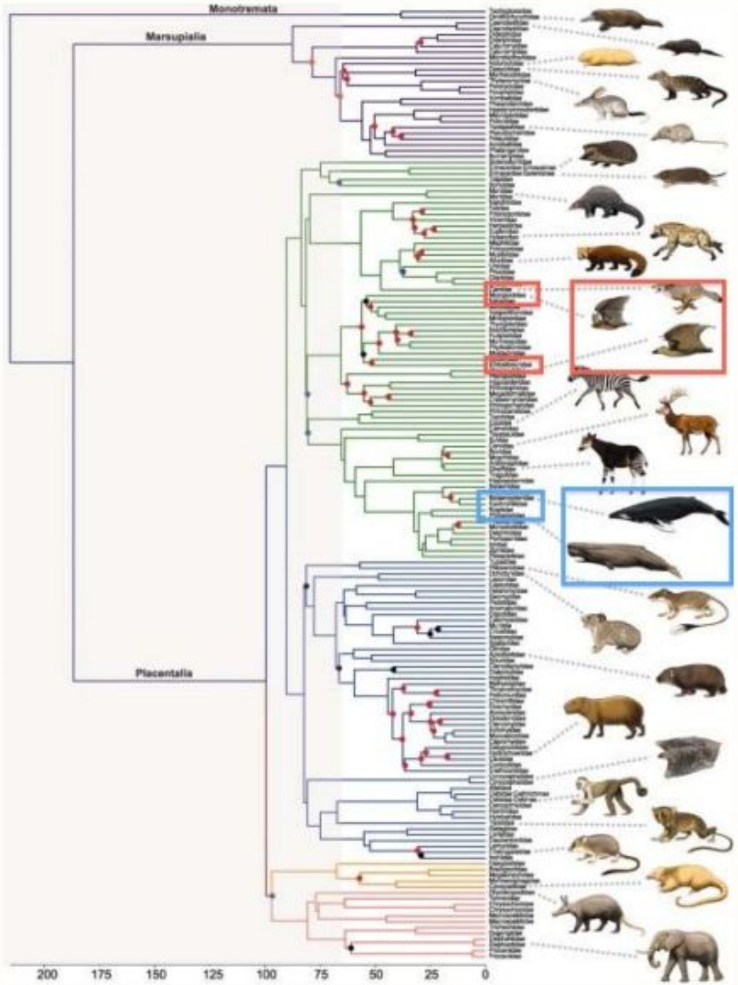
Convergent evolution



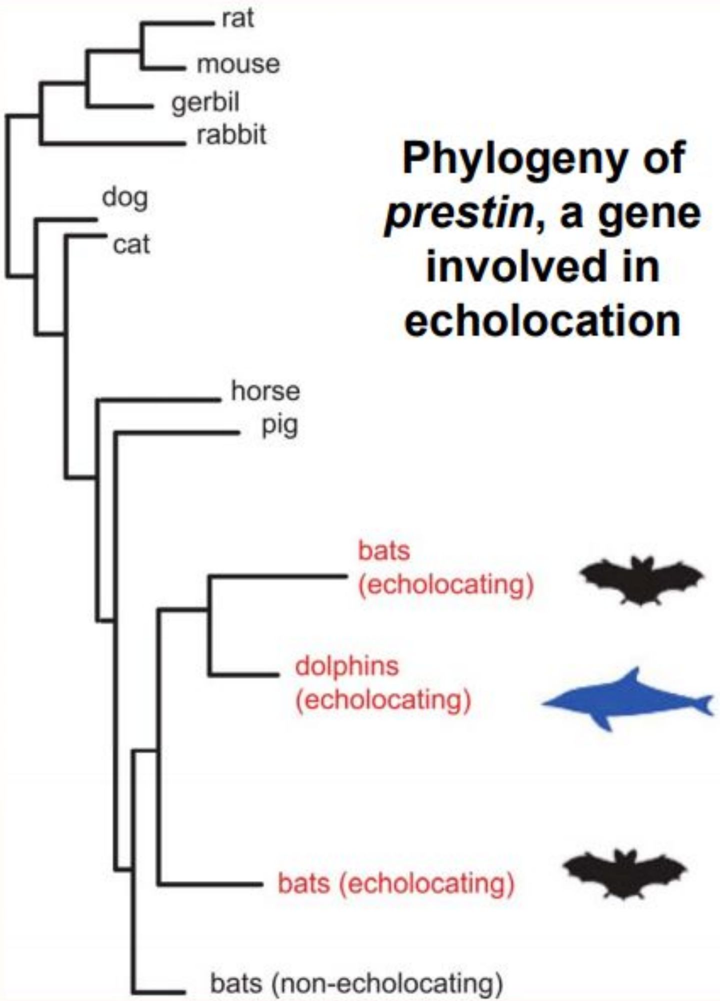
Cactus



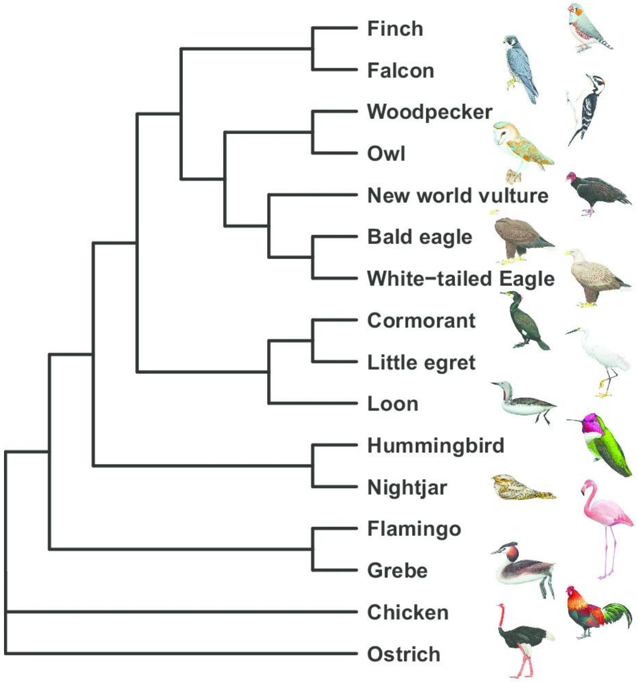
Euphorbia



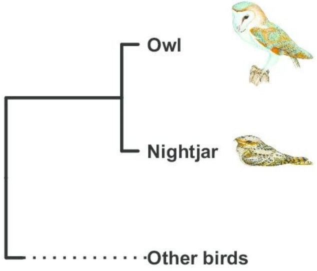
Phylogeny of *prestin*, a gene involved in echolocation



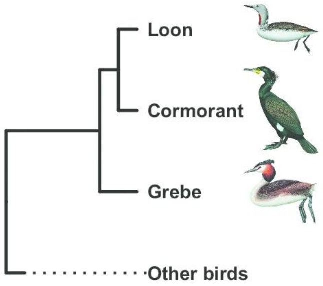
H_0: species phylogeny



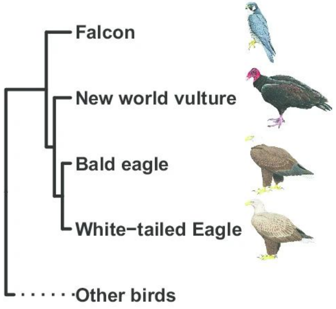
H_{noc} : nocturnal convergence



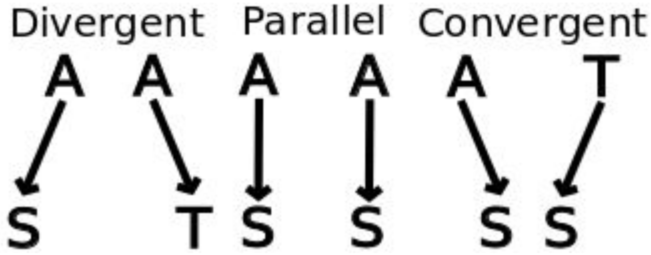
H_{foot} : foot−propelled diving convergence



H_{rap} : diurnal raptorial convergence

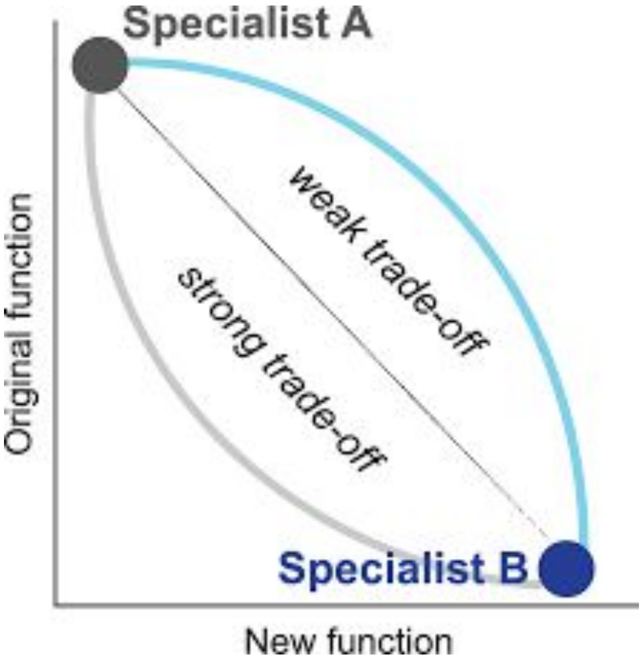


Parallel evolution



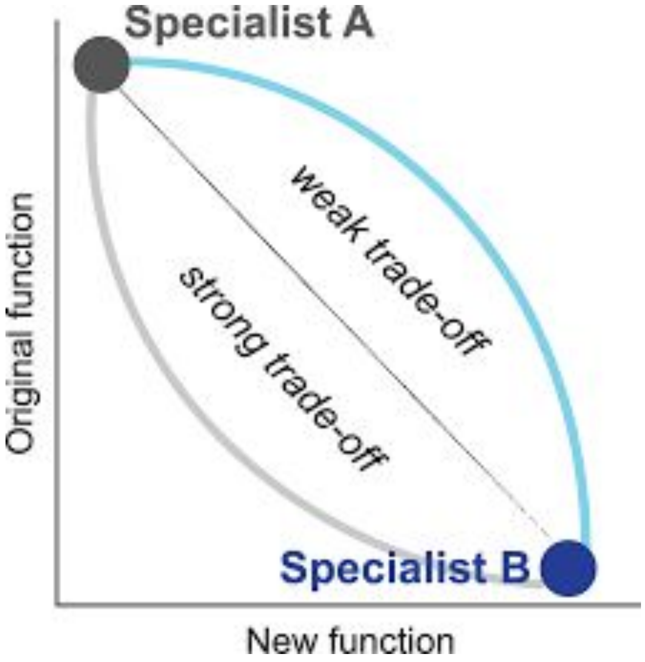
Escape from adaptive conflict

Gene with two functions in conflict



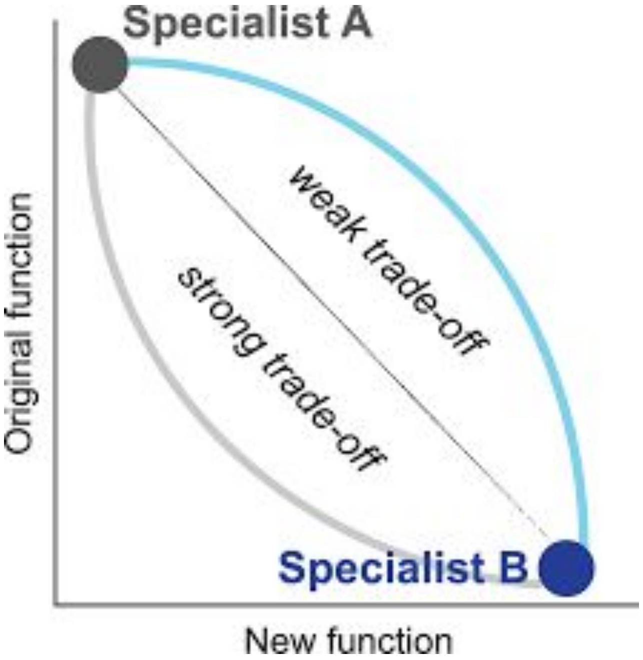
Escape from adaptive conflict

Gene with two functions in conflict



Escape from adaptive conflict

Gene with two functions in conflict

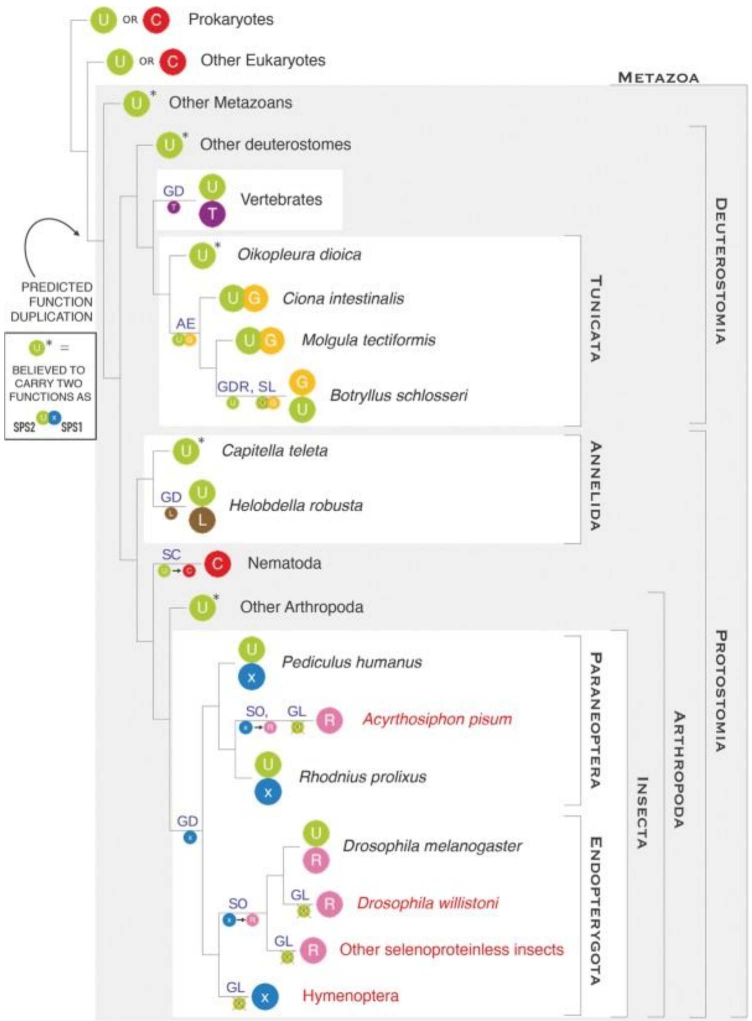


› Genome Res. 2015 Sep;25(9):1256-67. doi: 10.1101/gr.190538.115. Epub 2015 Jul 20.

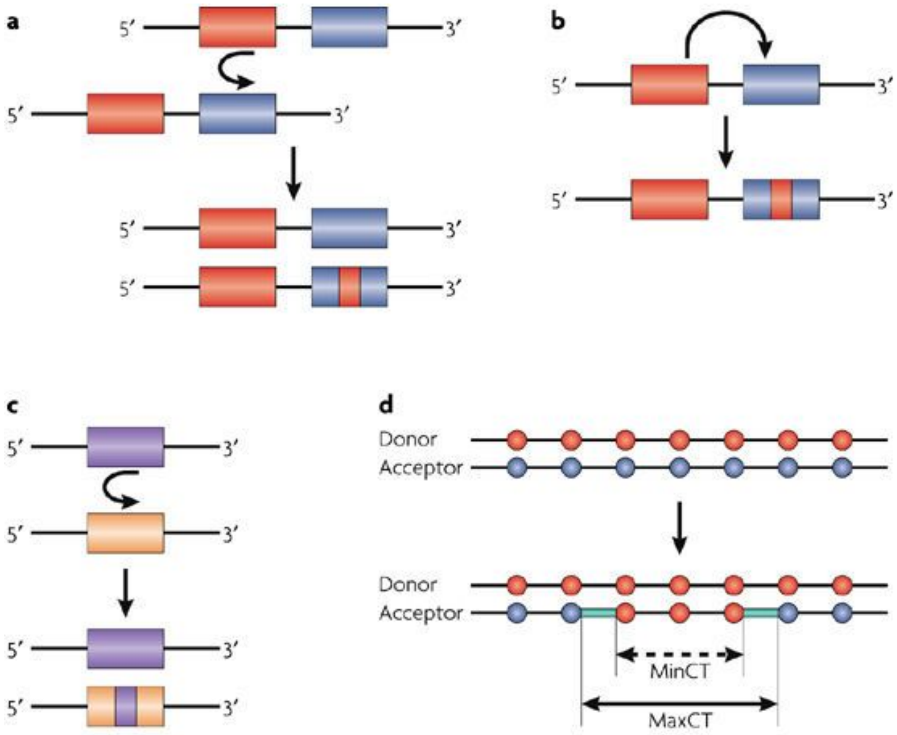
Evolution of selenophosphate synthetases: emergence and relocation of function through independent duplications and recurrent subfunctionalization

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

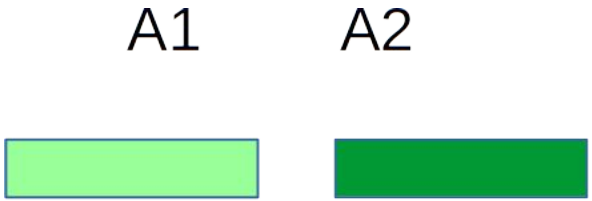


Concerted evolution (gene conversion)

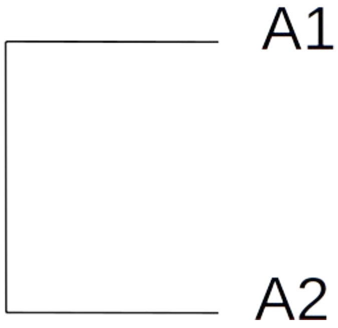


Nature Reviews | Genetics

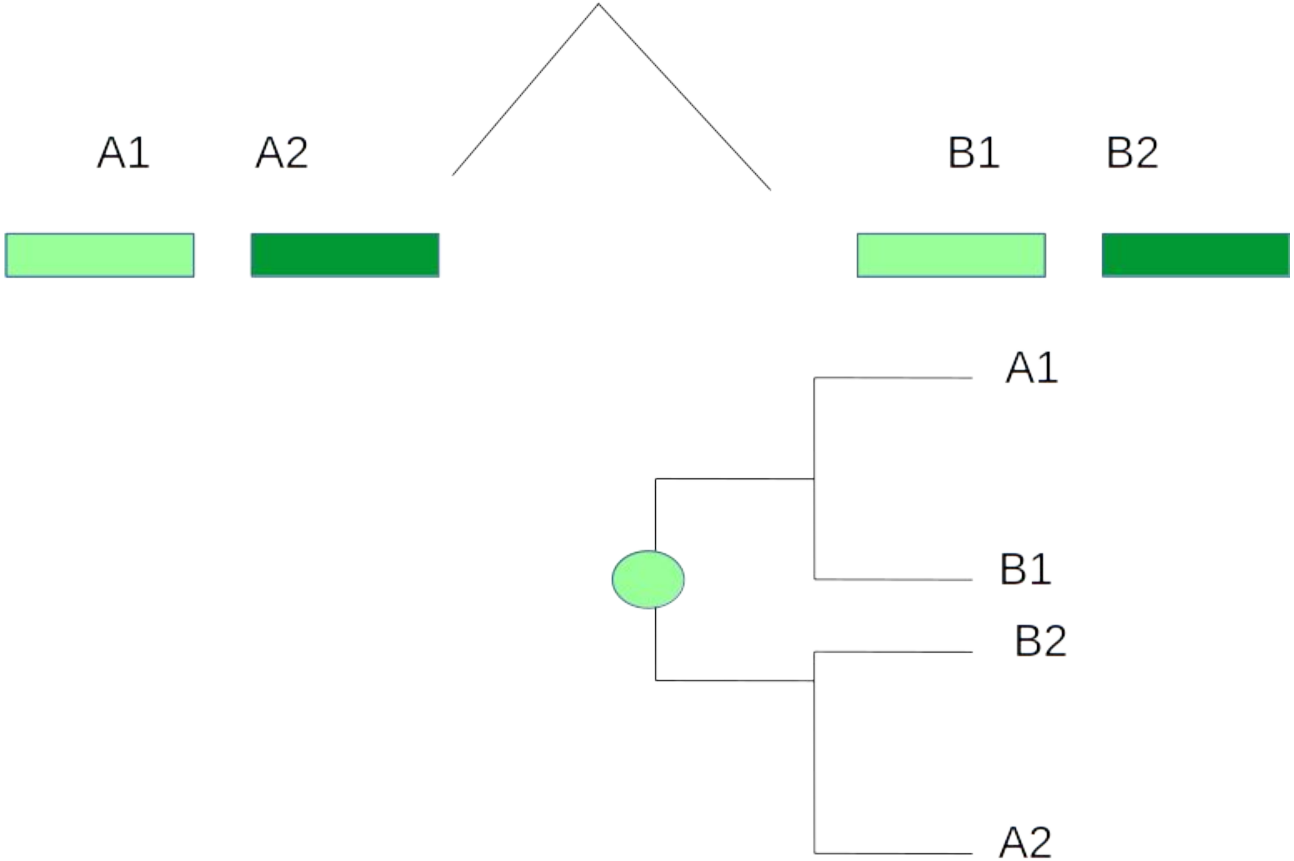
Duplication



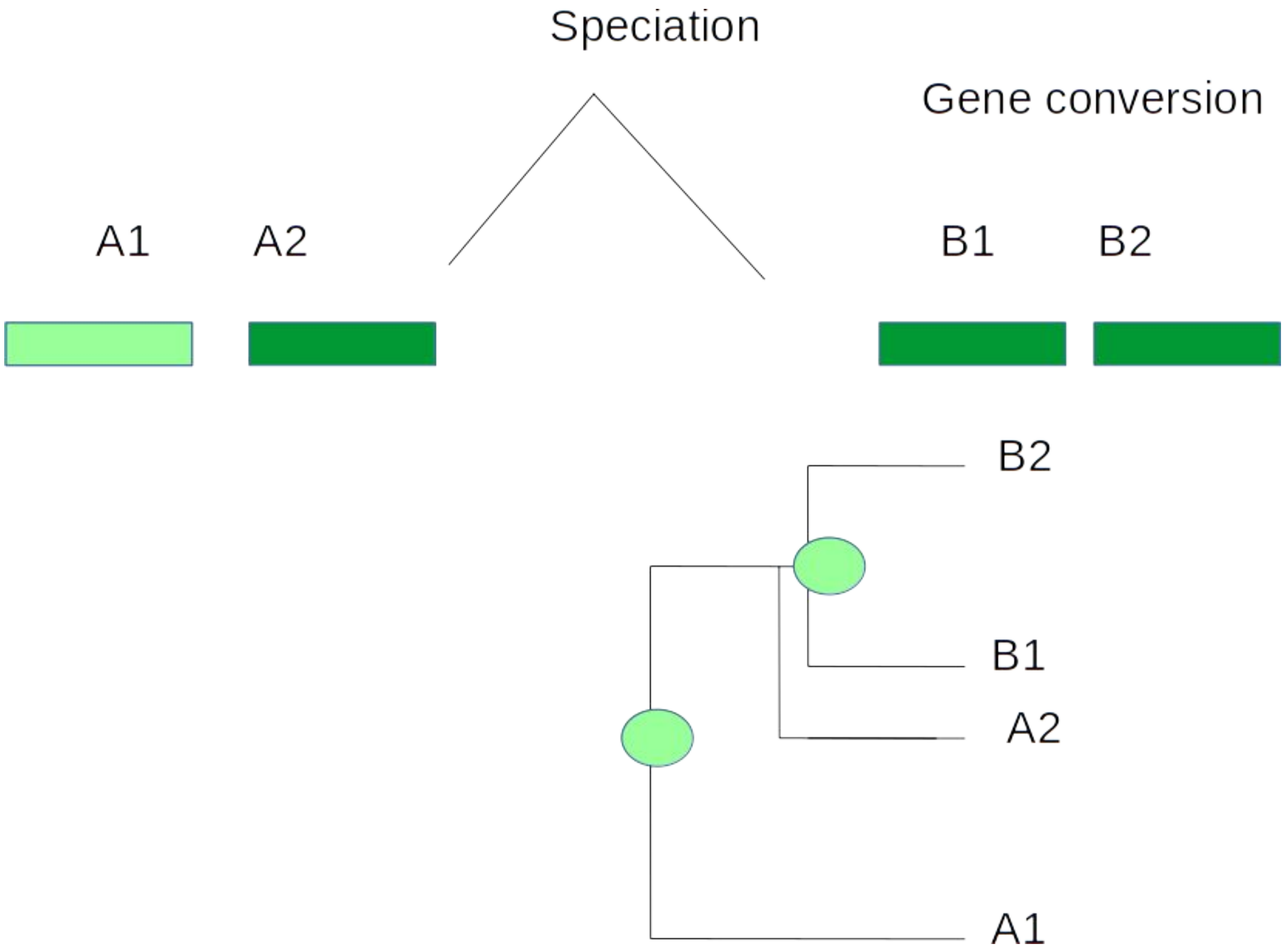
Divergence



Speciation



Orthologs are closer than ancient paralogs



Speciation

Gene conversion

A1

A2

B1

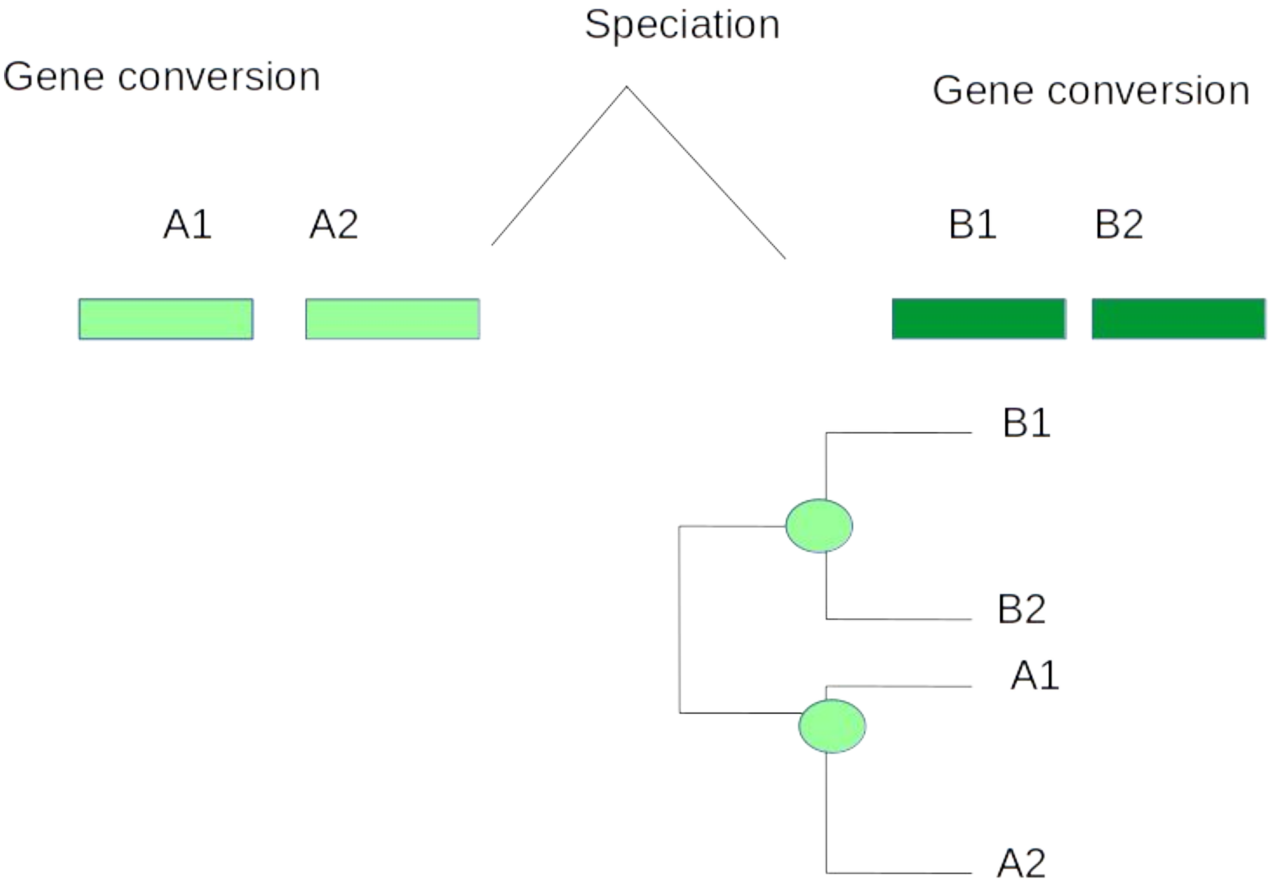
B2

B2

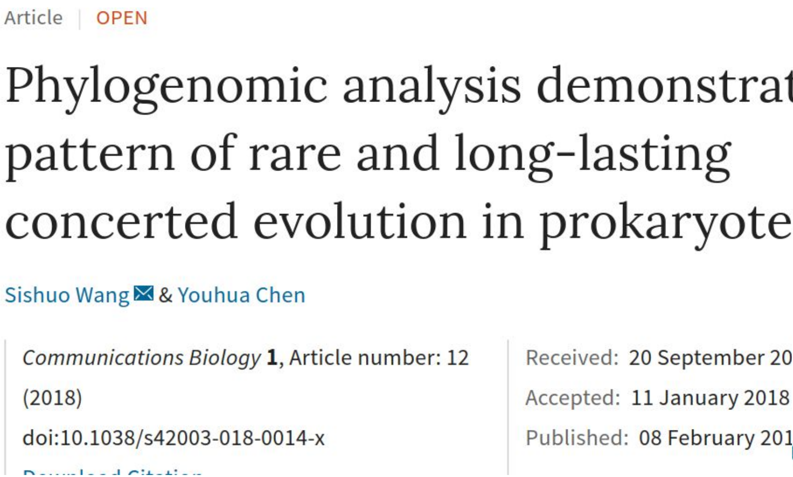
B1

A2

A1



Paralogs are closer than orthologs, apparent parallel duplication



Article | OPEN

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)

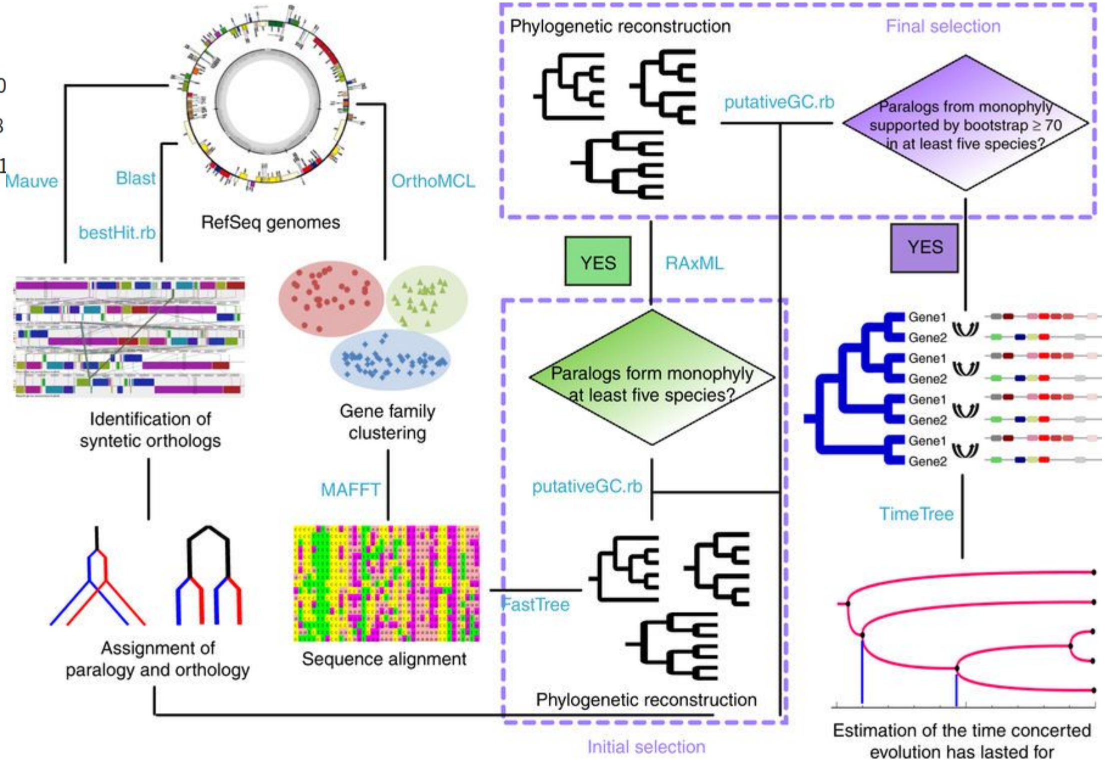
doi:10.1038/s42003-018-0014-x

Download Citation

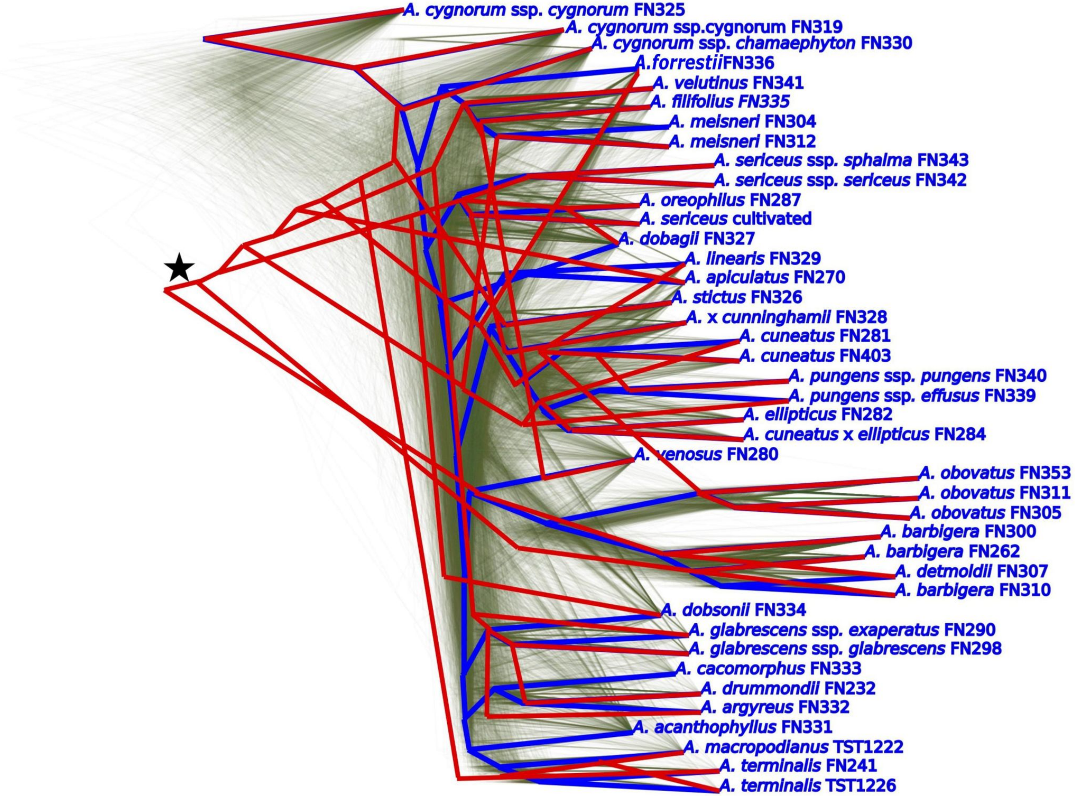
Received: 20 September 20

Accepted: 11 January 2018

Published: 08 February 201


RefSeq genomes
Mauve
Blast
bestHit.rb
OrthoMCL
Identification of syntetic orthologs
Gene family clustering
MAFFT
Assignment of paralogy and orthology
Sequence alignment
FastTree
Phylogenetic reconstruction
Initial selection
putativeGC.rb
Paralogs form monophyly at least five species?
YES
RAxML
Phylogenetic reconstruction
putativeGC.rb
Final selection
Paralogs from monophyly supported by bootstrap ≥ 70 in at least five species?
YES
Gene1
Gene2
Gene1
Gene2
Gene1
Gene2
Gene1
Gene2
TimeTree
Estimation of the time concerted evolution has lasted for

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