Functional and Comparative Genomics

ESCI upf.
School of International Studies

Comparative and Functional Genomics

Session 4
Phylogenomics

ESCI upf.
School of International Studies

Phylogenomics

Species tree reconstruction.

Genome-wide phylogenetic analysis (phylome).

Gene tree vs species tree.

Non-vertical processes of evolution

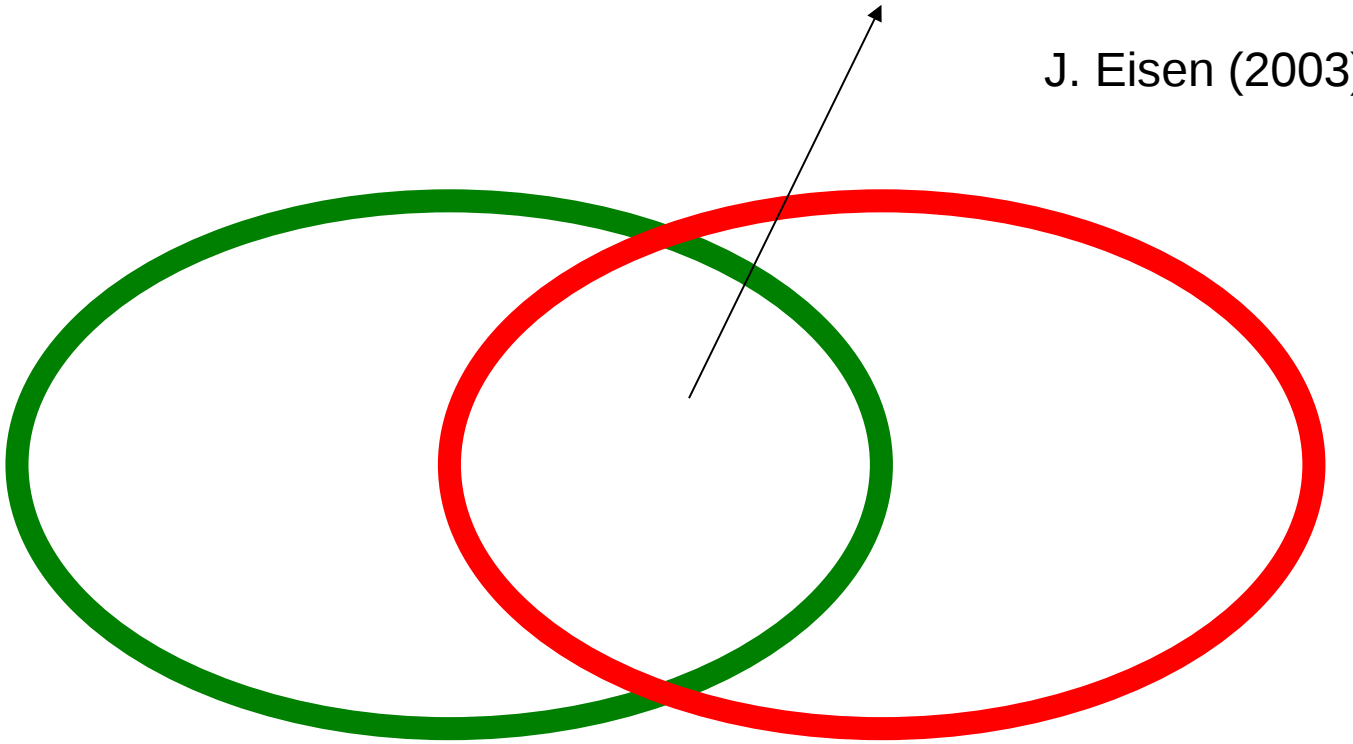
horizontal gene transfer.

Hybridization

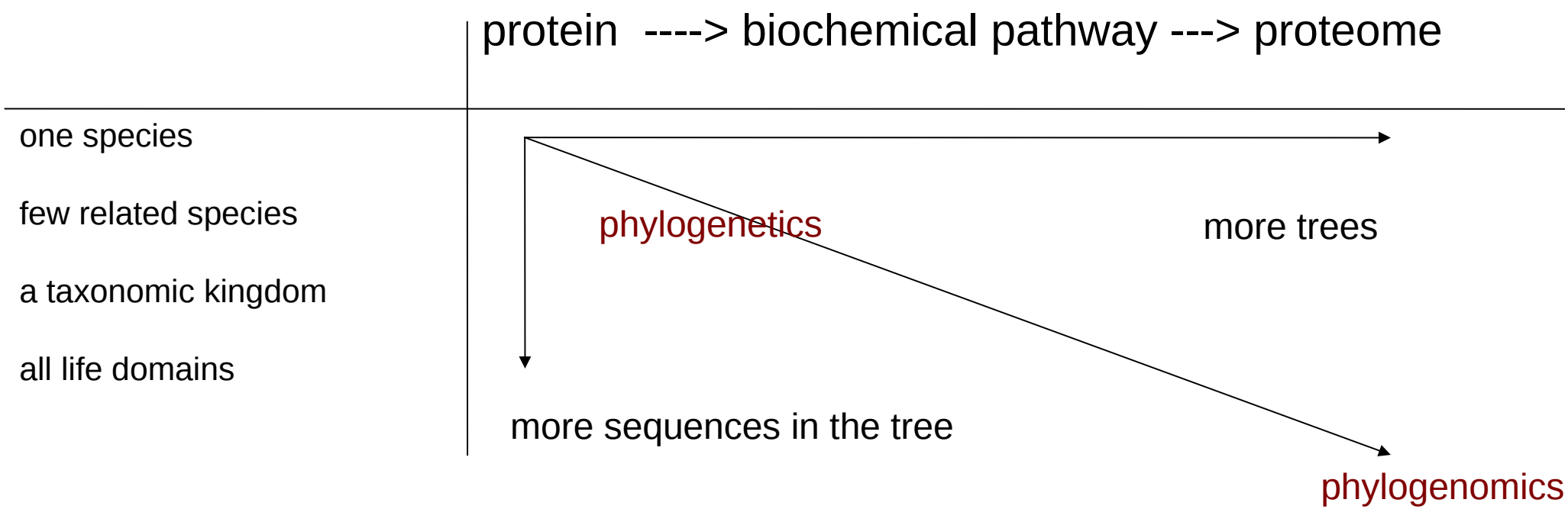
Whole genome duplication.

Phylogenomics is the intersection between Genomics and Evolution.

J. Eisen (2003)



That is, looking at genomes from an evolutionary perspective, often using **phylogenetics**


protein ----> biochemical pathway ---> proteome
one species
few related species
a taxonomic kingdom
all life domains
phylogenetics
more trees
more sequences in the tree
phylogenomics

Phylogenomics:

Opportunities:

- necessary to provide an evolutionary framework to the deluge of data generated

- useful to obtain biological knowledge from sequence data

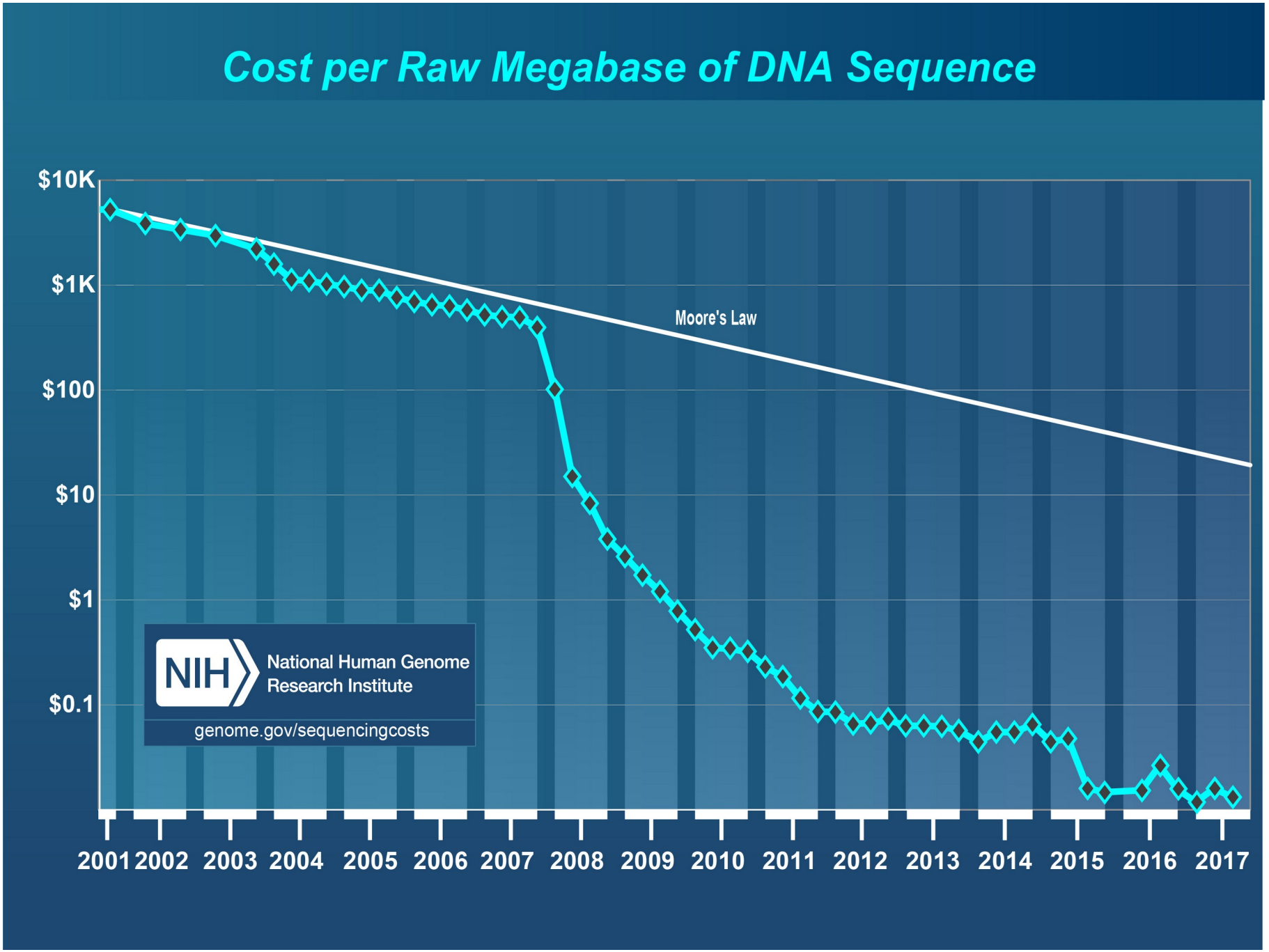
- The more data, the more powerful

Challenges:

- Computationally demanding

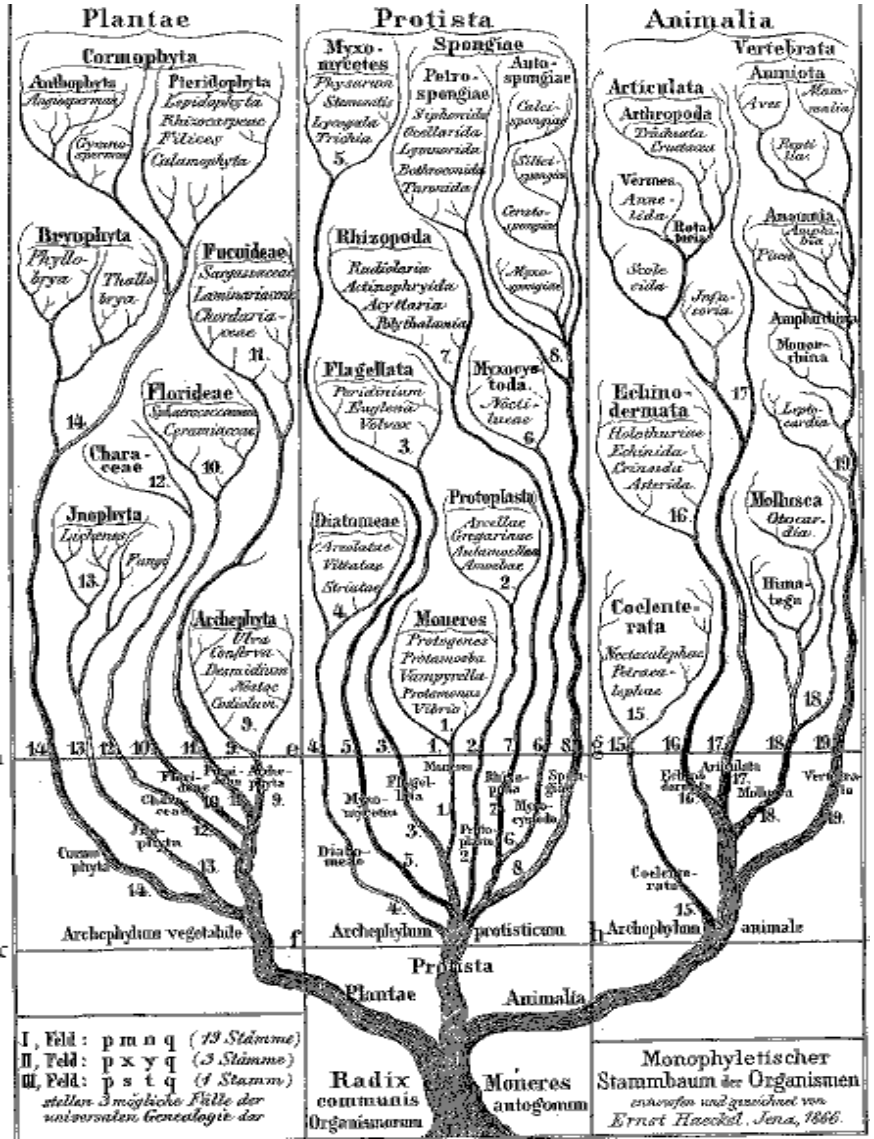
- Need for automation

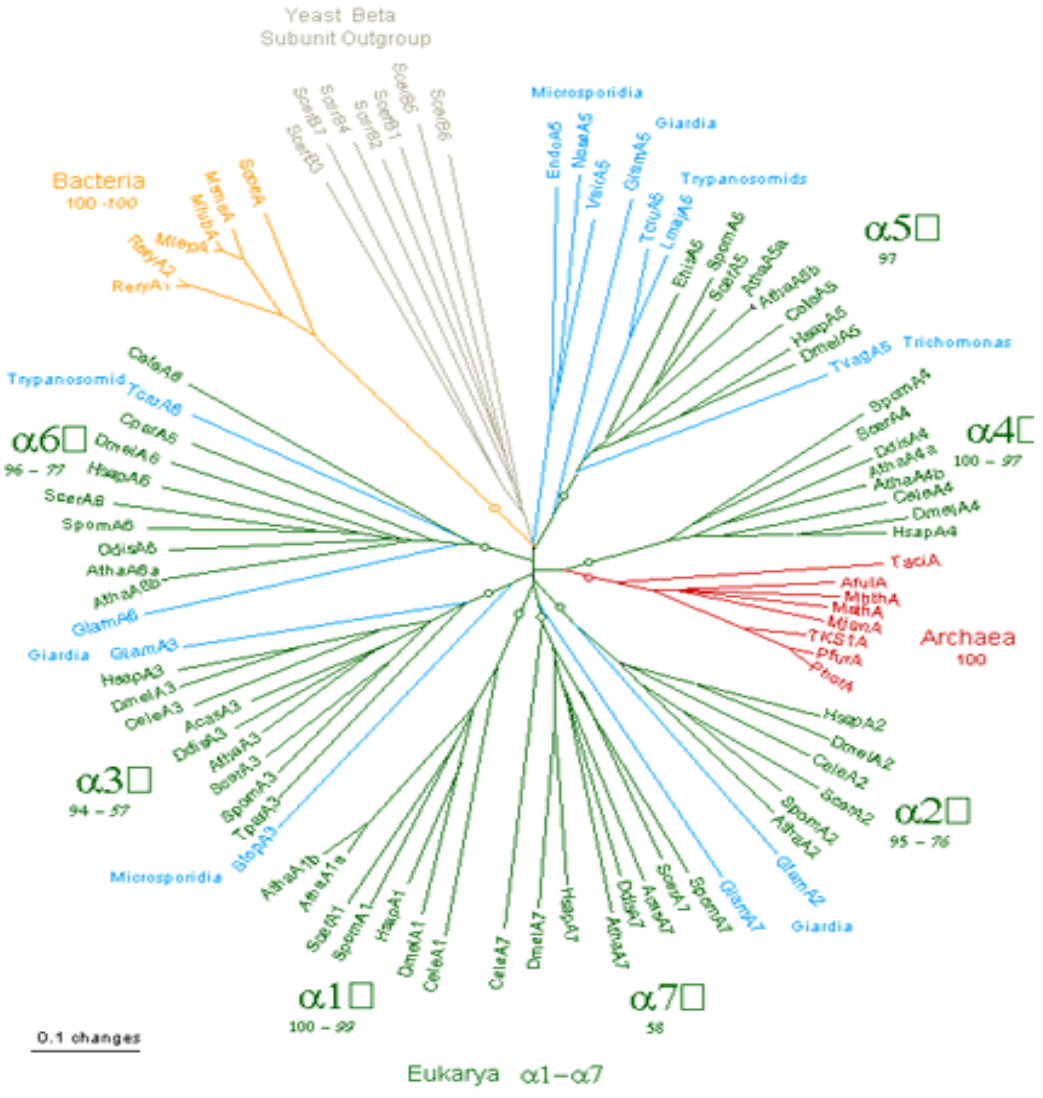
- Need for proper scalability, (e.g alignments, from a certain point, the more data, the more noisy)

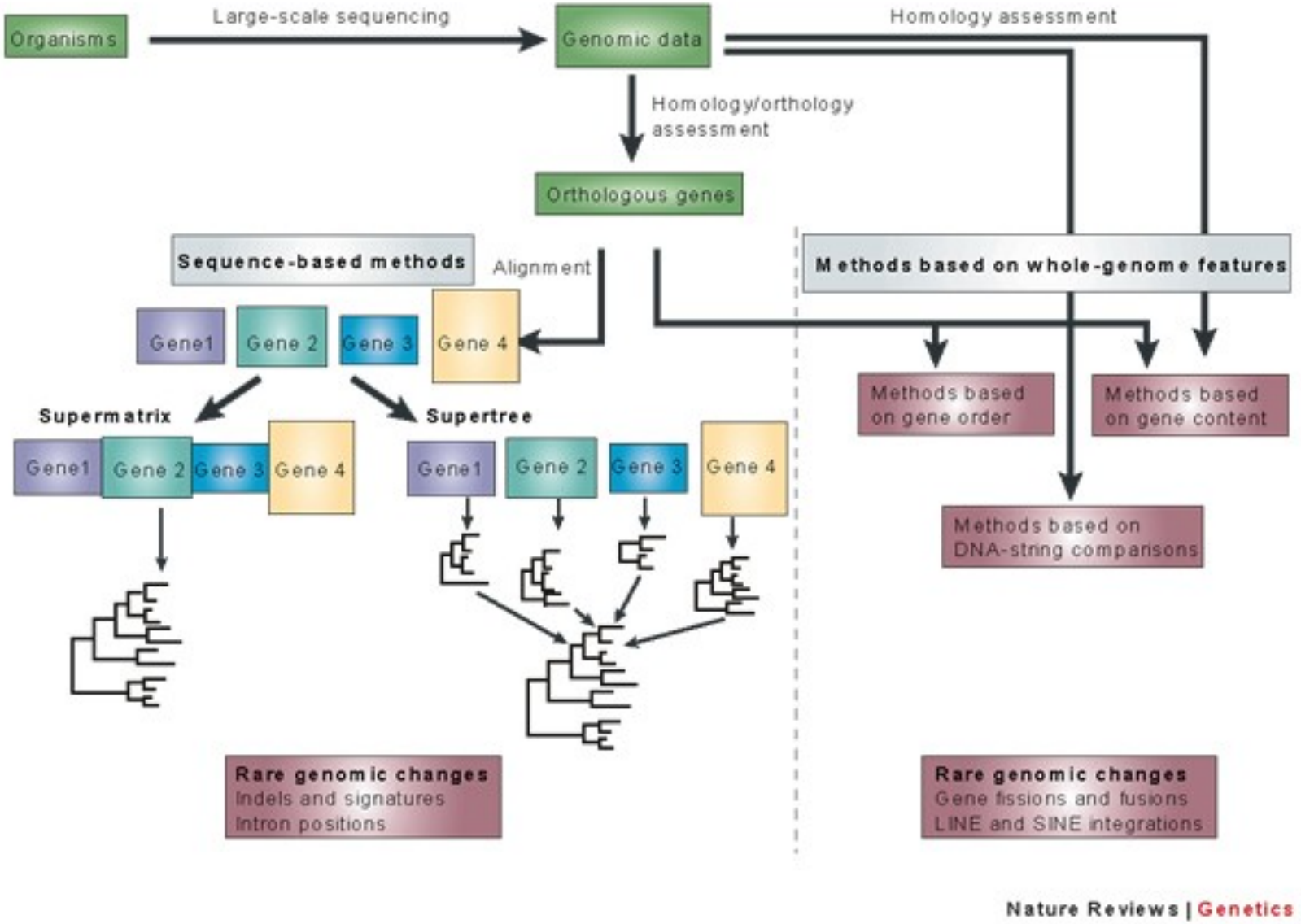


https://www.genome.gov/sequencingcostsdata/

Phylogenomics and species tree reconstruction

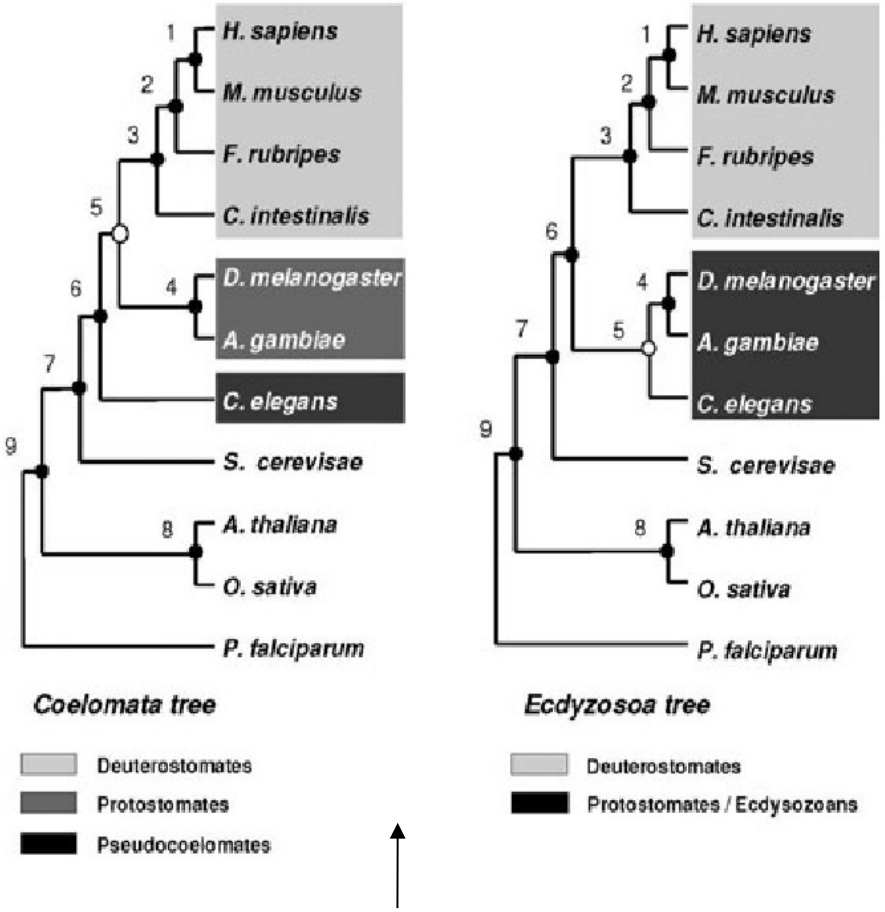




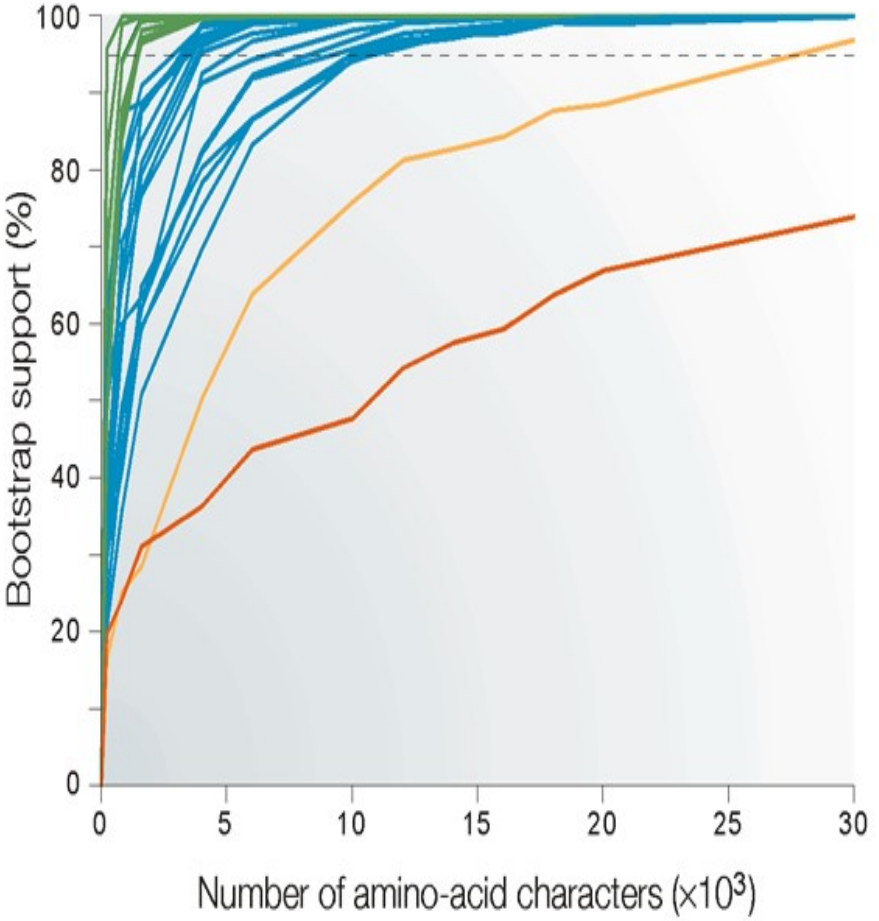

Organisms
Large-scale sequencing
Genomic data
Homology assessment
Homology/orthology assessment
Orthologous genes
Sequence-based methods
Alignment
Methods based on whole-genome features
Gene1
Gene 2
Gene 3
Gene 4
Supermatrix
Supertree
Methods based on gene order
Methods based on gene content
Methods based on DNA-string comparisons
Rare genomic changes
Indels and signatures
Intron positions
Rare genomic changes
Gene fissions and fusions
LINE and SINE integrations
Nature Reviews | Genetics

Delsuc. et al.

concatenated alignments allow using more sites (information) to resolve a phylogeny



Dopazo H. et al. (2004)

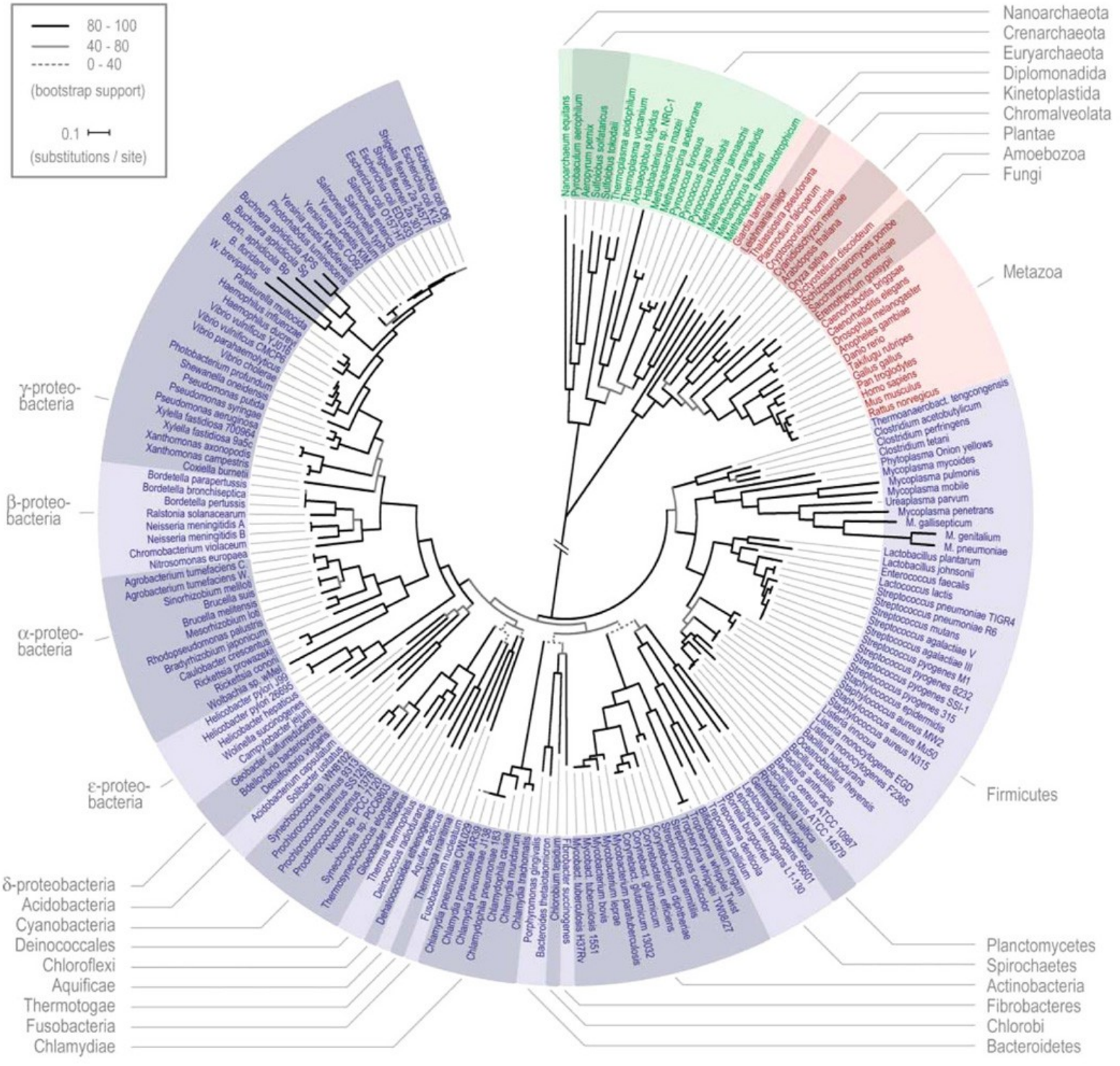


Delsuc et. al (2005)

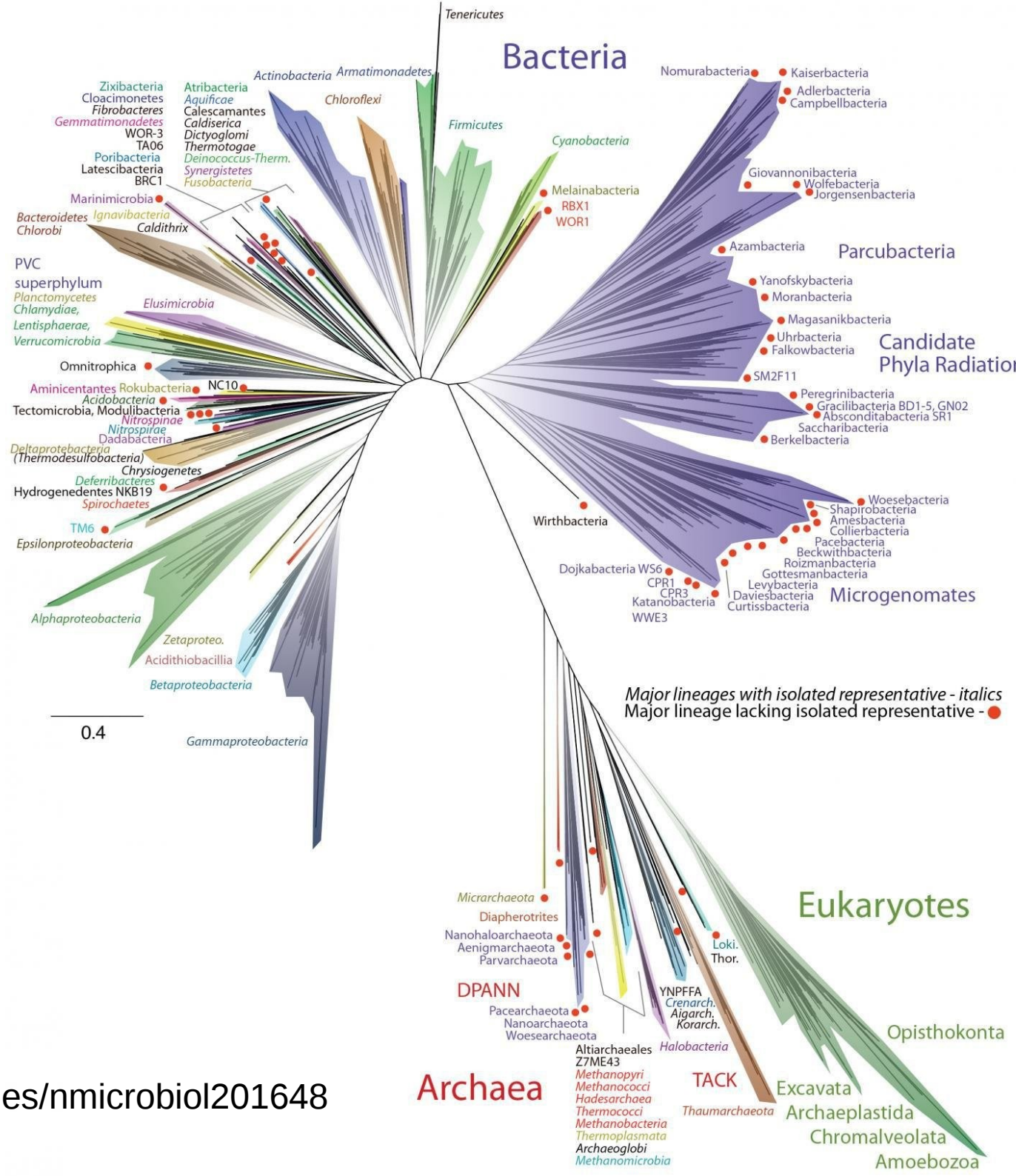
Nature Reviews | Genetics

Cicarelli (2005)

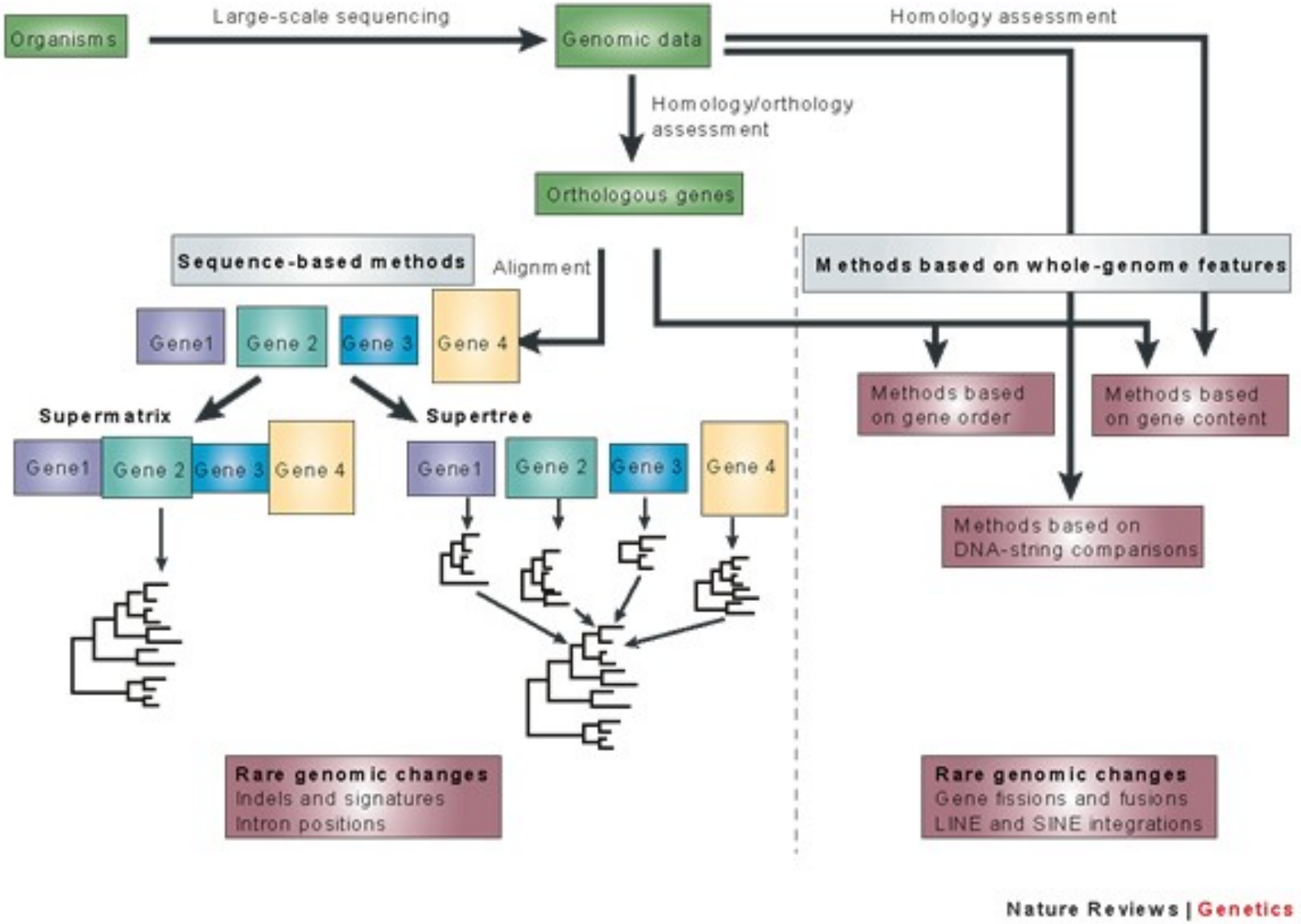
195 genomes
31 genes



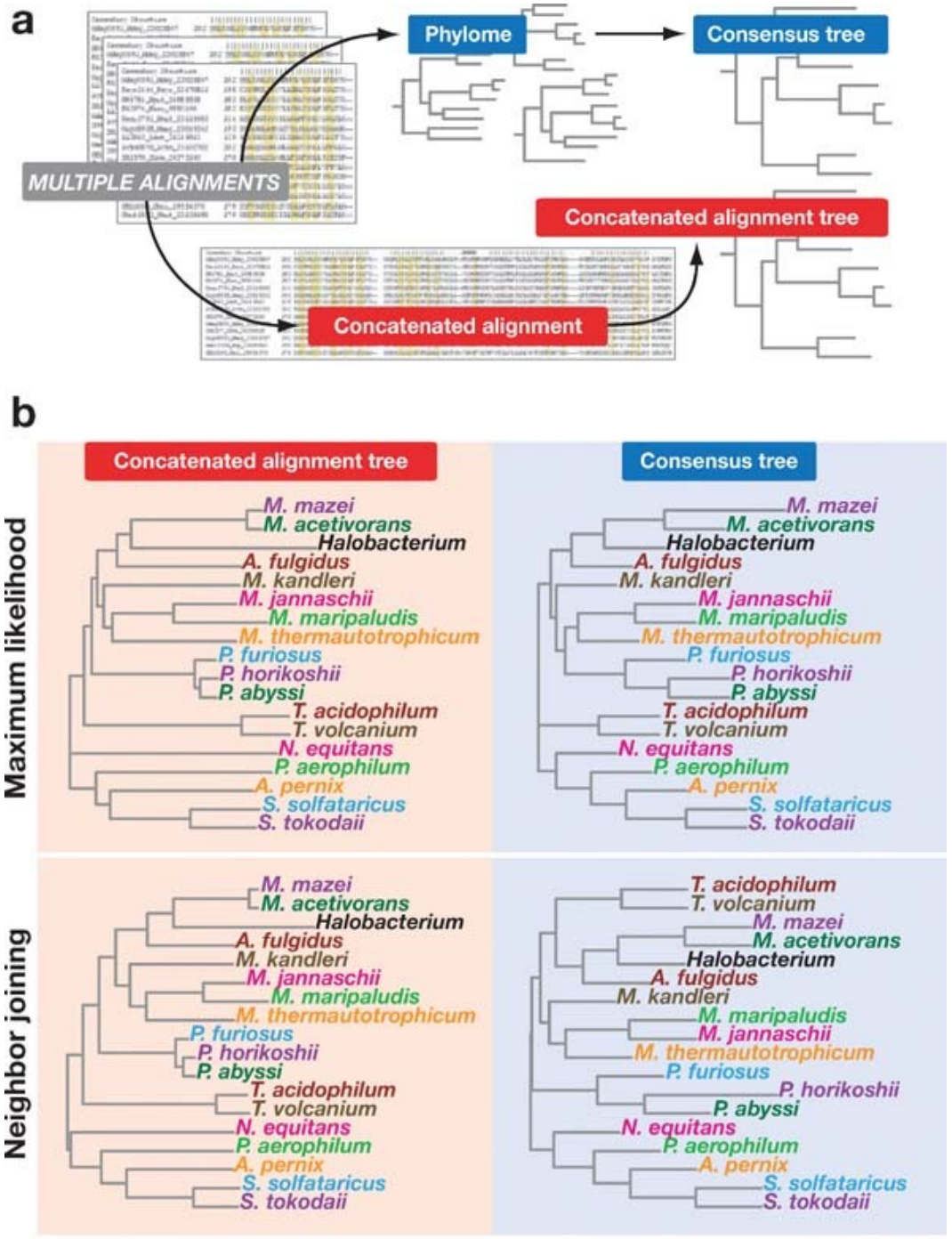
30,437 genomes
16 ribosomal genes



Hug. et. al. (2016)
https://www.nature.com/articles/nmicrobiol201648

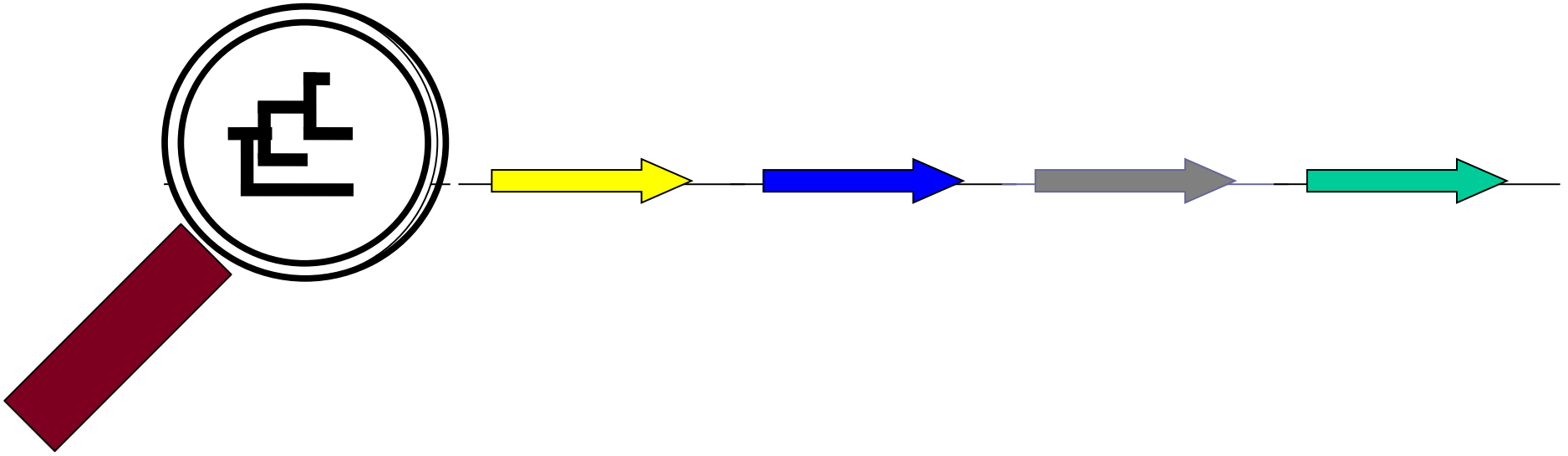

Organisms
Large-scale sequencing
Genomic data
Homology assessment
Homology/orthology assessment
Orthologous genes
Sequence-based methods
Alignment
Methods based on whole-genome features
Gene1
Gene 2
Gene 3
Gene 4
Supermatrix
Supertree
Methods based on gene order
Methods based on gene content
Methods based on DNA-string comparisons
Rare genomic changes
Indels and signatures
Intron positions
Rare genomic changes
Gene fissions and fusions
LINE and SINE integrations
Nature Reviews | Genetics

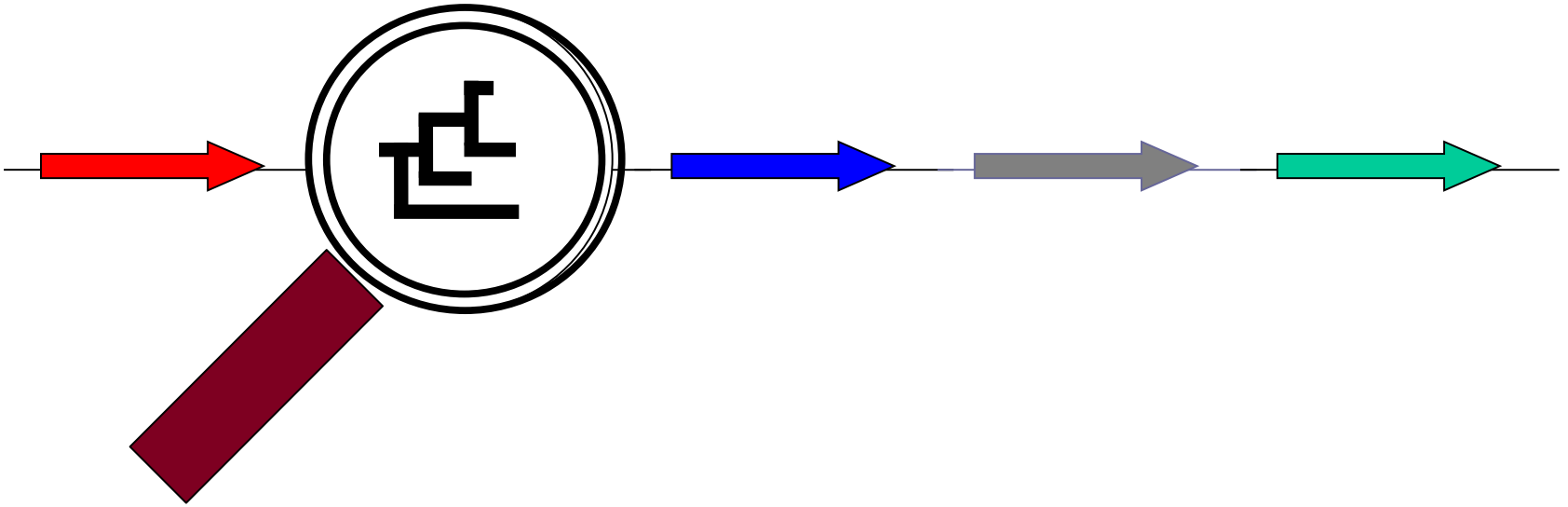
Delsuc. et al.

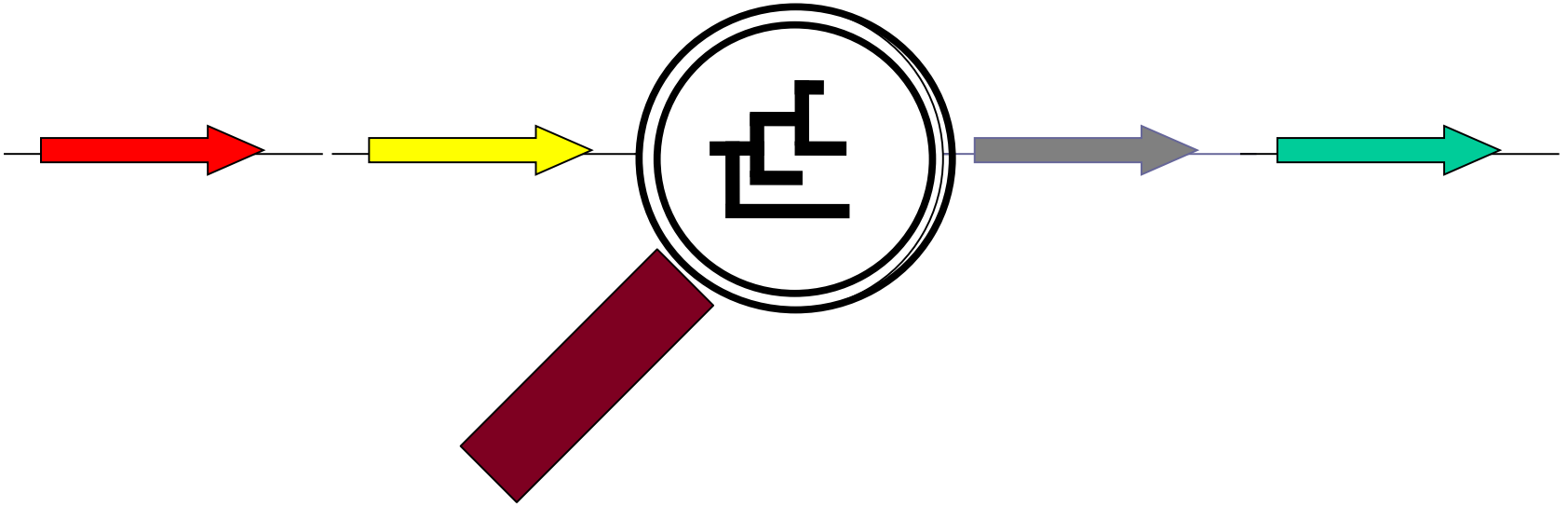


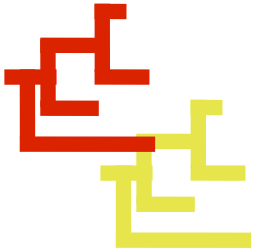
Snel B, et al. 2005.
Annu. Rev. Microbiol. 59:191–209

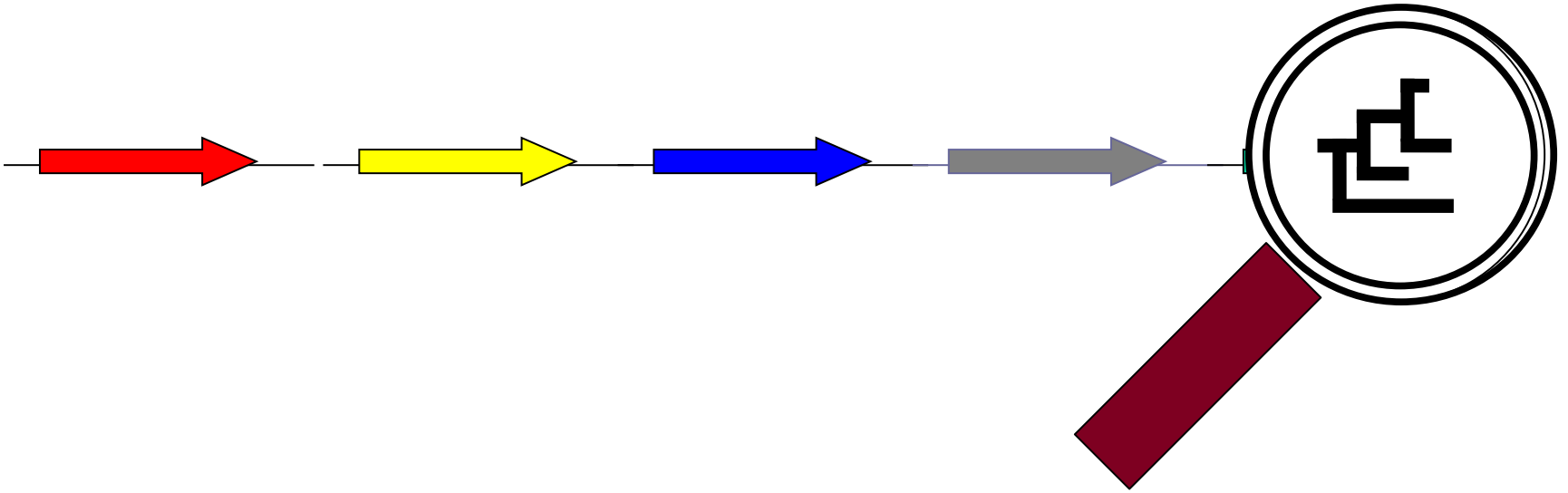
Genome-wide phylogenetic analysis (phylome).

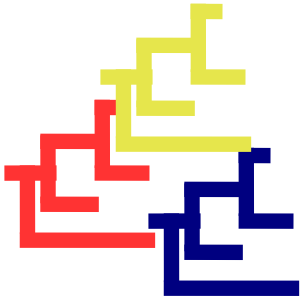










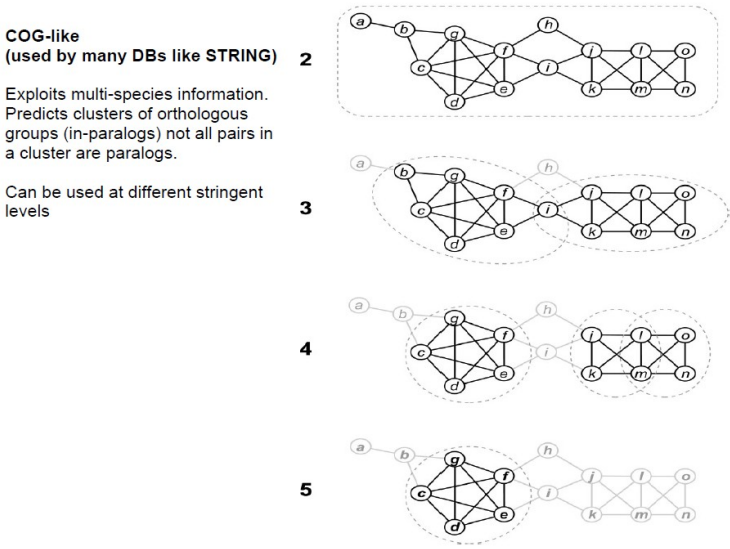


Phylome:

Complete collection of evolutionary histories of all genes encoded in a given genome

Phylome reconstruction

A) Family-based approach: first build families, then reconstruct one tree per family (Ensembl, EggNOG..)



1- Build families
2- Alignment per family
3- Phylogenetic reconstruction per alignment

B) Gene-based approach: sequentially use each gene of interest as a seed to build a gene tree.
(PhylomeDB)

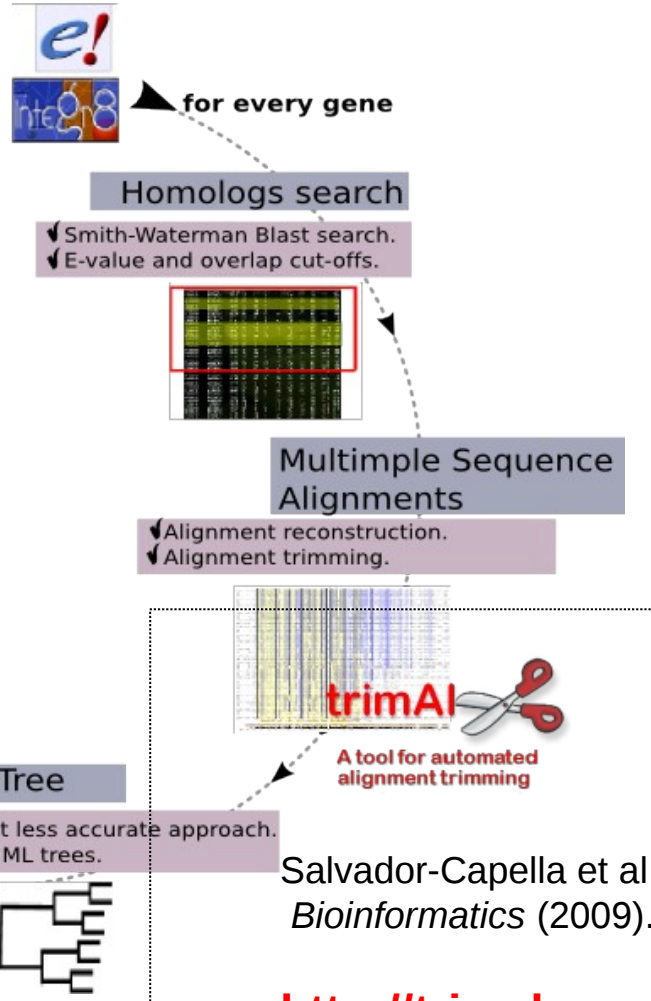
1- Search homologs
2- Alignment per gene + homologs
3- Phylogenetic reconstruction per alignment

Our pipeline:

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

PHYLOME

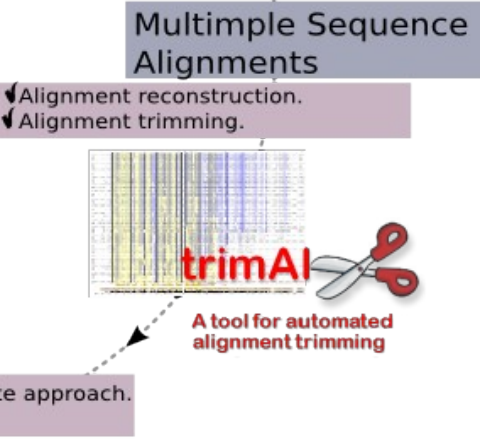
ETE: Environment for Tree Exploration
ete.cgenomics.org

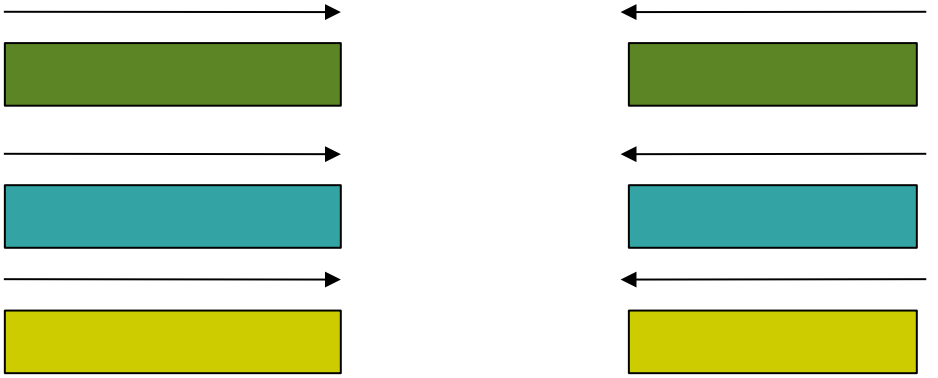


Salvador-Capella et al
Bioinformatics (2009).

http://trimal.cgenomics.org

Pipeline described in Huerta-Cepas et al NAR (2011)



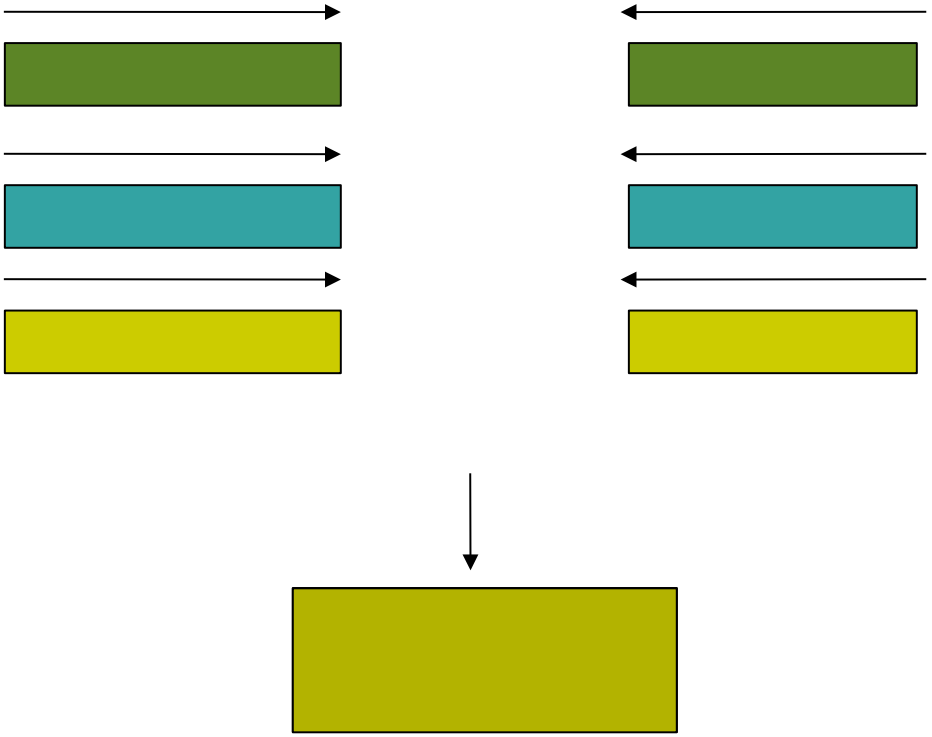


The set of homologous Sequences are aligned by 3 different aligners in forward and reverse modes (Head or Tails approach)

Http://trimal.cgenomics.com

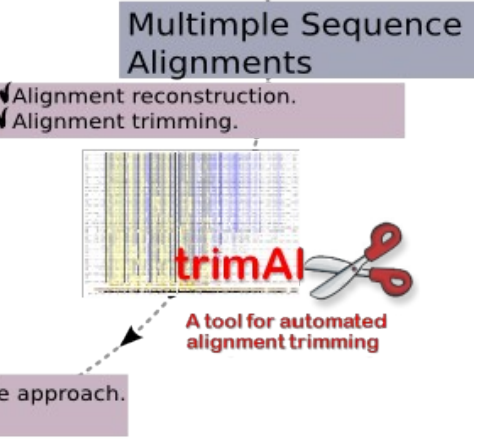
The set of homologous
Sequences are aligned by 3 different aligners in forward and reverse modes (Head or Tails approach)

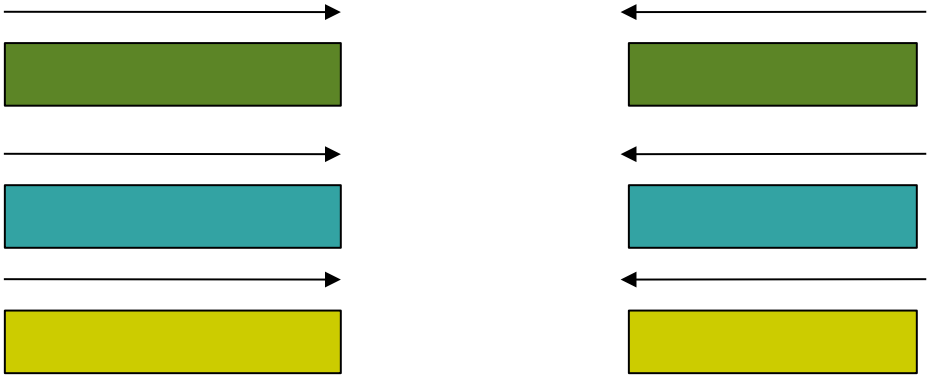
A consensus is built



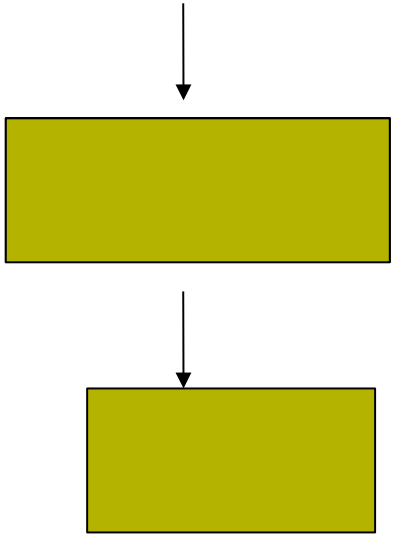
```
sw_DSBA_PSESM/1   ---MRNLIISAALVAASLFGMSAQAAEPIESGKQYV-ELTSAVPV
sw_DSBA_SALTY/1   ---MKKIWLA---LAGMVLAFSASAAQISD-GKQYI-TLDKP--V
sw_DSBA_ENTAM/3   AKWINSIFKSVVLTAALALPFTAS--AFTE-GTDYM-VLEKP---
sw_DSBA_LEGPN/1   -----------------LMPMTALATQFIE-GKDYQTVASAQ-LS
cons                                : ::*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

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

Ht





The set of homologous Sequences are aligned by 3 different aligners in forward and reverse modes (Head or Tails approach)



A consensus is buit

The consensus is trimmed (trimAl) based on:

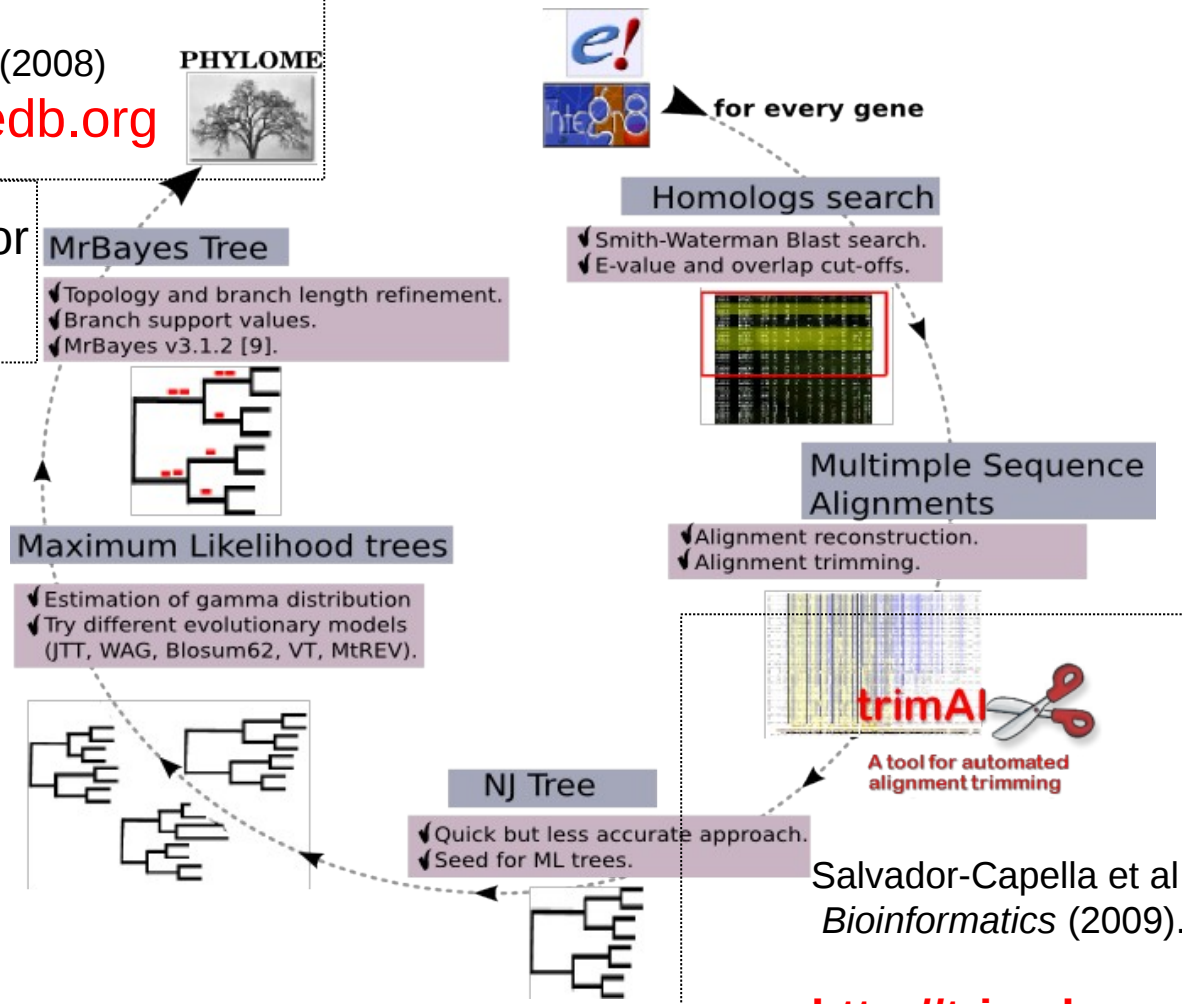
- consistency across the 6 alignments
- gap content

Http://trimal.cgenomics.com

Our pipeline:

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

ETE: Environment for Tree Exploration
ete.cgenomics.org



Salvador-Capella et al
Bioinformatics (2009).

http://trimal.cgenomics.org

Pipeline described in Huerta-Cepas et al NAR (2011)

Www.phylomedb.org

phylome DB

Collections All phylomes Downloads Help FAQ About

[Login]

Search in PhylomeDB

(i.e. ENSG00000139618, YBL058W, TP53)

Search

RandomTree!

BLAST search

Latest Phylomes

Arxula adeninivorans	2014
Beta vulgaris	2013
Clogmia albipunctata	2013
Penicillium digitatum	2012
Schistosoma mansoni	2012

see all phylomes

PhylomeDB uses

trimAl A tool for automated alignment trimming · Jalview a multiple alignment editor

PhylomeDB cross linking

UniProt · Ensembl · SGD · CGD · Génolevures · AcypiCyc · Treefam

Welcome to PhylomeDB 4!

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, as well as a **web API** that faciliates cross linking trees from external sources. 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.

What's new in phylomeDB 4?

Latest story

New Zygomycete phylome: the human pathogen Lichtheimia corymbifera

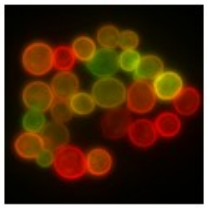
Mon, 09/15/2014 - 21:09

PhylomeDB extends its repertoire of fungal phylomes with that of a genome of a poorly sample clade, that of the basal group zygomycetes. In this case the phylome (245) of the human pathogenic mucorales Lichtemia corymbifera has served to reveal extensive past gene duplications in this group. Lichtheimia species are the second most important cause of mucormycosis in Europe. The sequencing of its genome and the comparison with other Zygomycete species, particularly of Rhizopus delemar, the main

Popular Phylome Collections

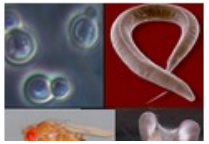
Human

Fungi



Plants

Model Species



Latest News

New Zygomycete phylome: the human pathogen Lichtheimia corymbifera

Mon, 09/15/2014 - 21:09

New hemyascomycete phylome: Blastobotrys (Arxula) adeninivorans, a yeast of biotechnological interest.

Mon, 05/19/2014 - 11:01

Help us to improve PhylomeDB: complete our survey.

Thu, 02/20/2014 - 16:22

show all

PhylomeDB Twitter

Tweets Follow

phylomedb @phylomedb 23 Oct

New birds, crocs, and fungal phylomes to come soon at phylomeDB. stay tuned!

Expand

phylomedb @phylomedb 15 Sep

New Zygomycete phylome: the human fungal pathogen Lichteimia corymbifera phylomedb.org/?q=node/537

Expand

phylomedb @phylomedb 26 Aug

NOTICE: PhylomeDB will be down due to MAINTENANCE

Search in PhylomeDB

(i.e. ENSG00000139618, YBL058W, TP53)

Search

RandomTree!

BLAST search

Latest Phylomes

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

see all phylomes

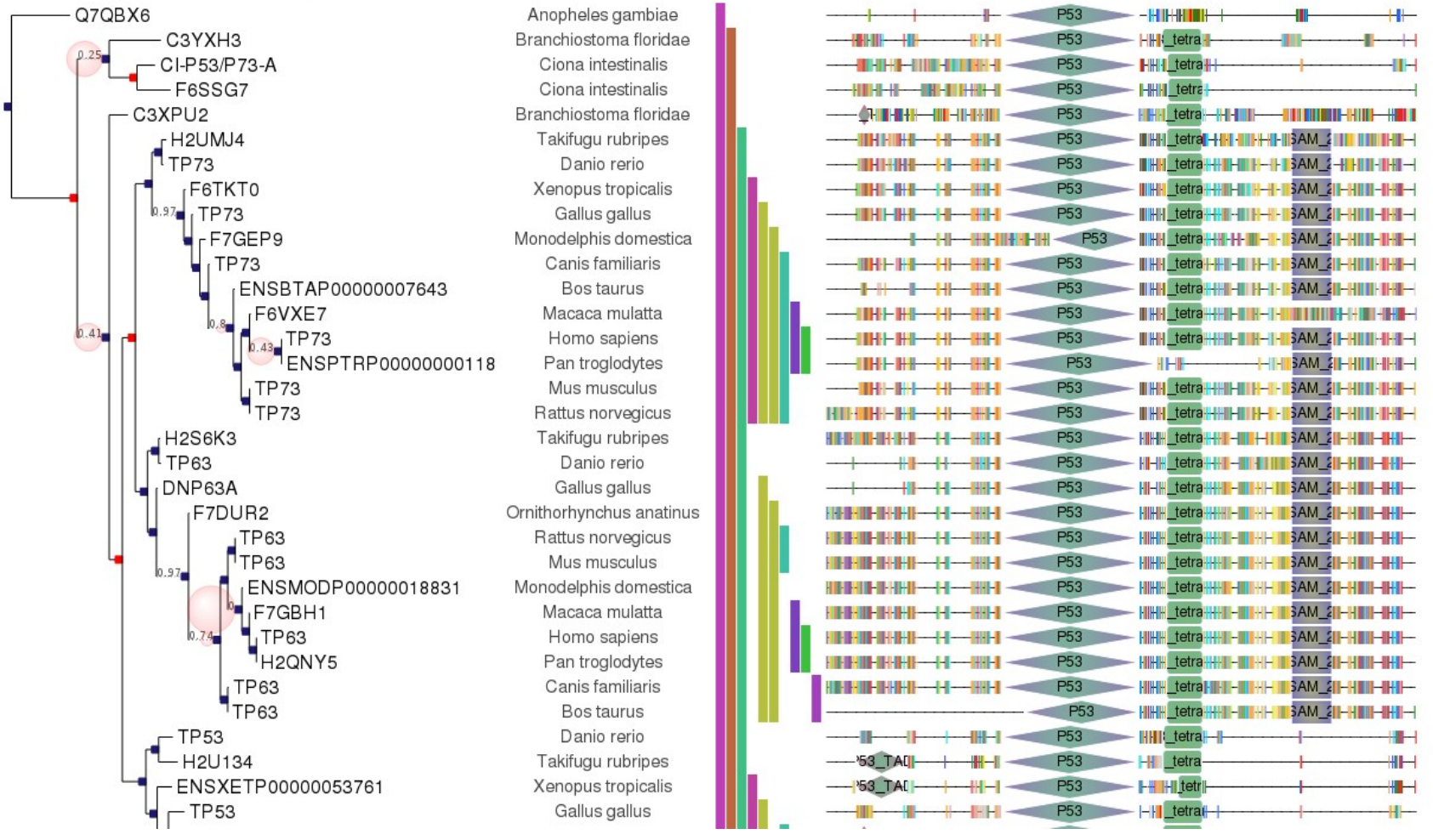
PhylomeDB uses

trimAl Jalview

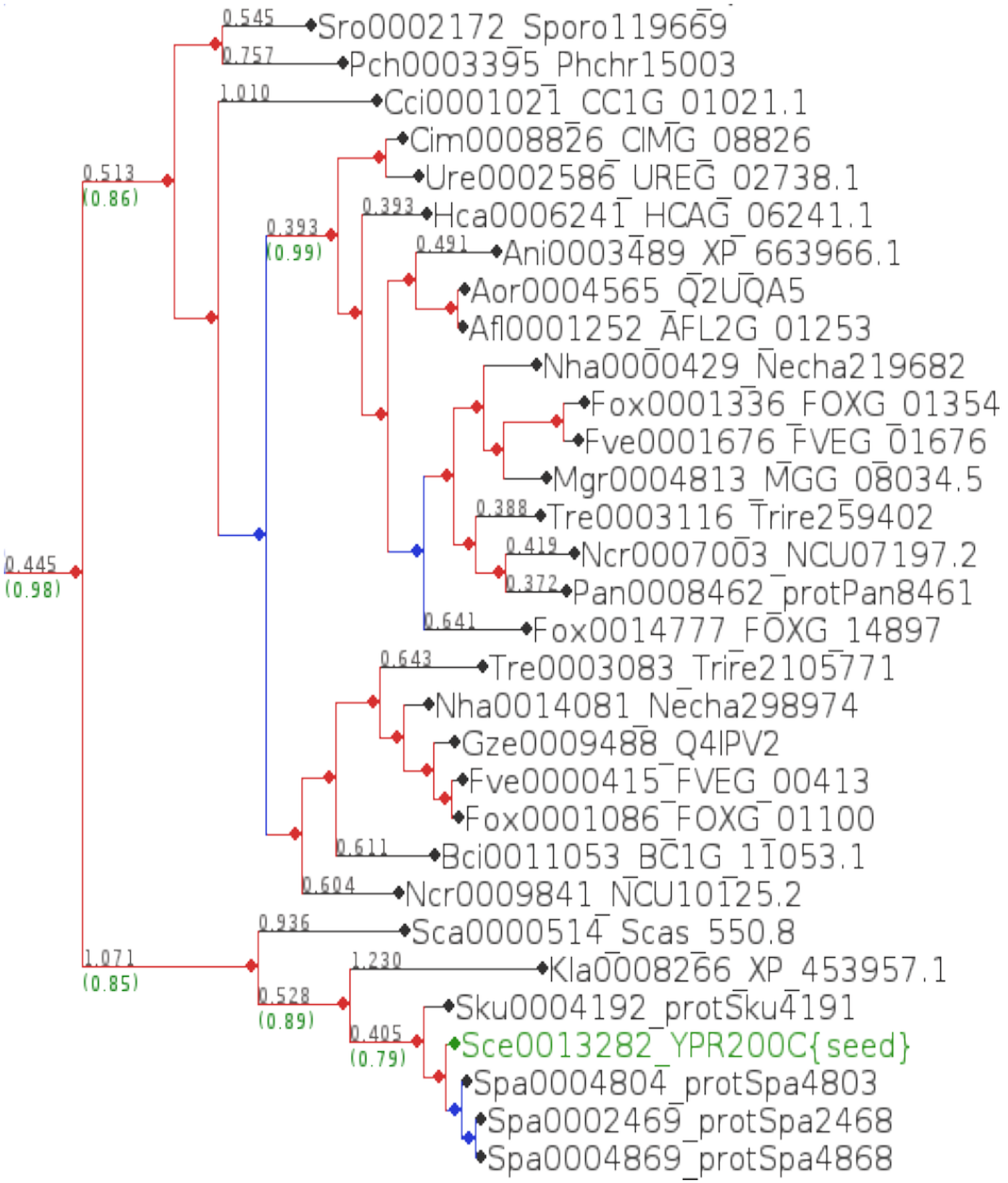
TP53 tree in phylome 218

AS seed in Rat phylome | JTT (lk:-18130.4) | -- in collateral trees --

Tree features | Search | Clear search | Image | Hard link | Download OrthoXML | See alignments | Download data.tar.gz



These phylomes can now be interrogated in many ways



- Families that show a particular topology
-Detect and date duplication events

- Genes that have accelerated evolutionary rates at a particular lineage (positive/relaxed selection)

- Families expanded at particular lineages

- Footprints of horizontal gene transfer, lineage sorting, gene conversion and other evolutionary processes

- Search for co-evolving genes

- predict functional properties

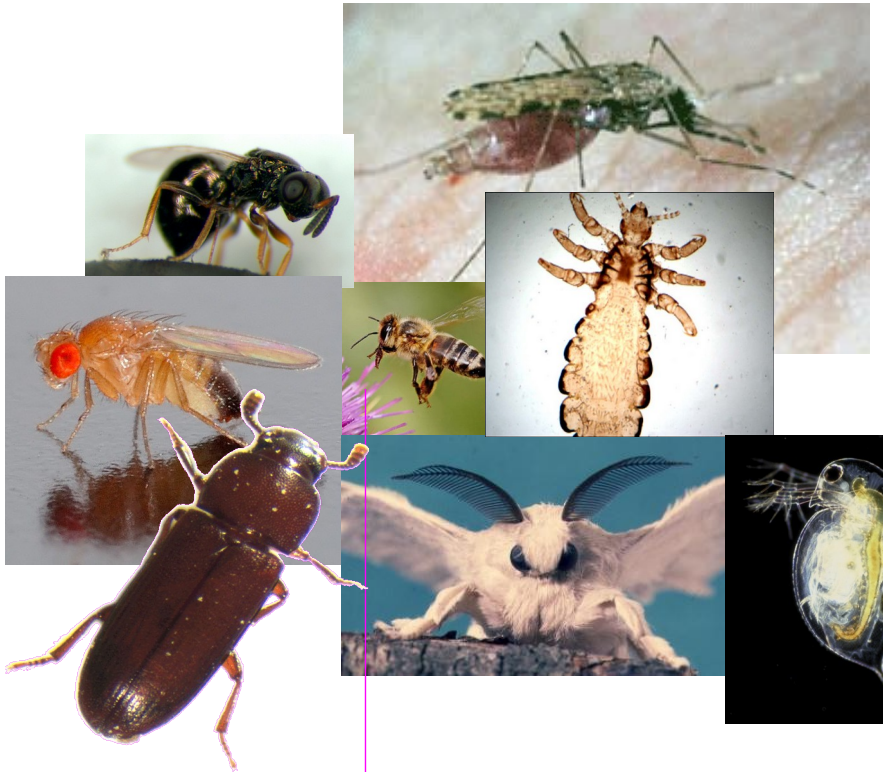
- across-species prediction of orthology and paralogy

-

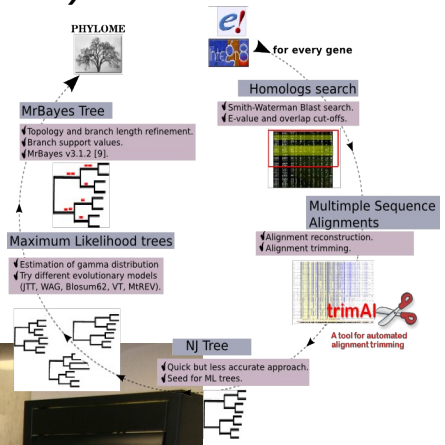
Large-scale phylogenetics to assist in the annotation of newly sequenced genomes: the *Acyrthosiphon pisum* genome .

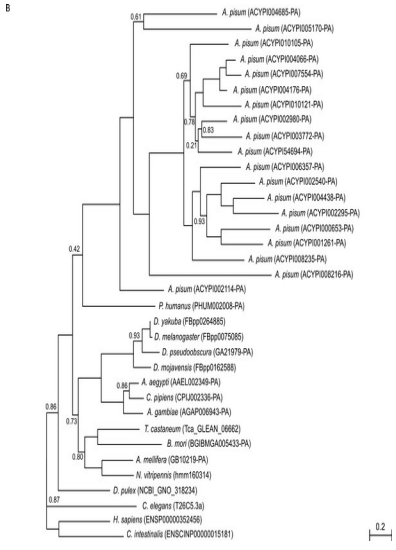
AphidBase

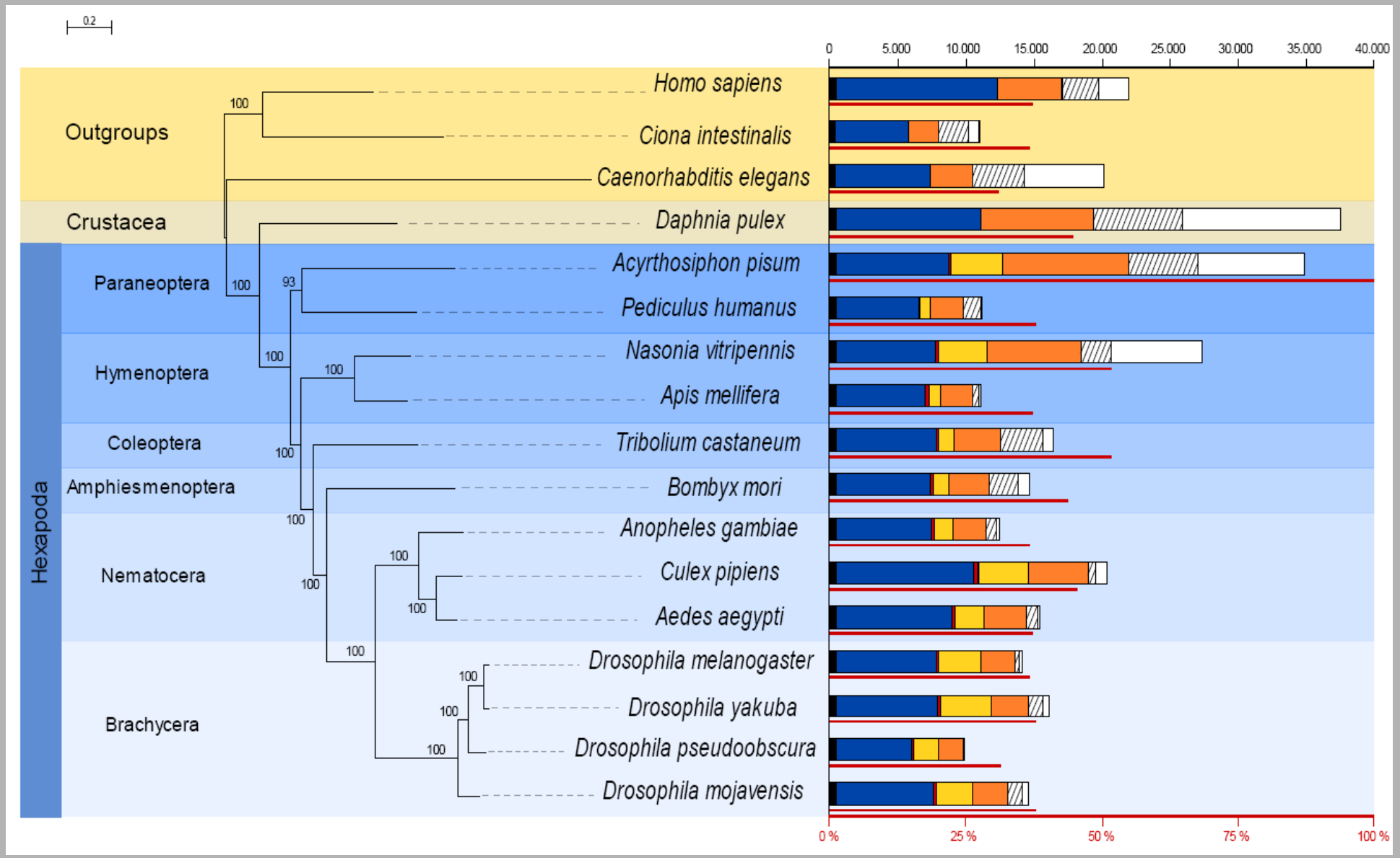
Acyr_1.0 (34,600 genes)



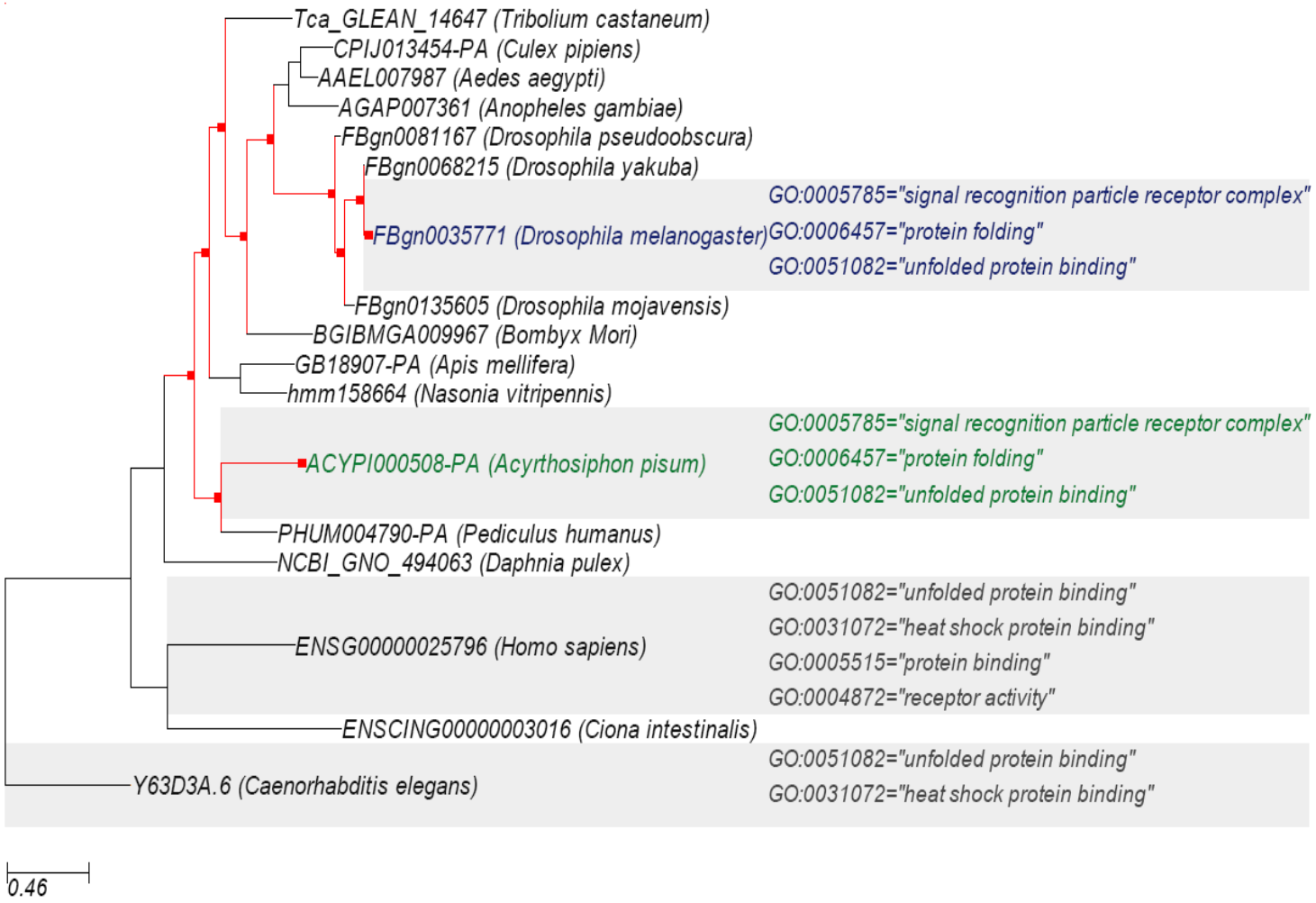
13 other sequenced arthropods and 3 out-groups







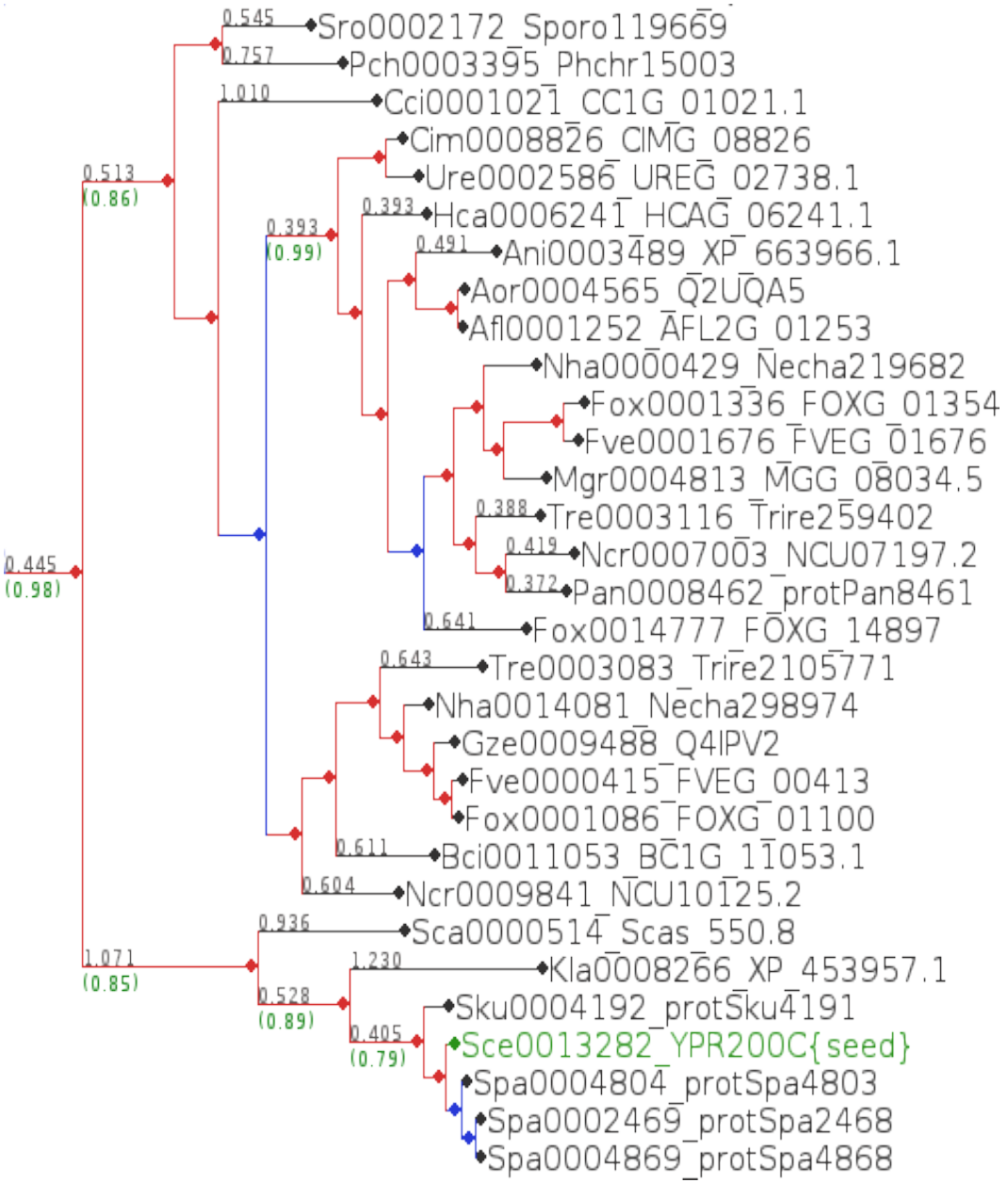
Tree based on the alignment concatenation of 197, widespread proteins



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)

These phylomes can now be interrogated in many ways



- Families that show a particular topology
-Detect and date duplication events

- Genes that have accelerated evolutionary rates at a particular lineage (positive/relaxed selection)

- **Families expanded at particular lineages**

- Footptints of horizontal gene transfer, lineage sorting, gene conversion and other evolutionary processes

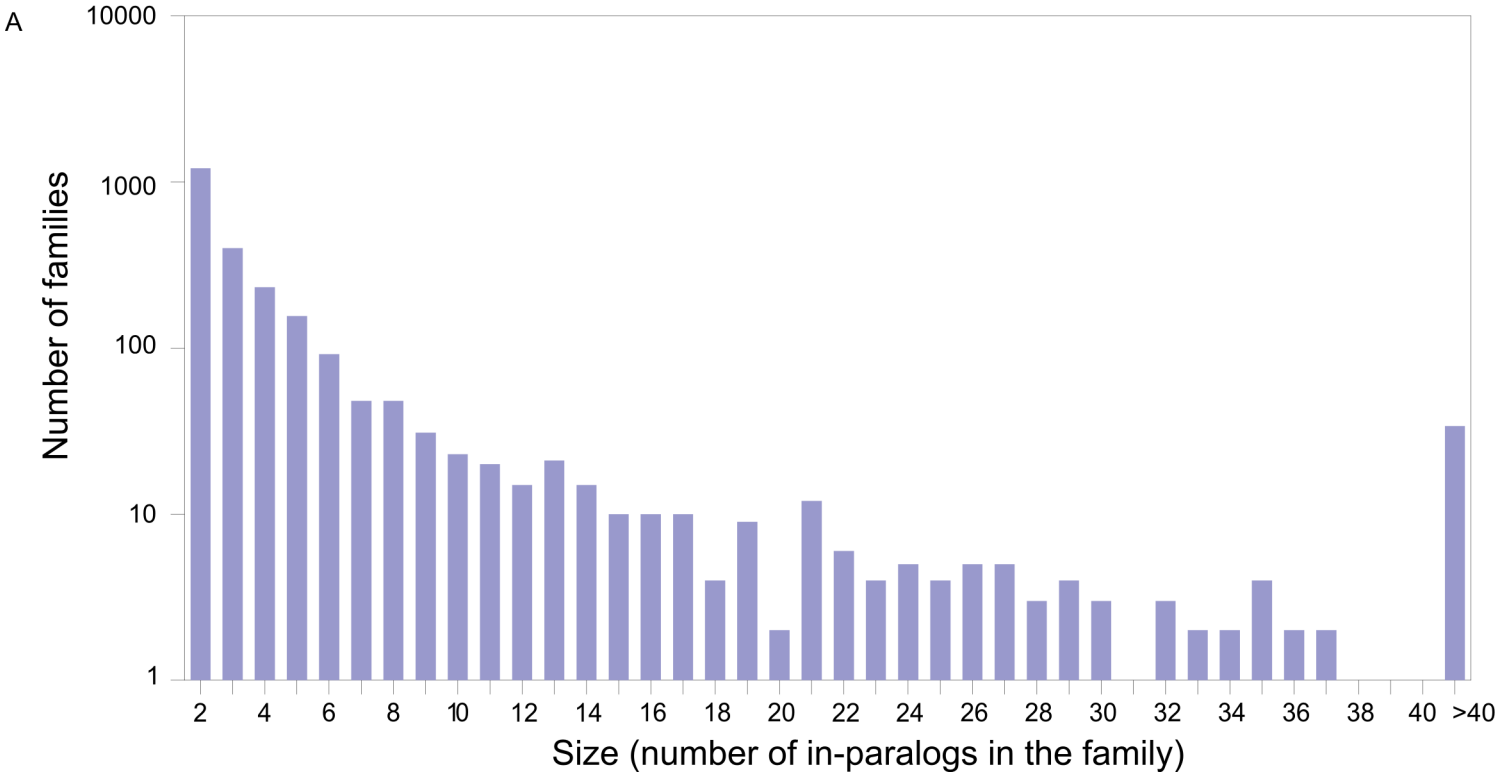
- Search for co-evolving genes

- predict functional properties

- accross-species prediction of orthology and paralogy

-

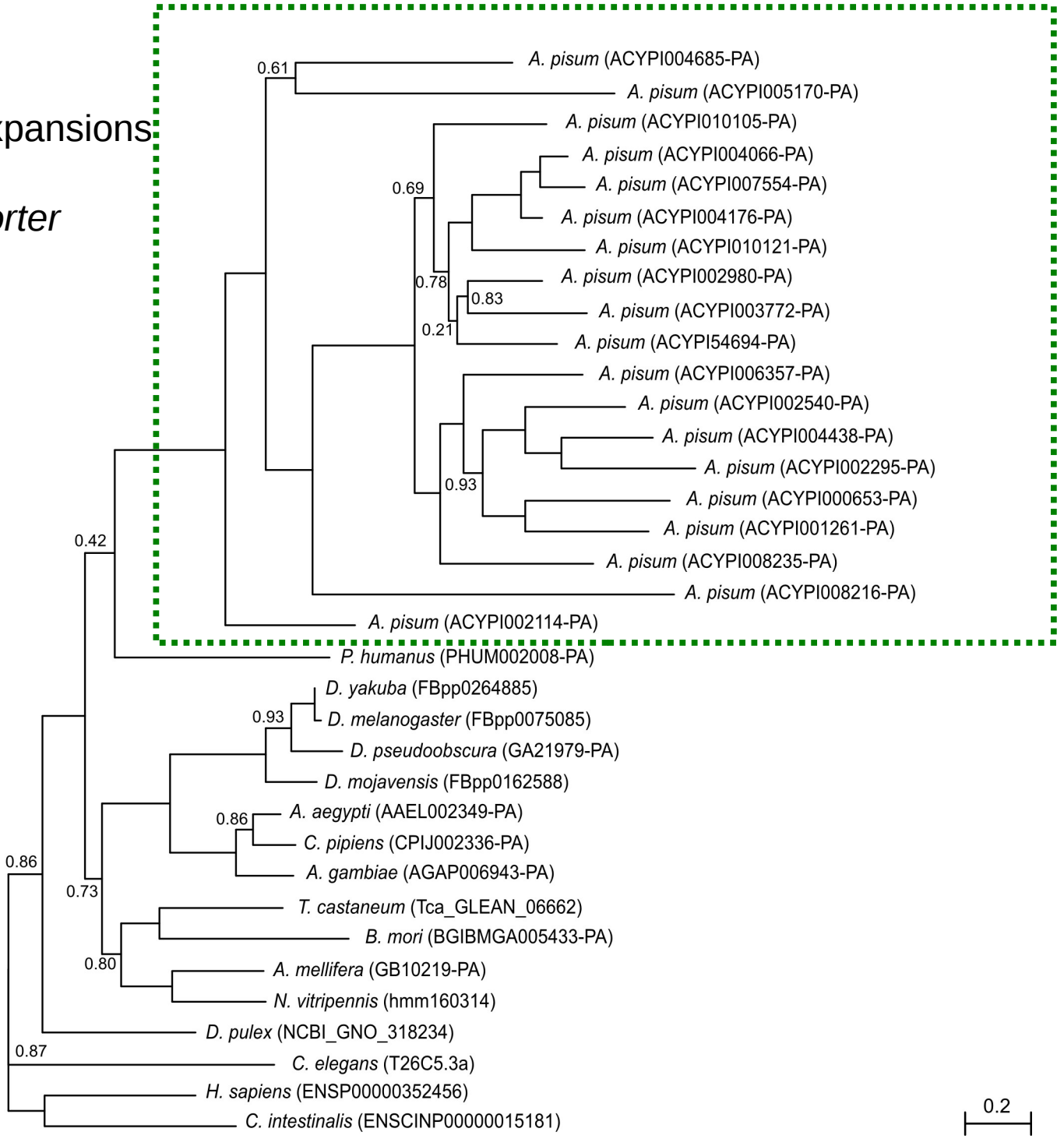
A wave of lineage-specific expansions in the pea aphid



B

Lineage-specific expansions

Acetyl-CoA transporter




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

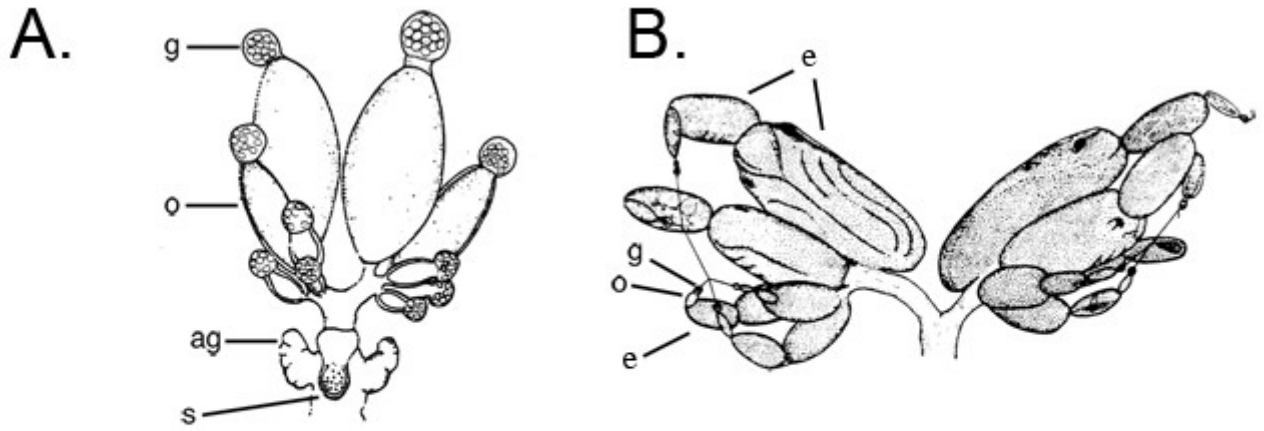
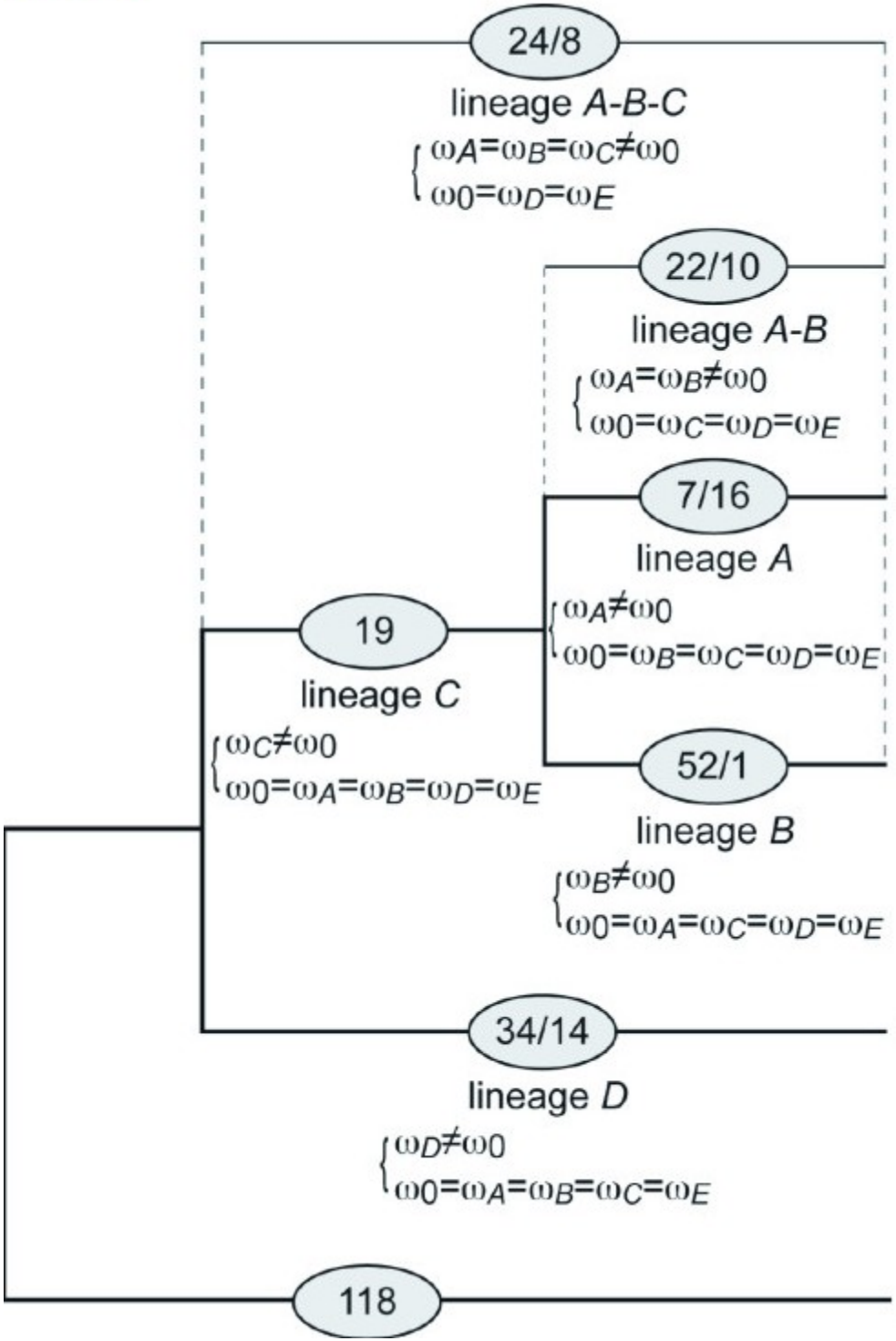


Figure 2. Viviparous and oviparous development. Oviparous (A) and viviparous (B) ovaries differ not only as to whether they possess embryos, accessory glands and spermathecae, but also in the relative size of germaria and oocytes. Abbreviations: g is germarium, o is oocyte, e is viviparous embryo, ag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

Pictures taken from www.princeton.edu (Gregory Davis)

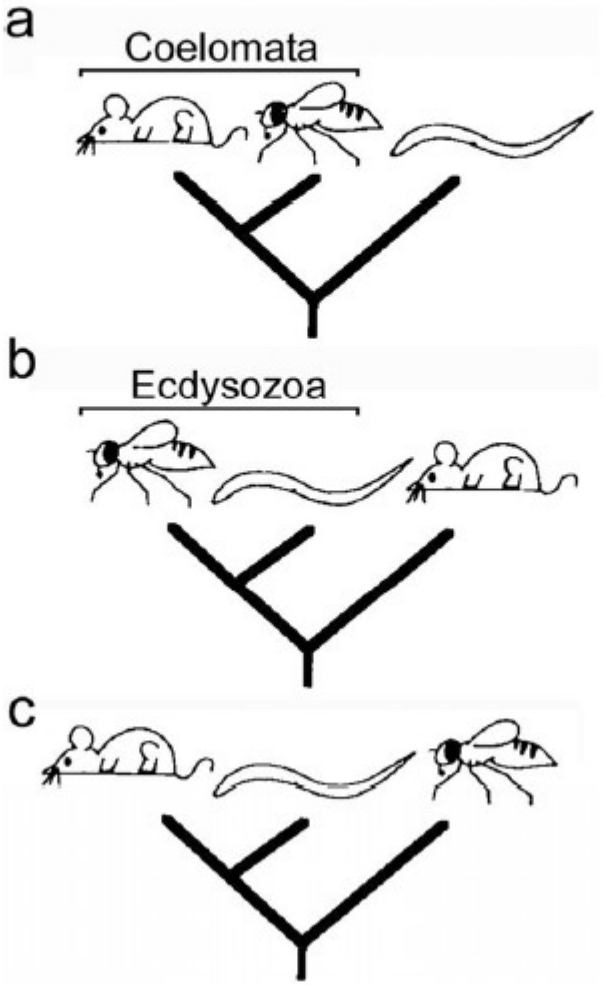
Figure 2.

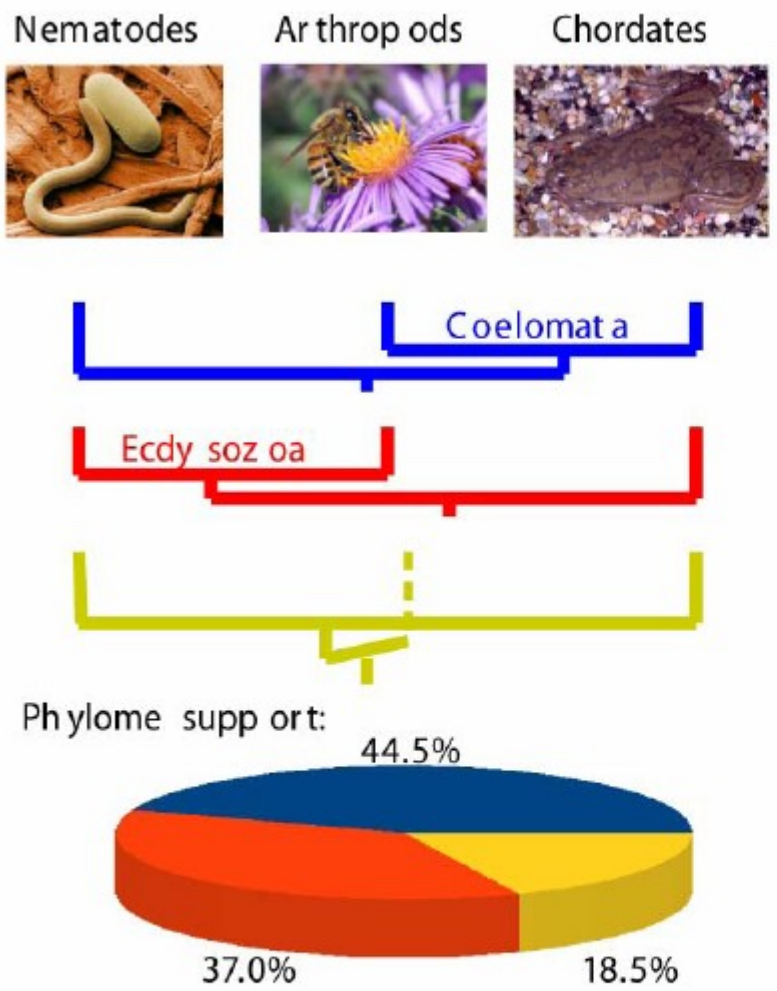


Going beyond topologies

Gene tree vs species trees

Uncertainty in species trees and topological variability in gene trees





What percentage of gene trees from the human phylome support each topology?

Similar results for

Primates
Rodents
laurasatheria

Huerta-Cepas et. al. (2007)

Possible sources of gene tree vs species tree discordance

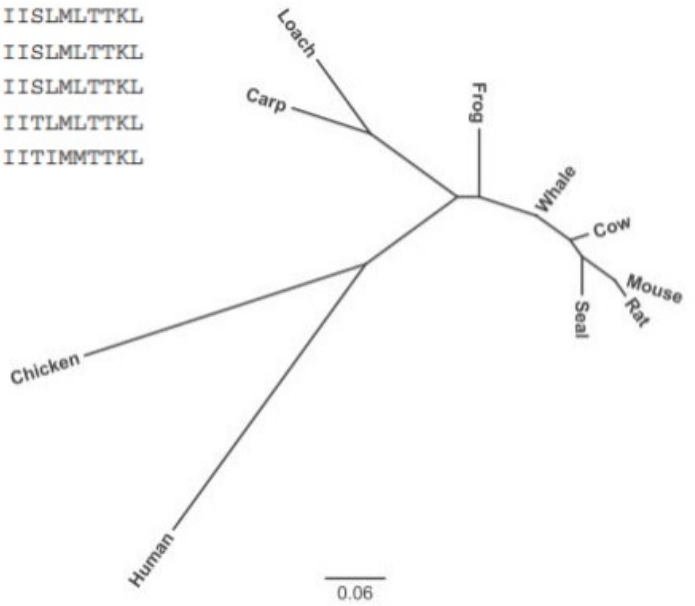
Analytical factors

They lead to failure in accurately inferring a gene tree; these can be either due to **stochastic error** (e.g., insufficient sequence length or taxon samples) or due to **systematic error** (e.g., observed data far depart from model assumptions)

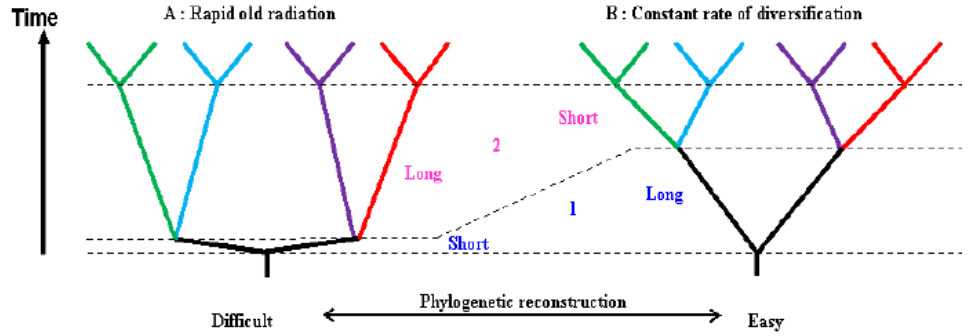
Biological factors

They lead to gene trees that are topologically distinct from each other and from the species tree. Known factors include **stochastic lineage sorting**, **hidden paralogy**, **horizontal gene transfer**, **recombination** and **natural selection**

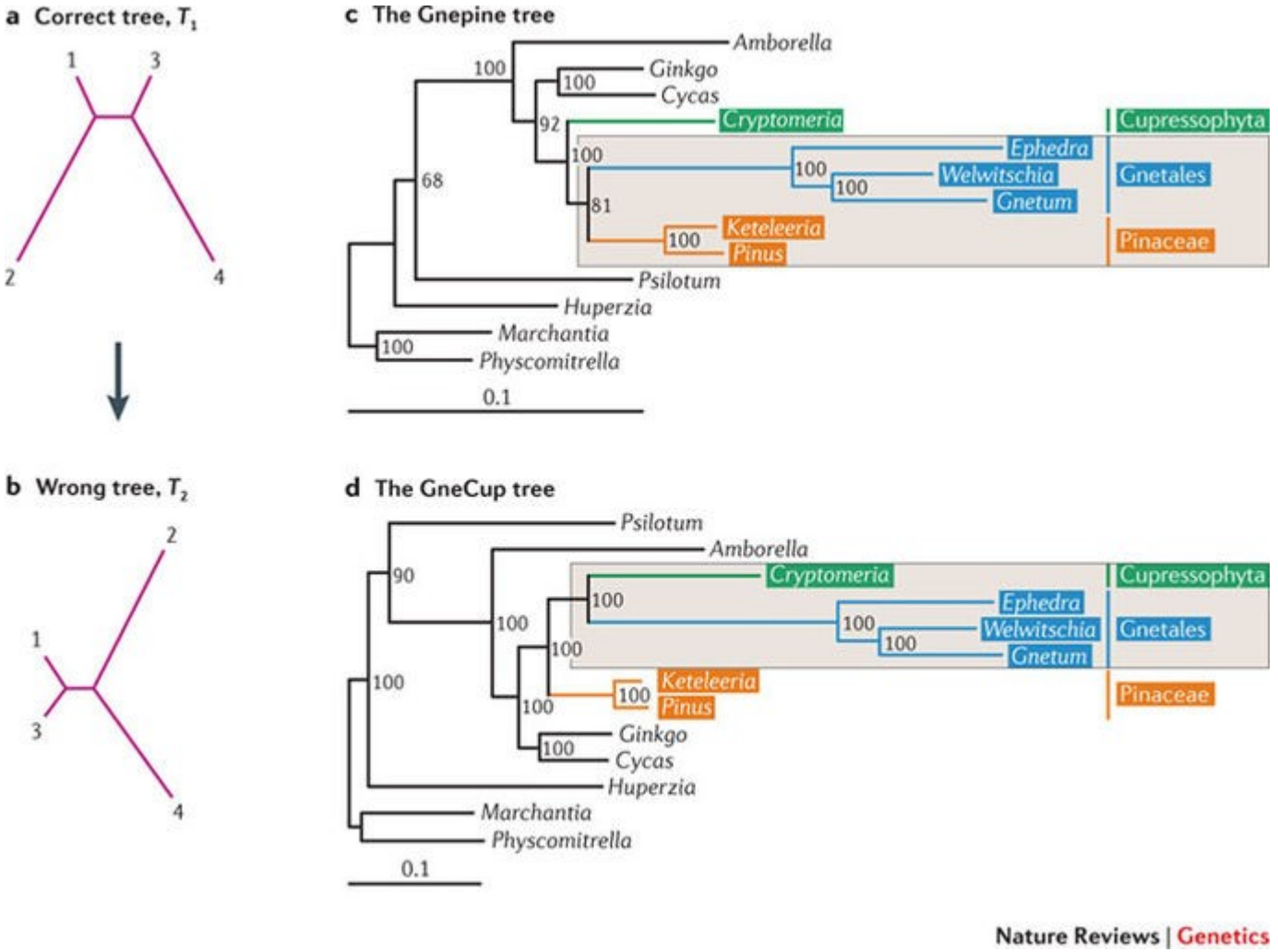
```
10 50
Cow         MAYPMQLGFQ DATSPIMEEL LHFHDHTLMI VFLISSLVLY IISLMLTTKL
Carp        MAHPTQLGFK DAAMPVMEEL LHFHDHALMI VLLISTLVLY IITAMVSTKL
Chicken     MANHSQLGFQ DASSPIMEEL VEFHDHALMV ALAICSLVLY LLTLMLMEKL
Human       MAHAAQVGLQ DATSPIMEEL ITFHDHALMI IFLICFLVLY ALFLTLTTKL
Loach       MAHPTQLGFQ DAASPVMEEL LHFHDHALMI VFLISALVLY VIITTVSTKL
Mouse       MAYPFQLGLQ DATSPIMEEL MNFHDHTLMI VFLISSLVLY IISLMLTTKL
Rat         MAYPFQLGLQ DATSPIMEEL TNFHDHTLMI VFLISSLVLY IISLMLTTKL
Seal        MAYPLQMGLQ DATSPIMEEL LHFHDHTLMI VFLISSLVLY IISLMLTTKL
Whale       MAYPFQLGFQ DAASPIMEEL LHFHDHTLMI VFLISSLVLY IITLMLTTKL
Frog        MAHPSQLGFQ DAASPIMEEL LHFHDHTLMA VFLISTLVLY IITIMMTTKL
```



Alignment can be seen as a sampling
It has stochastic variation, so **sampling "errors"**
Can be expected, particularly at:
- Short sequences
- Short internodes (i.e. in fast radiations)



Systematic errors (i.e. long branch attraction artifact)



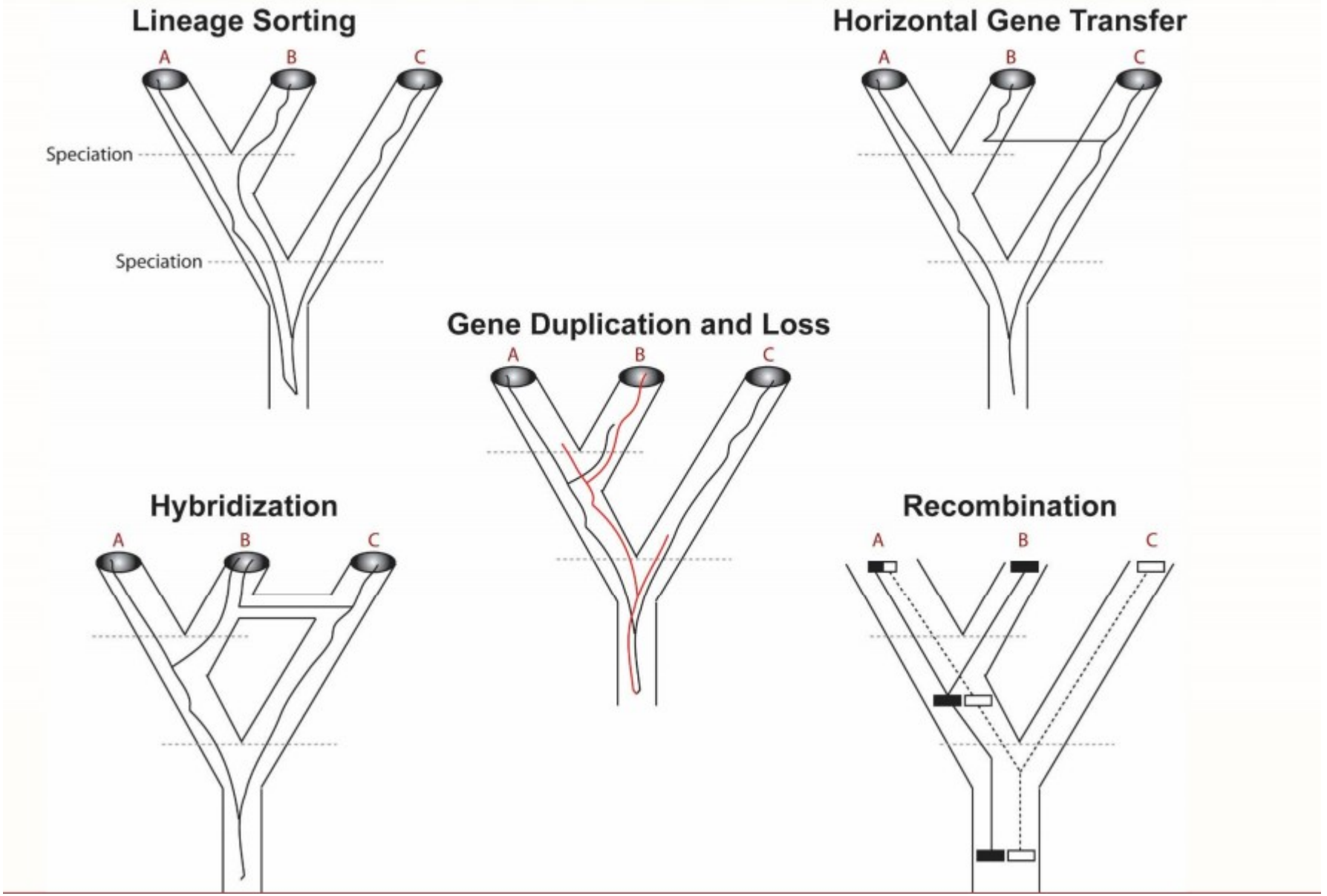
Possible sources of gene tree vs species tree discordance

Analytical factors

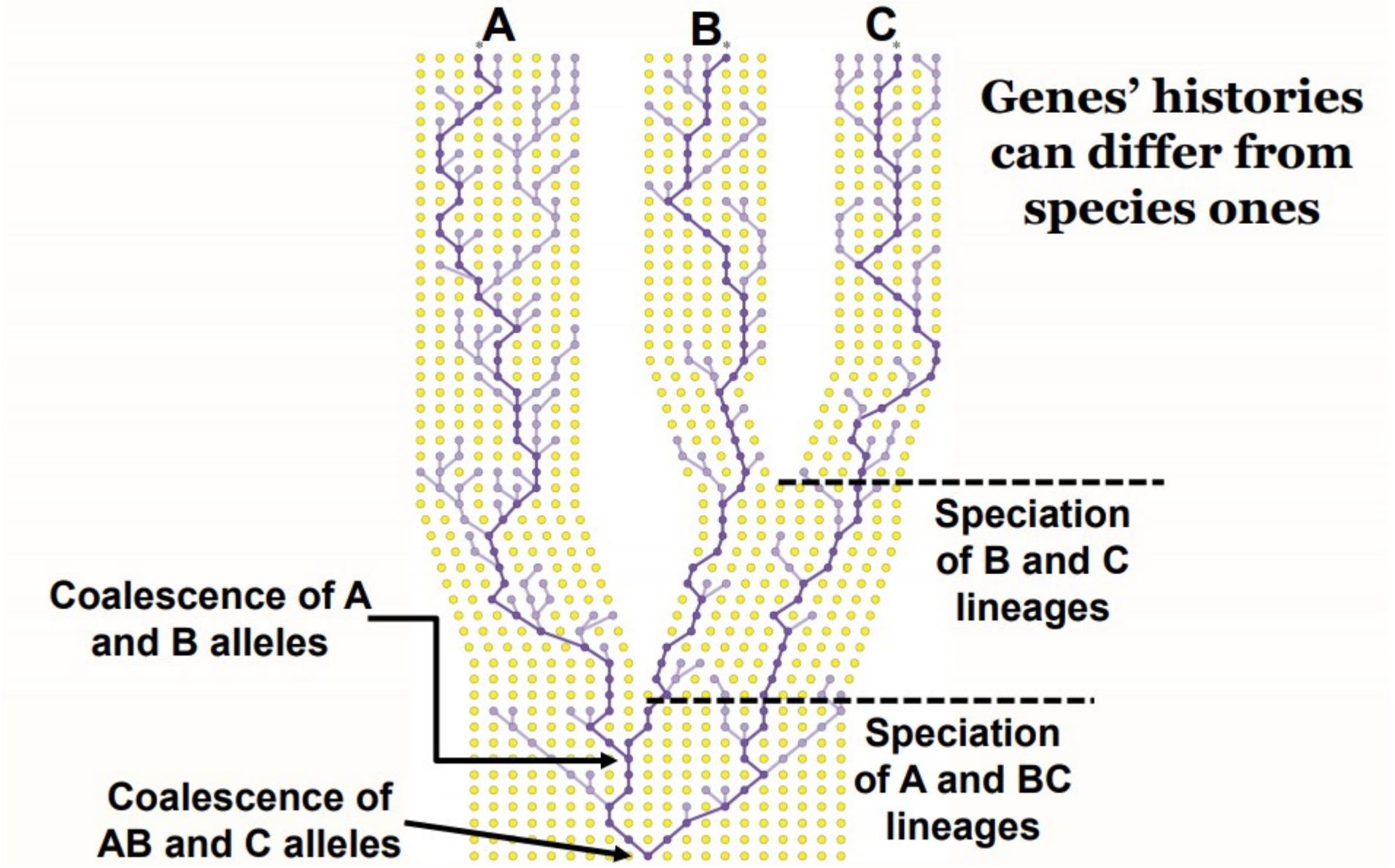
They lead to failure in accurately inferring a gene tree; these can be either due to **stochastic error** (e.g., insufficient sequence length or taxon samples) or due to **systematic error** (e.g., observed data far depart from model assumptions)

Biological factors

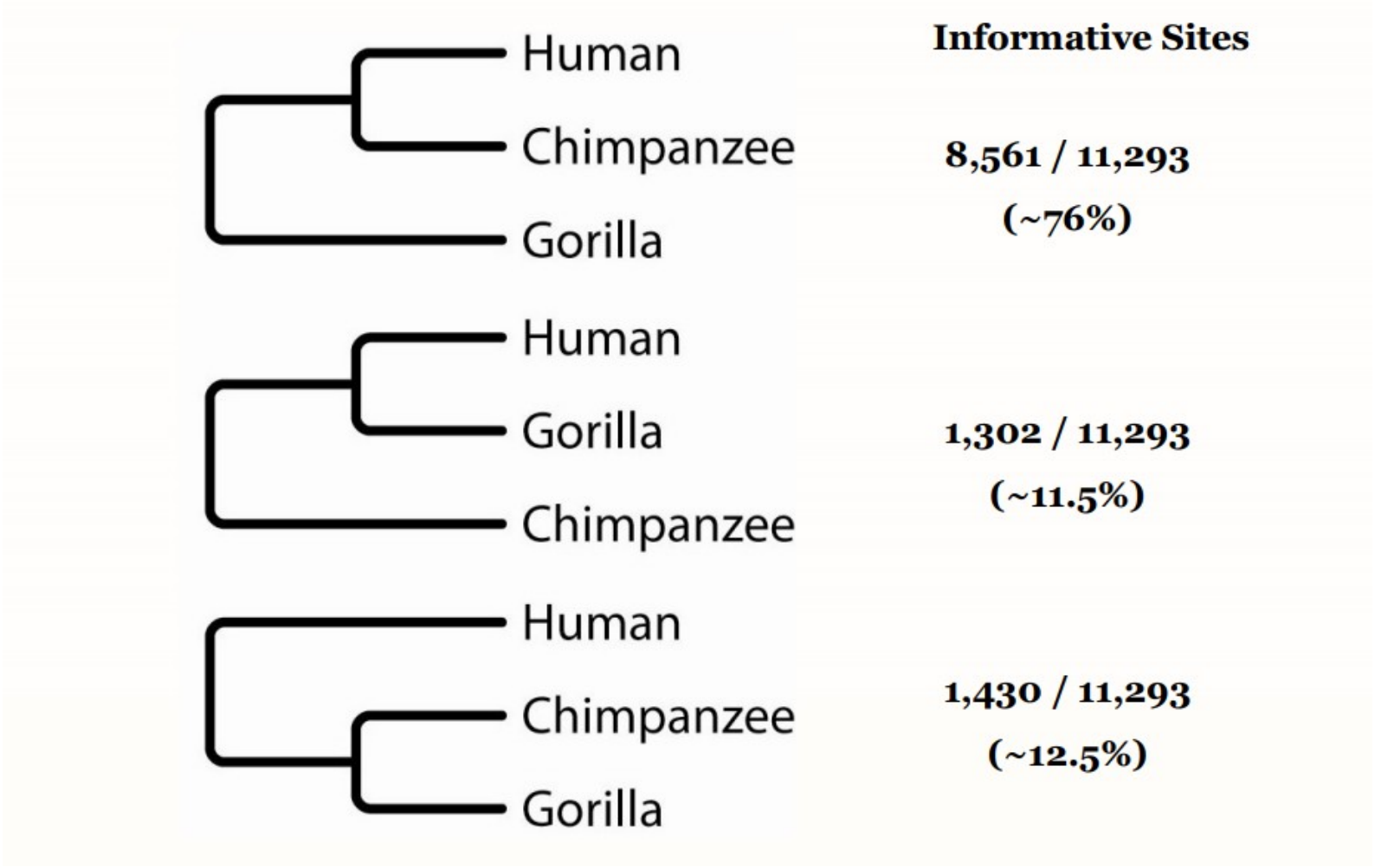
They lead to gene trees that are topologically distinct from each other and from the species tree. Known factors include **stochastic lineage sorting**, **hidden paralogy**, **horizontal gene transfer**, **recombination** and **natural selection**

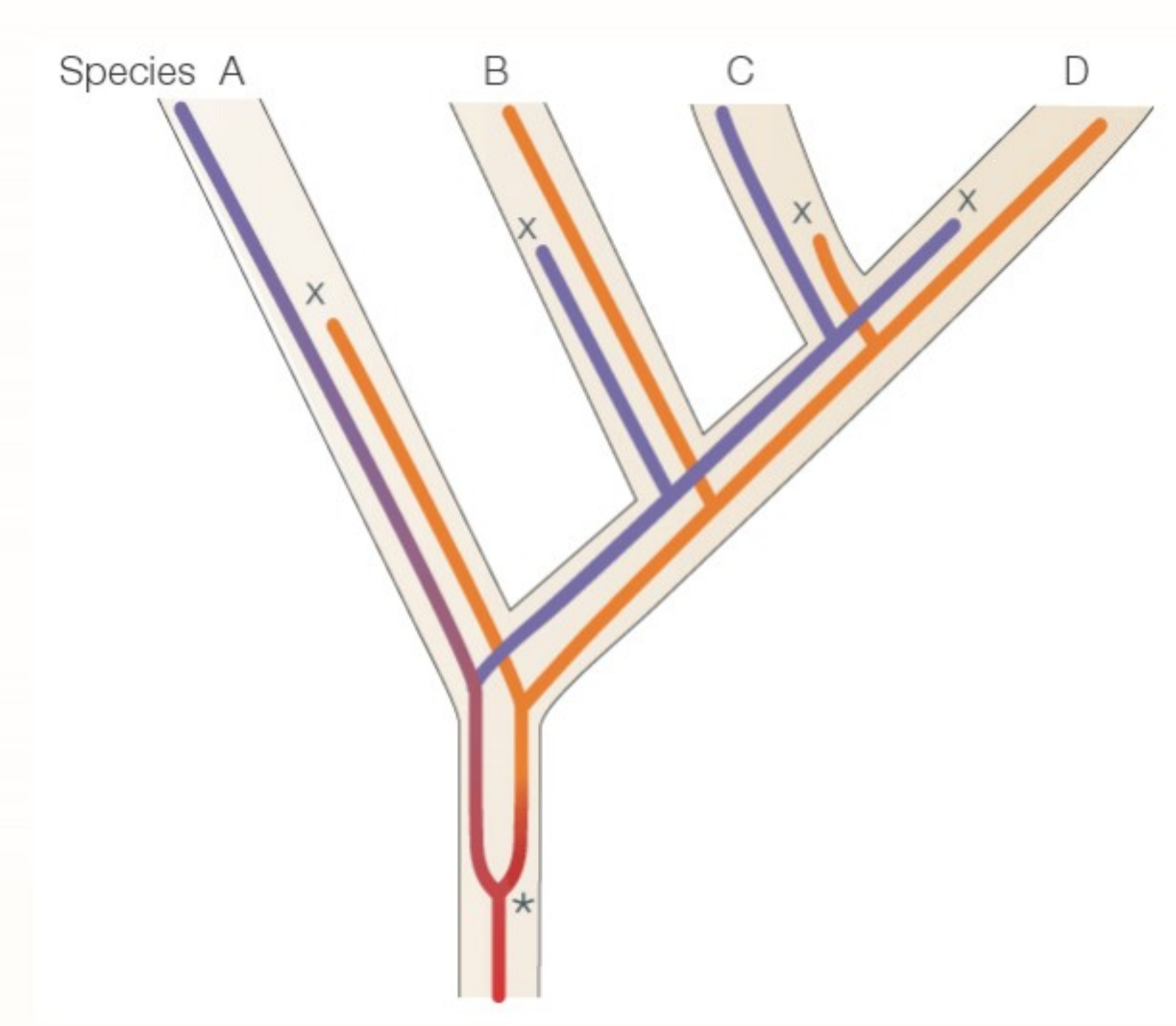

Lineage Sorting
A
B
C
Speciation
Speciation
Horizontal Gene Transfer
A
B
C
Gene Duplication and Loss
A
B
C
Hybridization
A
B
C
Recombination
A
B
C

Incomplete lineage sorting



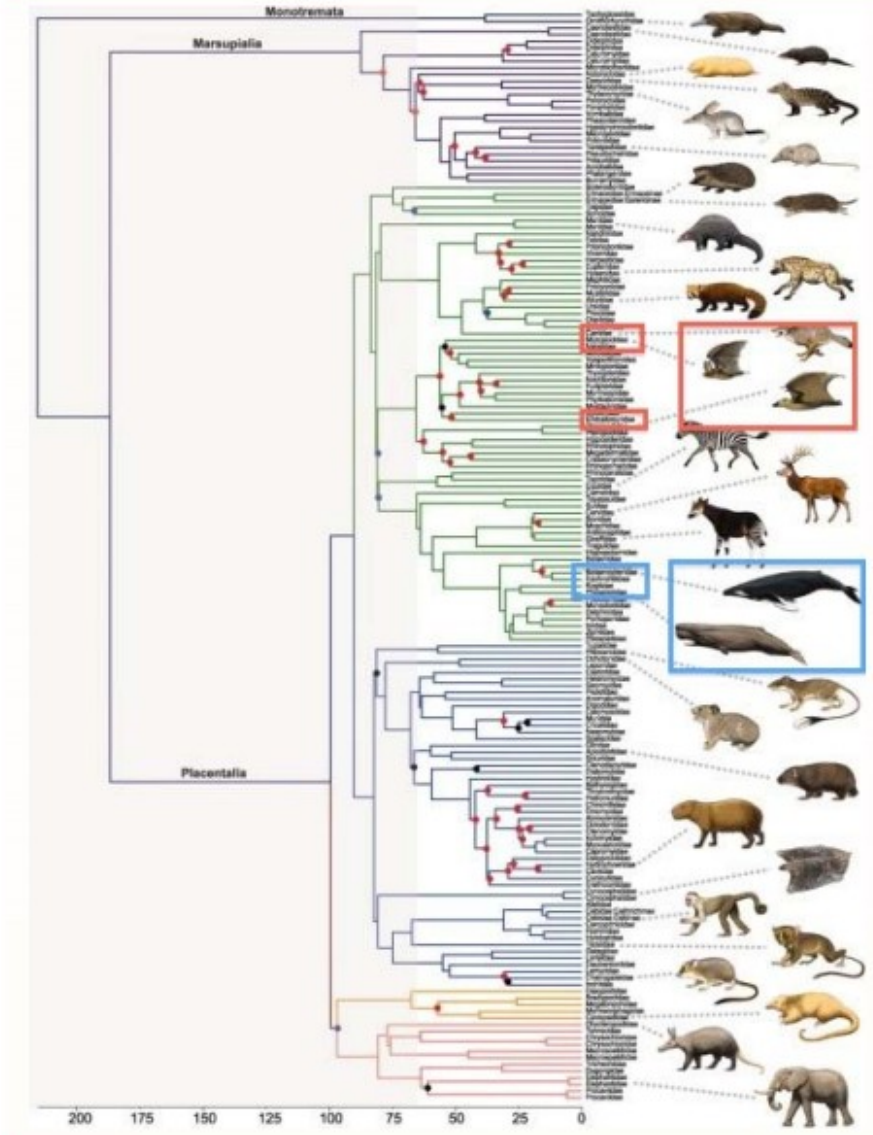
Incomplete lineage sorting in the primate lineage:

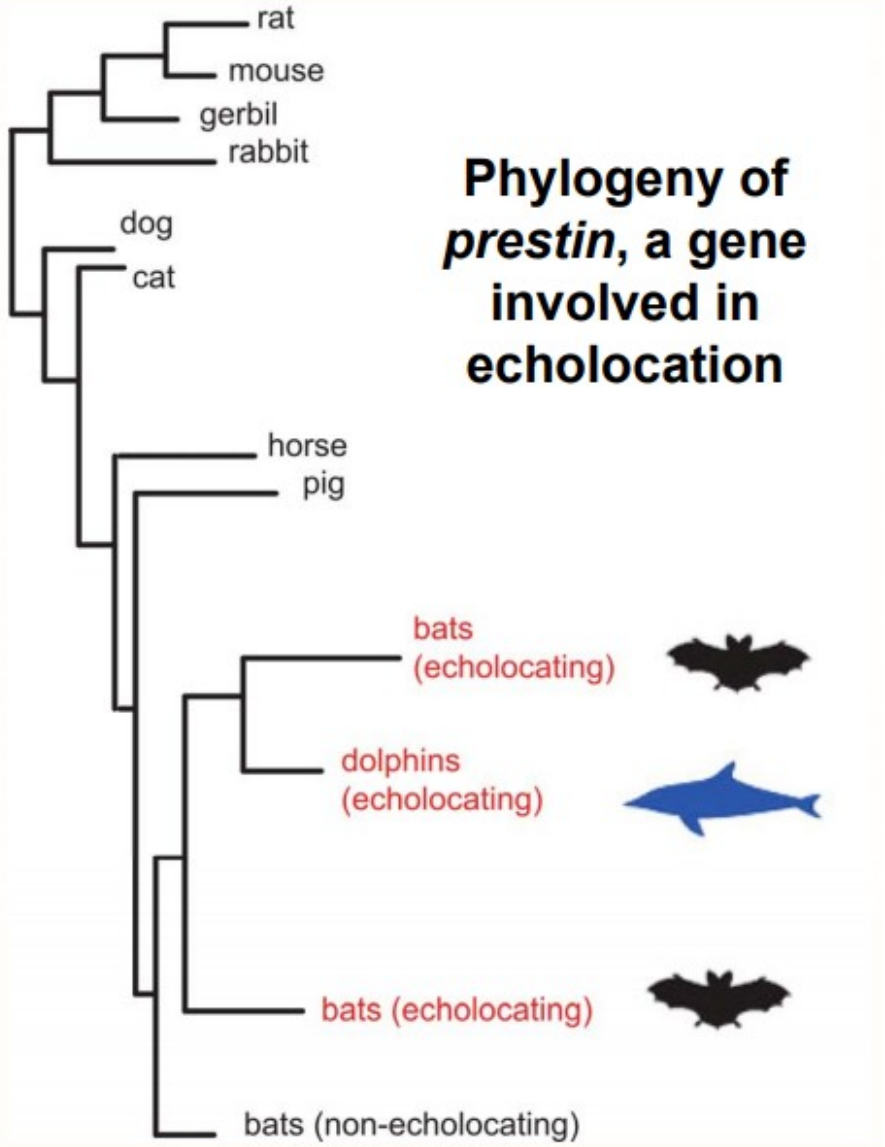




Hidden paralogy caused by differential gene loss following duplication

Positive selection can cause tree patterns different from the gene tree by convergent mutations



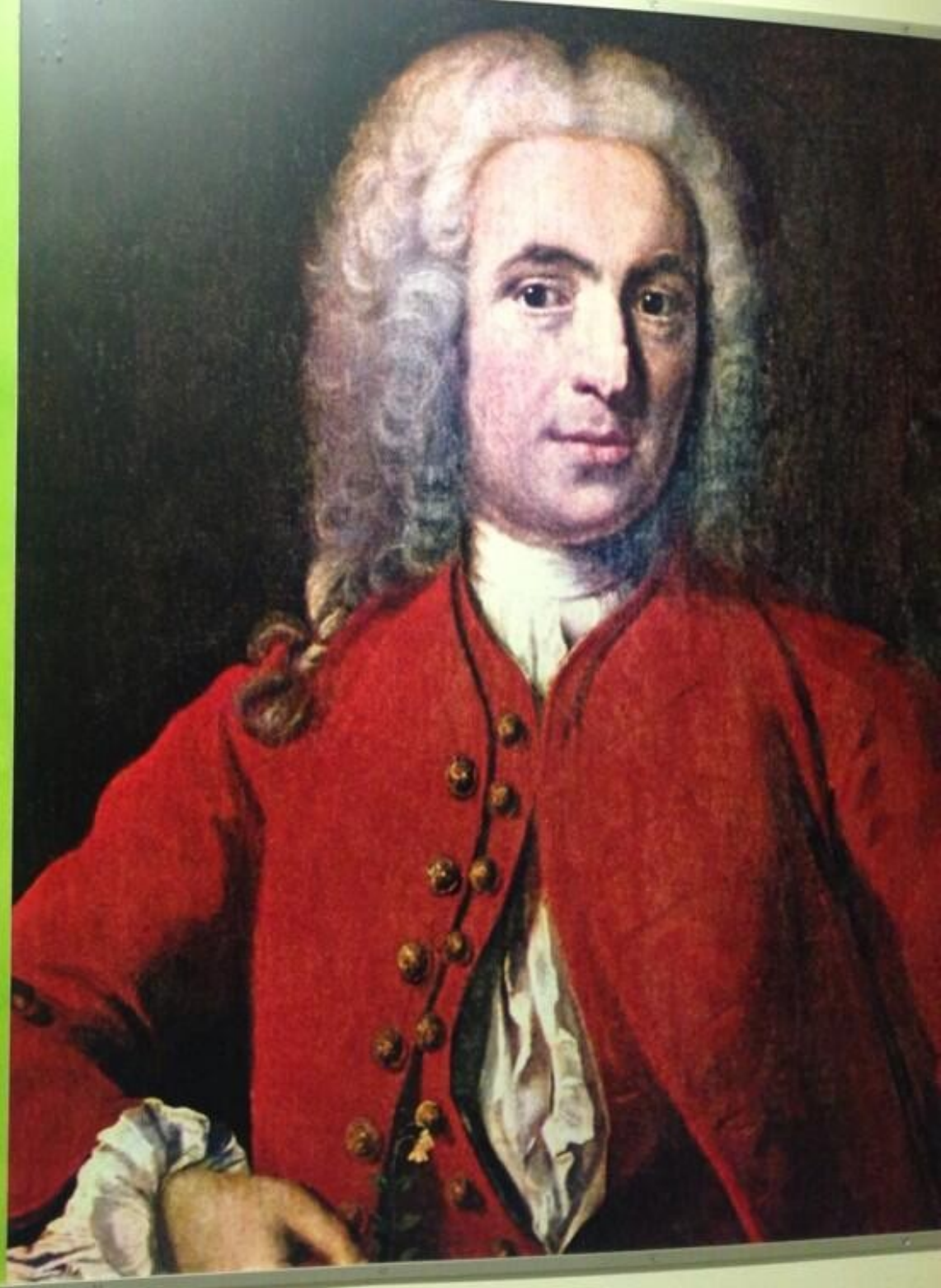


Phylogeny of *prestin*, a gene involved in echolocation

Non-vertical evolution

Welcome to my hometown

Carl von Linné, the Father of Taxonomy

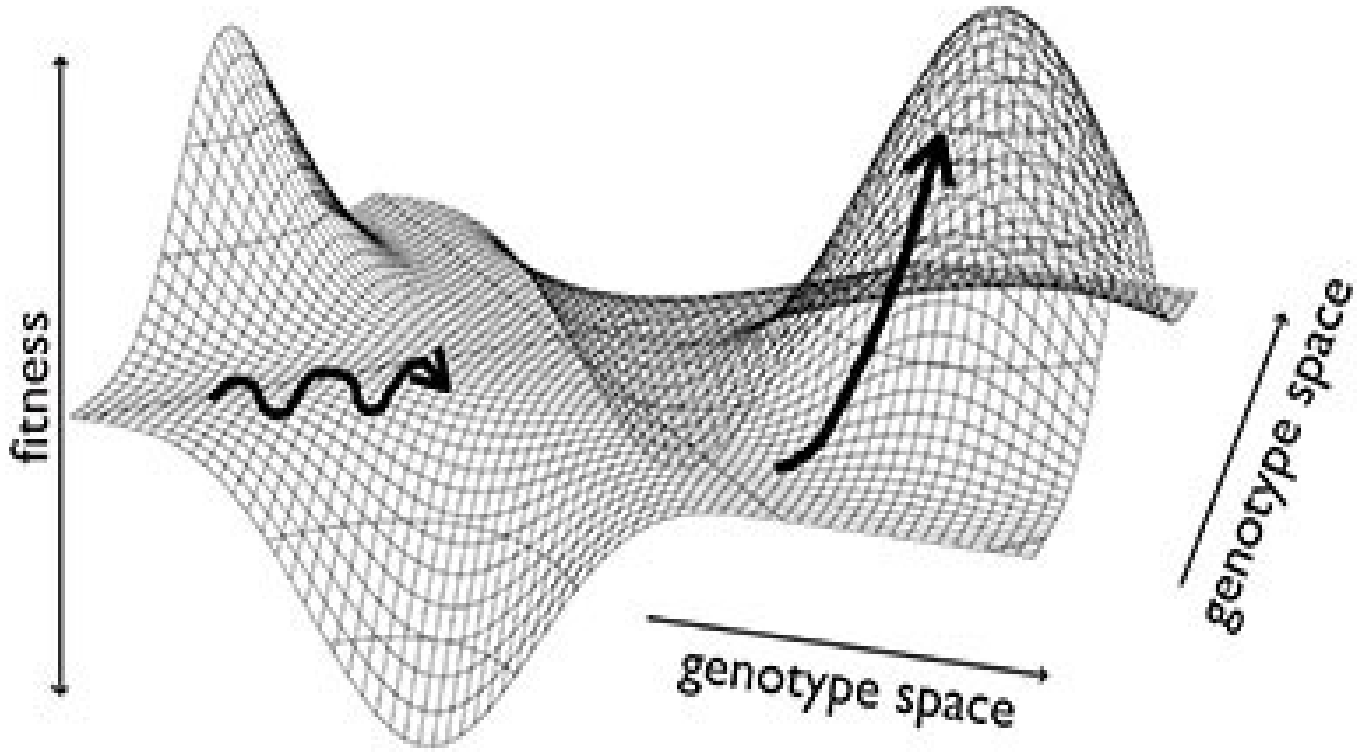


Huxley to Darwin (Nov 1859)

The only objections that have occurred to me are 1st that you have loaded yourself with an unnecessary difficulty in adopting 'Natura non facit saltum' so unreservedly. I believe she does make small jumps

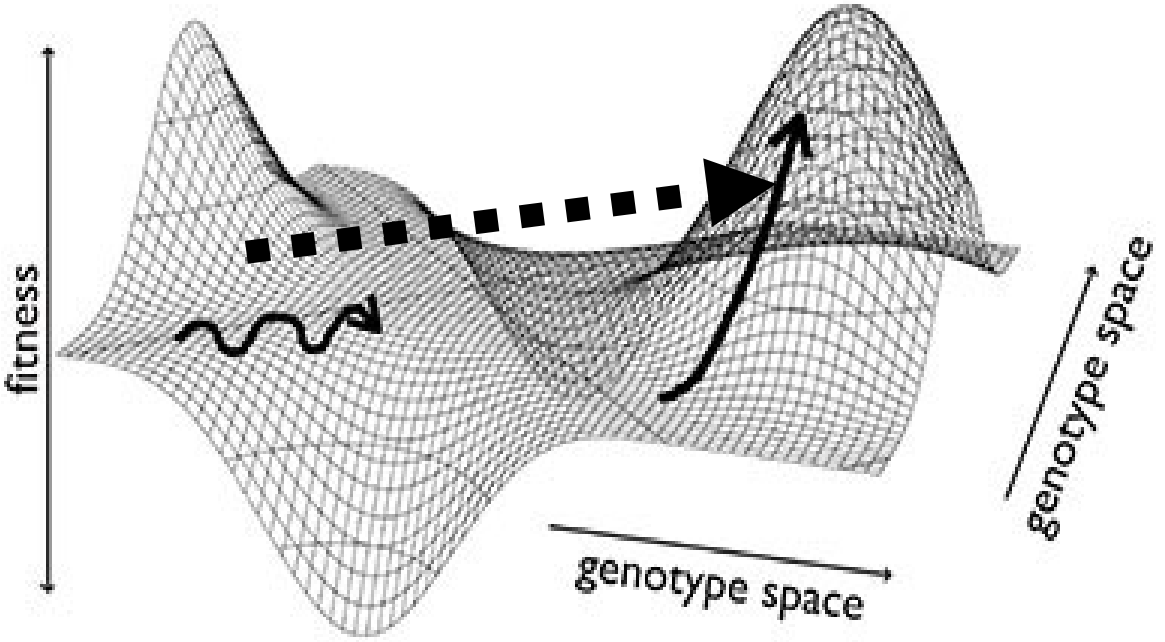
Darwin Correspondence Database
http://www.darwinproject.ac.uk/entry-2544

Mutation, selection and drift:

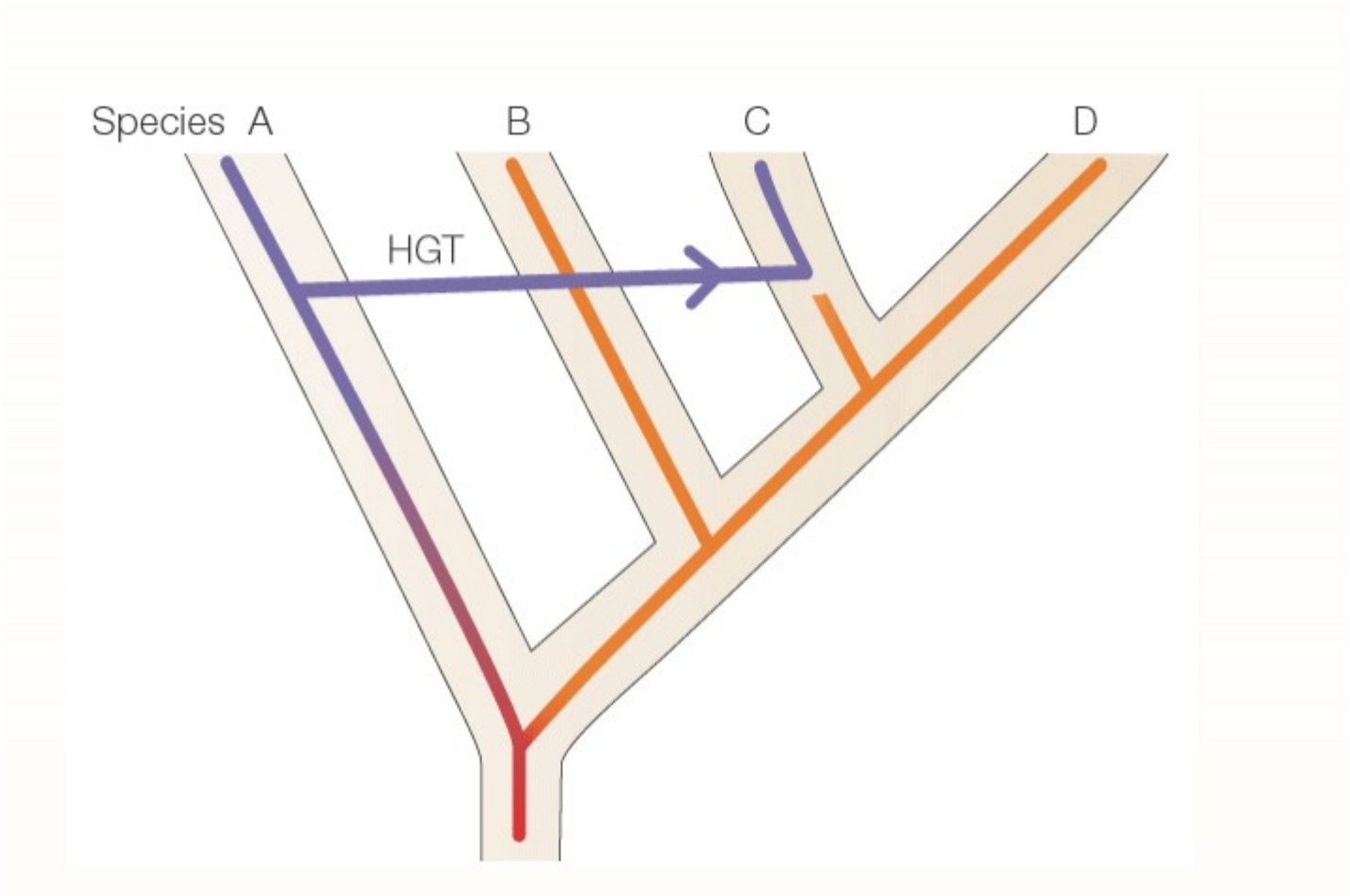


“Quantum leaps” in evolution:

- Gene **(Genome) duplication**
- (endo) symbiosis
- **hybridization**
- Lateral gene transfer



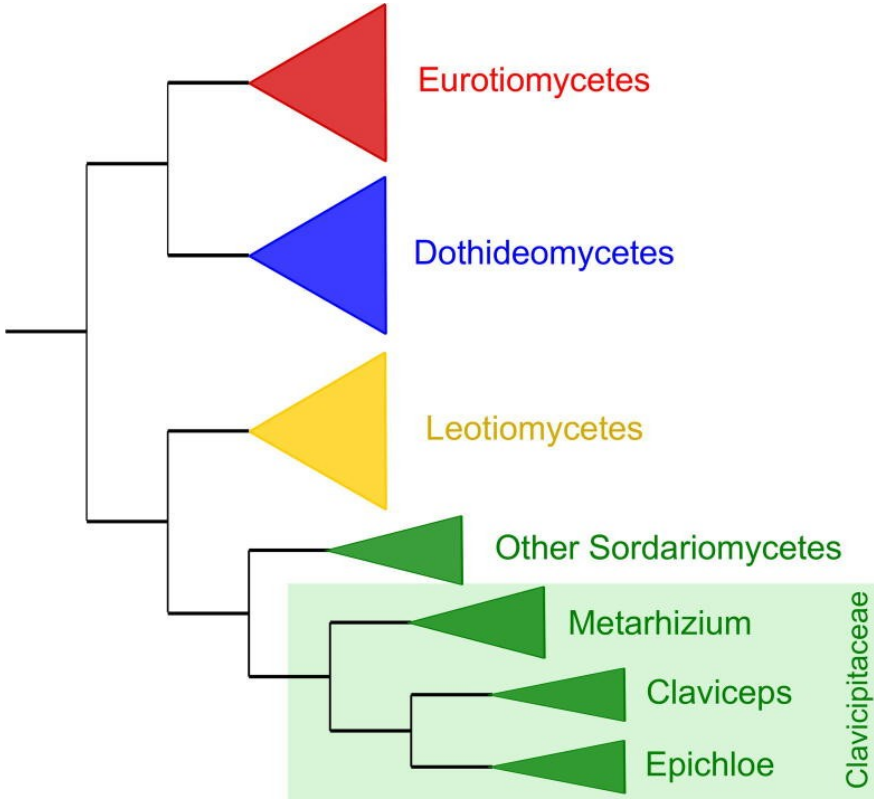
Horizontal Gene Transfer



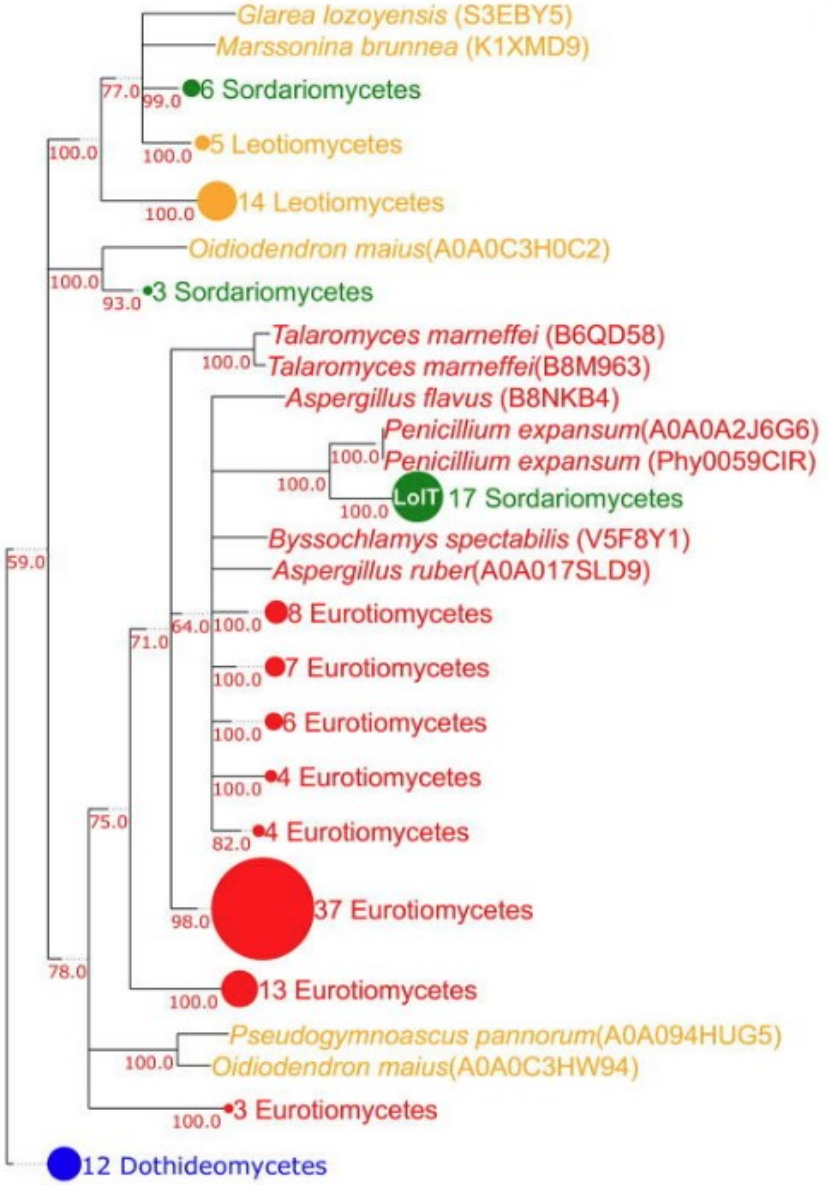
Fungal Genet Biol. 2016 Jan; 86: 71–80.
doi: 10.1016/j.fgb.2015.12.006
PMCID: PMC4712825
PMID: 26688468

Horizontal acquisition of toxic alkaloid synthesis in a clade of plant associated fungi

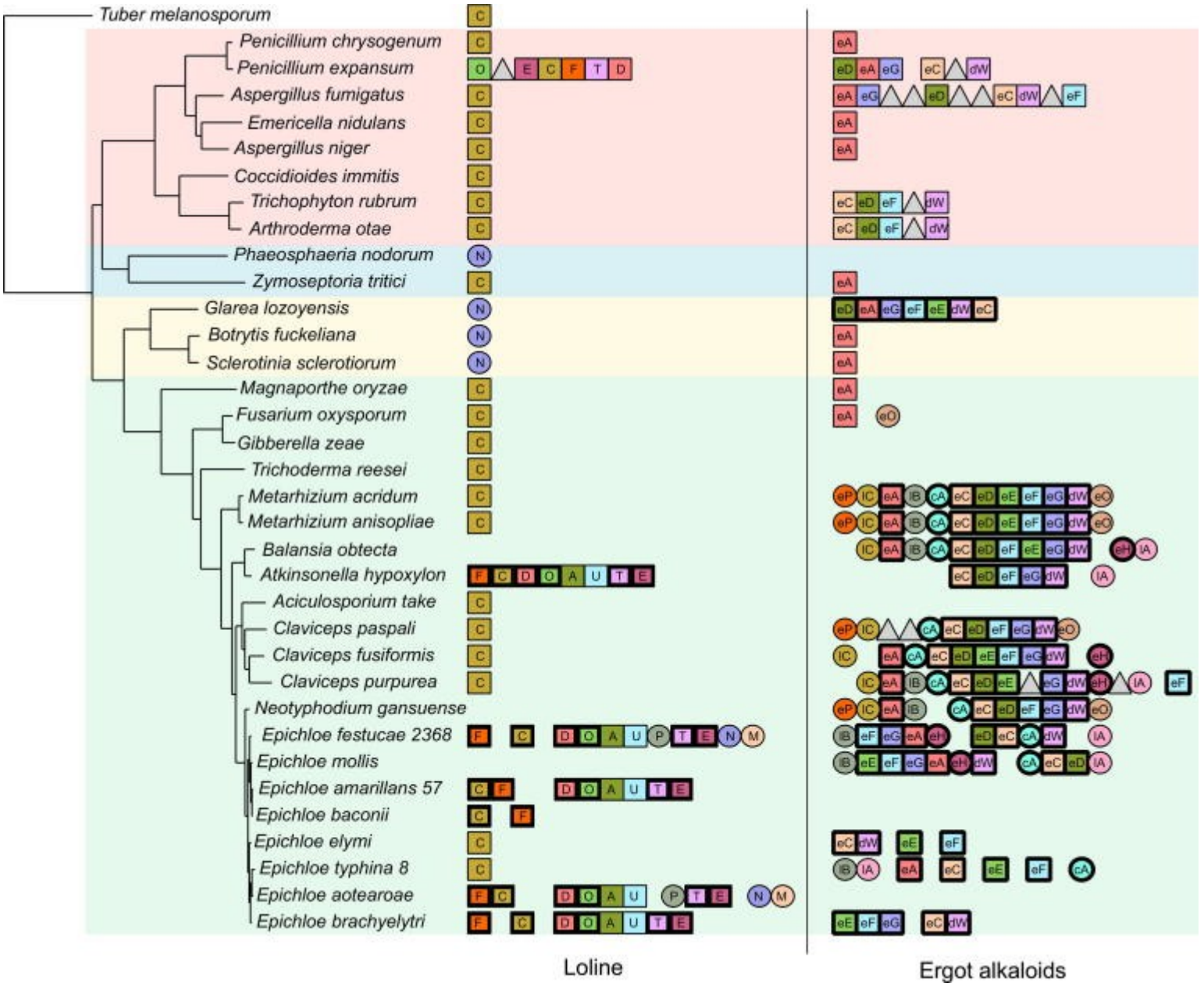
Marina Marcet-Houben[a,b] and Toni Gabaldón[a,b,c,*]



Species tree

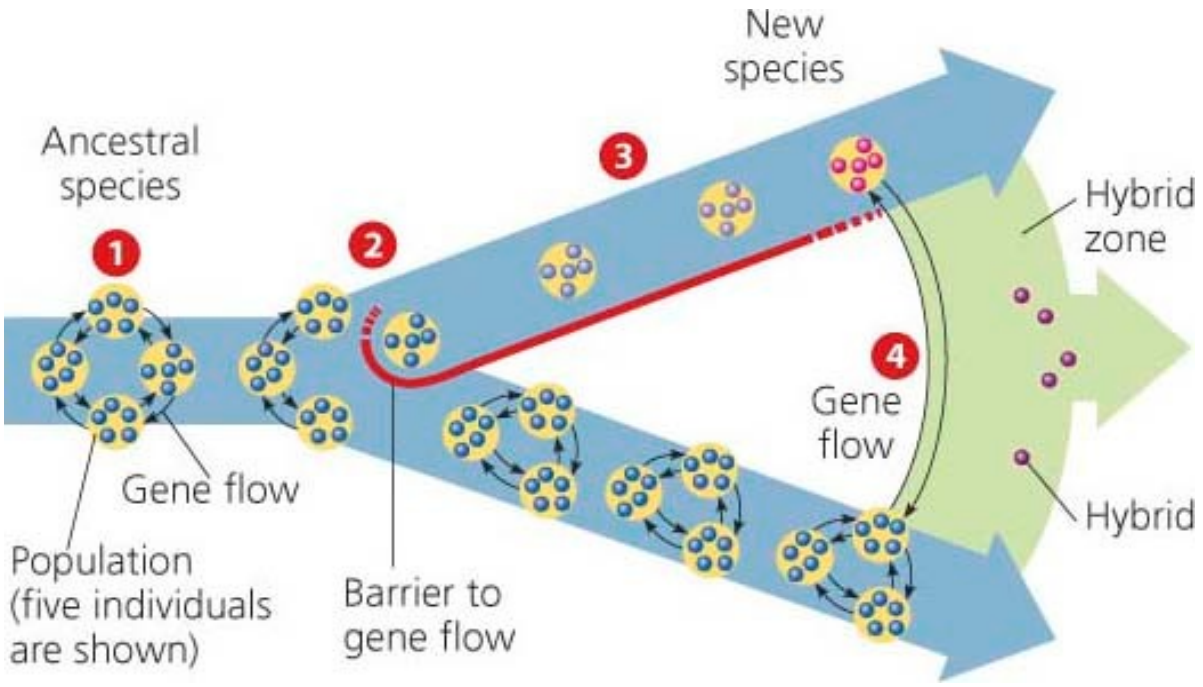


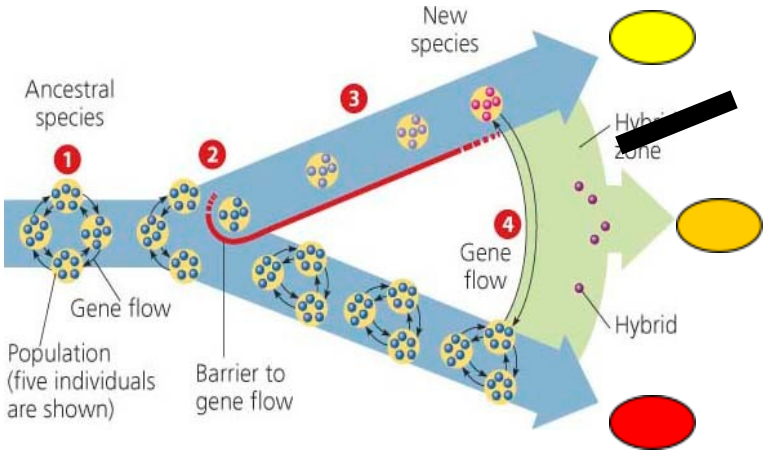
Gene tree

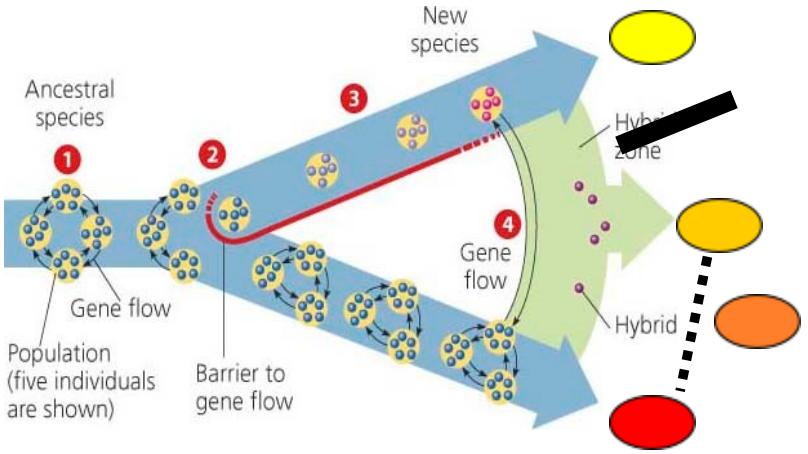

Tuber melanosporum
Penicillium chrysogenum
Penicillium expansum
Aspergillus fumigatus
Emericella nidulans
Aspergillus niger
Coccidioides immitis
Trichophyton rubrum
Arthroderma otae
Phaeosphaeria nodorum
Zymoseptoria tritici
Glarea lozoyensis
Botrytis fuckeliana
Sclerotinia sclerotiorum
Magnaporthe oryzae
Fusarium oxysporum
Gibberella zeae
Trichoderma reesei
Metarhizium acridum
Metarhizium anisopliae
Balansia obtecta
Atkinsonella hypoxylon
Aciculosporium take
Claviceps paspali
Claviceps fusiformis
Claviceps purpurea
Neotyphodium gansuense
Epichloe festucae 2368
Epichloe mollis
Epichloe amarillans 57
Epichloe baconii
Epichloe elymi
Epichloe typhina 8
Epichloe aotearoae
Epichloe brachyelytri
Loline
Ergot alkaloids

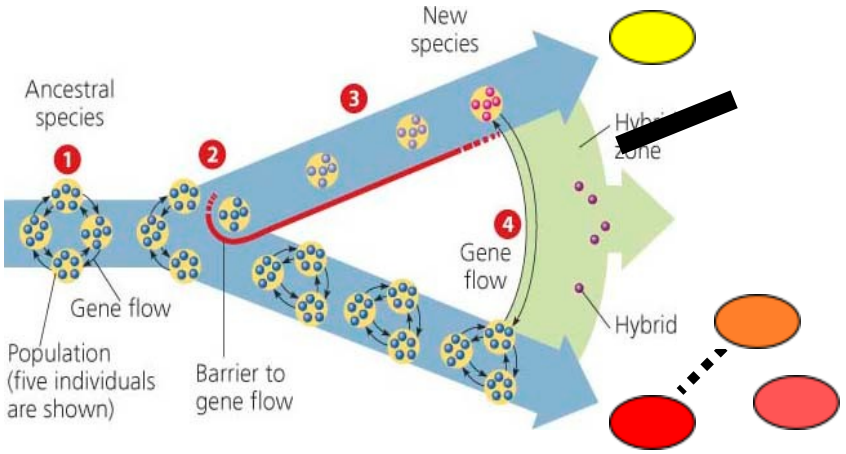
Hybrid concept is based on the concept of species (reproductive isolation)

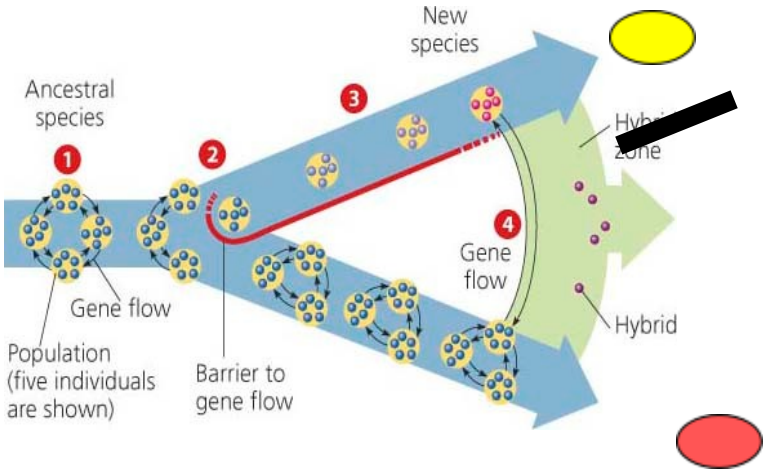
Hybridization




New species
Ancestral species
3
2
1
4
Gene flow
Gene flow
Hybrid
Population (five individuals are shown)
Barrier to gene flow

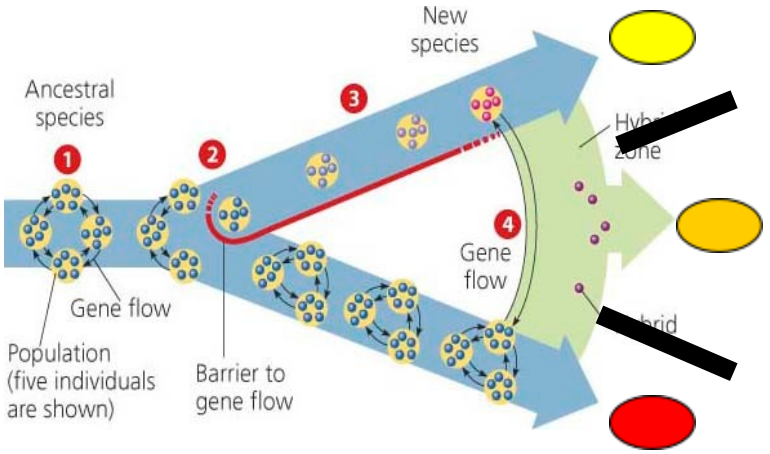

New species
Ancestral species
3
2
1
4
Gene flow
Hybrid
Gene flow
Population (five individuals are shown)
Barrier to gene flow

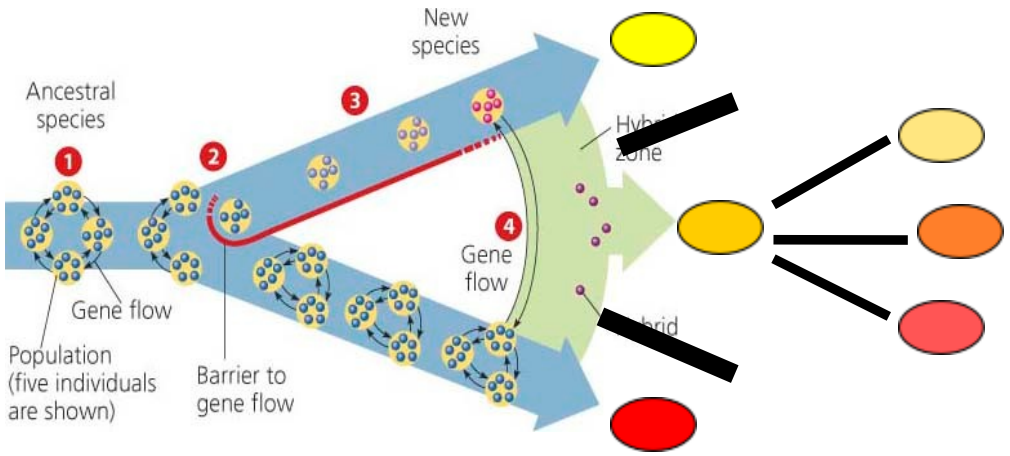

New species
Ancestral species
3
2
1
4
Gene flow
Hybrid
Gene flow
Population (five individuals are shown)
Barrier to gene flow

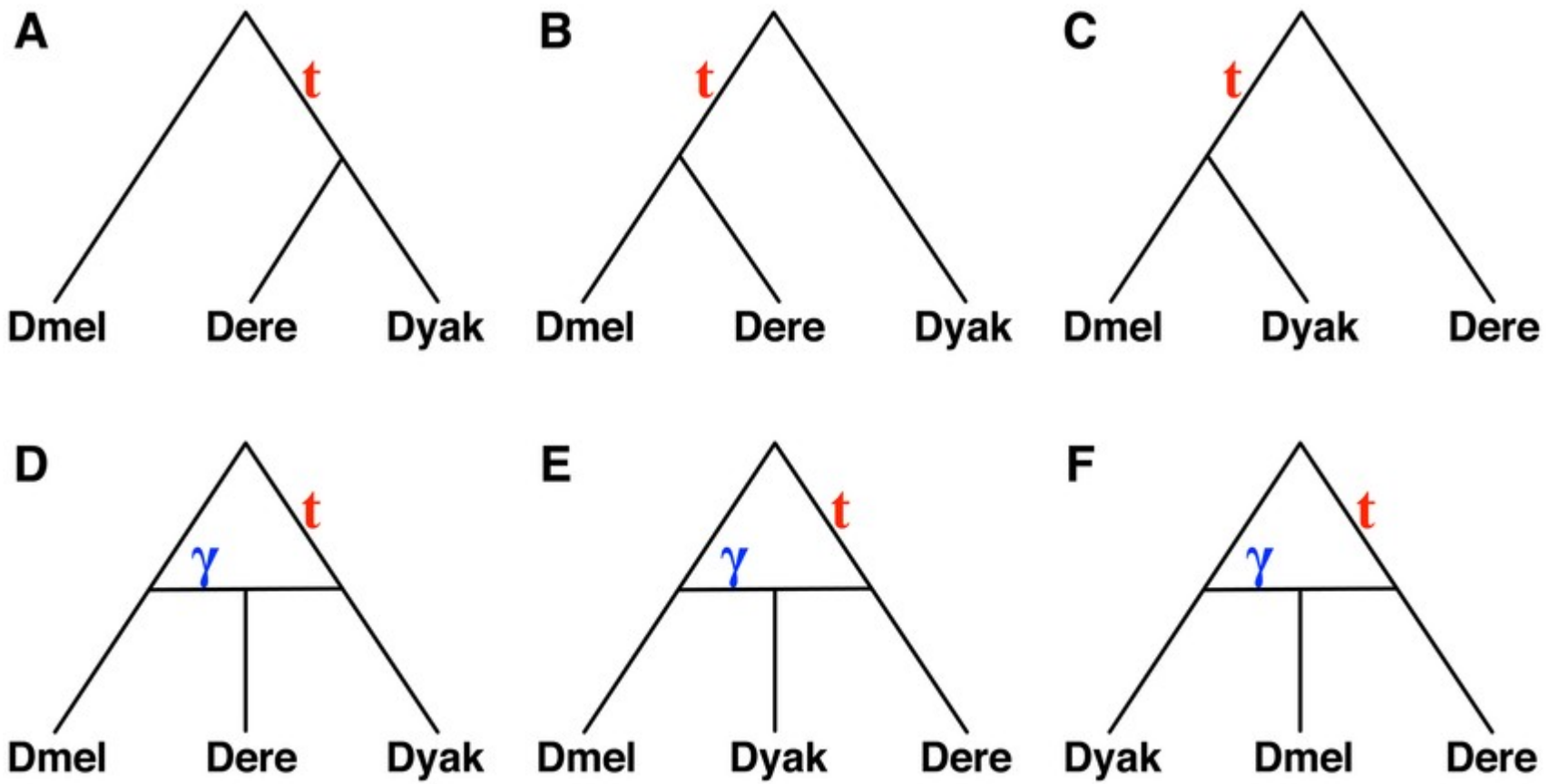


Introgression

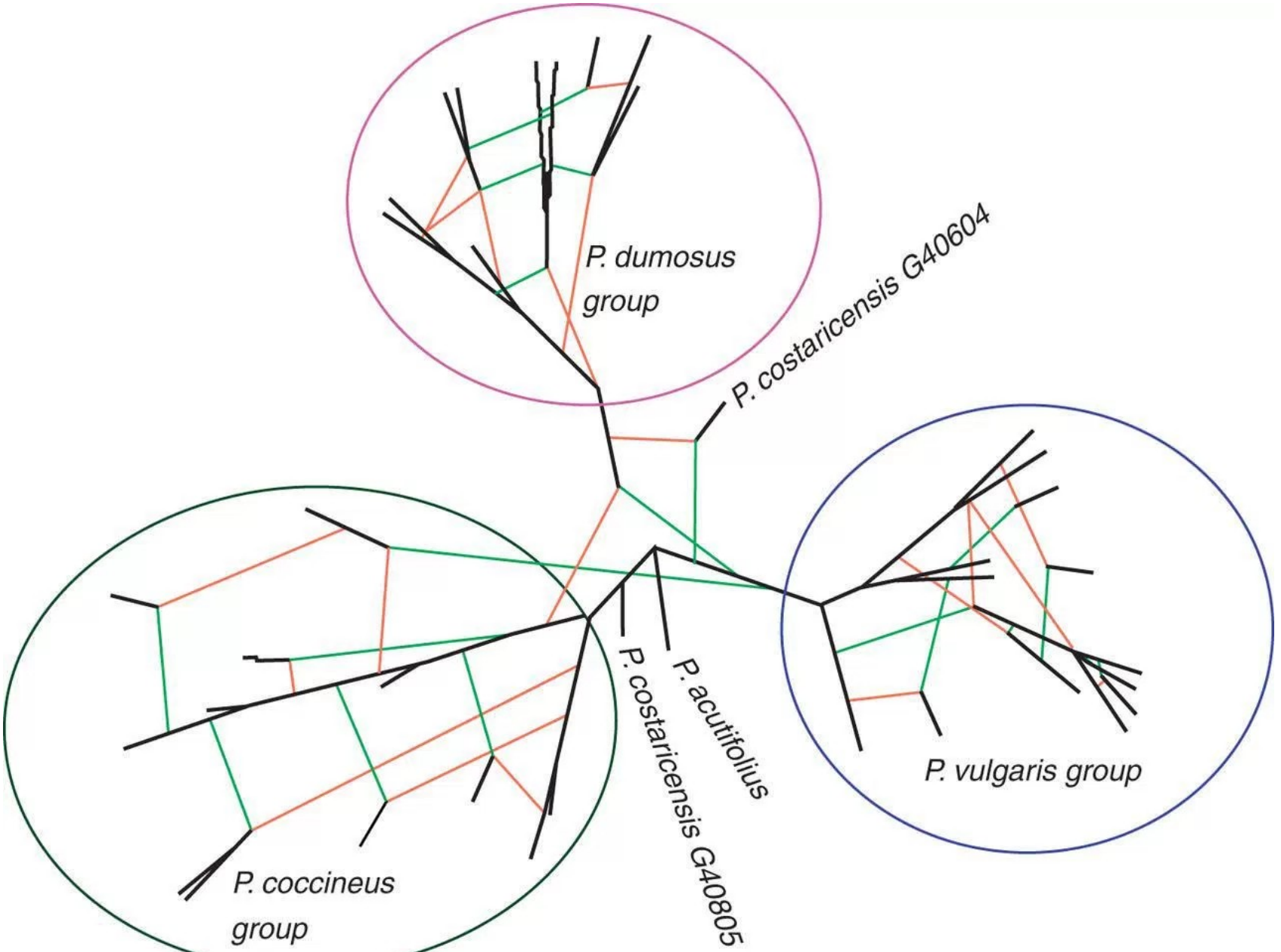
When part of the other separated species genome becomes part of the genome of the species


New species
Ancestral species
3
2
1
4
Gene flow
Gene flow
Population (five individuals are shown)
Barrier to gene flow

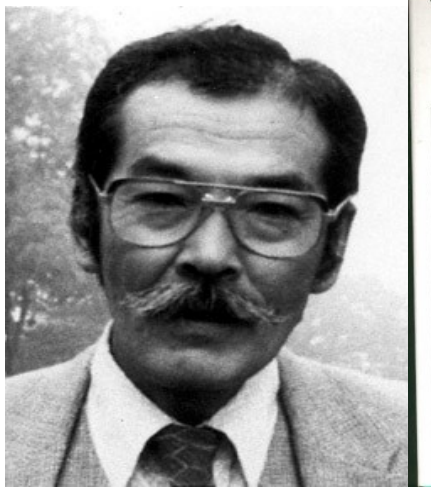

New species
Ancestral species
3
2
1
4
Gene flow
Gene flow
Population (five individuals are shown)
Barrier to gene flow

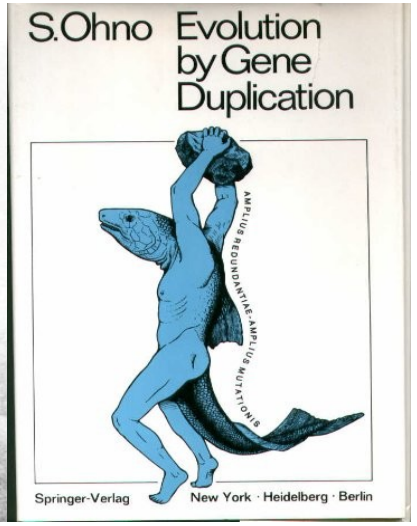

A
t
Dmel
Dere
Dyak
B
t
Dmel
Dere
Dyak
C
t
Dmel
Dyak
Dere
D
t
γ
Dmel
Dere
Dyak
E
t
γ
Dmel
Dyak
Dere
F
t
γ
Dyak
Dmel
Dere

Hybridization can lead to networks rather than trees

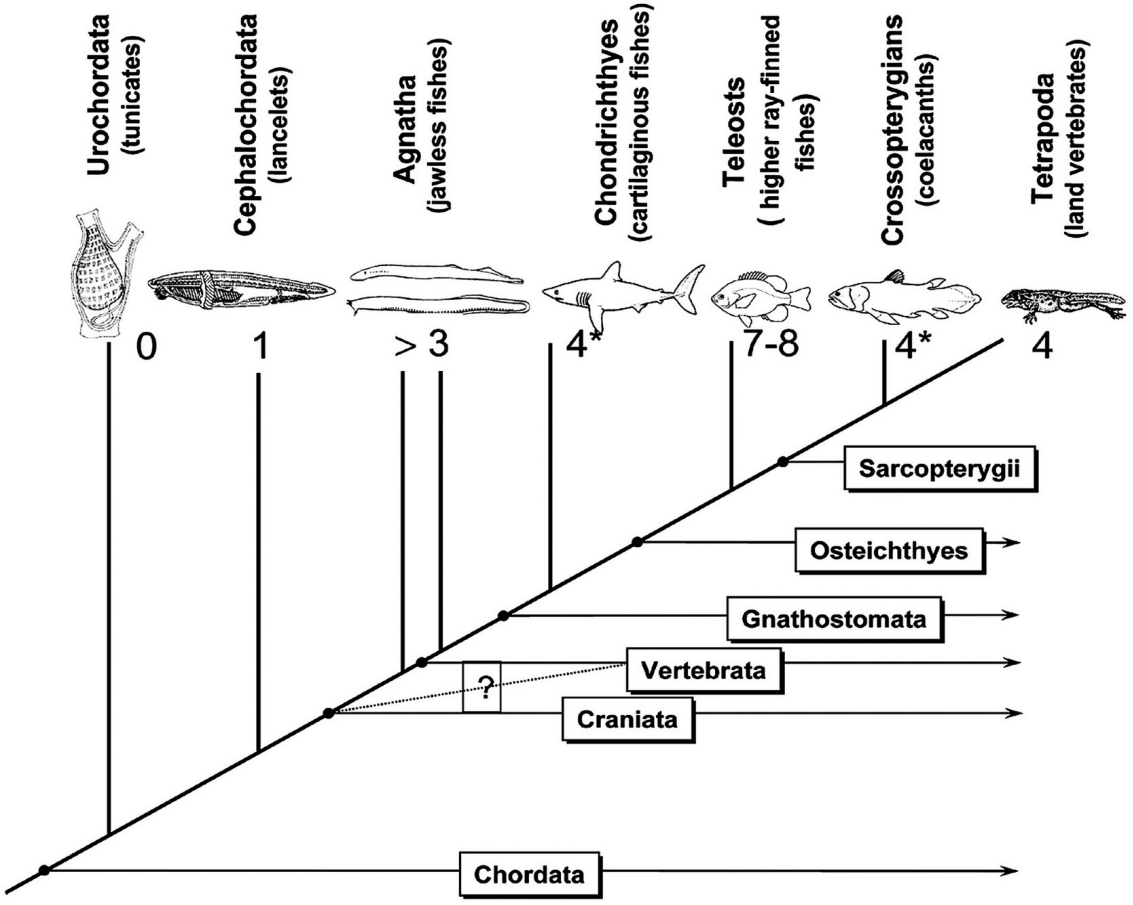


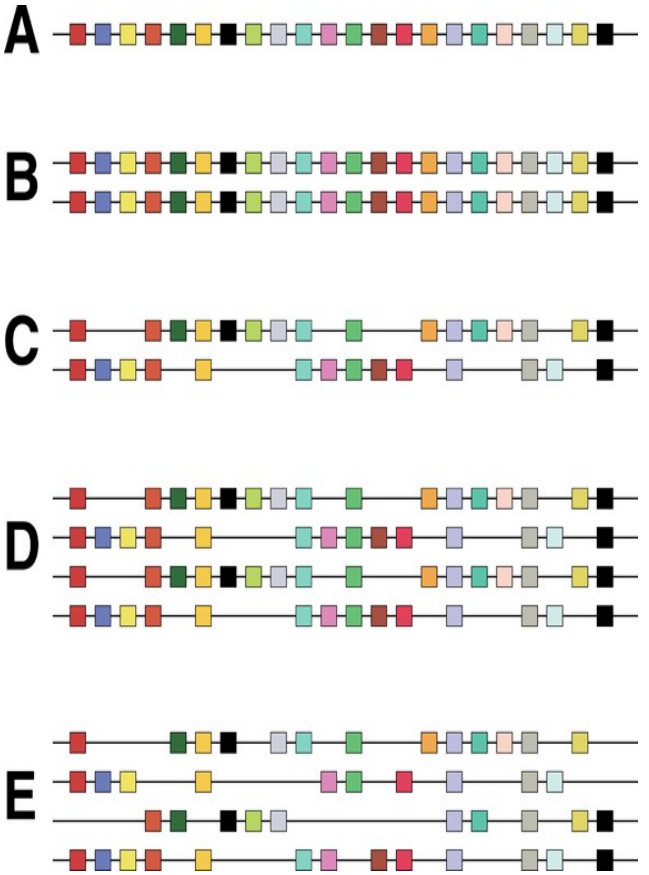
Whole genome duplications

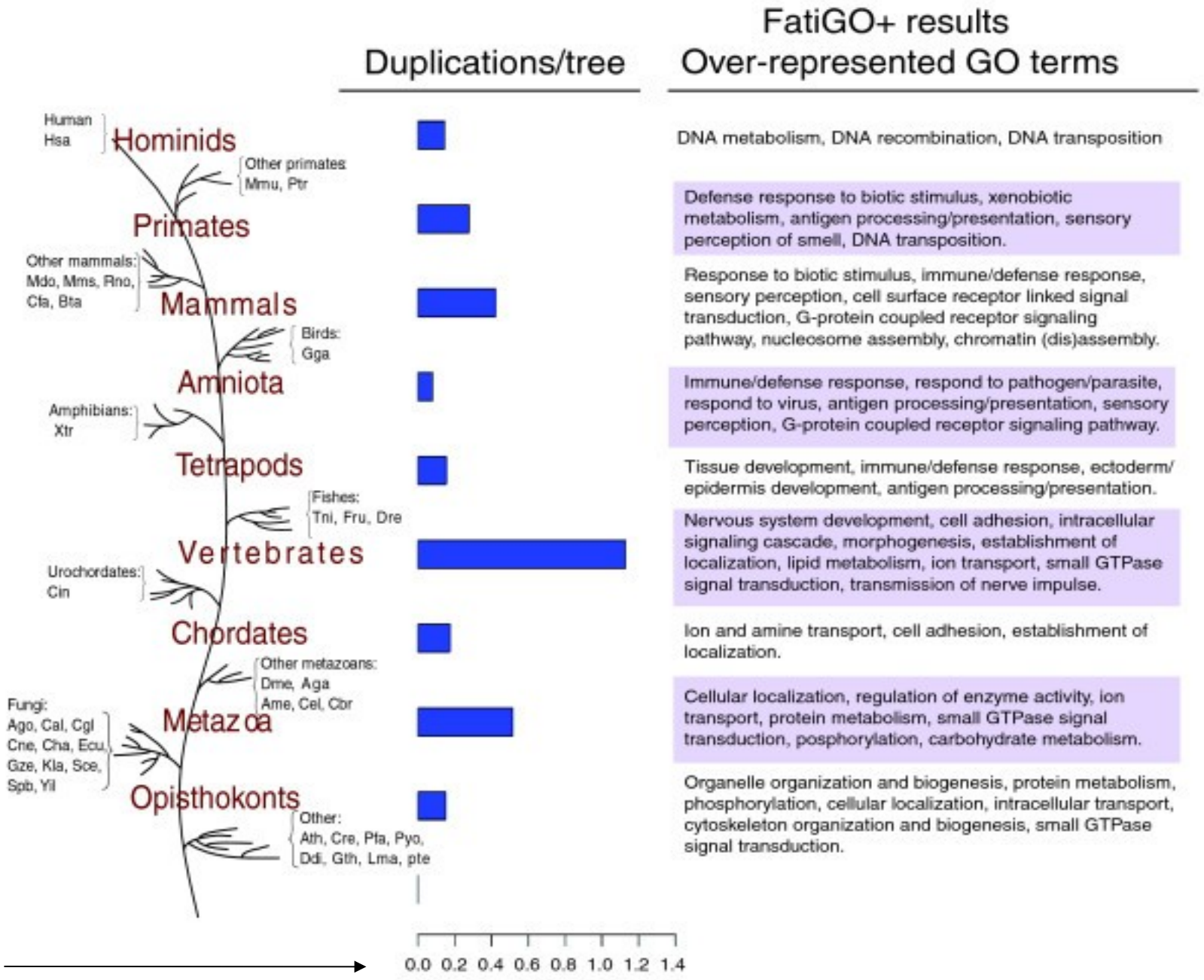




Susumu Ohno (1970)



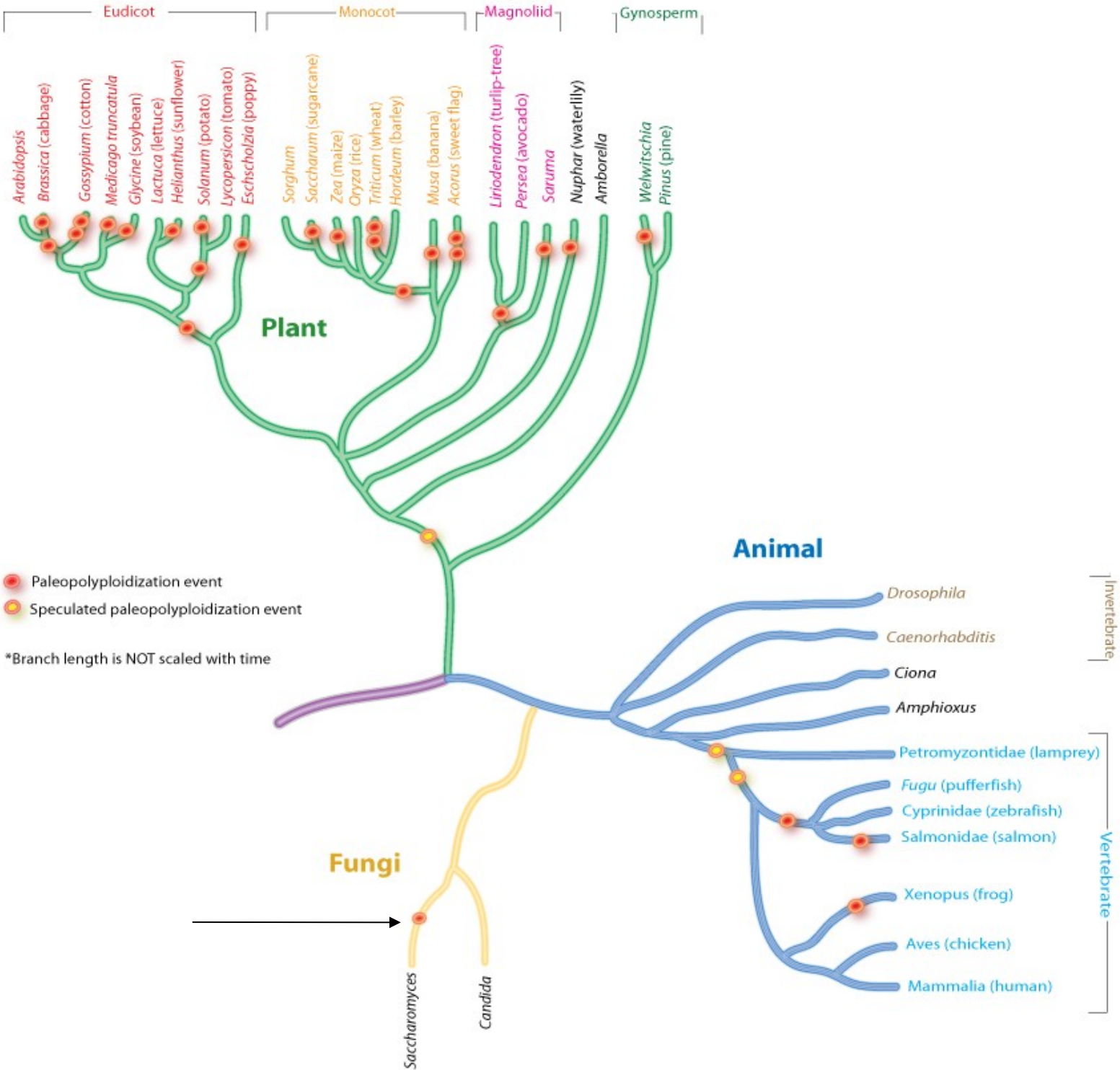


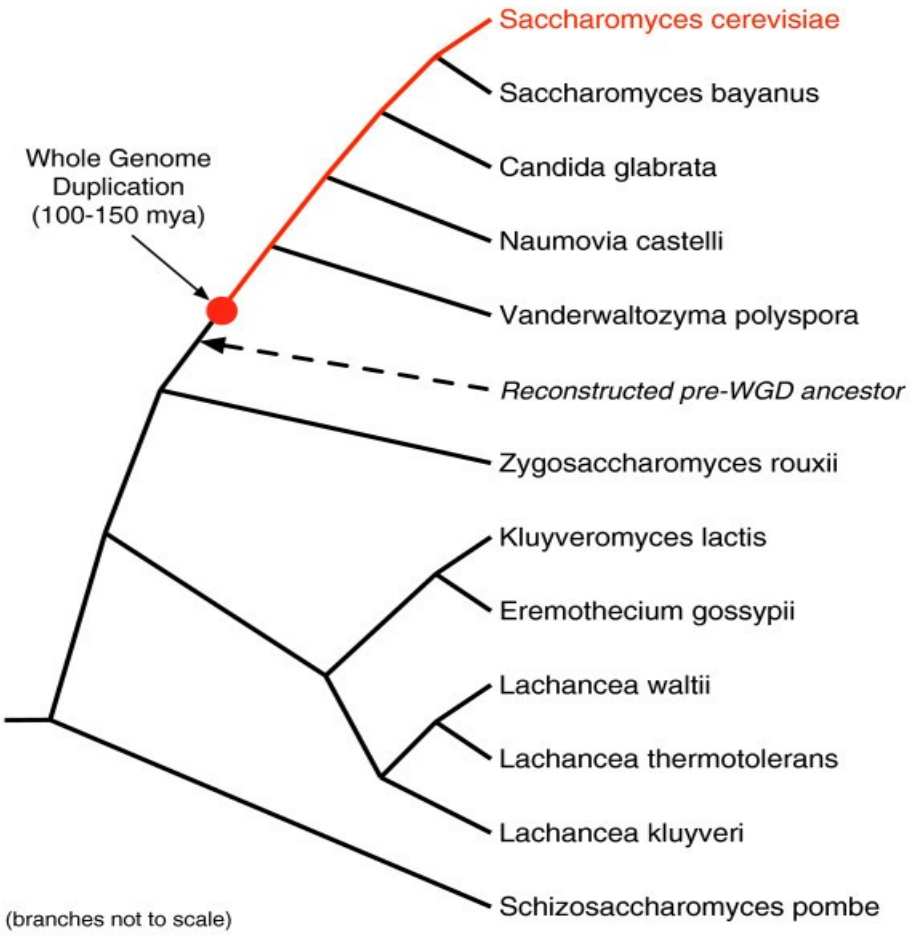


Duplication rate: Duplications per gene in that lineage

Huerta-Cepas et. al. (2007)

Known Paleopolyploidy in Eukaryotes





Article

Nature **428**, 617-624 (8 April 2004) | doi:10.1038/nature02424; Received 17 December 2003; Accepted 19 January 2004; Published online 7 March 2004

Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*

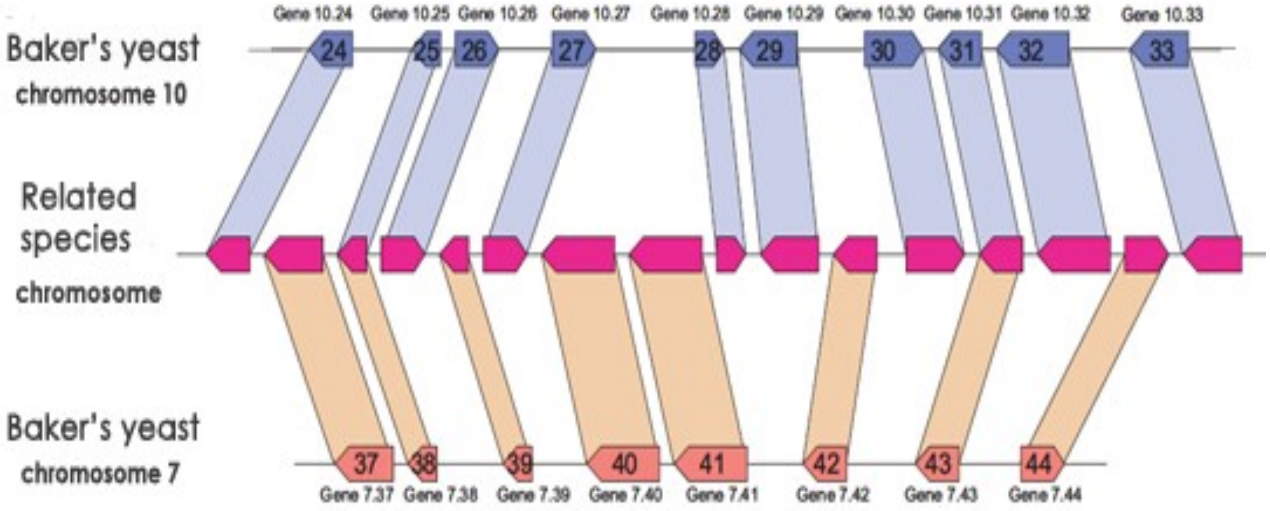
Manolis Kellis[1,2], Bruce W. Birren[1] & Eric S. Lander[1,3]

Letters to Nature

Nature **387**, 708-713 (12 June 1997) | ; Received 17 December 1996; Accepted 20 April 1997

Molecular evidence for an ancient duplication of the entire yeast genome

Kenneth H. Wolfe[1] & Denis C. Shields[1]



Hybridizations can lead to whole genome duplications