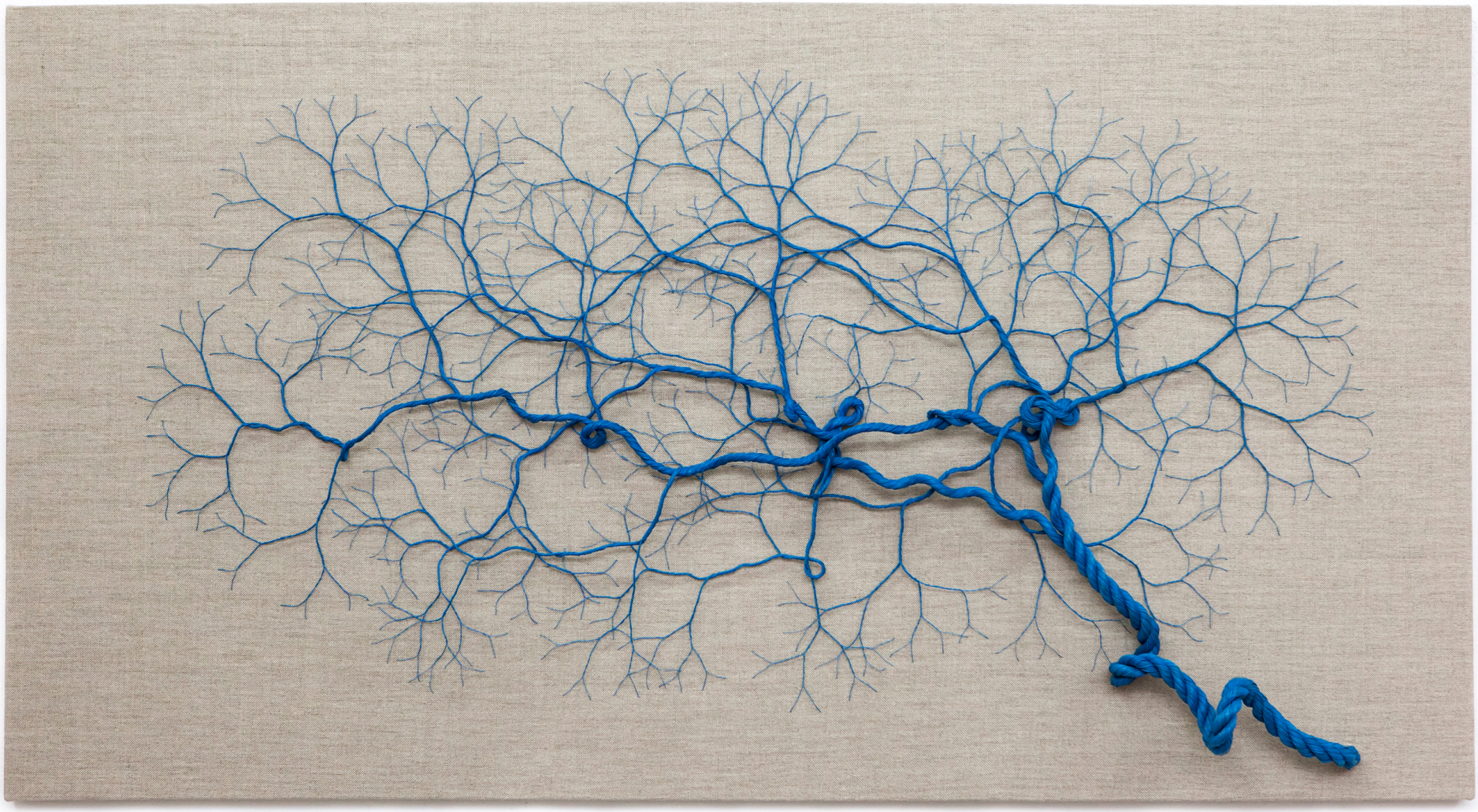
Phylogenomics — *"why we are doing it all wrong"*

with an emphasis on Horizontal Gene Transfer

Gergely Szöllősi

Model-Based Evolutionary Genomics Research Group
Okinawa Institute of Science and Technology
Okinawa, Japan

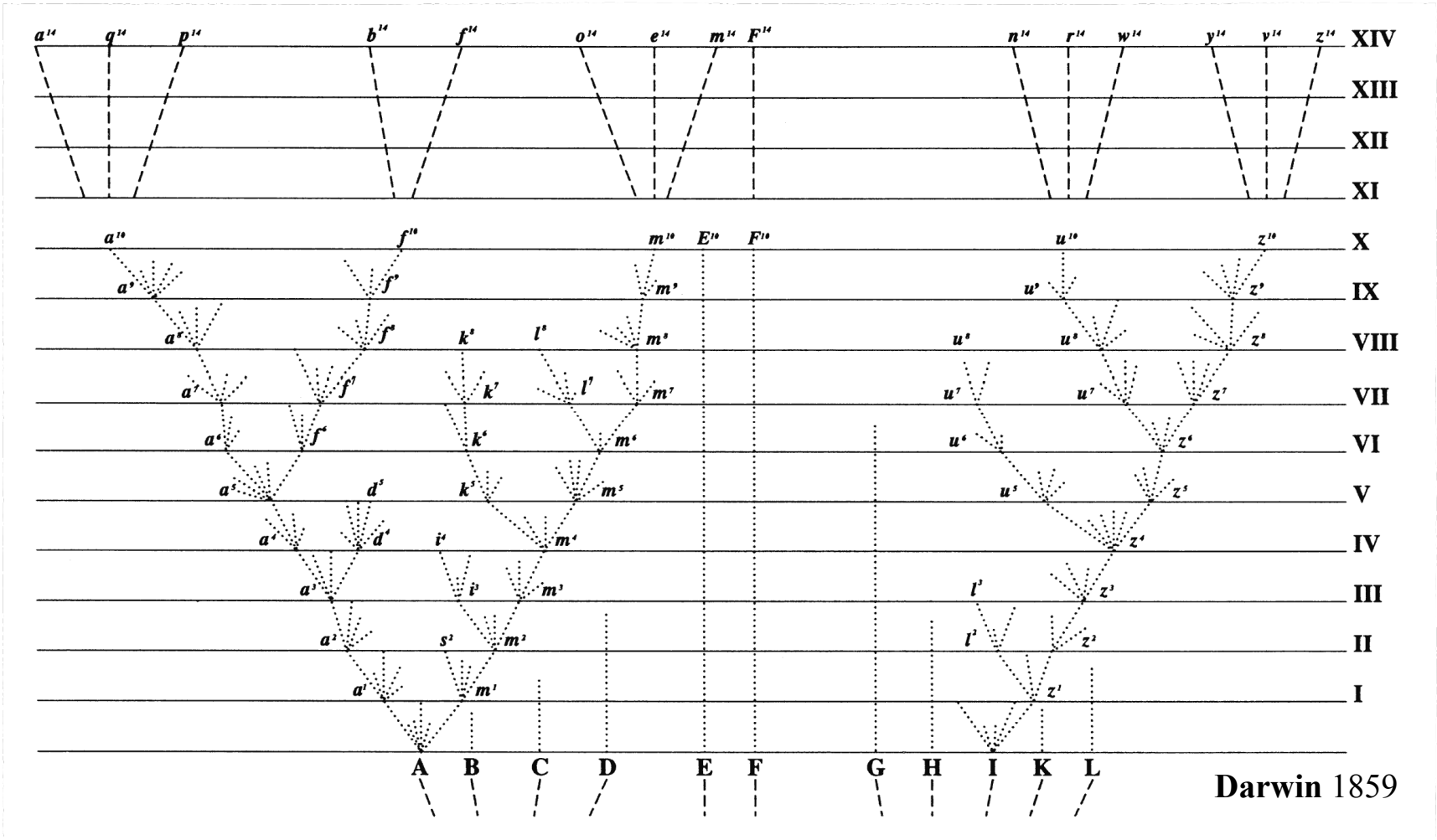
gergely.szollosi@oist.jp



Biology is a historical science

The structure of biological systems, from molecules to ecosystems, is the result of a long evolutionary process that began over 3 billion years ago. The idea that **studying the pattern and process of evolution is the key to understanding living systems** is now becoming widely accepted.

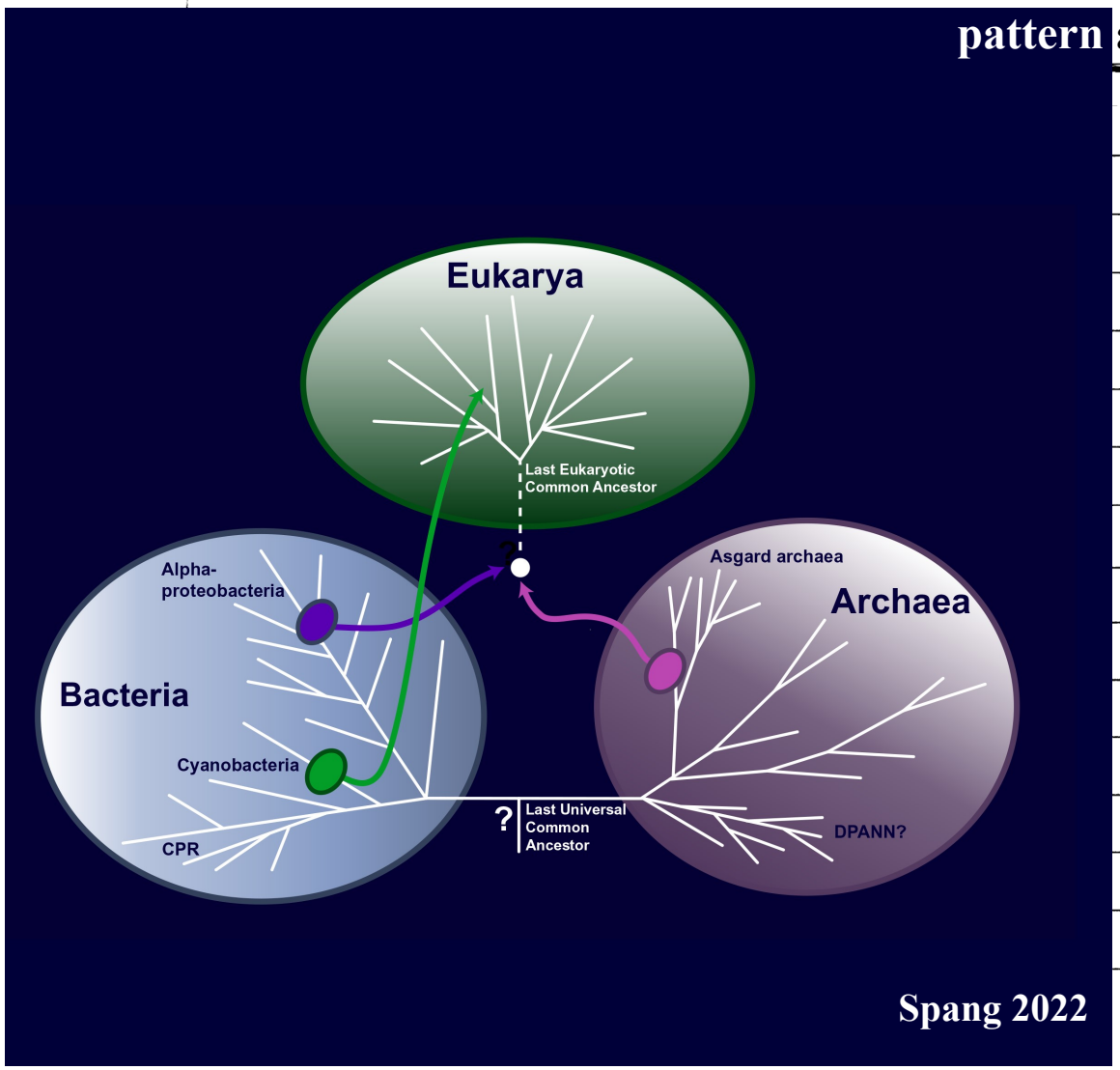
process

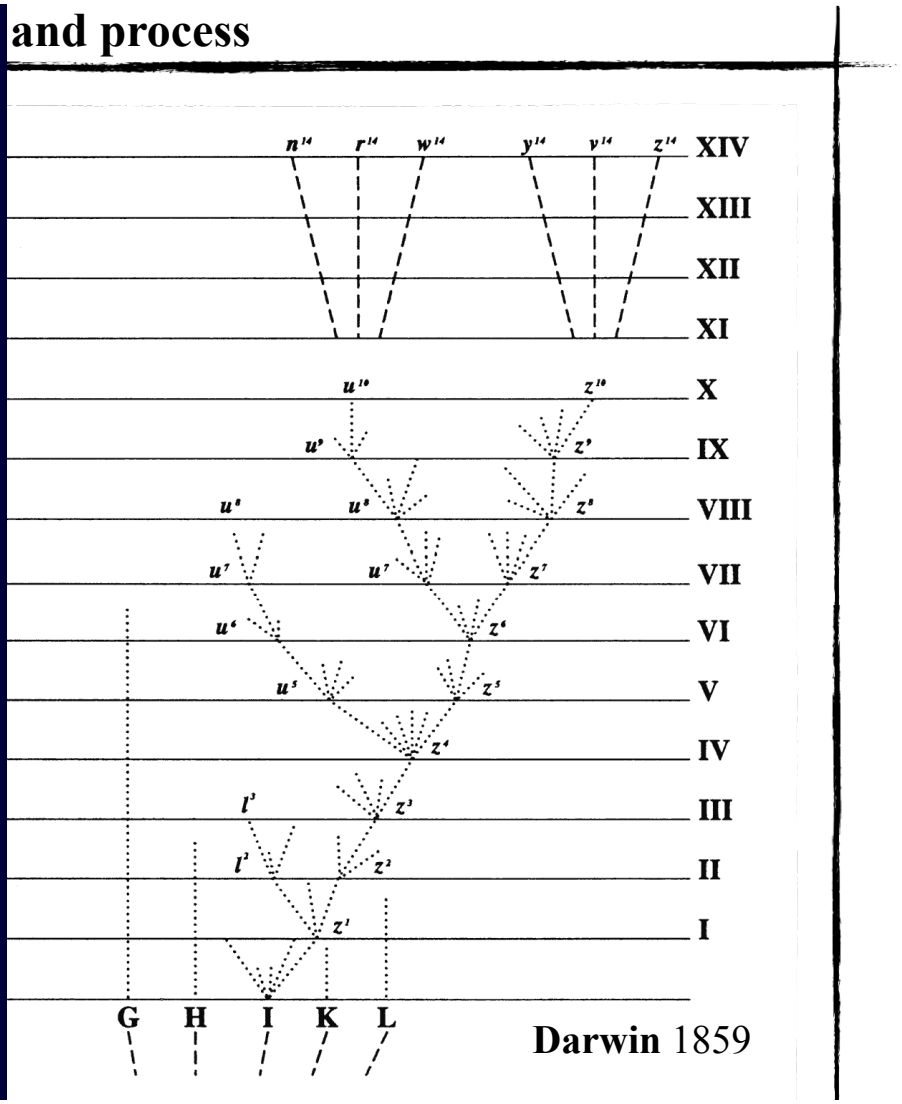


Darwin 1859

Biology is a historical science

The structure of biological systems, from molecules to ecosystems, is the result of a long evolutionary process that began over 3 billion years ago. The idea that **studying the pattern and process of evolution is the key to understanding living systems** is now becoming widely accepted.





J. Theoret. Biol. (1965) **8**, 357–366

Molecules as Documents of Evolutionary History

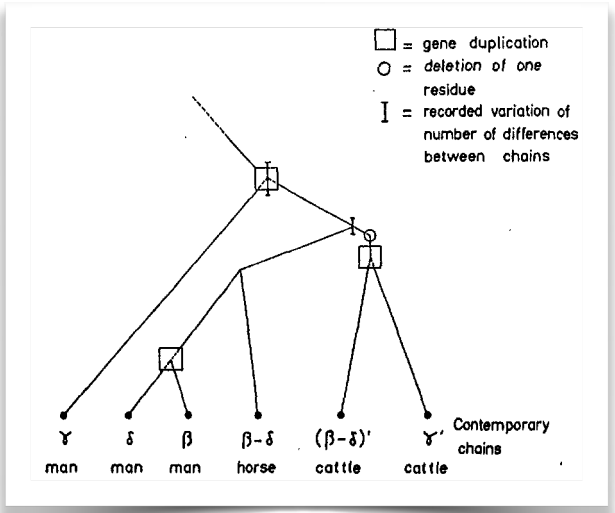
EMILE ZUCKERKANDL AND LINUS PAULING

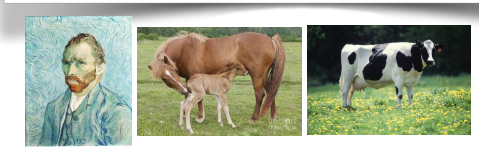
Gates and Crellin Laboratories of Chemistry,
California Institute of Technology, Pasadena, California, U.S.A.

We [..] ask the questions where in the now living systems the greatest amount of their past history has survived and how it can be extracted.[..]

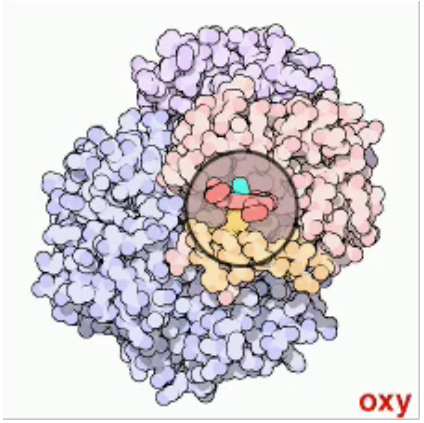
Best fit are the "semantides", i.e. the different types of macromolecules that carry the genetic information or a very extensive translation thereof. [..]

Using Hegel's expression, we may say that there is no other system that is better aufgehoben (constantly abolished and simultaneously preserved).





hemoglobin



J. Theoret. Biol. (1965) **8**, 357–366

Molecules as Documents of Evolutionary History

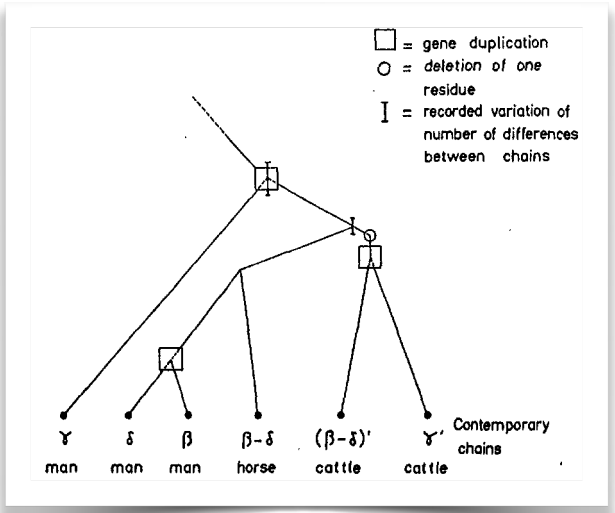
EMILE ZUCKERKANDL AND LINUS PAULING

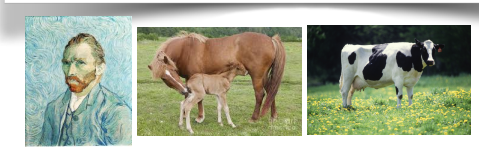
Gates and Crellin Laboratories of Chemistry,
California Institute of Technology, Pasadena, California, U.S.A.

We [..] ask the questions where in the now living systems the greatest amount of their past history has survived and how it can be extracted.[..]

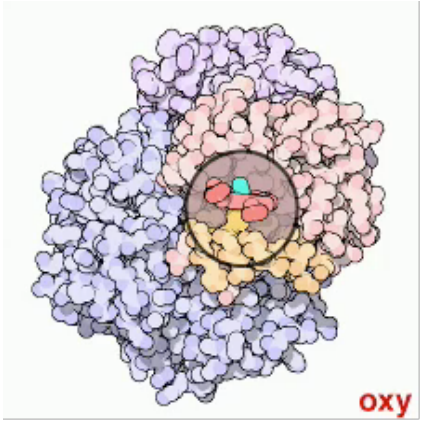
Best fit are the "semantides", i.e. the different types of macromolecules that carry the genetic information or a very extensive translation thereof. [..]

Using Hegel's expression, we may say that there is no other system that is better aufgehoben (constantly abolished and simultaneously preserved).





hemoglobin



Prelude

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 11, pp. 5088–5090, November 1977
Evolution

Phylogenetic structure of the prokaryotic domain: The primary kingdoms

(archaebacteria/eubacteria/urkaryote/16S ribosomal RNA/molecular phylogeny)

CARL R. WOESE AND GEORGE E. FOX*

Department of Genetics and Development, University of Illinois, Urbana, Illinois 61801

Communicated by T. M. Sonneborn, August 18, 1977

Ribonuclease T1 oligonucleotide fingerprints from SSU rRNA (1000-2000 nt long)

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 11, pp. 5088–5090, November 1977
Evolution

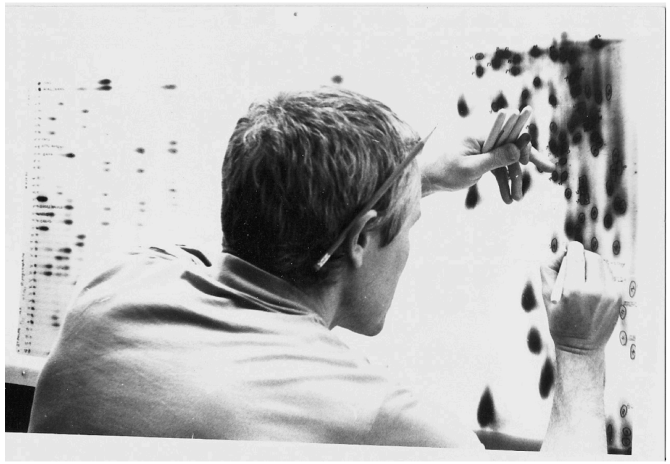
Phylogenetic structure of the prokaryotic domain: The primary kingdoms

(archaebacteria/eubacteria/urkaryote/16S ribosomal RNA/molecular phylogeny)

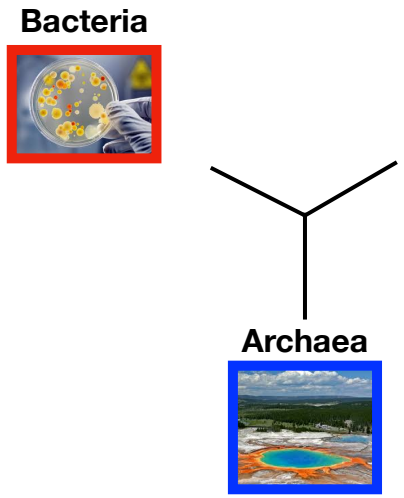
CARL R. WOESE AND GEORGE E. FOX*

Department of Genetics and Development, University of Illinois, Urbana, Illinois 61801

Communicated by T. M. Sonneborn, August 18, 1977



Photograph of Carl working on a fingerprint, circa 1976. (Courtesy of Ken Luehrsen.)



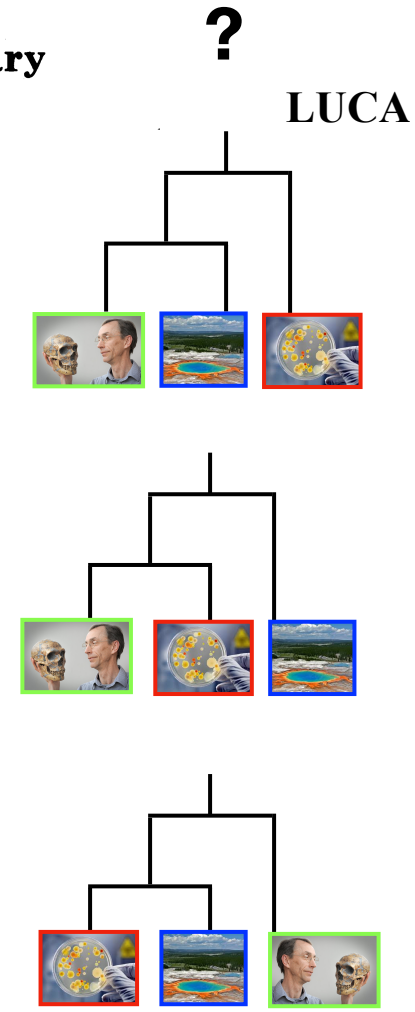
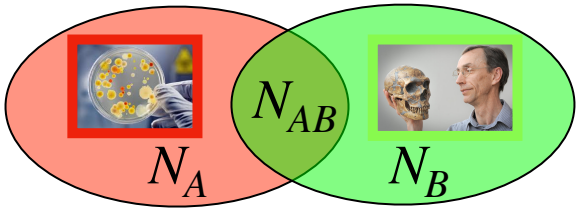


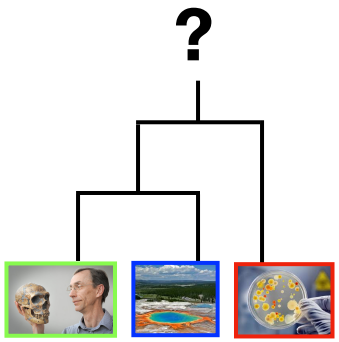
Table 1. Association coefficients (S_{AB}) between representative members of the three primary kingdoms

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. *Saccharomyces cerevisiae*, 18S	—	0.29	0.33	0.05	0.06	0.08	0.09	0.11	0.08	0.11	0.11	0.08	0.08
2. *Lemna minor*, 18S	0.29	—	0.36	0.10	0.05	0.06	0.10	0.09	0.11	0.10	0.10	0.13	0.07
3. L cell, 18S	0.33	0.36	—	0.06	0.06	0.07	0.07	0.09	0.06	0.10	0.10	0.09	0.07
4. *Escherichia coli*	0.05	0.10	0.06	—	0.24	0.25	0.28	0.26	0.21	0.11	0.12	0.07	0.12
5. *Chlorobium vibrioforme*	0.06	0.05	0.06	0.24	—	0.22	0.22	0.20	0.19	0.06	0.07	0.06	0.09
6. *Bacillus firmus*	0.08	0.06	0.07	0.25	0.22	—	0.34	0.26	0.20	0.11	0.13	0.06	0.12
7. *Corynebacterium diphtheriae*	0.09	0.10	0.07	0.28	0.22	0.34	—	0.23	0.21	0.12	0.12	0.09	0.10
8. *Aphanocapsa* 6714	0.11	0.09	0.09	0.26	0.20	0.26	0.23	—	0.31	0.11	0.11	0.10	0.10
9. Chloroplast (*Lemna*)	0.08	0.11	0.06	0.21	0.19	0.20	0.21	0.31	—	0.14	0.12	0.10	0.12
10. *Methanobacterium thermoautotrophicum*	0.11	0.10	0.10	0.11	0.06	0.11	0.12	0.11	0.14	—	0.51	0.25	0.30
11. *M. ruminantium* strain M-1	0.11	0.10	0.10	0.12	0.07	0.13	0.12	0.11	0.12	0.51	—	0.25	0.24
12. *Methanobacterium* sp., Cariaco isolate JR-1	0.08	0.13	0.09	0.07	0.06	0.06	0.09	0.10	0.10	0.25	0.25	—	0.32
13. *Methanosarcina barkeri*	0.08	0.07	0.07	0.12	0.09	0.12	0.10	0.10	0.12	0.30	0.24	0.32	—

The 16S (18S) ribosomal RNA from the organisms (organelles) listed were digested with T1 RNase and the resulting digests were subjected to two-dimensional electrophoretic separation to produce an oligonucleotide fingerprint. The individual oligonucleotides on each fingerprint were then sequenced by established procedures (13, 14) to produce an oligonucleotide catalog characteristic of the given organism (3, 4, 13–17, 22, 23; unpublished data). Comparisons of all possible pairs of such catalogs defines a set of association coefficients (S_{AB}) given by: $S_{AB} = 2N_{AB}/(N_A + N_B)$, in which N_A, N_B, and N_{AB} are the total numbers of nucleotides in sequences of hexamers or larger in the catalog for organism A, in that for organism B, and in the interreaction of the two catalogs, respectively (13, 23).

$$S_{AB} = \frac{N_{AB}}{N_A N_B}$$





Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 6661–6665, September 1989
Evolution

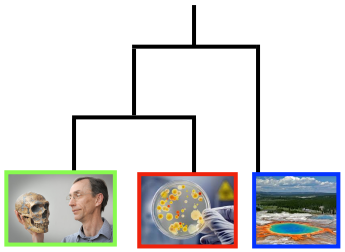
Evolution of the vacuolar H^+-ATPase: Implications for the origin of eukaryotes

(proton pump/membrane ATPase/vacuoles/archaebacteria/eocyte)

JOHANN PETER GOGARTEN*, HENRIK KIBAK*, PETER DITTRICH*†, LINCOLN TAIZ*, EMMA JEAN BOWMAN*, BARRY J. BOWMAN*, MORRIS F. MANOLSON‡§, RONALD J. POOLE‡, TAKAYASU DATE¶, TAIRO OSHIMA‖, JIN KONISHI‖, KIMITOSHI DENDA‖, AND MASASUKE YOSHIDA‖

*Department of Biology, University of California–Santa Cruz, Santa Cruz, CA 95064; ‡Department of Biology, McGill University, 1205 Avenue Docteur Penfield, Montreal, PQ H3A 1B1, Canada; ‖Department of Life Science, Tokyo Institute of Technology, Nagatsuta, Midori-ku, Yokohama 227, Japan; and ¶Department of Biochemistry, Kanazawa Medical School, Uchinada, Ishikawa 920-02, Japan

Communicated by Paul D. Boyer, May 22, 1989



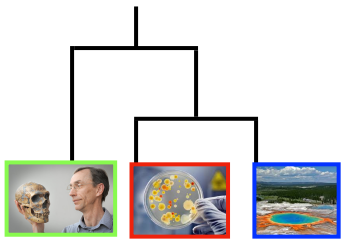
Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 9355–9359, December 1989
Evolution

Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes

NAOYUKI IWABE*, KEI-ICHI KUMA*, MASAMI HASEGAWA†, SYOZO OSAWA‡, AND TAKASHI MIYATA*

*Department of Biology, Faculty of Science, Kyushu University, Fukuoka 812, Japan; †Institute of Statistical Mathematics, Minatoku, Tokyo 106, Japan; ‡Department of Biology, Faculty of Science, Nagoya University, Nagoya 464-01, Japan

Communicated by Motoo Kimura, August 22, 1989

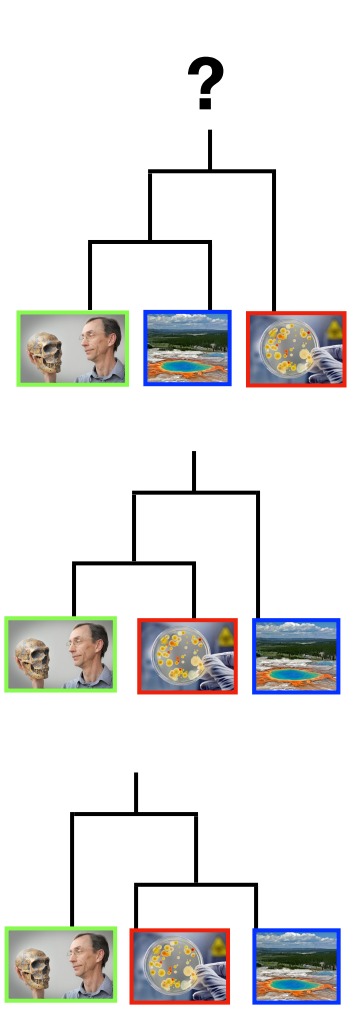


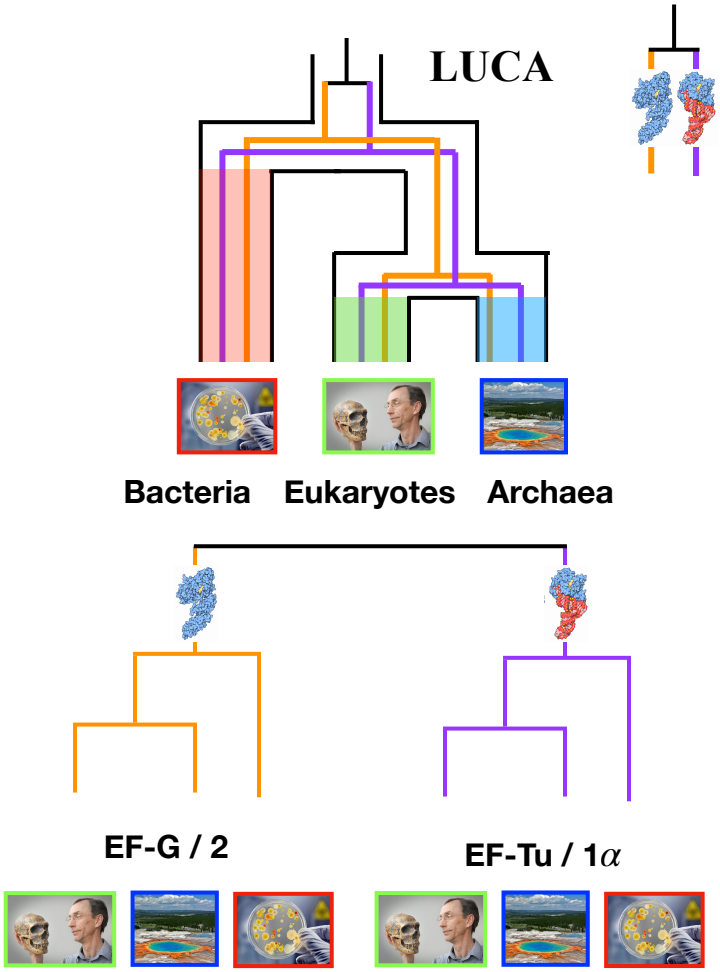
Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes

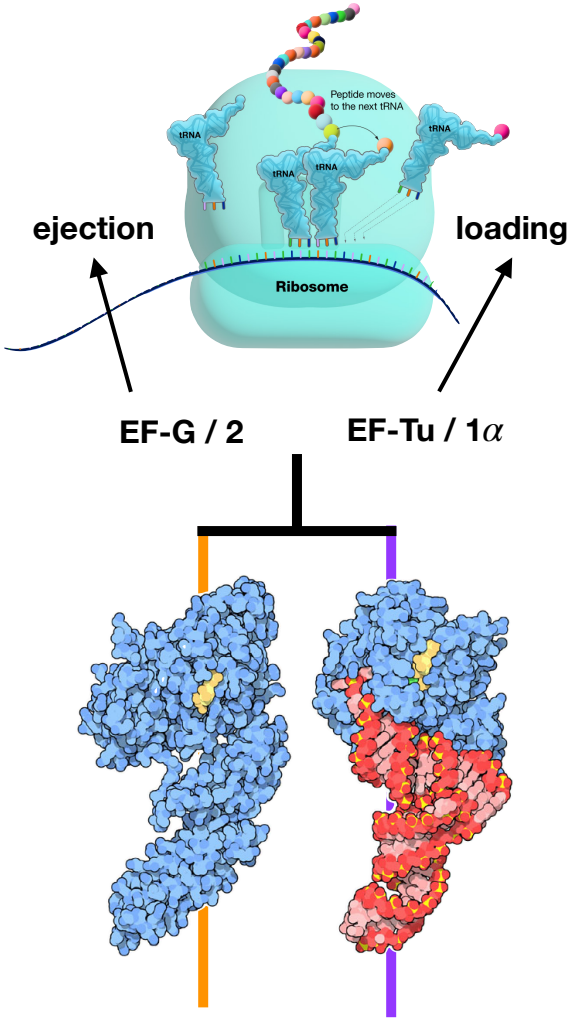
NAOYUKI IWABE*, KEI-ICHI KUMA*, MASAMI HASEGAWA†, SYOZO OSAWA‡, AND TAKASHI MIYATA*

*Department of Biology, Faculty of Science, Kyushu University, Fukuoka 812, Japan; †Institute of Statistical Mathematics, Minatoku, Tokyo 106, Japan; and ‡Department of Biology, Faculty of Science, Nagoya University, Nagoya 464-01, Japan

Communicated by Motoo Kimura, August 22, 1989





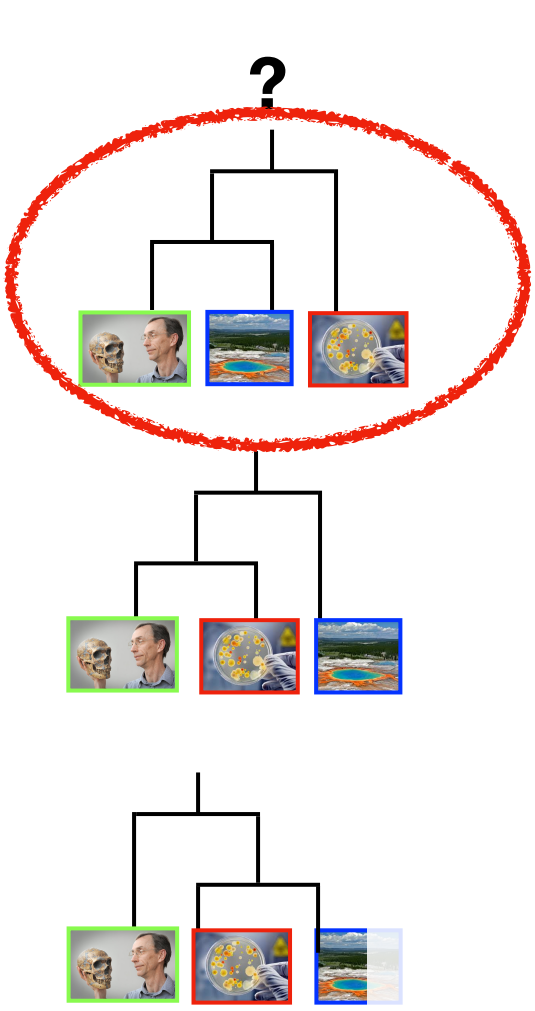


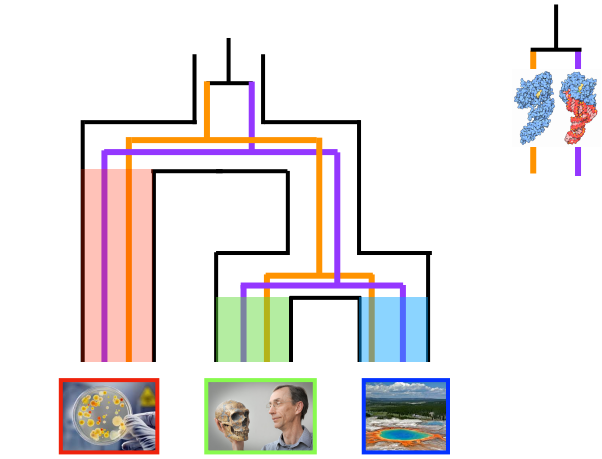
Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes

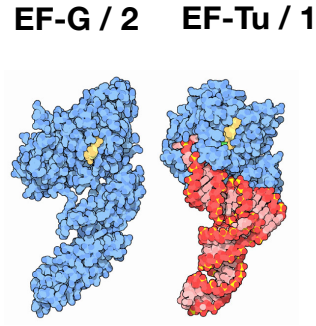
NAOYUKI IWABE*, KEI-ICHI KUMA*, MASAMI HASEGAWA†, SYOZO OSAWA‡, AND TAKASHI MIYATA*

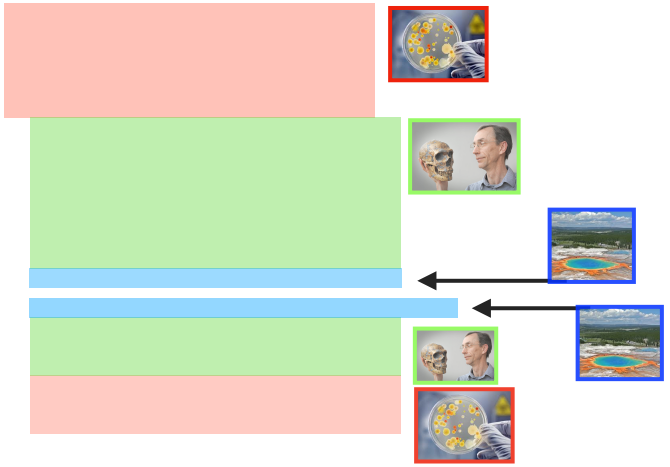
*Department of Biology, Faculty of Science, Kyushu University, Fukuoka 812, Japan; †Institute of Statistical Mathematics, Minatoku, Tokyo 106, Japan; and ‡Department of Biology, Faculty of Science, Nagoya University, Nagoya 464-01, Japan

Communicated by Motoo Kimura, August 22, 1989









Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 4576–4579, June 1990
Evolution

Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya

(Euryarchaeota/Crenarchaeota/kingdom/evolution)

CARL R. WOESE*†, OTTO KANDLER‡, AND MARK L. WHEELIS§

*Department of Microbiology, University of Illinois, 131 Burrill Hall, Urbana, IL 61801; ‡Botanisches Institut der Universität München, Menzinger Strasse 67, 8000 Munich 19, Federal Republic of Germany; and §Department of Microbiology, University of California, Davis, CA 95616

Contributed by Carl R. Woese, March 26, 1990

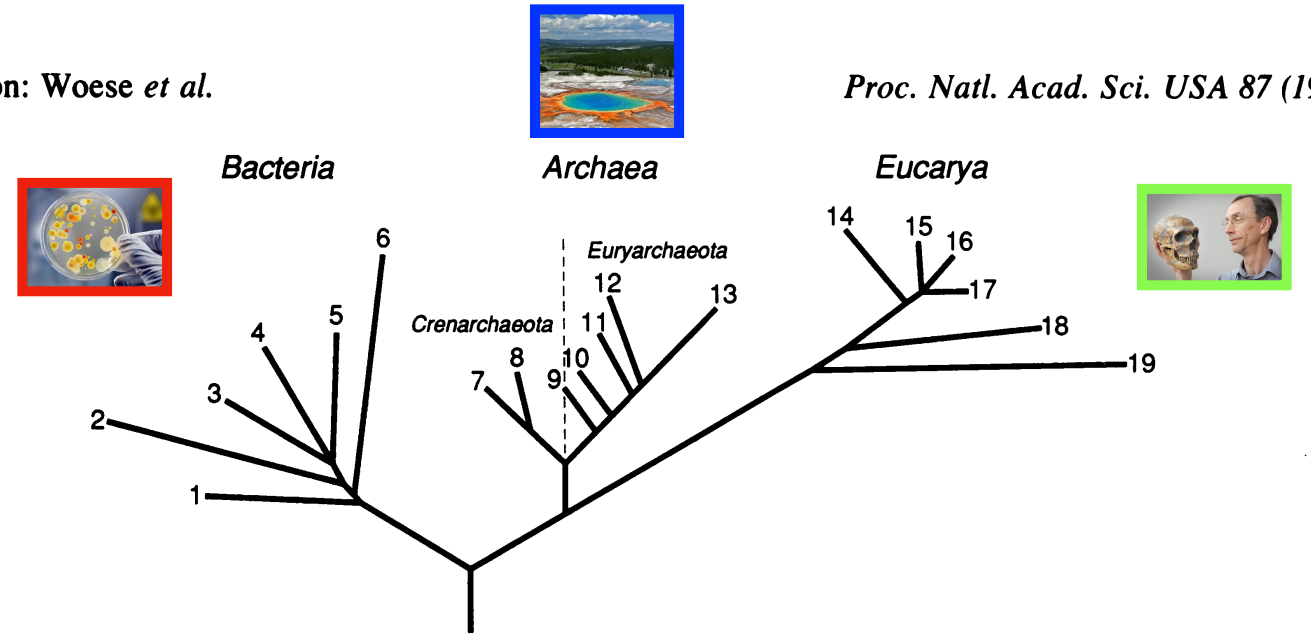


FIG. 1. Universal phylogenetic tree in rooted form, showing the three domains. Branching order and branch lengths are based upon rRNA sequence comparisons (and have been taken from figure 4 of ref. 2). The position of the root was determined by comparing (the few known) sequences of pairs of paralogous genes that diverged from each other before the three primary lineages emerged from their common ancestral condition (27). [This rooting strategy (28) in effect uses the one set of (aboriginally duplicated) genes as an outgroup for the other.] The numbers on the branch tips correspond to the following groups of organisms (2). Bacteria: 1, the Thermotogales; 2, the flavobacteria and relatives; 3, the cyanobacteria; 4, the purple bacteria; 5, the Gram-positive bacteria; and 6, the green nonsulfur bacteria. Archae: the kingdom Crenarchaeota: 7, the genus *Pyrodictium*; and 8, the genus *Thermoproteus*; and the kingdom Euryarchaeota: 9, the Thermococcales; 10, the Methanococcales; 11, the Methanobacteriales; 12, the Methanomicrobiales; and 13, the extreme halophiles. Eucarya: 14, the animals; 15, the ciliates; 16, the green plants; 17, the fungi; 18, the flagellates; and 19, the microsporidia.

We are very good at reconstructing gene trees..

DATA

observations
sequences
(alignment)

sites

$A_i, A_{i+1}, A_{i+2}, \ldots$

gene from species 1.	…	A	G	T	C	G	…
gene from species 2.	…	A	G	T	C	G	…
gene from species 3.	…	A	G	A	A	T	…
gene from species 4.	…	T	T	A	A	T	…

phylogenetic likelihood:

DATA MODEL

$$P(\text{data} \mid \text{model}) = \prod_i P(A_i \mid G, Q)$$

(Felsenstein 1981)

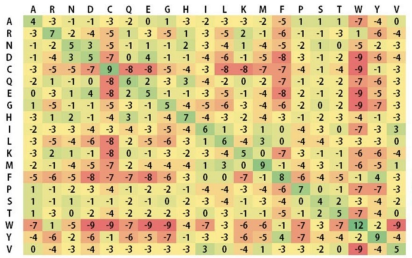
Which **gene tree** produced my **sequences**?

e.g. RAxML, IQ-TREE, MrBayes, PhyloBayes

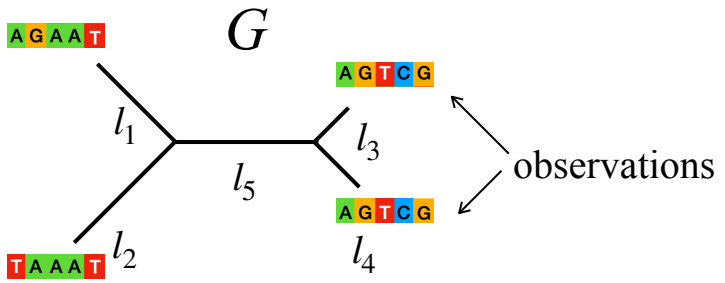
MODEL

matrix of substitution rates

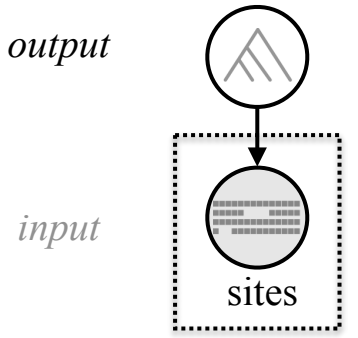
$$Q = \begin{array}{c} \begin{matrix} A & T & G & C \end{matrix} \\ \begin{pmatrix} * & \kappa & 1 & 1 \\ \kappa & * & 1 & 1 \\ 1 & 1 & * & \kappa \\ 1 & 1 & \kappa & * \end{pmatrix} \end{array} \begin{matrix} A \\ T \\ G \\ C \end{matrix}$$



gene tree



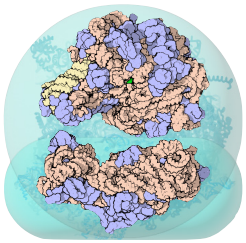
species tree unaware



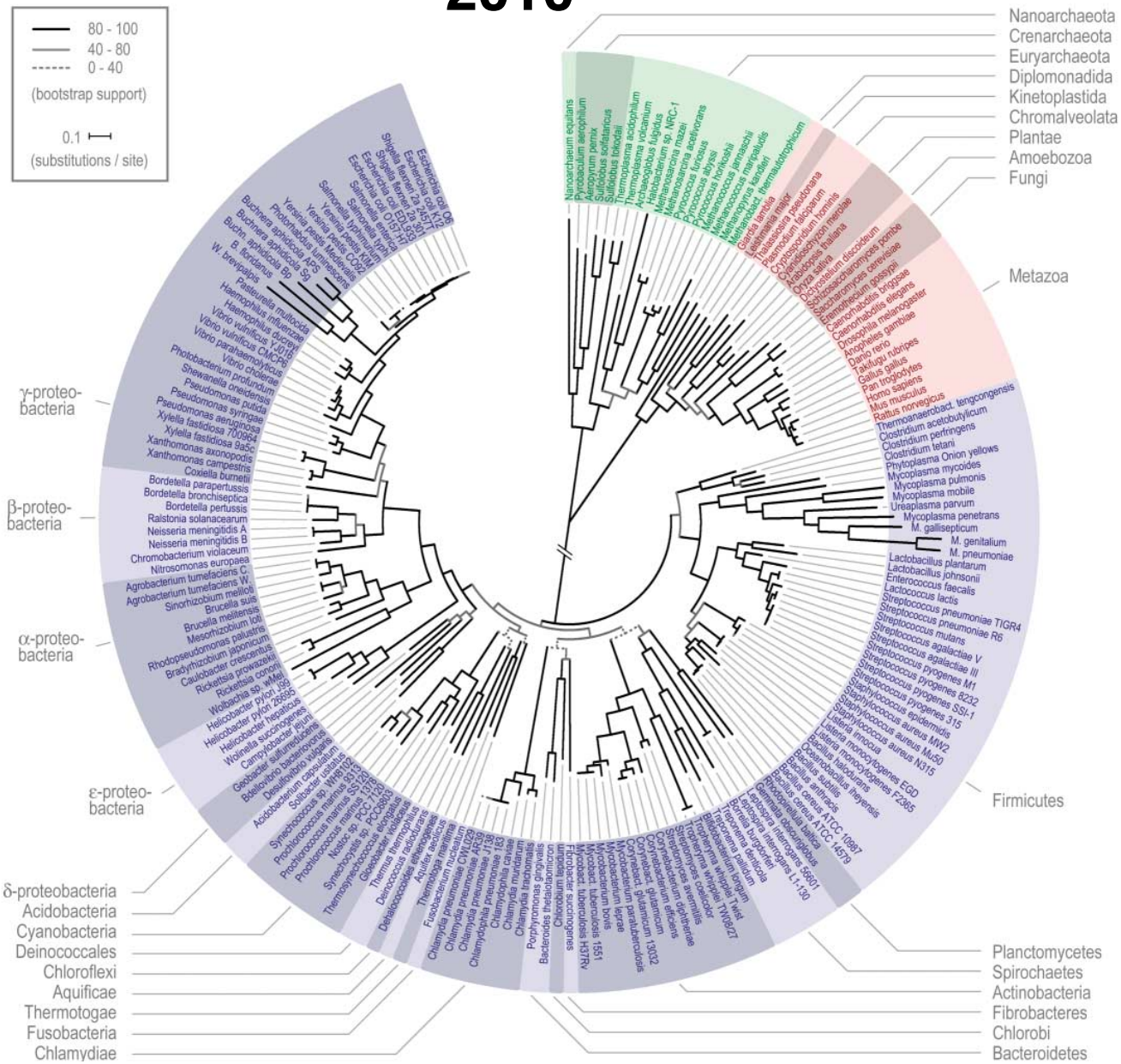
The golden age of molecular evolution

2010

31
genes



191
genomes



(Ciccarelli 2006)

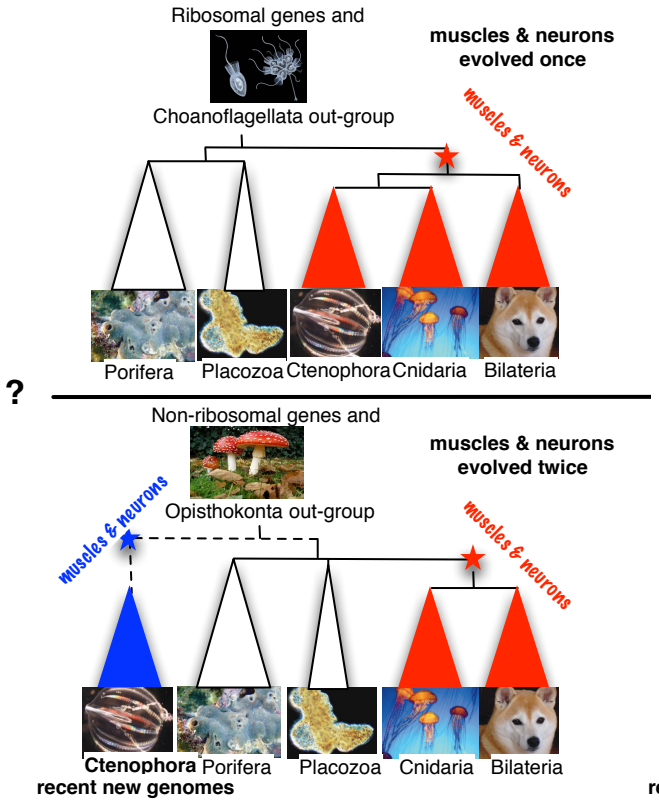
gergely.szollosi@oist.jp

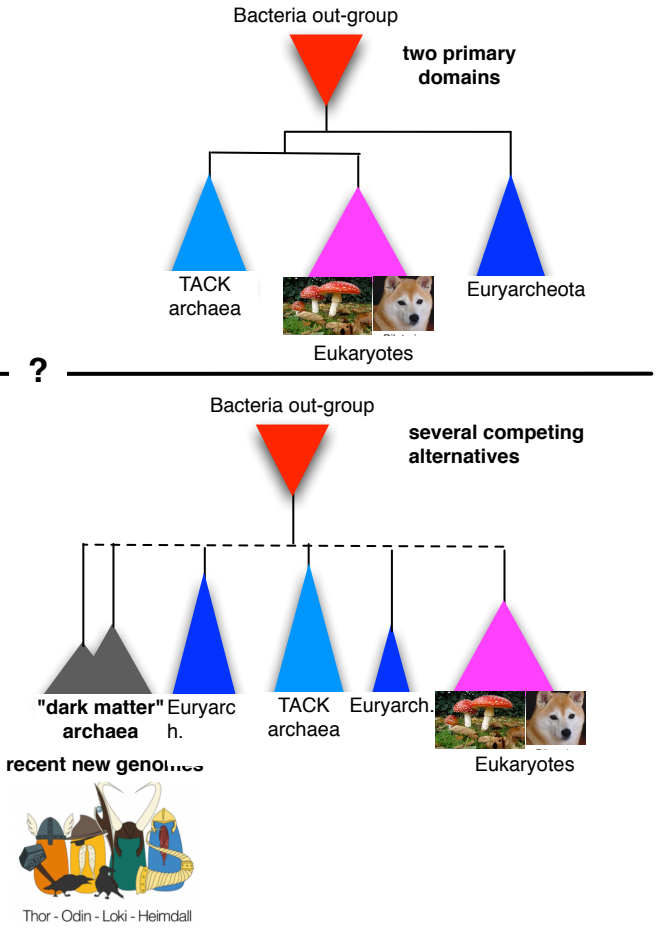
New genomes, old questions

New genomes instead of bringing into sharper focus major evolutionary events such as the origin of eukaryotes or the diversification of major animal lineages have instead reignited old debates.

Phillipe 2009
..
Noshenko 2013
Pisani 2015
Szánthó 2023

Dunn 2008
..
Ryan 2013
Moroz 2014
Schultz 2023





Zaremba-Niedzwiedzka - 2017
Imachi - 2020
Rodrigues-Oliveira - 2023

Lake 1988
..
Cox 2008
Foster 2009
Williams 2013
Williams 2017

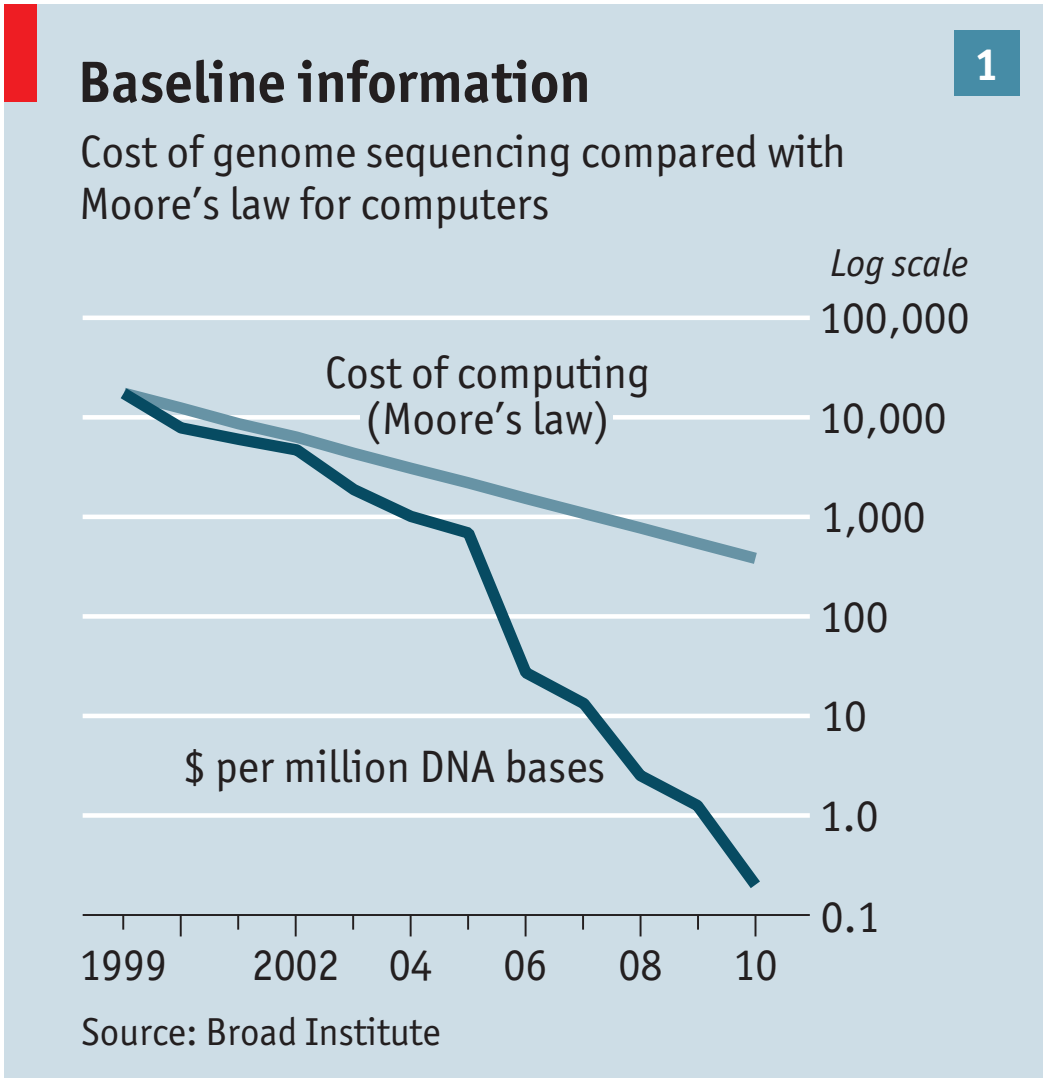
Woese 1977
..
Rinke 2013
Forterre 2015
Da Chuna 2017

Biology 2.0

1996

12 Giga FLOPS

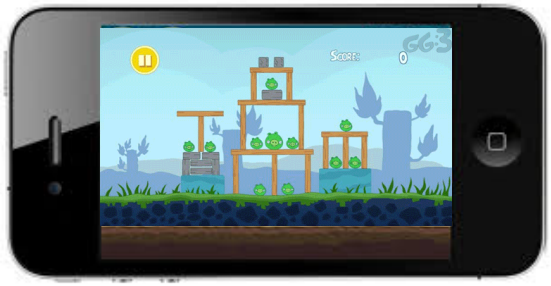
(Floating Point Operations Per Second)



100 x

2016

200-300 Giga FLOPS



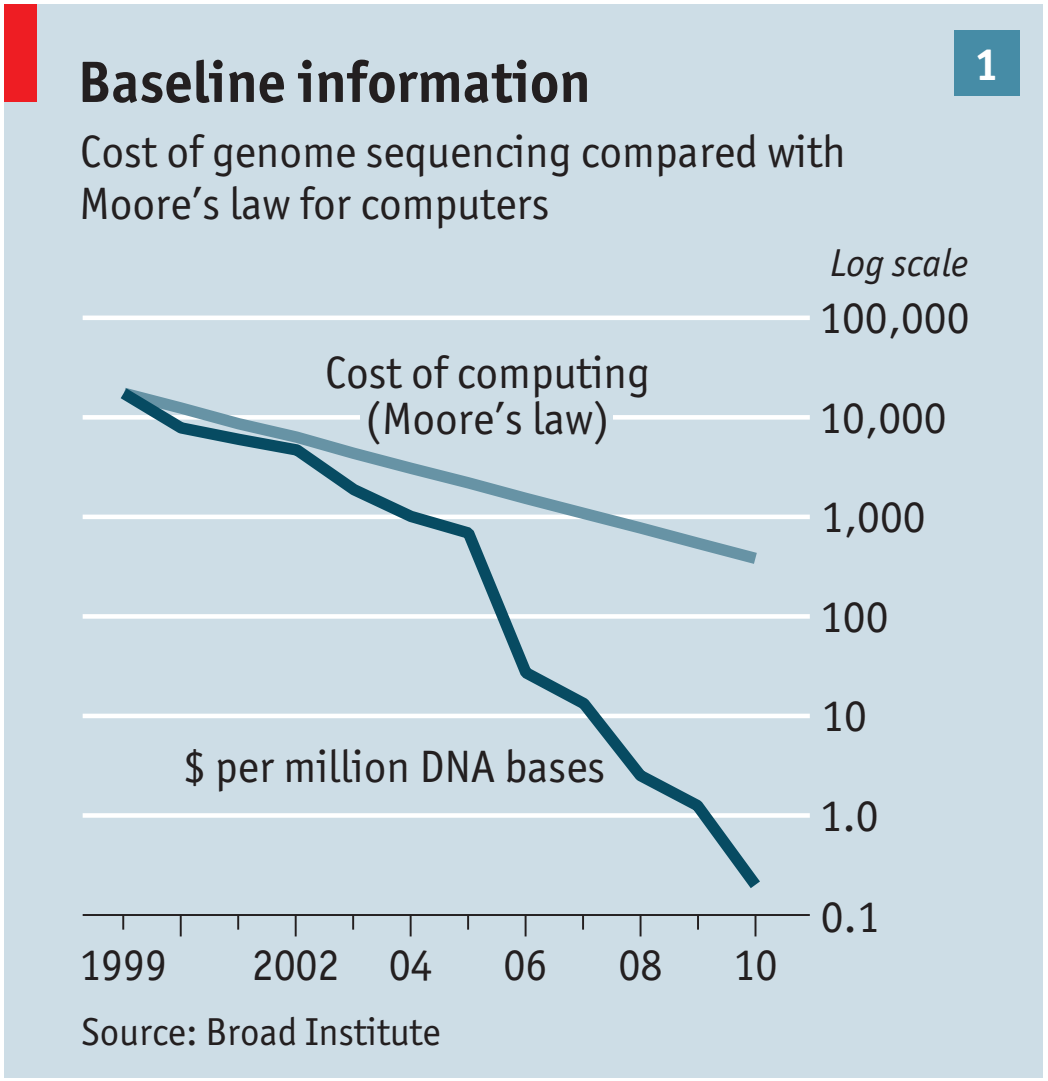
10 000 x

Biology 2.0

1996

12 Giga FLOPS

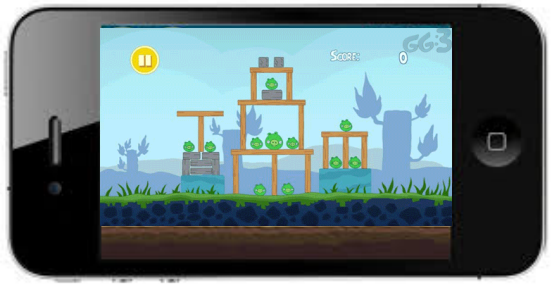
(Floating Point Operations Per Second)



100 x

2016

200-300 Giga FLOPS

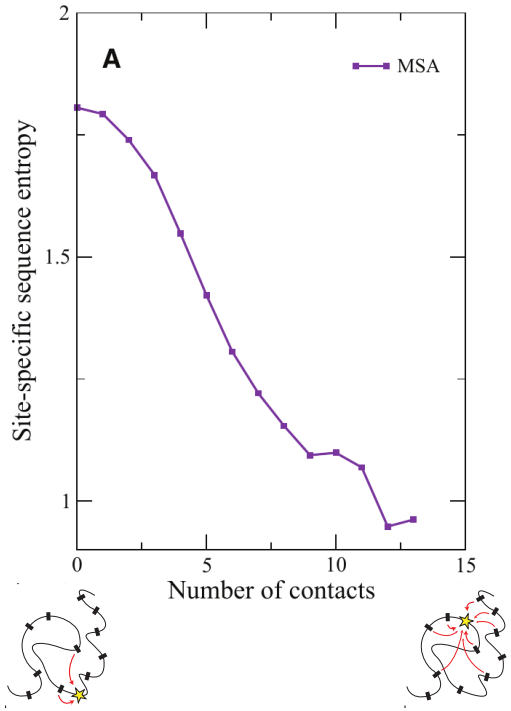


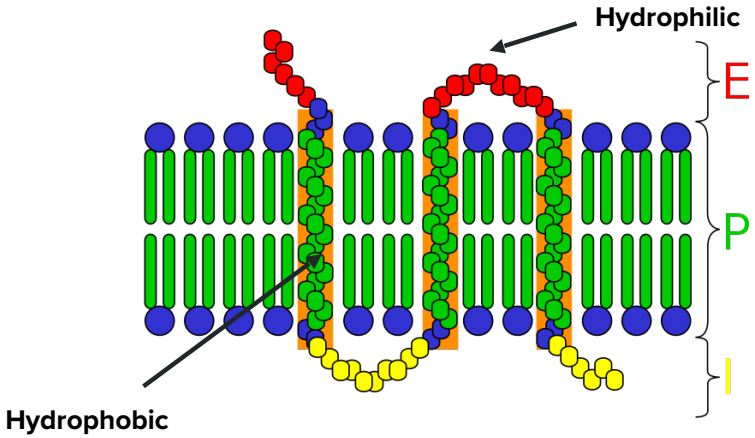
10 000 x

Prelude

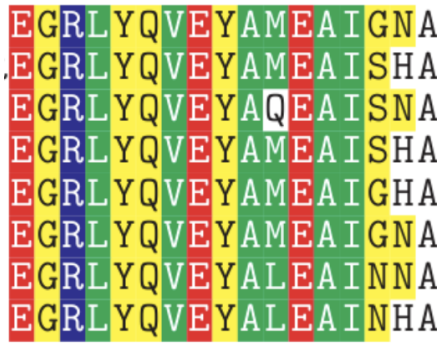
We are very good at reconstructing gene trees..

..but only as good as the model of sequence evolution used!

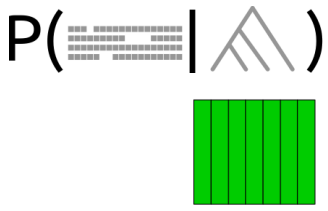




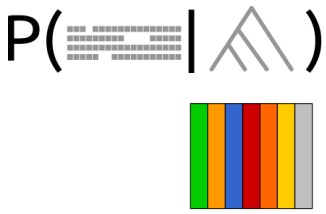
data

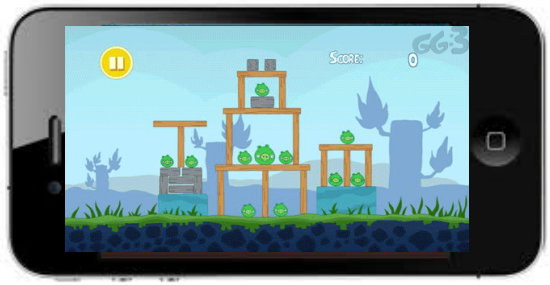


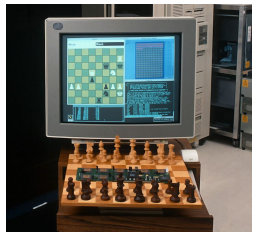
site homogenous composition **model**



site heterogenous composition **model**





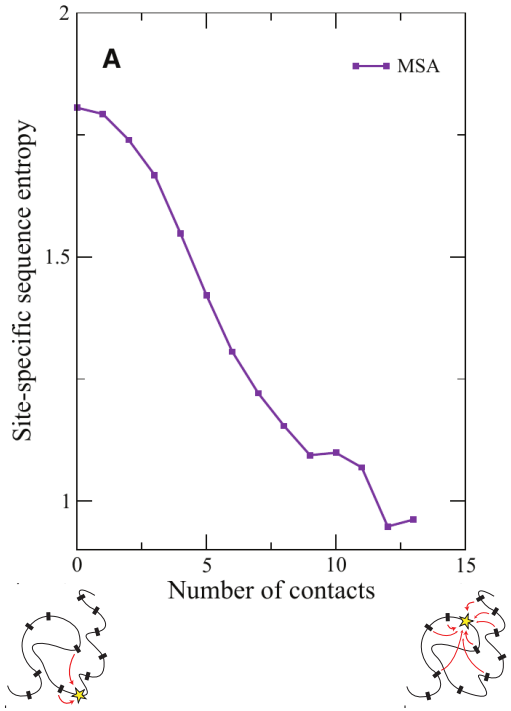


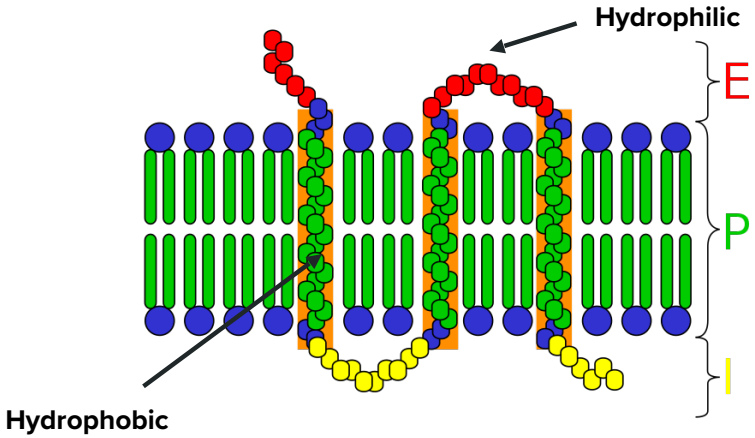
~100 × fewer leaves

Prelude

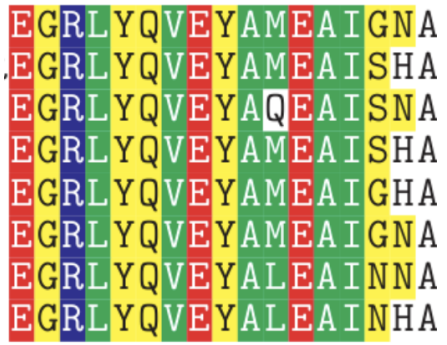
We are very good at reconstructing gene trees..

..but only as good as the model of sequence evolution used!

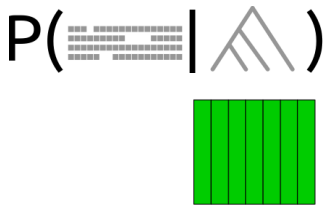




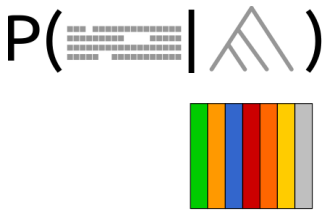
data

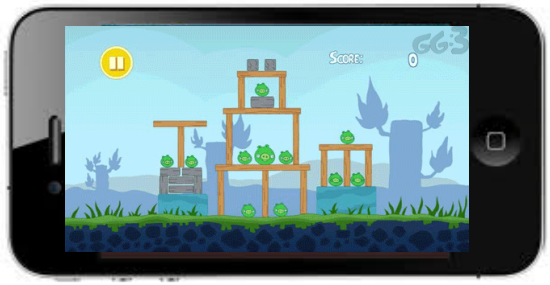


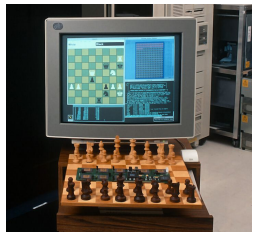
site homogenous composition **model**



site heterogenous composition **model**



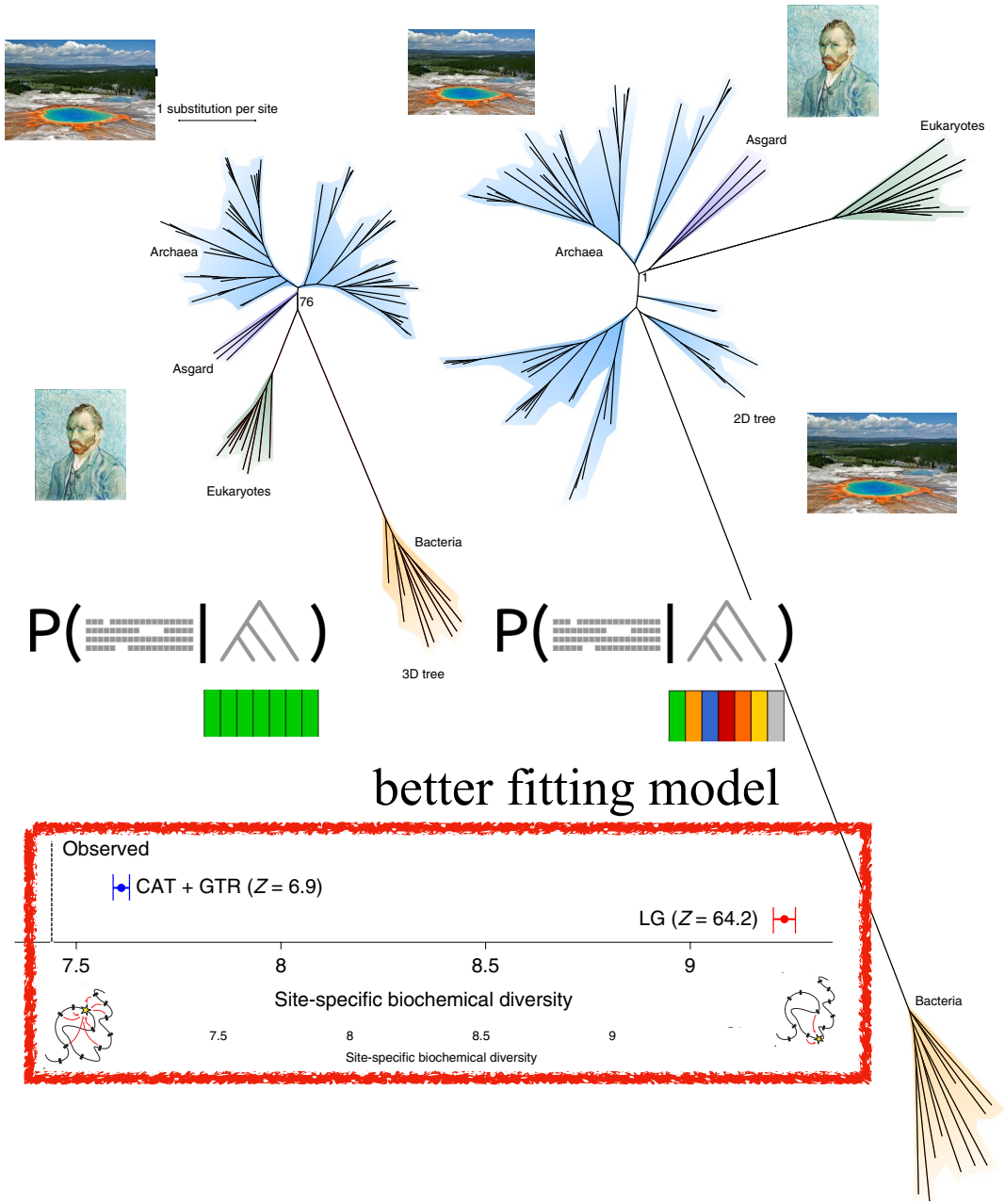




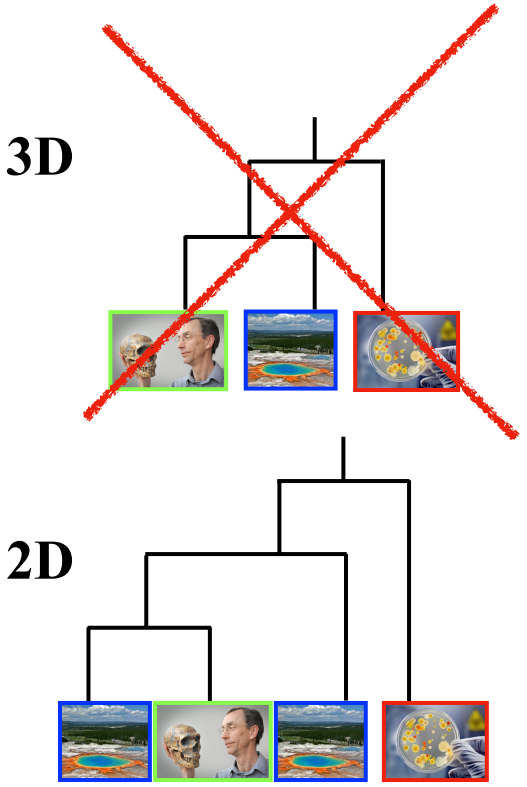
~100 × fewer leaves

Prelude

We are very good at reconstructing gene trees..
..but only as good as the model of sequence evolution used!

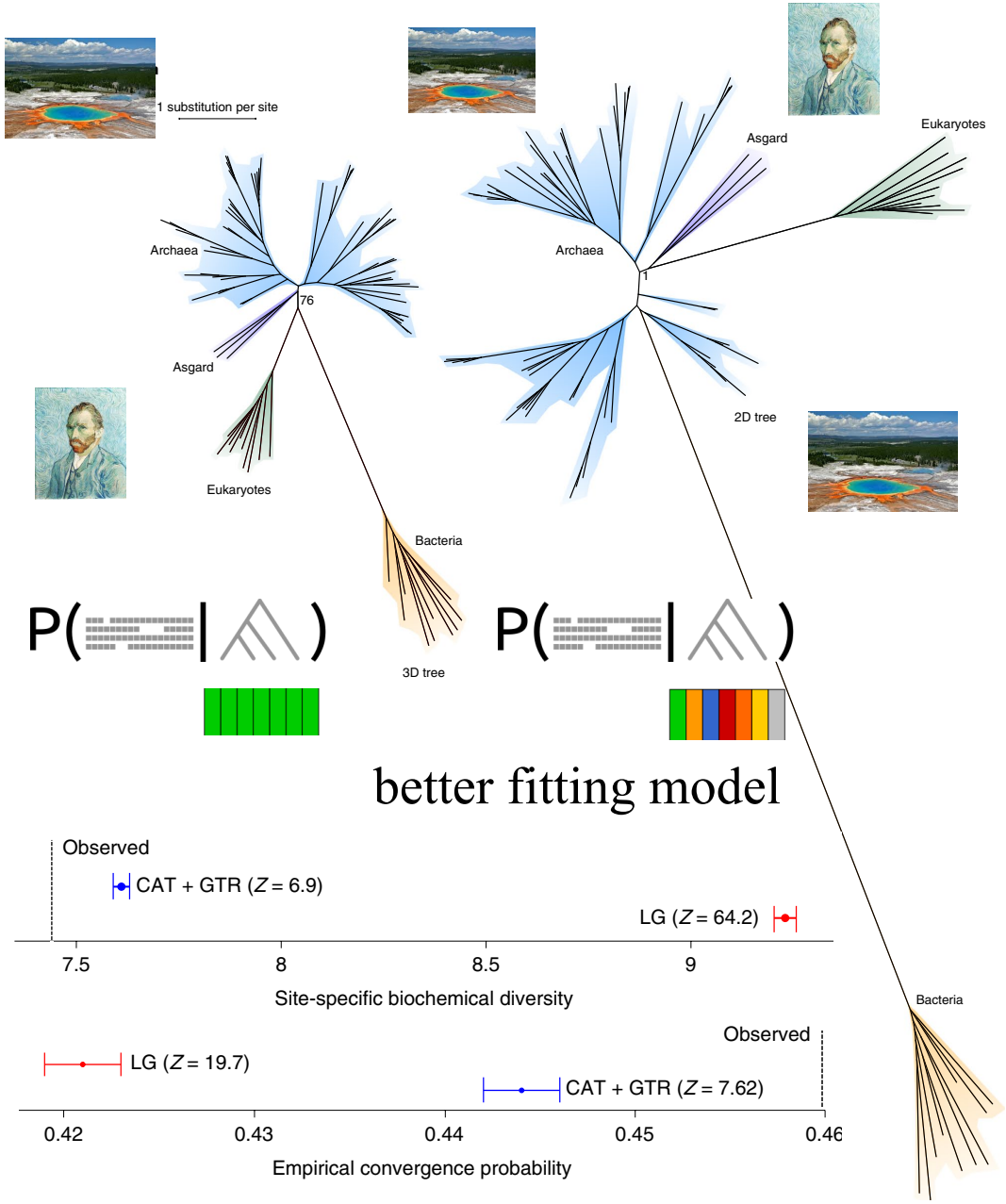


same data,
better fitting model

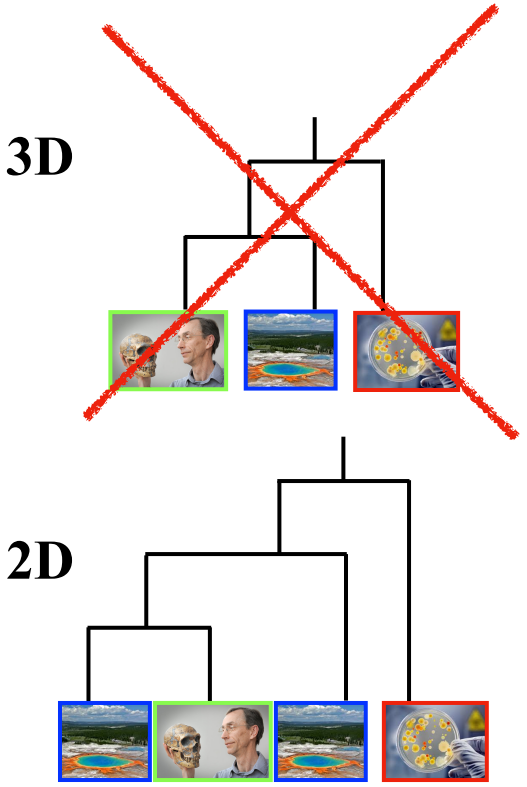


Williams, Cox, Foster, **Szöllősi**, Embley ***Systematic Biology*** (2022)
Phylogenomics provides robust support for a two-domains tree of life

We are very good at reconstructing gene trees..
..but only as good as the model of sequence evolution used!



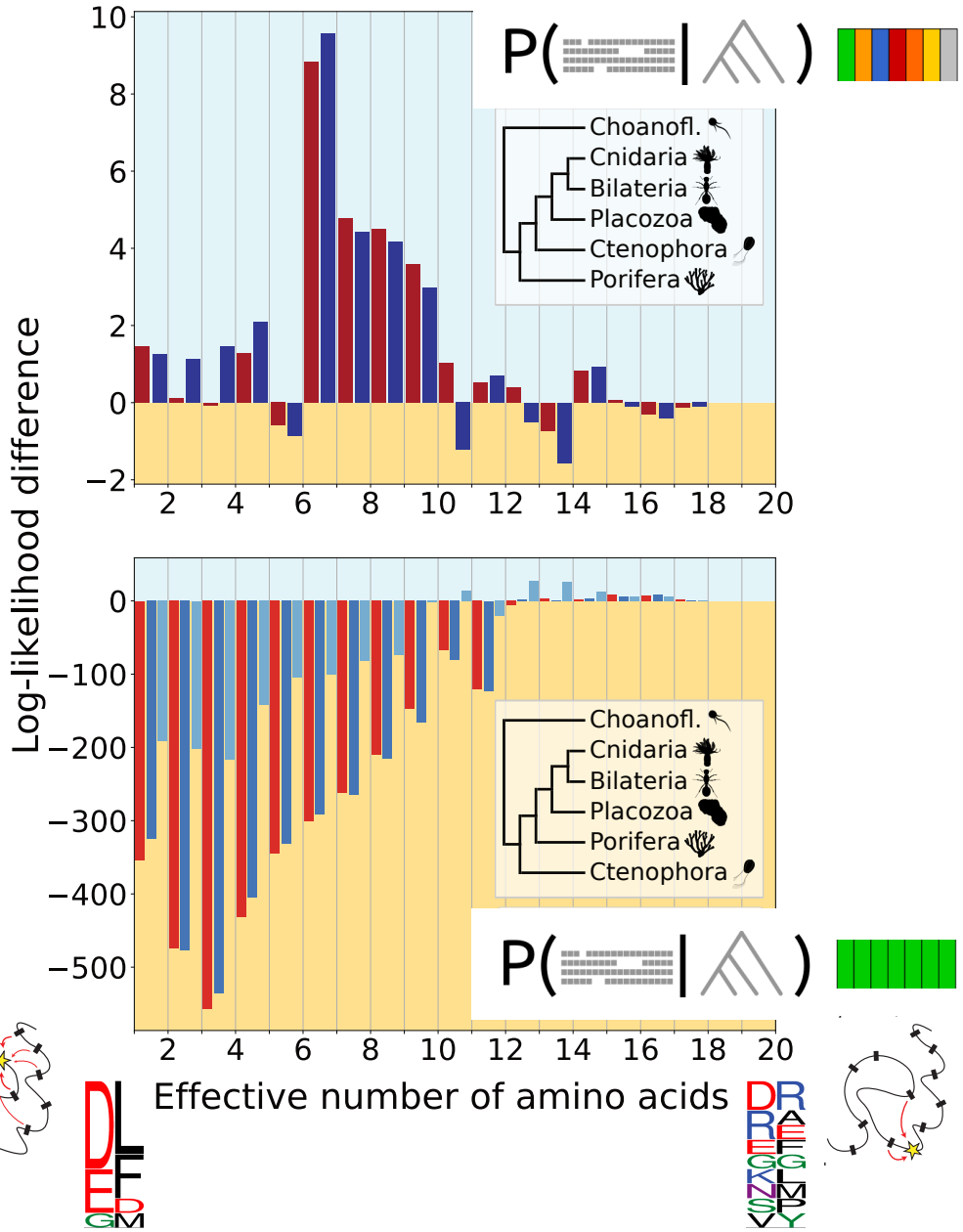
same data,
better fitting model



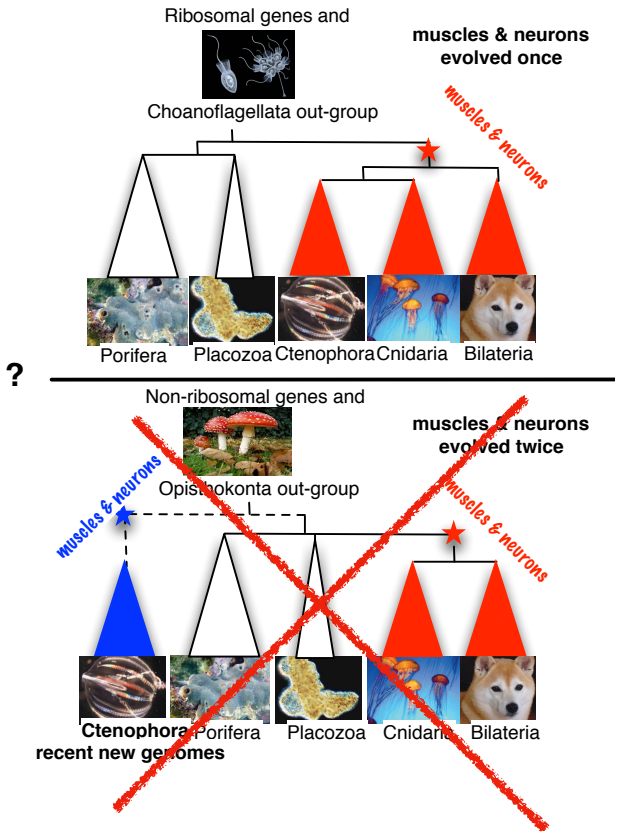
Williams, Cox, Foster, **Szöllősi**, Embley ***Systematic Biology*** (2022)
Phylogenomics provides robust support for a two-domains tree of life

We are very good at reconstructing gene trees..

..but only as good as the model of sequence evolution used!



same data,
better fitting model



Szánthó, Lartillot, **Szöllősi** & Schrempf ***Systematic Biology*** (2023)
Compositionaly constrained sites drive long branch attraciton

Phylogenomics — *"why we are doing it all wrong"*
with an emphasis on Horizontal Gene Transfer

Species tree aware & unaware methods

"*phylogenomics — why we are doing it all wrong*"

models of gene family evolution with D&L

joint reconstruction with DL of the mammalian ToL

HGT in the context of species tree-aware methods

models of gene family evolution with D,T&L

just how much HGT?

HGT as information

***A**malgamated **L**ikelihood **E**stimation*

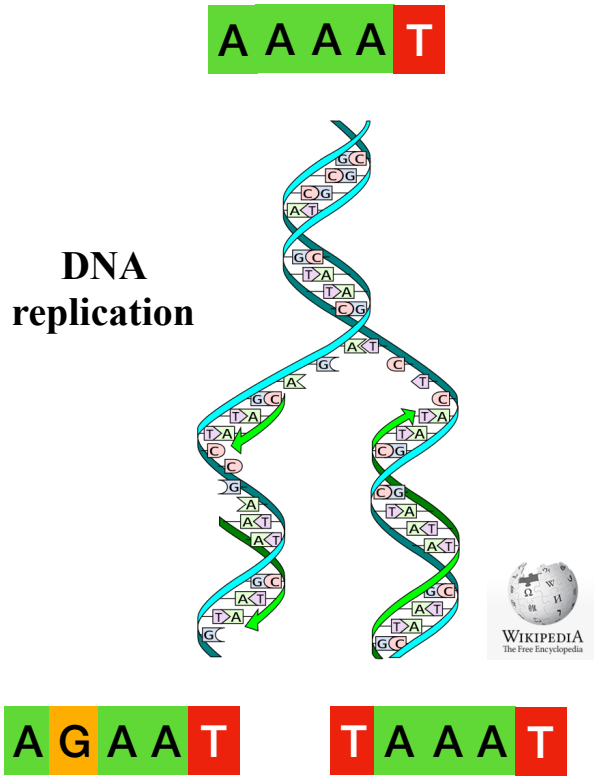
faster models of gene family evolution with D,T&L

outgroup-free rooting

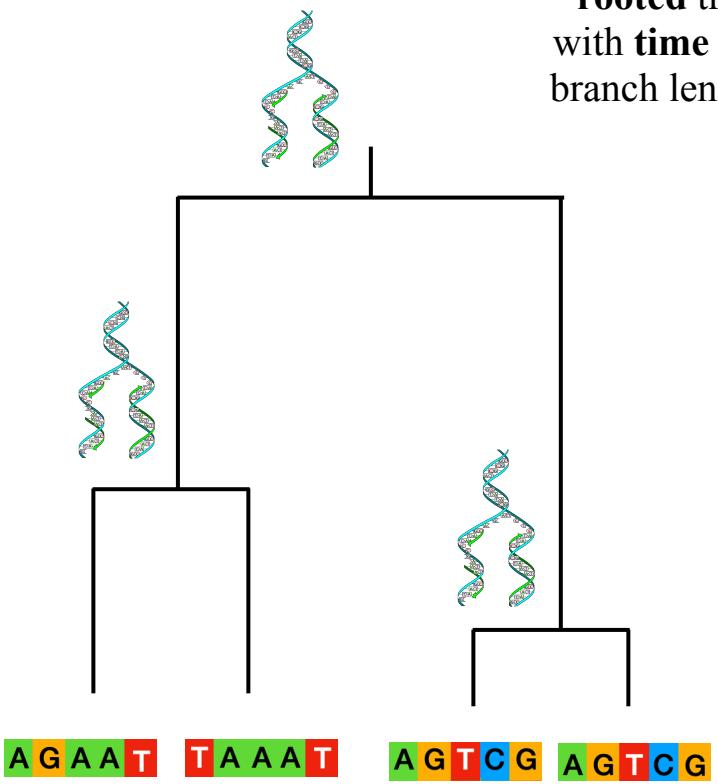
HGT from the dead

What process generates genes trees?

Independent of the details of reproduction the story of two homologous pieces of DNA can (locally) always be traced back to a single replication event. For a set of sequences this implies the existence of a **bifurcating gene tree** along which the sequences evolved



rooted tree with **time like** branch lengths



gergely.szollosi@oist.jp

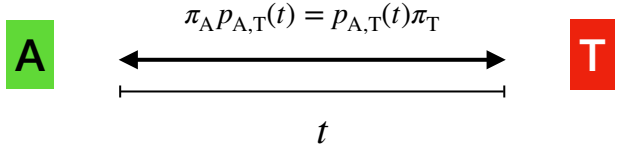
We are very good at reconstructing gene trees..
..with two practical caveats

1. computational constraints on calculating

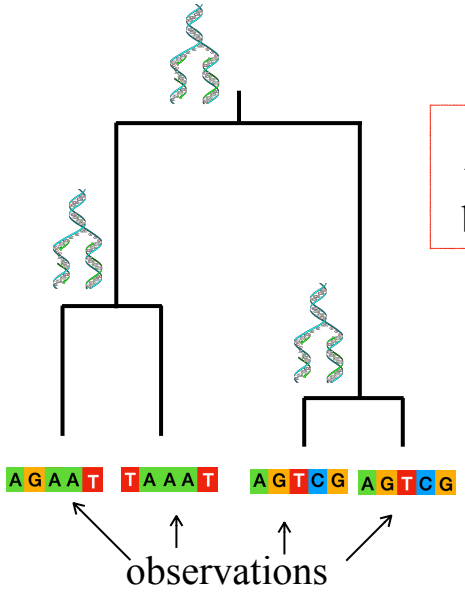
the phylogenetic likelihood $P(\text{data} \mid \text{tree}) = \prod_i P(A_i \mid G, Q)$

restrict us to the subset of **site independent** and **time reversible** substitution models:

stationary frequencies: $\{\pi_A, \pi_T, \pi_G, \pi_C\} : Q\pi = 0$

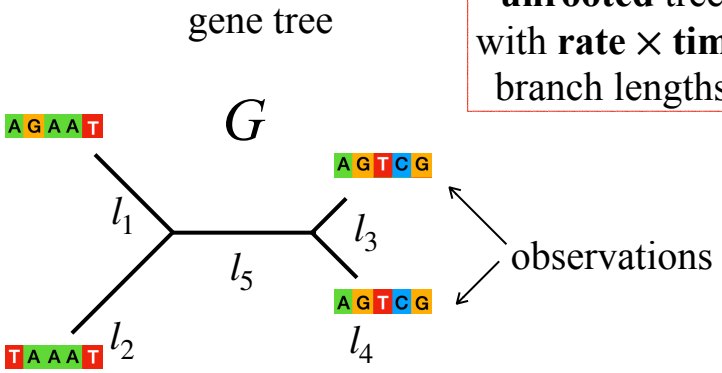


under which the likelihood is independent of root position.



rooted tree with **time like** branch lengths

2. without knowledge of the root and additional calibrations on evolutionary rates (e.g. fossils) **rate and time are confounded**

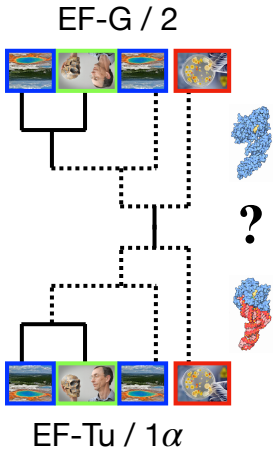


unrooted tree with **rate × time** branch lengths

gergely.szollosi@oist.jp

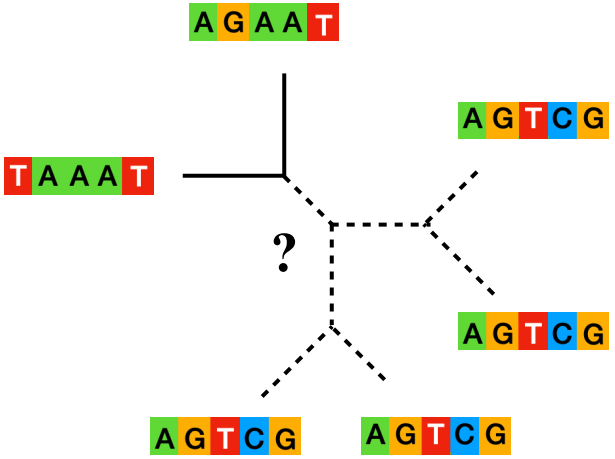
Two more fundamental issues!

I. Individual genes alone contain limited signal, e.g., in modern datasets the number of sequences often approaches or exceeds the number of sites



The **M**ost **L**ikely tree is most likely wrong ..

unrooted tree
with **rate** × **time**
branch lengths



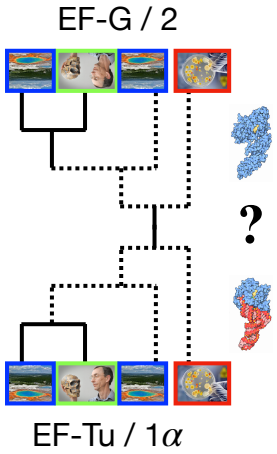
bifurcations
are **DNA**
replication events

observations
are **genes**
(related sequences)

gergely.szollosi@oist.jp

Two more fundamental issues!

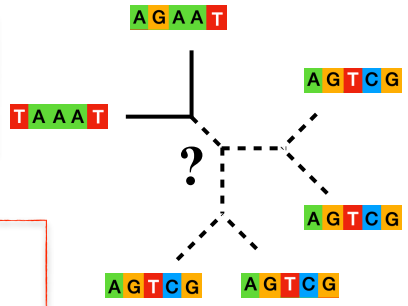
I. Individual genes alone contain limited signal, e.g., in modern datasets the number of sequences often approaches or exceeds the number of sites



Most likely tree

unrooted tree with **rate** × **time** branch lengths

observations are **genes** (related sequences)



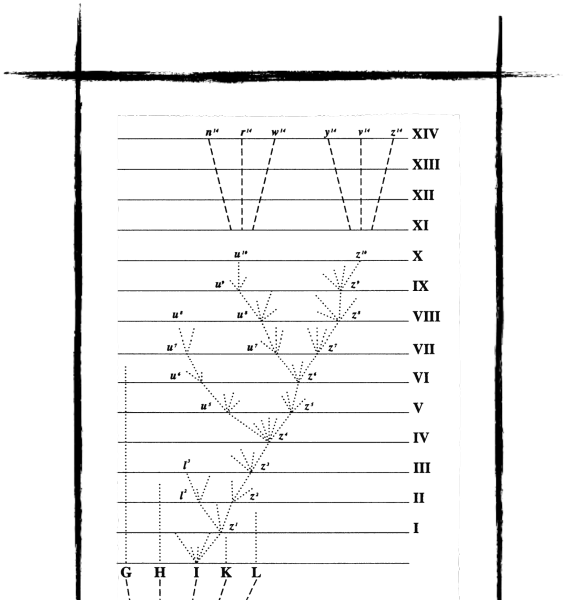
bifurcations are **DNA replication** events

S.J. Carlson, 1999

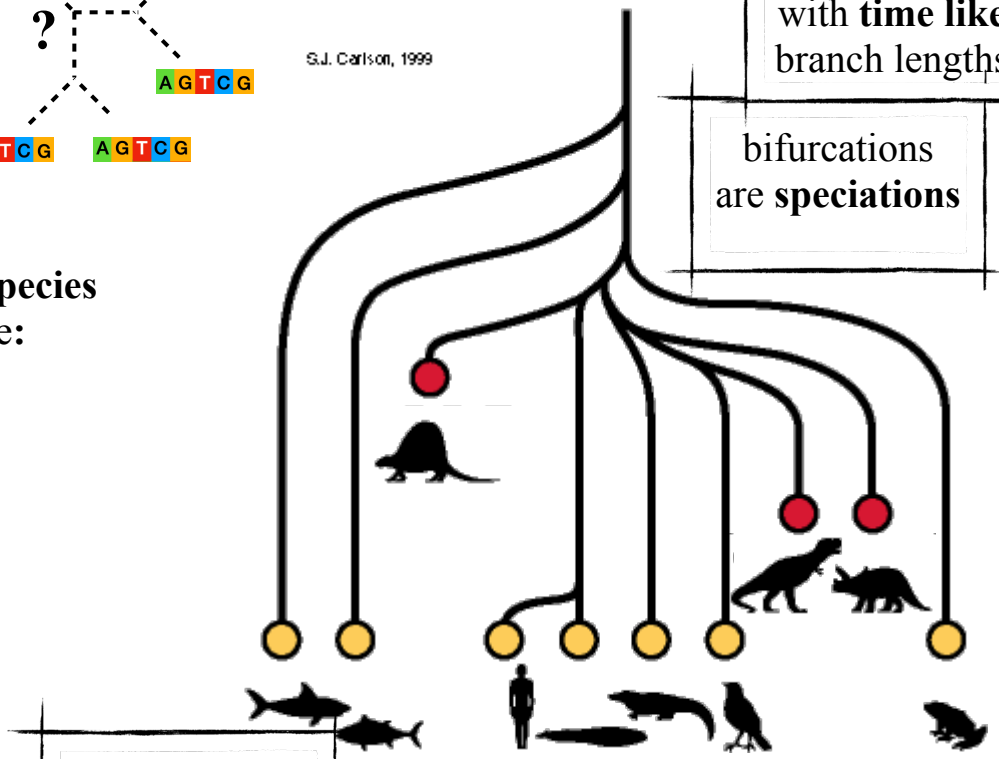
rooted tree with **time like** branch lengths

bifurcations are **speciations**

II. What we are **interested in is the tree is species** or gene trees in the context of the species tree**:**



observations are **genomes**

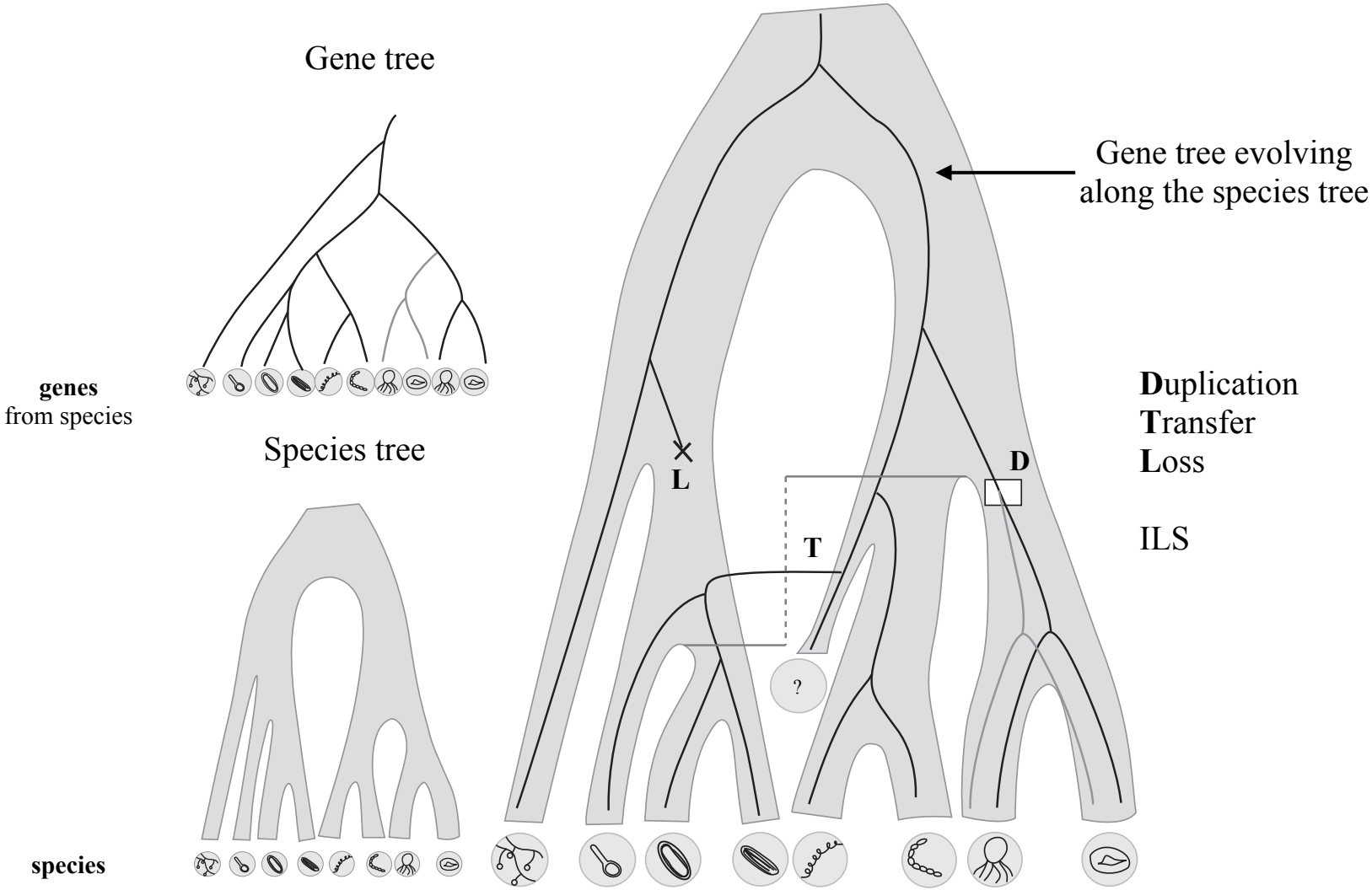


today's vertebrates

gergely.szollosi@oist.jp

The problem is gene trees are not the species tree

A gene tree is a deformation of the species tree through the prism of genome evolution and population genetics processes.



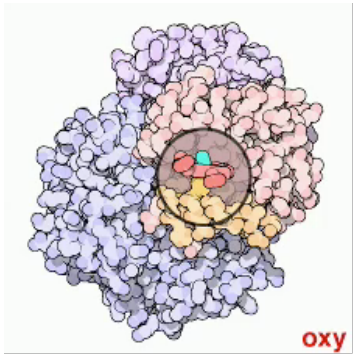
Daubin & Szöllősi 2016

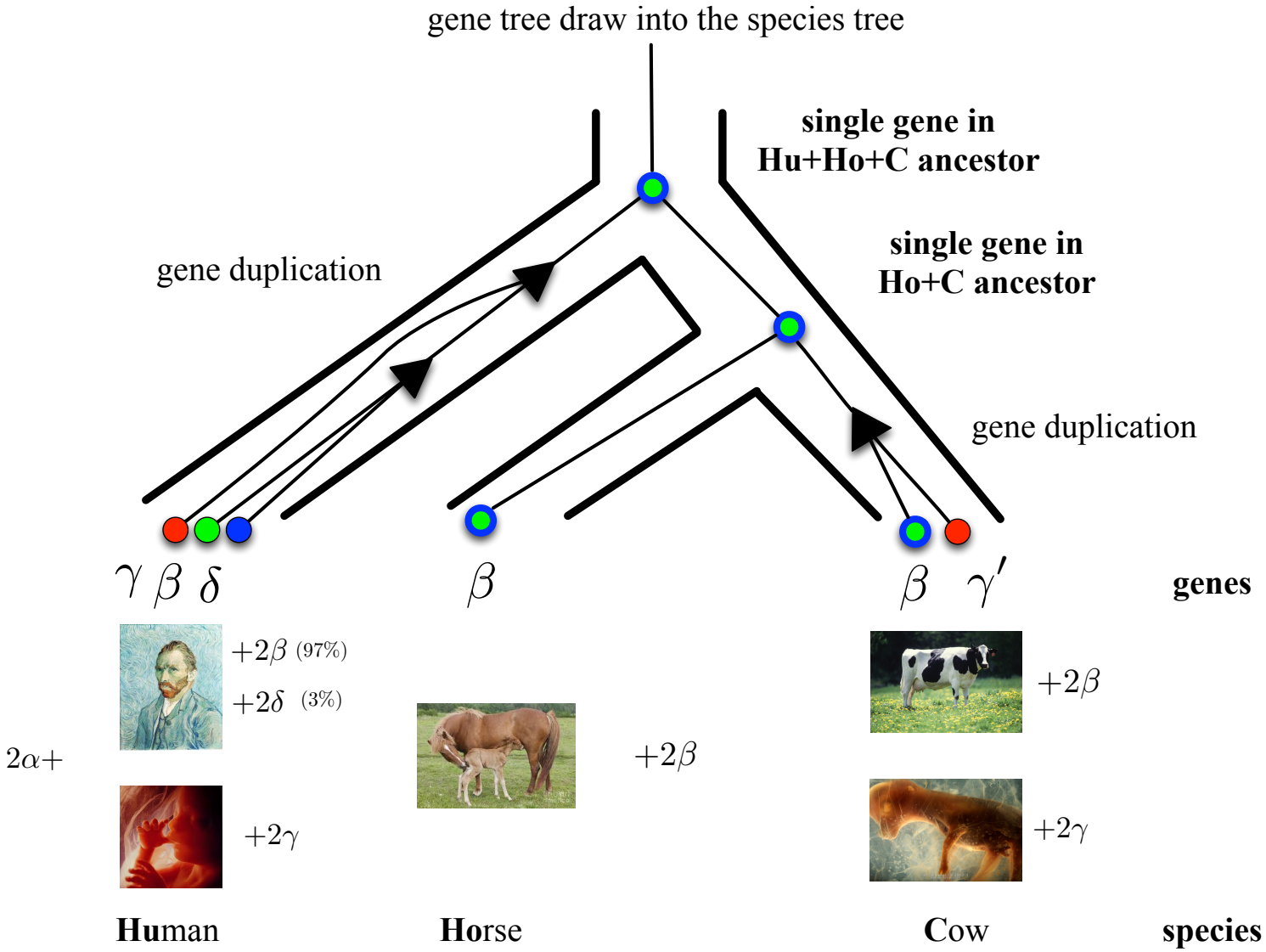
gergely.szollosi@oist.jp

Gene trees tell evolutionary stories in the context of the species tree

Gene duplication often results in new or modified function

hemoglobin

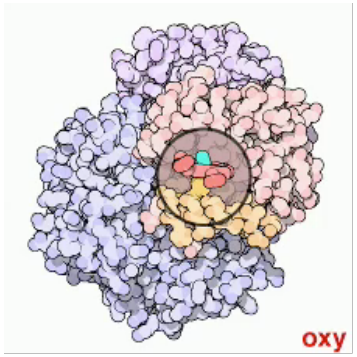


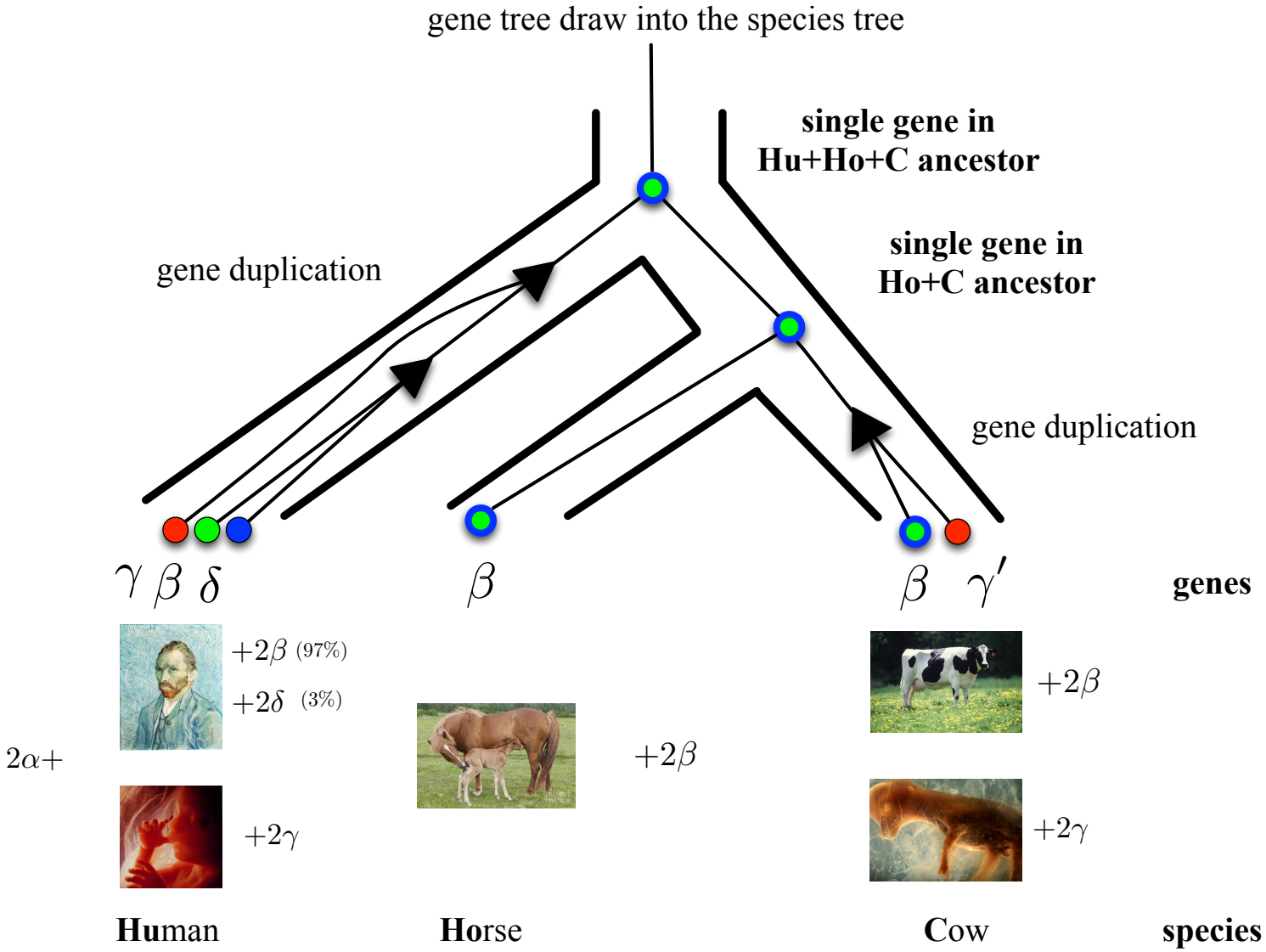


Gene trees tell evolutionary stories in the context of the species tree

Gene duplication often results in new or modified function

hemoglobin



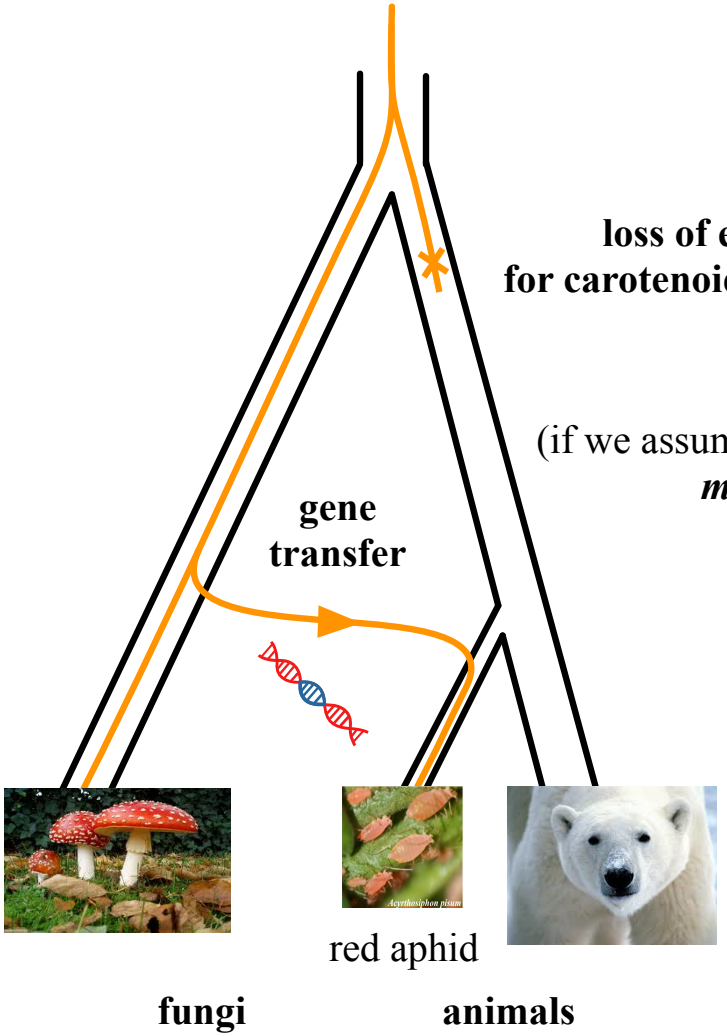


Gene trees tell evolutionary stories in the context of the species tree

Horizontal gene transfer is common among unicellular organisms, but examples are know even among animals.

pea aphids

Moran & Jarvik 2010 Science



The history of the 1%

Carefully choosing gene present in a single copy in each organism and introducing external information we can equate gene trees with the species tree ..

MARKER GENES

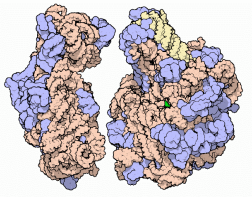

TAAAT


AGAAT


AGTCG


AGTCG

e.g. ribosomal genes



16S rRNA

observations are **genomes** ✔ **rooted** tree

bifurcations are **speciations** ✔ with **time like** branch lengths

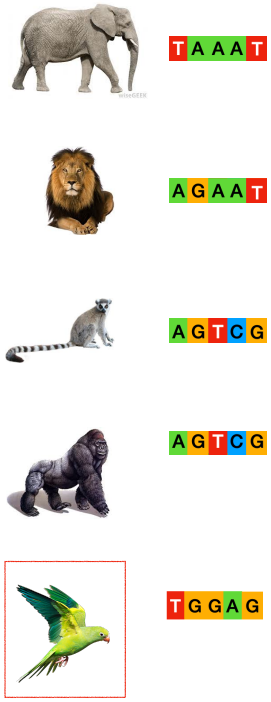
gergely.szollosi@oist.jp

"*phylogenomics — why we are doing it all wrong*"

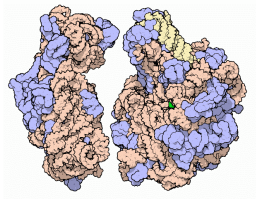
The history of the 1%

Carefully choosing gene present in a single copy in each organism and introducing external information we can equate gene trees with the species tree ..

MARKER GENES

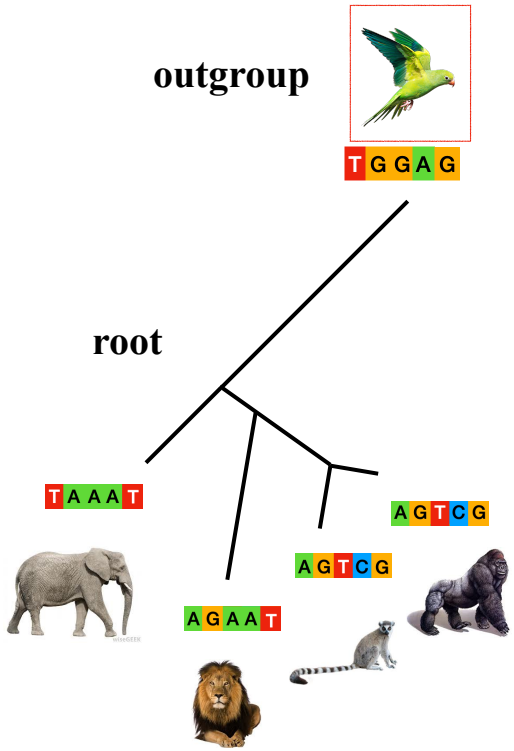


e.g. ribosomal genes



16S rRNA

OUTGROUP ROOTING



ingroup

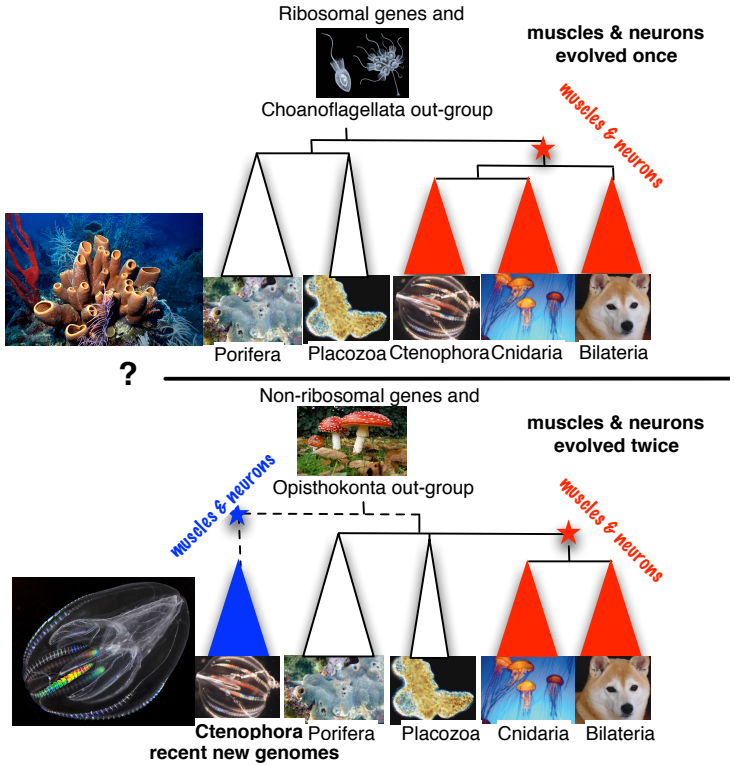
observations are **genomes** ✔

bifurcations are **speciations** ✔

rooted tree ✔

with **time like** branch lengths

can be problematic..



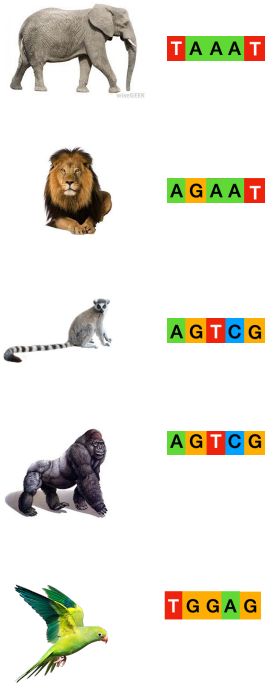
gergely.szollosi@oist.jp

"*phylogenomics — why we are doing it all wrong*"

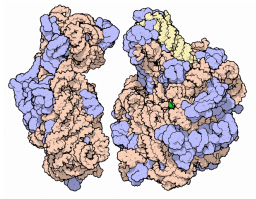
The history of the 1%

Carefully choosing gene present in a single copy in each organism and introducing external information we can equate gene trees with the species tree ..

MARKER GENES

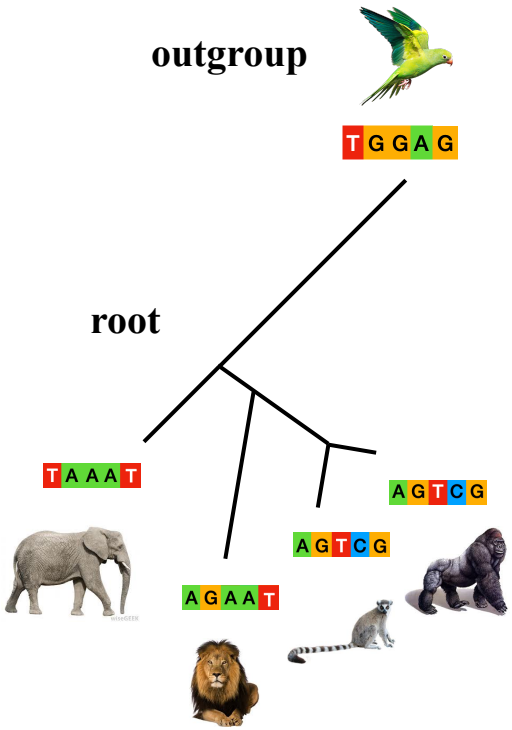


e.g. ribosomal genes



16S rRNA

OUTGROUP ROOTING



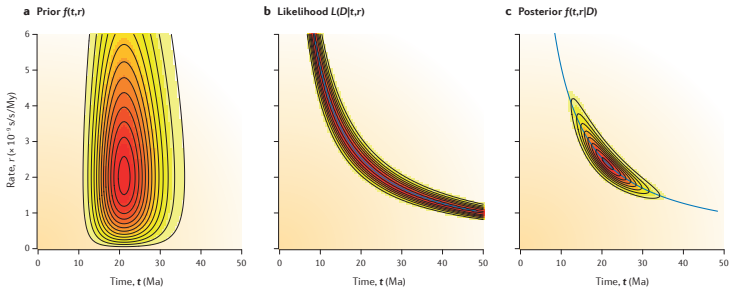
observations are **genomes** ✔

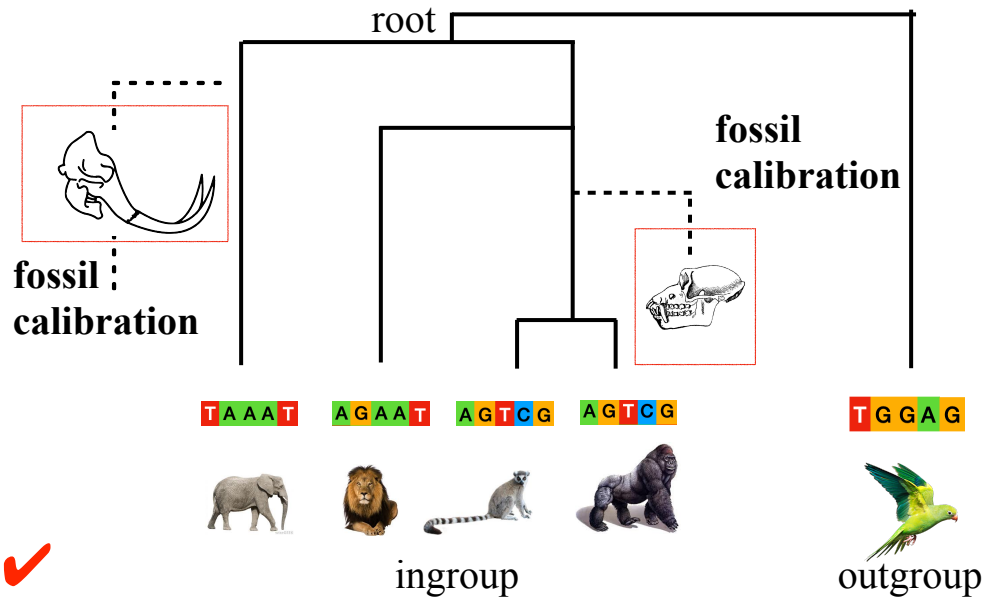
bifurcations are **speciations** ✔

rooted tree ✔

with **time like** branch lengths ✔

RELAXED MOLECULAR CLOCK





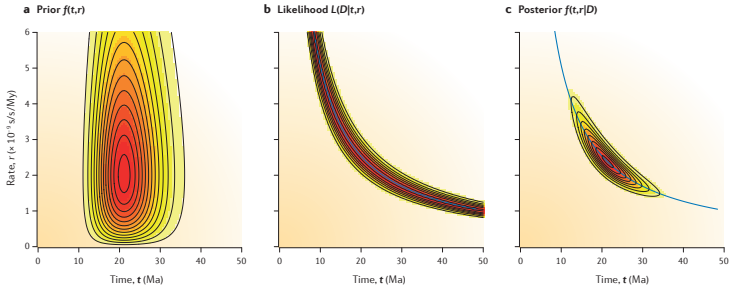
gergely.szollosi@oist.jp

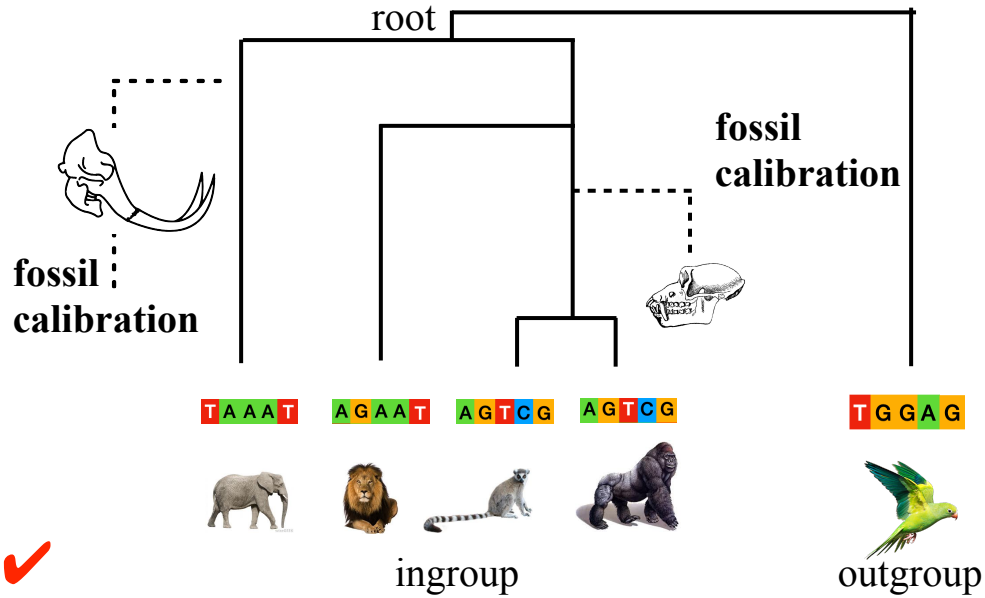
The history of the 1%

Carefully choosing gene present in a single copy in each organism and introducing external information we can equate gene trees with the species tree ..

can be problematic..

RELAXED MOLECULAR CLOCK





16S rRNA

observations are **genomes** ✔

bifurcations are **speciations** ✔

rooted tree ✔

with **time like** branch lengths ✔

gergely.szollosi@oist.jp

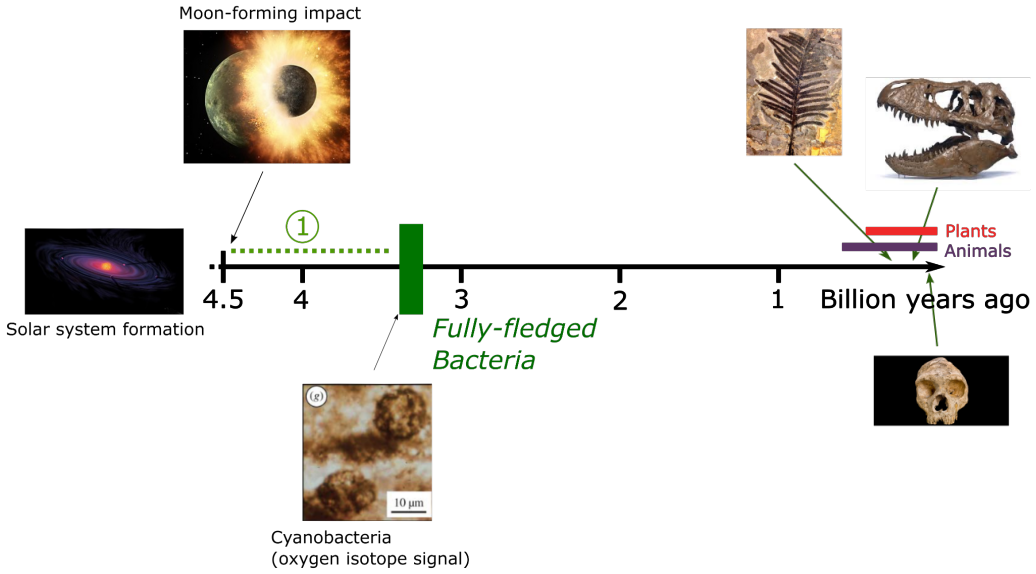
"*phylogenomics — why we are doing it all wrong*"

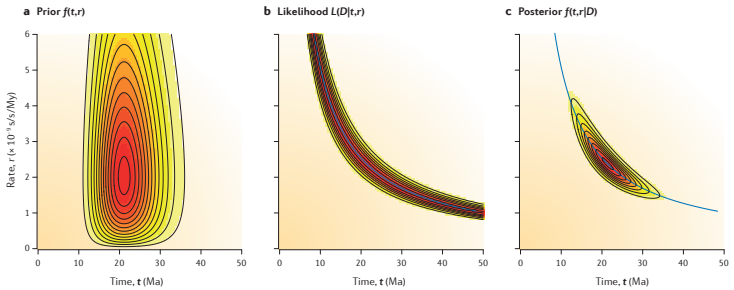
The history of the 1%

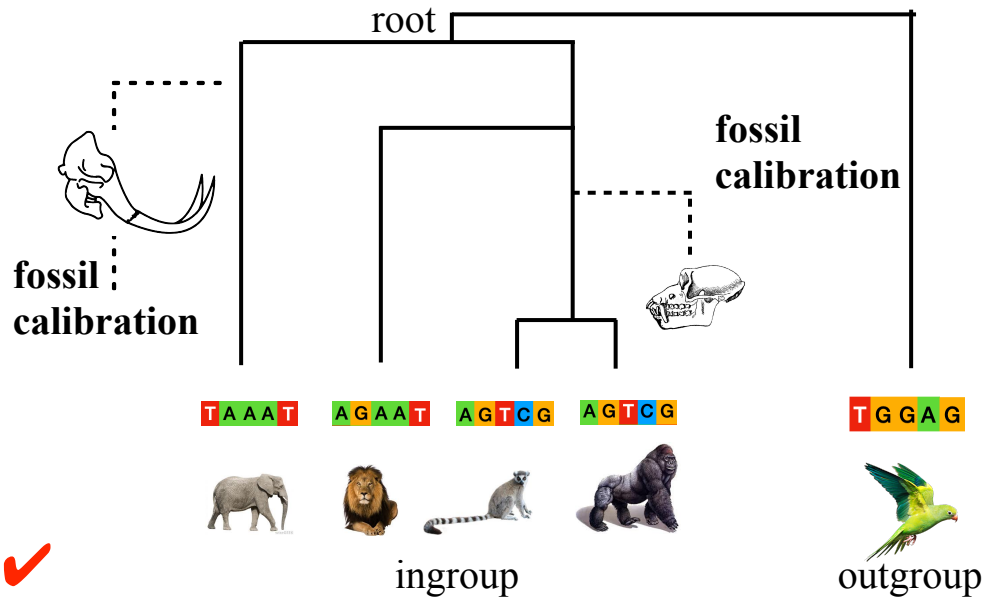
Carefully choosing gene present in a single copy in each organism and introducing external information we can equate gene trees with the species tree ..

can be problematic..

RELAXED MOLECULAR CLOCK







observations are **genomes** ✔

rooted tree ✔

16S rRNA

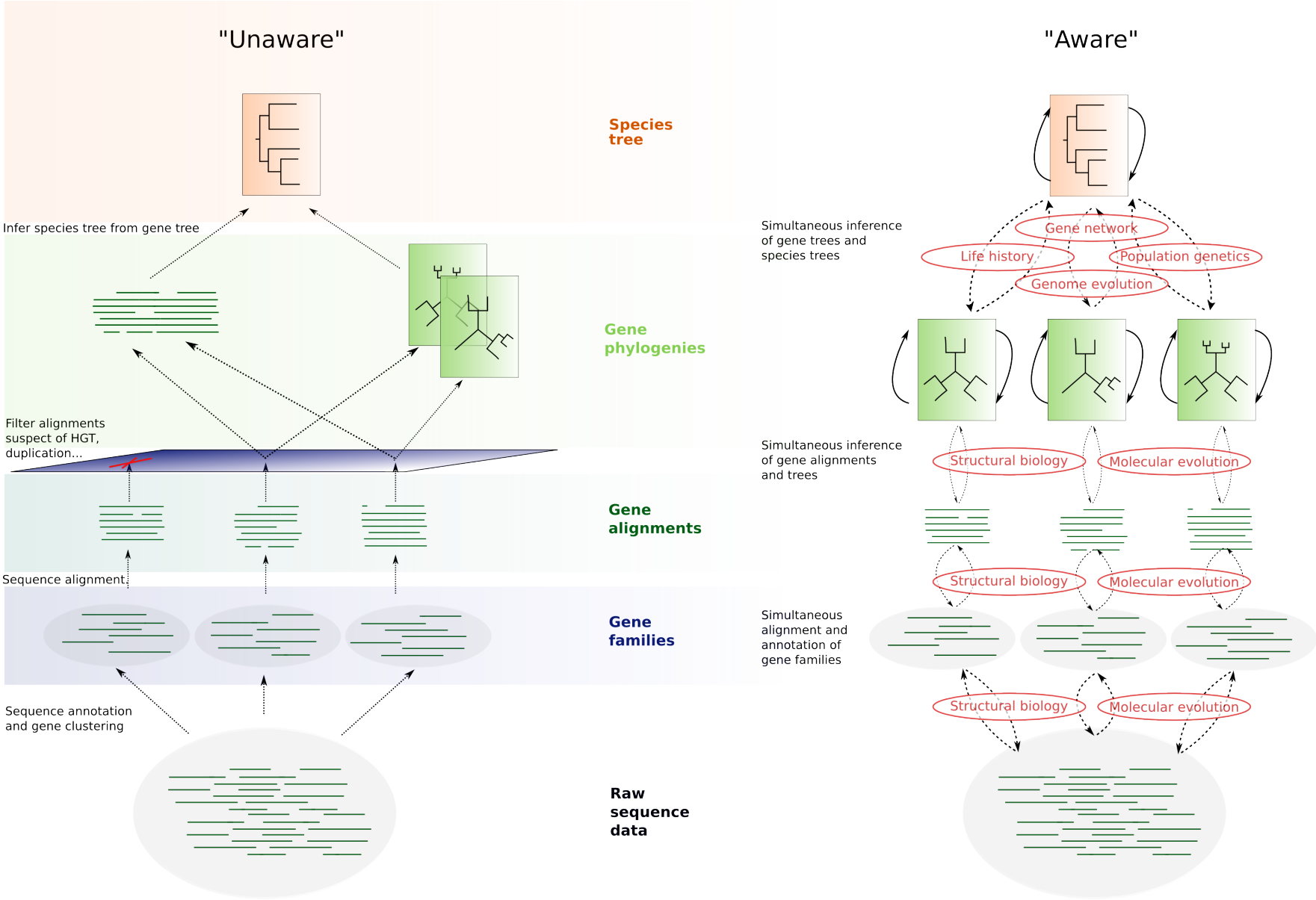
bifurcations are **speciations** ✔

with **time like** branch lengths ✔

gergely.szollosi@oist.jp

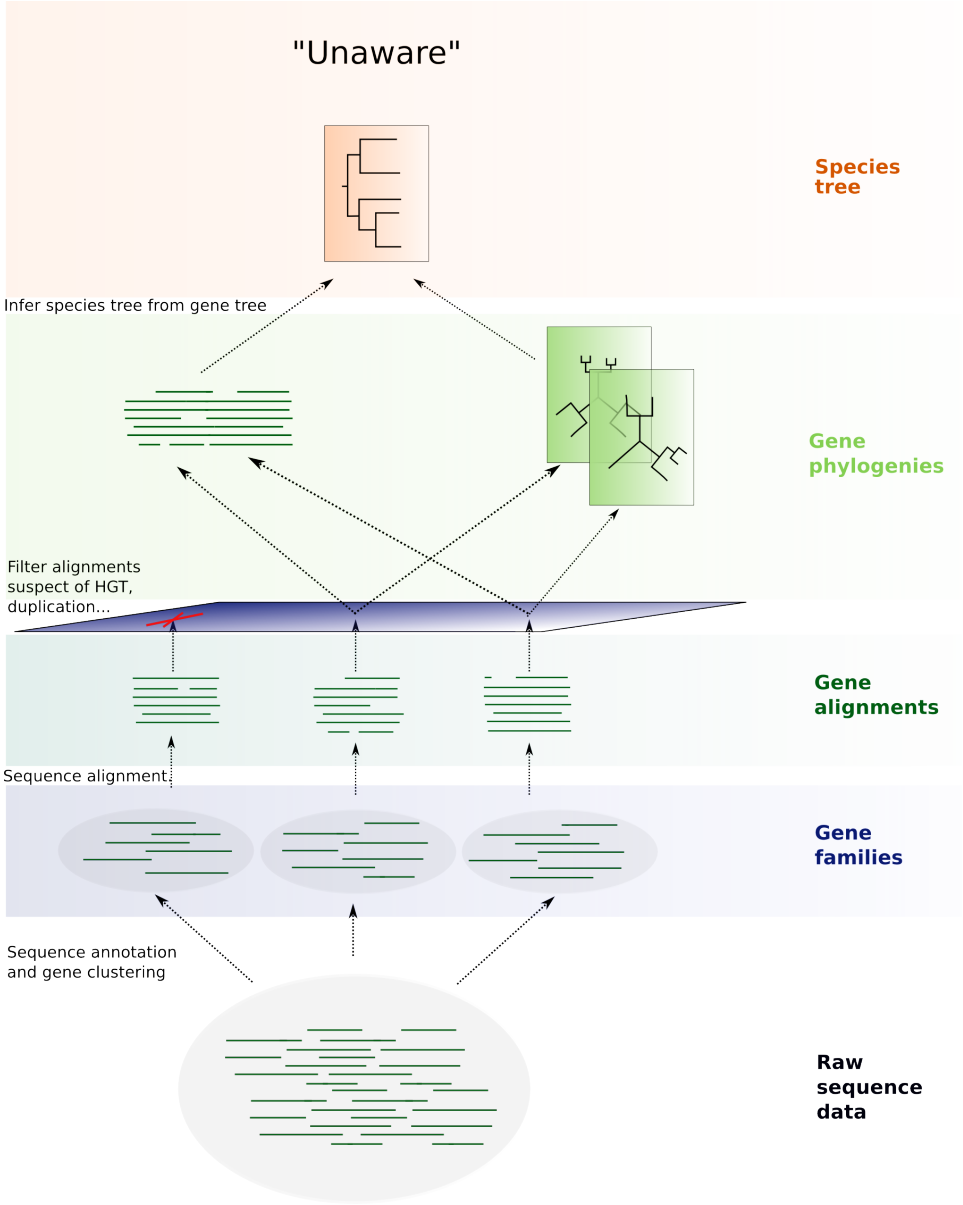
“*phylogenomics — why we are doing it all wrong*”

Phylogenetic awareness



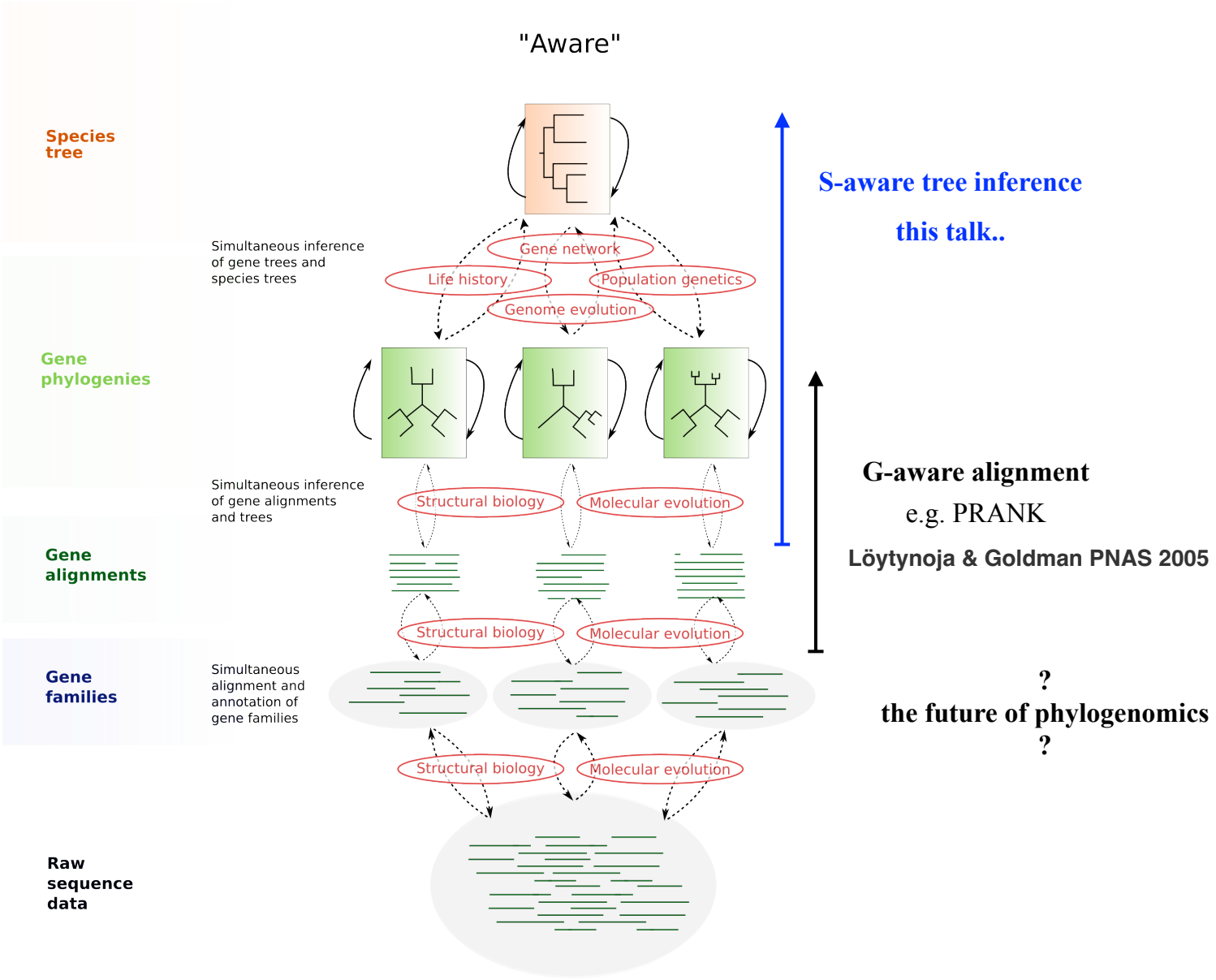
Daubin & Boussau 2011 TrEE

In the “unaware” path (the traditional way of inferring the species tree) each stage of the phylogenetic inference is independent from the steps up- and downstream.



Daubin & Boussau 2011 TrEE

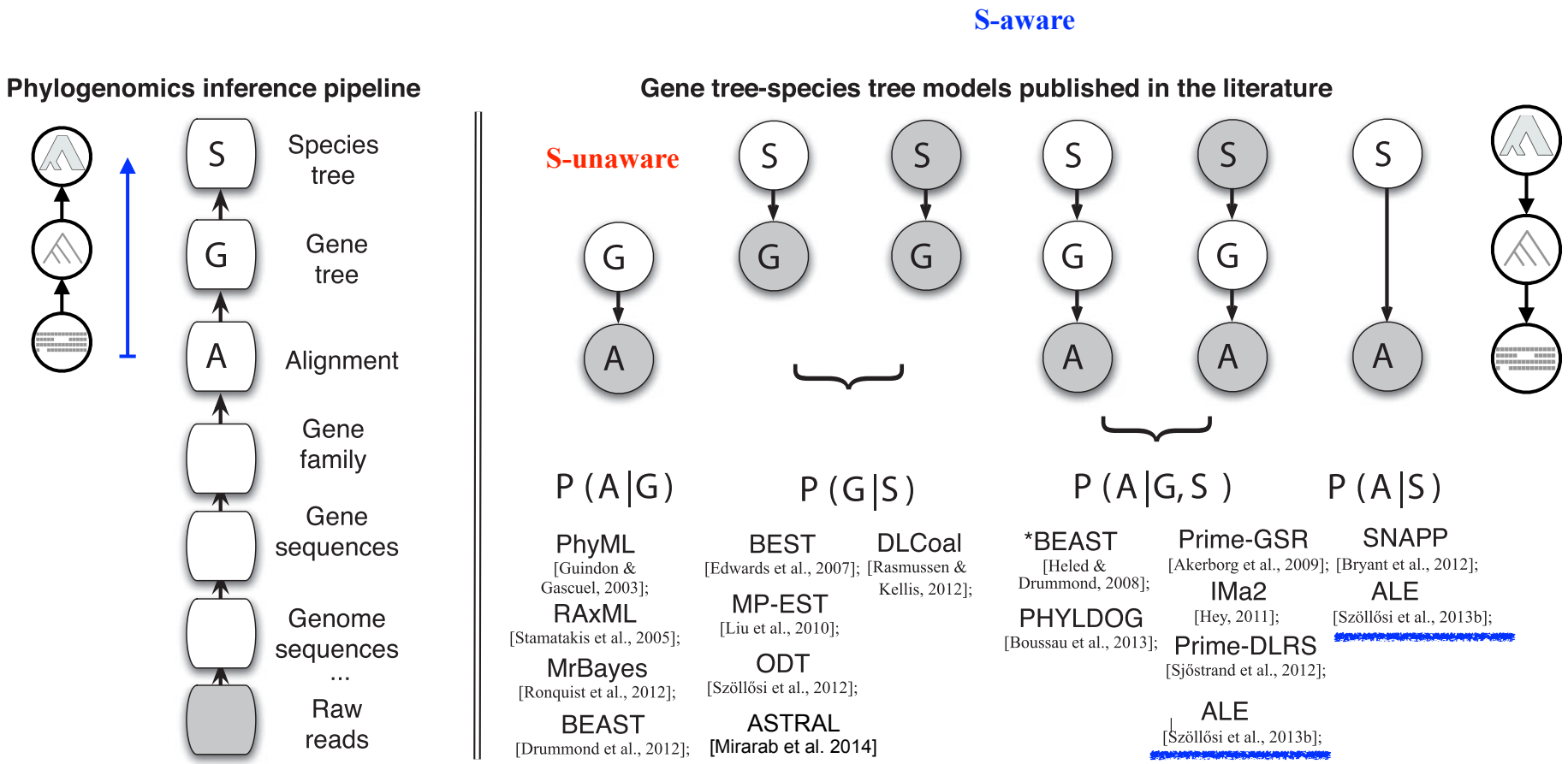
In contrast, the “aware” path models the dependency between each step using knowledge from different fields of biology..



Daubin & Boussau 2011 TrEE

"*phylogenomics — why we are doing it all wrong*"

Species tree-awareness

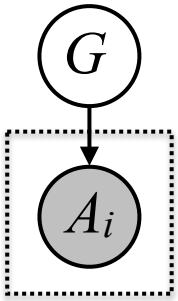


Szöllősi,.., Boussau 2015 Syst. Biol.

Molecular phylogenetics infers gene trees based on sequence..

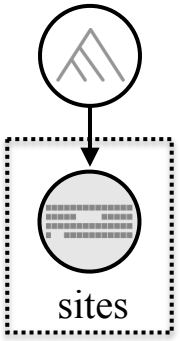
"sequence only" inference

species tree-unaware

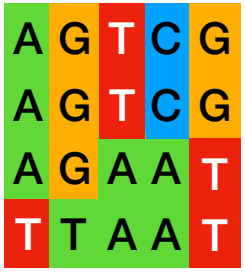


gene tree

sequences

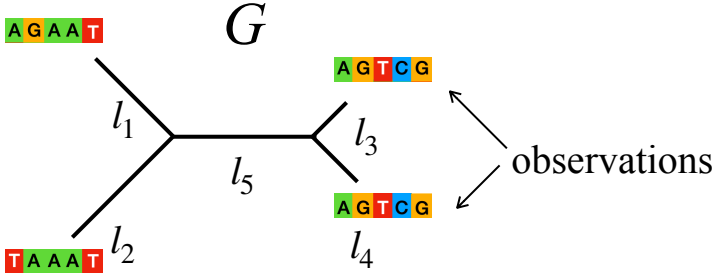


$A_i, A_{i+1}, A_{i+2}, \ldots$



Which **gene tree** produced my **sequences**?

gene tree

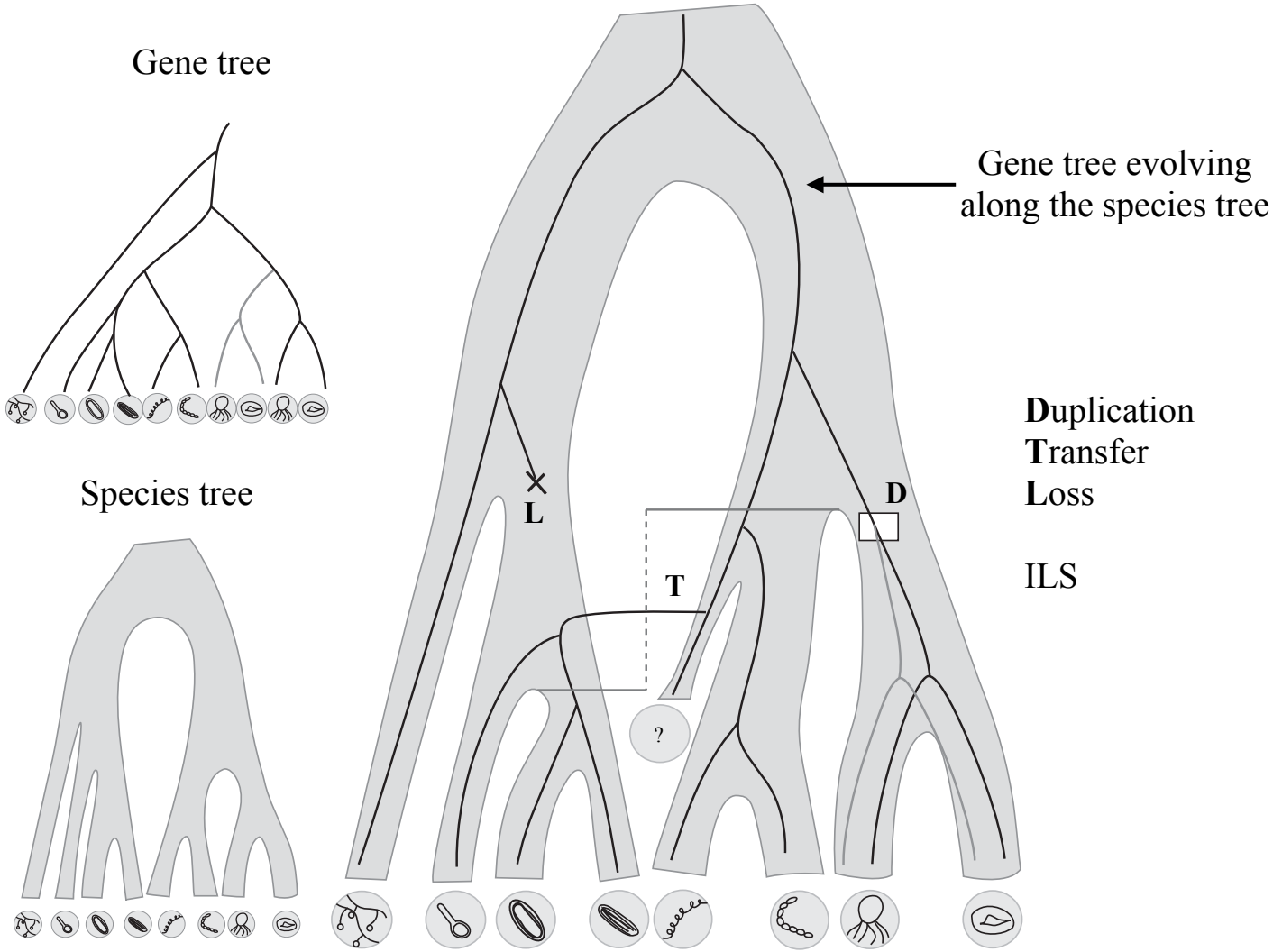


RAxML

MrBayes PhyloBayes

The problem is gene trees are not species trees

A gene tree is a deformation of the species tree through the prism of genome evolution and population genetics processes.

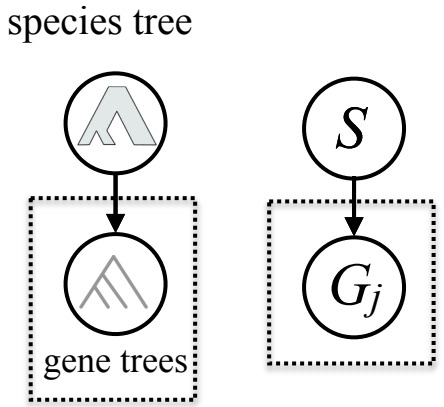


Daubin & Szöllősi 2016

.. but gene trees are generated along the species tree

A gene tree is a deformation of the species tree through the prism of genome evolution and population genetics processes. As a result some gene trees are more likely than others given a particular species tree, informing us about the species tree along which they were generated.

Daubin & Boussau 2011 Trends Ecol. Evol.



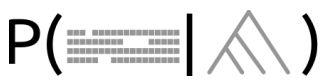
ssolo@elte.hu

The solution is to model how gene trees are generated along the species tree

A gene tree is a deformation of the species tree through the prism of genome evolution and population genetics processes. As a result some gene trees are more likely than others given a particular species tree, informing us about the species tree along which they were generated.

Which **gene tree** produced my **sequences**?

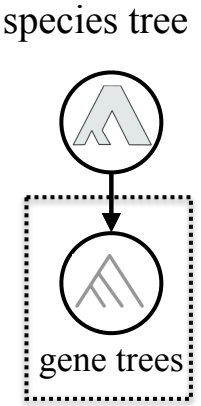
DATA **MODEL**

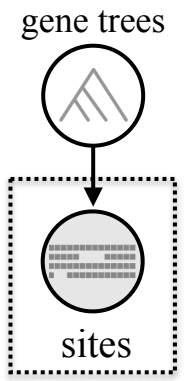


Which **species tree** produced my **gene trees**?

DATA **MODEL**

Genomes as documents of evolutionary history
Daubin & Boussau 2011 Trends Ecol. Evol.





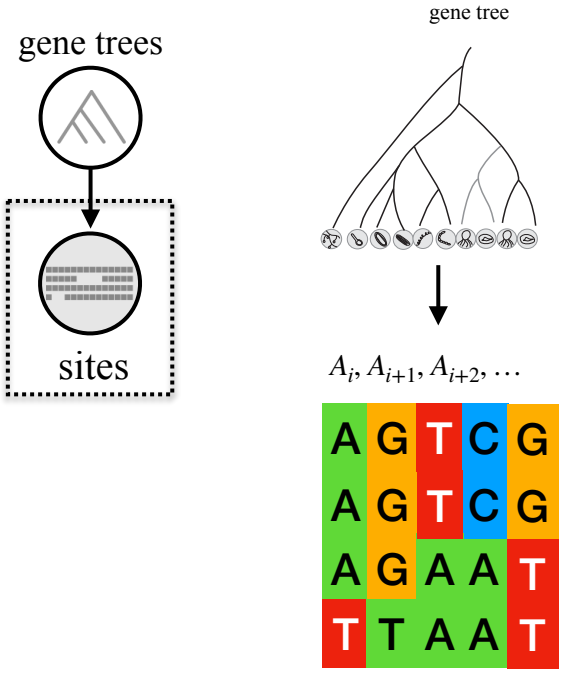
gergely.szollosi@oist.jp

The solution is to model how gene trees are generated along the species tree

A species tree induces a probability distribution over gene trees (some gene trees are more likely than others given a particular species tree), and in return, the inferred distribution of gene trees informs about the species tree along which they were generated.

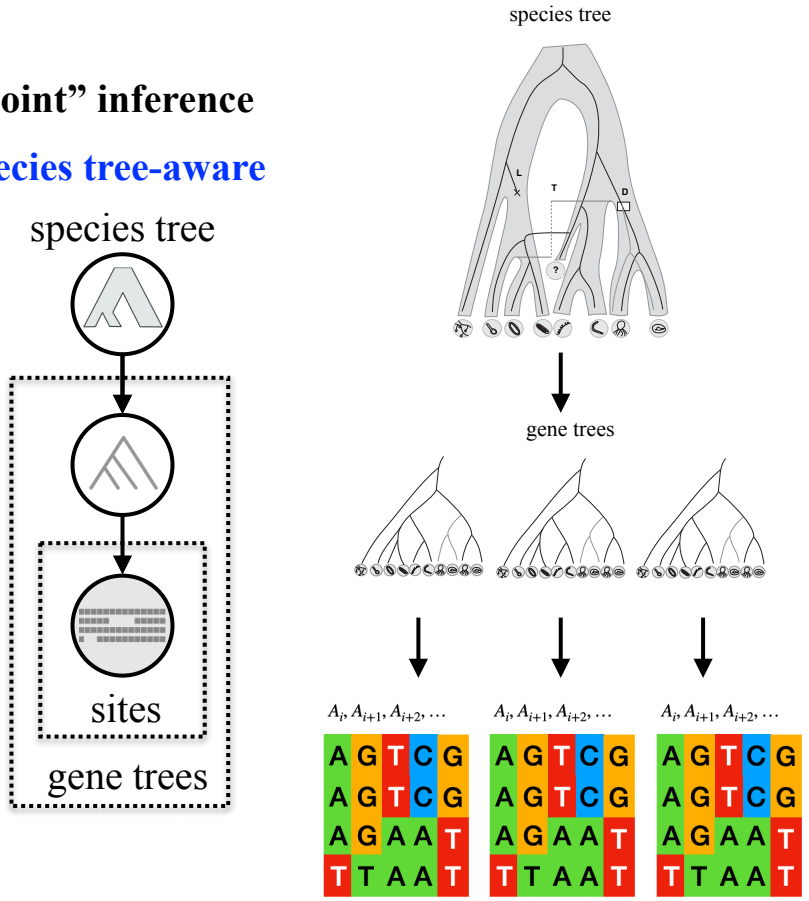
"sequence only" inference

species tree-unaware



"joint" inference

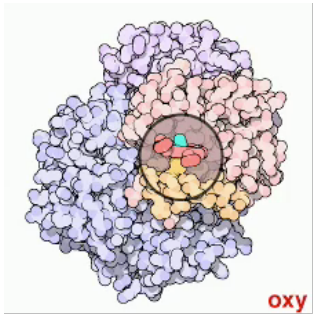
species tree-aware



The stories gene families can be complicated

The story of each gene family consist of a unique series of evolutionary events that often results in a change of copy number and shifts in function.

Human hemolglobin is composed of

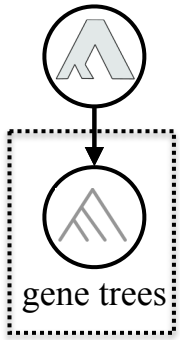


molecular machine

$2\alpha + 2\beta$ **chains.**

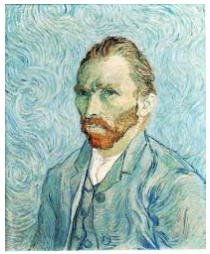
DL

species tree



Human **Cow** **Horse**

$2\alpha+$



adult

$+2\beta$ (97%)

$+2\delta$ (3%)

adult

$+2\{\beta\delta\}$

adult and fetus

$+2\{\beta\delta\}$

fetus

$+2\gamma$

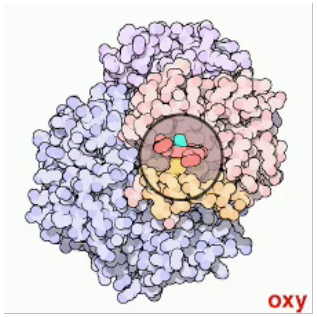
fetus

$+2\gamma$

The stories gene families can be complicated

The story of each gene family consist of a unique series of evolutionary events that often results in a change of copy number and shifts in function.

Human hemolglobin is composed of

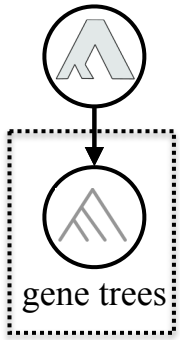


molecular machine

$2\alpha + 2\beta$ **chains.**

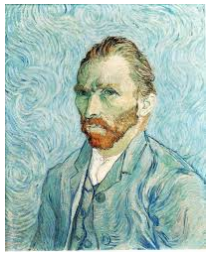
DL

species tree



Human **Cow** **Horse**

$2\alpha+$



adult

$+2\beta$ (97%)

$+2\delta$ (3%)

adult

$+2\{\beta\delta\}$

adult and fetus

$+2\{\beta\delta\}$

fetus

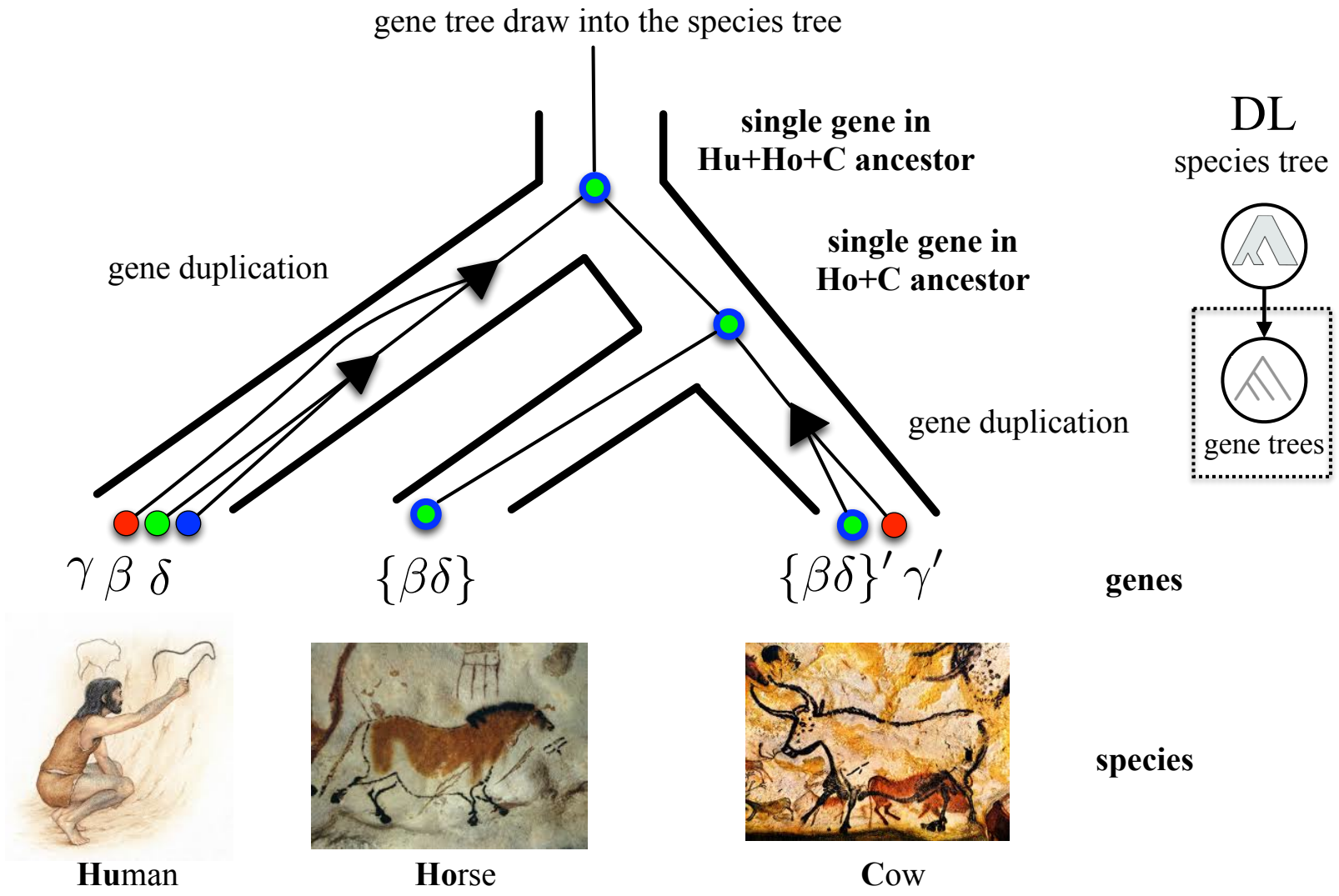
$+2\gamma$

fetus

$+2\gamma$

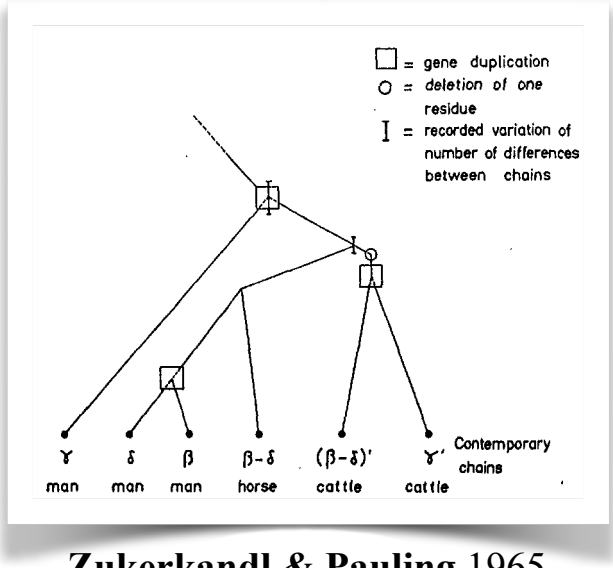
The stories gene families can be complicated

The story of each gene family consist of a unique series of evolutionary events that often results in a change of copy number and shifts in function.

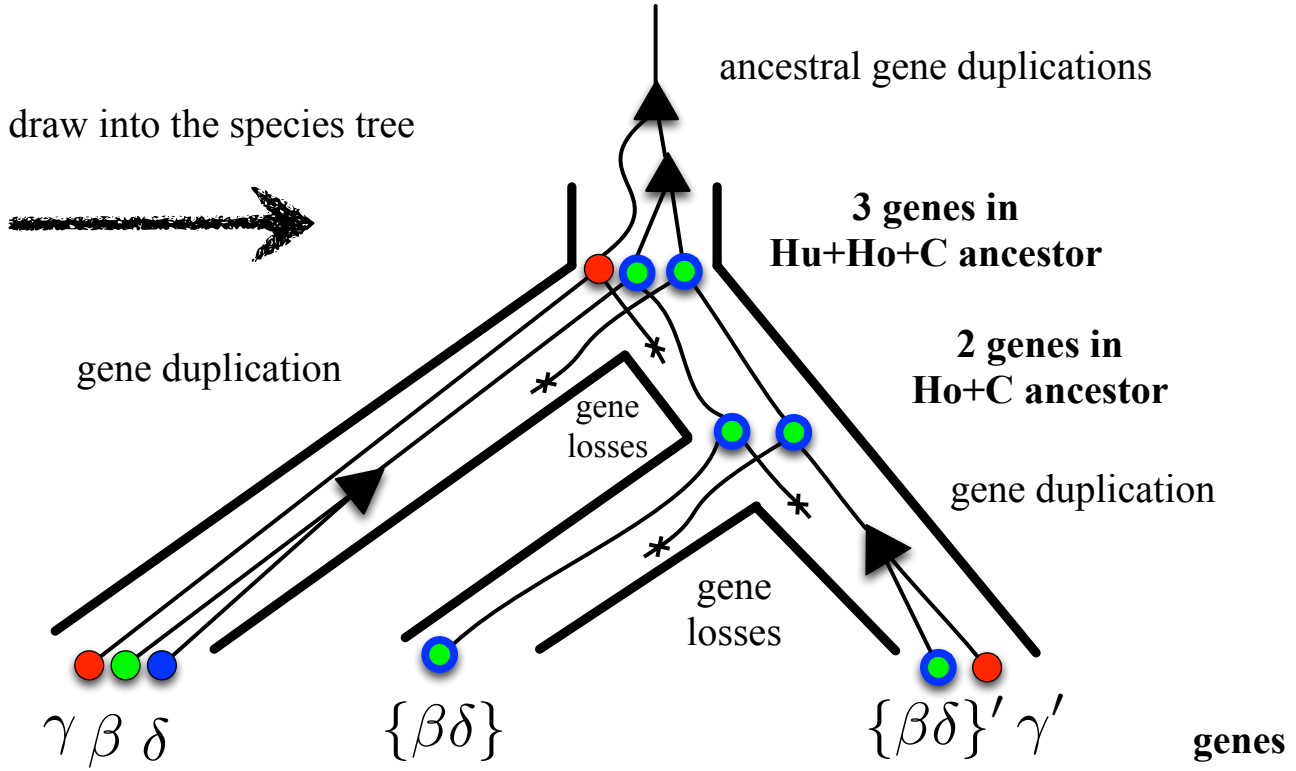


The story of individual gene families is often blurred

Errors in gene trees will result in conflicts with the species tree that imply spurious evolutionary events.



Zukerkandl & Pauling 1965



Human

Horse

Cow

species

The solution is to model how gene trees are generated along the species tree

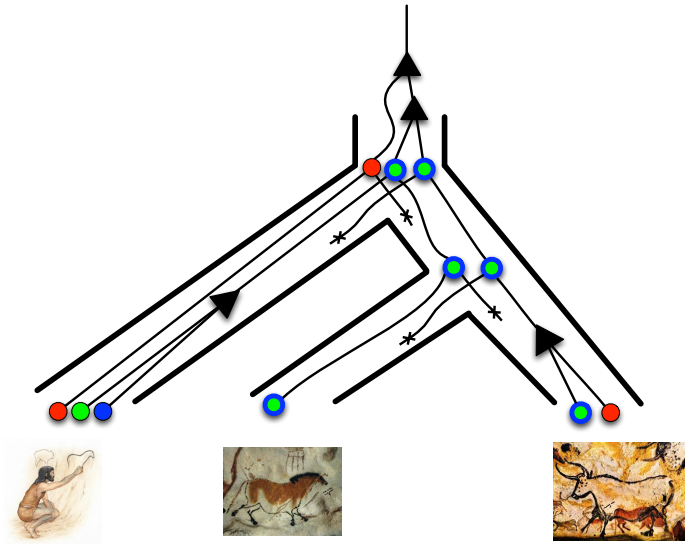
Given a model of gene family evolution a species tree induces a probability distribution over gene trees, thus **some gene trees are more probable than others given a particular species tree**.

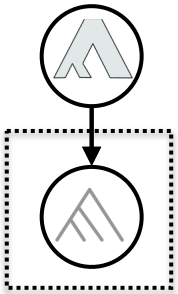
less probable gene tree

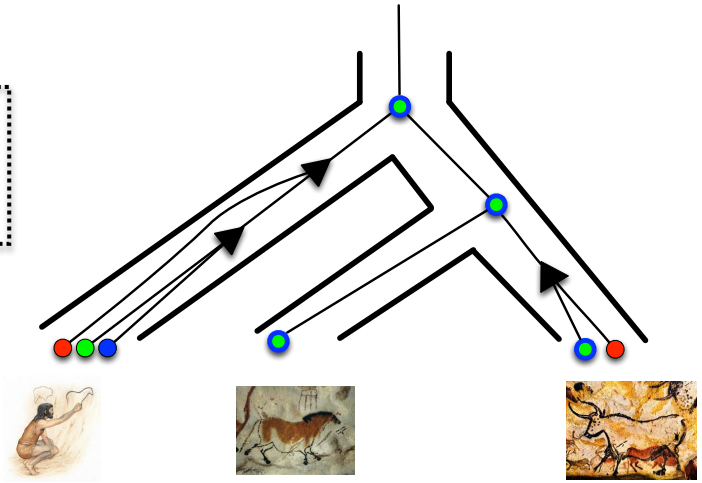
gene tree species tree

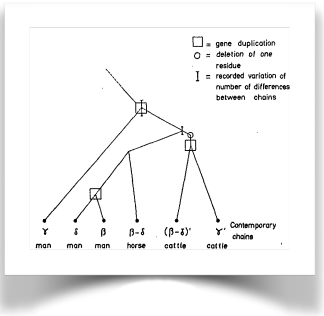
$P(\text{gene tree} \mid \text{species tree})$

more probable gene tree





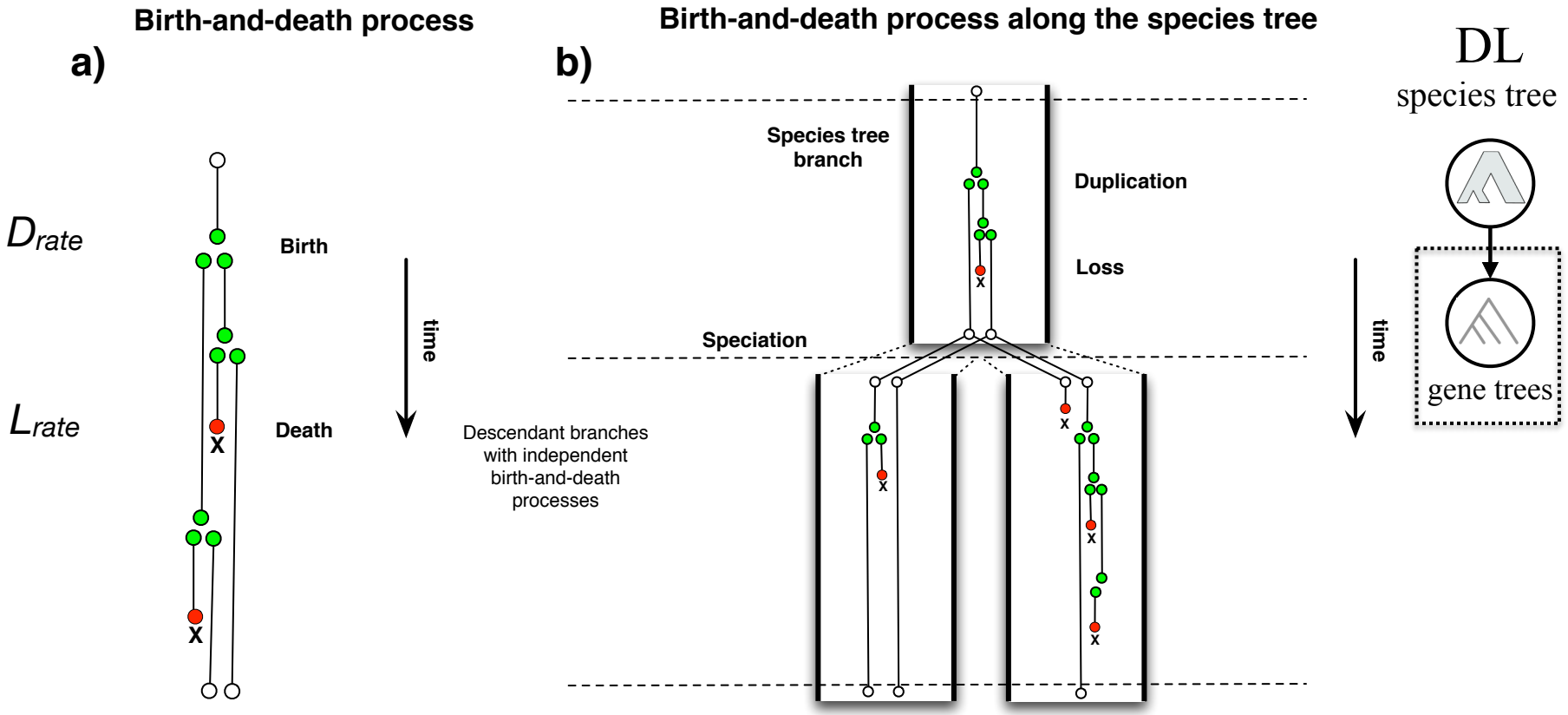




Zukerkandl & Pauling 1965

The solution is to model how gene trees are generated along the species tree

Given a model of gene family evolution a species tree induces a probability distribution over gene trees.



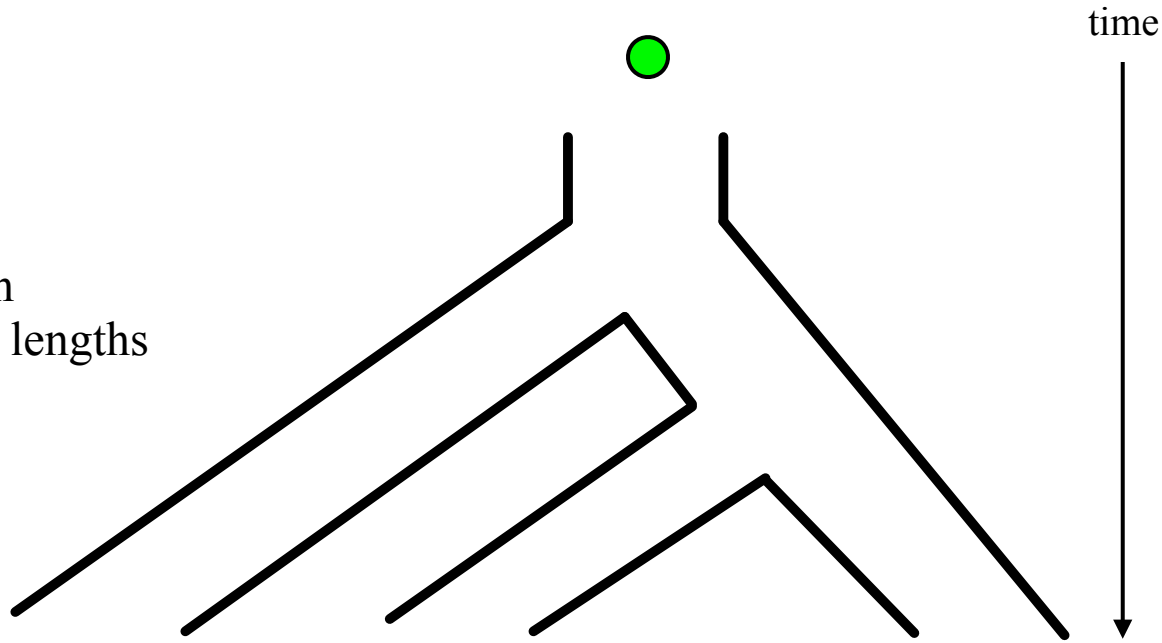
The solution is to model how gene trees are generated along the species tree

Given a model of gene family evolution a species tree induces a probability distribution over gene trees.

D_{rate}

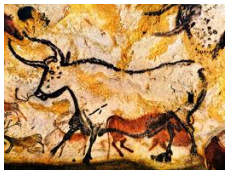
L_{rate}

Species tree with
time like branch lengths



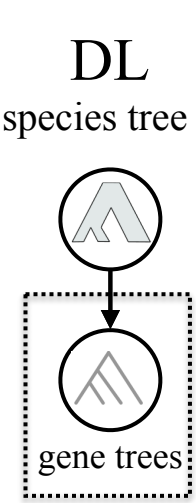
Human

Horse



Cow

species



implemented in ALE:

http://github.com/ssolo/ALE

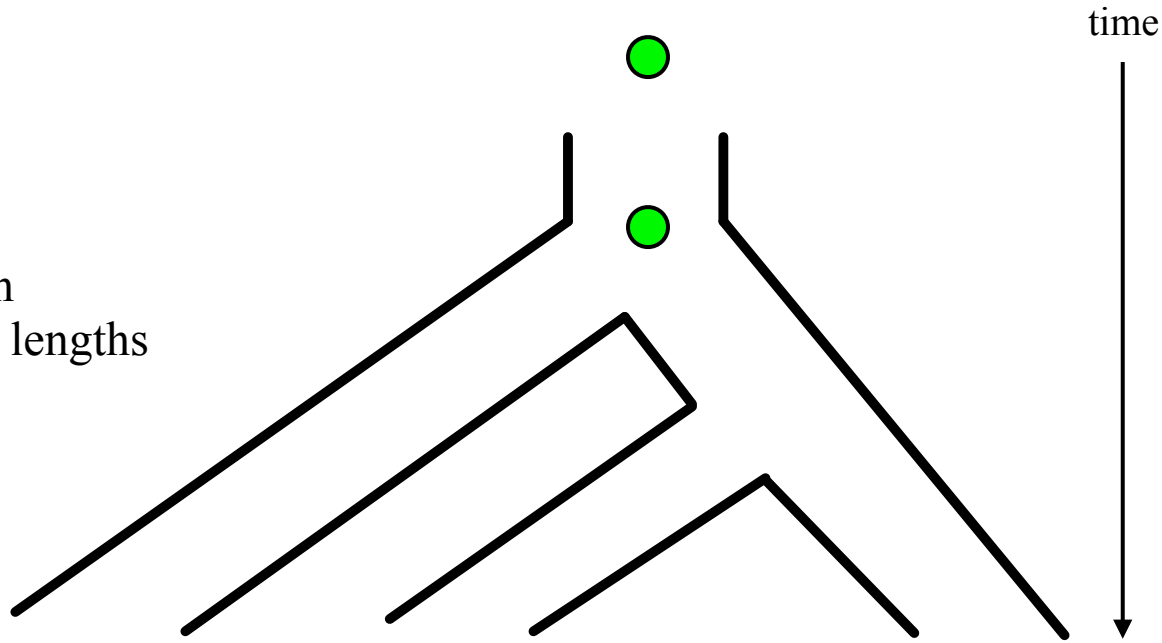
The solution is to model how gene trees are generated along the species tree

Given a model of gene family evolution a species tree induces a probability distribution over gene trees.

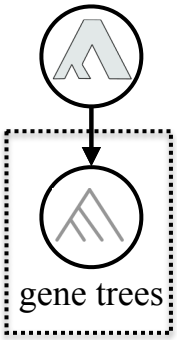
D_{rate}

L_{rate}

Species tree with
time like branch lengths

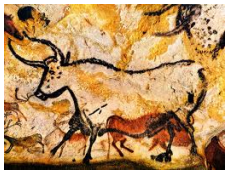


DL
species tree



Human

Horse



Cow

species

implemented in ALE:

http://github.com/ssolo/ALE

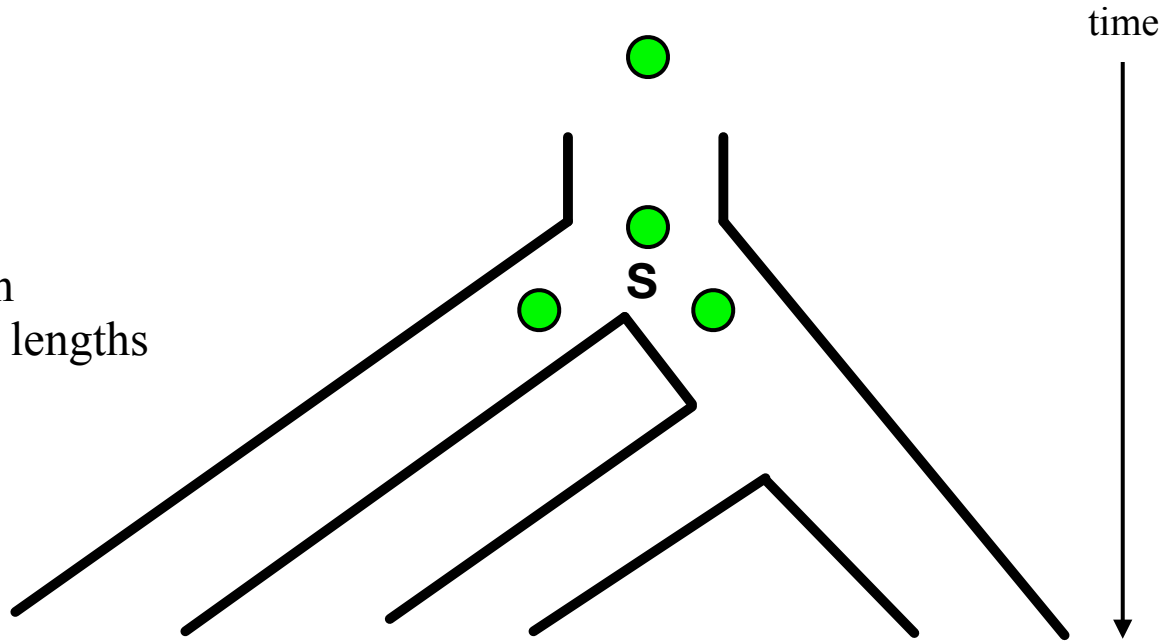
The solution is to model how gene trees are generated along the species tree

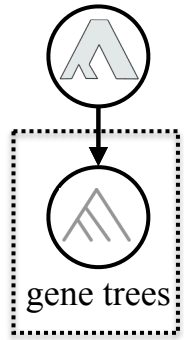
Given a model of gene family evolution a species tree induces a probability distribution over gene trees.

D_{rate}

L_{rate}

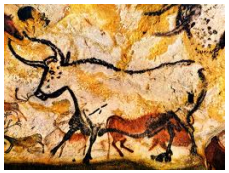
Species tree with
time like branch lengths





Human

Horse



Cow

species

implemented in ALE:

http://github.com/ssolo/ALE

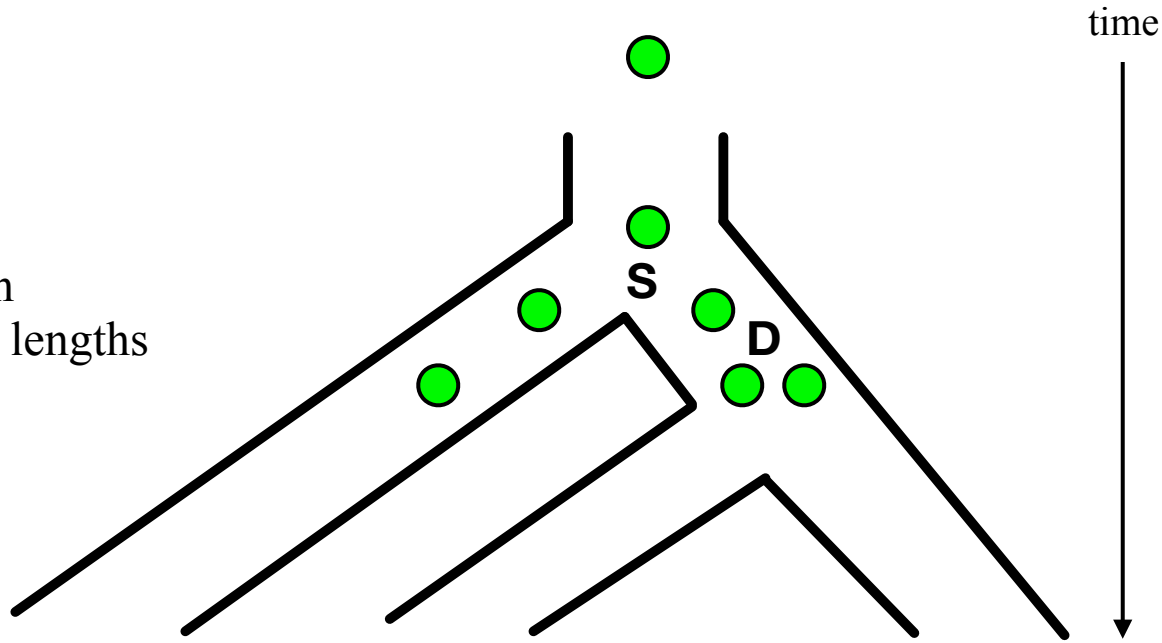
The solution is to model how gene trees are generated along the species tree

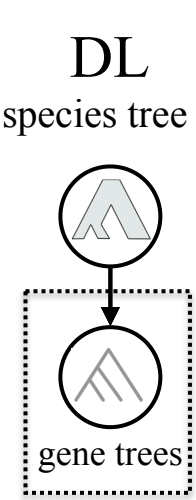
Given a model of gene family evolution a species tree induces a probability distribution over gene trees.

D_{rate}

L_{rate}

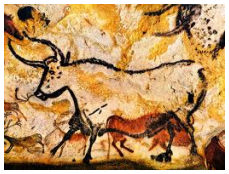
Species tree with
time like branch lengths





Human

Horse



Cow

species

implemented in ALE:

http://github.com/ssolo/ALE

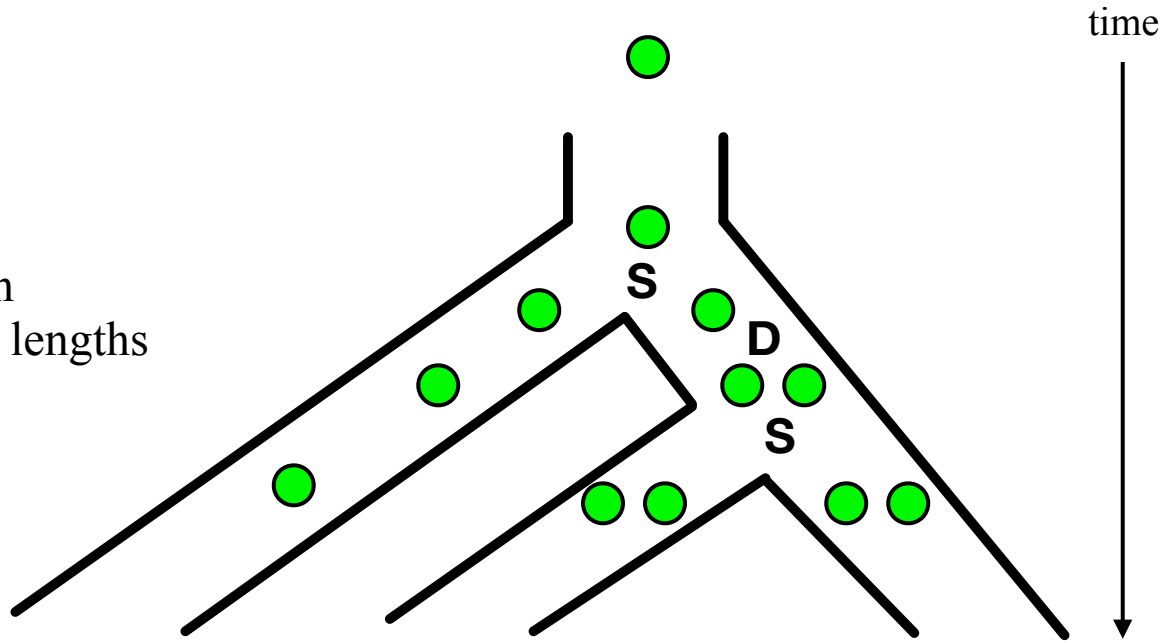
The solution is to model how gene trees are generated along the species tree

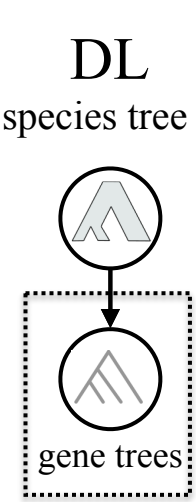
Given a model of gene family evolution a species tree induces a probability distribution over gene trees.

D_{rate}

L_{rate}

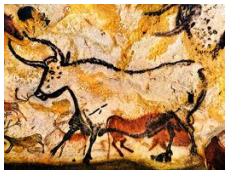
Species tree with
time like branch lengths





Human

Horse



Cow

species

implemented in ALE:

http://github.com/ssolo/ALE

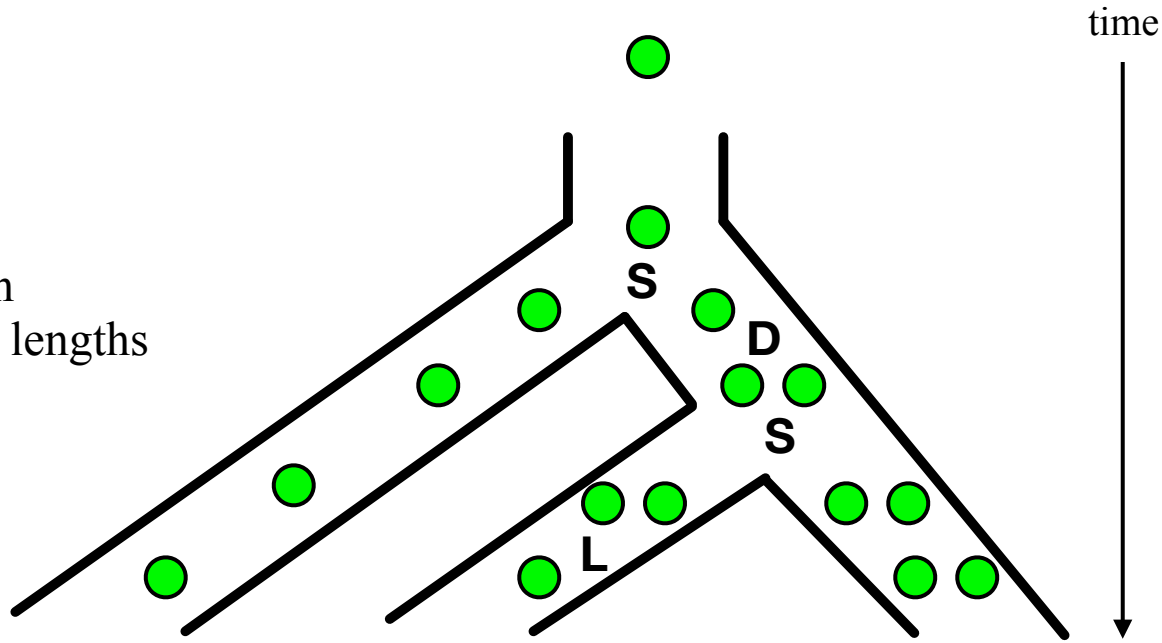
The solution is to model how gene trees are generated along the species tree

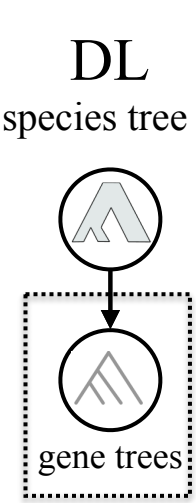
Given a model of gene family evolution a species tree induces a probability distribution over gene trees.

D_{rate}

L_{rate}

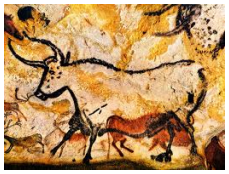
Species tree with
time like branch lengths





Human

Horse



Cow

species

implemented in ALE:

http://github.com/ssolo/ALE

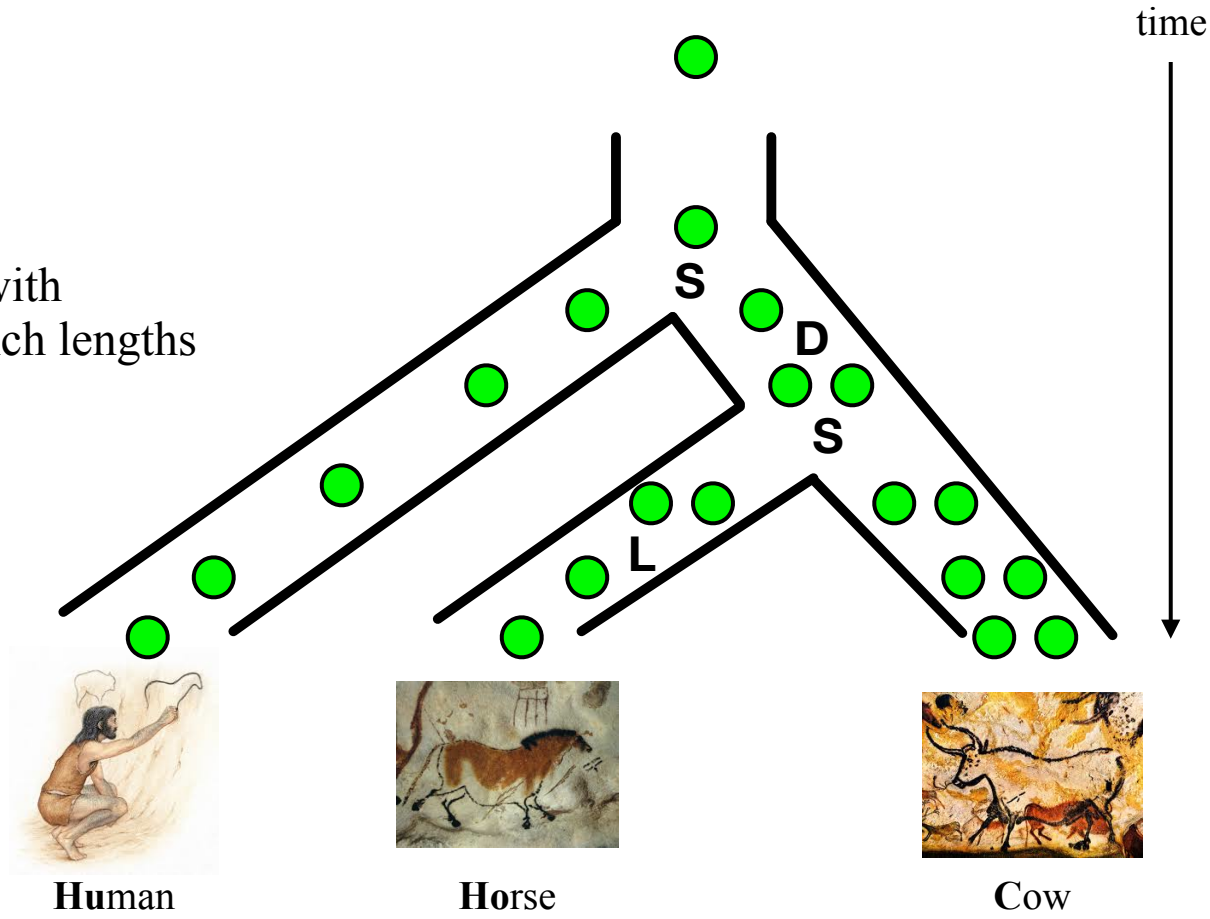
The solution is to model how gene trees are generated along the species tree

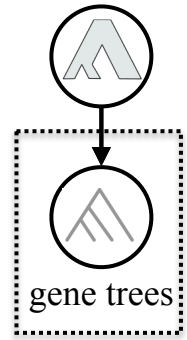
Given a model of gene family evolution a species tree induces a probability distribution over gene trees.

D_{rate}

L_{rate}

Species tree with
time like branch lengths





implemented in ALE:

http://github.com/ssolo/ALE

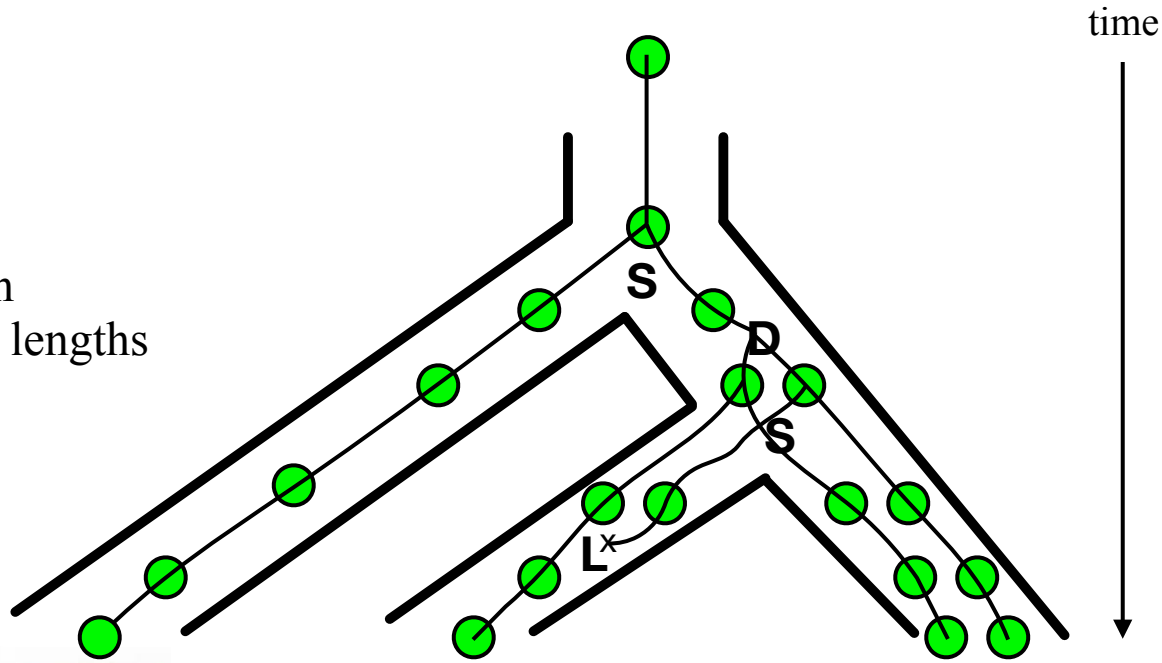
The solution is to model how gene trees are generated along the species tree

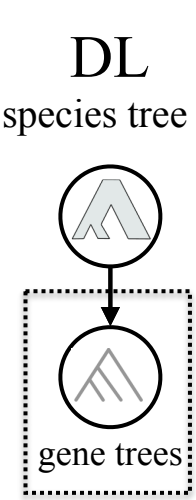
Given a model of gene family evolution a species tree induces a probability distribution over gene trees.

D_{rate}

L_{rate}

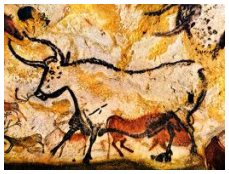
Species tree with
time like branch lengths





Human

Horse



Cow

species

implemented in ALE:

http://github.com/ssolo/ALE

The solution is to model how gene trees are generated along the species tree

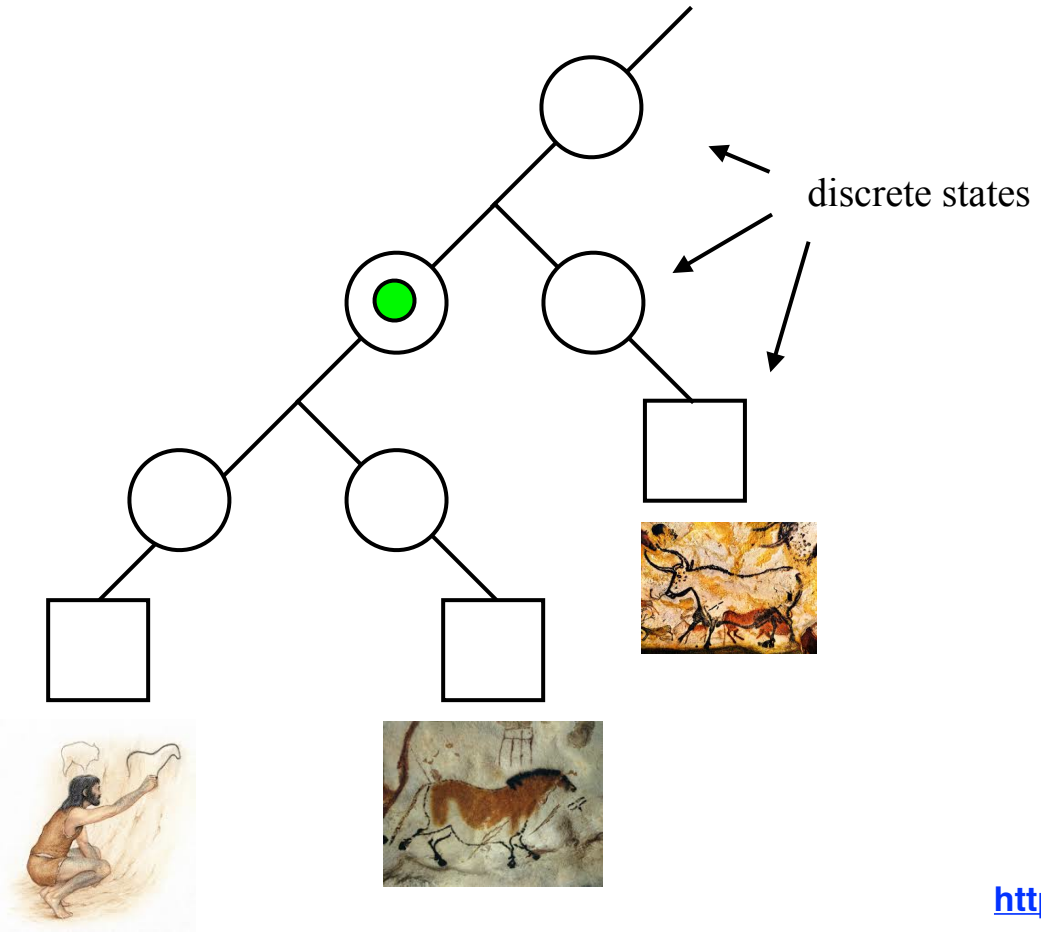
Given a model of gene family evolution a species tree induces a probability distribution over gene trees.

D_{rate}

L_{rate}

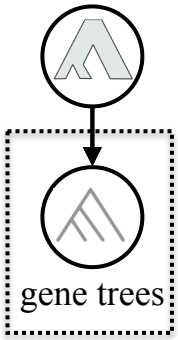
Rooted species tree

$P_S + P_D + P_L = 1$



DL

species tree



implemented in ALE:

http://github.com/ssolo/ALE

The solution is to model how gene trees are generated along the species tree

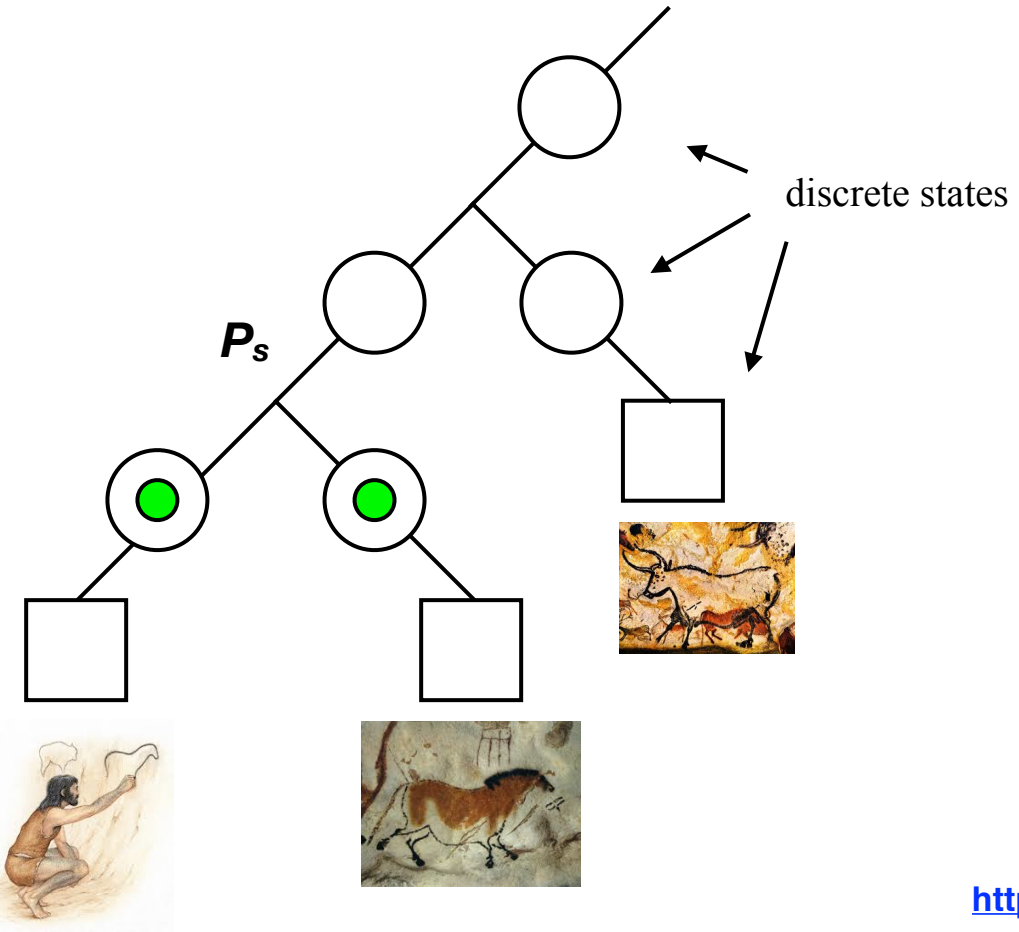
Given a model of gene family evolution a species tree induces a probability distribution over gene trees.

D_{rate}

L_{rate}

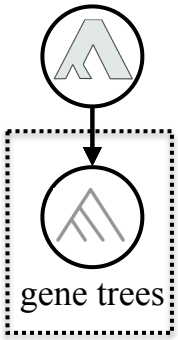
Rooted species tree

$P_S + P_D + P_L = 1$



DL

species tree



implemented in ALE:

http://github.com/ssolo/ALE

The solution is to model how gene trees are generated along the species tree

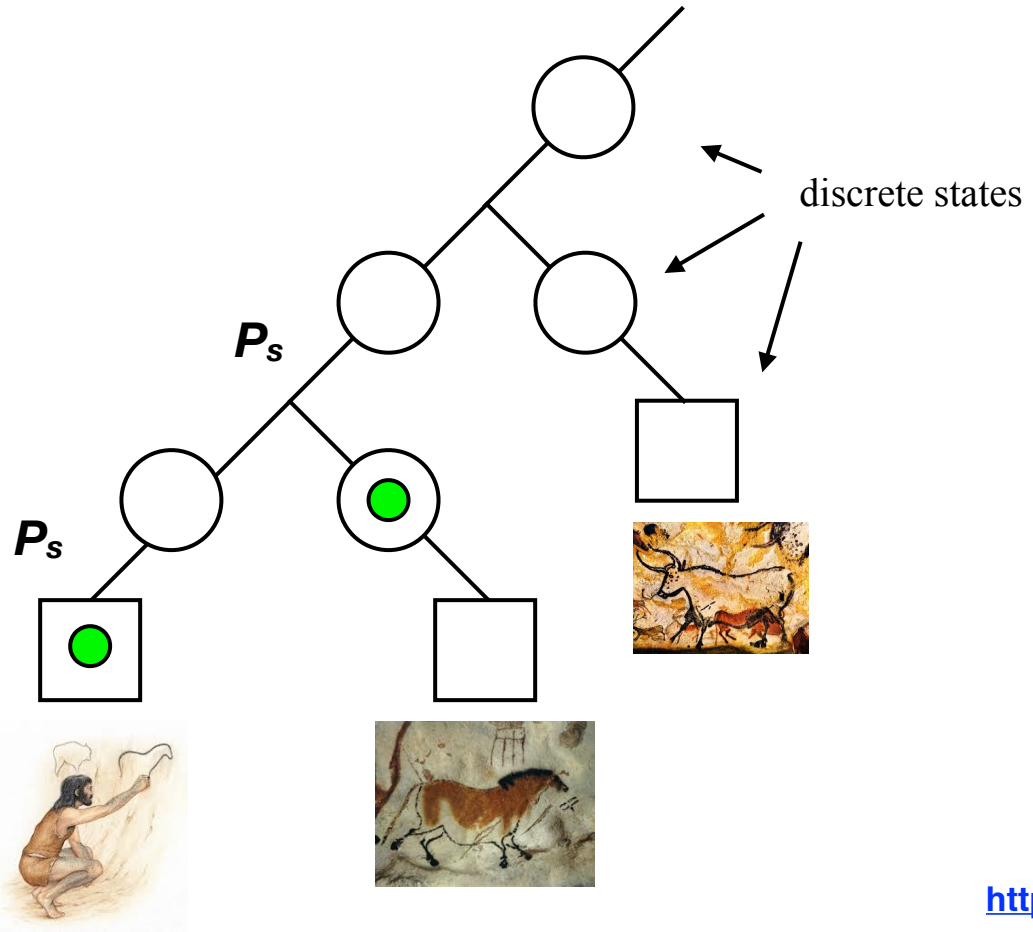
Given a model of gene family evolution a species tree induces a probability distribution over gene trees.

D_{rate}

L_{rate}

Rooted species tree

$P_S + P_D + P_L = 1$



DL

species tree

gene trees

implemented in ALE:

http://github.com/ssolo/ALE

The solution is to model how gene trees are generated along the species tree

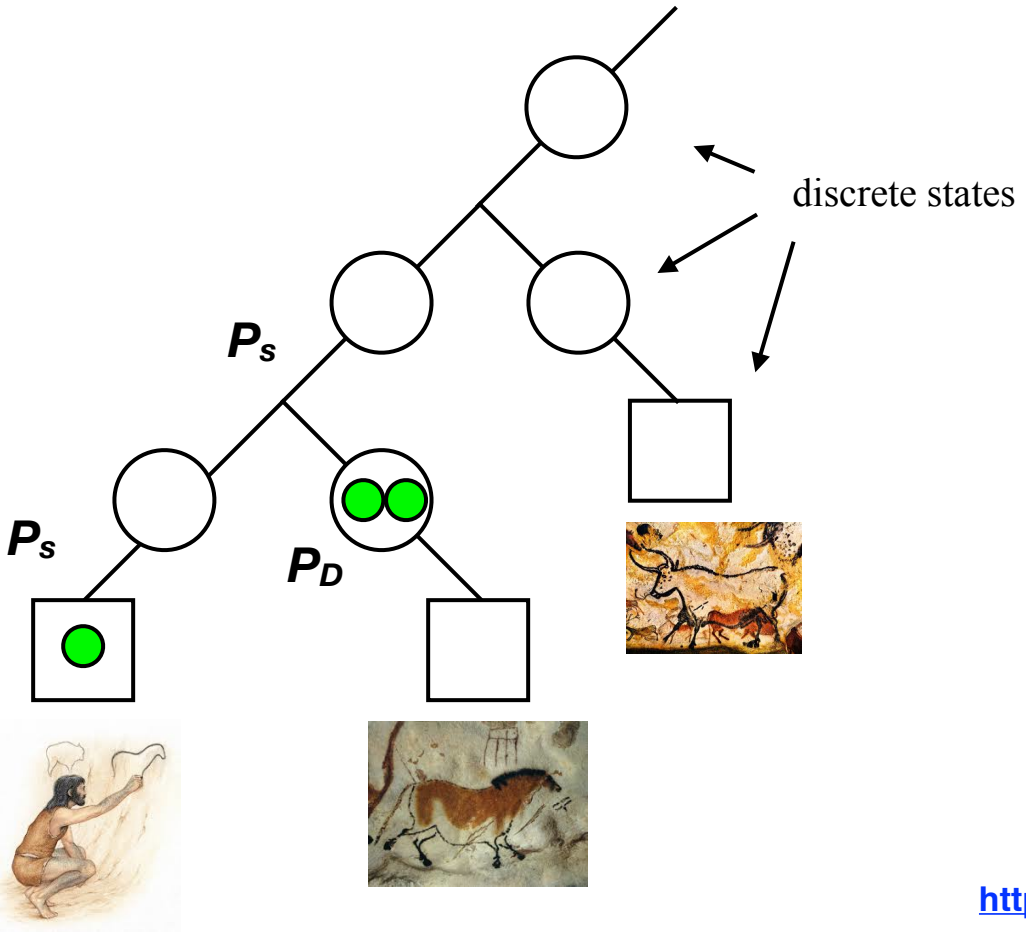
Given a model of gene family evolution a species tree induces a probability distribution over gene trees.

D_{rate}

L_{rate}

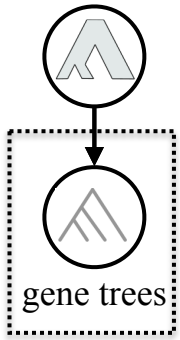
Rooted species tree

$P_S + P_D + P_L = 1$



DL

species tree



implemented in ALE:

http://github.com/ssolo/ALE

The solution is to model how gene trees are generated along the species tree

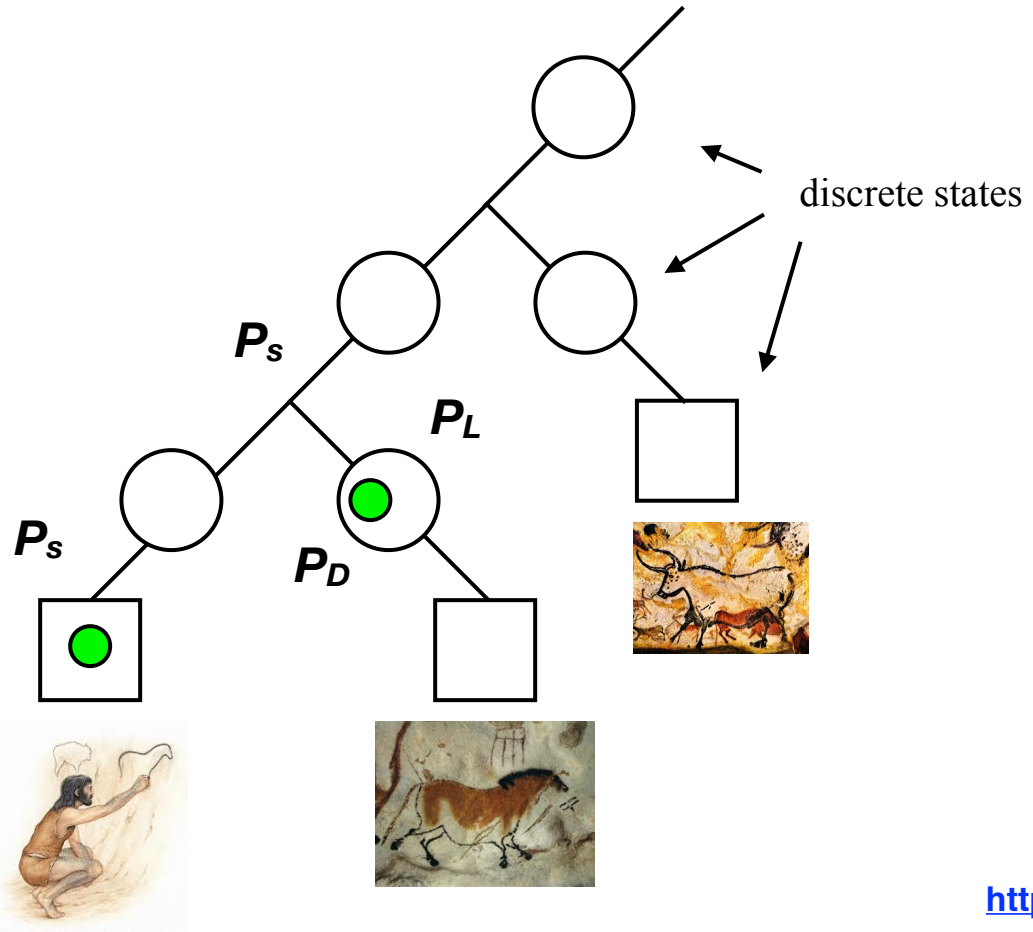
Given a model of gene family evolution a species tree induces a probability distribution over gene trees.

D_{rate}

L_{rate}

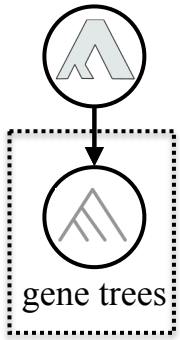
Rooted species tree

$P_S + P_D + P_L = 1$



DL

species tree



implemented in ALE:

http://github.com/ssolo/ALE

The solution is to model how gene trees are generated along the species tree

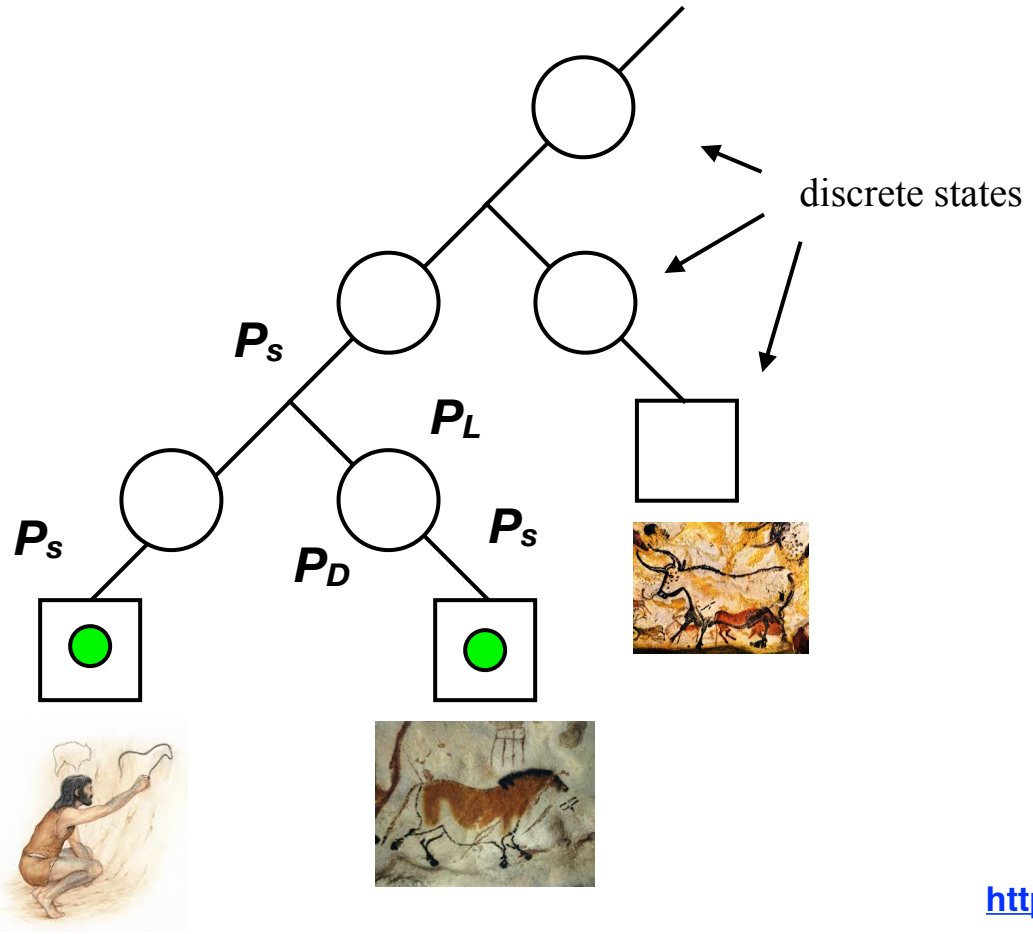
Given a model of gene family evolution a species tree induces a probability distribution over gene trees.

D_{rate}

L_{rate}

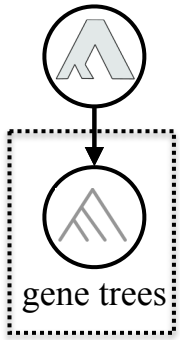
Rooted species tree

$P_S + P_D + P_L = 1$



DL

species tree



implemented in ALE:

http://github.com/ssolo/ALE

The solution is to model how gene trees are generated along the species tree

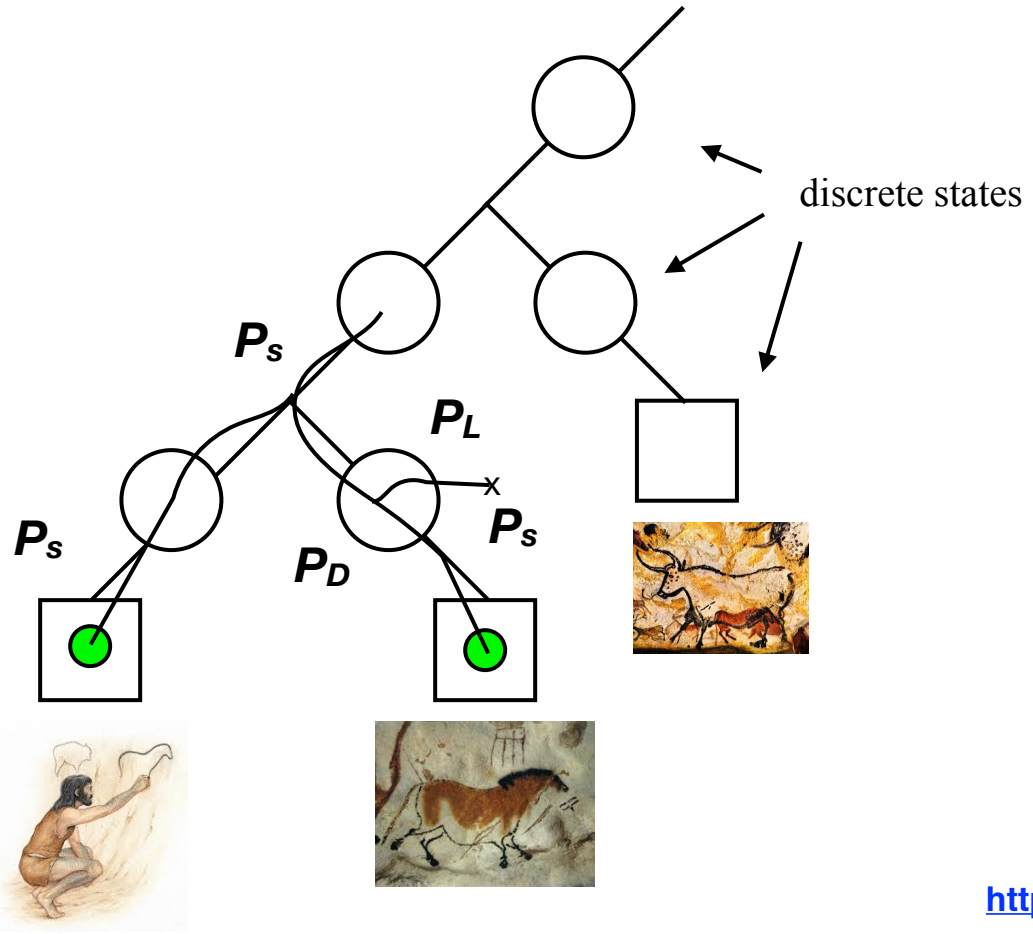
Given a model of gene family evolution a species tree induces a probability distribution over gene trees.

D_{rate}

L_{rate}

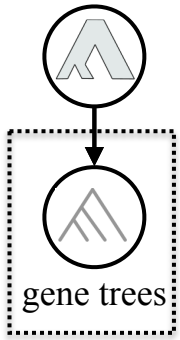
Rooted species tree

$P_S + P_D + P_L = 1$



DL

species tree



implemented in ALE:

http://github.com/ssolo/ALE

We are very good at reconstructing gene trees..

Calculating the likelihood of the sequences A given the gene tree G

$$P(\text{tree} \mid \text{alignment}) = \prod_i P(A_i \mid G, Q)$$

requires summing over all possible substitution paths.

DATA

observations
sequences
(alignment)

sites

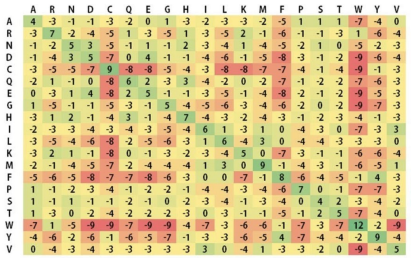
$A_i, A_{i+1}, A_{i+2}, \ldots$

gene from species 1.	…	A	G	T	C	G	…
gene from species 2.	…	A	G	T	C	G	…
gene from species 3.	…	A	G	A	A	T	…
gene from species 4.	…	T	T	A	A	T	…

MODEL

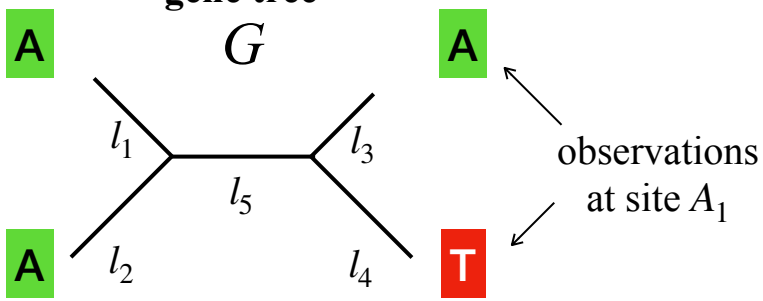
matrix of substitution rates

$$Q = \begin{array}{c} \\ A \\ T \\ G \\ C \end{array} \begin{array}{c} \begin{array}{cccc} A & T & G & C \end{array} \\ \begin{pmatrix} * & \kappa & 1 & 1 \\ \kappa & * & 1 & 1 \\ 1 & 1 & * & \kappa \\ 1 & 1 & \kappa & * \end{pmatrix} \end{array}$$



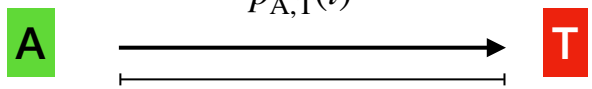
gene tree

G



(Felsenstein 1981)

$$P(t) = e^{tQ} = \{p_{i,j}(t)\}$$



We are very good at reconstructing gene trees..

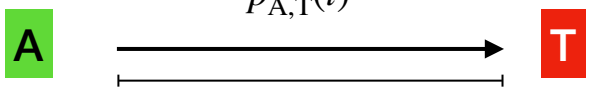
Calculating the likelihood of the sequences A given the gene tree G

$$P(\text{alignment} \mid \text{tree}) = \prod_i P(A_i \mid G, Q)$$

requires summing over all possible substitution paths.

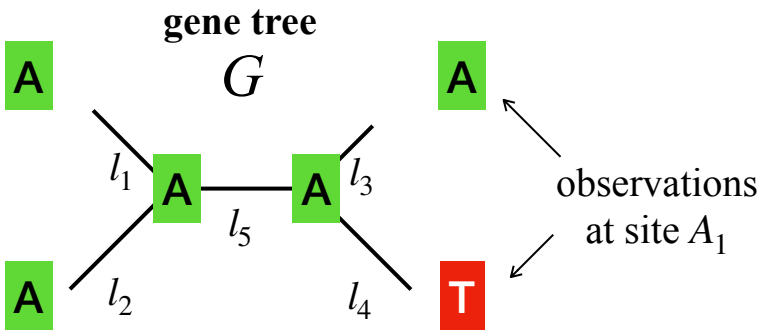
sum over subs. along branch conditional on states on top and bottom

$$P(t) = e^{tQ} = \{p_{i,j}(t)\}$$

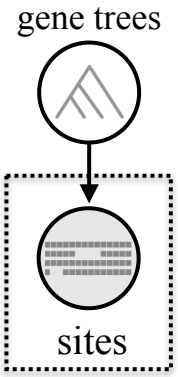


sum over ancestral states

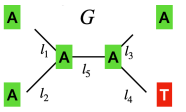
Felsenstein 1981



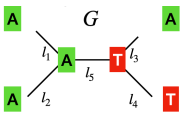
(Felsenstein 1981)



$$P(A_1 = \text{AAAT} \mid \text{tree}) = P_{AA}(l_1) \times P_{AA}(l_2) \times P_{AA}(l_5) \times P_{AA}(l_3) \times P_{AT}(l_4) +$$



$$P_{AA}(l_1) \times P_{AA}(l_2) \times P_{AA}(l_5) \times P_{TA}(l_3) \times P_{TT}(l_4) + ..$$



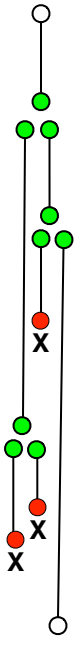
The solution is to model how gene trees are generated along the species tree

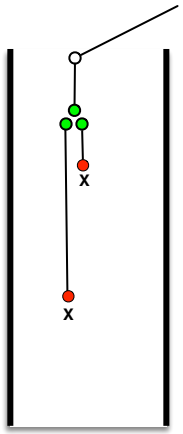
Calculating the likelihood $P(\text{gene tree} \mid \text{species tree})$ requires

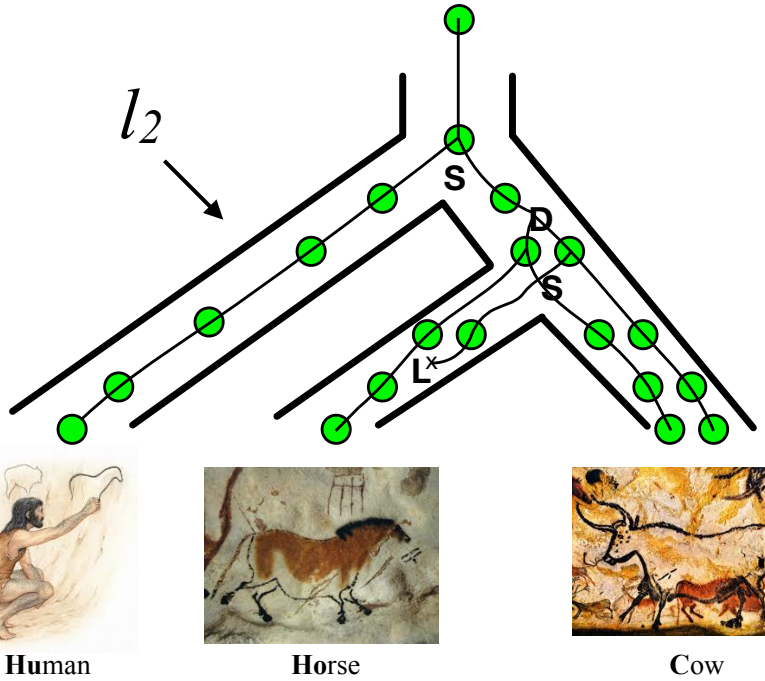
summing over all possible ***gene birth and death events*** along a given ***species tree.***

sum over gene birth and death events
along a branch conditional on reconciliation

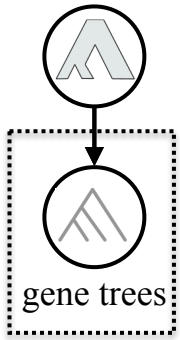
$P_{11}(l_{2,}\mathbf{Hu})$ $P_{10}(\mathbf{Ho})$







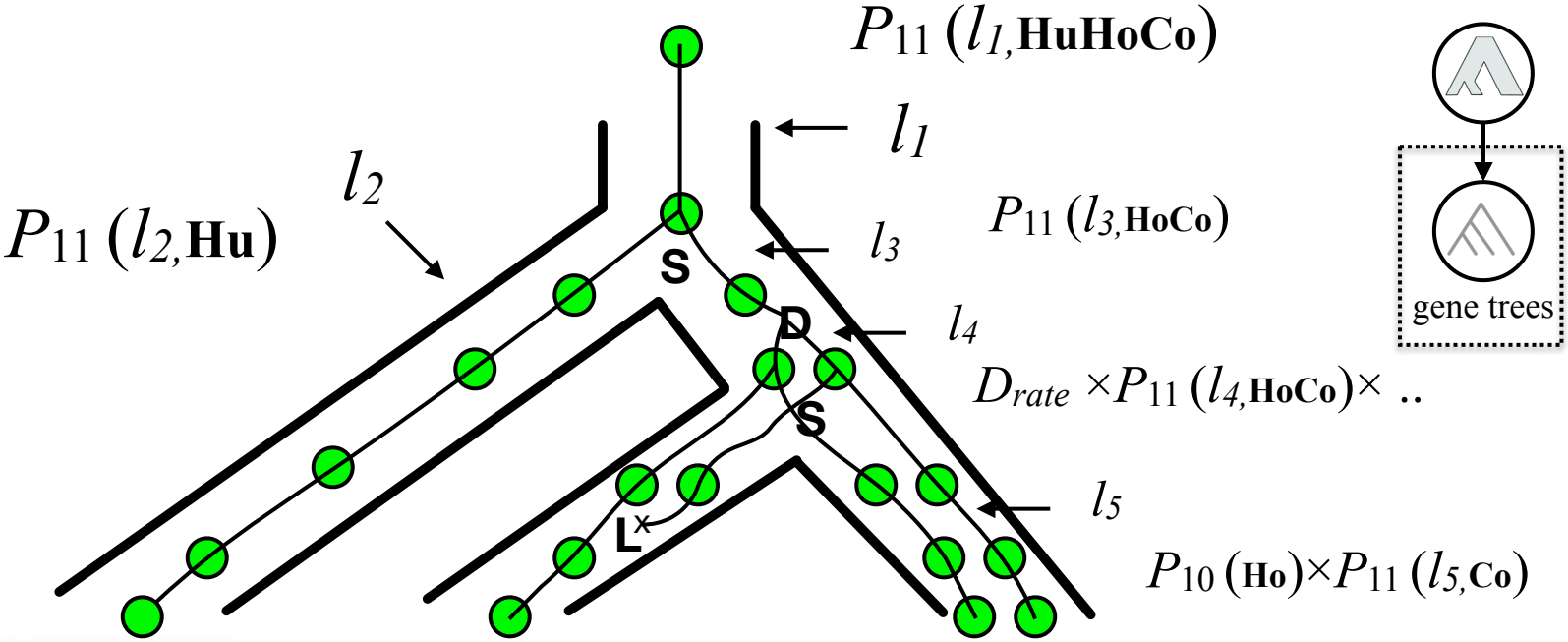
DL
species tree



The solution is to model how gene trees are generated along the species tree

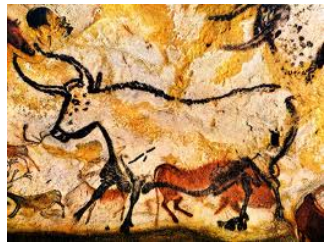
Calculating the likelihood $P(\text{gene tree} \mid \text{species tree})$ requires

summing over all possible ***gene birth and death events*** along a given ***species tree.***



Human

Horse

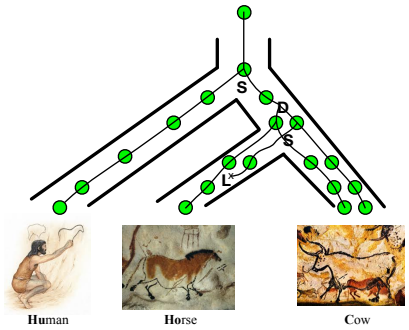


Cow

.. but gene trees are generated along the species tree

Calculating the likelihood $P(\text{gene tree} \mid \text{species tree})$ requires

summing over all possible ***gene birth and death events*** along a given ***species tree.***



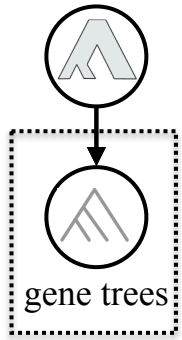
calculation complexity

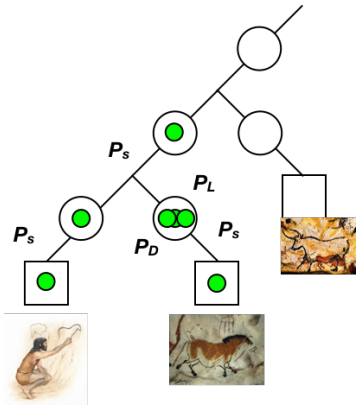
~10 × log(#species) × #genes

parameters (ML or Bayes)

D&L rates
branch lengths, root

DL
species tree





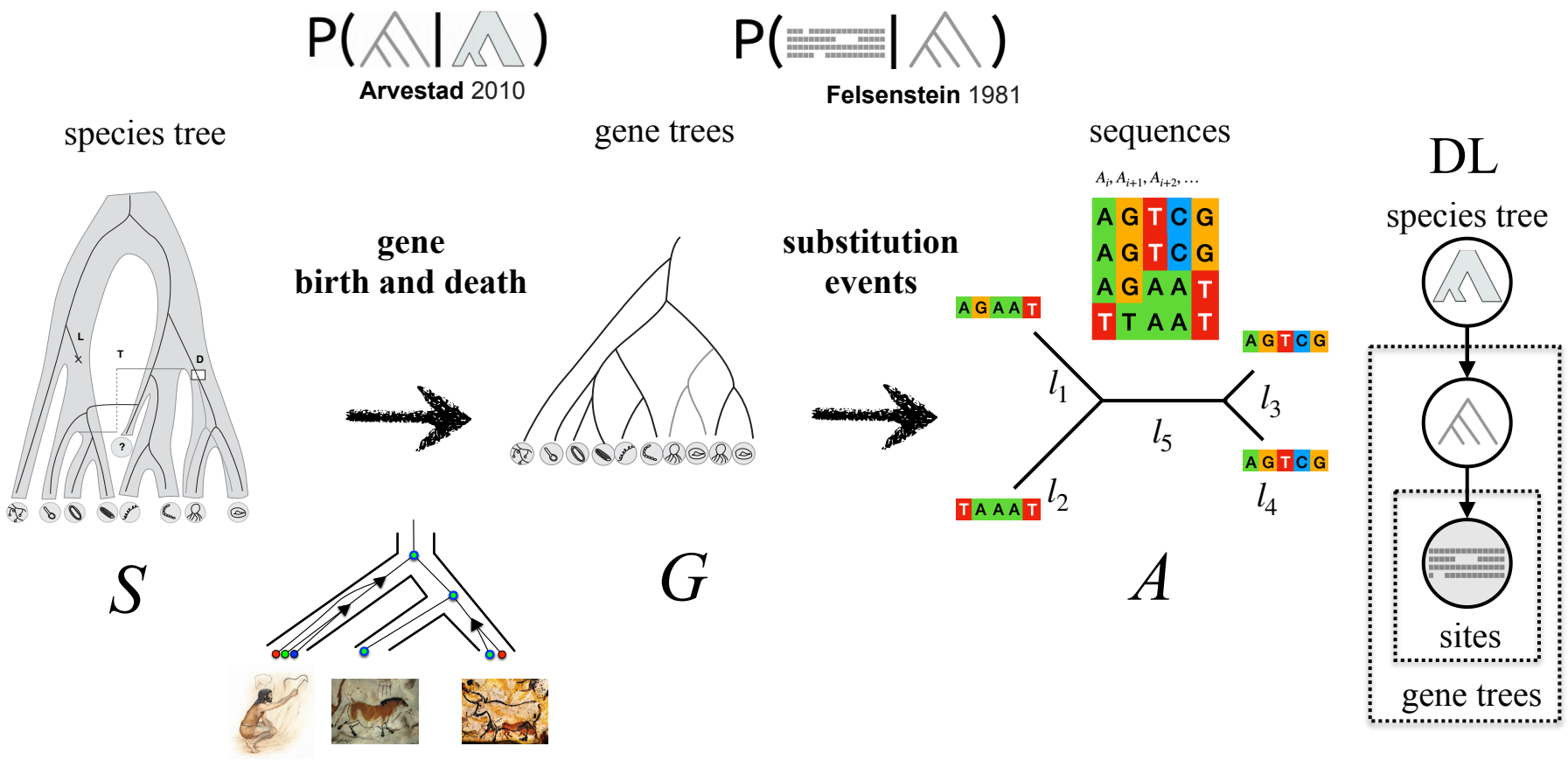
log(#species) × #genes

D&L rates
root

Gene trees and species trees can be jointly reconstructed

Estimating genes and species history can be achieved through a hierarchical structure, on top of which a species tree is inferred from gene trees through models of gene family evolution, themselves inferred from sequence alignments through models of sequence evolution.

Hierarchical generative model:



Gene trees and species trees can be jointly reconstructed

Estimating genes and species history can be achieved through a hierarchical structure, on top of which a species tree is inferred from gene trees through models of gene family evolution, themselves inferred from sequence alignments through models of sequence evolution.

parallel computation scheme

$$\mathcal{L}(\{G_j\}, S, \text{rates}|\{A_{ij}\}) :$$

server:
calculate

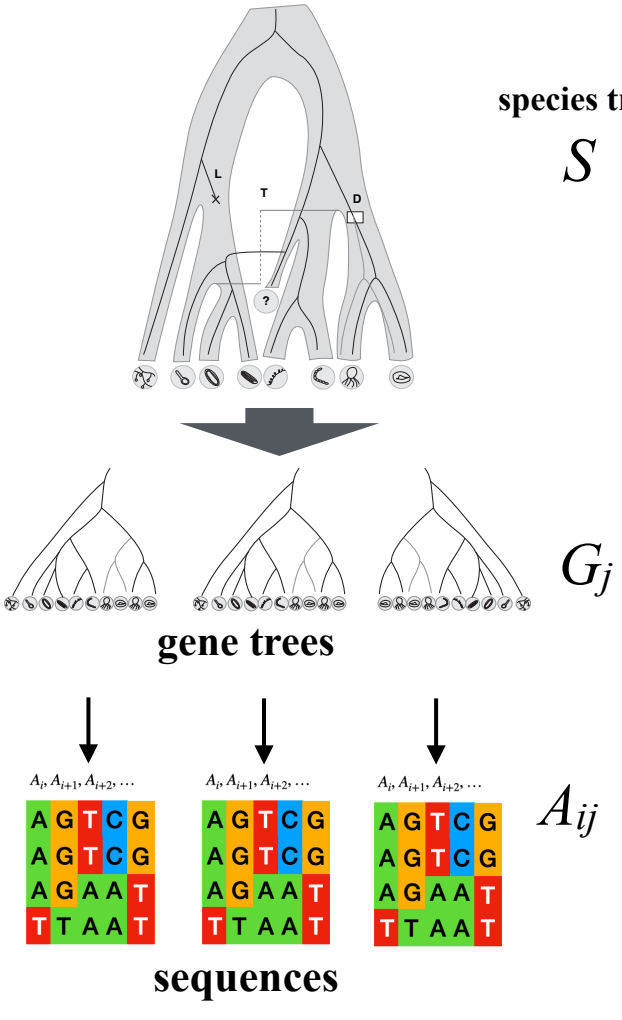
$$\prod_j$$

optimise S
and estimate rates

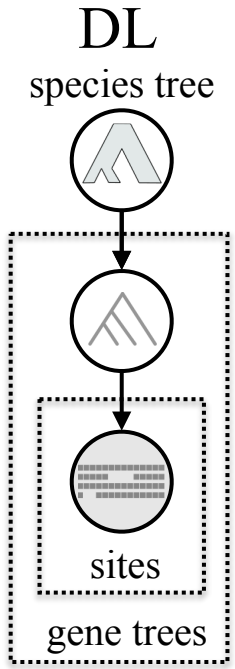
clients:
calculate

$$\prod_i p(A_{ij}|G_j) \times \; p(G_j|S, \text{rates})$$

optimise (or integrate over) G_j



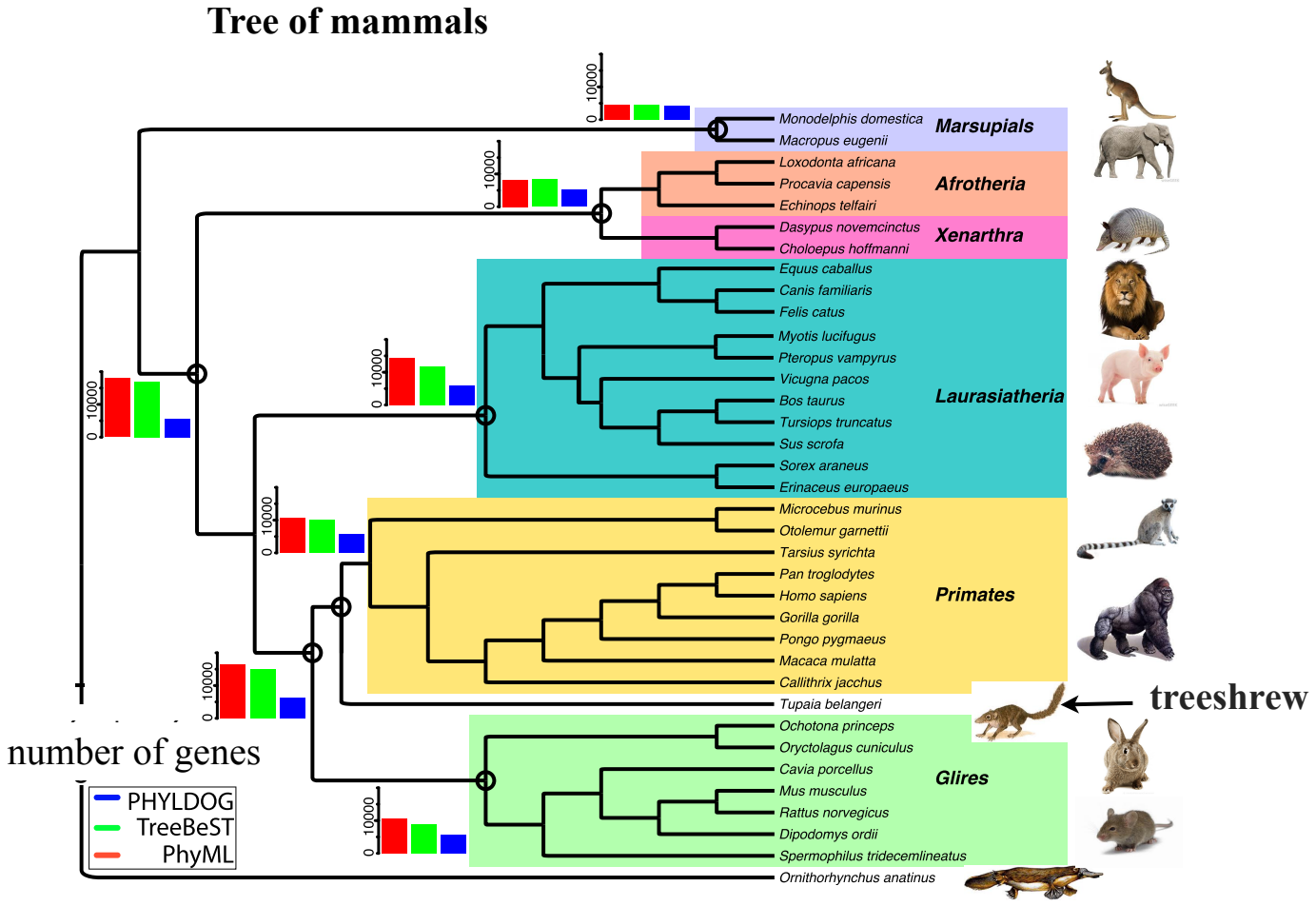
Daubin & Boussau 2011

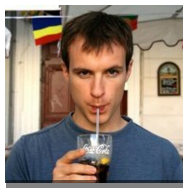


Boussau, **Szöllősi**, Duret, Gouy, Tannier & Daubin *Genome Res.* (2013)
Genome-scale coestimation of species and gene trees

Genome-scale reconstruction of the tree of mammals

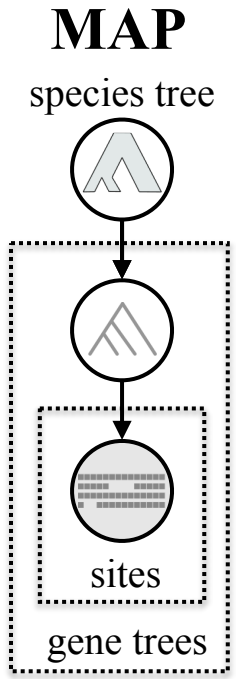
Using 6966 gene families from 36 mammals we jointly reconstructed the species tree and gene trees.





Bastien Boussau

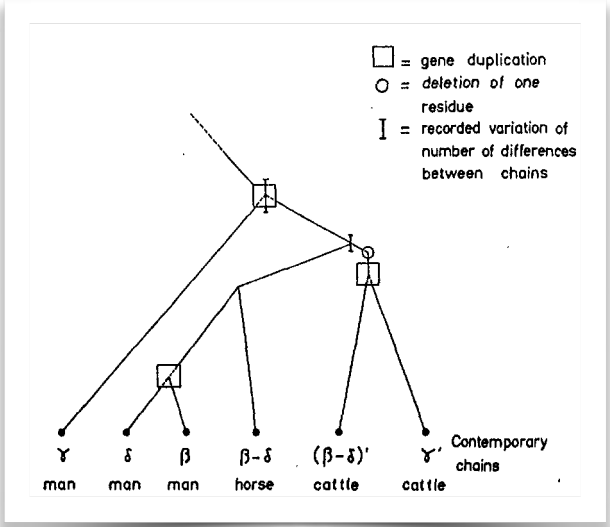
LBBE



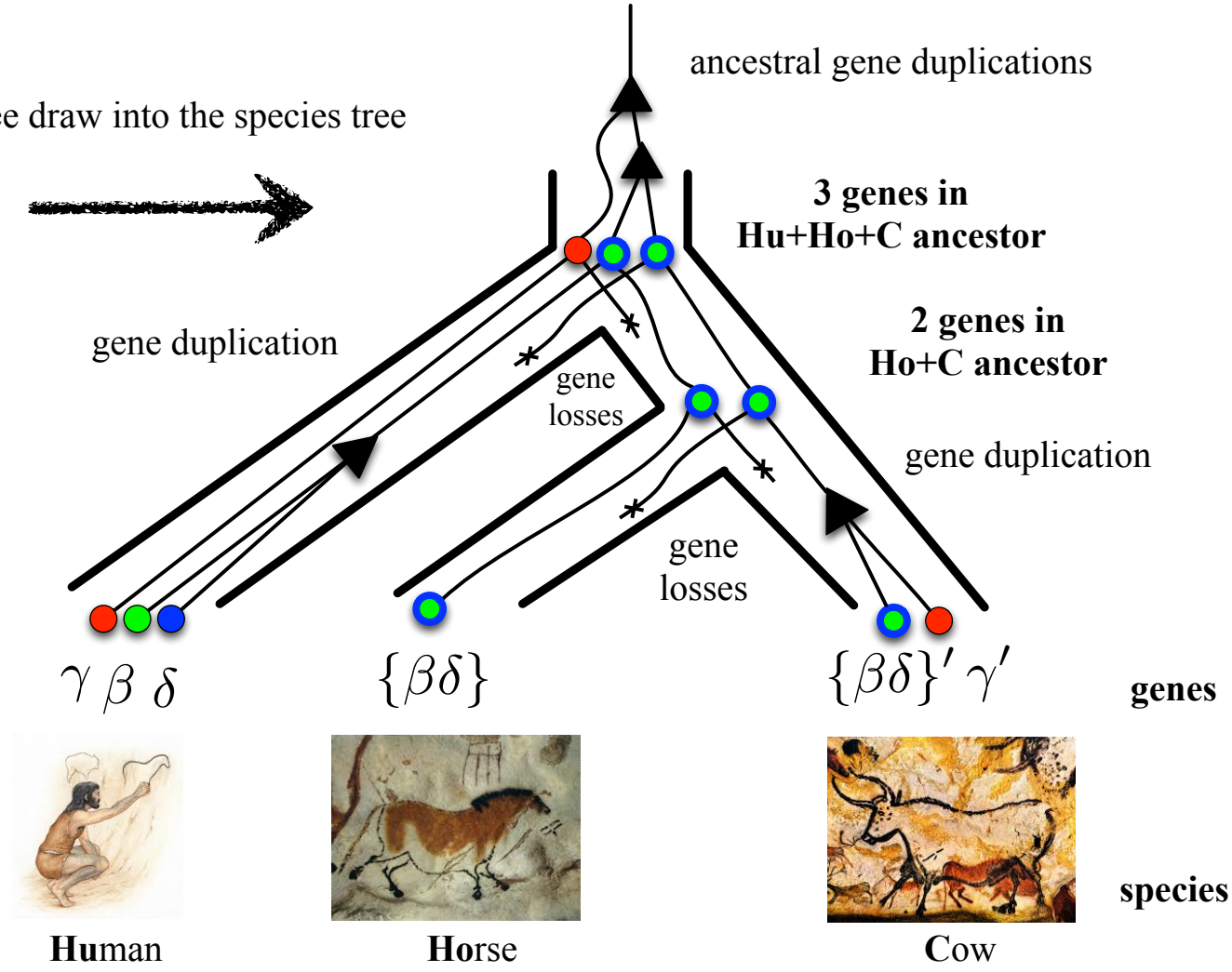
Boussau, **Szöllősi**, Duret, Gouy, Tannier & Daubin *Genome Res.* (2013)
Genome-scale coestimation of species and gene trees

The story of individual gene families is often blurred

Errors in gene trees will result in conflicts with the species tree that imply spurious evolutionary events.



Zukerkandl & Pauling 1965

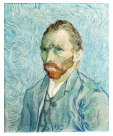


joint reconstruction with D&L

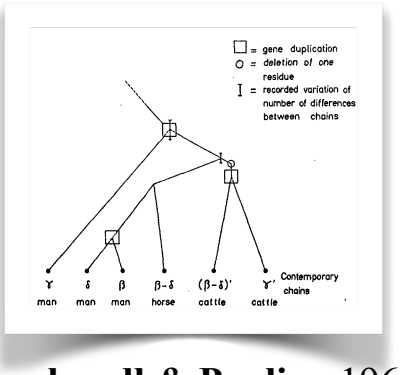
The first ever gene tree

Horse_beta-delta

Human_delta



Human_beta



Zukerkandl & Pauling 1965

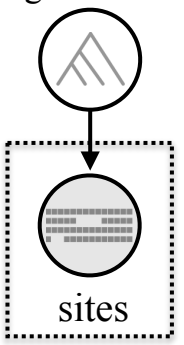
Human_gamma

Cow_gamma

Cow_beta-delta

S-unaware

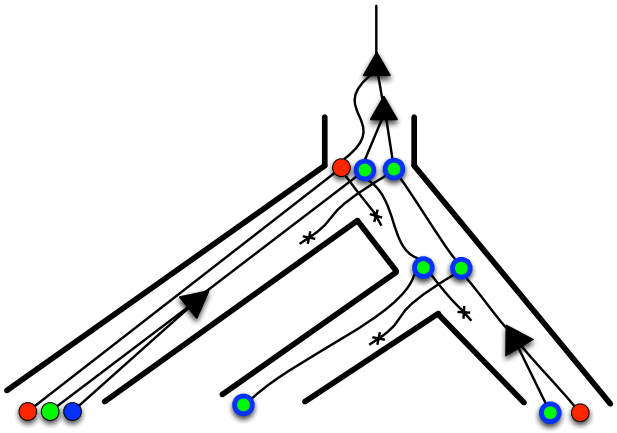
gene tree



The story of individual gene families is often blurred

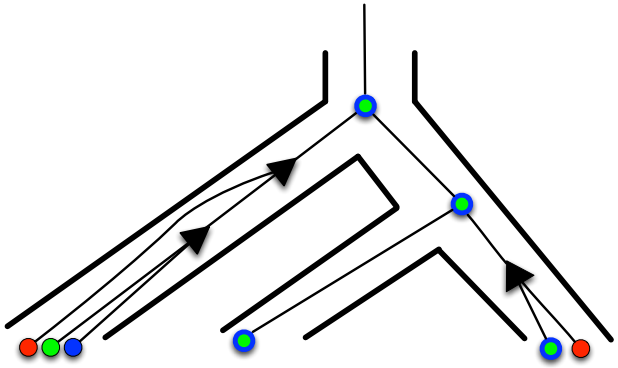
Errors in gene trees will result in conflicts with the species tree that imply spurious evolutionary events.

gene tree with errors



S-unaware trees

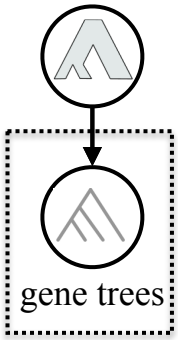
correct gene tree

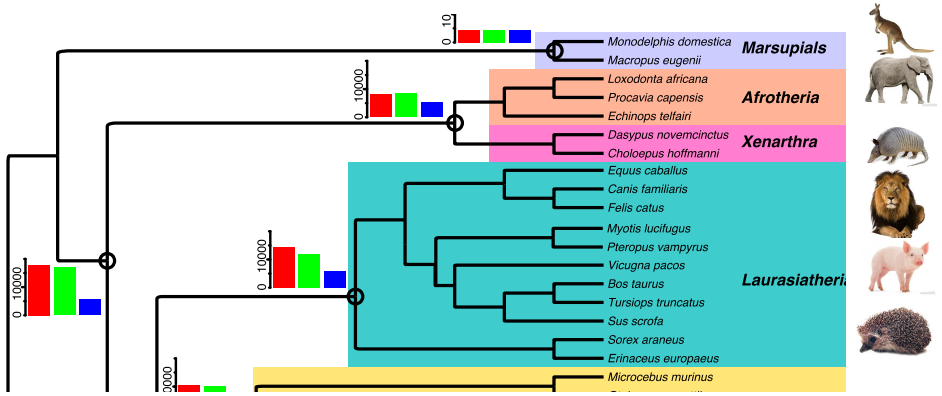


S-aware trees

DL

species tree

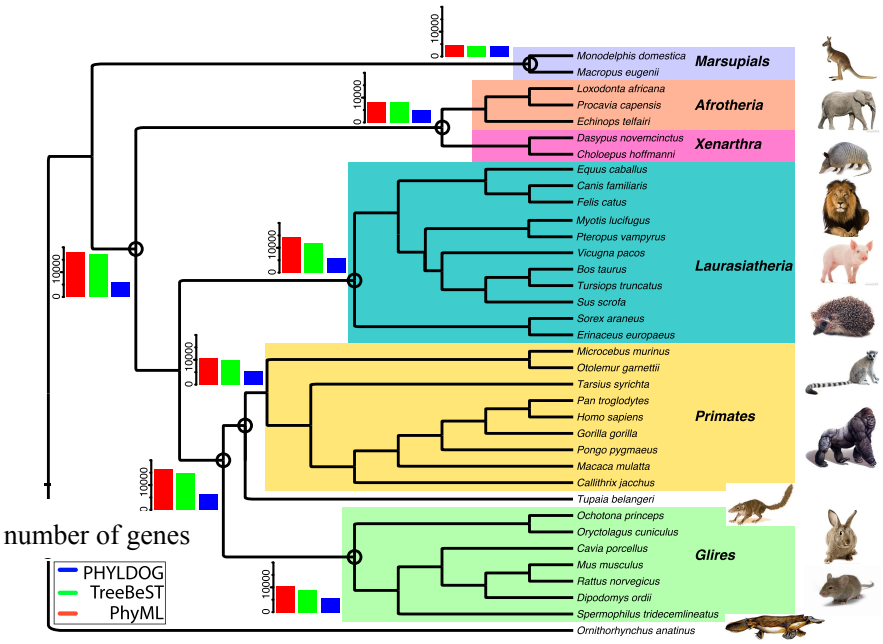




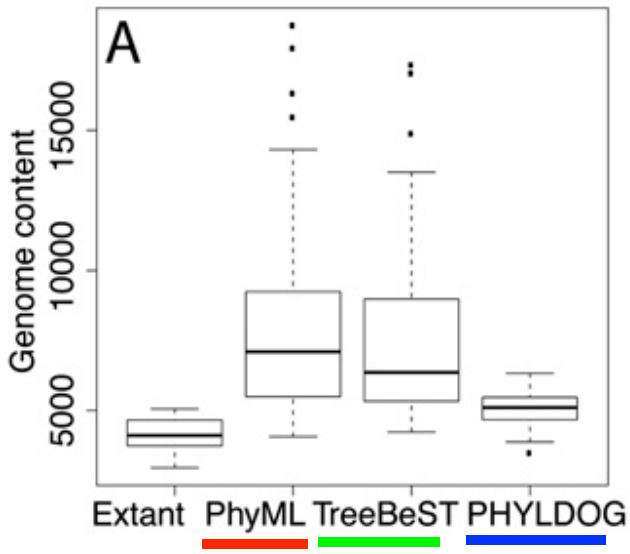
Genome-scale reconstruction of the tree of mammals

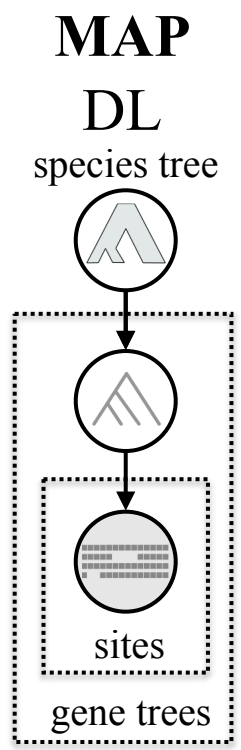
Using 6966 gene families from 36 mammals we jointly reconstructed the species tree and gene trees.

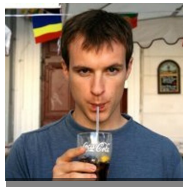
Tree of mammals



Joint reconstruction gives more realistic gene content







Bastien Boussau

LBBE

Boussau, **Szöllősi**, Duret, Gouy, Tannier & Daubin *Genome Res.* (2013)
Genome-scale coestimation of species and gene trees

Genome-scale reconstruction of the tree of mammals

Using 6966 gene families from 36 mammals we jointly reconstructed the species tree and gene trees.

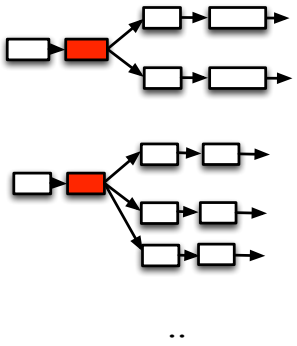
ancestral gene order can be reconstructed
using gene trees drawn into the species tree

correct gene trees

two neighbours

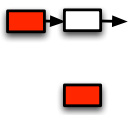
gene tree errors

three or more neighbours



..

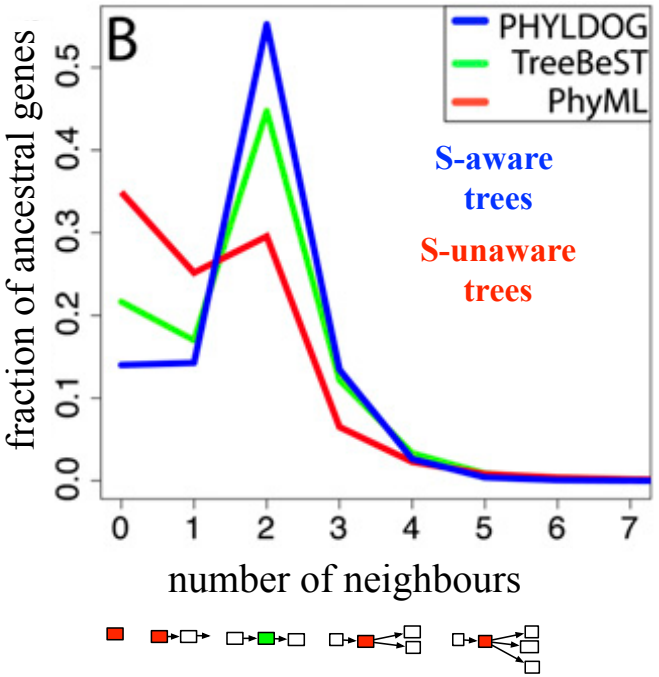
one or zero neighbours

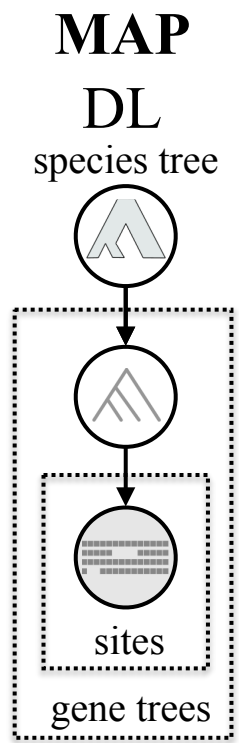


Eric Tannier

LBBE

Joint reconstruction imply more realistic ancestral gene order



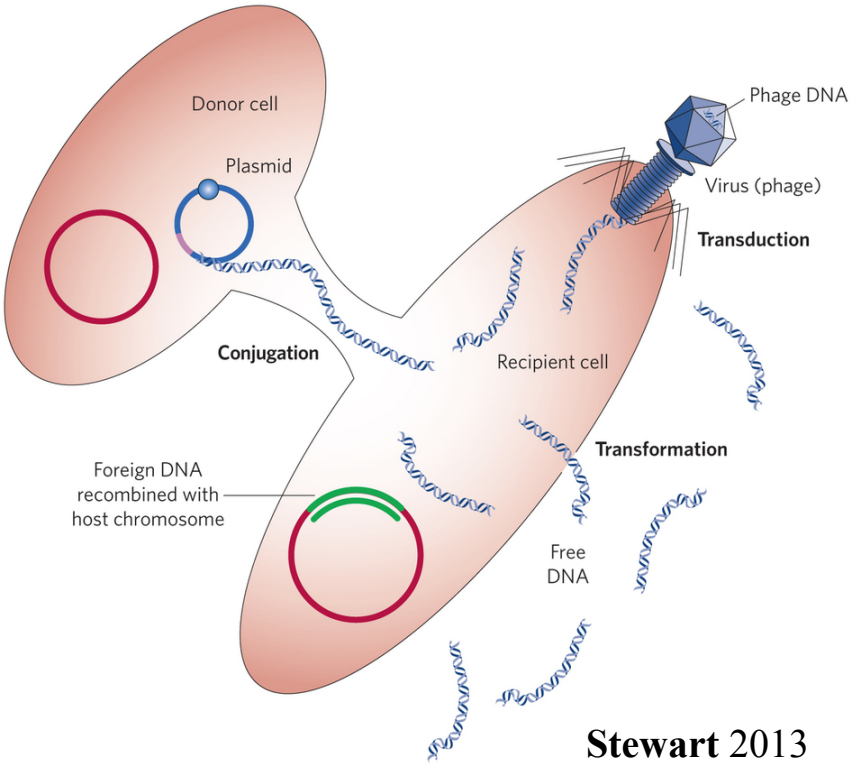


Boussau, **Szöllősi**, Duret, Gouy, Tannier & Daubin *Genome Res.* (2013)
Genome-scale coestimation of species and gene trees
Bérard, Gallien, Boussau, **Szöllősi**, Daubin, Tannier *Bioinformatics* (2013)
Evolution of gene neighborhoods within reconciled phylogenies

Horizontal gene transfer

DNA from outside the cell can be incorporated in to the genome and passed on vertically.

Horizontal Gene Transfer

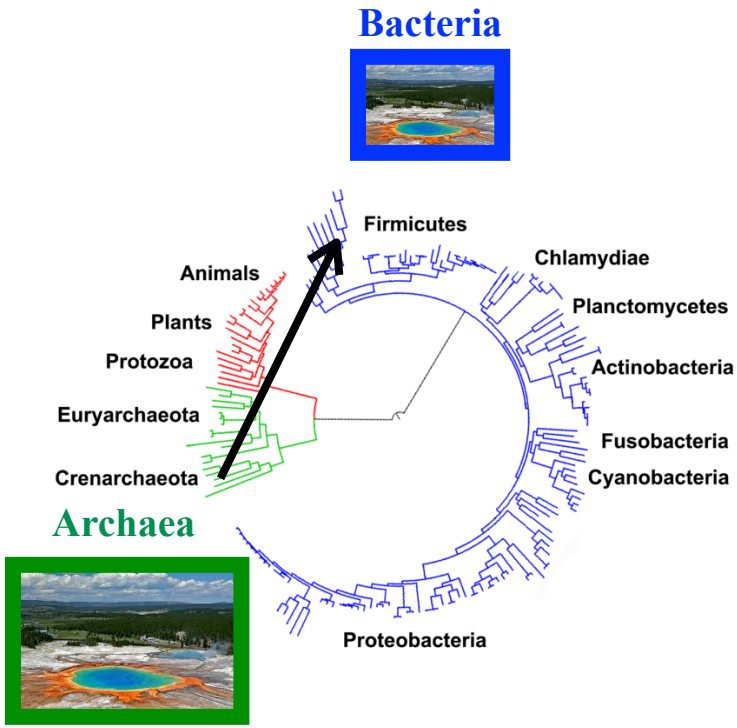


Stewart 2013

classic examples:

antibiotic resistance

thermophilic enzymes



Horizontal gene transfer

Horizontal gene transfer is common among unicellular organisms, but examples are know even among animals.

Carotenoids

plants, alga and fungi; *bacteria and archaea;*

they produce it

colour of flamingos and the human eye's macula lutea

they eat it

except!

a species of aphid living on peas

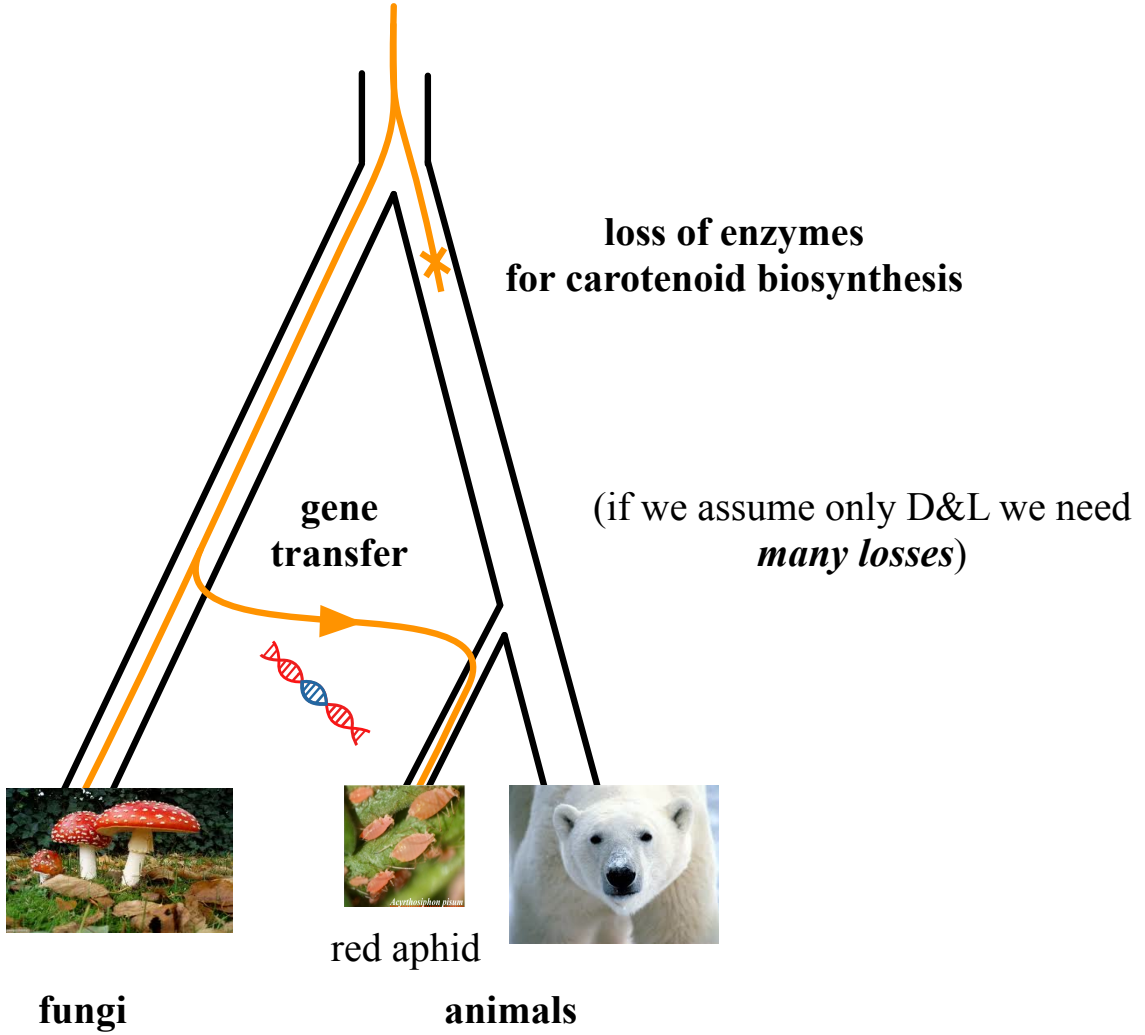
Moran & Jarvik 2010 Science

Horizontal gene transfer

Horizontal gene transfer is common among unicellular organisms, but examples are know even among animals.

pea aphids

Moran & Jarvik 2010 Science

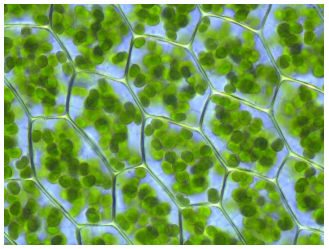


Horizontal gene transfer

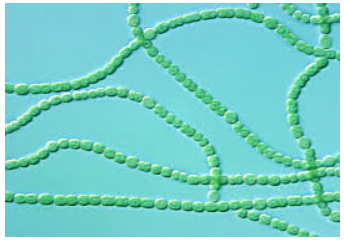
Horizontal gene transfer is common among unicellular organisms, but examples are know even among animals.

photosynthesis

chloroplasts of
algae and green plants



cyanobacteria



Eastern emerald elysia
(US East Coast)

except!

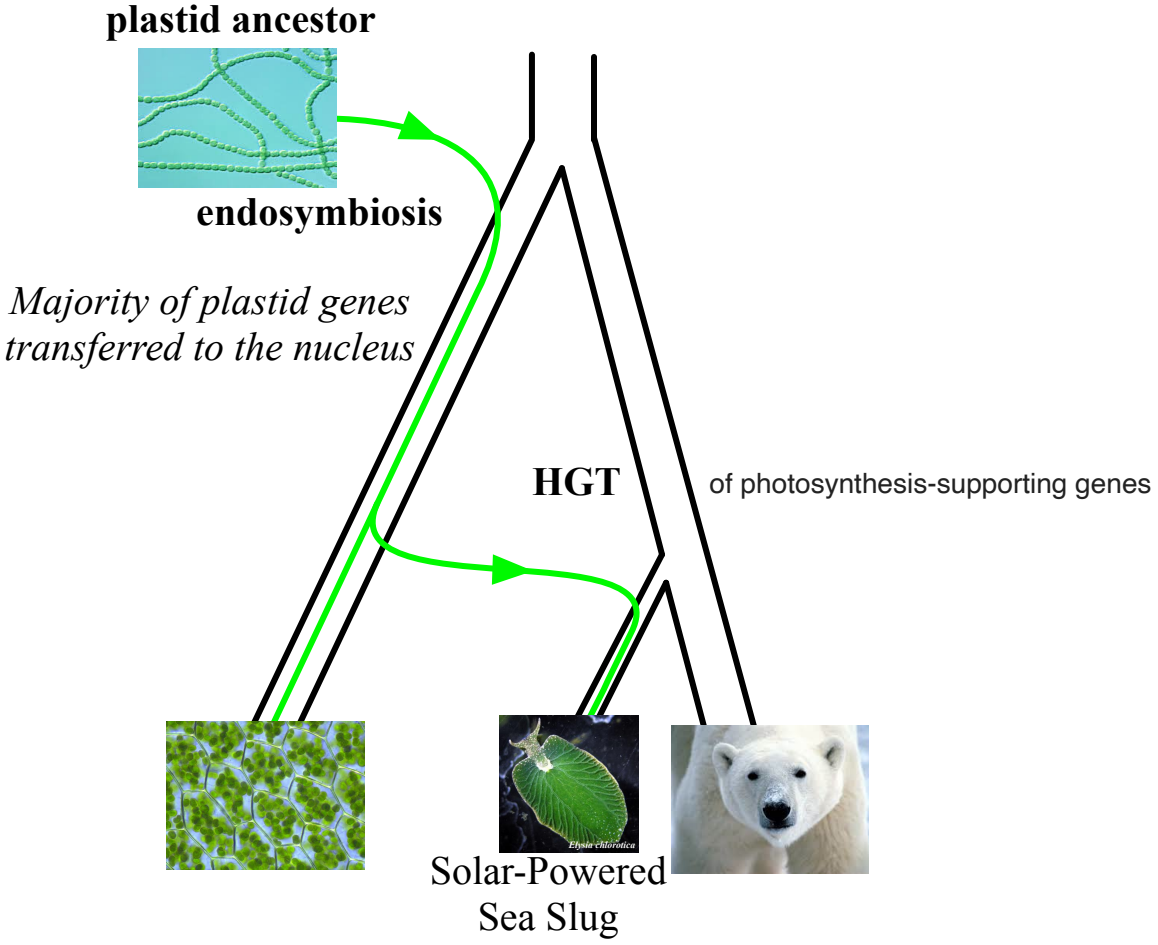
Rumpho *et al.* 2008 PNAS

Horizontal gene transfer

Horizontal gene transfer is common among unicellular organisms, but examples are know even among animals.

Eastern emerald elysia
(US East coast)

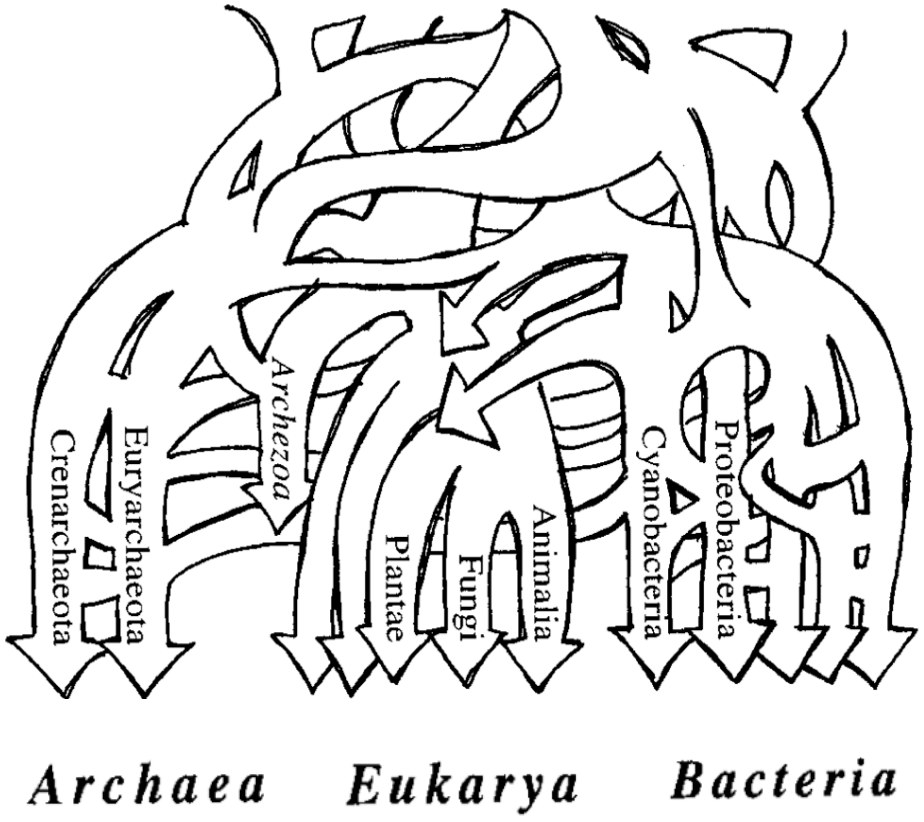
Rumpho *et al.* 2008



Horizontal gene transfer as noise

Gene transfers result in apparently contradicting gene phylogenies, fungi can seem closely related to aphids. A potentially high rate of transfer esp. early in the evolution of life, suggests that the vertical signal may be drowned in noise.

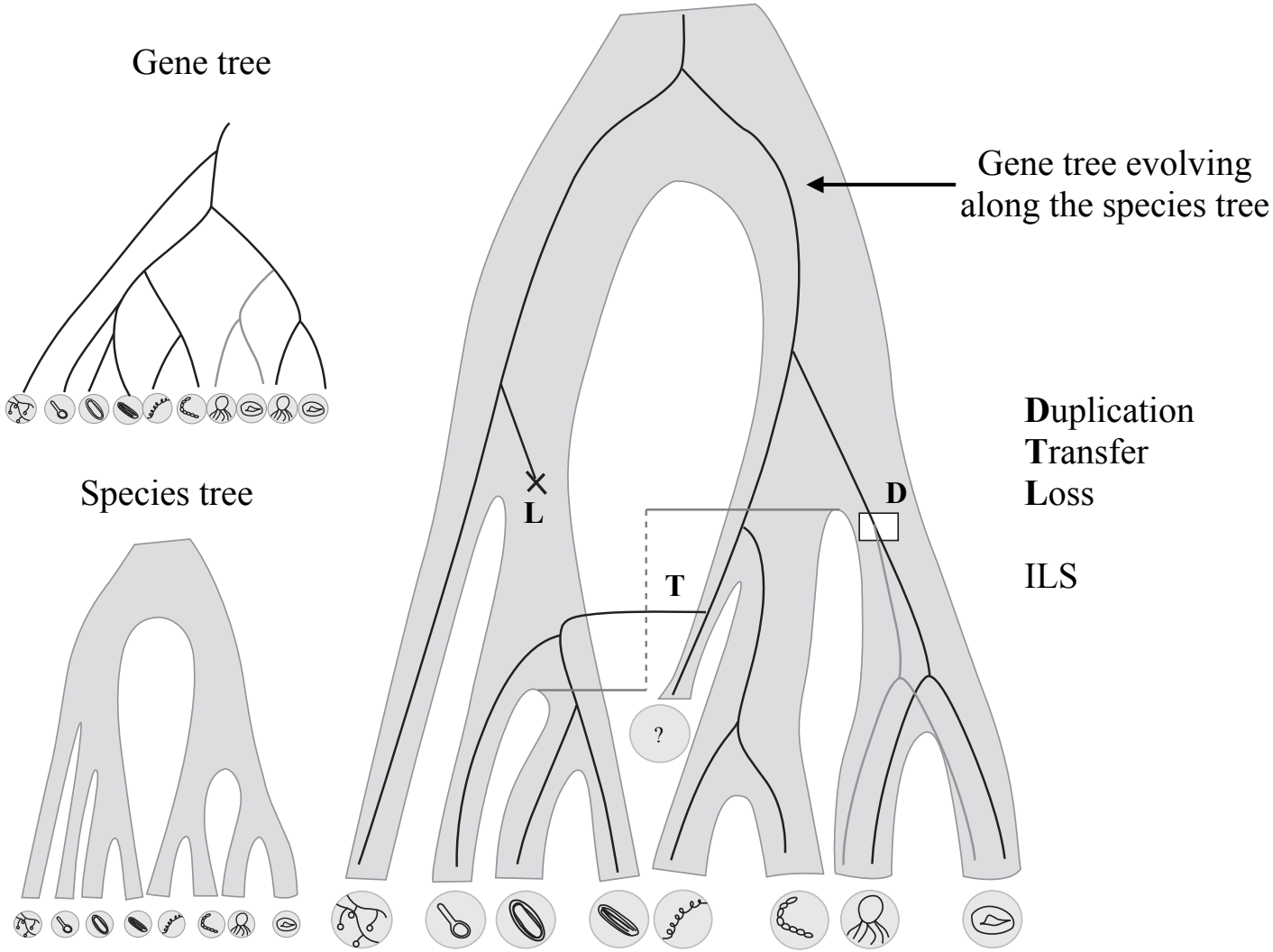
LUCA



Doolittle 1999

The problem is gene trees are not species trees

A gene tree is a deformation of the species tree through the prism of genome evolution and population genetics processes.



Daubin & Szöllősi 2016