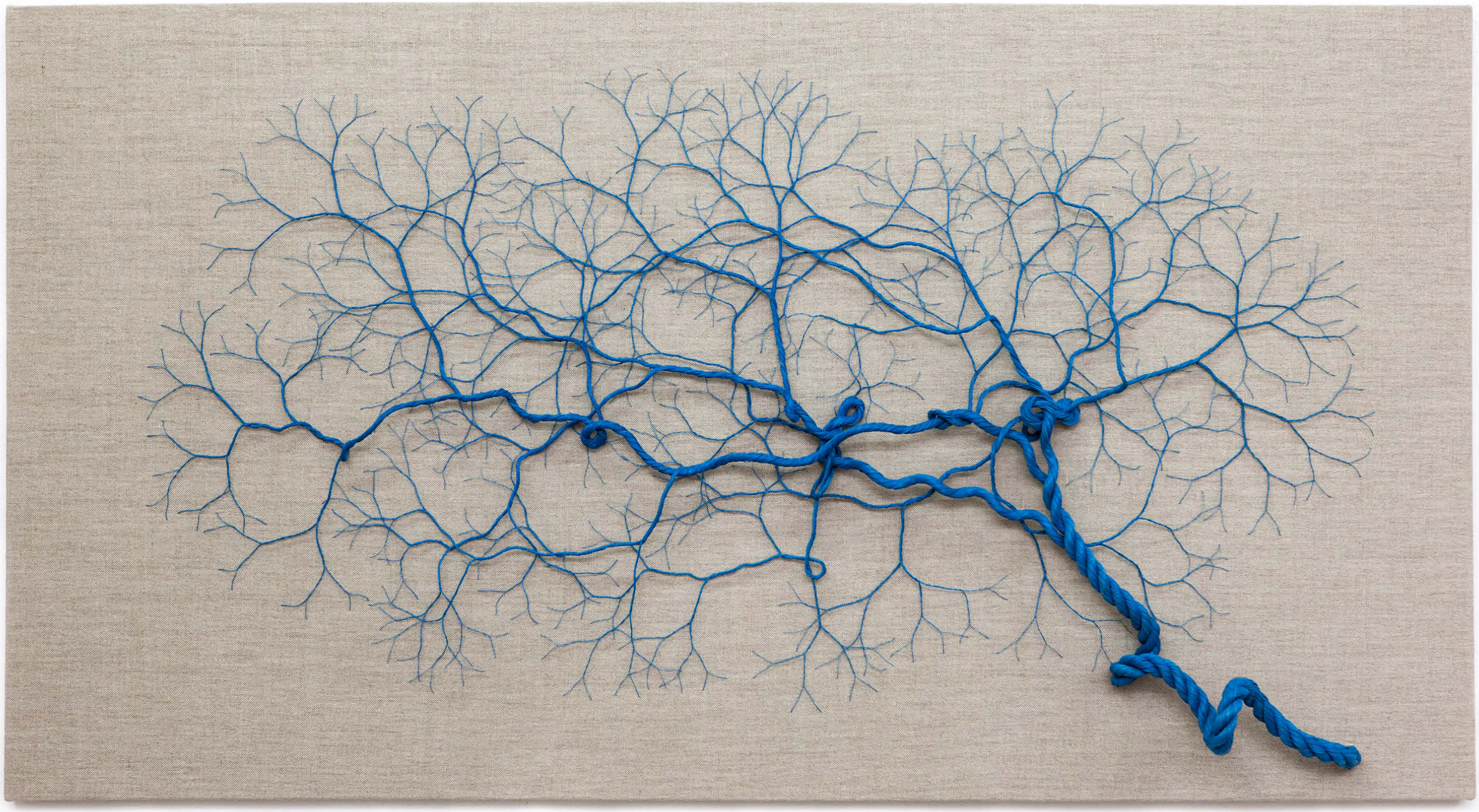
# The Inference of Gene Trees with Species Trees with an Emphasis on Horizontal Gene Transfer


Gergely Szöllősi

MTA-ELTE "Lendület"
Evolutionary Genomics Research Group
Budapest, Hungary

ssolo@elte.hu

# Biology 2.0

1996

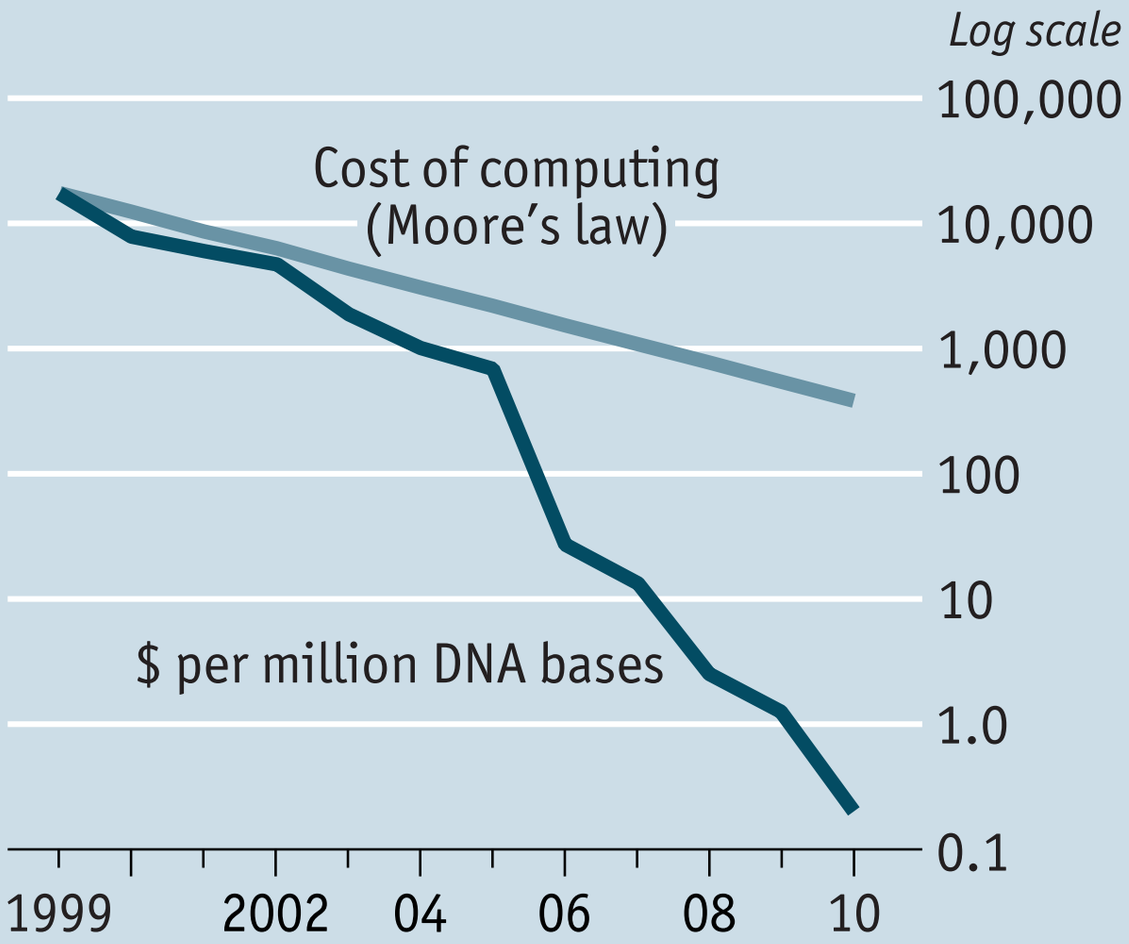
12 Giga FLOPS

(Floating Point Operations Per Second)

## Baseline information

1

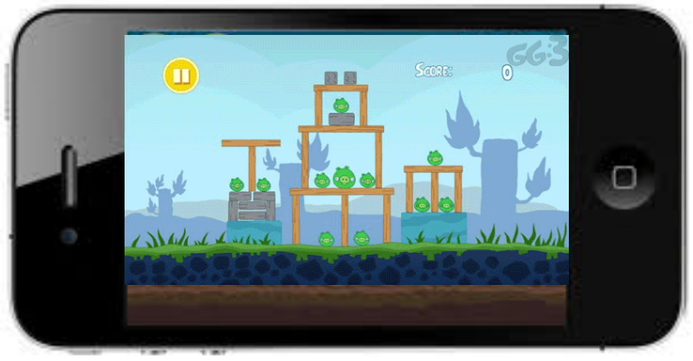
Cost of genome sequencing compared with Moore's law for computers




Source: Broad Institute

2016

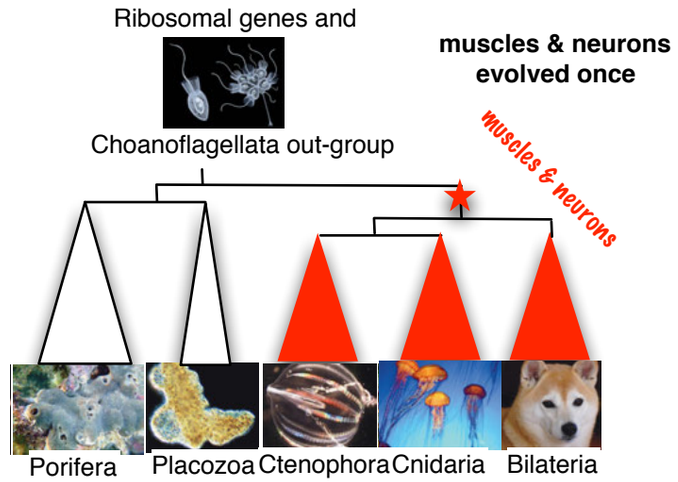
200-300 Giga FLOPS



# New genomes, old questions

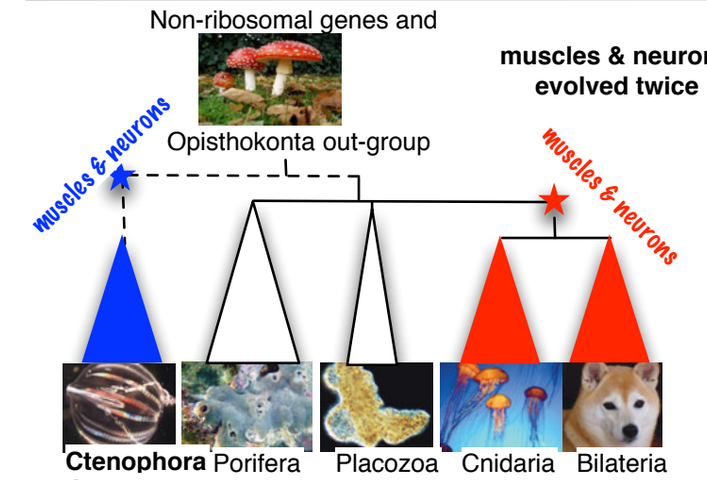
New genomes instead of bringing into sharper focus major evolutionary events such as the origin of eukaryotes or the diversification of major animal lineages have instead reignited old debates.

**Noshenko 2013, Pisani 2015, ..**

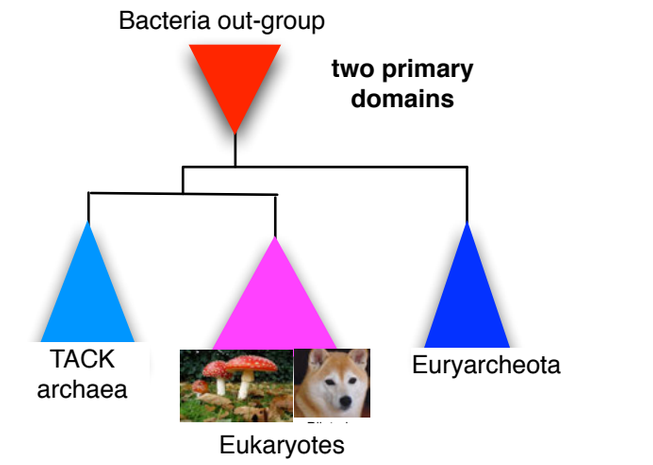



?

**Ryan 2013, Moroz 2014, ..**




?



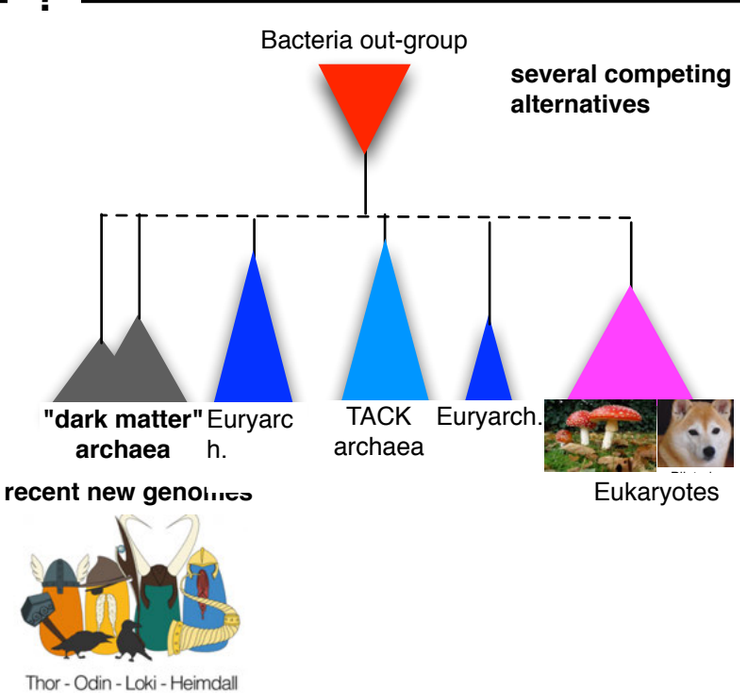

**Williams 2013**
**Williams 2017**

**Rinke 2013**

?

**Raymann 2015**

**Da Chuna 2017**




**Zaremba-Niedzwiedzka - 2017**

# The Inference of Gene Trees with Species Trees 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

just how much HGT?

HGT as information

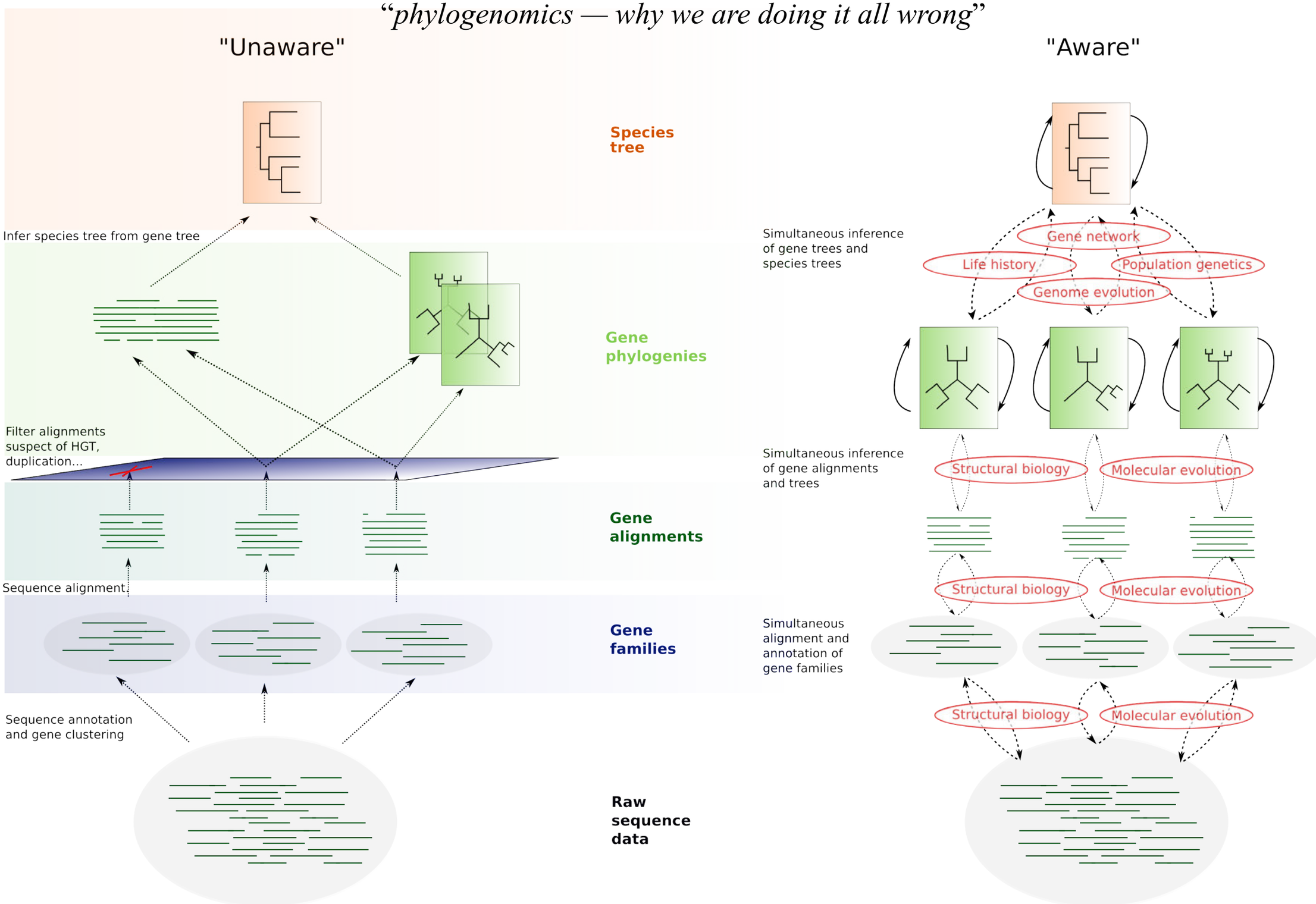
models of gene family evolution with D,T&L

HGT from the dead

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

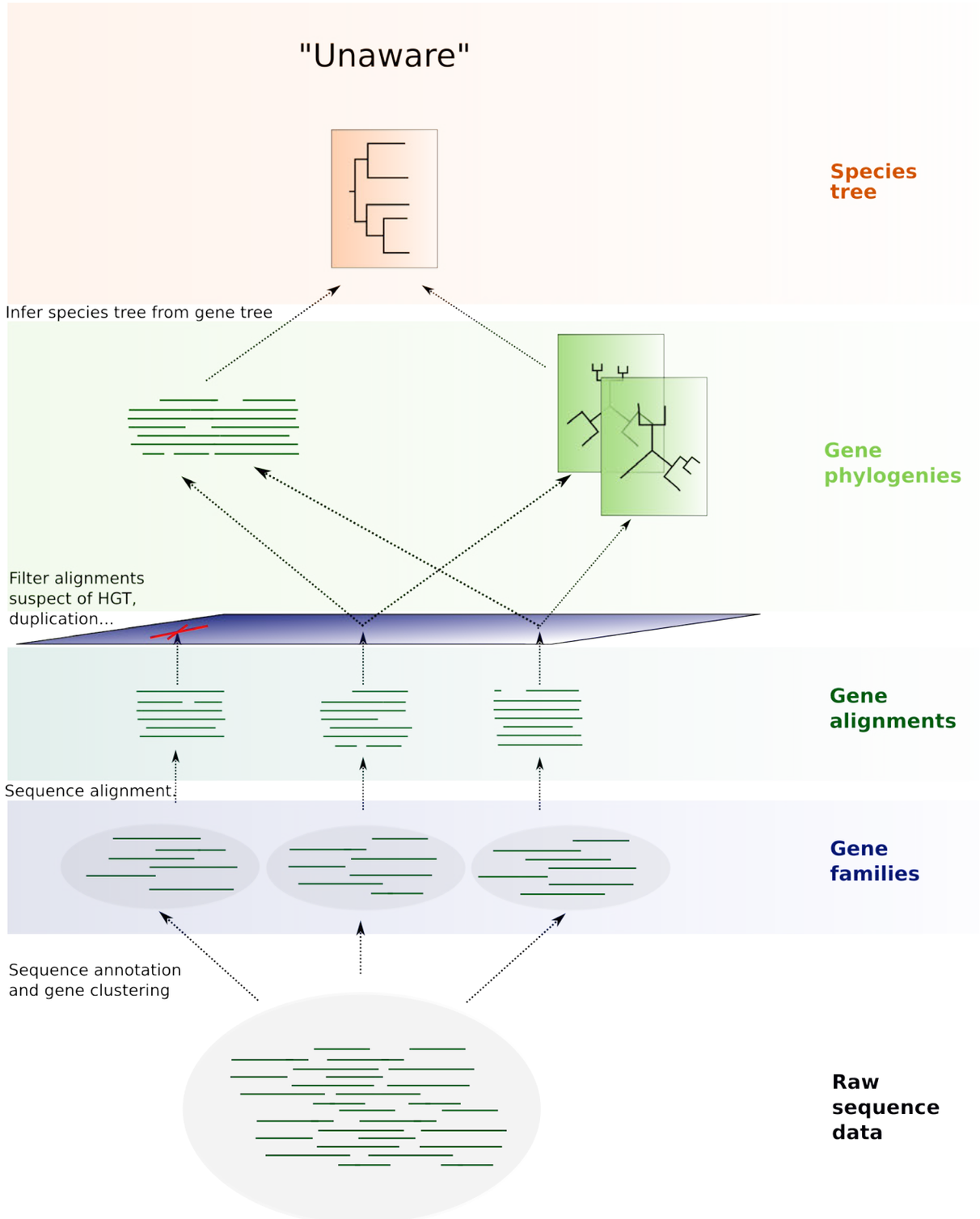
# Phylogenetic awareness

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



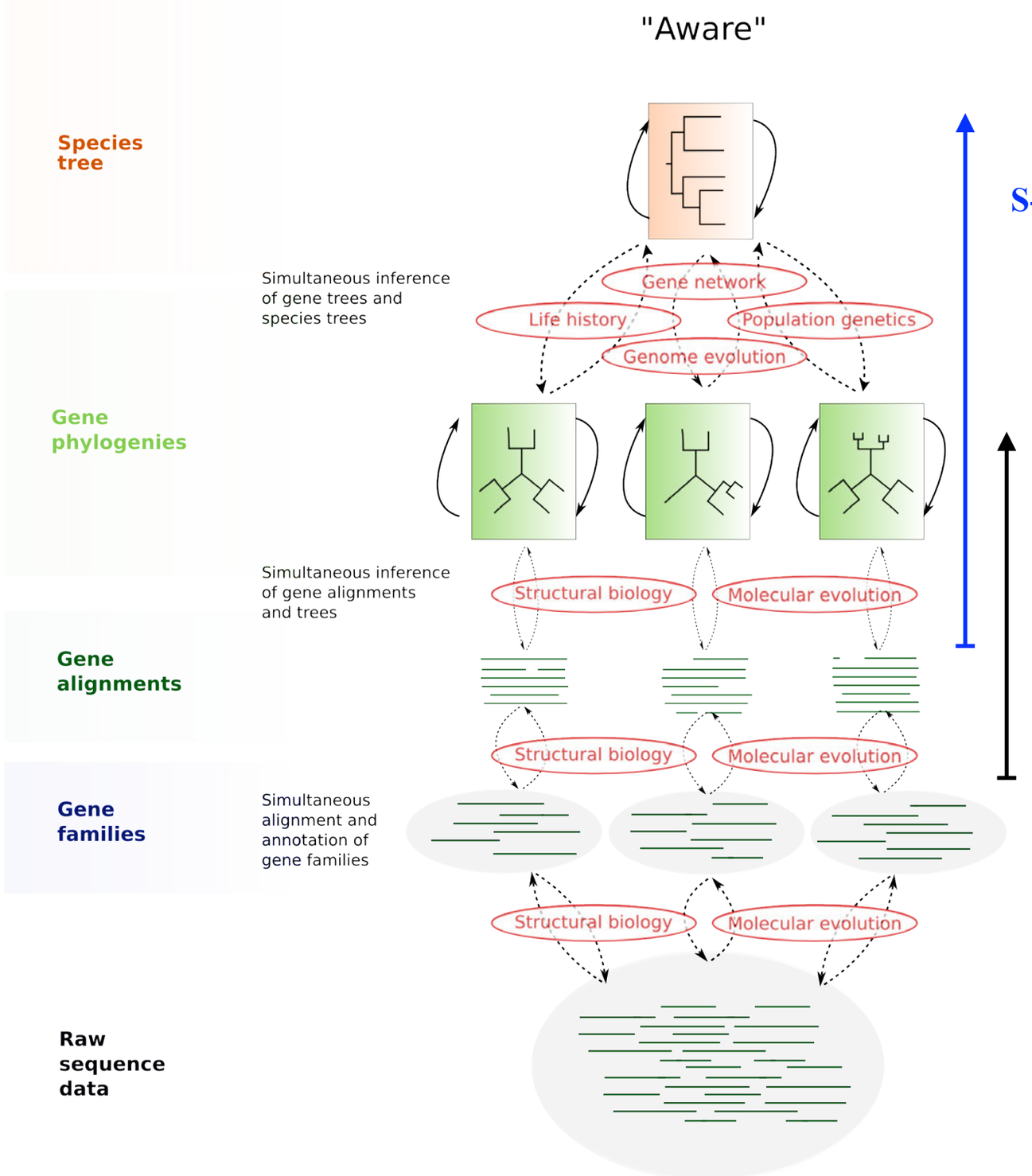

**Daubin & Boussau** 2011 TrEE

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..




**S-aware tree inference**

**this talk..**

**G-aware alignment**

e.g. PRANK

**Löytynoja & Goldman PNAS 2005**

**?**

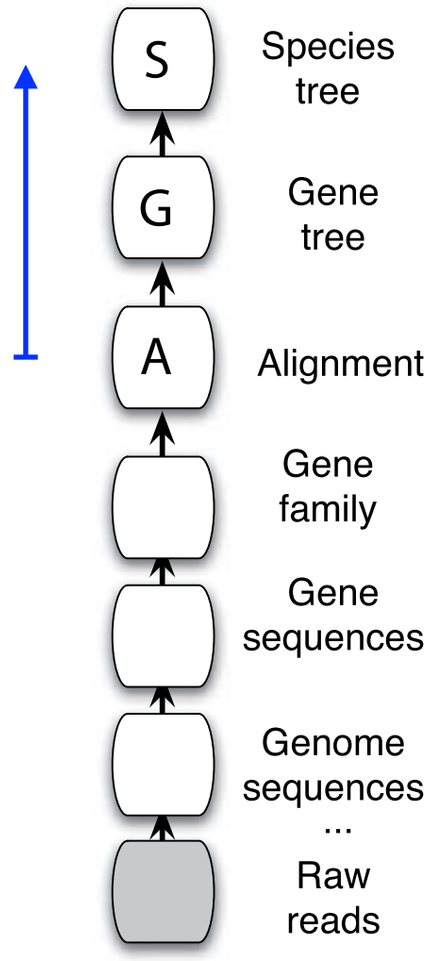
**the future of phylogenomics**

**?**

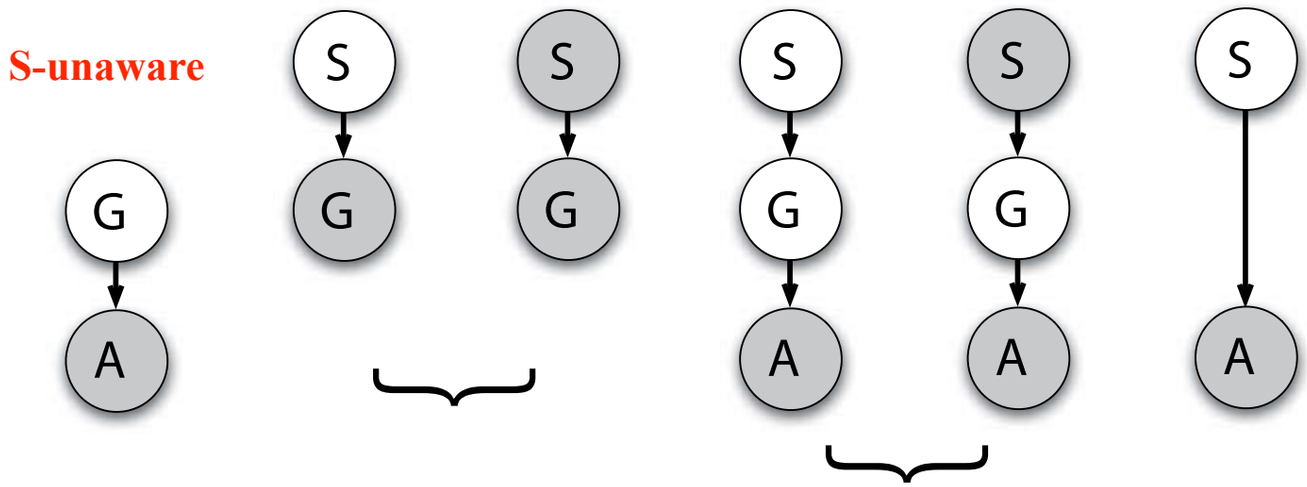
**Daubin & Boussau** 2011 TrEE

# Species tree-awareness

**Phylogenomics inference pipeline**




**Gene tree-species tree models published in the literature**




P(A|G)

PhyML [Guindon & Gascuel, 2003];
RAxML [Stamatakis et al., 2005];
MrBayes [Ronquist et al., 2012];
BEAST [Drummond et al., 2012];

P(G|S)

BEST [Edwards et al., 2007];
MP-EST [Liu et al., 2010];
ODT [Szöllősi et al., 2012];
ASTRAL [Mirarab et al. 2014]
DLCoal [Rasmussen & Kellis, 2012];

P(A|G,S)

*BEAST [Heled & Drummond, 2008];
PHYLDOG [Boussau et al., 2013];

P(A|S)

Prime-GSR [Akerborg et al., 2009];
IMa2 [Hey, 2011];
Prime-DLRS [Sjőstrand et al., 2012];
ALE [Szöllősi et al., 2013b];
SNAPP [Bryant et al., 2012];
ALE [Szöllősi et al., 2013b];

**Szöllősi,.., Boussau** 2015 Syst. Biol.

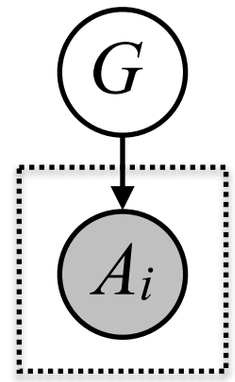
# Molecular phylogenetics infers gene trees based on sequence..

"sequence only" inference

**species tree-unaware**

gene tree

sequences




```
TATCAAGTC..
TATCAAGAC..
TATCACGAC..
TACCAGGAC..
TACCAGGTT..
TACCAGGAT..
```

# Today's sequences are a results of a series of replication events

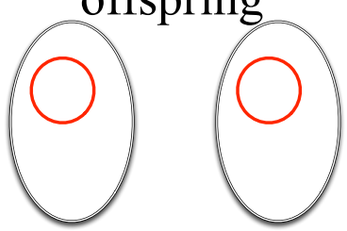
Independent of the details of reproduction the story of two homologous pieces of DNA can (locally) always be traced back to a single replication event.

**asexual reproduction**

parent

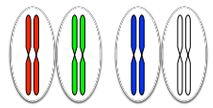
cell division

offspring



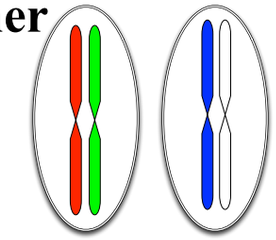
**sexual reproduction**

grandparents

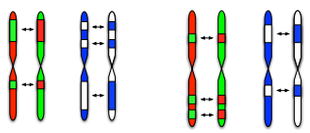


parents

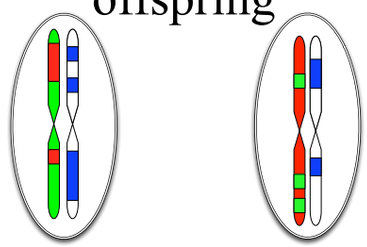
**father** **mother**



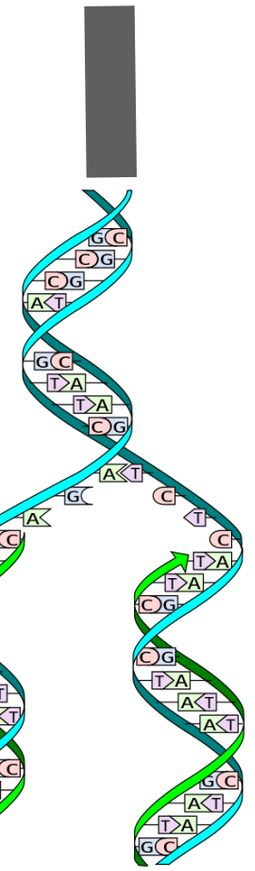
recombination

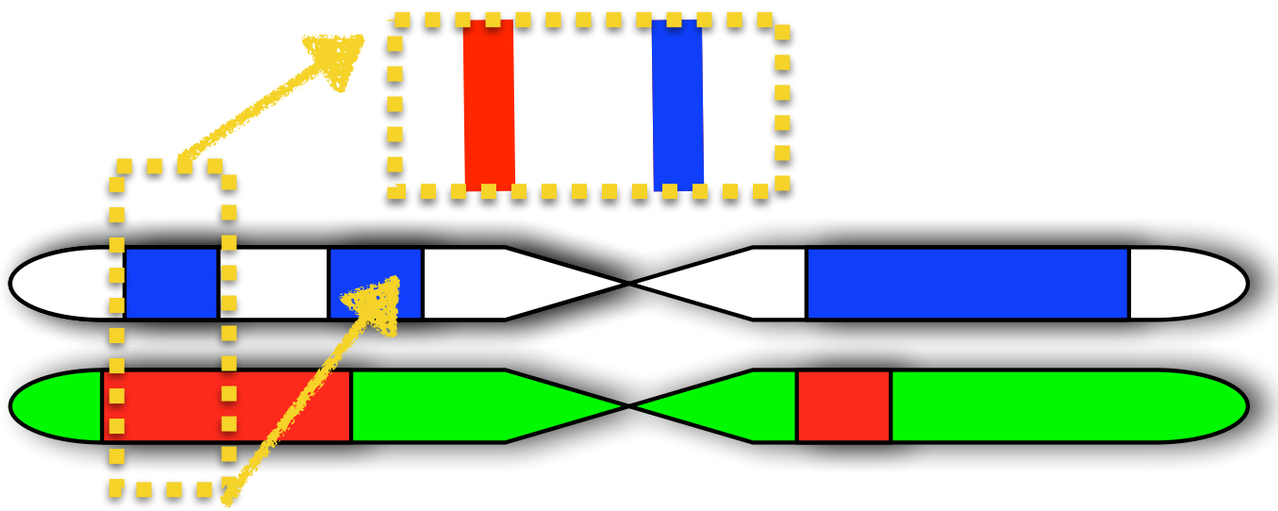


offspring



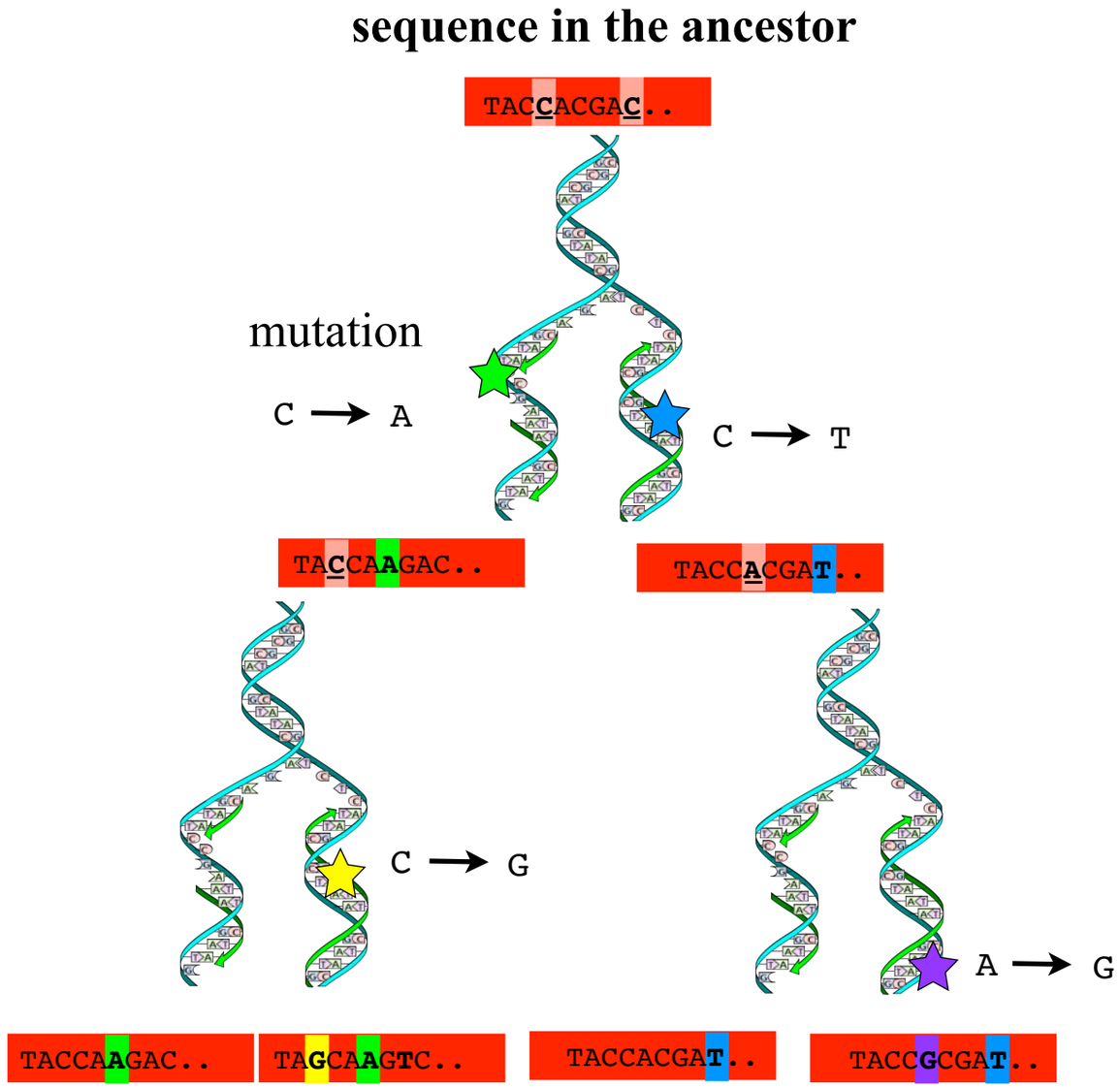
**DNA replication**





# Errors that occur during replication are inherited

Errors introduced into sequences during replication (e.g. point mutation) are inherited. The fate of each error depends on its effect on phenotype and a variety of other factors.



# Errors that occur during replication are inherited

Errors introduced into sequences during replication (e.g. point mutation) are inherited. The fate of each error depends on its effect on phenotype and a variety of other factors.

**DNA**

| | | |
|---|---|---|
| **neutral** |  | CA**C** → CA**A** |
| **?** |  | C**A**C → C**G**C |
| **deleterious** |  | TA**G** → TA**C** |

**genetic code**

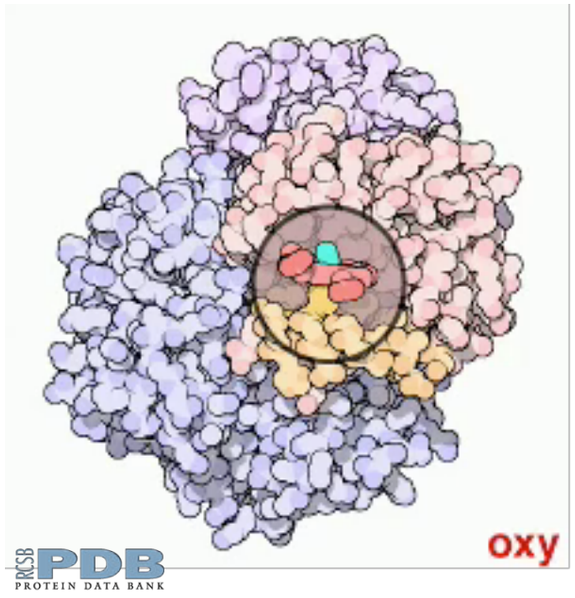
$4^3 = 64 \rightarrow 20$

codons → amino acids

**amino acids**

| | | |
|---|---|---|
| Valine | → | Valine |
| Valine | → | **Alanine** |
| Methionine **(START)** | → | **Isoleucine** |

**molecular machine**




(human hemoglobin)

empirical models of ***substitution events***

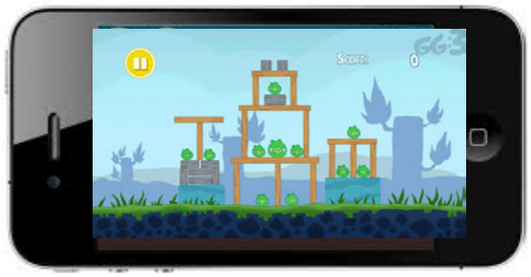
TN92, TN93, F81, HKY85, GTR, TKF91, TKF92, WAG, BLOSUM, PAM, JTT92, LG08, REV, MTREV, GY94, MG95, NY98, M0, M1, . . . M13, CAT (and CAT again), MKv, Dayhoff, JC69, K2P, K3P, ECM, DEC, BM, OU, EB, CATBP, GG98, TS98, G01, UCLN, UCG, RLC, ACLN, CIR, and WN

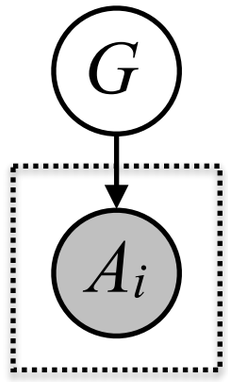
**CAT+GTR**

**Lartillot & Philippe 2004**

**LG**

**Le & Gascuel 2008**





# The story of homologs genes can be reconstructed

Given a model of sequence evolution the gene tree can be reconstructed, because it induces a probability distribution over site patterns (some site patterns are more likely than others given a particular gene tree), and in return, the site patterns we observe inform us about the gene tree along which they were generated.

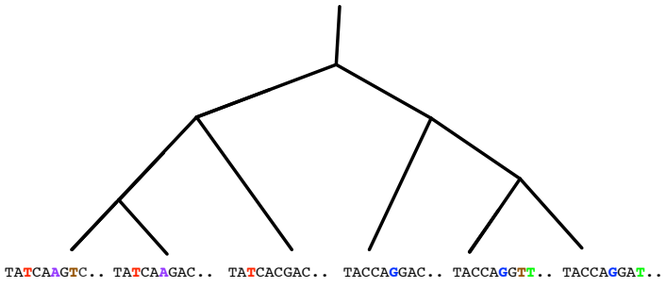
Likelihood of the sequences $A$ given the gene tree $G$:

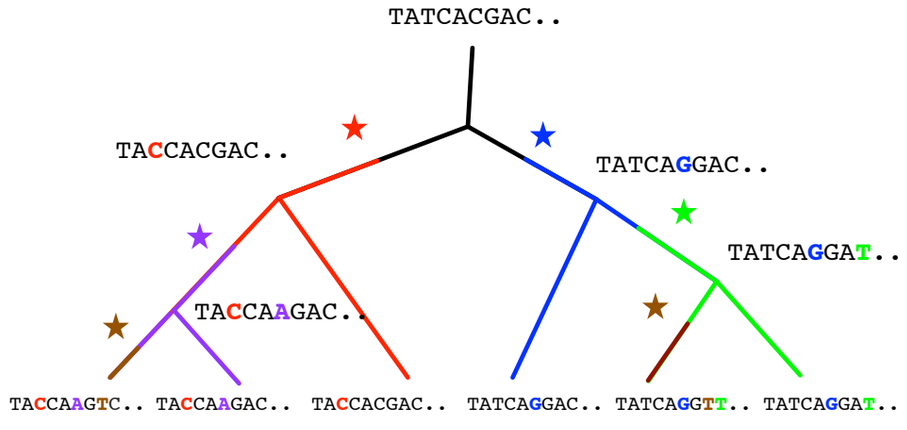
$$p(A|G)$$

**Felsenstein** 1981

$G$ — *substitution events* → $A$

**gene tree**

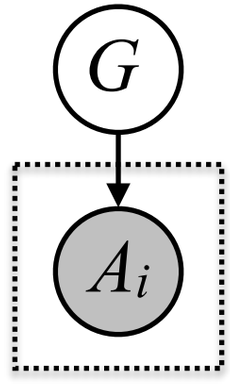



**sequences**
with many indep. sites

TATCAAGTC..
TATCAAGAC..
TATCACGAC..
TACCAGGAC..
TACCAGGTT..
TACCAGGAT..

$i = 1 \ldots L$



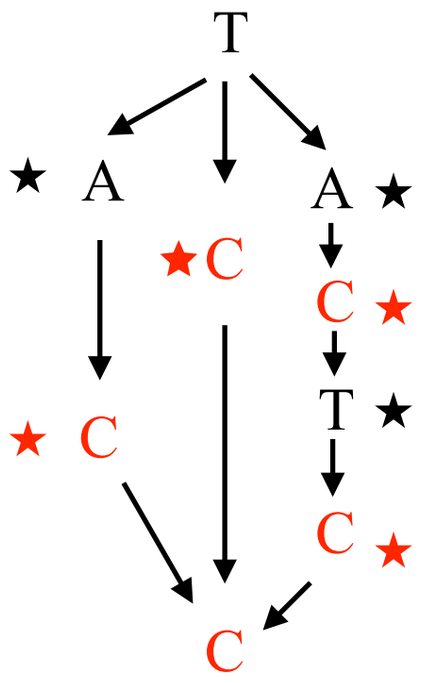
# The story of homologs genes can be reconstructed

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

$$p(A|G)$$

*requires summing over all possible substitution paths.*

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

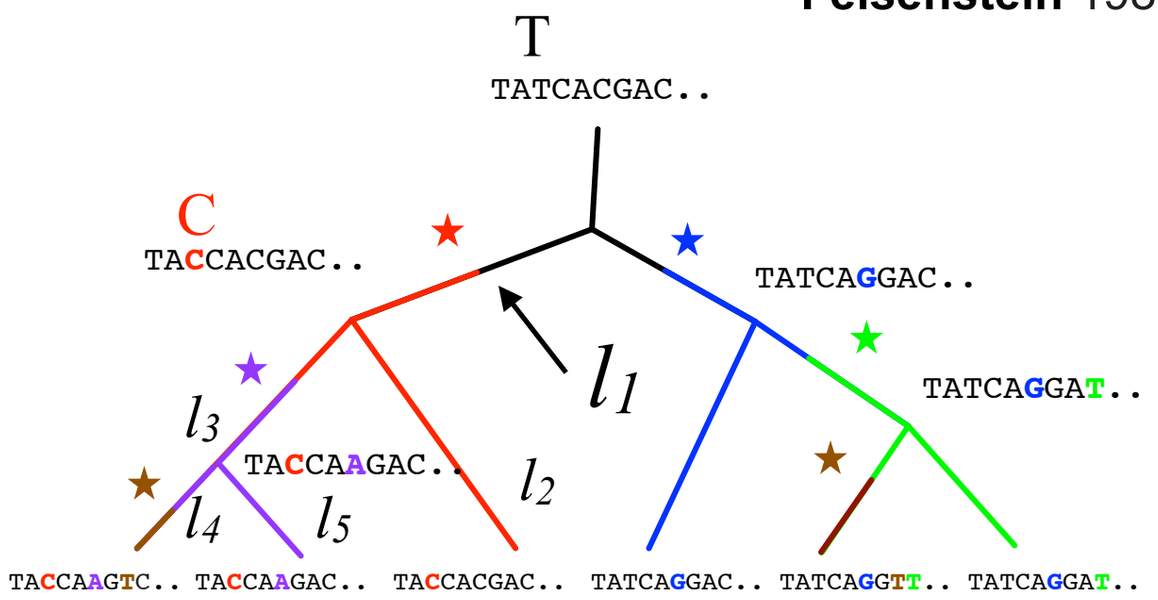


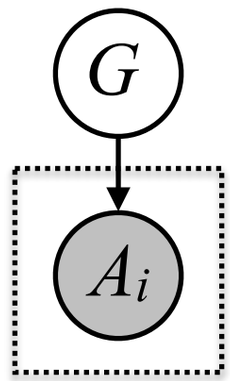

$P_{TC}(l_1)$

( for $l$=0.2 subs./site
1.6% will have 2 subs. )

**sum over ancestral states**

**Felsenstein** 1981



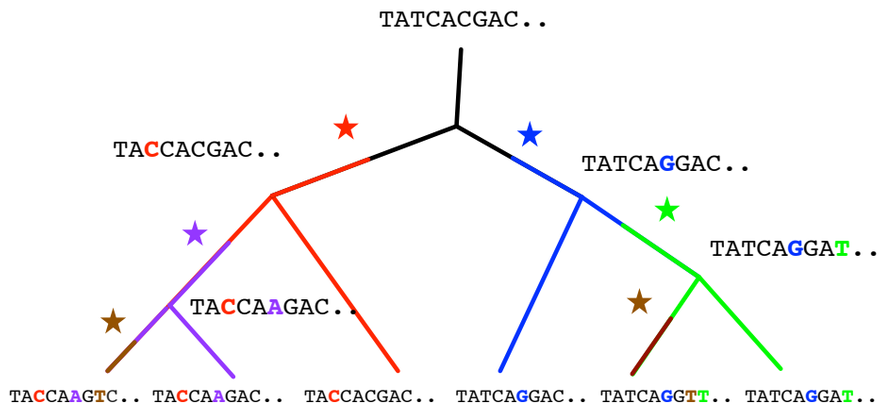

$$= .. \times P_{TC}(l_1) \times P_{CC}(l_2) \times P_{CC}(l_3) \times P_{CC}(l_4) \times P_{CC}(l_4)$$

# The story of individual gene families is often blurred

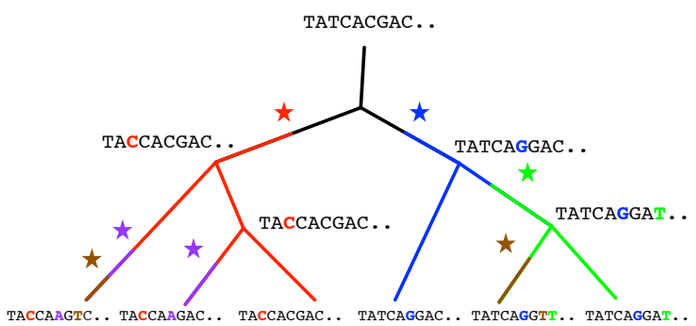
Individual genes alone contain limited signal, and as a result phylogenetic reconstruction almost always involves choosing between statistically equivalent or weakly distinguishable relationships.

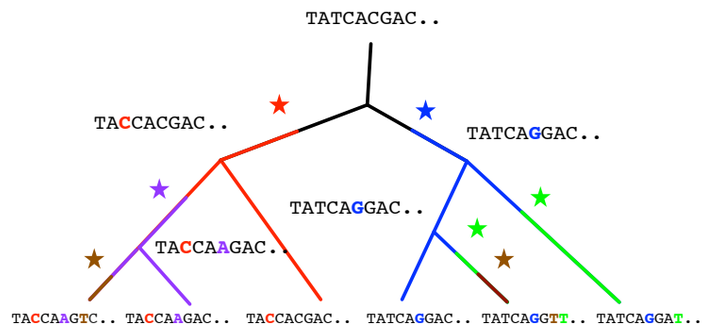
In the vicinity of the $G$-s that optimises the likelihood

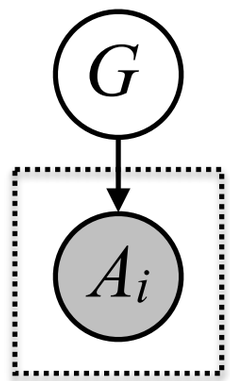
$$p(A|G)$$



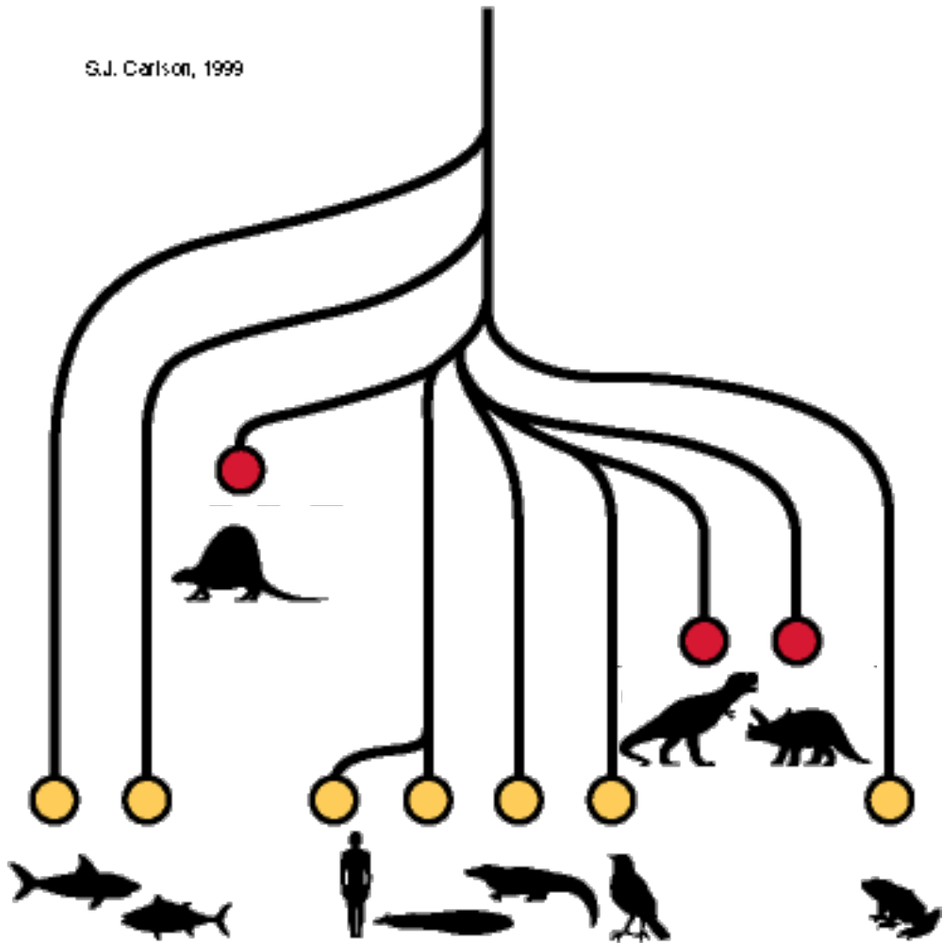

there are often many statistically **equivalent gene trees:**



# Anyway, we really care about the species tree..

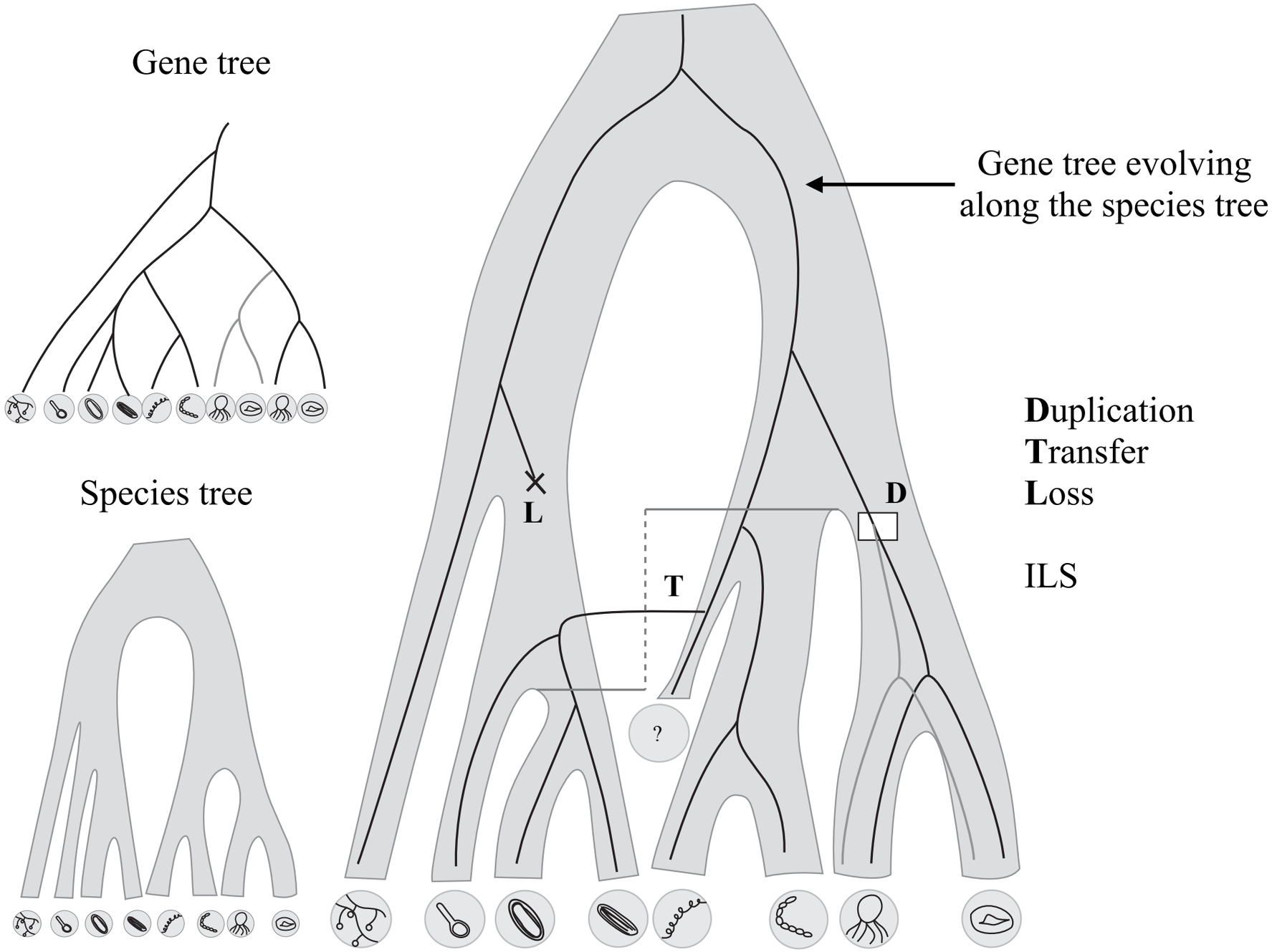



*S*

**today's vertebrates**

# 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.*

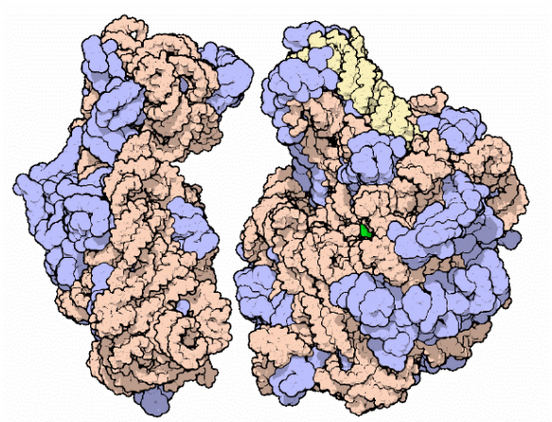



**Daubin & Szöllősi** 2016

# Phylogenomics — the history of the 1%

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

e.g.

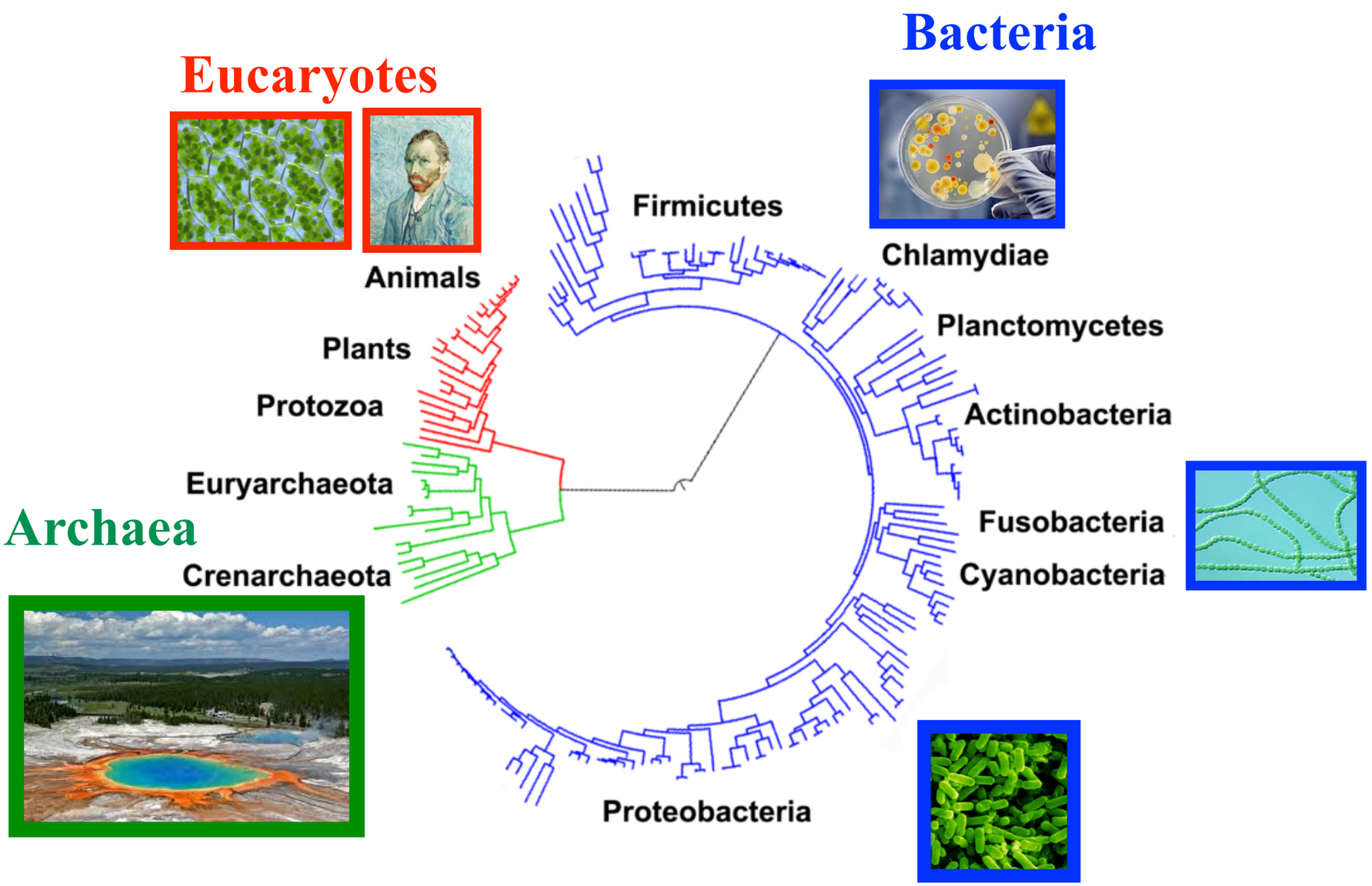


16S rRNA

**Carl Woese**, 1977

at most a few dozen

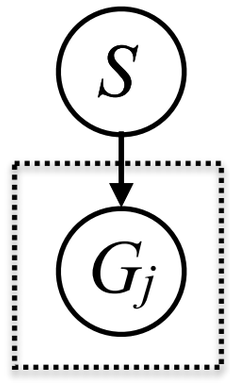
**Ciccarelli**, 2006



# .. 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.* 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.

**Daubin & Boussau** 2011 Trends Ecol. Evol.

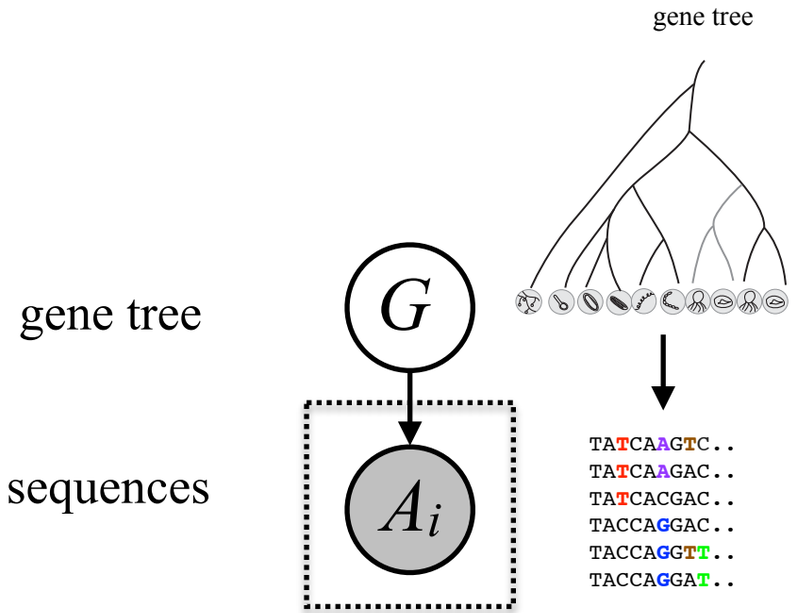


# The solution is to model how gene trees are generated along the species tree

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**

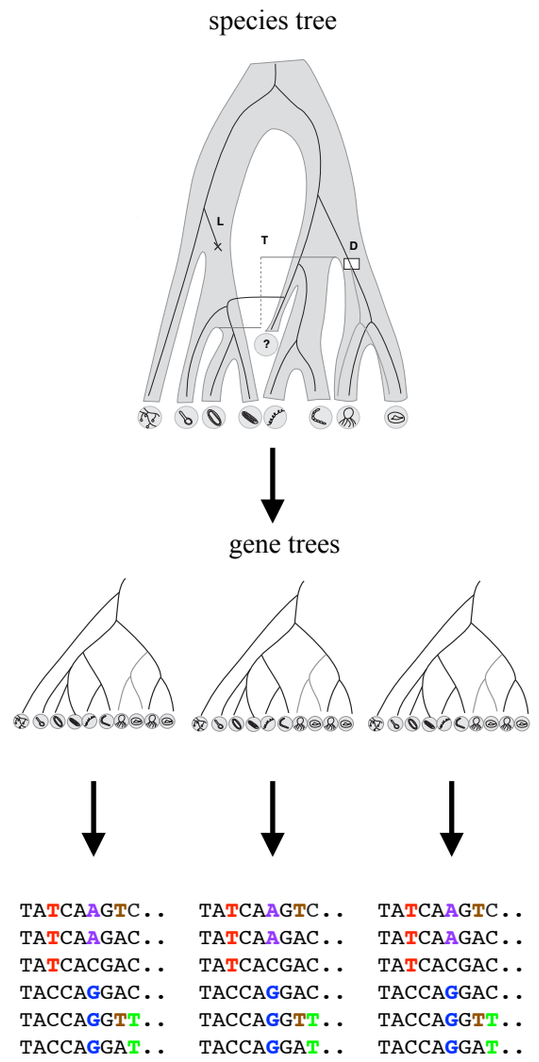
**species tree-unaware**




**“joint” inference**

**species tree-aware**

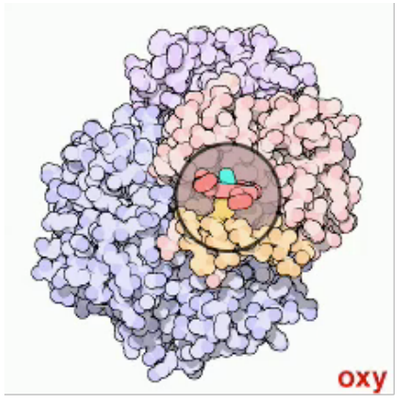


# The stories gene families can be complicated

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.

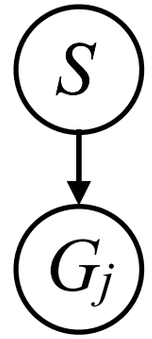
**Human hemolglobin is composed of**




molecular machine

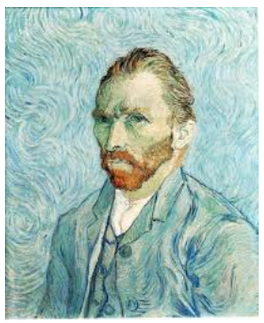
$2\alpha + 2\beta$ **chains.**

DL




$2\alpha+$

**Human**



adult

$+2\beta$ (97%)

$+2\delta$ (3%)

fetus

$+2\gamma$

**Cow**

adult

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

fetus

$+2\gamma$

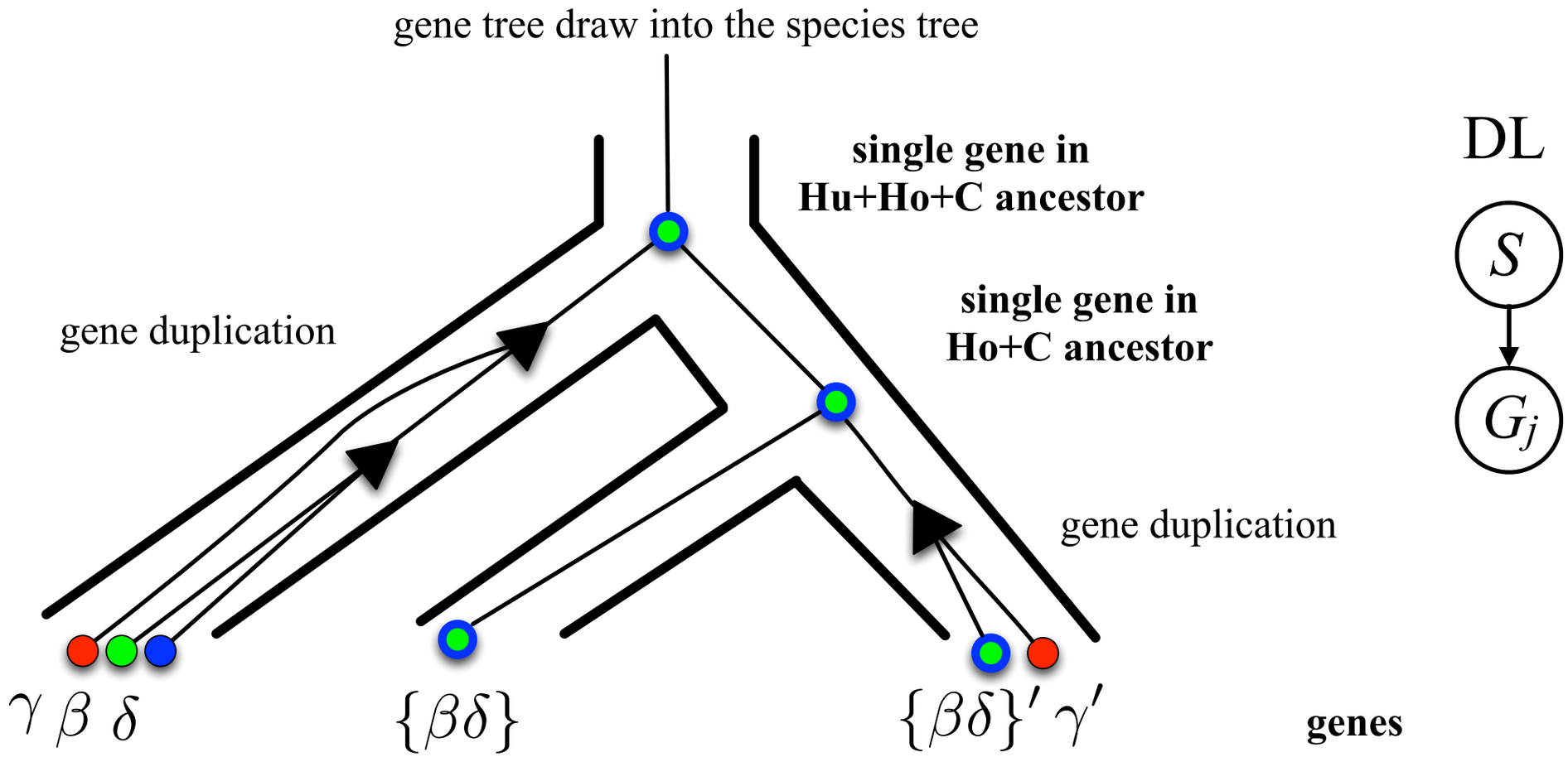
**Horse**

adult and fetus

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

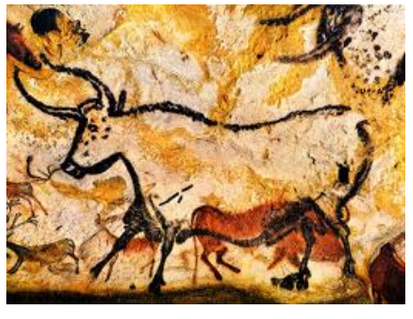
# The stories gene families can be complicated

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.




**Hu**man

**Ho**rse

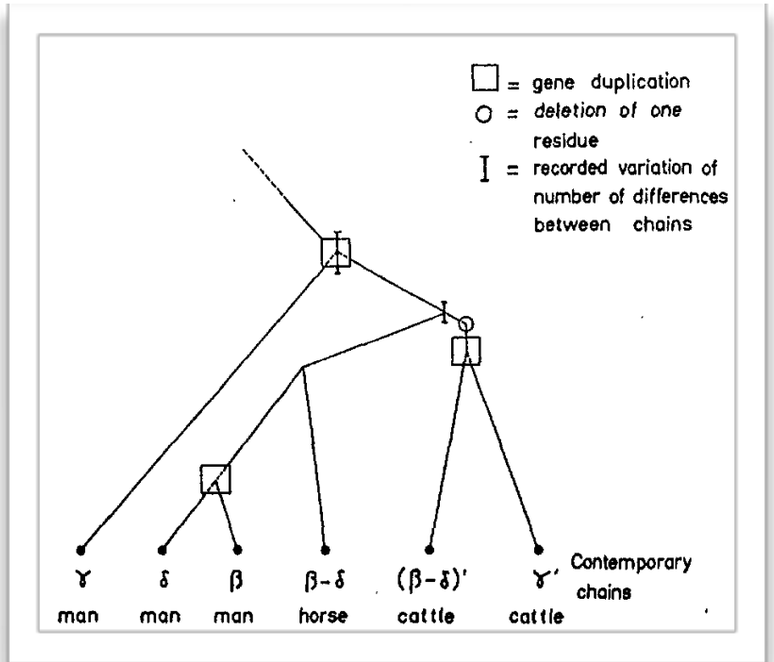


Cow

**speices**

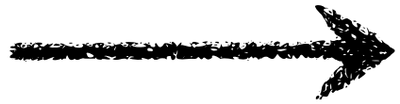
# The story of individual gene families is often blurred

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




**Zukerkandl & Pauling** 1965

gene tree draw into the species tree






**Hu**man

**Ho**rse

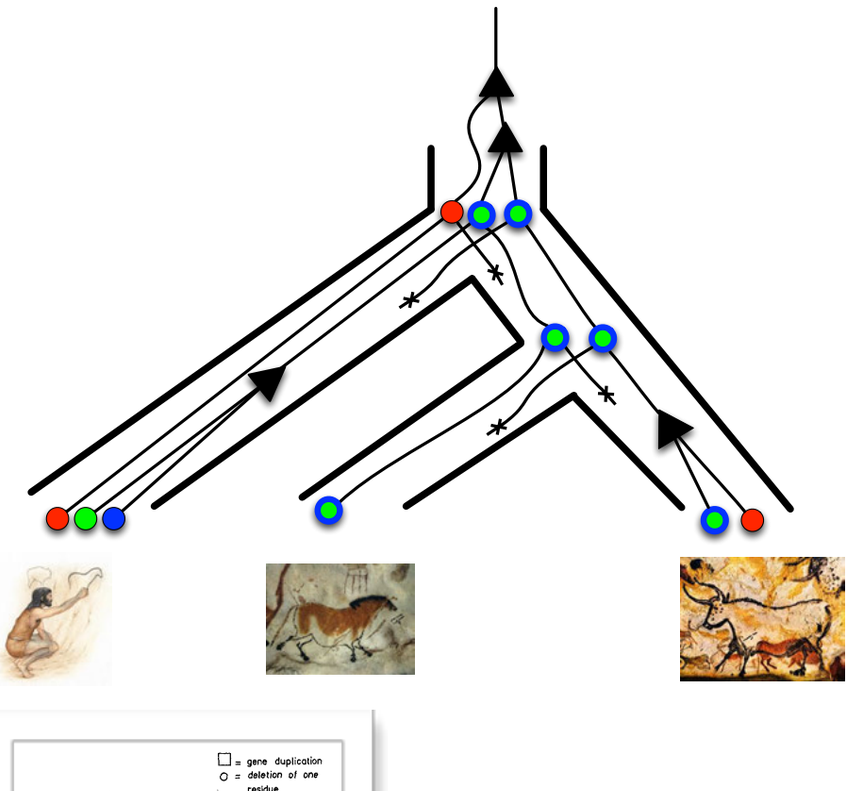
Cow

**species**

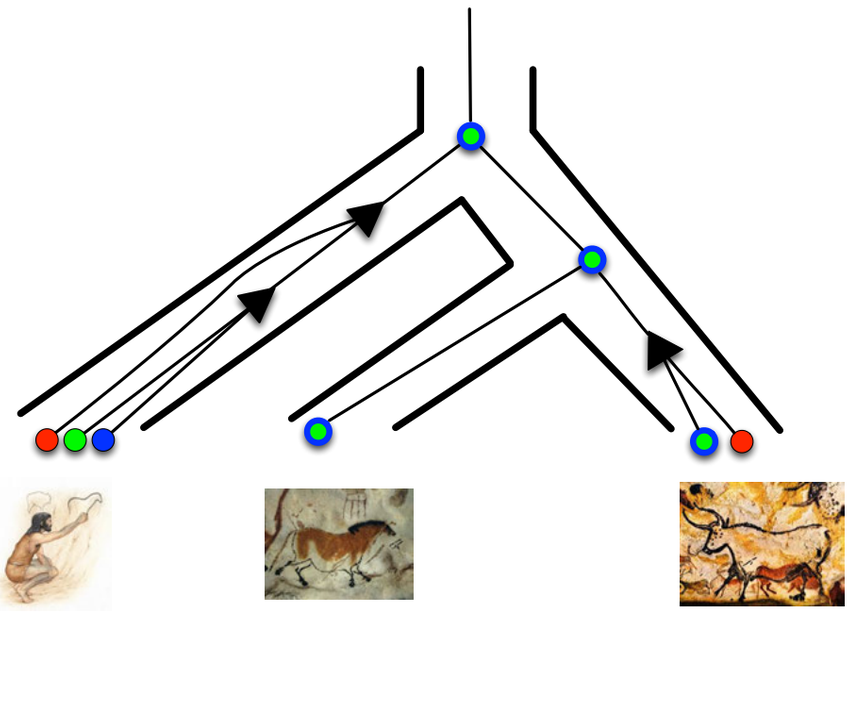
# .. but gene trees are generated along the species tree

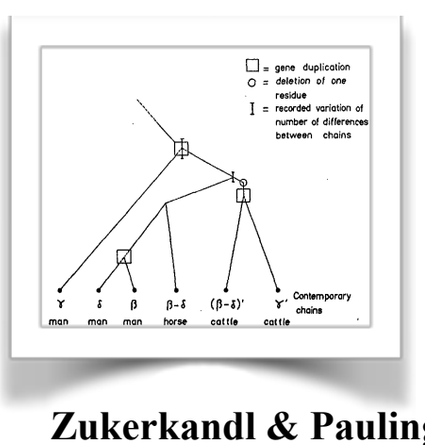
Given a model of gene family evolution a species tree induces a probability distribution over gene trees, thus **some gene trees are more likely than others given a particular species tree**.

## less likely gene tree



## more likely gene tree

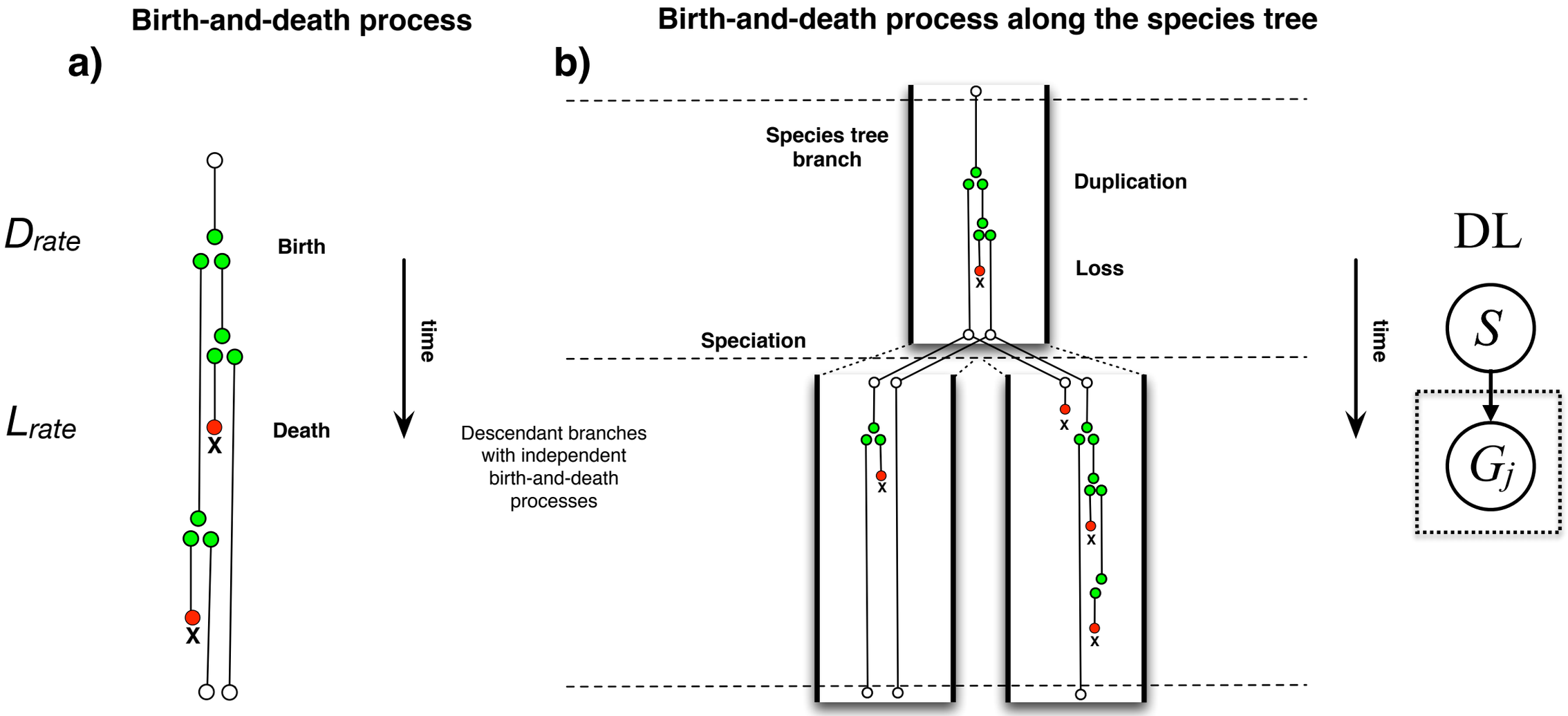





**Zukerkandl & Pauling** 1965

# .. but 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. For the DL process to calculate the likelihood of a ***gene tree*** we sum over all possible ***gene birth and death events*** along a given ***species tree.***



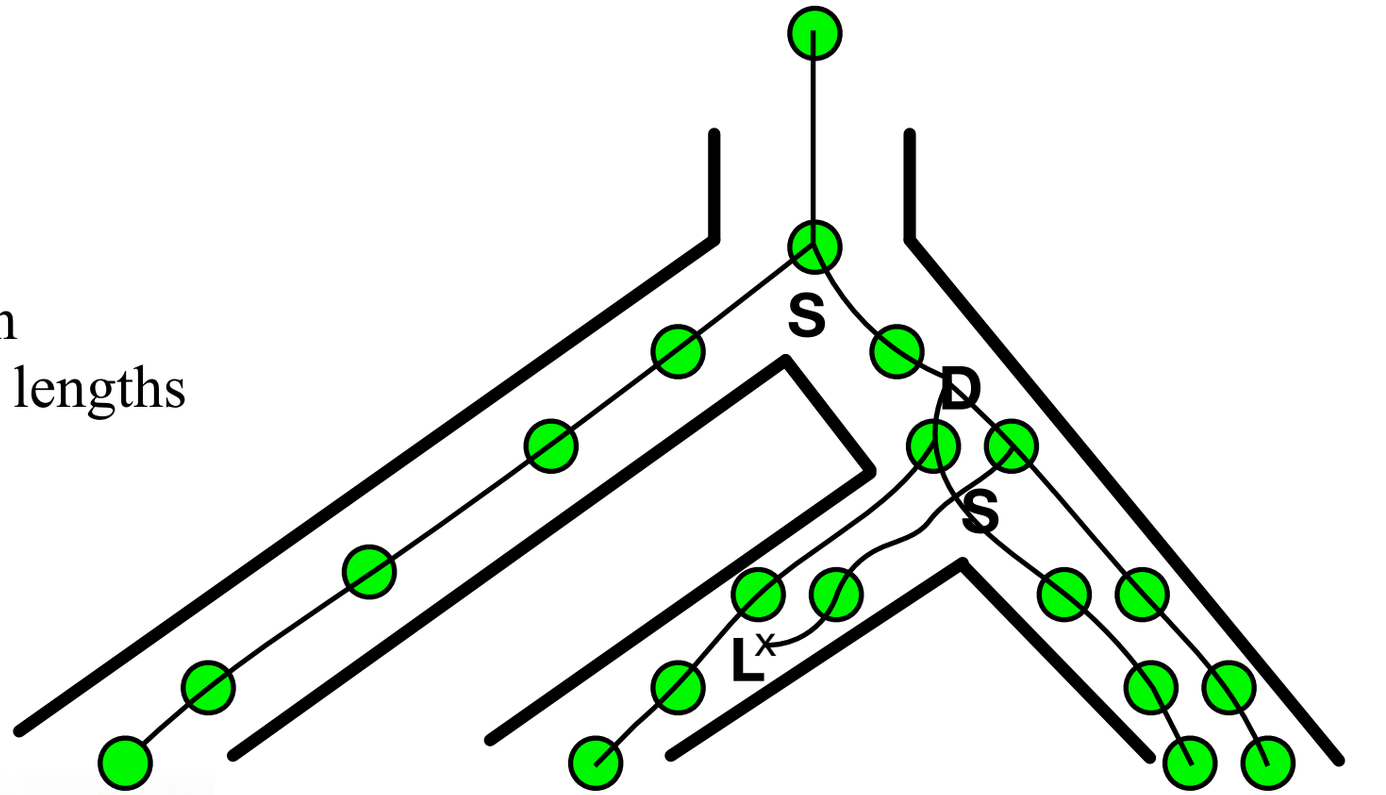
# .. but gene trees are generated along the species tree

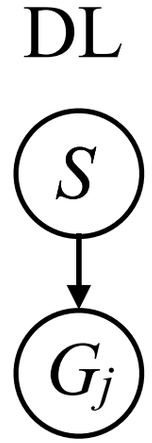
Given a model of gene family evolution a species tree induces a probability distribution over gene trees. For the DL process to calculate the likelihood of a ***gene tree*** we sum over all possible ***gene birth and death events*** along a given ***species tree.***

$D_{rate}$

$L_{rate}$

Species tree with
time like branch lengths




**Hu**man

**Ho**rse

Cow

**species**

implemented in ALE:

**http://github.com/ssolo/ALE**

# .. but 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. For the DL process to calculate the likelihood of a ***gene tree*** we sum over all possible ***gene birth and death events*** along a given ***species tree.***

$D_{rate}$

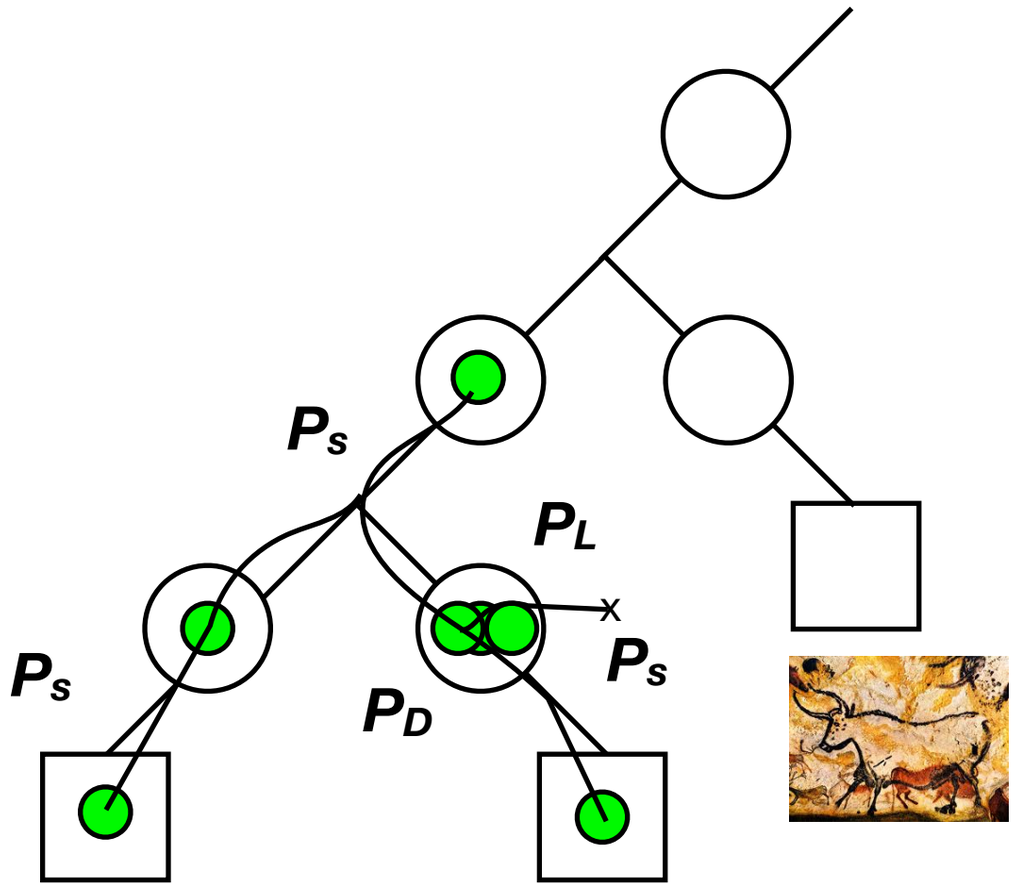
$L_{rate}$

Rooted species tree

$$P_S + P_D + P_L = 1$$

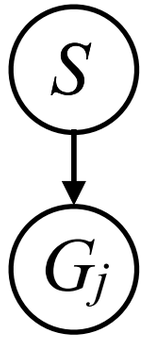
$$P_D = D_{rate} / D_{rate} + L_{rate}$$

$$P_L = L_{rate} / D_{rate} + L_{rate}$$






DL




implemented in ALE:

http://github.com/ssolo/ALE

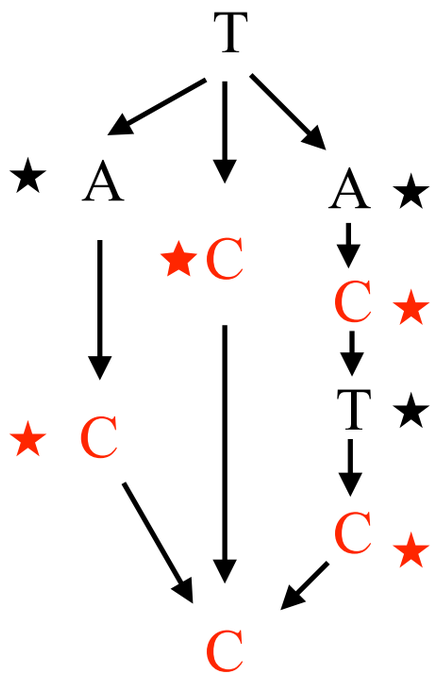
# The story of homologs genes can be reconstructed

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

$$p(A|G)$$

*requires summing over all possible substitution paths.*

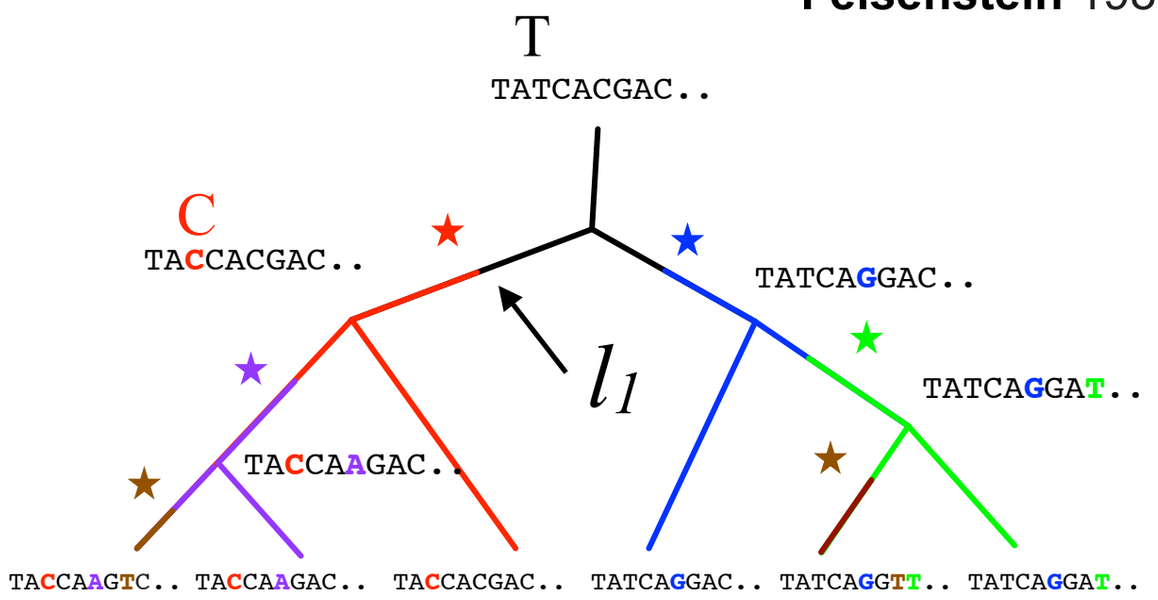
**sum over subs. along branch**
**conditional on states on top and bottom**

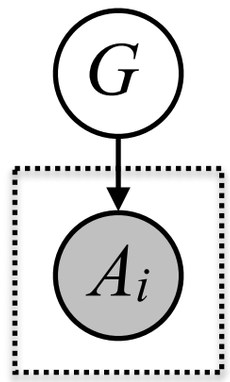



$$P_{\mathrm{TC}}(l_1)$$

**sum over ancestral states**

**Felsenstein** 1981



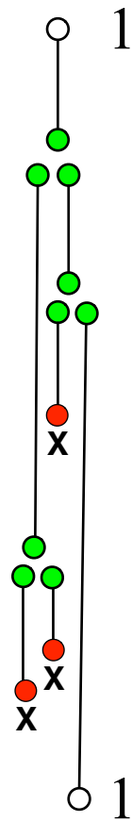

$$= .. \times P_{\mathrm{TC}}(l_1) \times P_{\mathrm{CC}}(l_2) \times P_{\mathrm{CC}}(l_3) \times P_{\mathrm{CC}}(l_4) \times ..$$

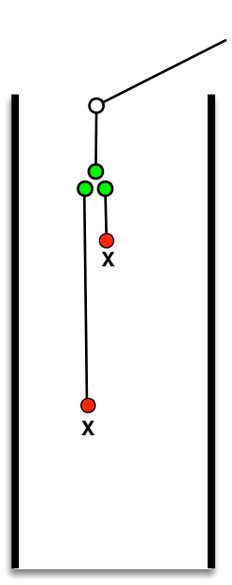
# .. but gene trees are generated along the species tree

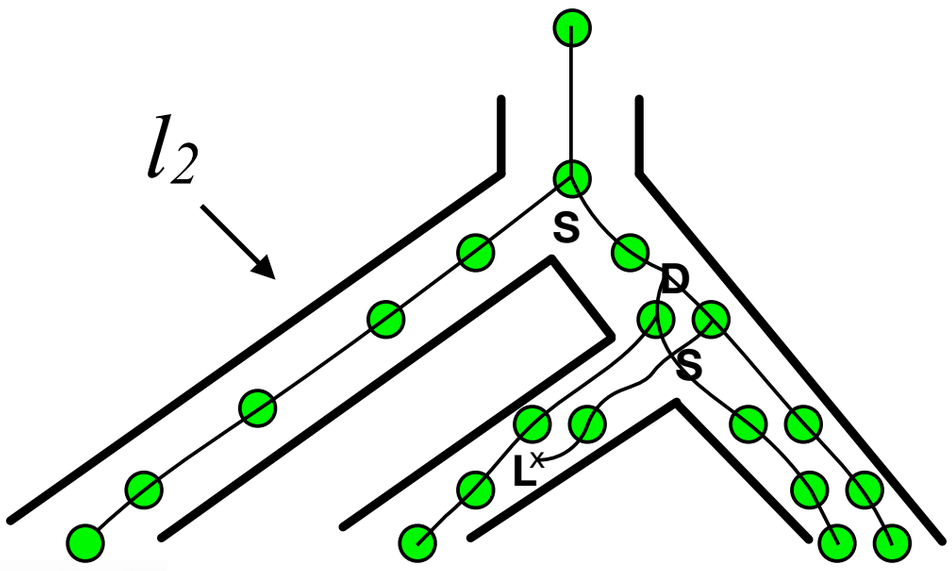
Given a model of gene family evolution a species tree induces a probability distribution over gene trees. For the DL process to calculate the likelihood of a ***gene tree*** we sum over all possible ***gene birth and death events*** along a given ***species tree.***

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

$P_{11}(l_2, \mathbf{Hu})$ $P_{10}(\mathbf{Ho})$

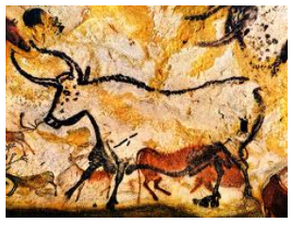


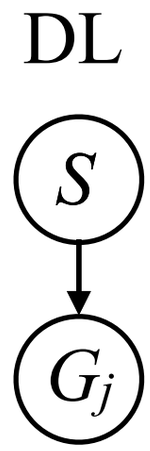




**Hu**man

**Ho**rse

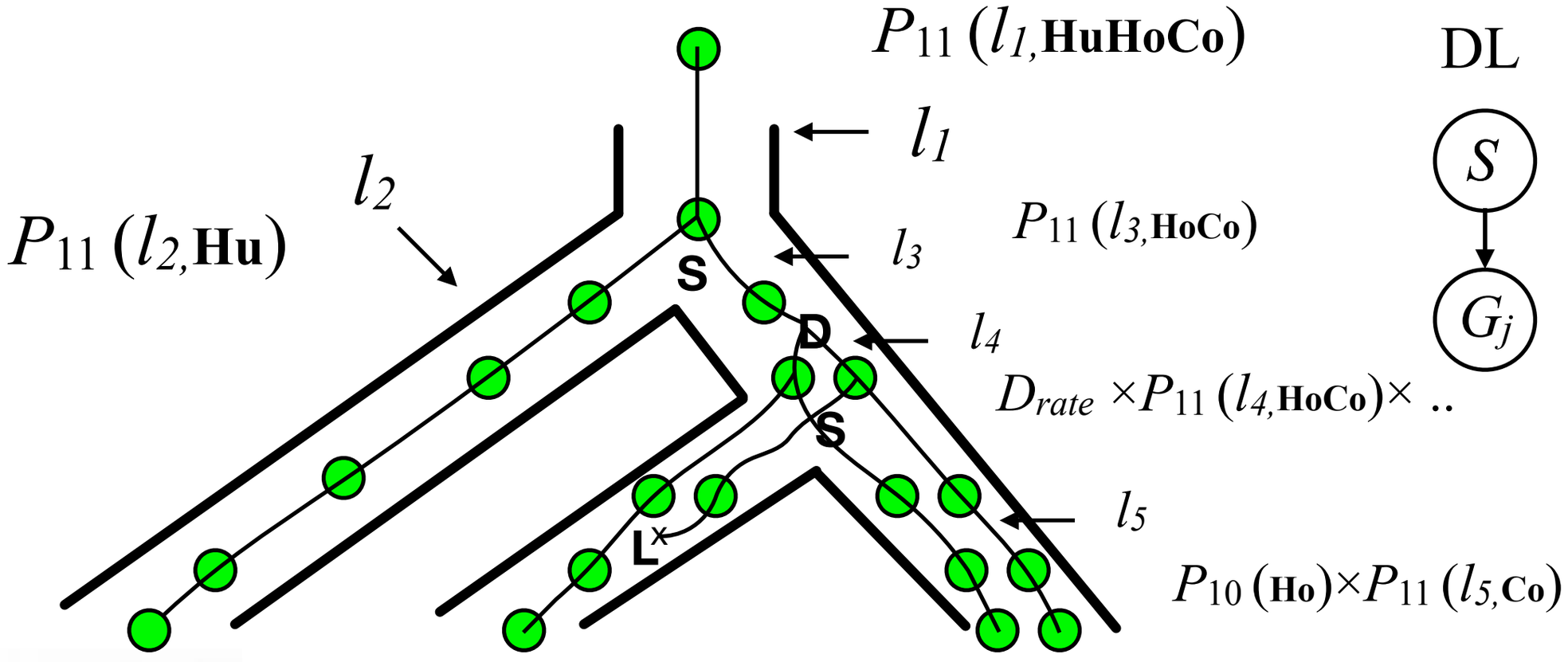


**C**ow



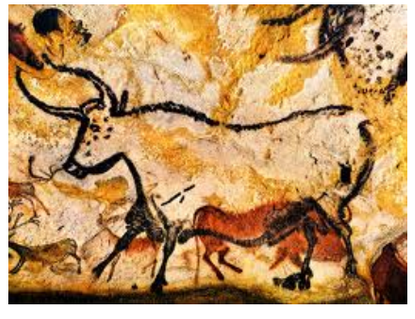
# .. but gene trees are generated along the species tree

Given a model of gene family evolution a species tree induces a probability distribution over gene trees. For the DL process to calculate the likelihood of a ***gene tree*** we sum over all possible ***gene birth and death events*** along a given ***species tree.***




**Hu**man

**Ho**rse



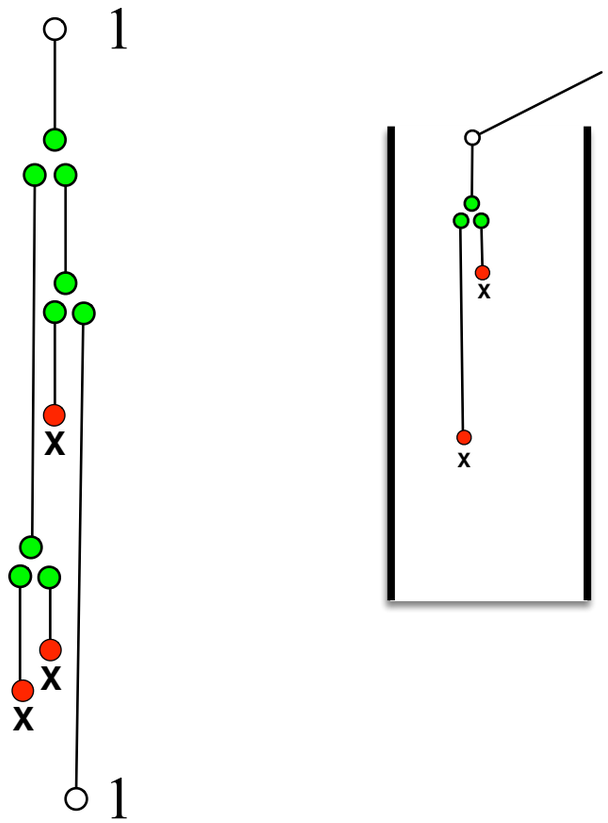
**C**ow

# .. but 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. For the DL process to calculate the likelihood of a ***gene tree*** we sum over all possible ***gene birth and death events*** along a given ***species tree.***

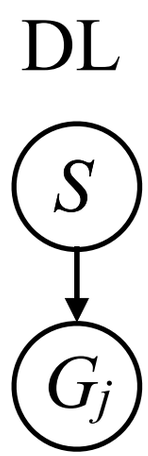
**sum over gene birth and death events along a branch conditional on reconciliation**

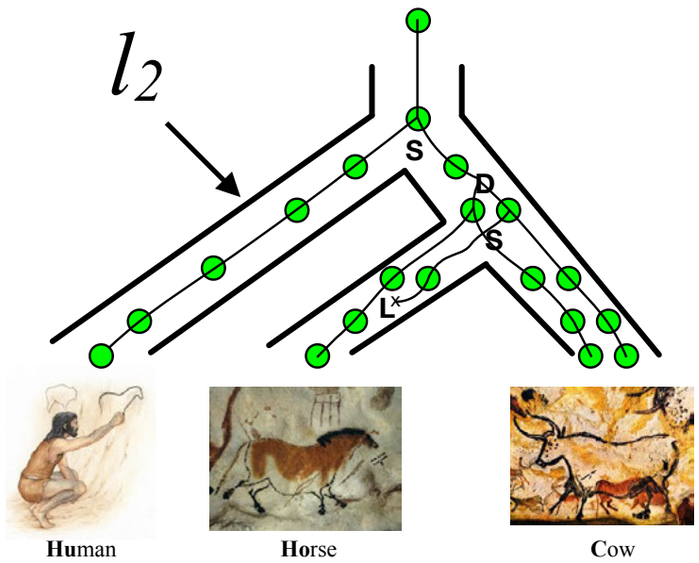
$P_{11}(l_{2,}\mathbf{Hu})$ $E(\mathbf{Ho})$




**sum over all reconciliations with species tree conditional on gene tree**

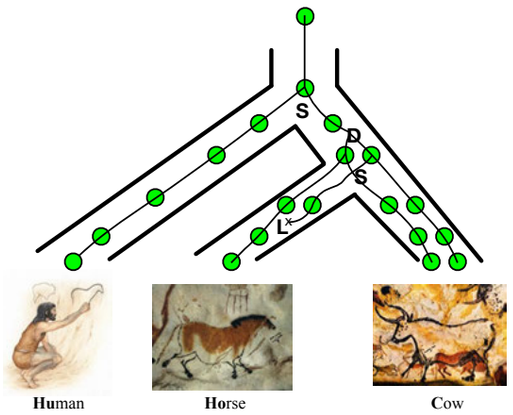
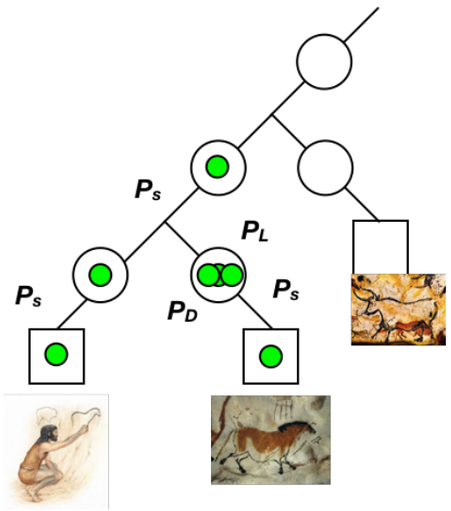
**Arvestad** 2010

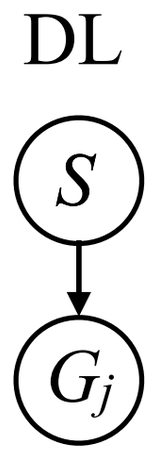


# .. but gene trees are generated along the species tree

Given a model of gene family evolution a species tree induces a probability distribution over gene trees. For the DL process to calculate the likelihood of a ***gene tree*** we sum over all possible ***gene birth and death events*** along a given ***species tree.***

| | **calculation complexity** | **parameters (ML or Bayes)** |
|---|---|---|
|  | 2 to 10 × log(#species) × #genes | D&L rates<br>branch lengths, root |
|  | log(#species) × #genes | D&L rates<br>root |



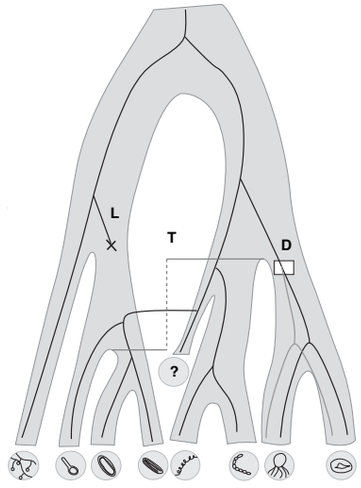
# Gene trees and species trees can be jointly reconstructed

*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:**

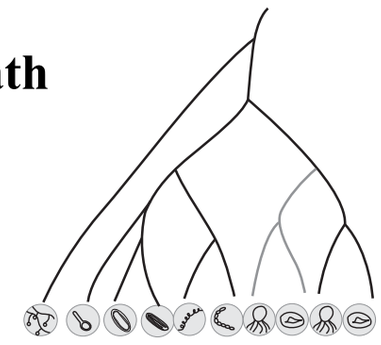
$p(G|S)$ **Arvestad** 2010 $p(A|G)$ **Felsenstein** 1981

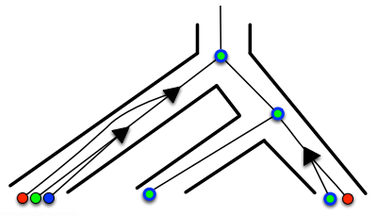
species tree gene trees sequences DL




$S$

**gene birth and death**

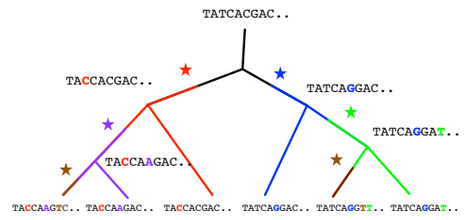




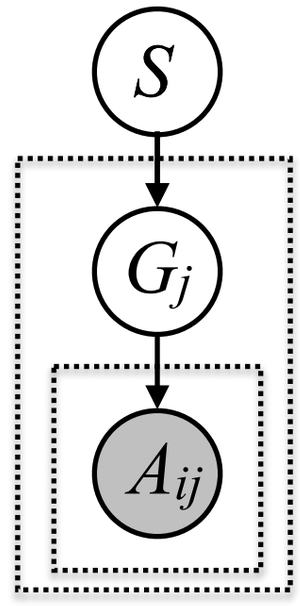
$G$

**substitution events**

TATCAAGTC..
TATCAAGAC..
TATCACGAC..
TACCAGGAC..
TACCAGGTT..
TACCAGGAT..




$A$



# 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

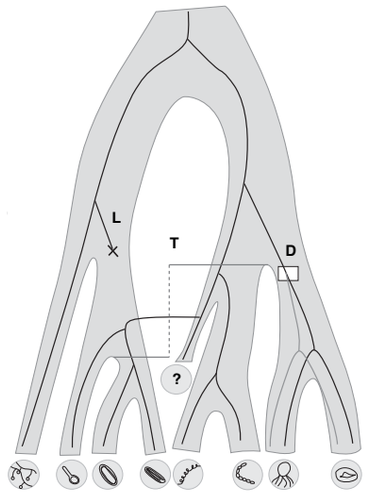
$$\prod_j$$

optimise $S$
and estimate rates

**clients:**
calculate

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

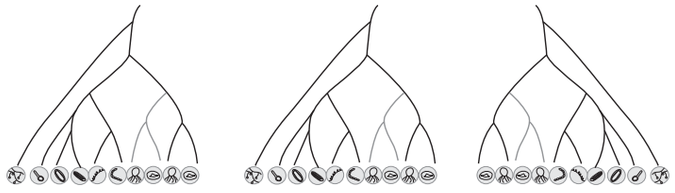
optimise (or integrate over) $G_j$




**species tree**

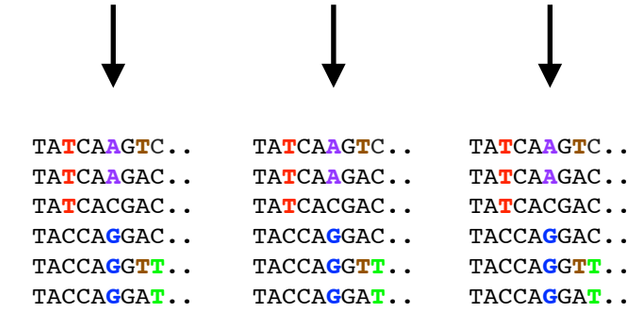
$S$

**Daubin & Boussau** 2011



$G_j$

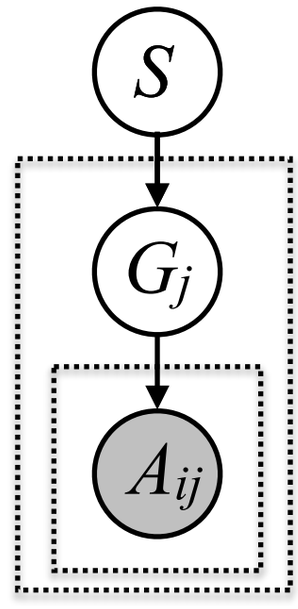
**gene trees**




$A_{ij}$

**sequences**

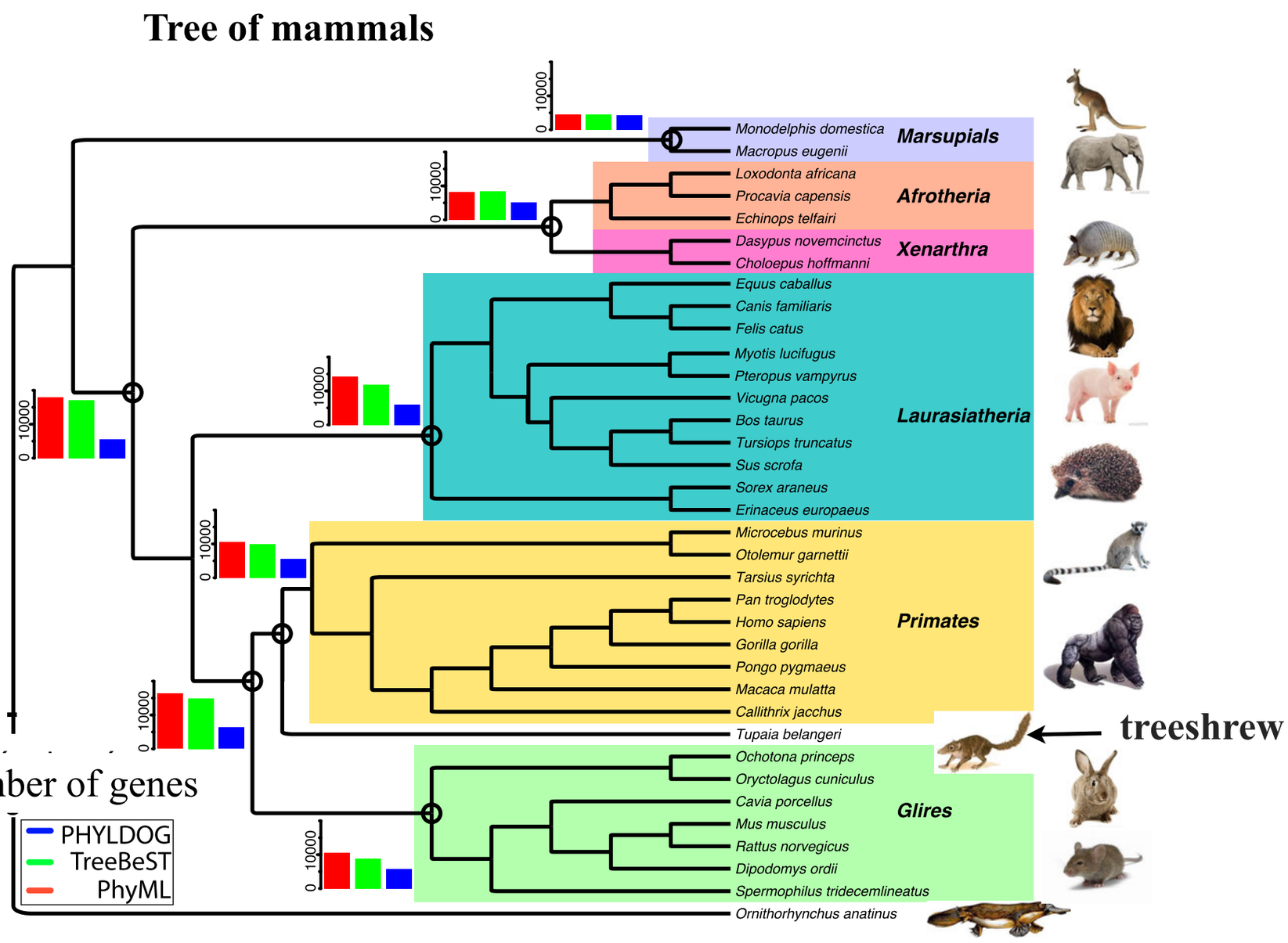
DL




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

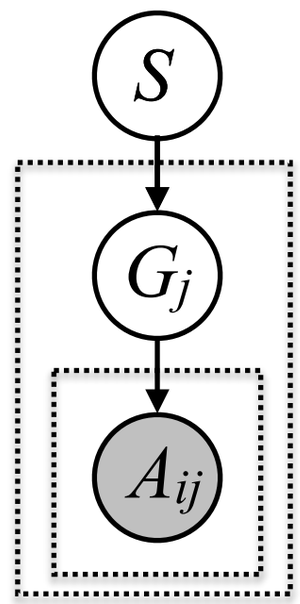
# Genome-scale reconstruction of the tree of mammals

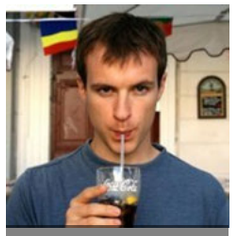
Using 6966 gene families from 36 mammals we jointly reconstructed the species tree and gene trees.




**MAP**

DL



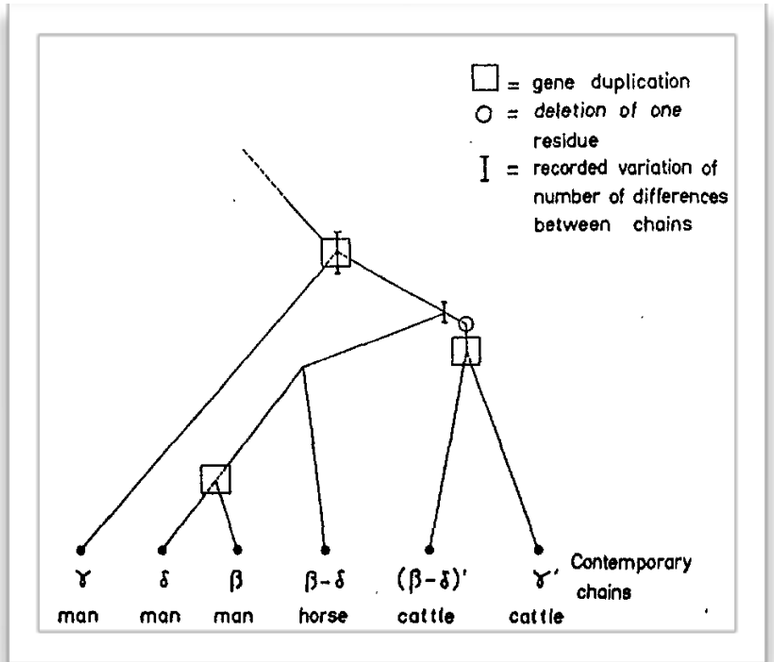



Bastien Boussau
LBBE

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

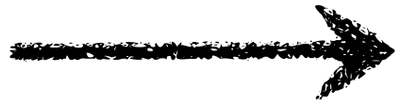
# The story of individual gene families is often blurred

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




**Zukerkandl & Pauling** 1965

gene tree draw into the species tree





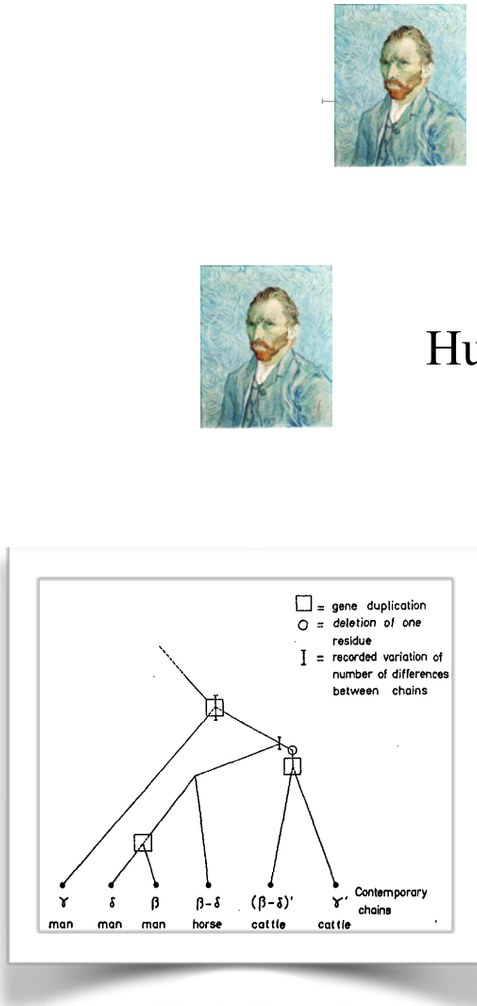

**Hu**man

**Ho**rse

Cow

**species**

# The first ever gene tree




**Zukerkandl & Pauling** 1965





Horse_beta-delta

Human_delta

Human_beta

Human_gamma

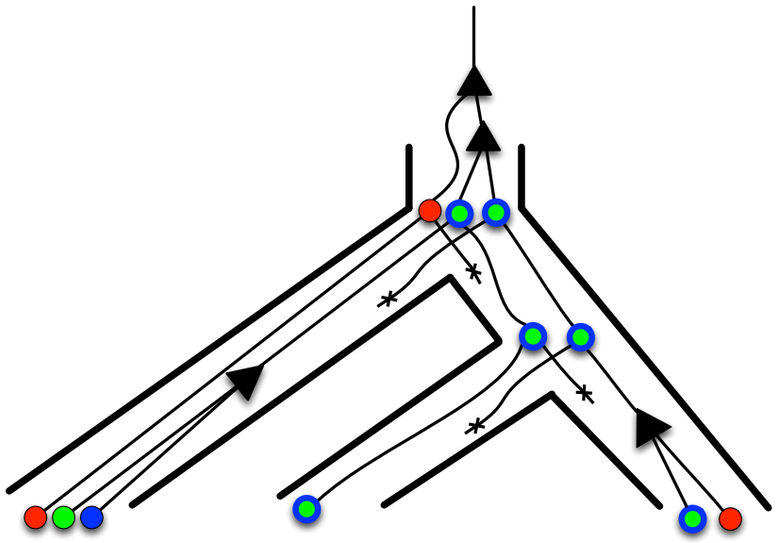
Cow_gamma

Cow_beta-delta

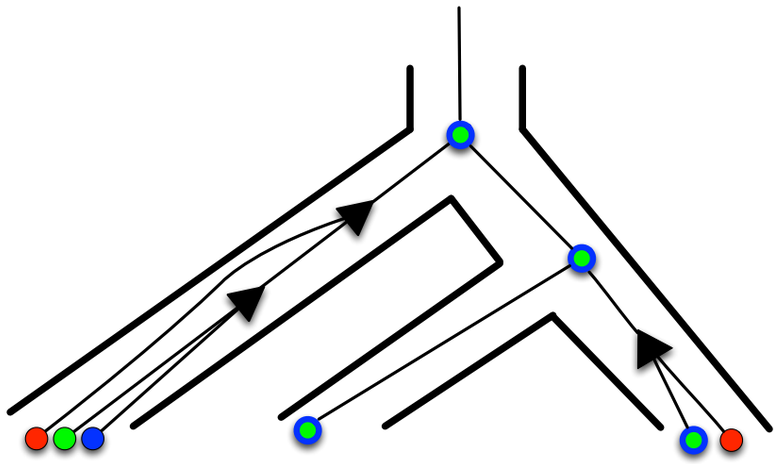
# The story of individual gene families is often blurred

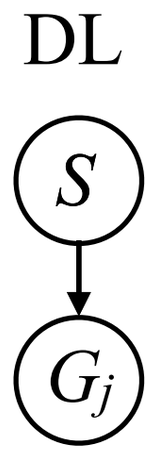
Errors in gene trees will result in conflicts with the species tree that imply spurious evolutionary events.

## gene tree with errors



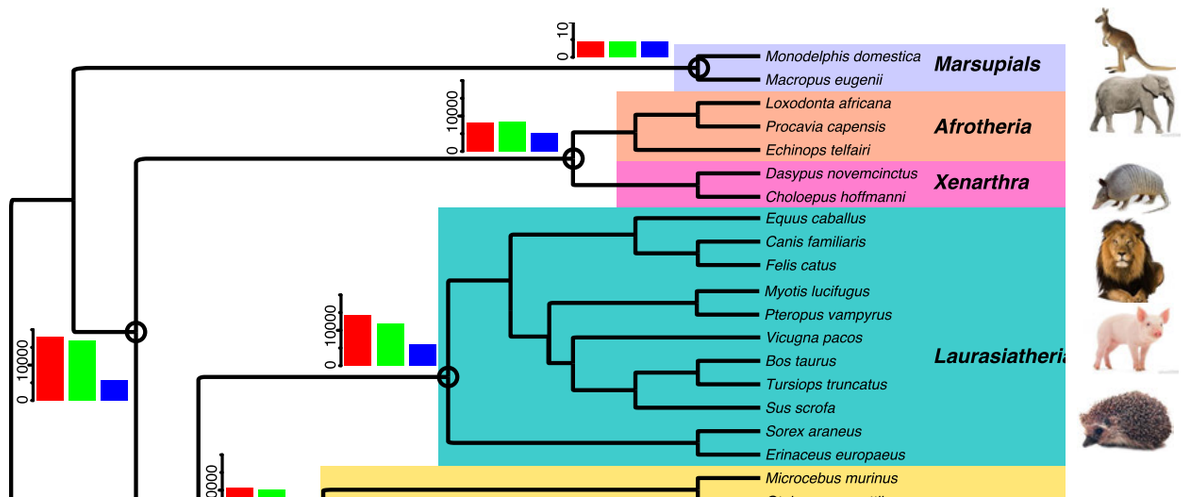
## correct gene tree






**S-unaware trees**

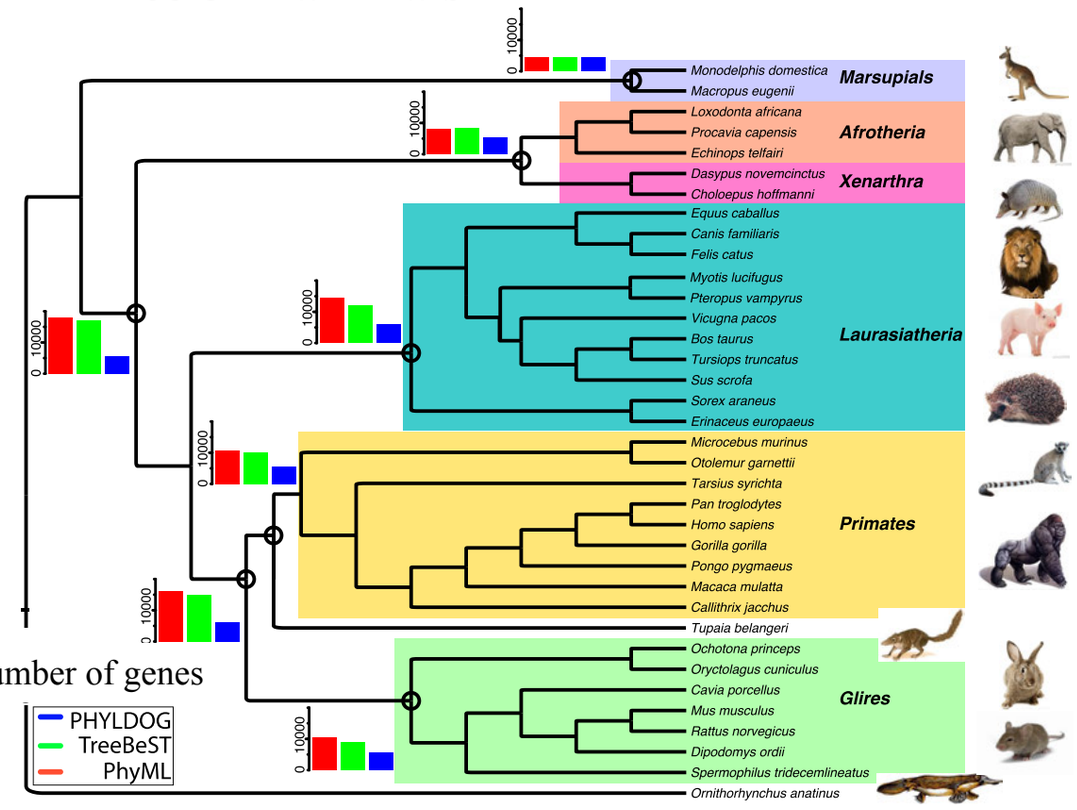
**S-aware trees**



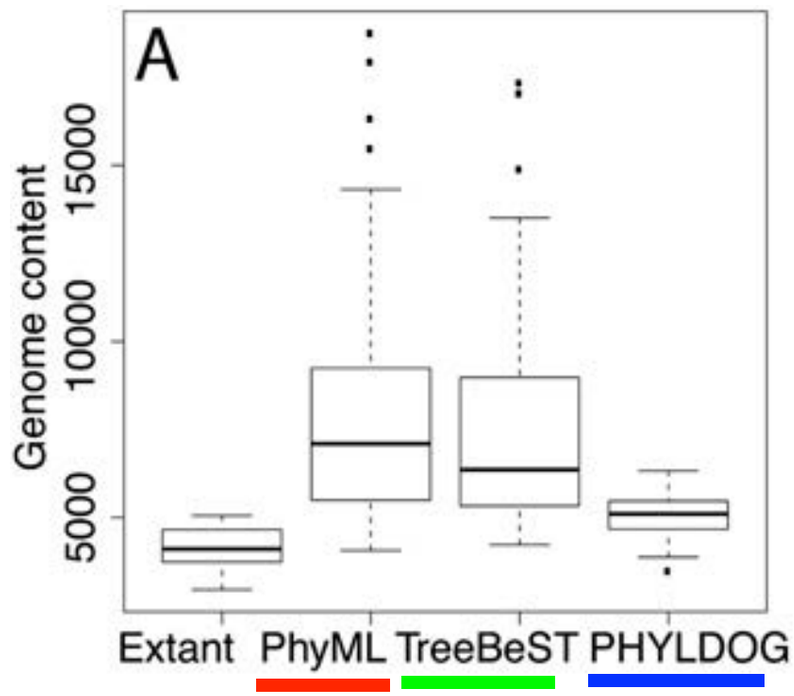
# Genome-scale reconstruction of the tree of mammals

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

## Tree of mammals




*Joint reconstruction gives more realistic gene content*

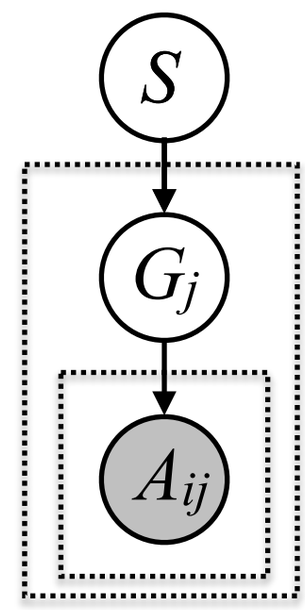


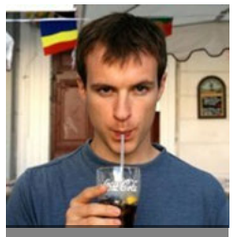

**S-unaware trees**

**S-aware trees**

**MAP**

DL



Bastien Boussau

LBBE

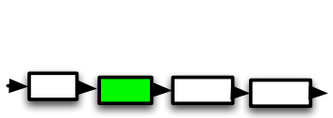
Boussau, **Szöllősi**, Duret, Gouy, Tannier & Daubin *Genome Res.* (2013)
*Genome-scale coestimation of species and gene trees*

# Genome-scale reconstruction of the tree of mammals

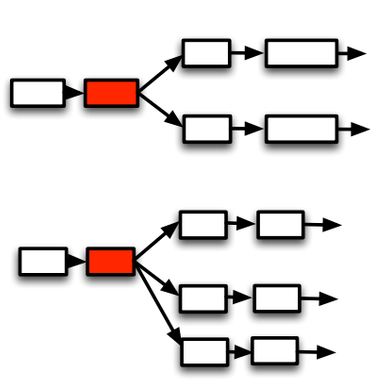
Using 6966 gene families from 36 mammals we jointly reconstructed the species tree and gene trees.

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

**correct gene trees** — two neighbours

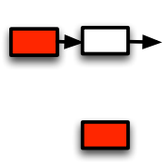


**gene tree errors** — three or more neighbours



..

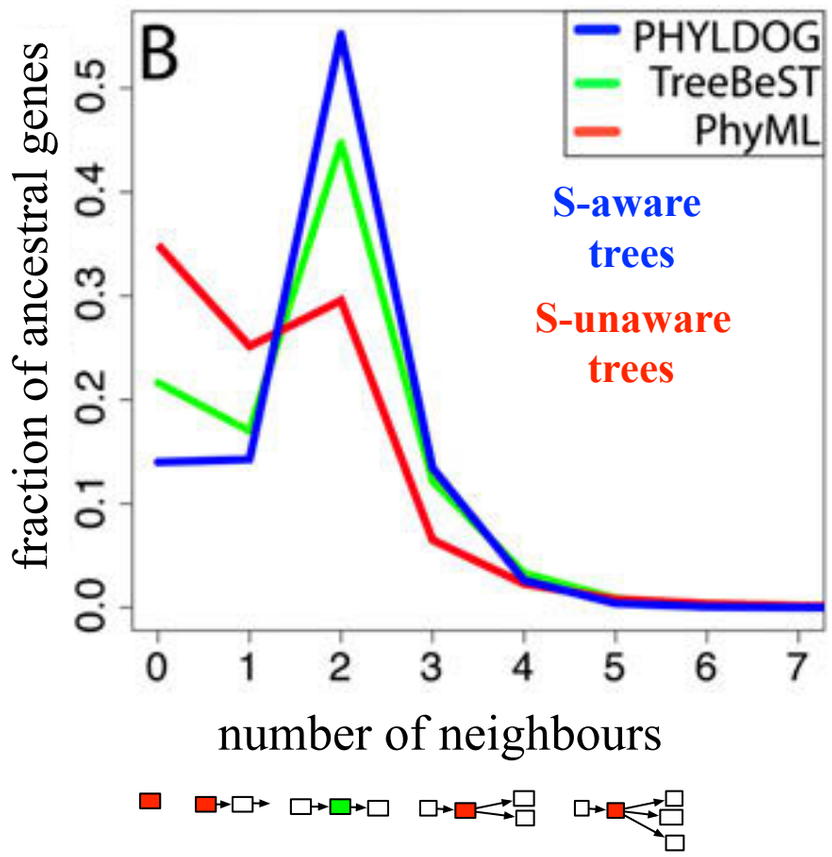
one or zero neighbours



Eric Tannier

LBBE

*Joint reconstruction imply more realistic ancestral gene order*




**MAP**

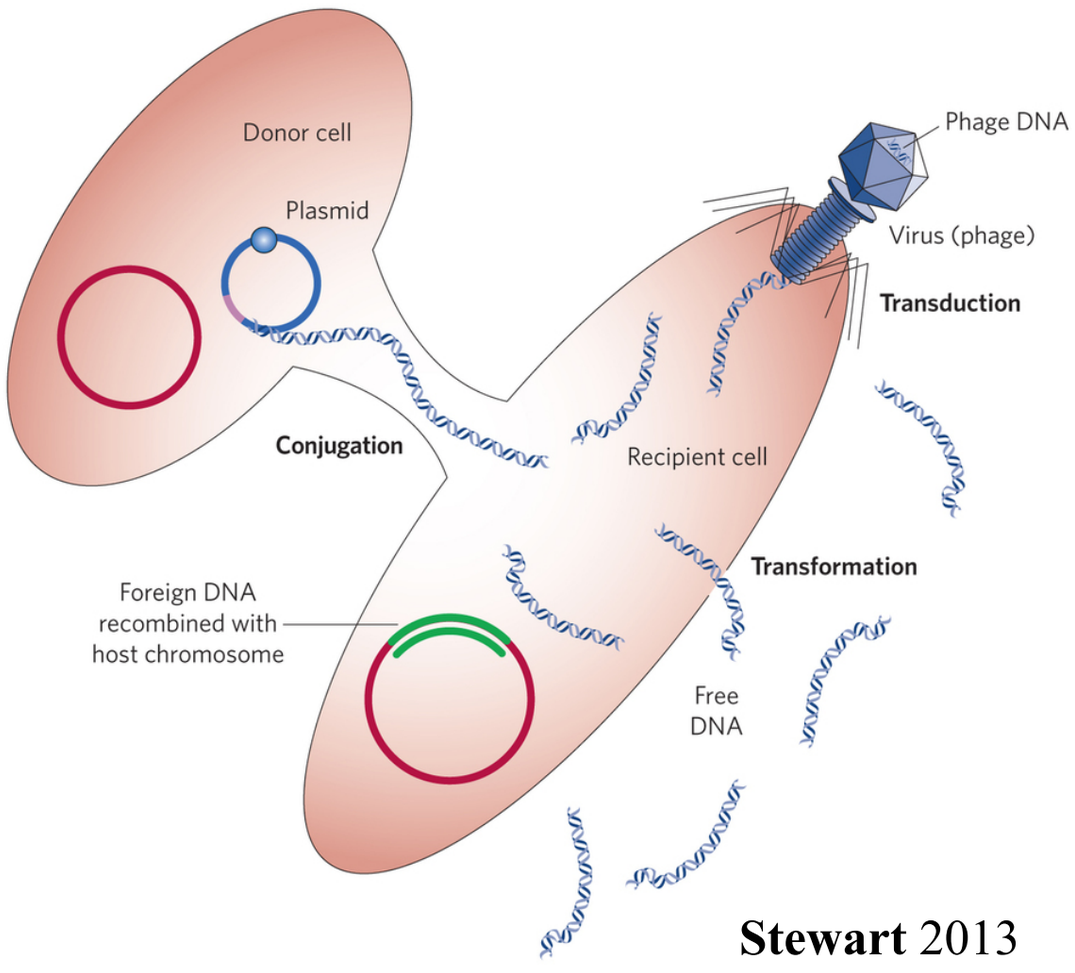
DL


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**



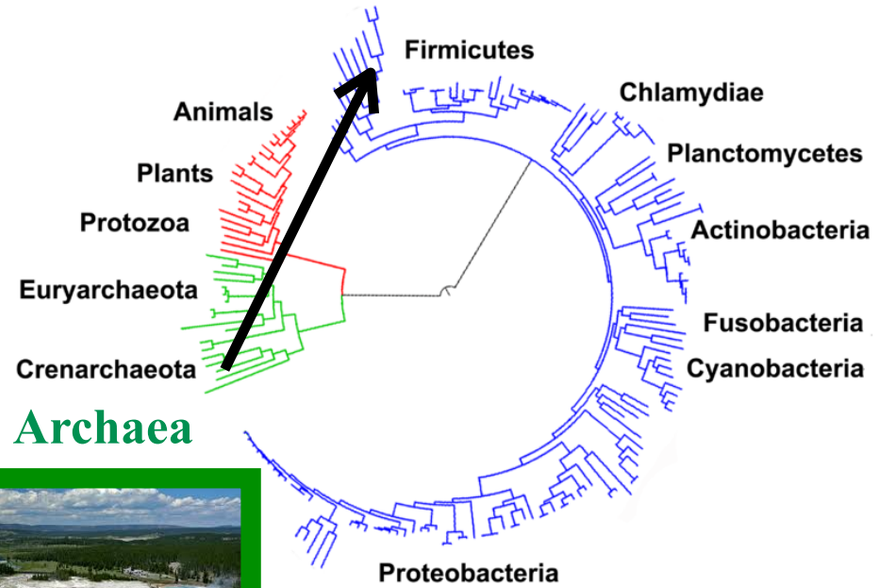

**Stewart** 2013

**classic examples:**

antibiotic resistance

thermophilic enzymes

**Bacteria**




**Archaea**

ssolo@elte.hu

# Horizontal gene transfer

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

## Carotenoids

**they produce it**

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

**they eat it**

*colour of flamingos and the human eye's macula lutea*

**except!**


*a species of aphid living on peas*

**Moran & Jarvik** 2010 Science

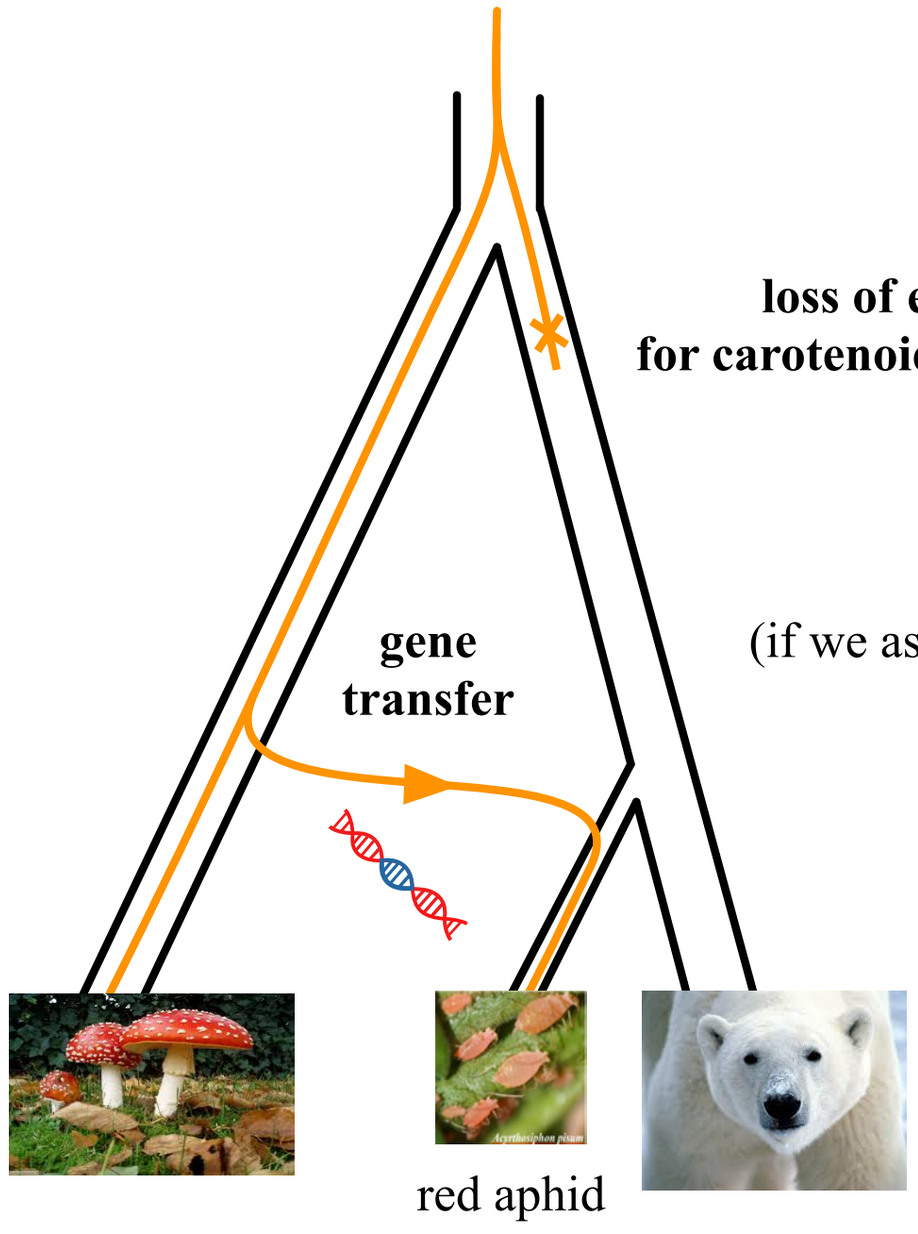
ssolo@elte.hu

# Horizontal gene transfer

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

*pea aphids*


**Moran & Jarvik** 2010 Science

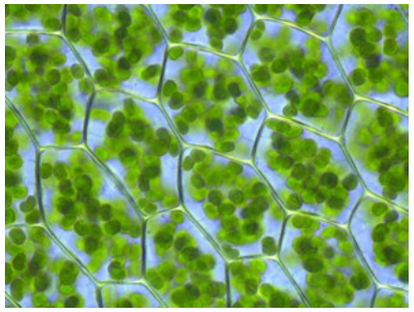


# Horizontal gene transfer

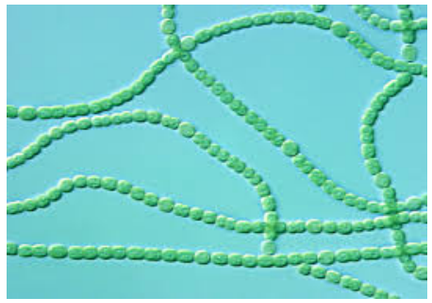
Horizontal gene transfer is common among unicellular organisms, but examples are know even among animals.

**photosynthesis**

chloroplasts of
algae and green plants



cyanobacteria



**except!**

*Eastern emerald elysia*
*(US East Coast)*

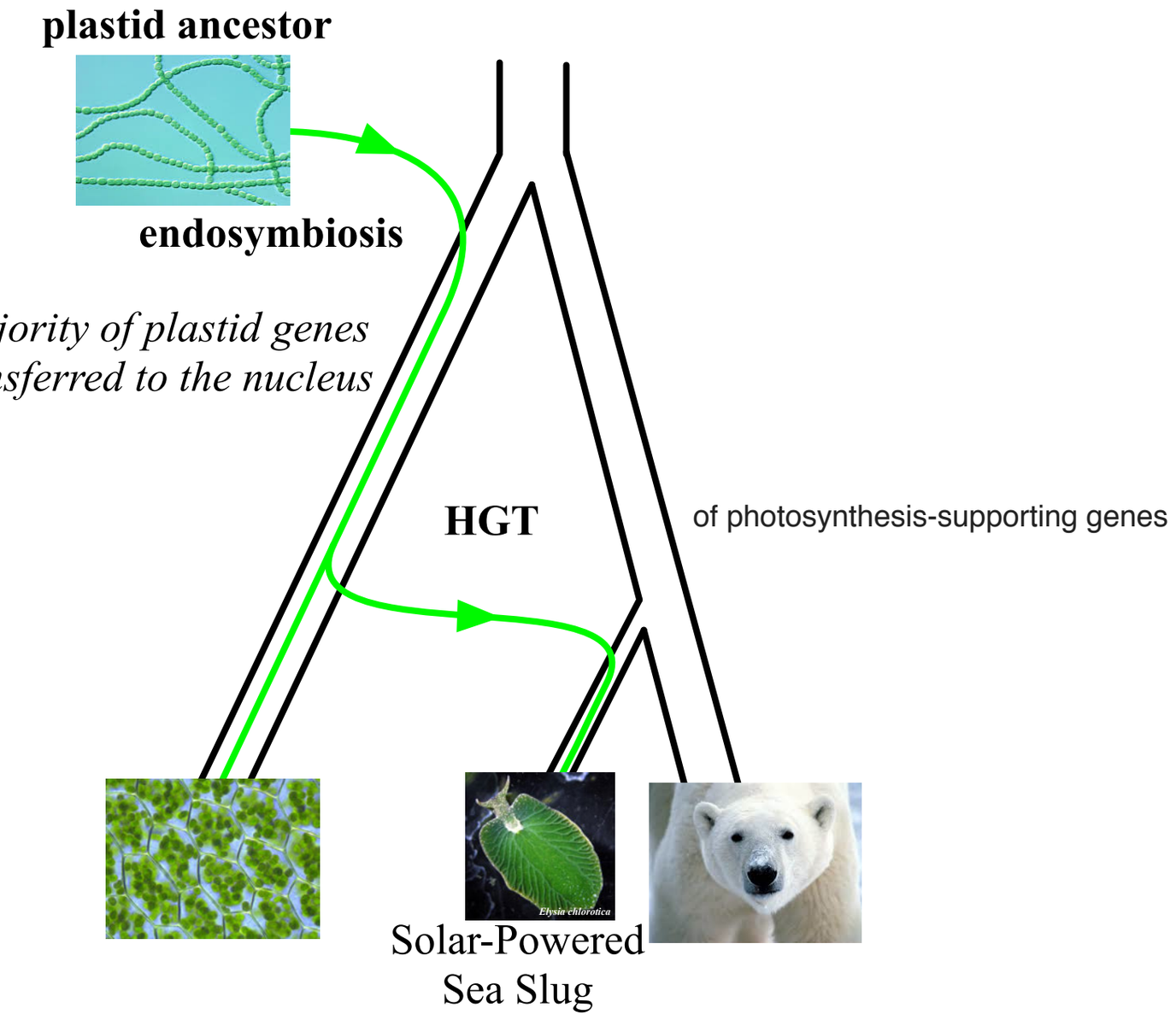

**Rumpho *et al.*** 2008 PNAS

# Horizontal gene transfer

Egysejtűek között gyakori a géntranszfer, de többsejtű élőlényeknél, köztük az állatok közt is ismertek példák.

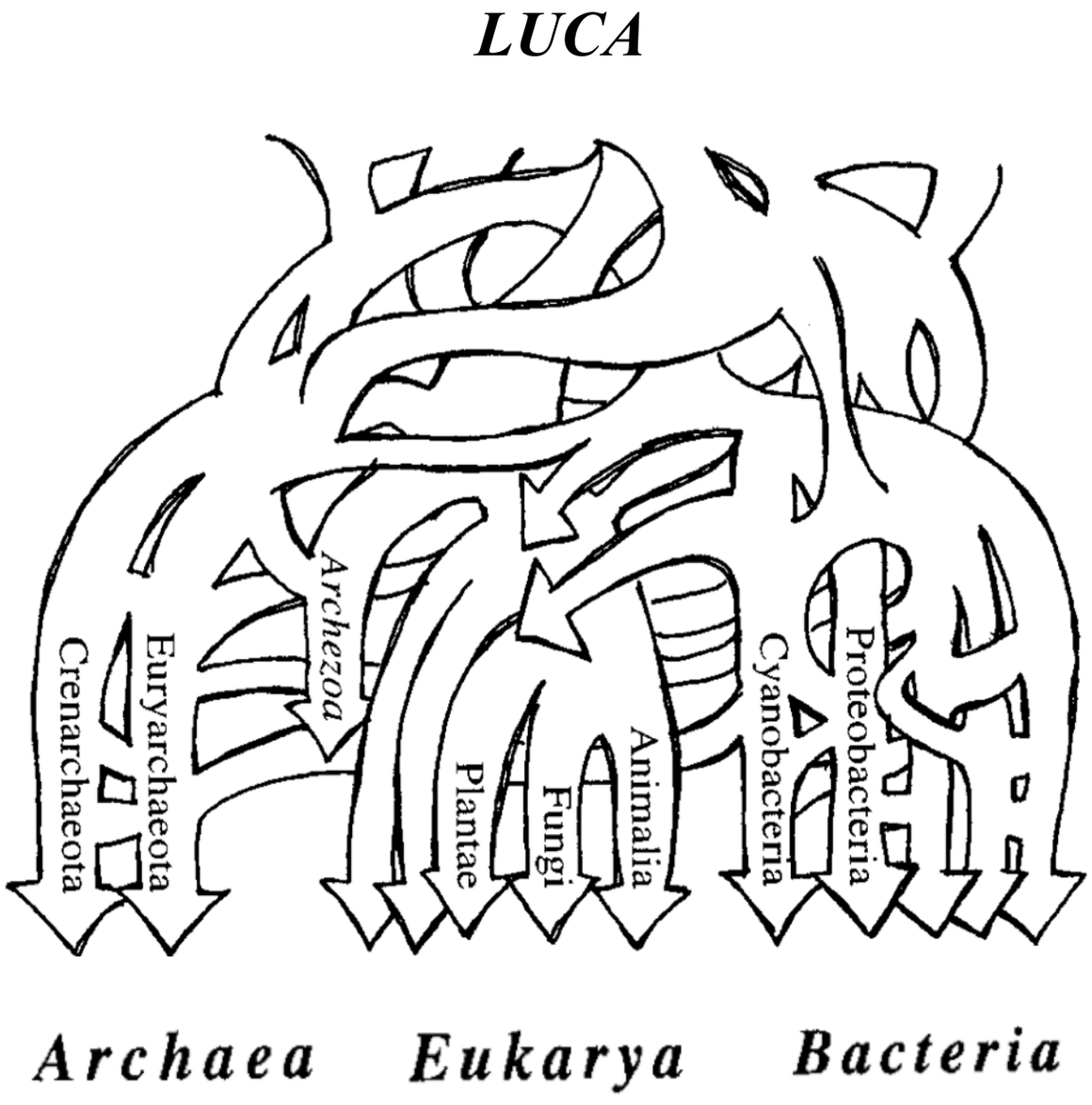
*Eastern emerald elysia*
*(US East coast)*


**Rumpho *et al.*** 2008



# Horizontal gene transfer as noise

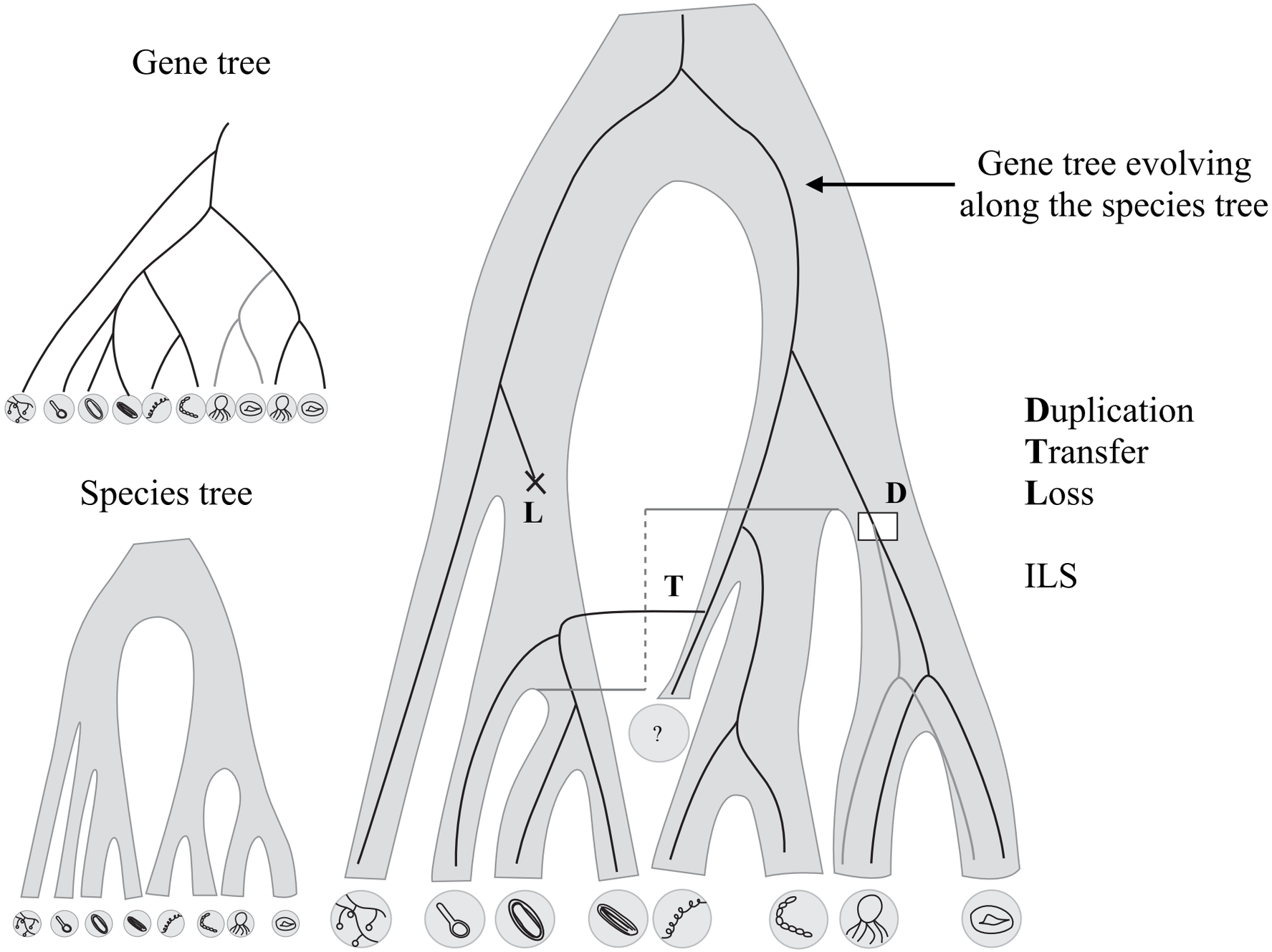
Gene transfers result in apparently contradicting gene phylogenies, fungi can seem closely related to aphids. A potentially high rate of transfer esp. early in the evolution of life, suggests that the vertical signal may be drowned in noise.




**Doolittle** 1999

# The problem is gene trees are not species trees

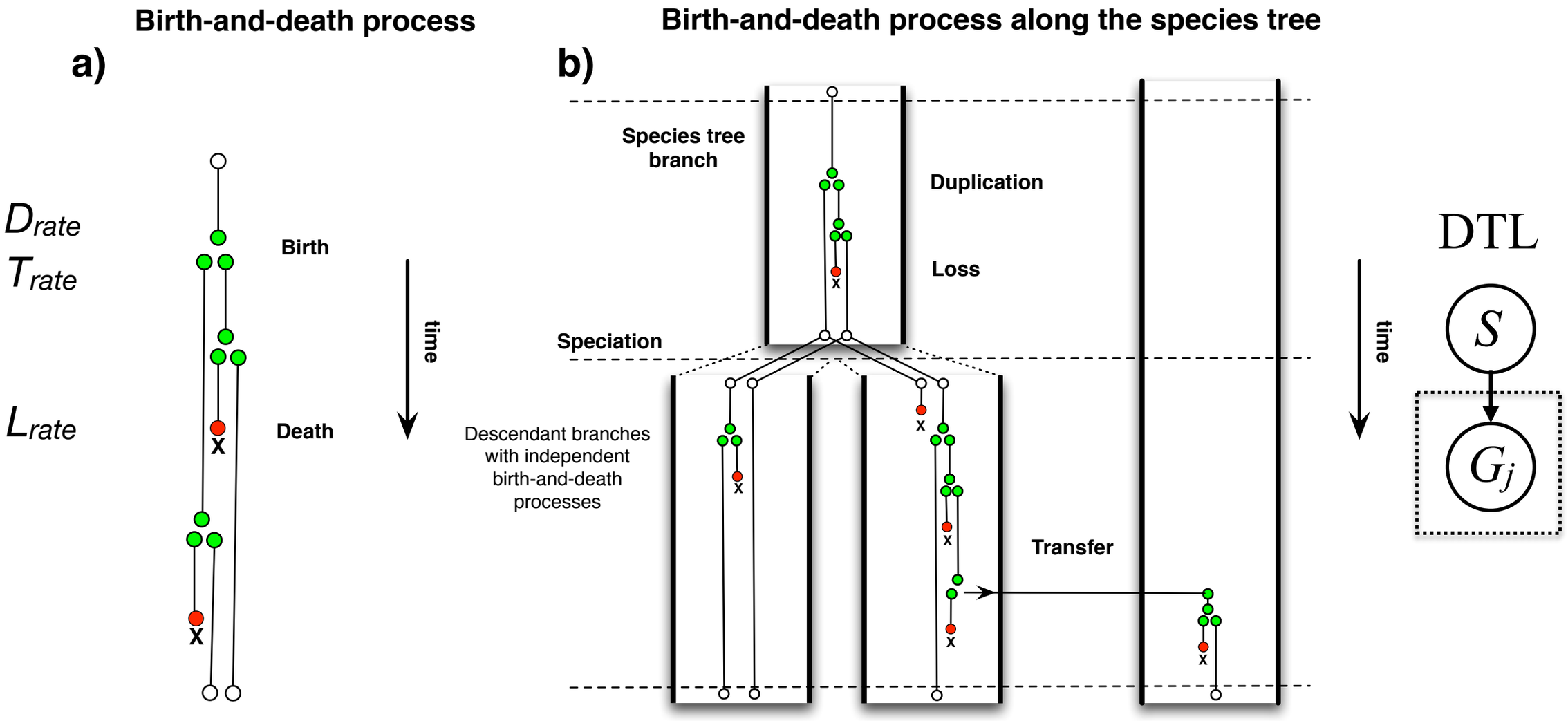
*A gene tree is a deformation of the species tree through the prism of genome evolution and population genetics processes.*




**Daubin & Szöllősi** 2016

# .. gene trees are generated along the species tree

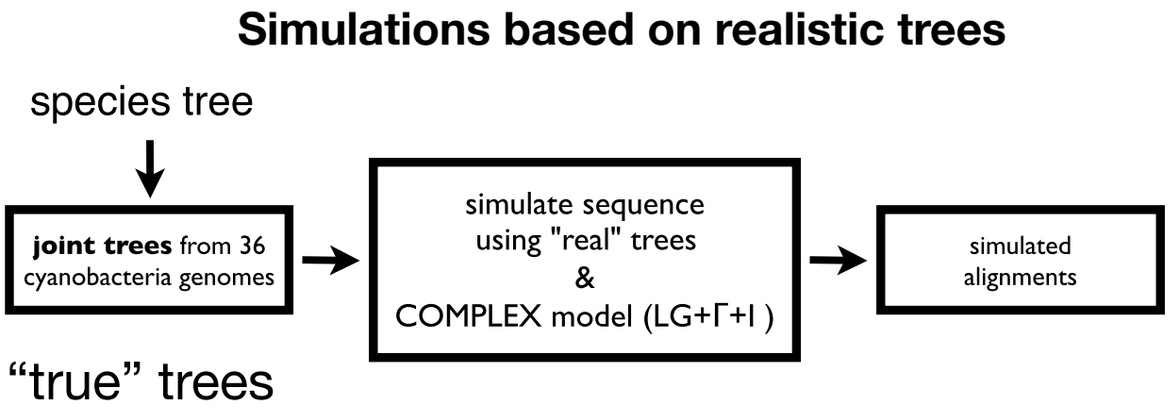
Given a model of gene family evolution a species tree induces a probability distribution over gene trees. For the DTL process to calculate the likelihood of a ***gene tree*** we sum over all possible ***gene birth and death events*** along a given ***species tree.***

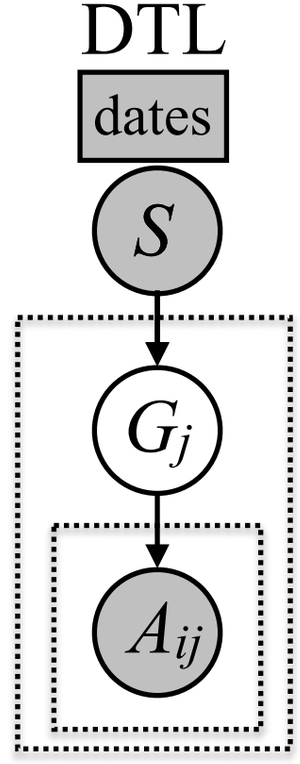



**Tofigh** *PhD thesis* 2009

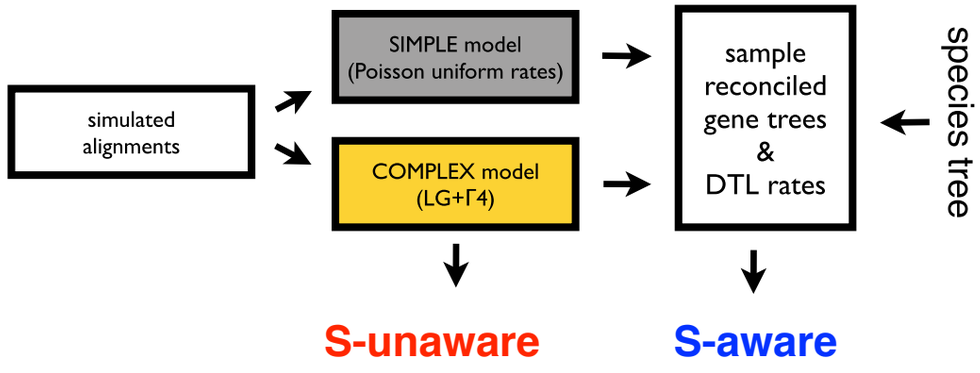
**Szöllősi**, Boussau, Abby, Tannier & Daubin ***PNAS*** (2012)
*Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations*

# “Realistic simulations” suggest S-aware methods are important



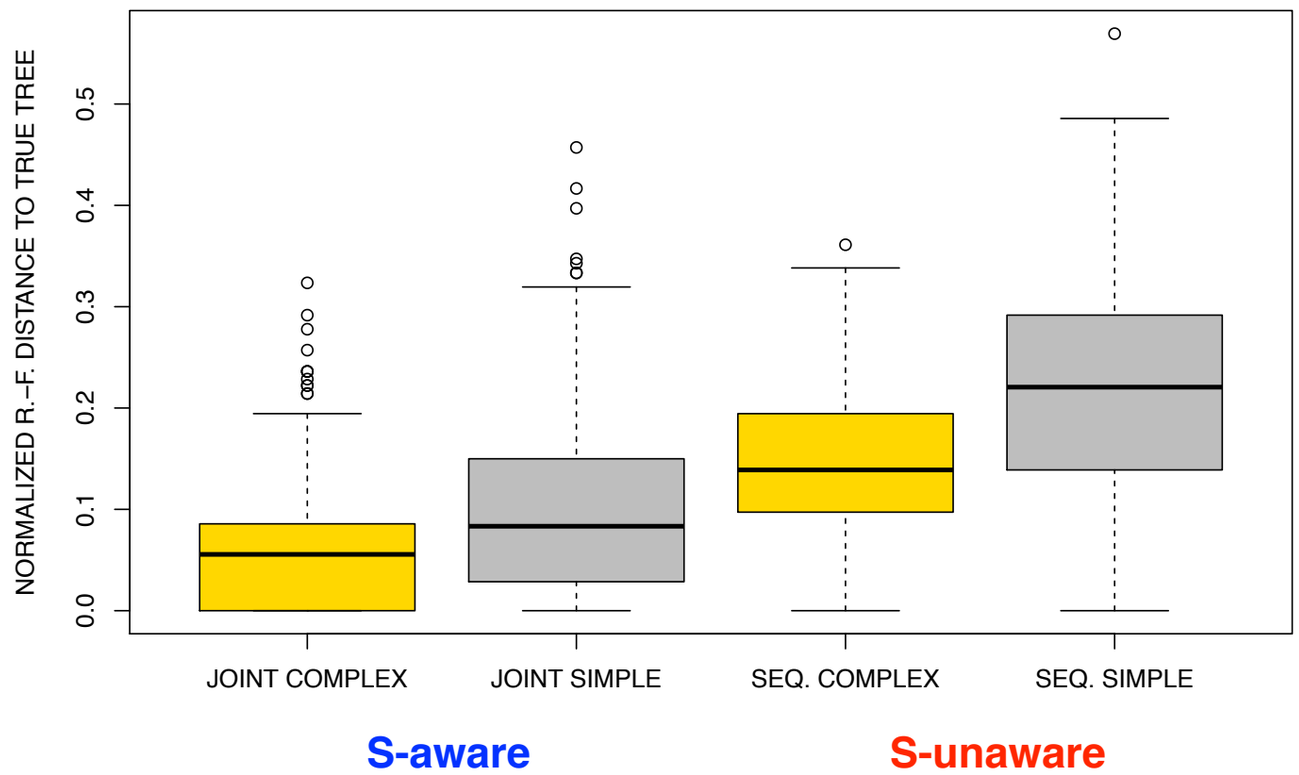



# “Realistic simulations” suggest S-aware methods are important

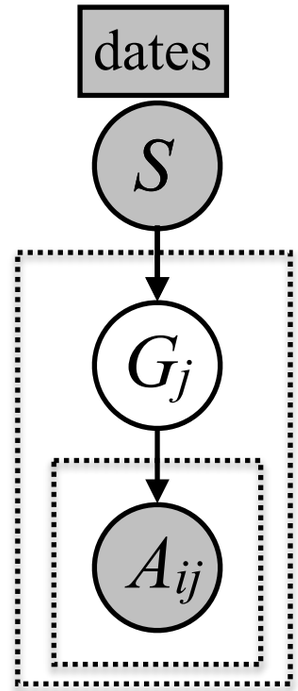



**Distance to true tree**

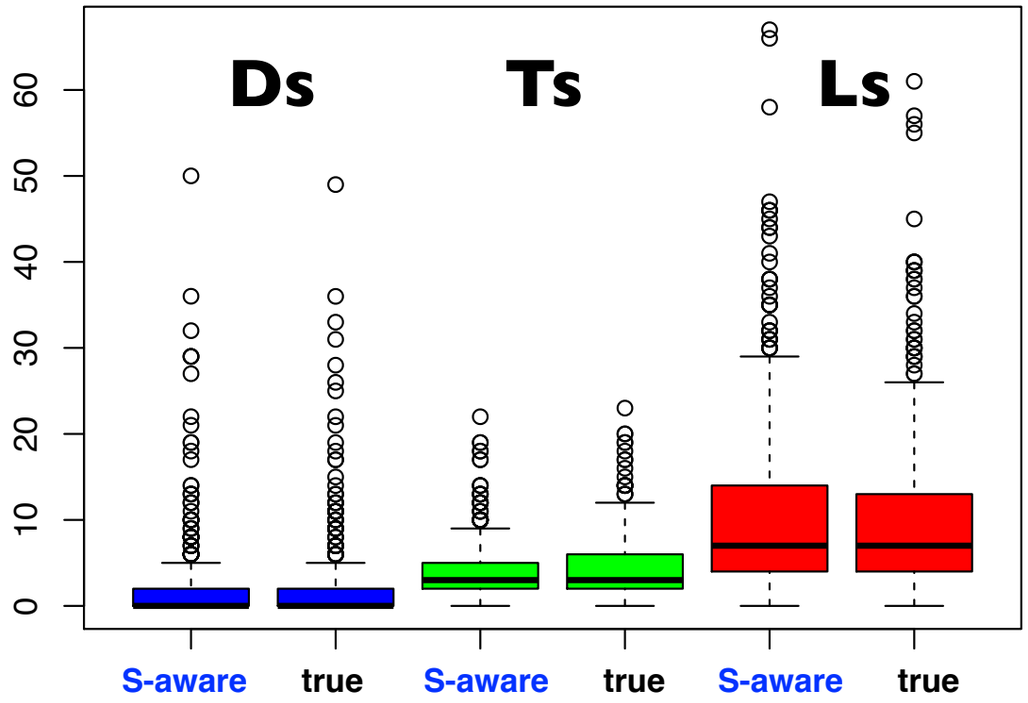
(based on 474 near universal single copy fams.)

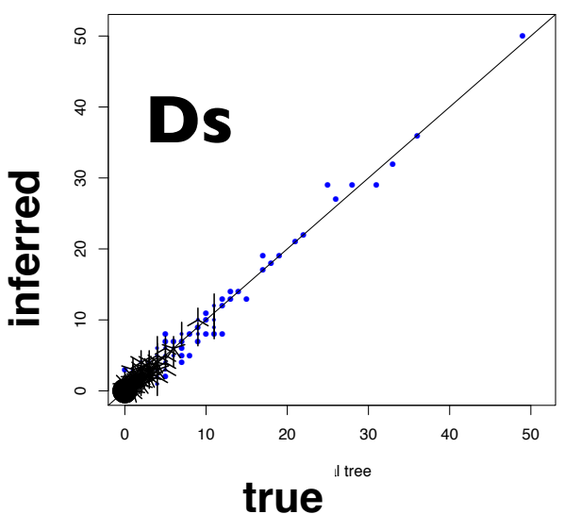


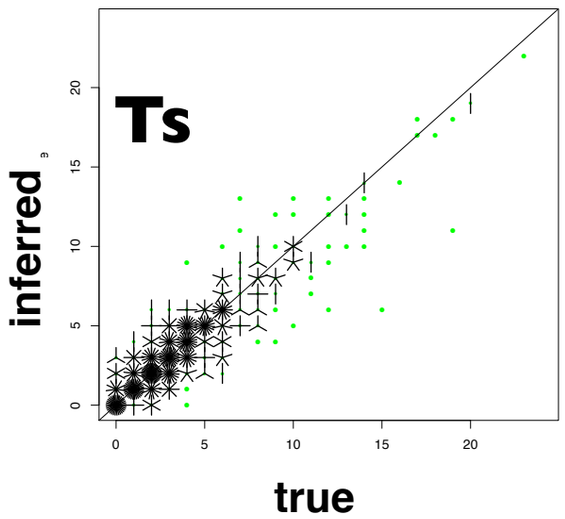

DTL

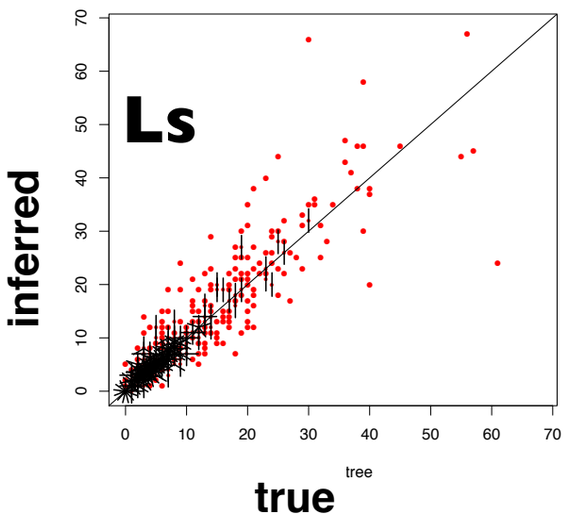


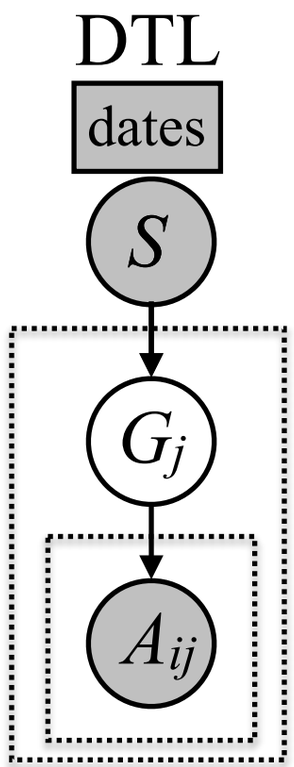
# S-aware methods accurately recover number of DTL events



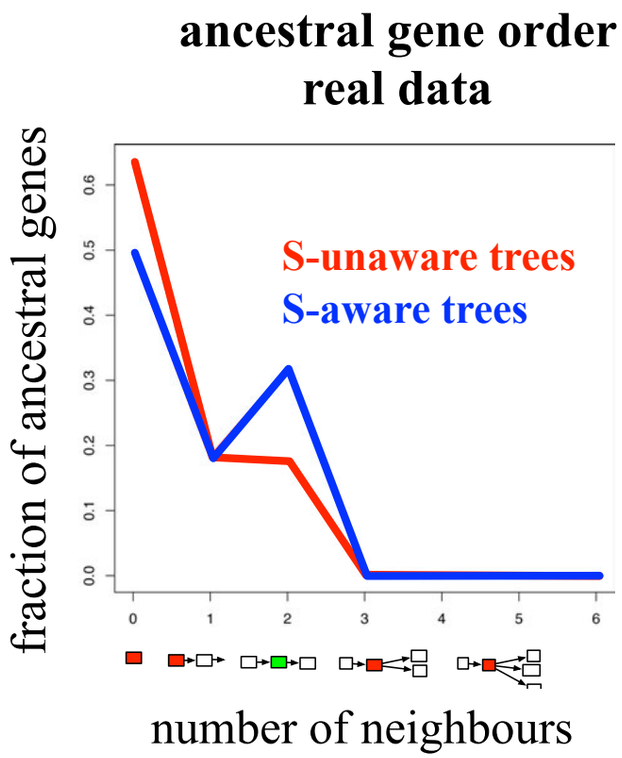



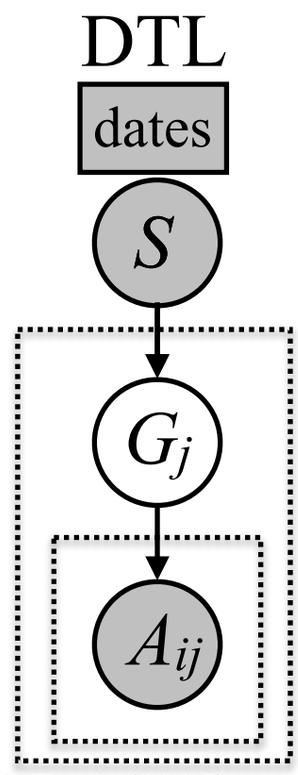



# Real data and experiments suggest S-aware methods are important

More accurate gene trees, ancestral sequences (simulations & experiment) and chromosomes (synteny):




implemented in ALE:

**http://github.com/ssolo/ALE**


Groussin, Hobbs, **Szöllősi**, Gribaldo, Arcus & Gouy ***Mol. Biol. Evol.*** (2015)
*Toward More Accurate Ancestral Protein Genotype–Phenotype Reconstructions with the Use of Species Tree-Aware Gene Trees*

**Szöllősi**, Tannier, Lartillot & Daubin ***Systematic Biology*** (2013)
*Lateral Gene Transfer from the Dead*

**Szöllősi**, Rosikiewicz, Boussau, Tannier & Daubin ***Systematic Biology*** (2013)
*Efficient exploration of the space of reconciled gene trees*

# Real data and experiments suggest S-aware methods are important

More accurate gene trees, ancestral sequences (simulations & experiment) and chromosomes (synteny):

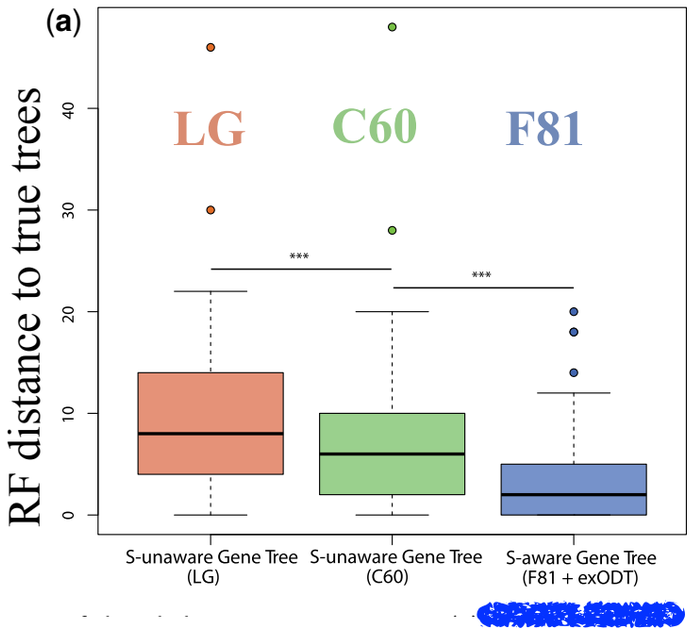
site heterogenous — C60

empirical exchangeabilities — LG

uniform exchangeabilities — F81

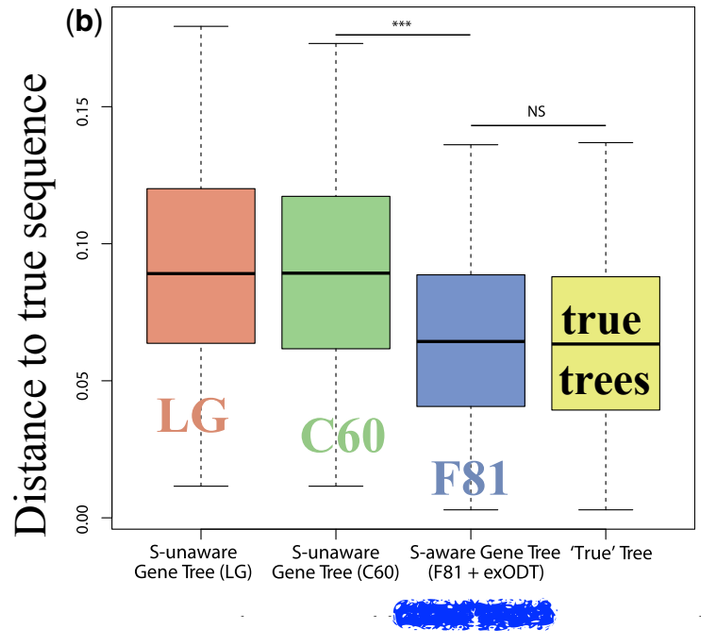
model complexity

**gene tree accuracy**
simulation

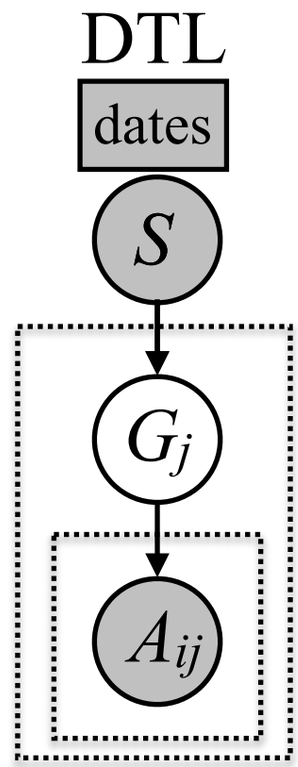



**S-aware trees**

**ancestral sequence accuracy**
simulation




**S-aware ancestral sequences**





implemented in ALE:
**http://github.com/ssolo/ALE**


Groussin, Hobbs, **Szöllősi**, Gribaldo, Arcus & Gouy ***Mol. Biol. Evol.*** (2015)
*Toward More Accurate Ancestral Protein Genotype–Phenotype Reconstructions with the Use of Species Tree-Aware Gene Trees*

**Szöllősi**, Tannier, Lartillot & Daubin ***Systematic Biology*** (2013)
*Lateral Gene Transfer from the Dead*

**Szöllősi**, Rosikiewicz, Boussau, Tannier & Daubin ***Systematic Biology*** (2013)
*Efficient exploration of the space of reconciled gene trees*

# Real data and experiments suggest S-aware methods are important

## *in vitro* biochemical essay

extant LeuB

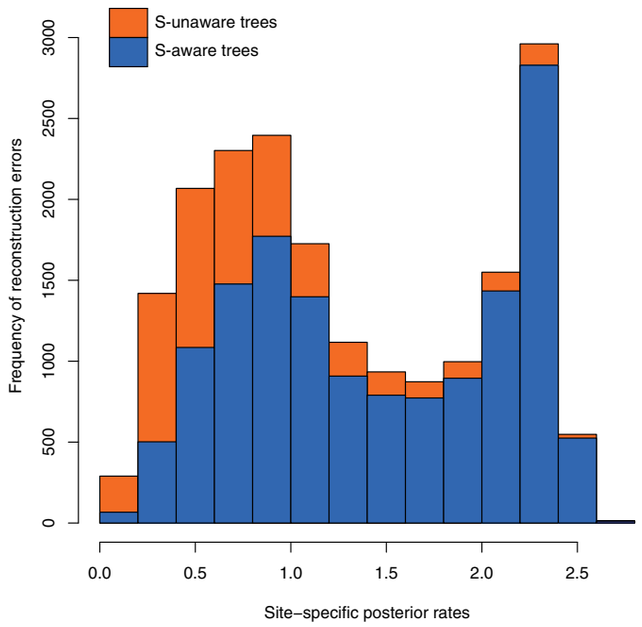
**3-isopropylmalate dehydrogenase E. C. 1.1.1.85**

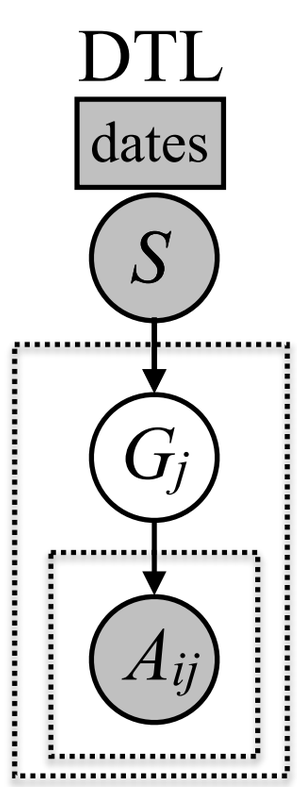
resurrected LeuB

| Enzyme | $K_M^{(IPM)}$ (mM) | $T_{opt}$ ($^\circ C$) | $\Delta G^{\ddagger}_{N-U}$ ($kJmol^{-1}$) |
|---|---|---|---|
| BPSYC | 0.2 | 47 | 94.9 |
| BSUB | 0.7 | 53 | 95.9 |
| BCVX | 1.1 | 69 | 100.7 |
| **S-aware Tree + LG** | **1.5** | **85** | **114.4** |
| **S-aware Tree + EX_EHO** | **1.6** | **85** | **110.9** |
| **S-unaware Tree + EX_EHO** | **6.8** | **78** | **91.4** |

Table 1: **Biophysical parameters for the ancestral LeuB enzyme of the Firmicutes ancestor.** Values obtained in this study for the ancestor of Firmicutes (bold characters) were inferred using either the LeuB sequence tree or the LeuB reconciled tree and either with the site-homogeneous LG model or with the site-heterogeneous EX_EHO model. Data for contemporary (first three lines) are shown for comparison)

## Why and how?





implemented in ALE:
**http://github.com/ssolo/ALE**

Groussin, Hobbs, **Szöllősi**, Gribaldo, Arcus & Gouy *Mol. Biol. Evol.* *Toward More Accurate Ancestral Protein Genotype–Phenotype Reconstructions with the Use of Species Tree-Aware Gene Trees* (2015)

# Real data and experiments suggest S-aware methods are important

## *in vitro* biochemical essay

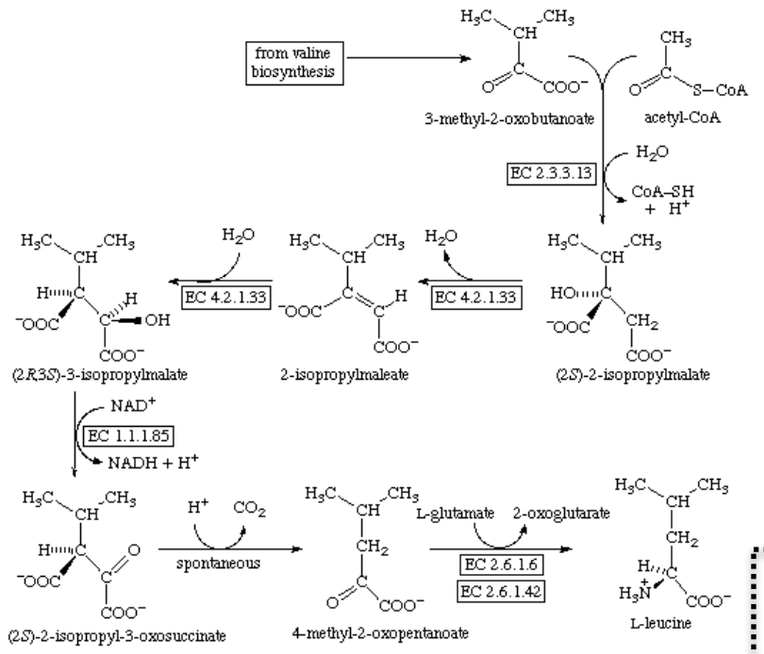
extant LeuB
**3-isopropylmalate dehydrogenase E. C. 1.1.1.85**

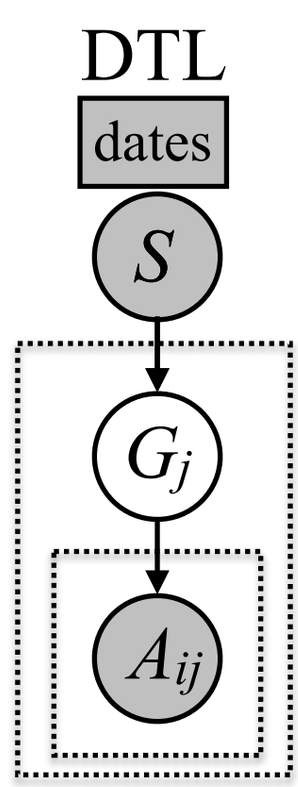
resurrected LeuB

| Enzyme | $K_M^{(IPM)}$ (mM) | $T_{opt}$ ($^\circ C$) | $\Delta G^{\ddagger}_{N-U}$ ($kJmol^{-1}$) |
|---|---|---|---|
| BPSYC | 0.2 | 47 | 94.9 |
| BSUB | 0.7 | 53 | 95.9 |
| BCVX | 1.1 | 69 | 100.7 |
| **S-aware Tree + LG** | **1.5** | **85** | **114.4** |
| **S-aware Tree + EX_EHO** | **1.6** | **85** | **110.9** |
| **S-unaware Tree + EX_EHO** | **6.8** | **78** | **91.4** |

Table 1: **Biophysical parameters for the ancestral LeuB enzyme of the Firmicutes ancestor.** Values obtained in this study for the ancestor of Firmicutes (bold characters) were inferred using either the LeuB sequence tree or the LeuB reconciled tree and either with the site-homogeneous LG model or with the site-heterogeneous EX_EHO model. Data for contemporary (first three lines) are shown for comparison)

## Leucine Biosynthesis



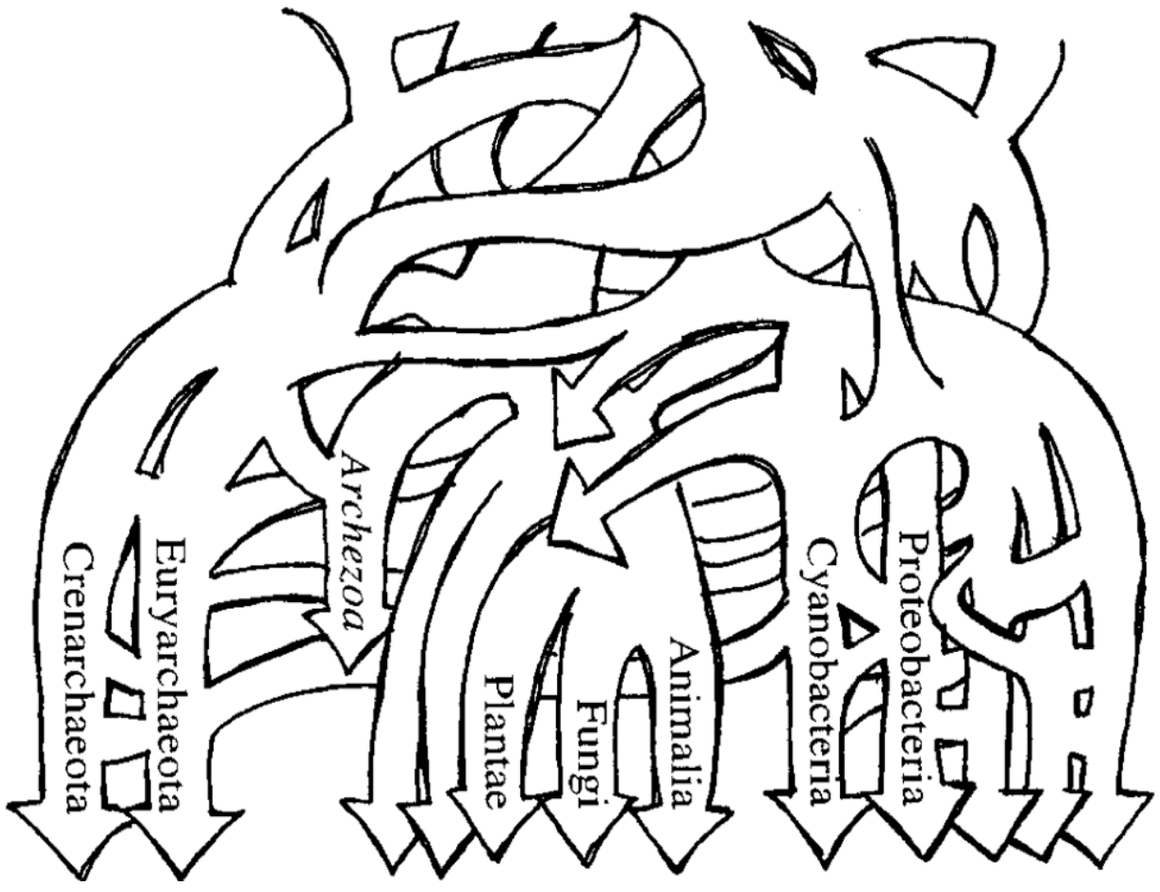





implemented in ALE:
http://github.com/ssolo/ALE

Groussin, Hobbs, **Szöllősi**, Gribaldo, Arcus & Gouy *Mol. Biol. Evol.* *Toward More Accurate Ancestral Protein Genotype–Phenotype Reconstructions with the Use of Species Tree-Aware Gene Trees* (2015)

**just how much HGT is there?**

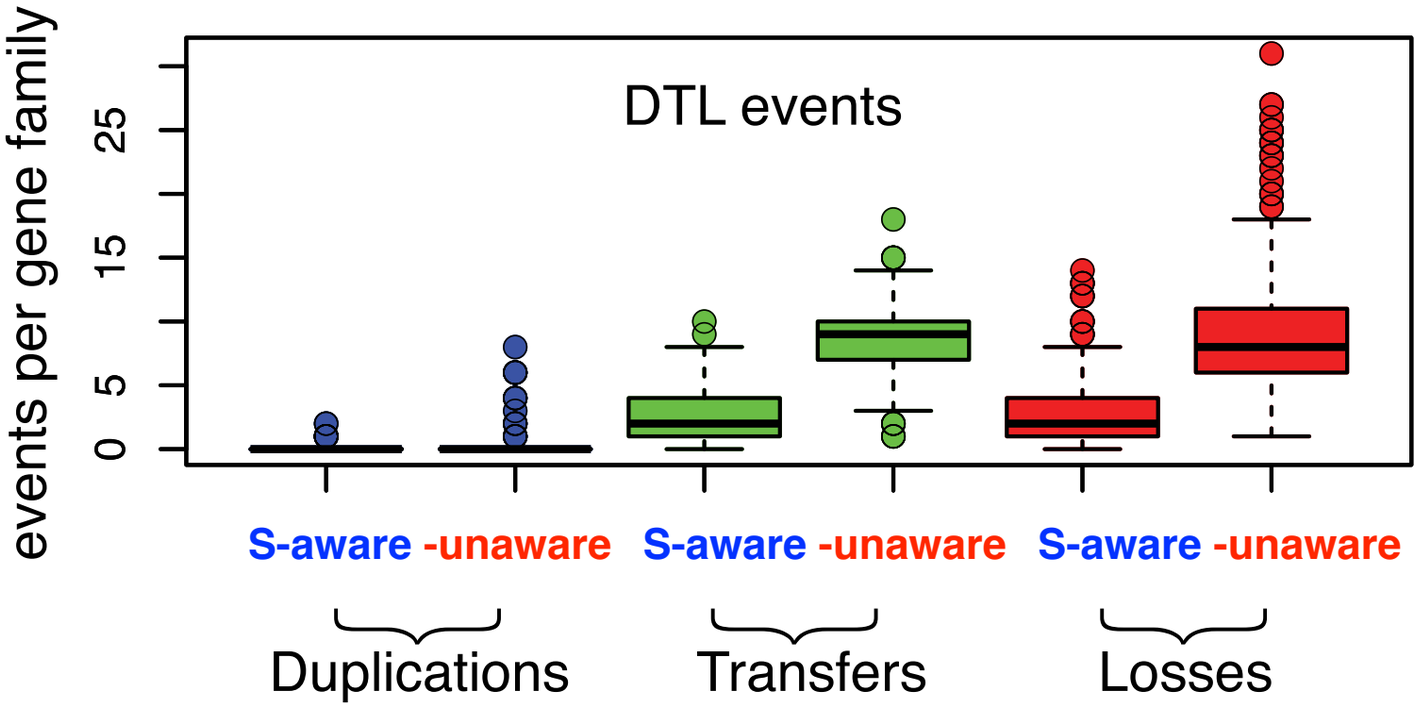


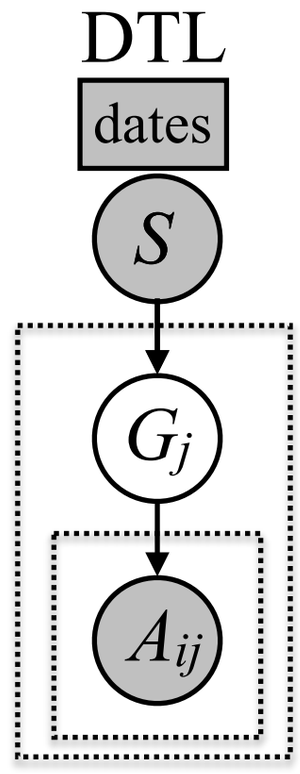

**Doolittle** 1999

# just how much HGT is there?

**Two out of three transfers** inferred based S-unaware gene trees **are the result phylogenetic errors.**

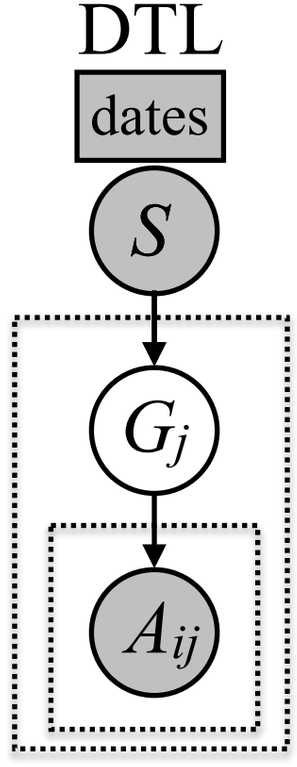
Two out of three losses inferred based S-unaware gene trees are the result phylogenetic errors.




**Szöllősi**, Rosikiewicz, Boussau, Tannier & Daubin ***Systematic Biology*** (2013)
*Efficient exploration of the space of reconciled gene trees*

# what about ancestral gene content?

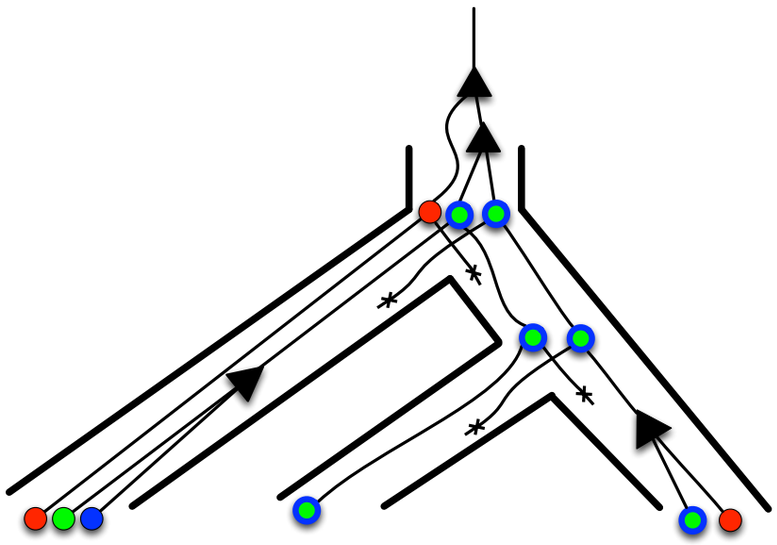



**Szöllősi**, Rosikiewicz, Boussau, Tannier & Daubin ***Systematic Biology*** (2013)
*Efficient exploration of the space of reconciled gene trees*

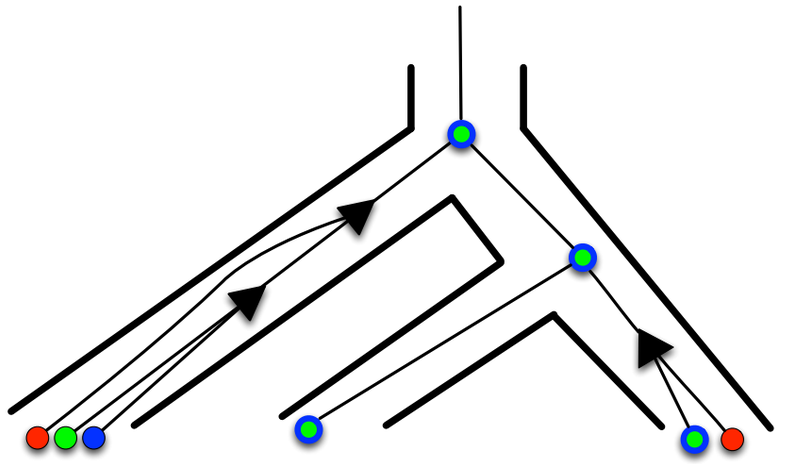
# what about ancestral gene content?

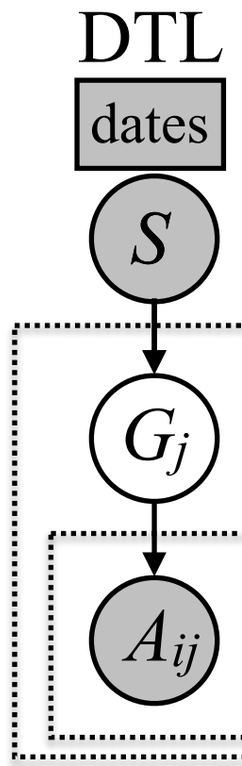
Spurious (**false positives**) Ds lead to an **overestimation** of ancestral genome contents, missing Ds (**false negatives**) lead to an **underestimation**.

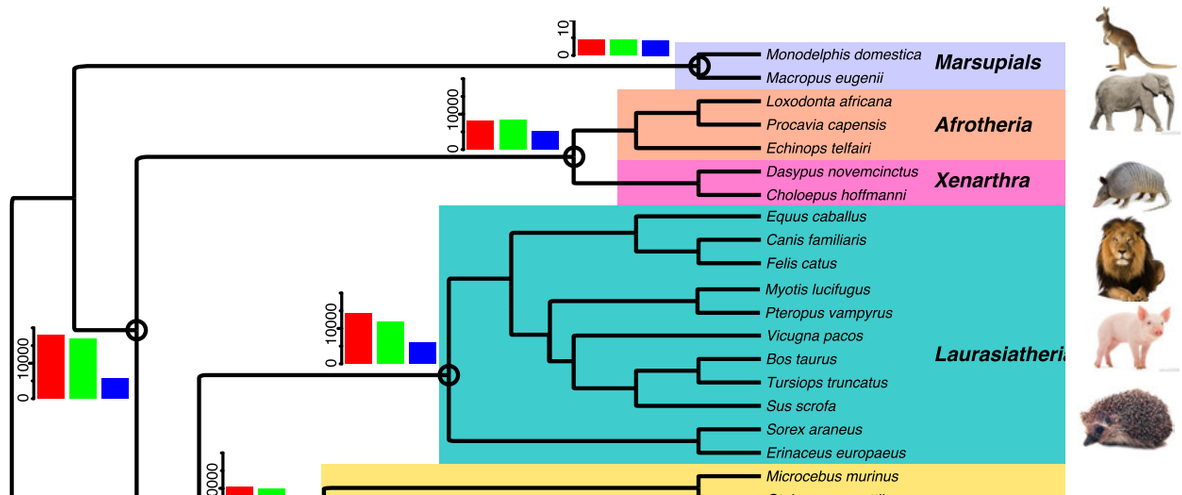
## gene tree with errors



## correct gene tree

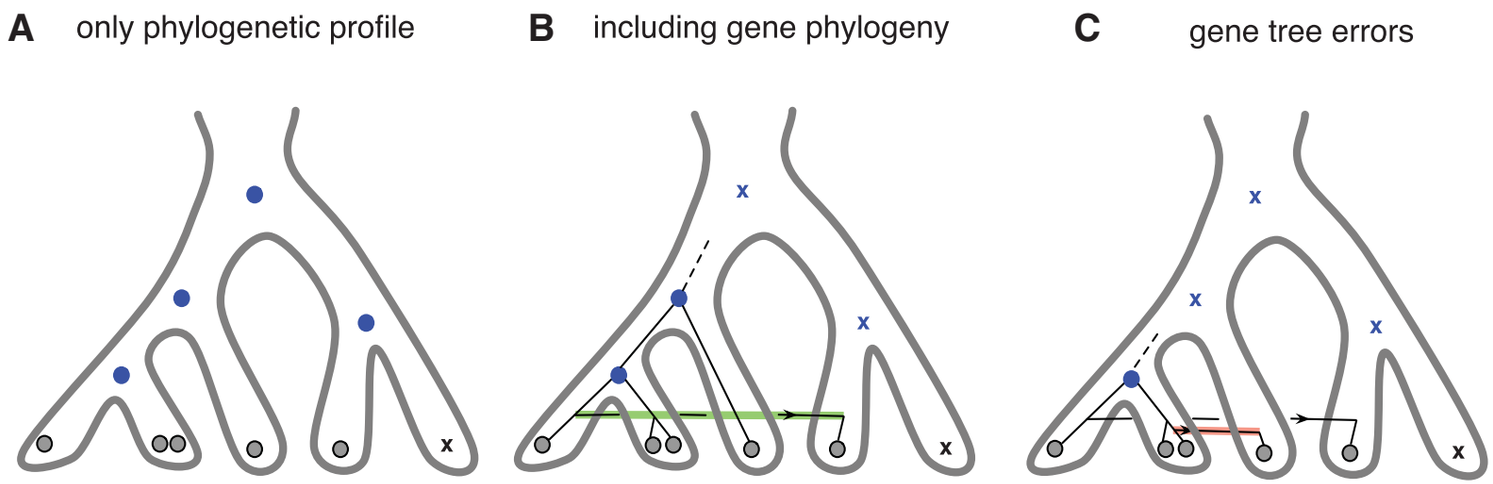




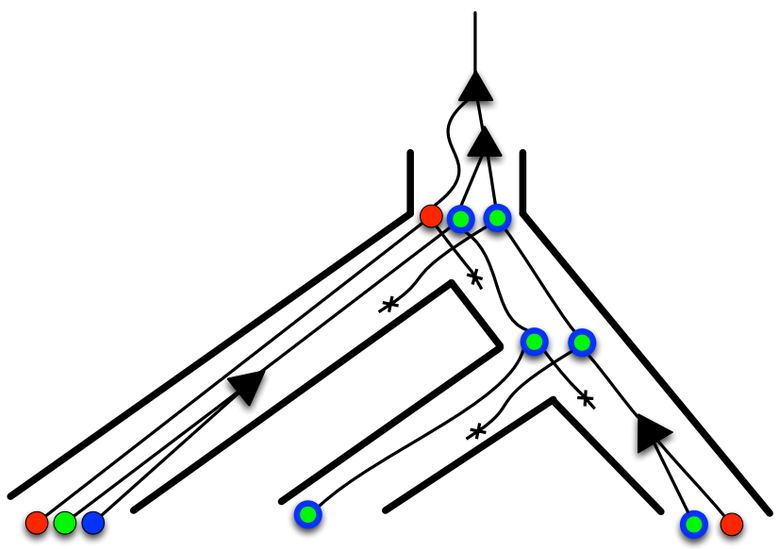



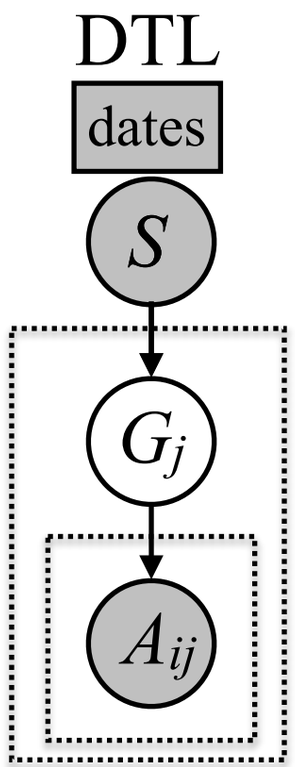
# what about ancestral gene content?

Spurious (**false positives**) Ts lead to an **underestimation** of ancestral genome contents, missing Ts (**false negatives**) lead to an **overestimation** at deep nodes.




Ignoring transfer, i.e. **considering only DL also leads to an overestimate**.





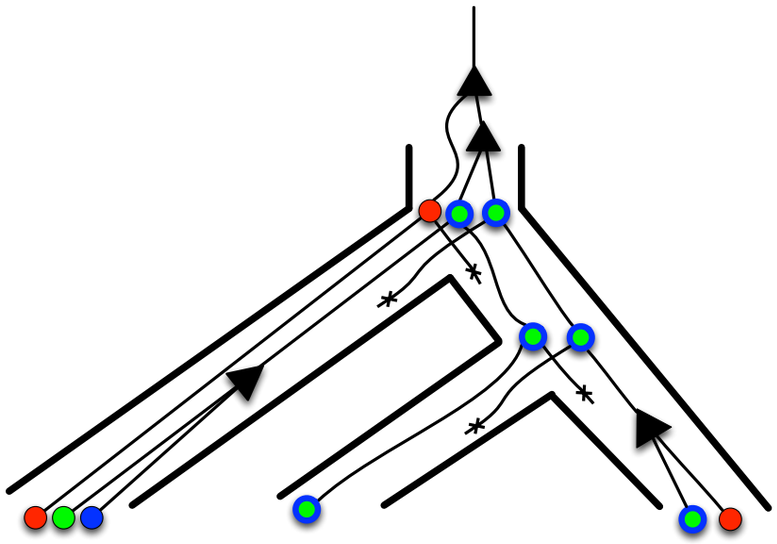
# what is the effect of spurious HGTs on ancestral gene content?

Spurious (**false positives**) Ds lead to an **overestimation** of ancestral genome contents, missing Ds (**false negatives**) lead to an **underestimation**.

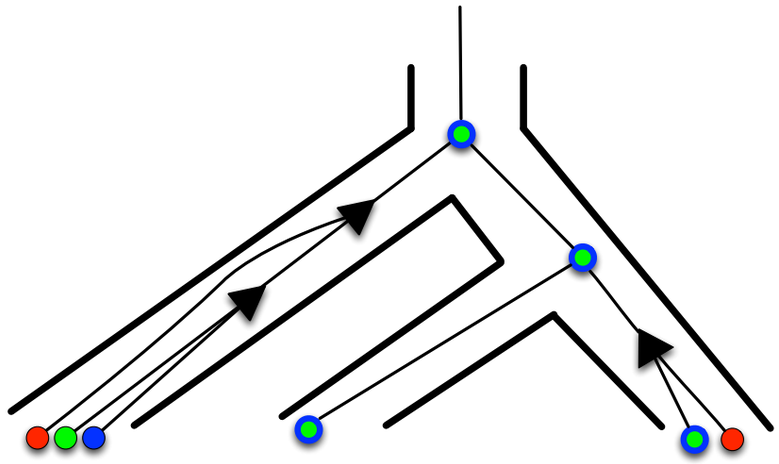
But, ignoring transfer, i.e. considering only DL also leads to an overestimate!

## gene tree with errors

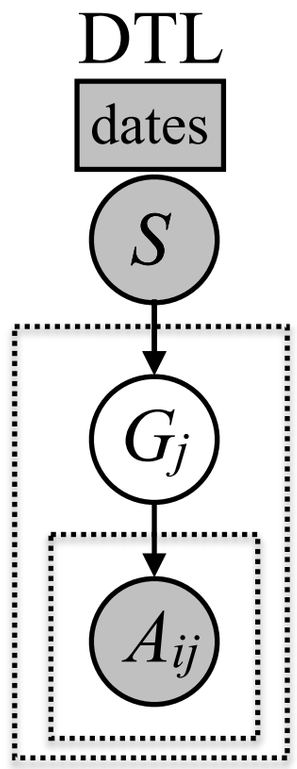
gene tree with "unmodelled" transfers



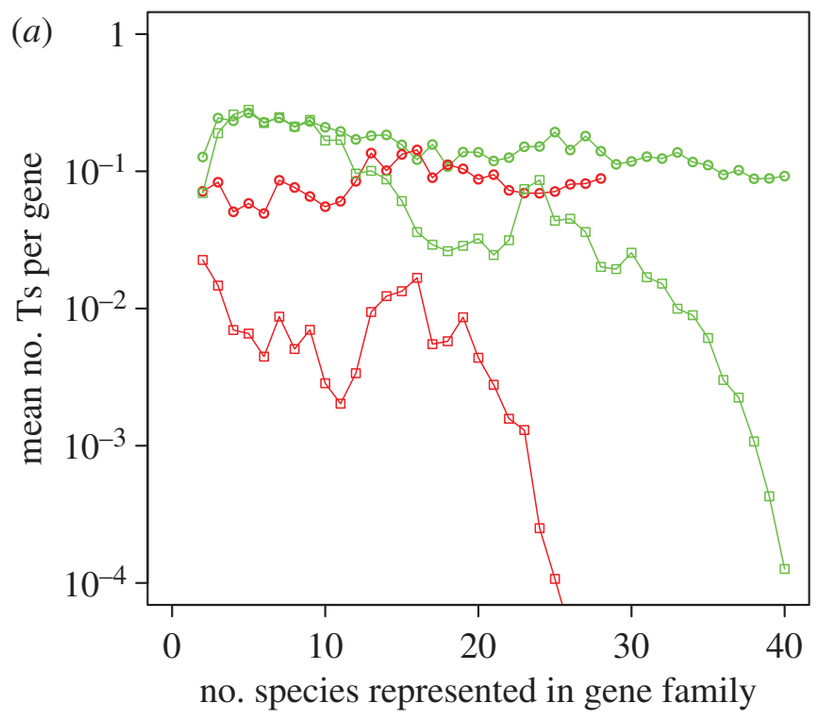
## correct gene tree

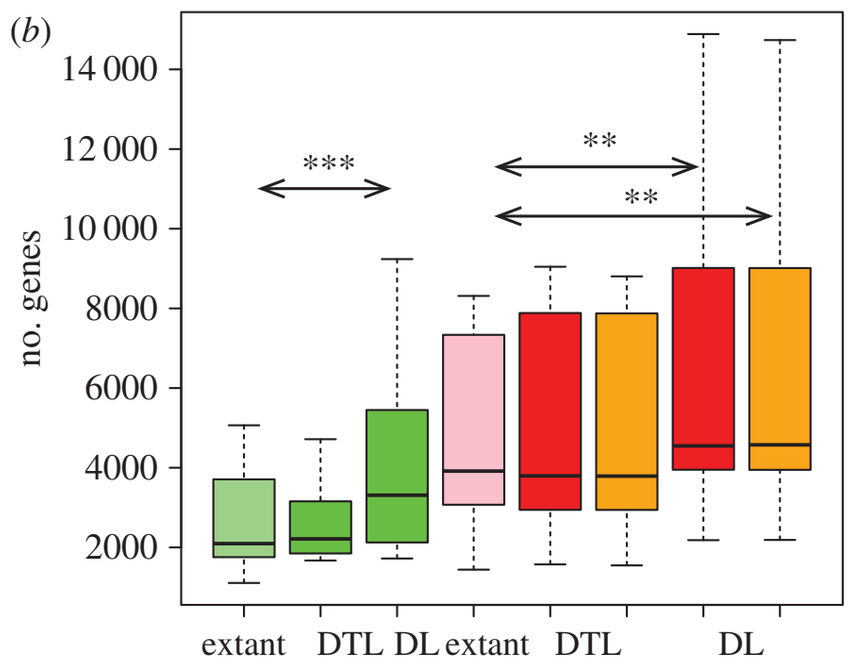


DTL

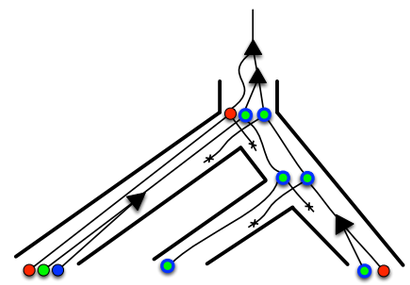


# Is HGT frequent only in prokaryotes?

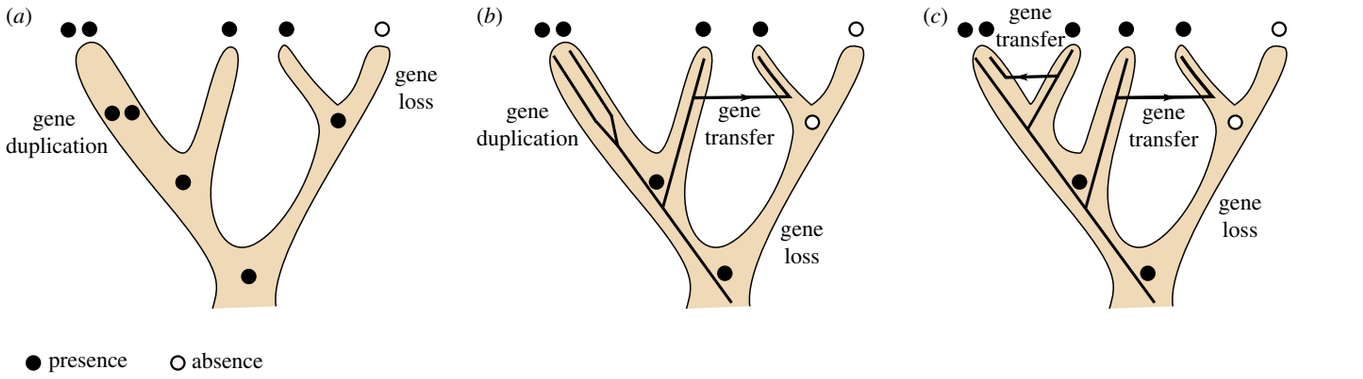

"unmodelled" transfers

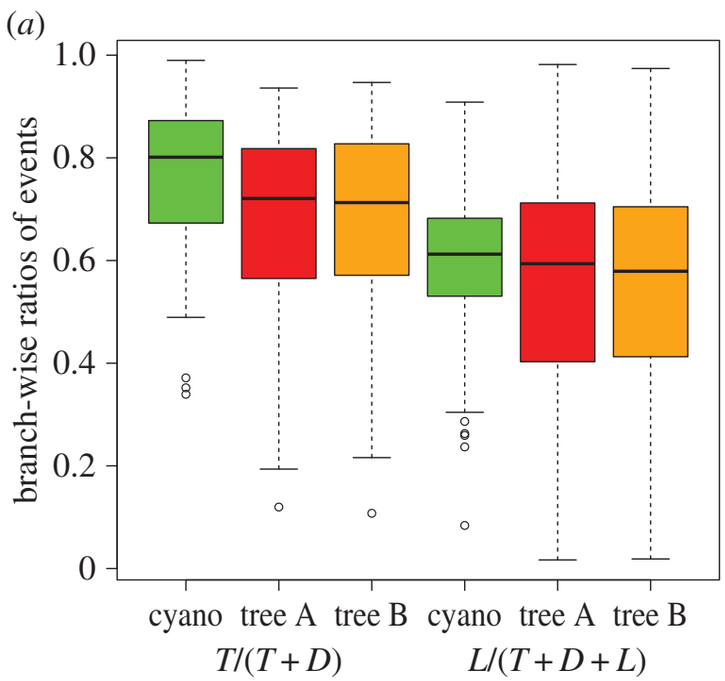


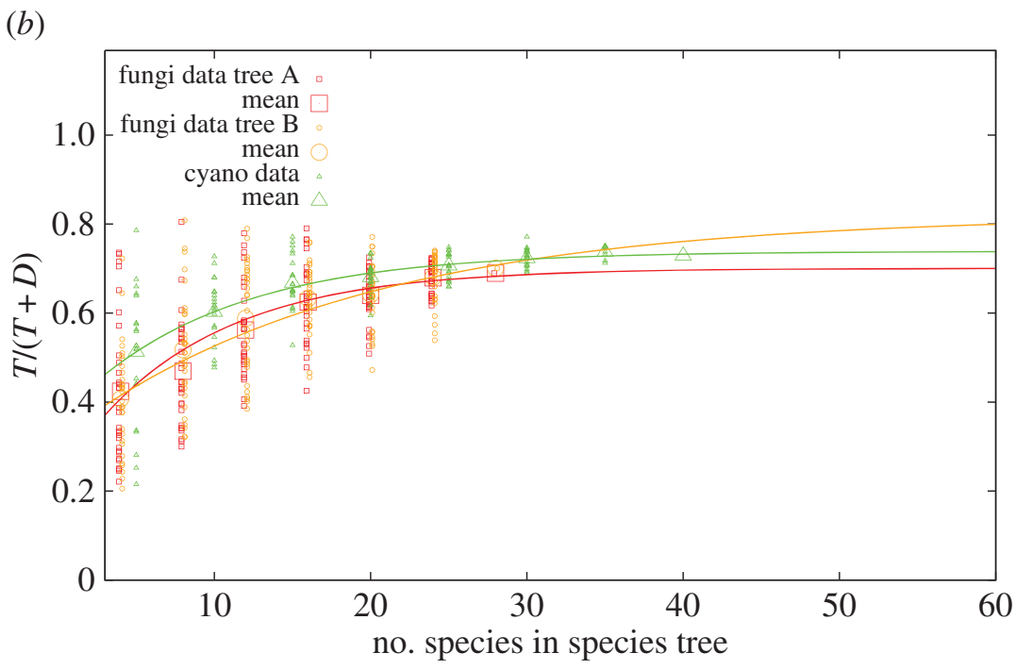
Ignoring transfer, i.e. considering only DL leads to an overestimate..

**Szöllősi**, Davin, Tannier, Daubin & Boussau *Phil.Trans.Roy.Soc.B* (2015)
*Genome-scale phylogenetic analysis finds extensive gene transfer among fungi.*

# There is extensive gene transfer among fungi



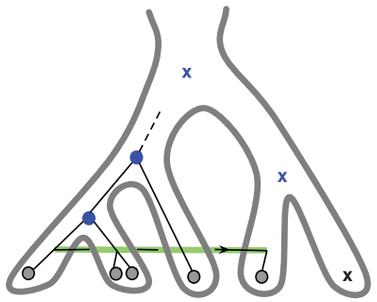




**Szöllősi**, Davin, Tannier, Daubin & Boussau *Phil.Trans.Roy.Soc.B* (2015)
*Genome-scale phylogenetic analysis finds extensive gene transfer among fungi.*

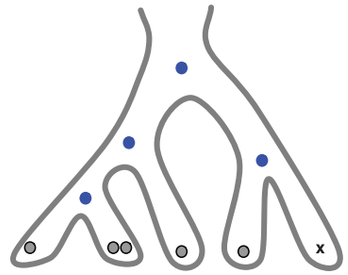
# There is extensive gene transfer among fungi

**B** including gene phylogeny

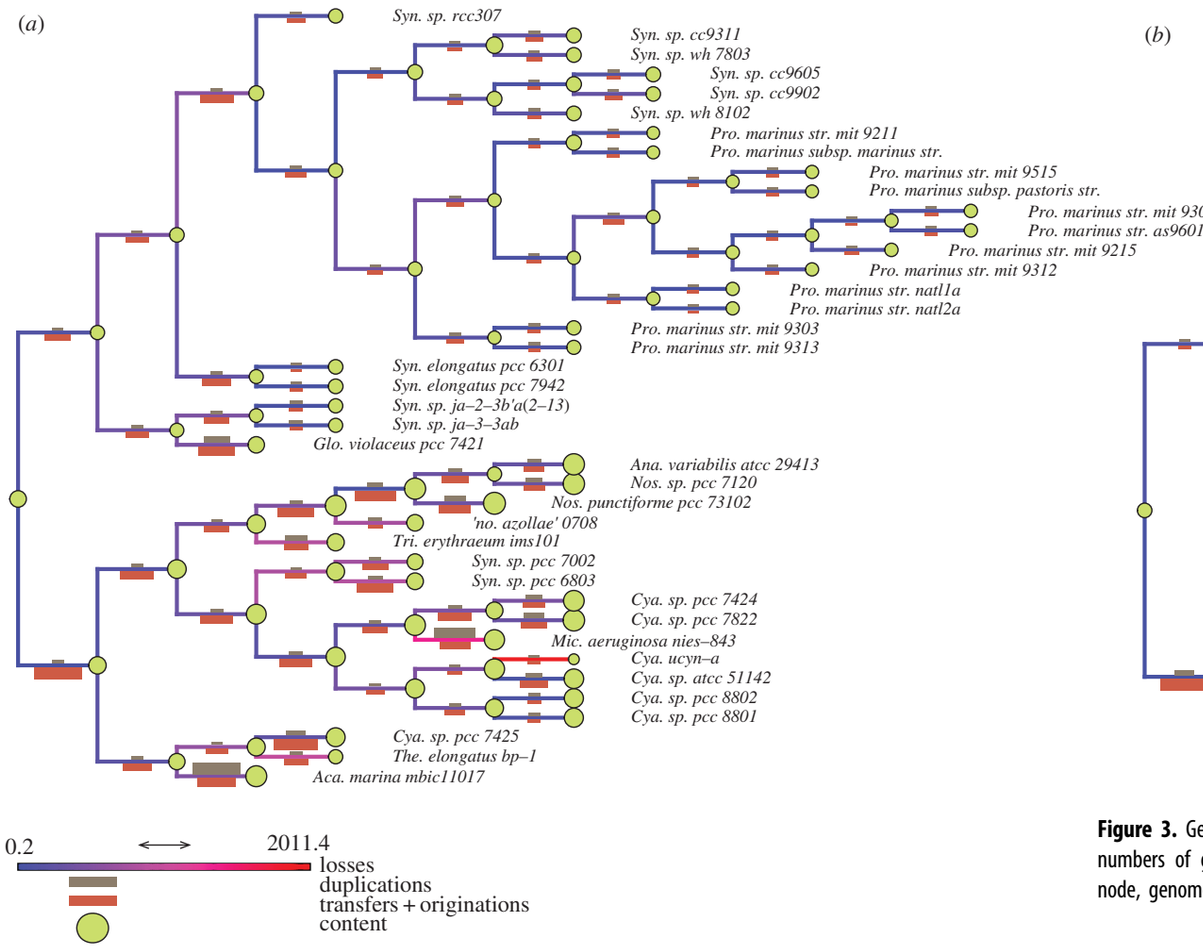



## 28 fungi genomes

**A** only phylogenetic profile




(*a*)




(*b*)

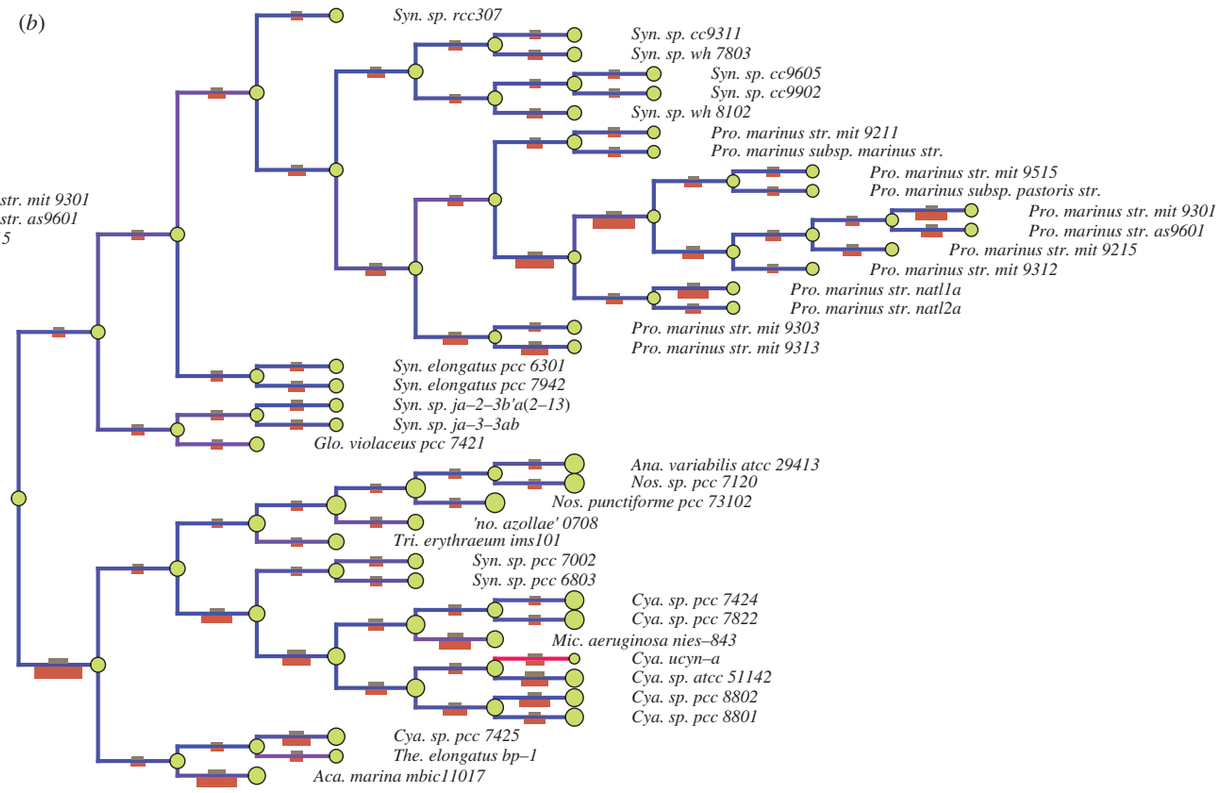



**Figure 3.** Genome evolution in cyanobacteria. Edges are colour-coded according to the inferred numbers of losses along the branches. Crimson bars represent numbers of gene gains (transfers + originations) arriving on the branch; taupe bars represent numbers of duplications happening on the branch. At each node, genome content size is represented as a green disc. (*a*) Inferences from *ALEml_undated*. (*b*) Inferences from *Count*.

**Szöllősi**, Davin, Tannier, Daubin & Boussau *Phil.Trans.Roy.Soc.B* (2015)
*Genome-scale phylogenetic analysis finds extensive gene transfer among fungi.*

# Stronger transfer highways in fungi

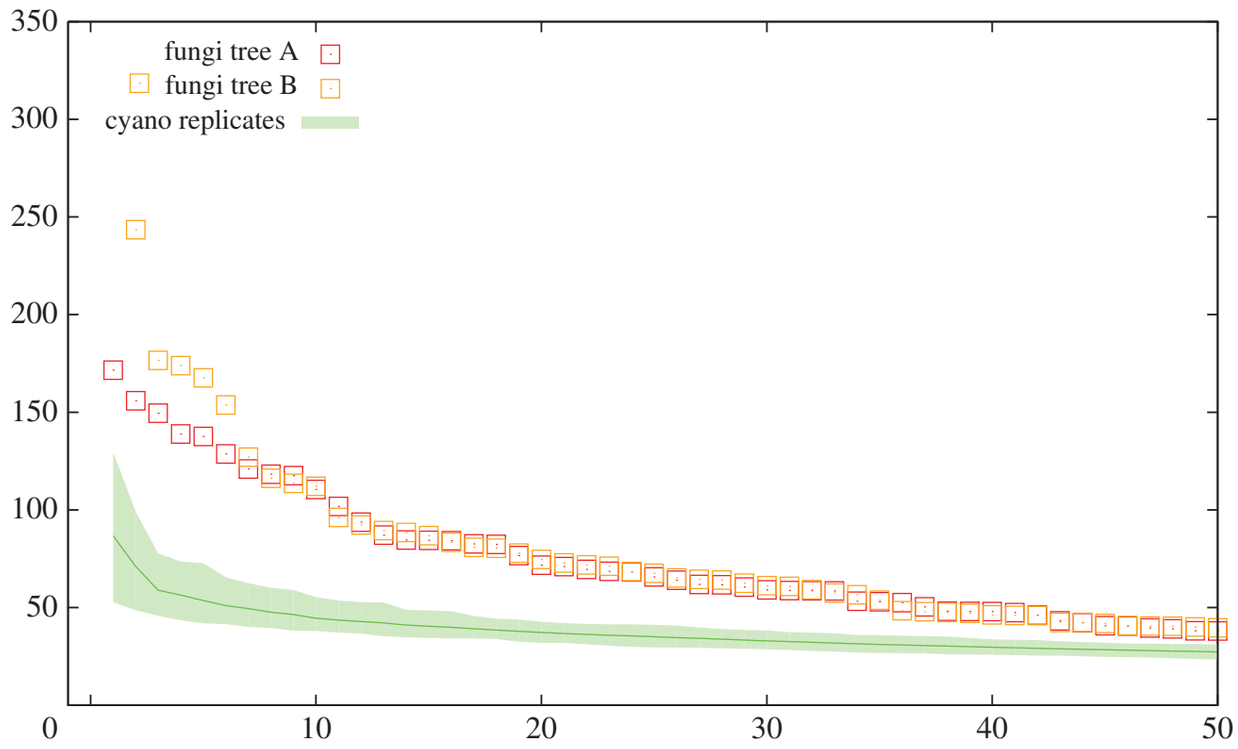


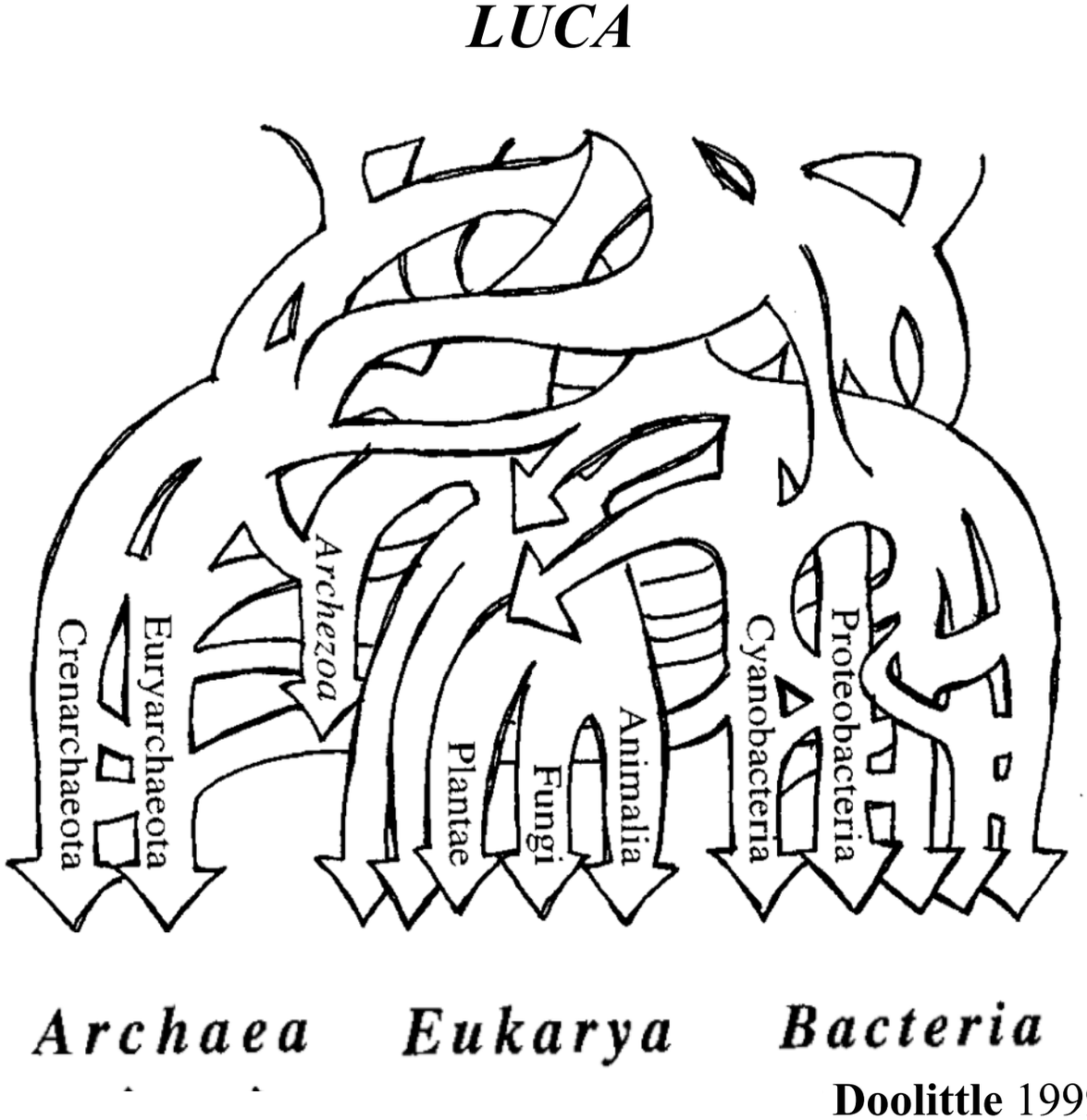

**Figure 6.** Stronger highways in fungi. Data points, red for tree A and orange for tree B, correspond to numbers of transfers between pairs of branches ('highways of transfers') in either fungi phylogeny plotted in decreasing order. The continuous line (green online) shows the mean number of transfers between pairs of branches among 25 replicates where a random set of 28 cyanobacterial genomes were chosen as in figure 5*b*. The shaded area shows the 95% CI. Fungi are in red and cyanobacteria in green throughout.

**Szöllősi**, Davin, Tannier, Daubin & Boussau *Phil.Trans.Roy.Soc.B* (2015)
*Genome-scale phylogenetic analysis finds extensive gene transfer among fungi.*

ssolo@elte.hu

# 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.

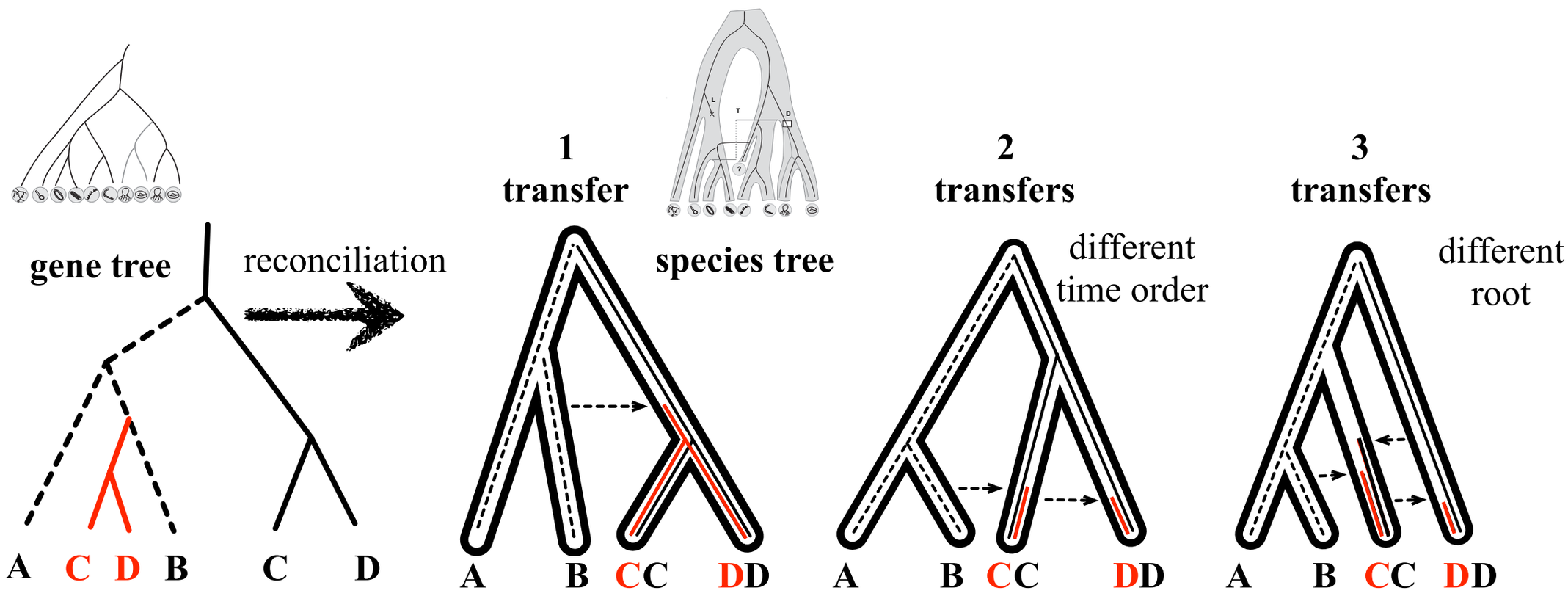



**Doolittle** 1999

ssolo@elte.hu

# Horizontal gene transfer as information

Transfer events, encoded in the topologies of gene trees can be thought of as "*molecular fossils*" that record the order of speciation events.




Vincent Daubin

LBBE

**Szöllősi**, Boussau, Abby, Tannier & Daubin ***PNAS*** (2012)
*Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations*

# 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

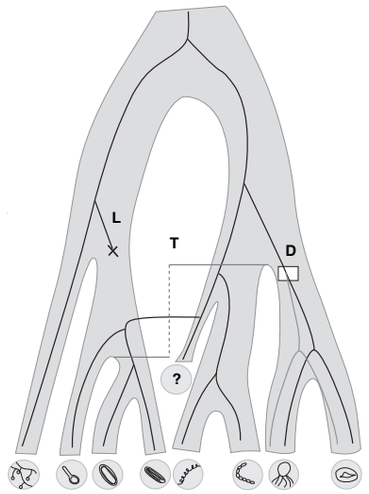
$$\prod_j$$

optimise $S$
and estimate rates

**clients:**
calculate

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

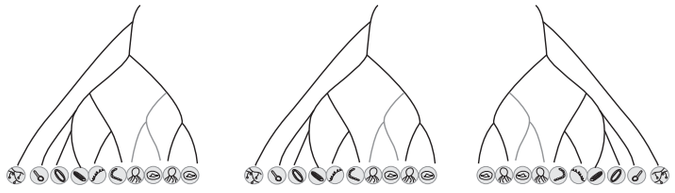
optimise (or integrate over) $G_j$




**species tree**

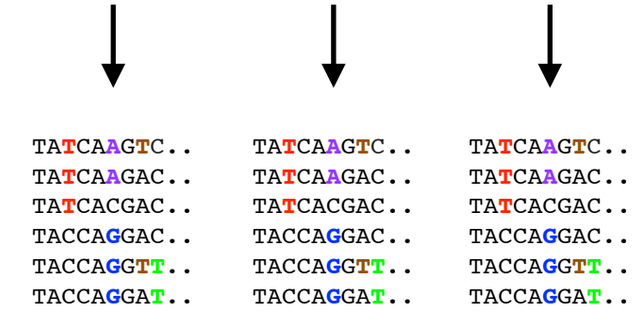
$S$

**Daubin & Boussau** 2011



$G_j$

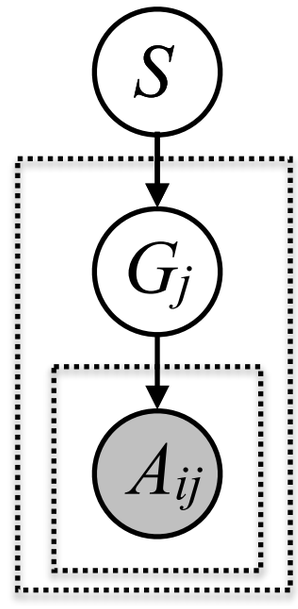
**gene trees**




$A_{ij}$

**sequences**

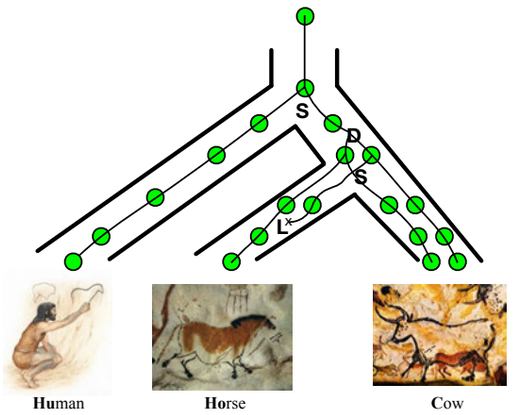
DL




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

# .. but gene trees are generated along the species tree

If we model the process generating gene trees along the species tree we can hope to infer better gene trees and species trees. To calculate the likelihood of a ***gene tree*** we sum over all possible ***gene birth and death events*** along a given ***species tree.***




**calculation complexity**

DL

2 to 10 × log(#species) × #genes

DTL

2 to 10 × #species$^2$ × #genes

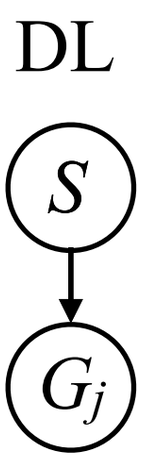
**parameters (ML or Bayes)**

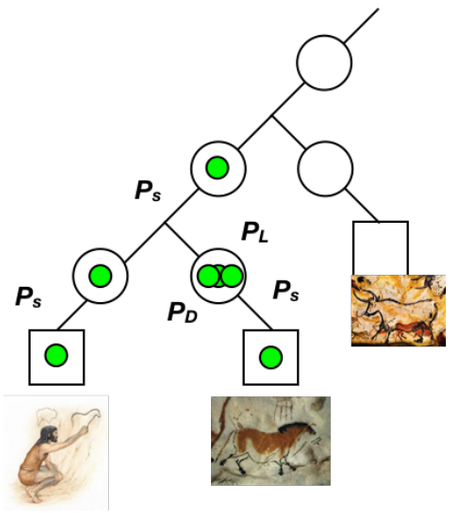
DL

D&L rates
branch lengths, root

DTL

D,T&L rates
dated tree




DL

log(#species) × #genes

"undated"
DTL

#species × log(#species) × #genes

DL

D&L rates, root

"undated"
DTL

D,T&L rates, root

ssolo@elte.hu

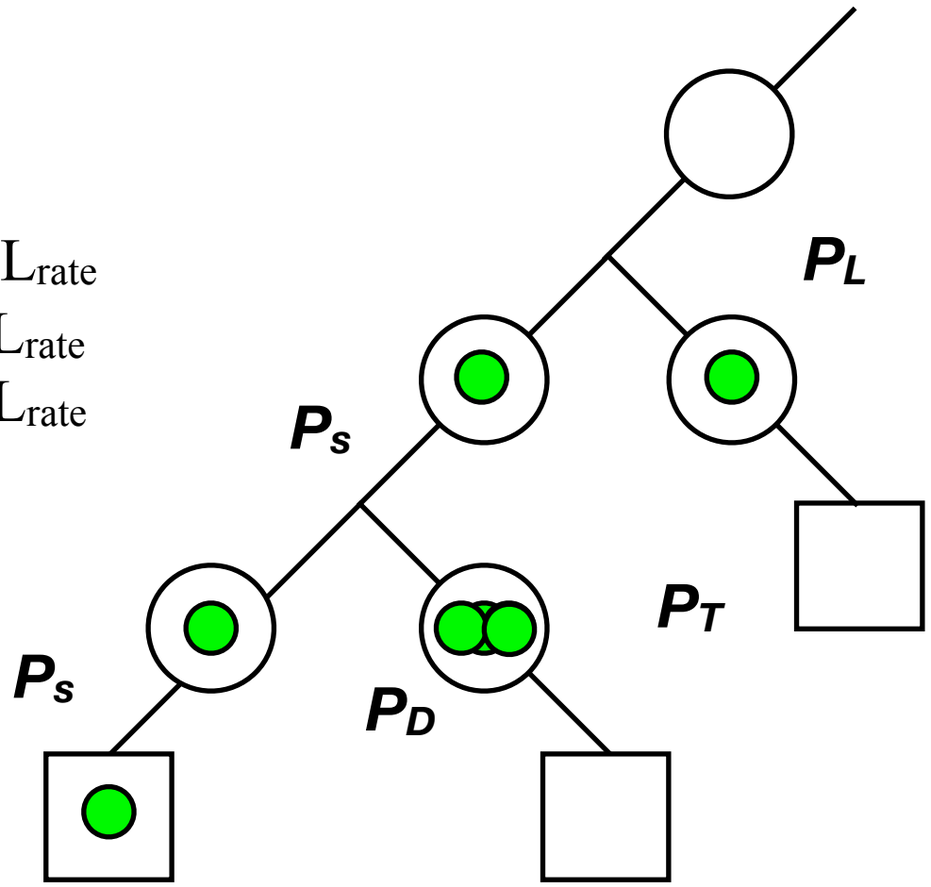
# “undated” DTL

$$P_S + P_D + P_T + P_L = 1$$

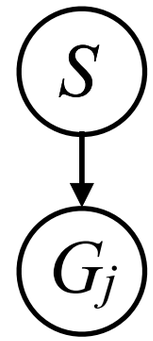
$$P_D = D_{rate} / (D_{rate} + T_{rate} + L_{rate})$$

$$P_T = T_{rate} / (D_{rate} + T_{rate} + L_{rate})$$

$$P_L = L_{rate} / (D_{rate} + T_{rate} + L_{rate})$$



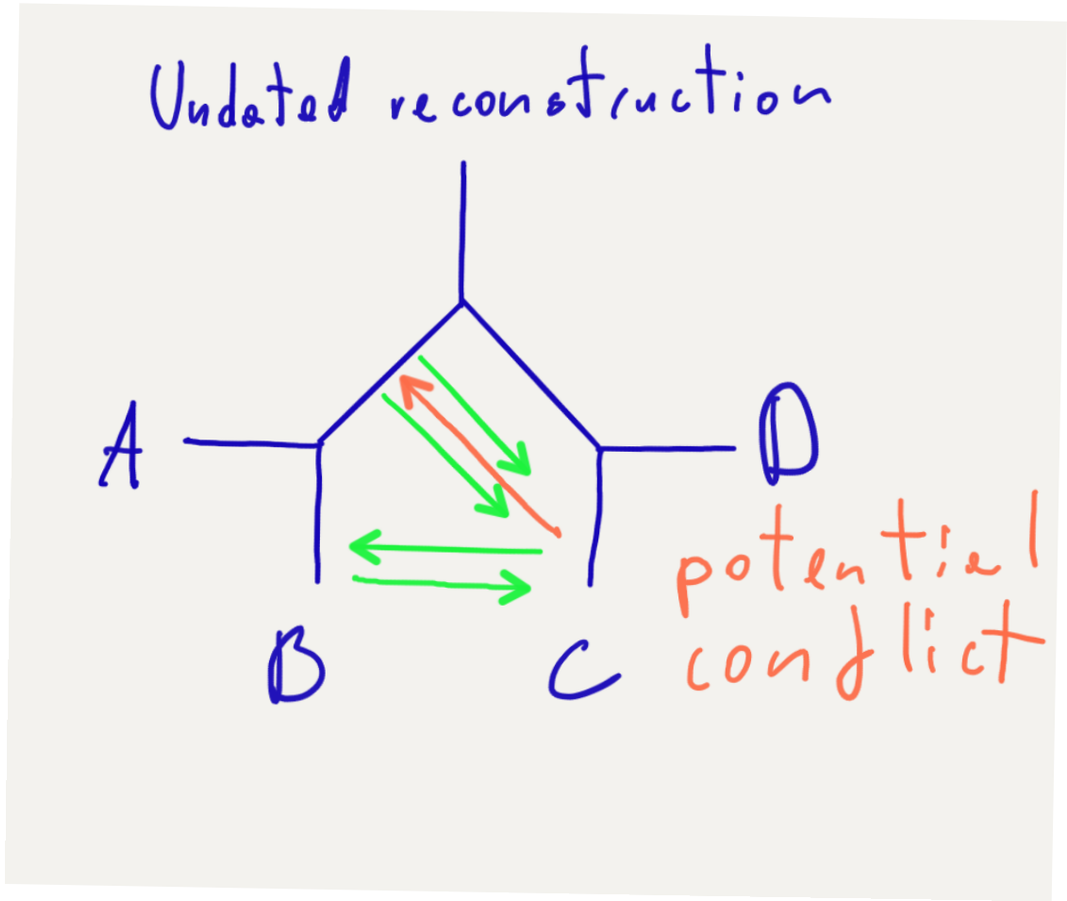

**undated** DTL



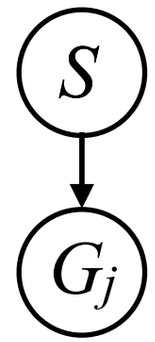

implemented in ALE:

**http://github.com/ssolo/ALE**

“undated”
DTL

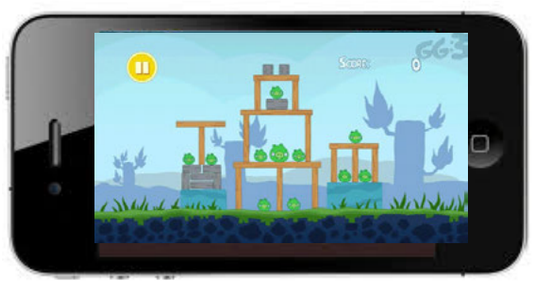



**undated**
DTL




DTL

“undated” DTL



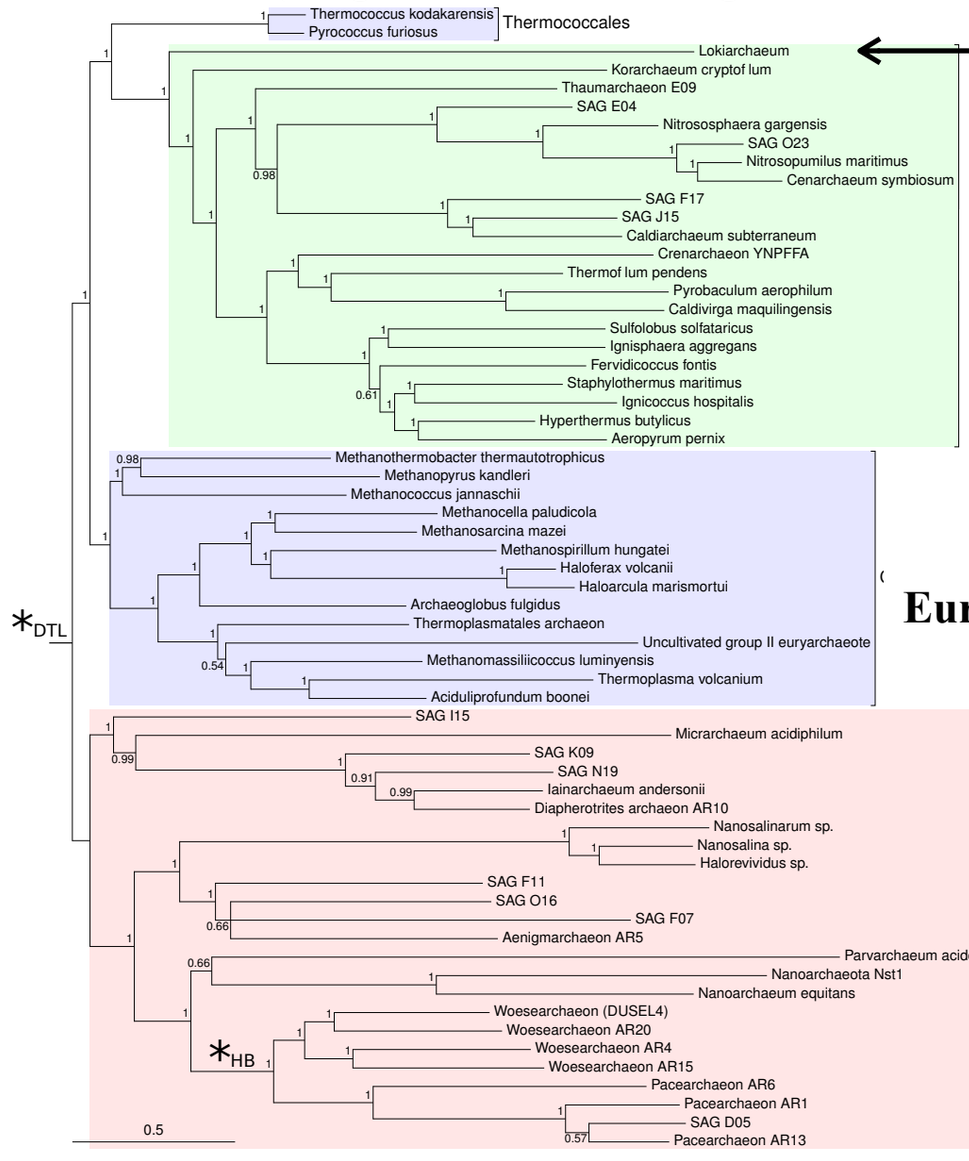
implemented in ALE:

**http://github.com/ssolo/ALE**

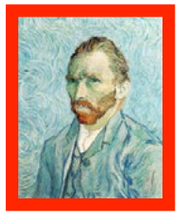
# Transfers can root the archaeal tree of life

Using ALE on gene families from 60 genomes allowed us to **root the Archaeal tree without an out-group** while at the same time reconstructing ancestral gene contents.

**Archaea** (60 genomes, including recent DPANN and Lokiarchaeum)




**TACKL**

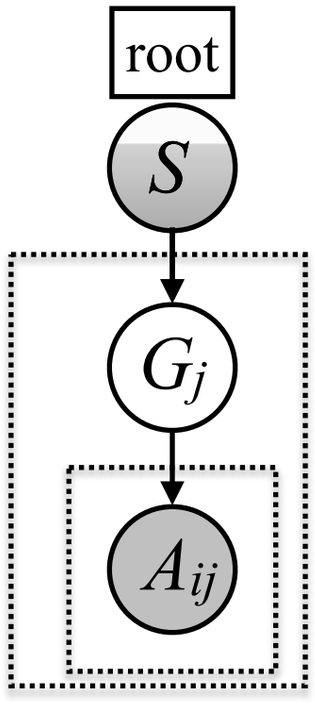



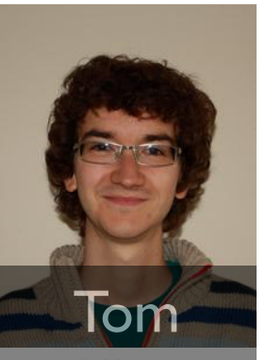
**core Euryarchaeota**

**DPANN**

**undated**

DTL



Tom Williams

U.Bristol

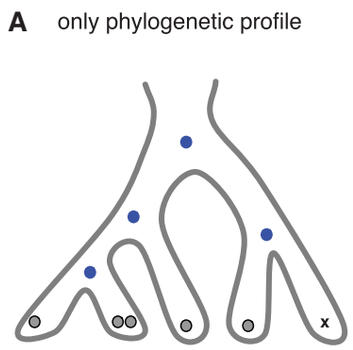
**The predominant mode of genome evolution is not gene loss, but ongoing lineage specific innovation.**

Williams, **Szöllősi**, .. & Embley ***PNAS*** (2017)

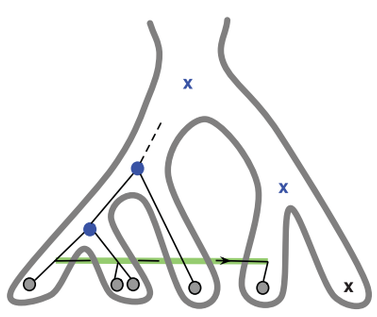
*Integrative modelling of gene and genome evolution roots the archaeal tree of life*

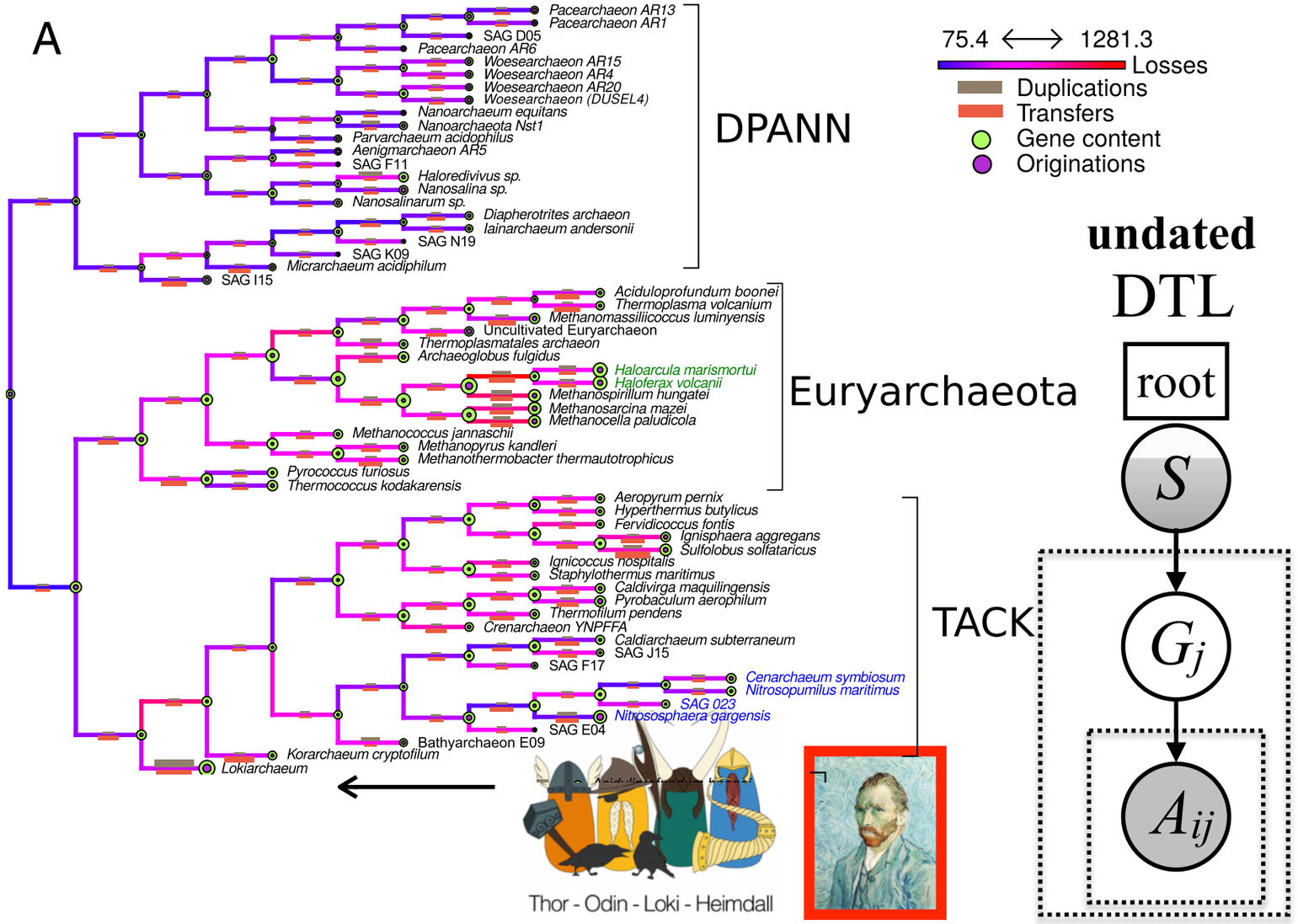
# Transfers can root the archaeal tree of life

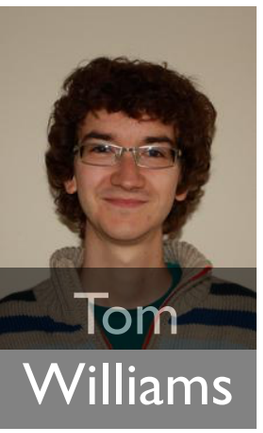
Using ALE on gene families from 60 genomes allowed us to **root the Archaeal tree without an out-group** while at the same time reconstructing ancestral gene contents.




**Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model.**
COUNT Csürös & Miklós 2009 ***MBE***




implemented in ALE:

**http://github.com/ssolo/ALE**

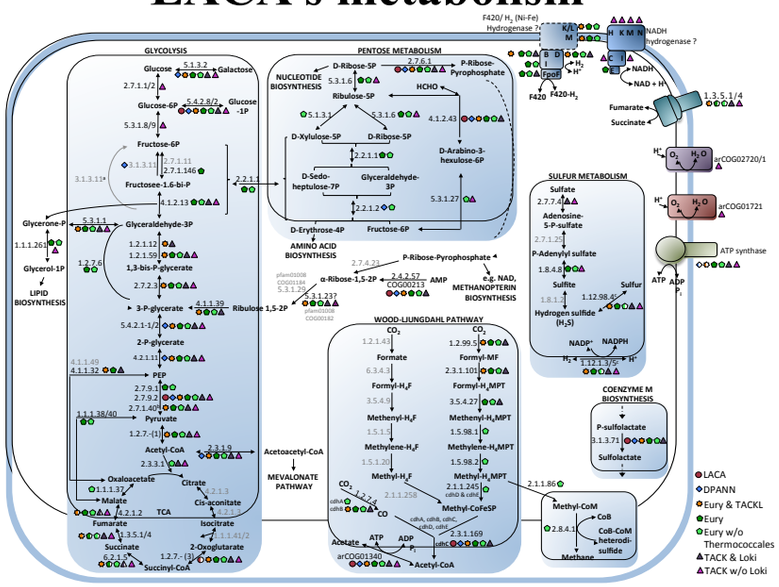
**The predominant mode of genome evolution is not gene loss, but ongoing lineage specific innovation.**

Williams, **Szöllősi**, .. & Embley ***PNAS*** (2017)
*Integrative modelling of gene and genome evolution roots the archaeal tree of life*

# Transfers can root the archaeal tree of life

Using ALE on gene families from 60 genomes allowed us to **root the Archaeal tree without an out-group** while at the same time reconstructing ancestral gene contents.

## LACA's metabolism

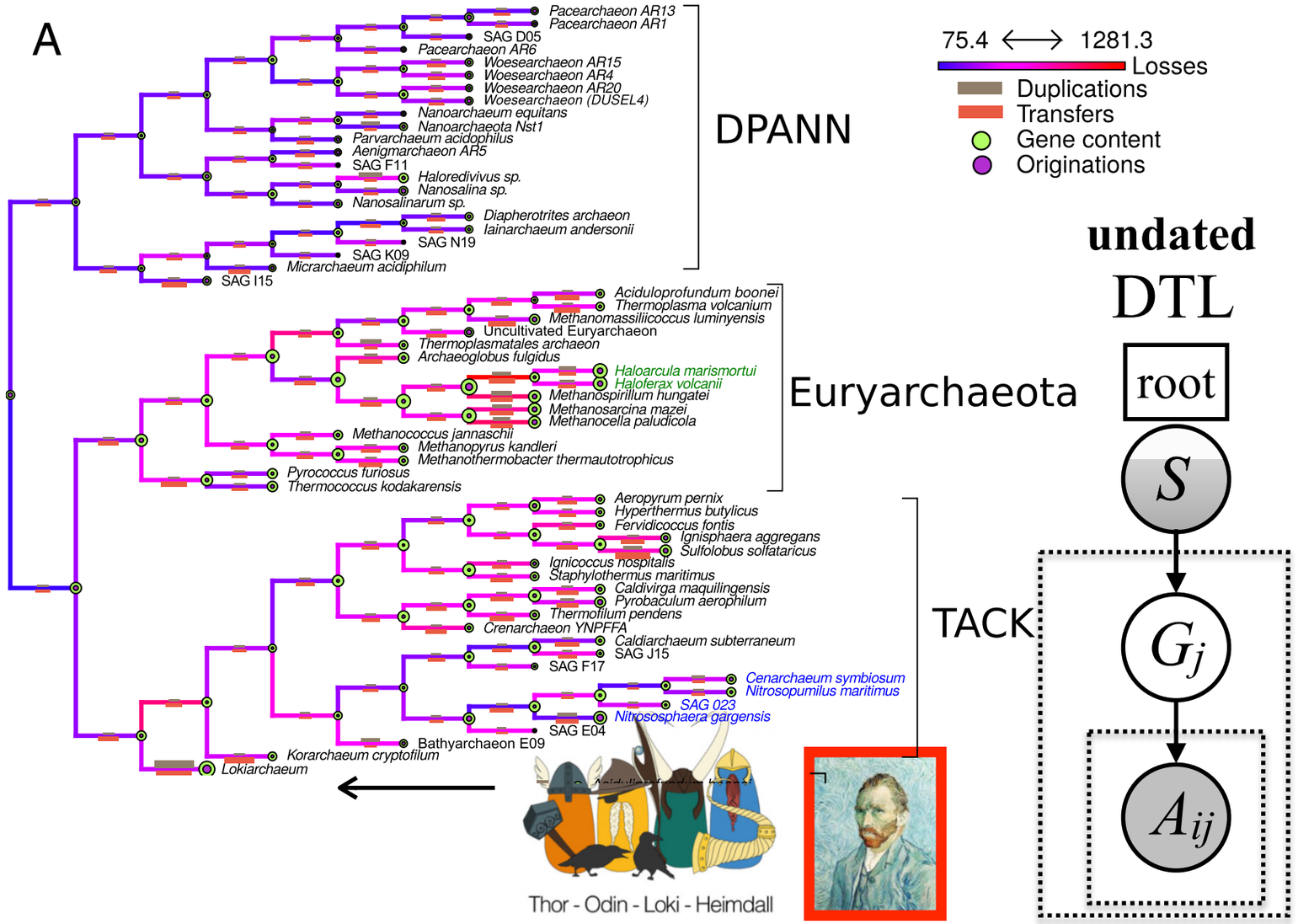



An **anaerobe** that could fix $CO_2$ to acetyl-CoA and generate acetate and ATP from it.


U.Bristol

implemented in ALE:

http://github.com/ssolo/ALE



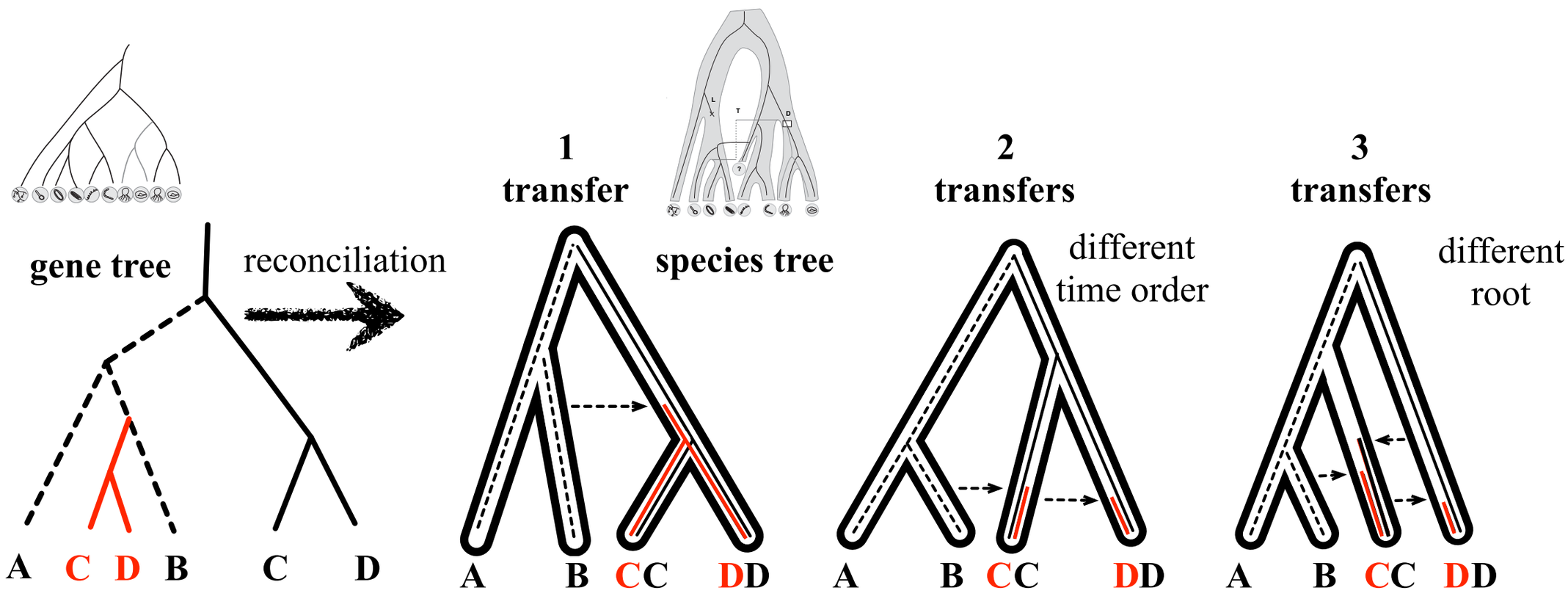

**The predominant mode of genome evolution is not gene loss, but ongoing lineage specific innovation.**

Williams, **Szöllősi**, .. & Embley *PNAS* (2017)
*Integrative modelling of gene and genome evolution roots the archaeal tree of life*

ssolo@elte.hu

# Horizontal gene transfer as information

Transfer events, encoded in the topologies of gene trees can be thought of as "*molecular fossils*" that record the order of speciation events.




Vincent Daubin

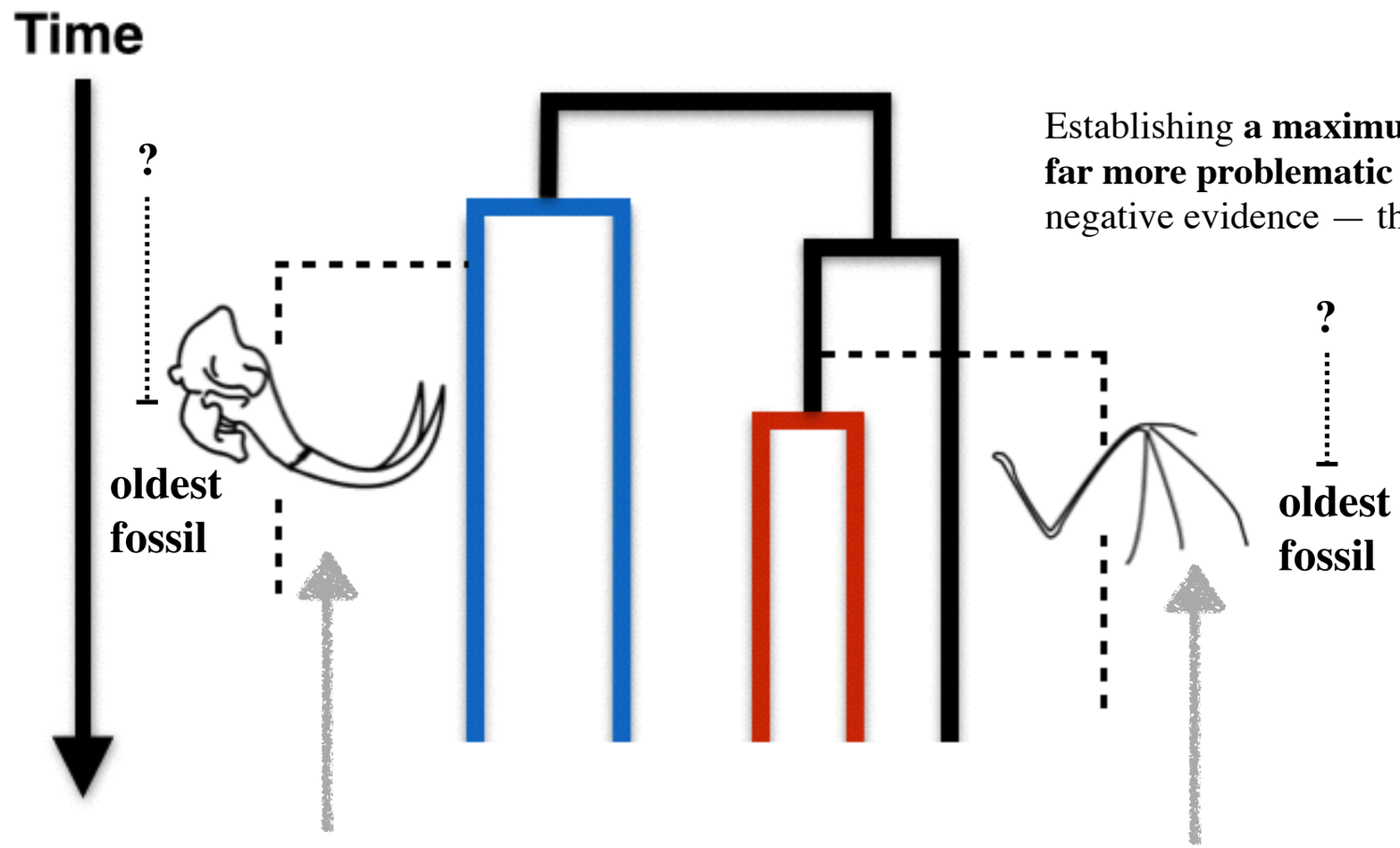
LBBE

**Szöllősi**, Boussau, Abby, Tannier & Daubin ***PNAS*** (2012)
*Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations*

ssolo@elte.hu

# Rocks

The geological record is the only source of information concerning absolute time



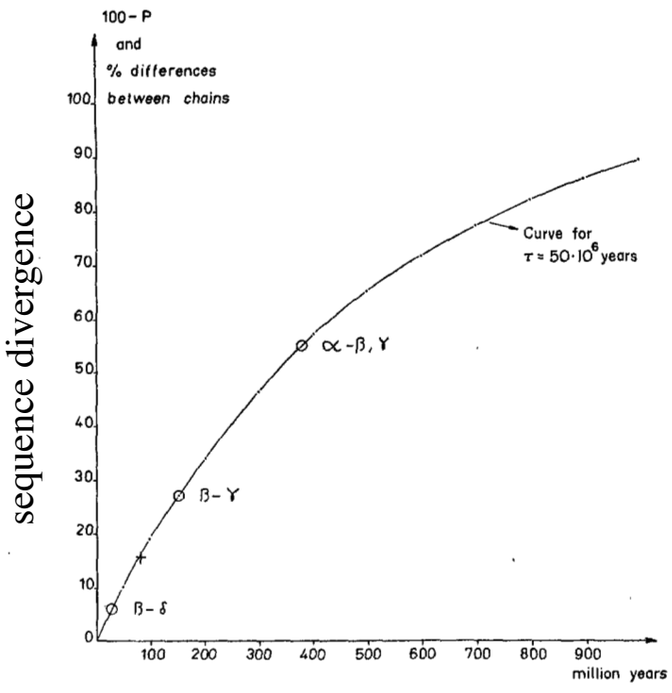

Establishing **a maximum constraint is far more problematic** — it relies on negative evidence — the absence of fossil.

The fossil record is **directly informative on the minimum ages** of clades based on the age of their oldest fossil representative

ssolo@elte.hu

# Molecular Clocks

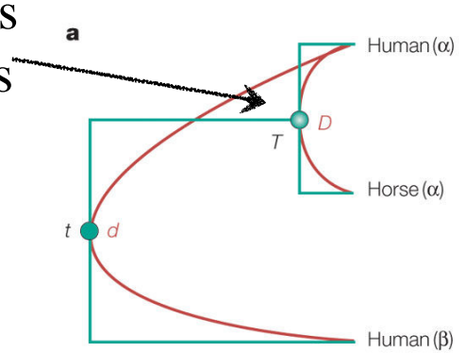
The molecular clock hypothesis reflects the observation that the **differences between homologous** amino acid **sequences** from different mammals **are roughly proportional to their time of divergence.**



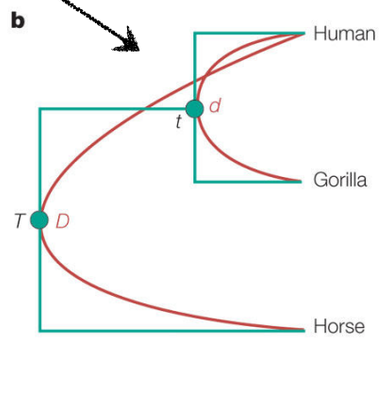

divergence time according to fossils

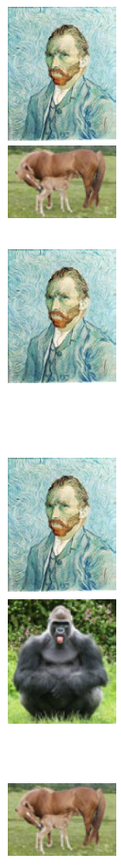
**Zukerkandl and Pauling** 1965

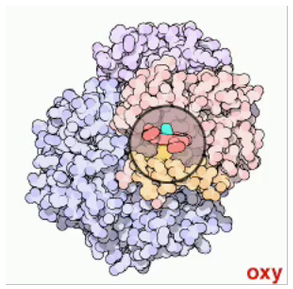
~130 million years
18 aa substitutions




1 & 2 substitutions
~11 million years





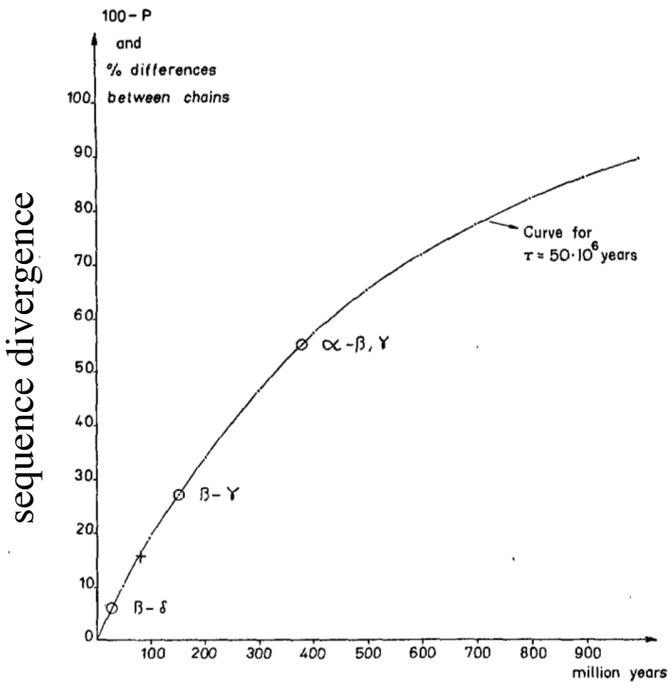

$2\alpha + 2\beta$

Nature Reviews | Genetics

ssolo@elte.hu

# Molecular Clocks

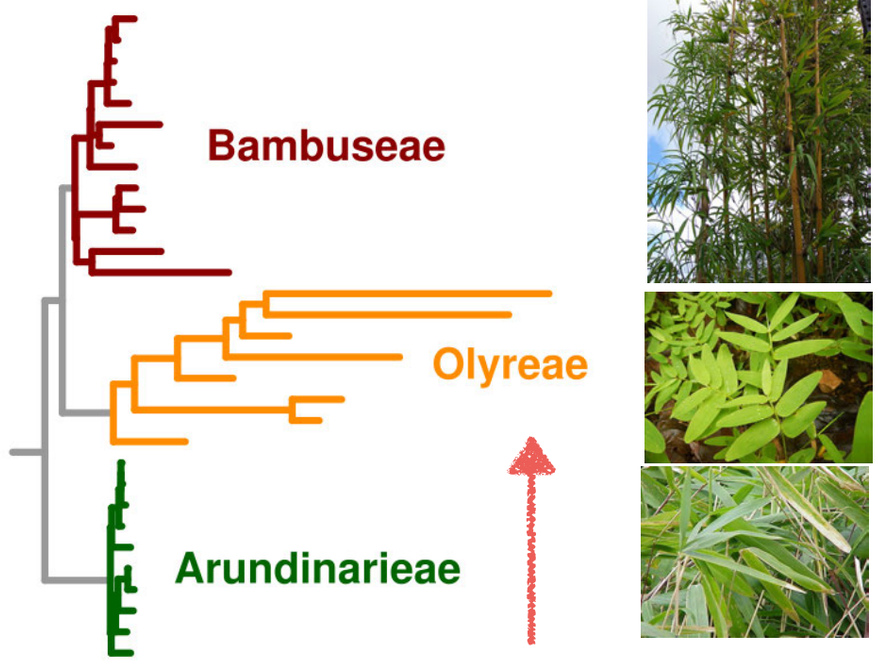
The molecular clock hypothesis reflects the observation that the **differences between homologous** amino acid **sequences** from different mammals **are roughly proportional to their time of divergence.**




divergence time according to fossils

**Zukerkandl and Pauling** 1965

**evolutionary rates vary**
**molecular clocks are local**




**Wysocki et al.** 2014 & Wikipedia

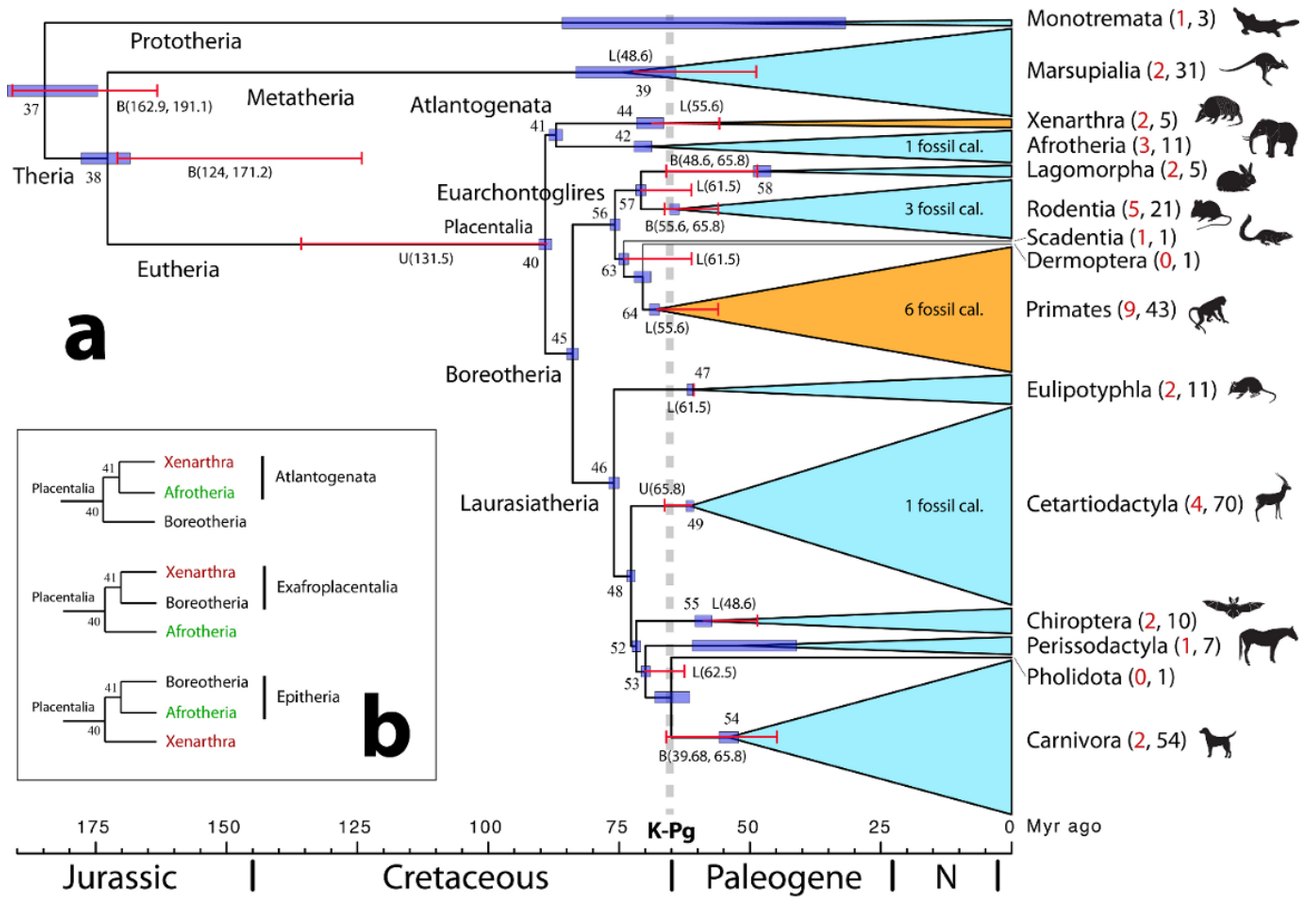
ssolo@elte.hu

# Rocks & Clocks

Inadequate modelling of the global violation of the molecular clock historically lead to great controversies..

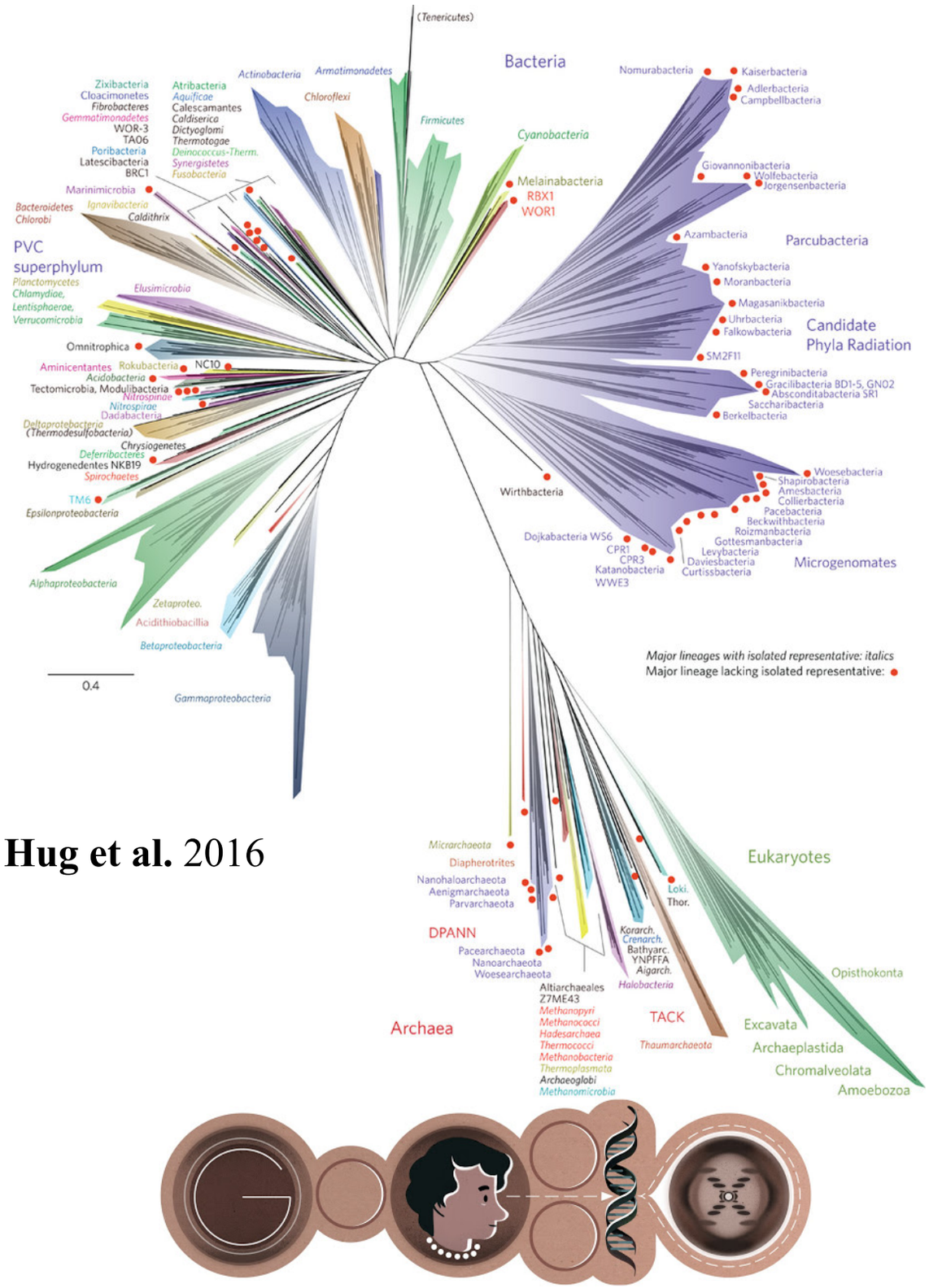

**Donoghue and Yang** 2015

… today, Bayesian RMC methods have resolved most, but not all controversies, using **sequence based local molecular clocks anchored by multiple fossil calibrations.**




**dos Reis et al.** 2012

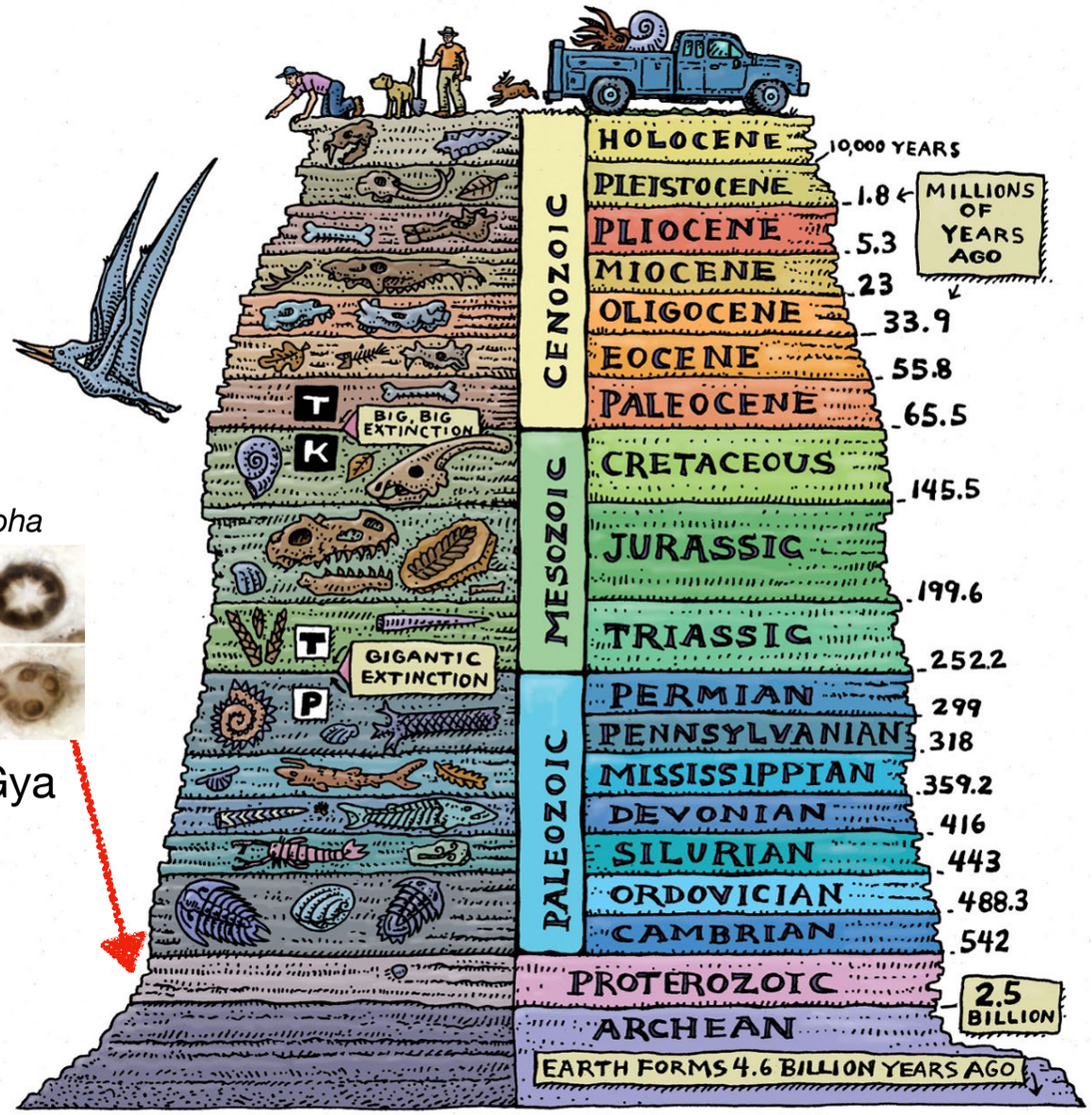
# Rocks & Clocks



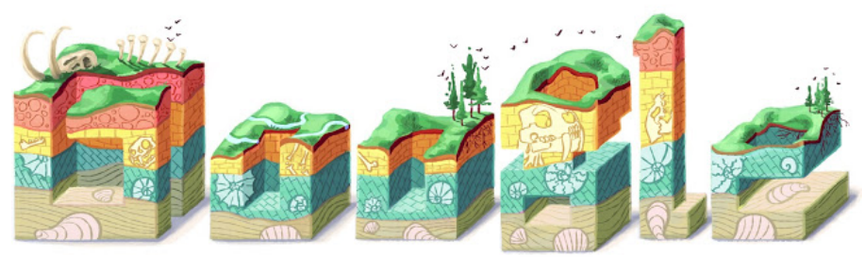

**Hug et al.** 2016

*Bangiomorpha*

~ 1-1.5 Gya



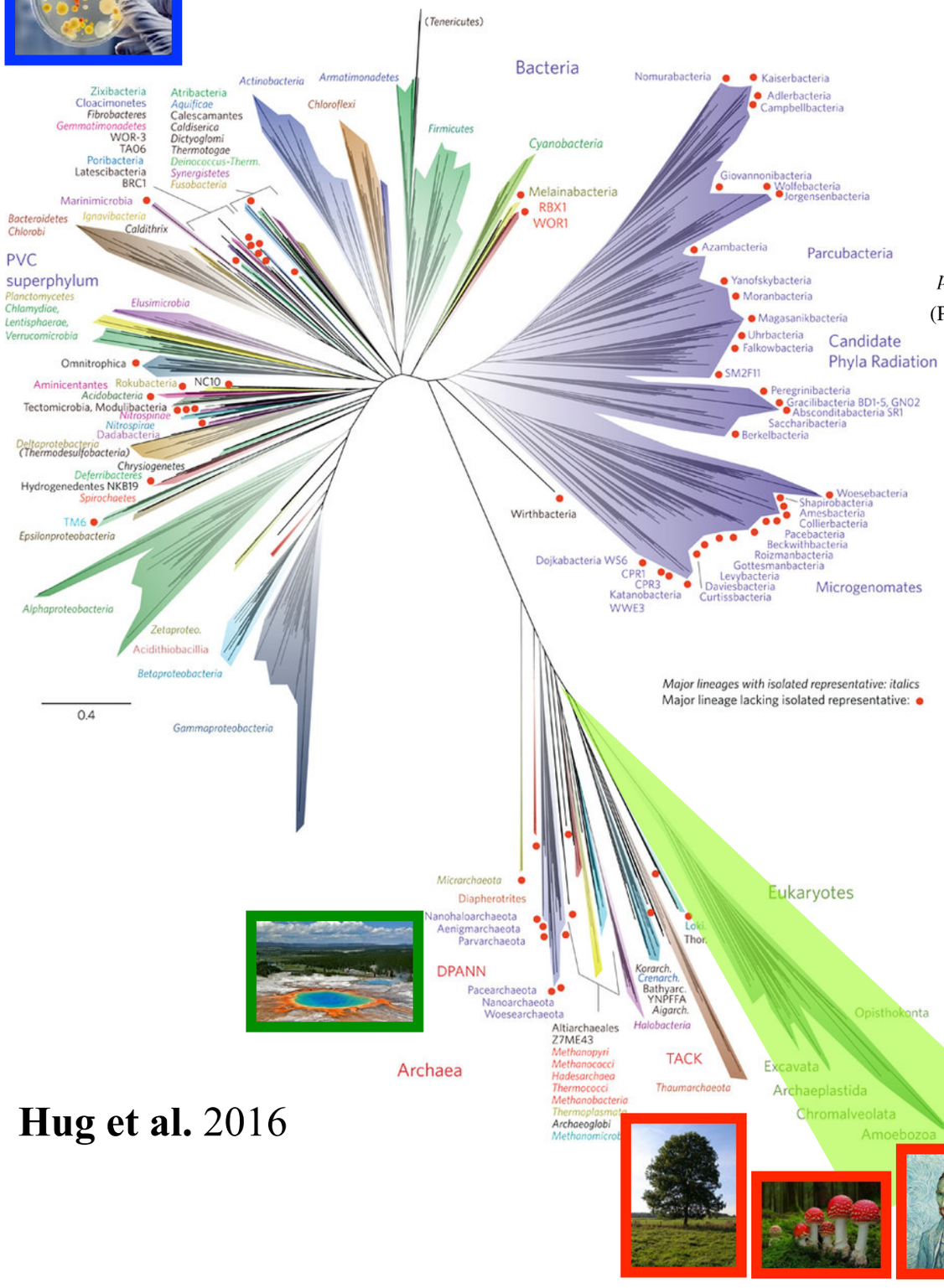


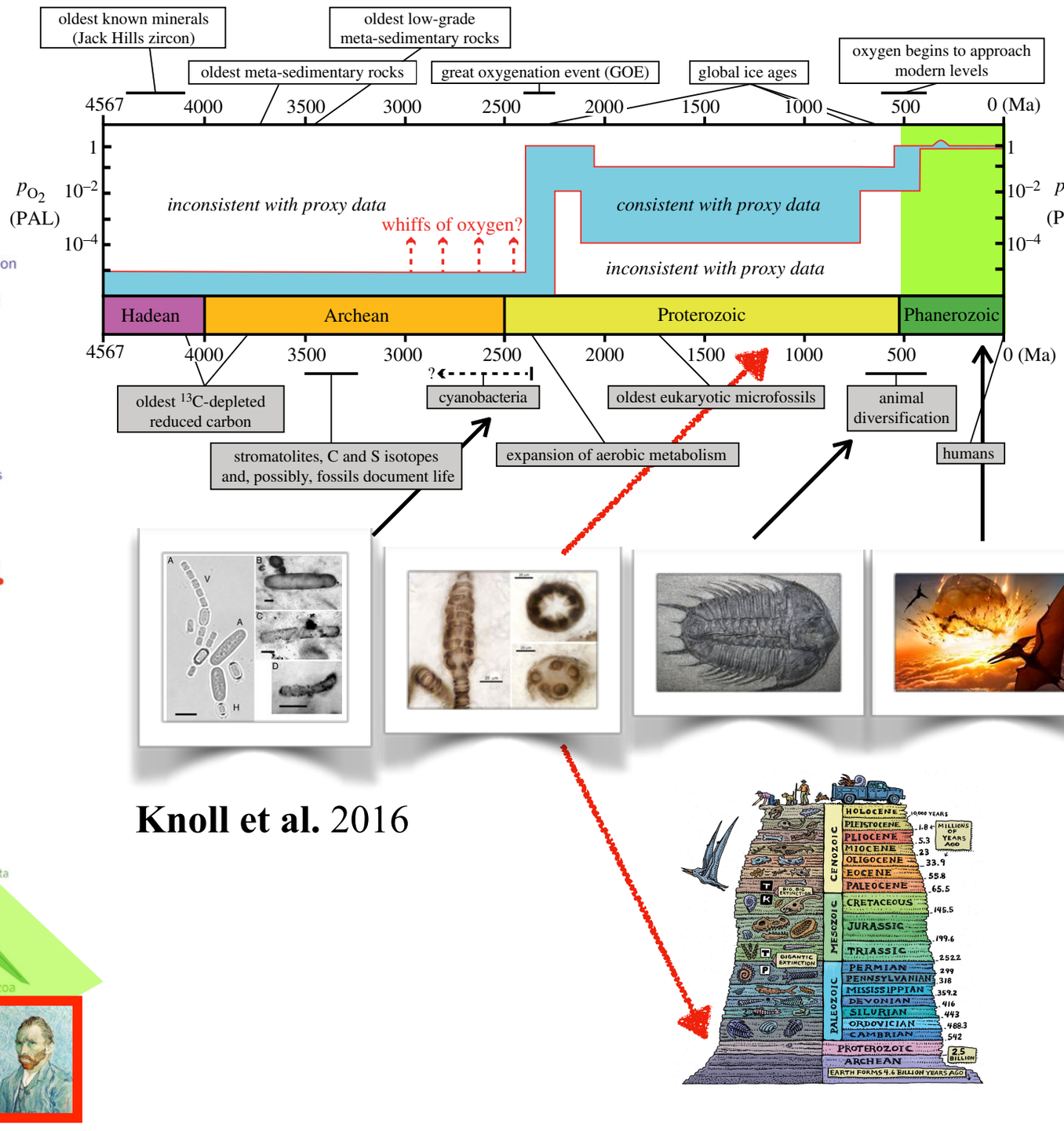
ssolo@elte.hu

# Rocks & Clocks

For majority of life and most it's history we lack sufficient fossils to anchor local clocks.




**Hug et al.** 2016









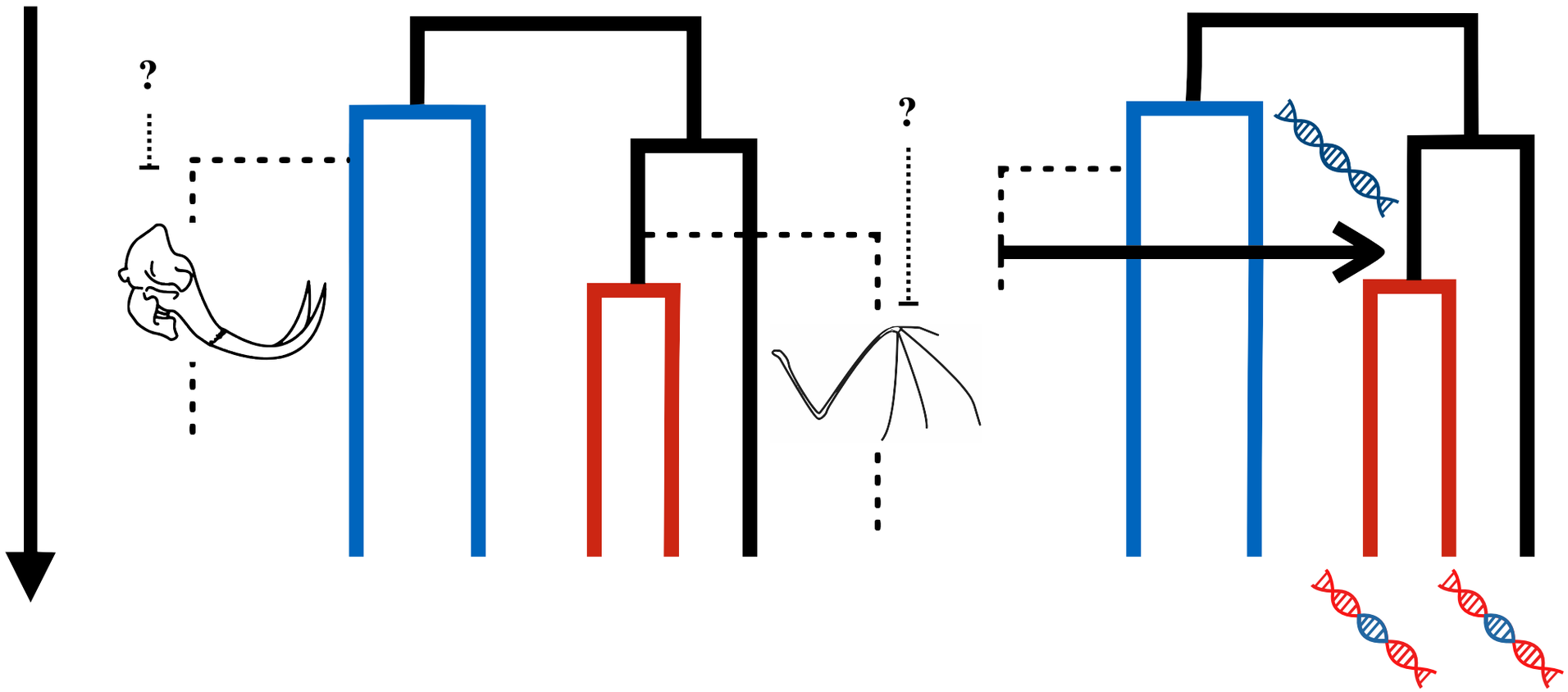
**Knoll et al.** 2016

ssolo@elte.hu

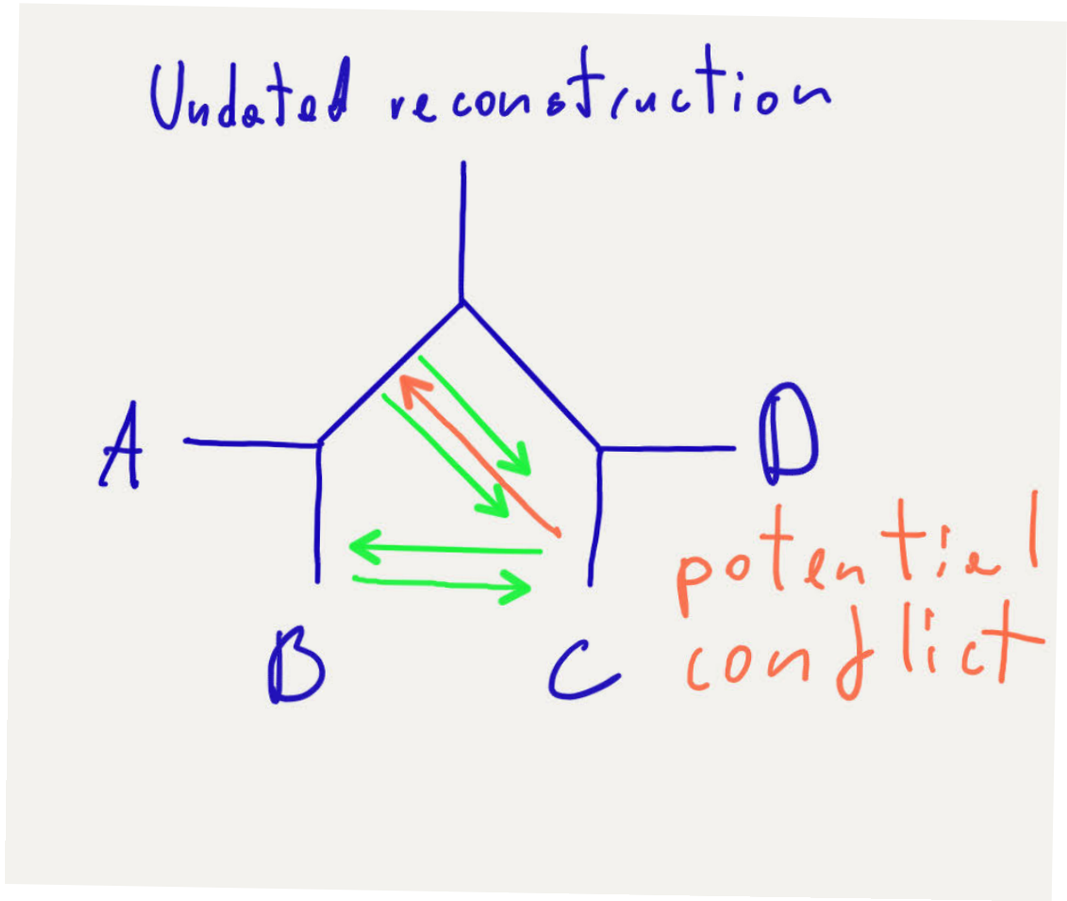
# … and genes from other species!

**Fossils** provide **direct evidence on minimum age**, but only **indirect evidence on maximum and relative ages**.

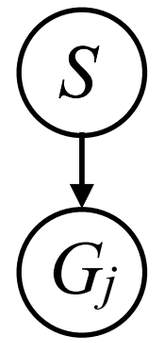
**Transfers** are not informative on absolute age, but do provide **direct evidence on relative ages**.



“undated”
DTL

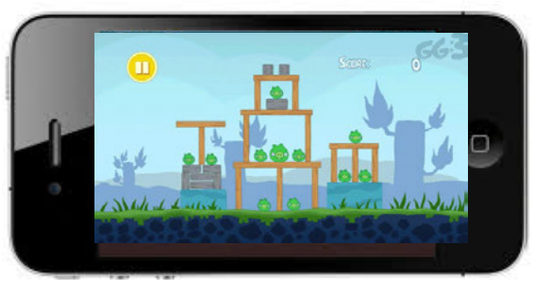



**undated**
DTL




DTL

“undated” DTL



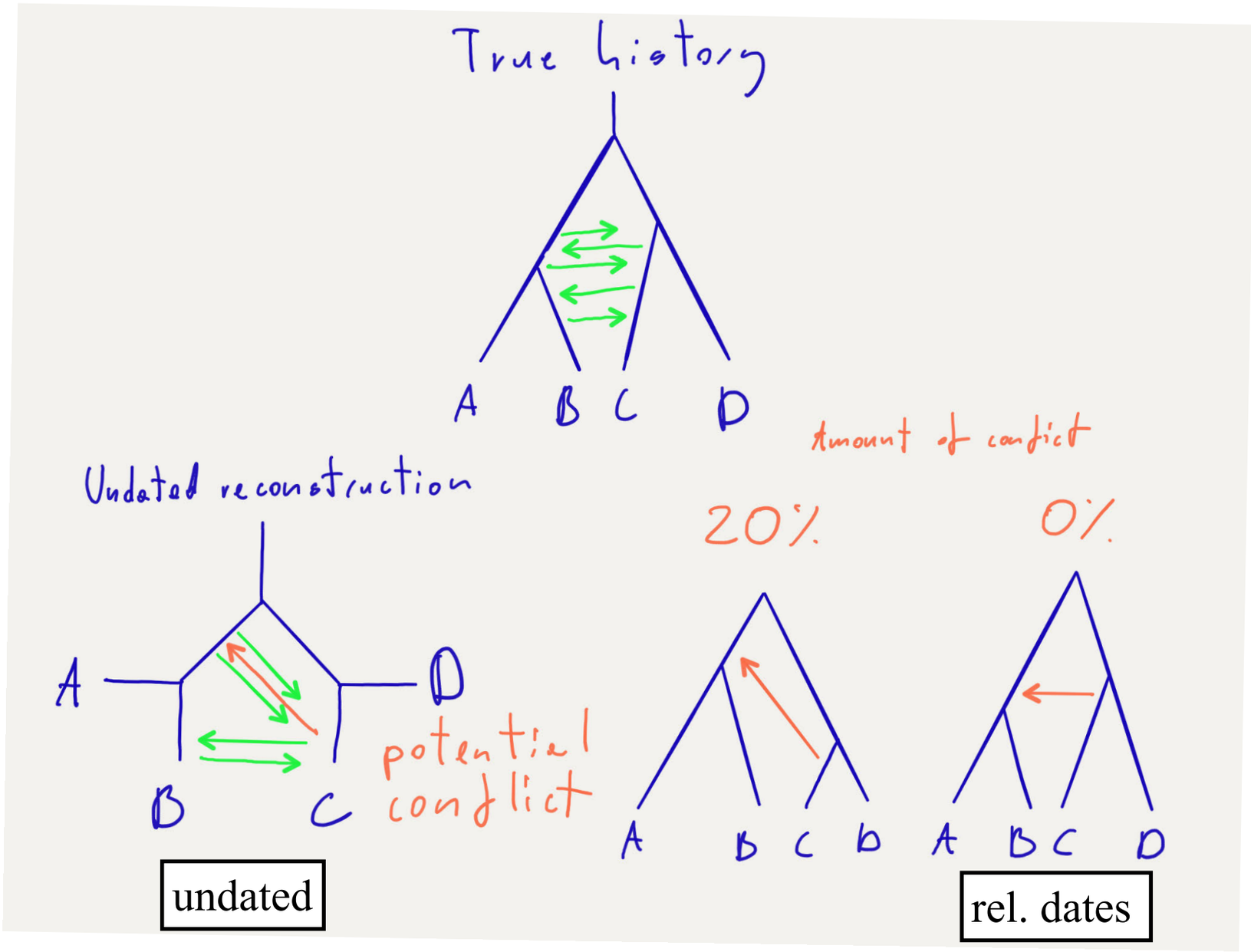
implemented in ALE:

**http://github.com/ssolo/ALE**

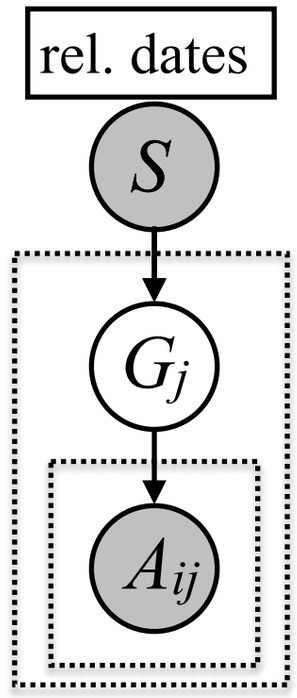
ssolo@elte.hu

# Relative age constrains from transfers

HGT events inferred by an “undated” version of the species tree-aware method ALE were input into the MaxTiC (**max**imal **ti**me **c**onsistency) optimisation method to obtain relative age constrains.



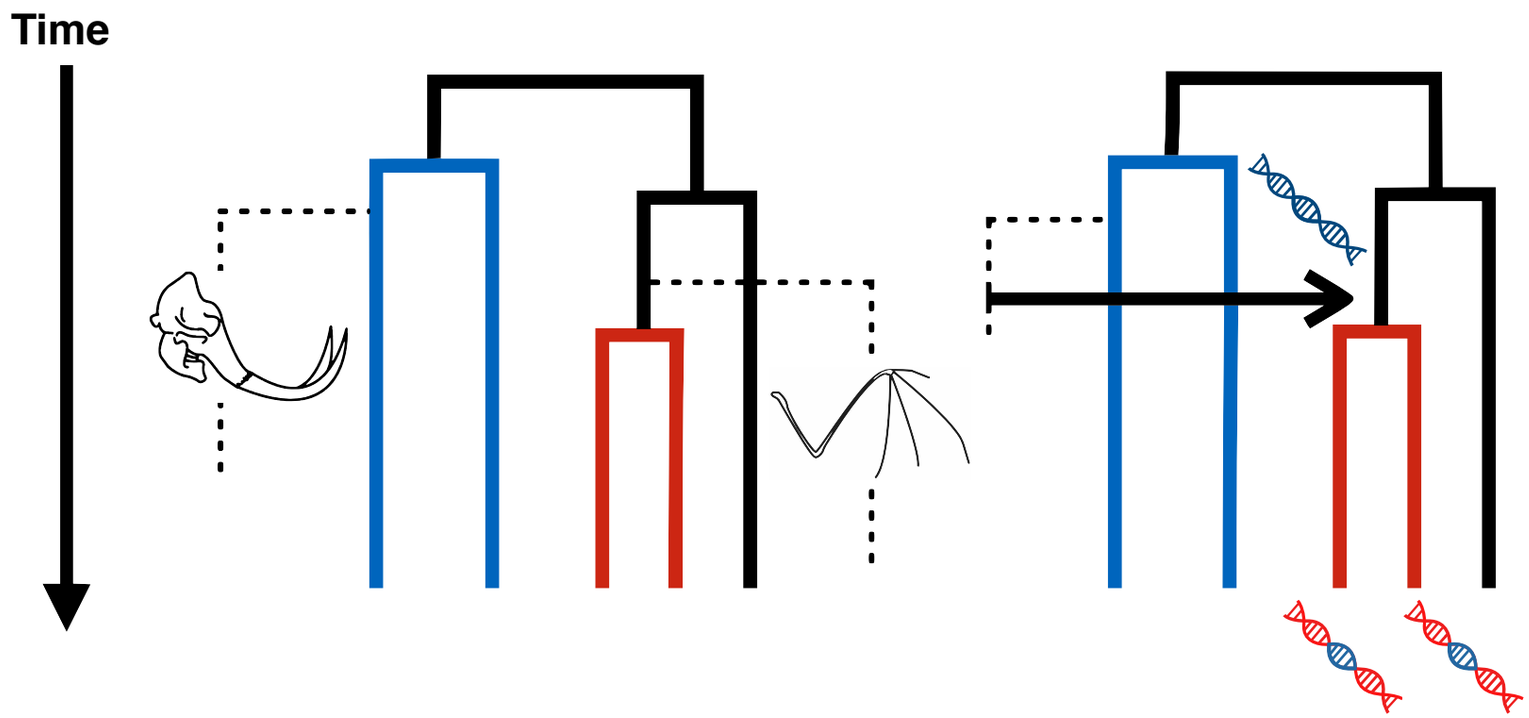

Eric Tannier
LBBE



ssolo@elte.hu

# Rocks, clocks and genes from other species

A direct comparison between fossils and transfers is not possible. Following Zuckerkandl and Pauling, we correlated both fossil and transfer based age estimates with sequence divergence:




40 genomes

↓

ALE

+

MaxTiC

↓

**3322 HGT events**

informative constraints

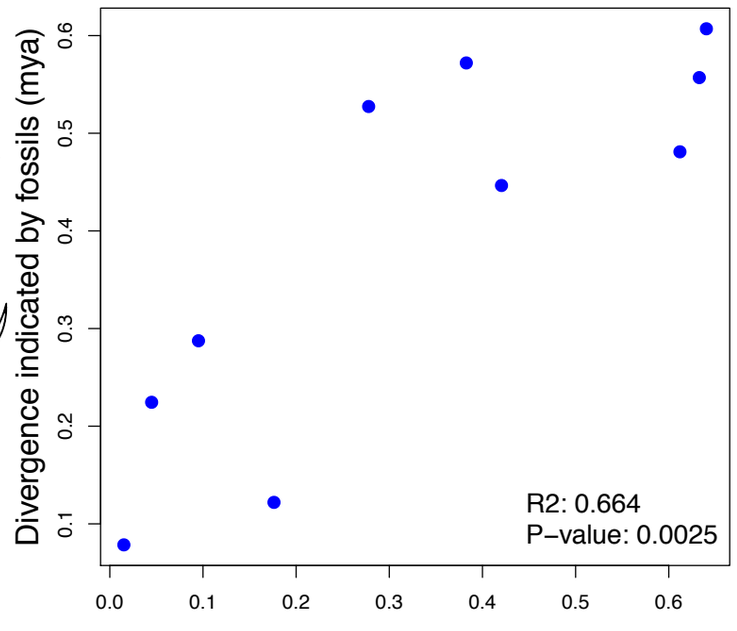
**12 calibrations with maximum age**

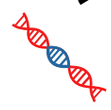
dos Reis et al. 2012

Adrian Davin

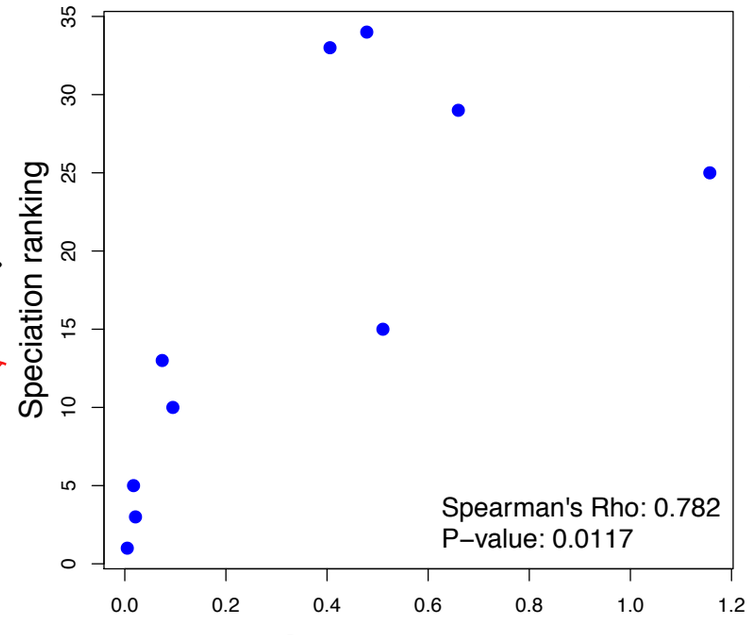
LBBE

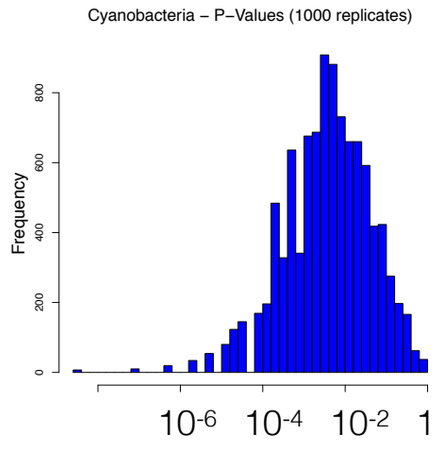
Fossils in mammals



Gene transfers in Cyanobacteria



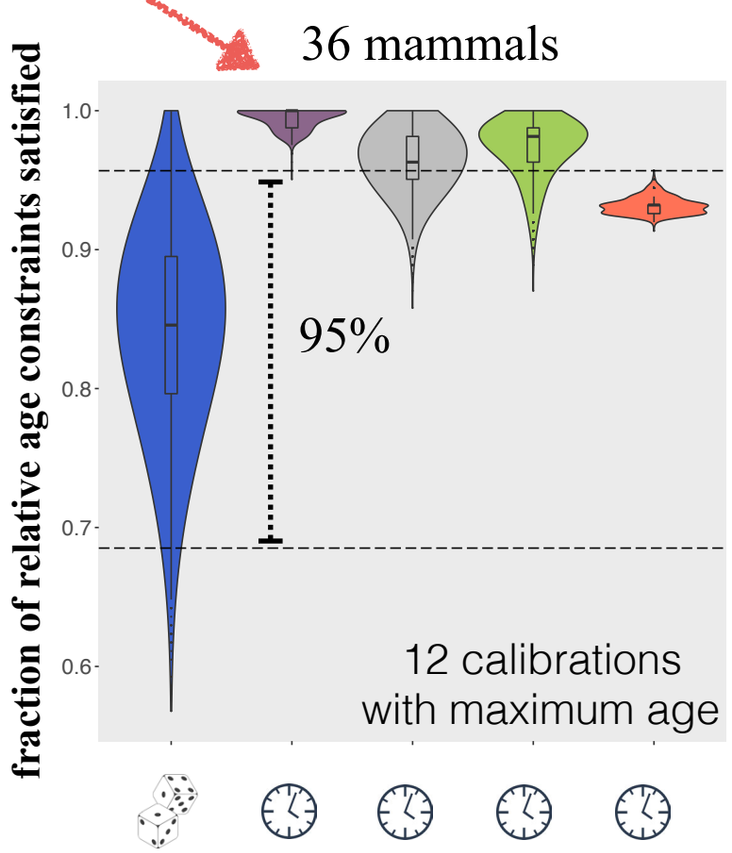



ssolo@elte.hu

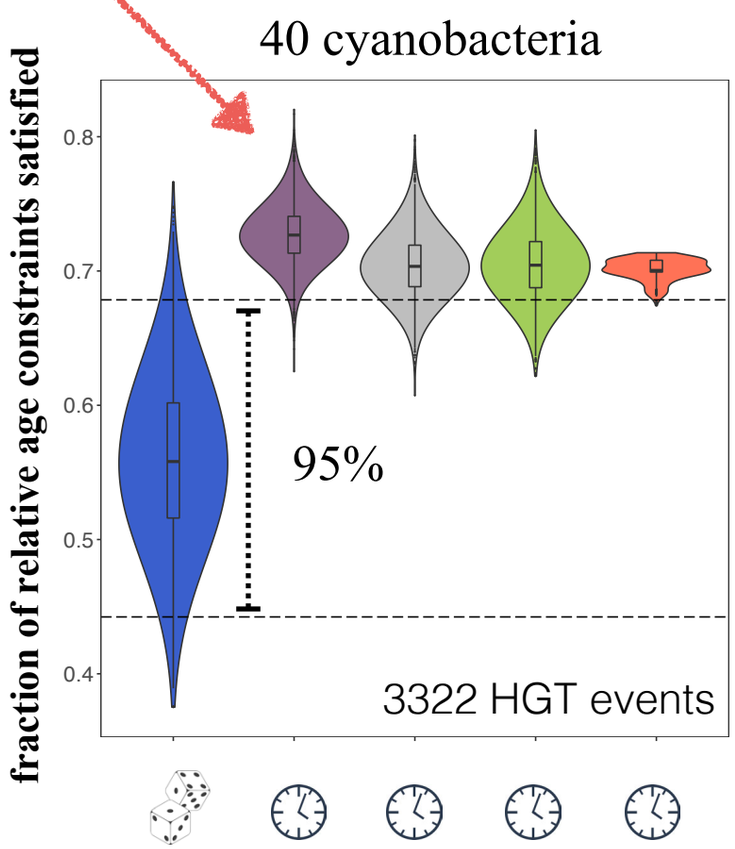
# Rocks, clocks and genes from other species

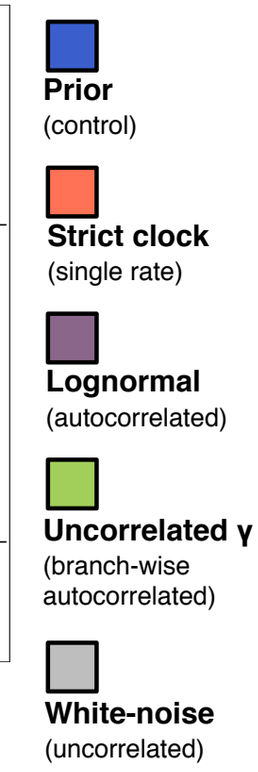
To directly compare relative age constraints, we measured how different relaxed molecular clock models, without fossil calibrations, are able to predict the relative timing of speciations implied by fossils and by transfers.

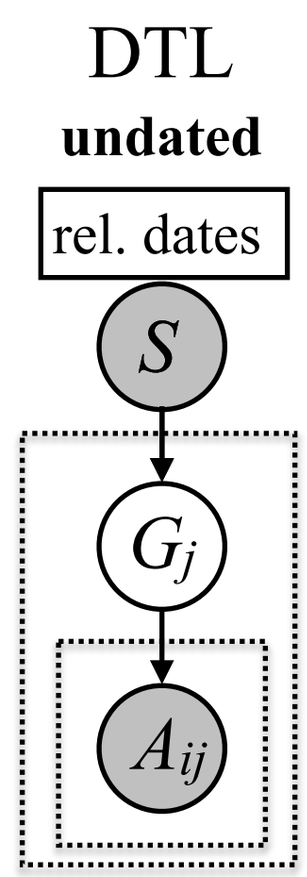
**faster local clock in rodents**

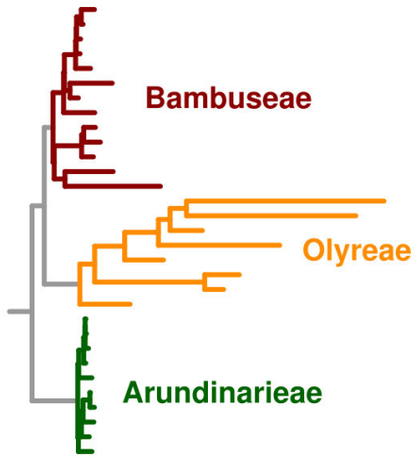



**faster local clock in prochlorococcus**




**faster local clock**

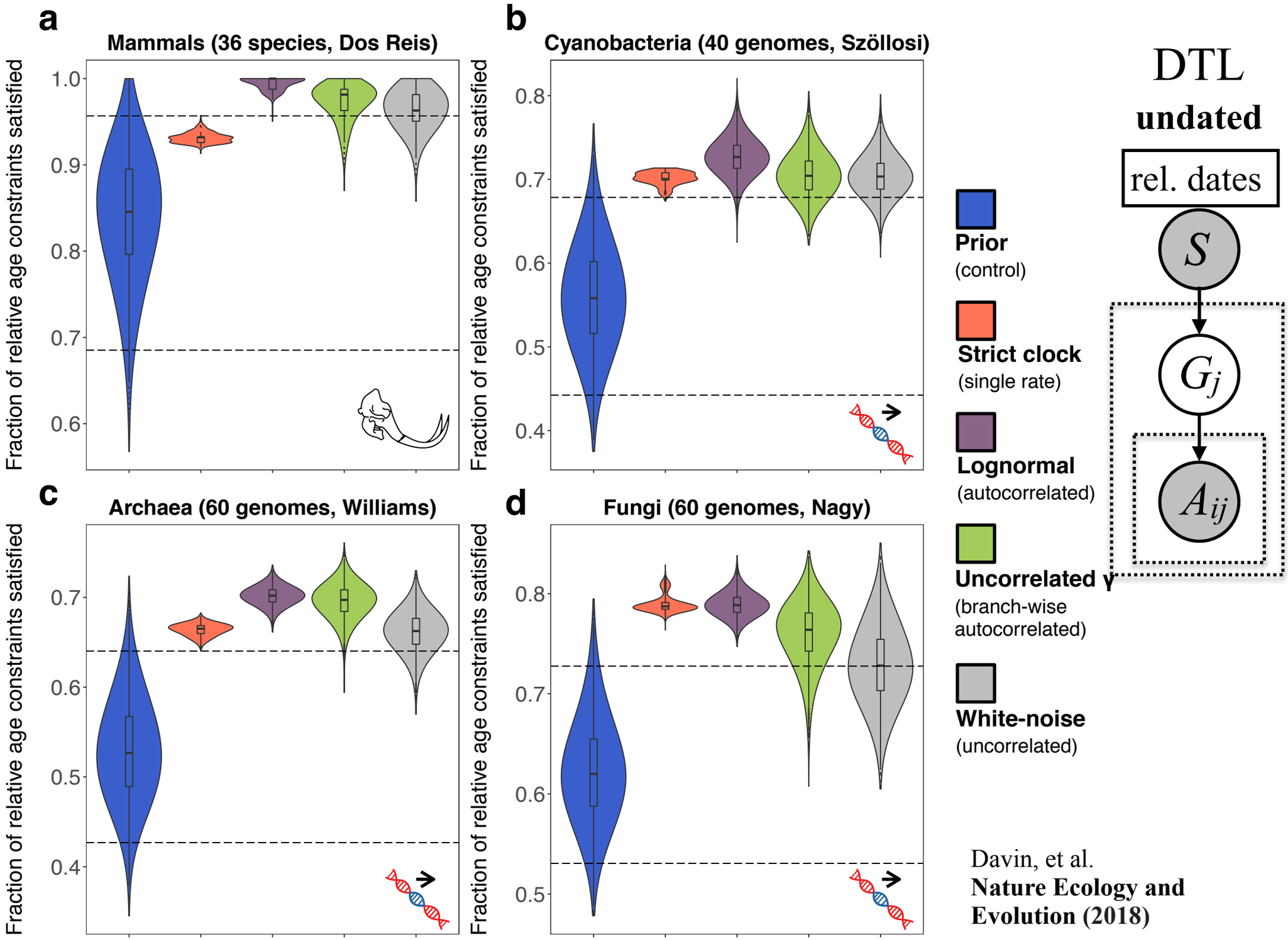
Adrian Davin

Davin, Tannier, Williams, Boussau, Daubin & **Szöllősi**, **Nature Ecology and Evolution (2018)**
*Gene Transfers Can Dated the Tree of Life*

ssolo@elte.hu

# Transfer based relative dates correlate with molecular clocks

To directly compare relative ages, we measured how different relaxed molecular clock models, without fossil calibrations, are able to predict the relative timing of speciations implied by fossils and by transfers.



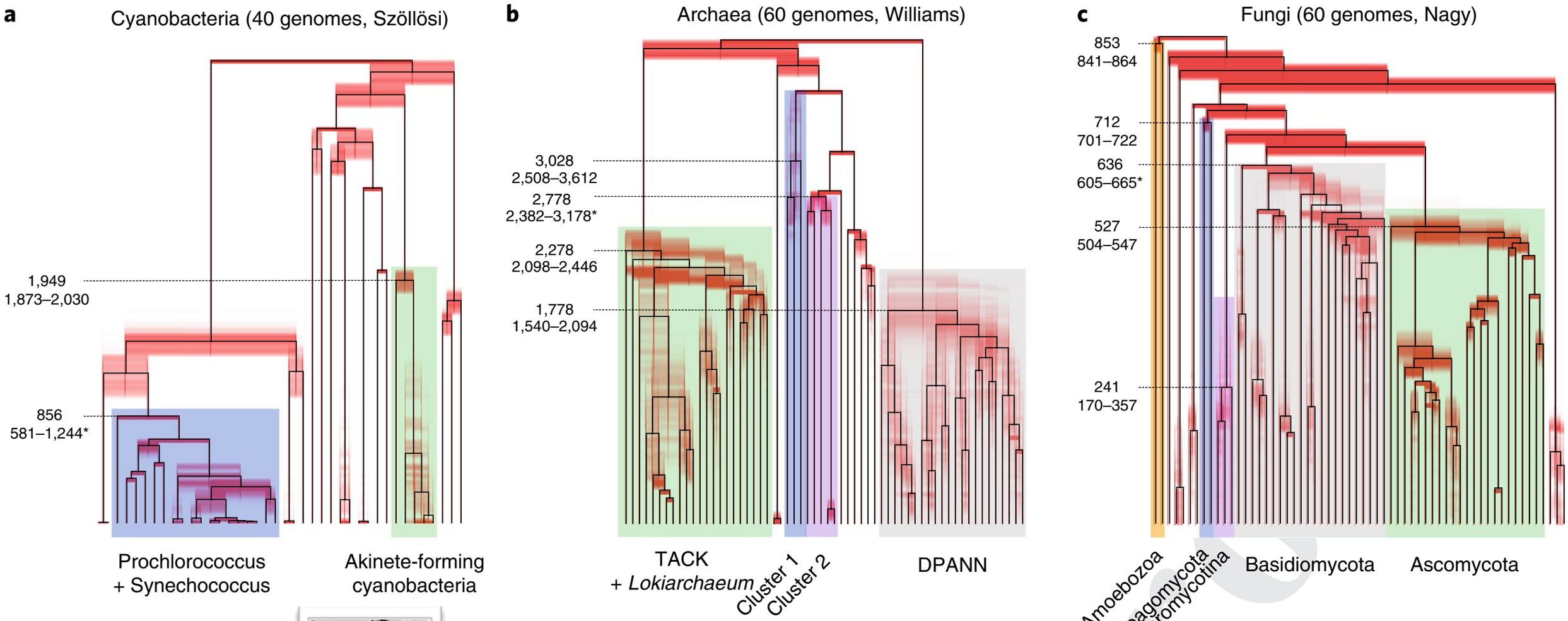

Adrian Davin

Davin, et al.
**Nature Ecology and Evolution (2018)**

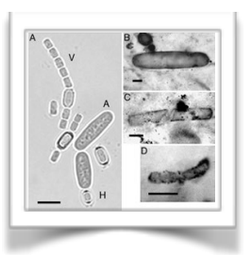
ssolo@elte.hu

# Rocks, clocks and genes from other species

The order of speciations according to LGT. 5000 chronograms with a speciation time order compatible with LGT-based constraints were sampled per data set.




Davin, et al.
**Nature Ecology and Evolution (2018)**



**microfossils ~ 2.1 Gya**

**Tomitani *et al.*** 2006


**biogenic methane ~ 3.5 Gya**

**0.6-0.8 Gya**

nature ecology & evolution

ARTICLES

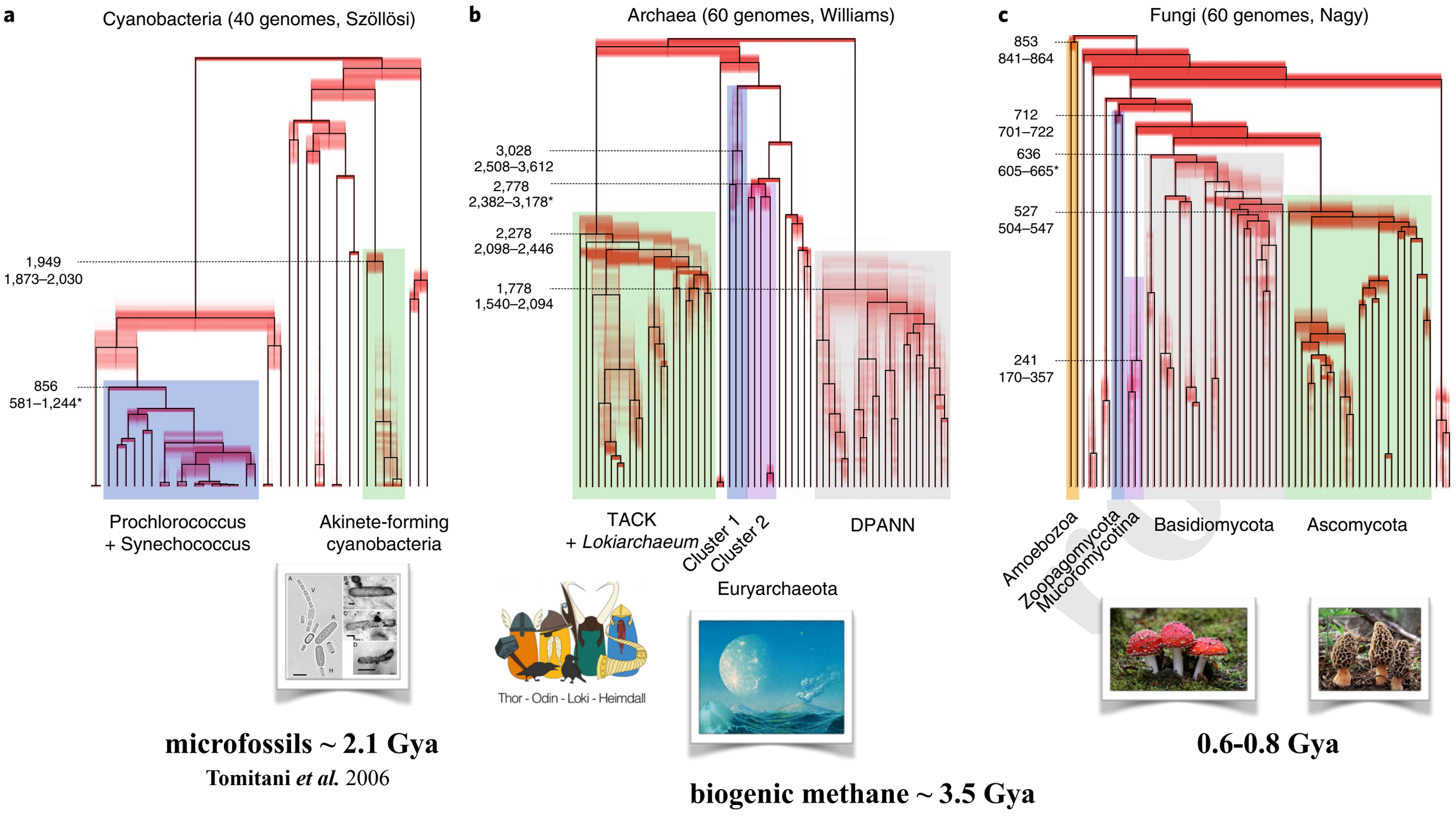
https://doi.org/10.1038/s41559-018-0525-3

ssolo@elte.hu

# Gene transfers can date the tree of life

http://rdcu.be/KrjA

Adrián A. Davín[1], Eric Tannier[1,2], Tom A. Williams[3], Bastien Boussau[1], Vincent Daubin[1*] and Gergely J. Szöllősi[4,5*]

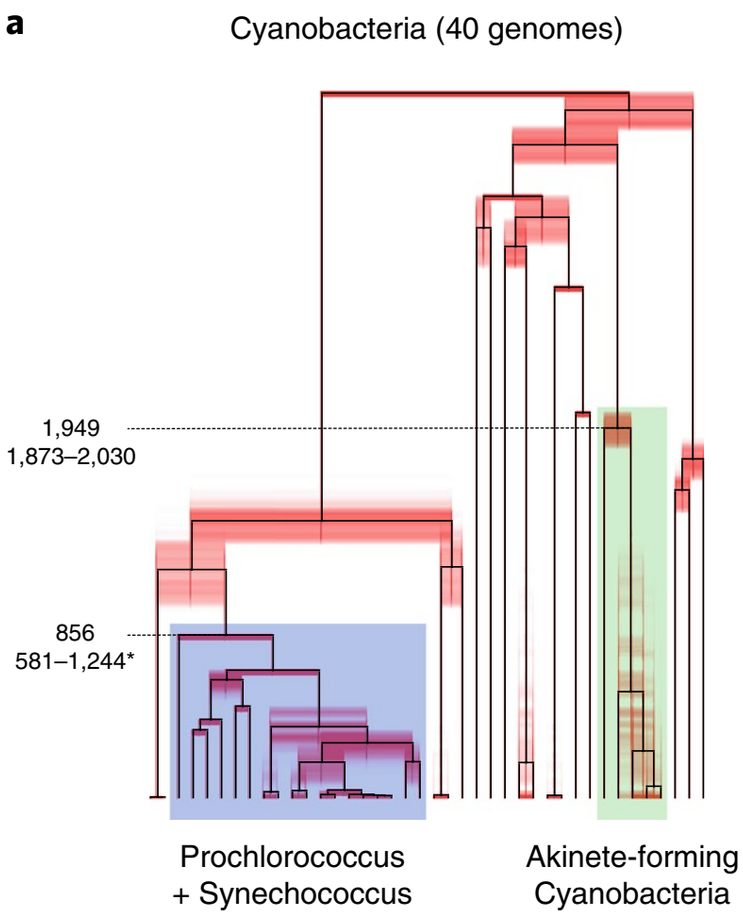


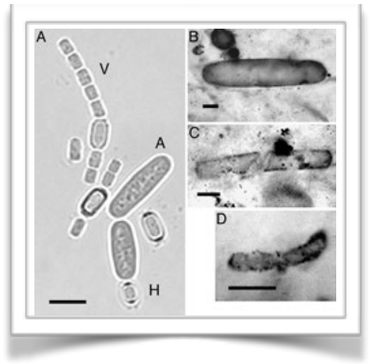
ssolo@elte.hu

# Do clocks & transfers predict rocks?

Combing **RMCs with relative constraints and fossil calibrations** we can infer dated trees in RevBayes

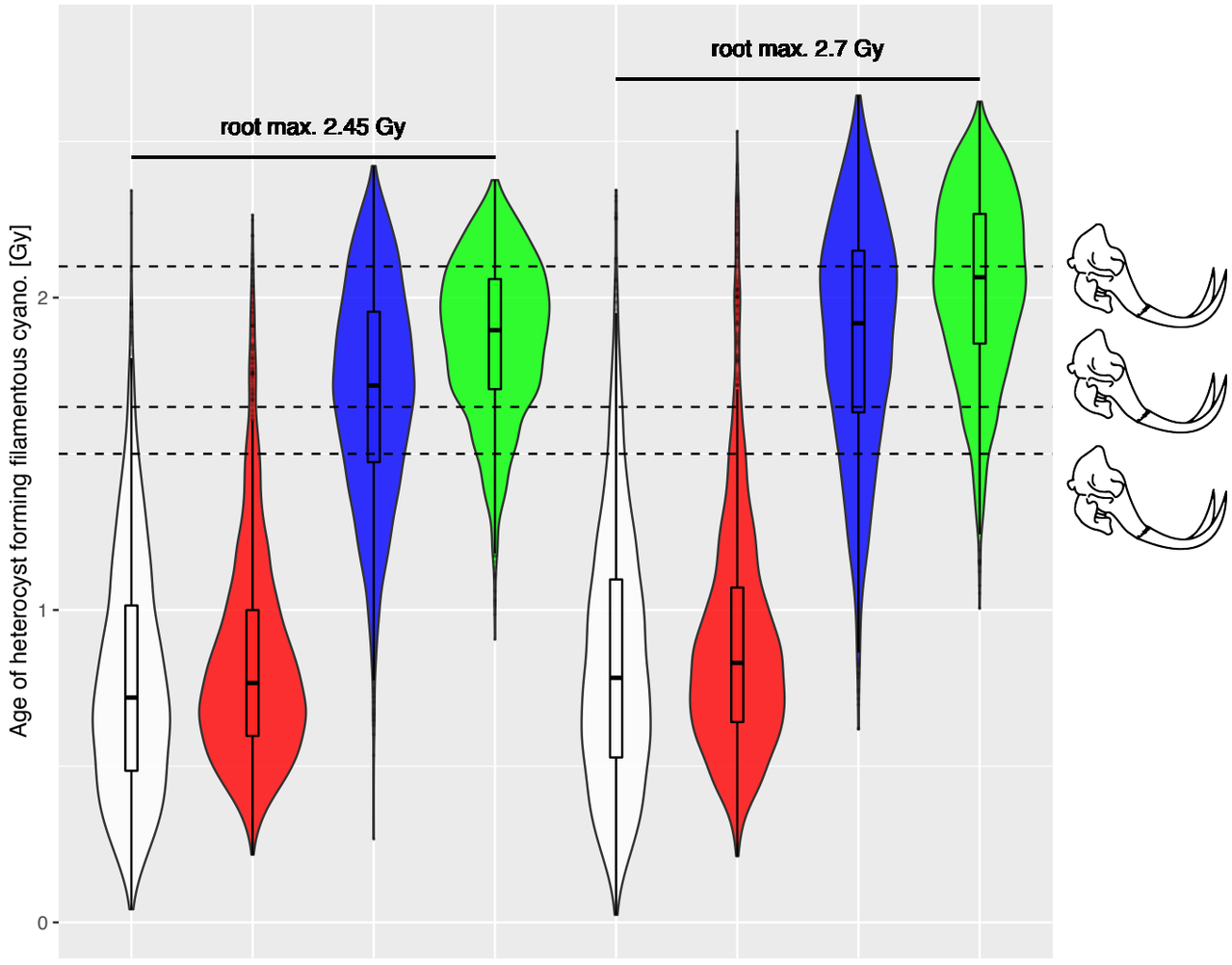
**a**

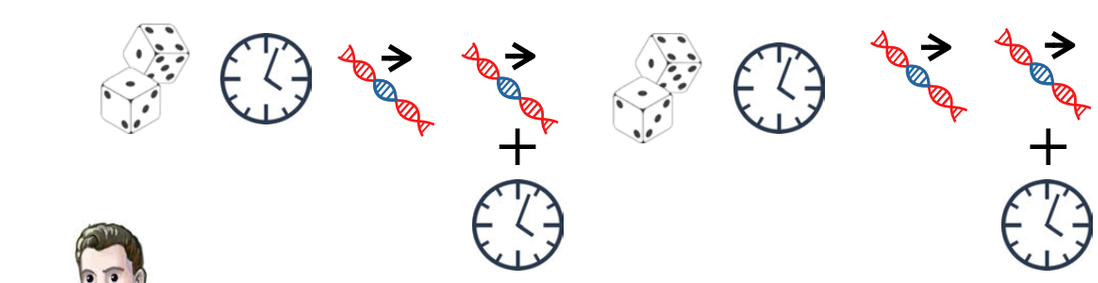



**microfossils**

**Tomitani *et al.*** 2006



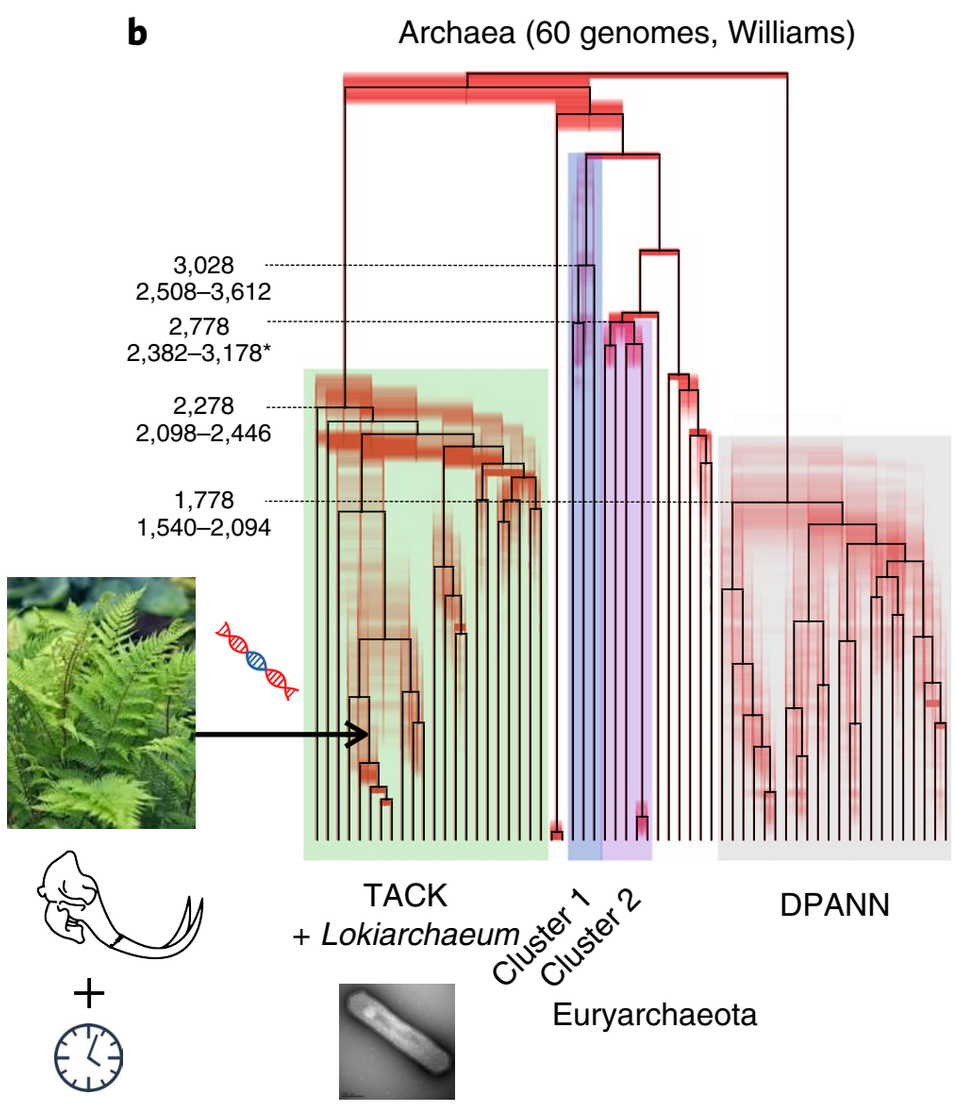
implemented in **RevBayes**

**https://revbayes.github.io**

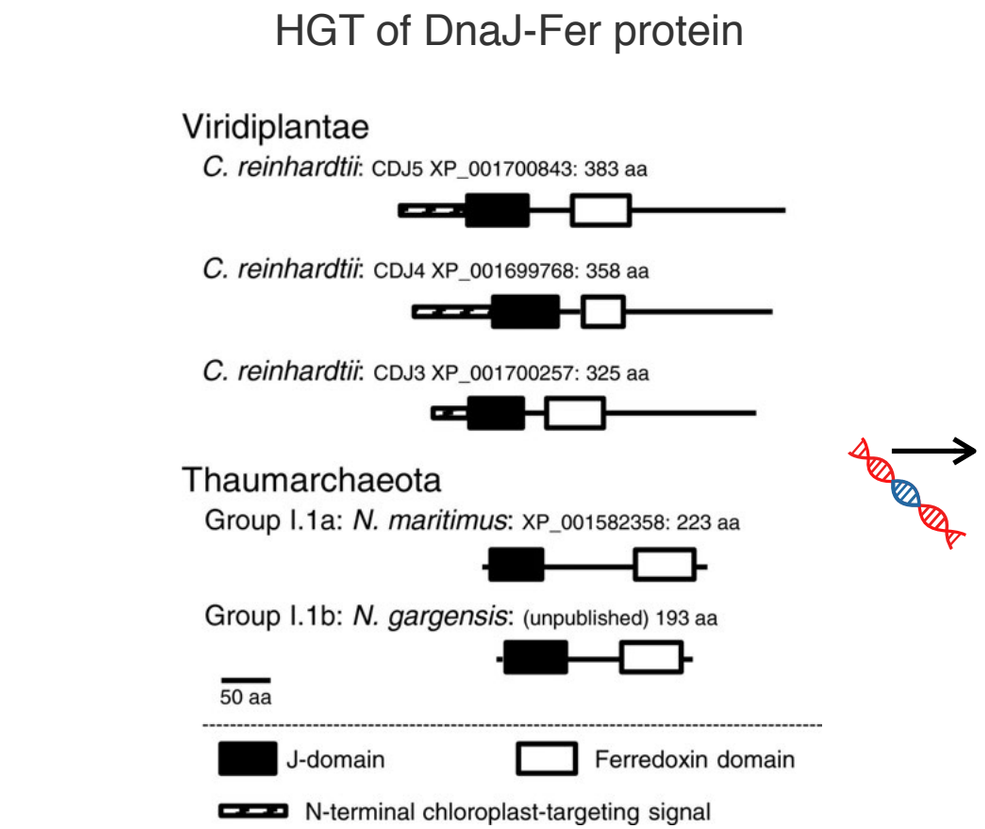
ssolo@elte.hu

# Do clocks & transfers predict rocks?

Combing **RMCs with relative constraints and fossil calibrations** we can infer dated trees in RevBayes




Thaumarchaeota are the dominant archaea in most soil systems where they constitute up to 5% of all prokaryotes




Petitjean et al. ***BMC Evol Biol.*** **(2012)**

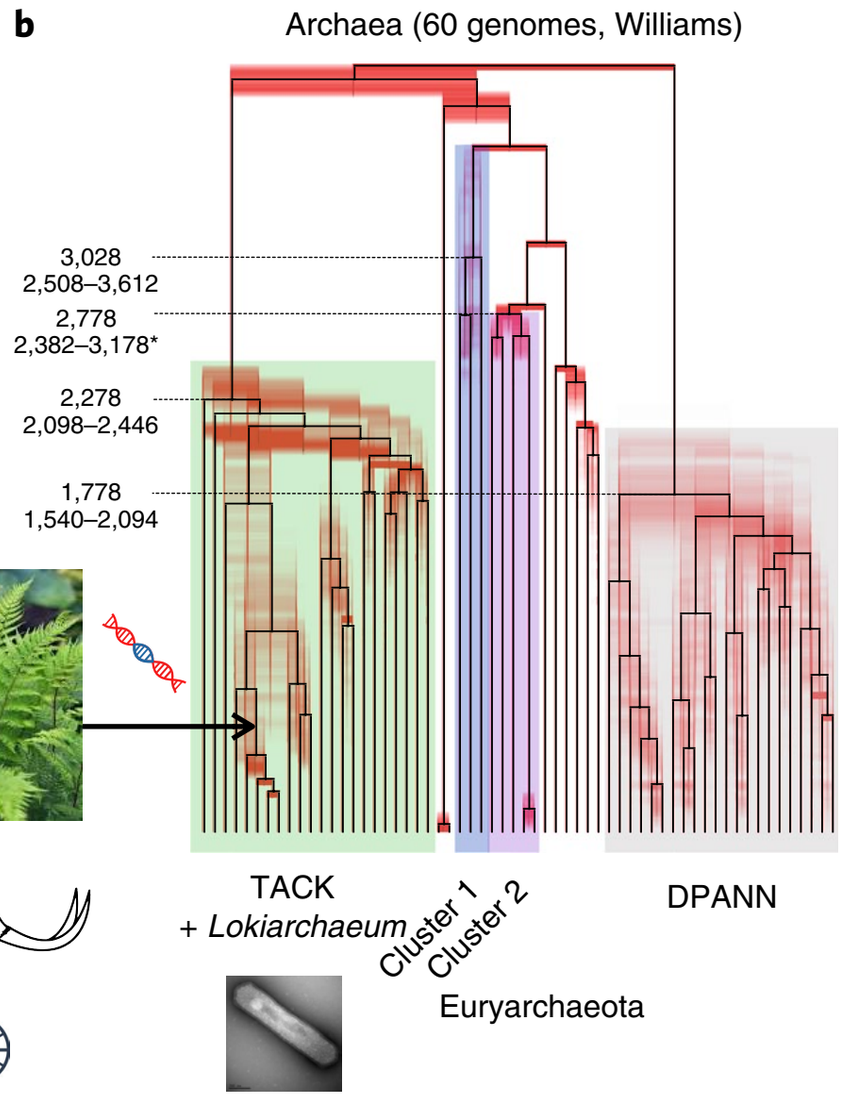
implemented in **RevBayes**

**https://revbayes.github.io**

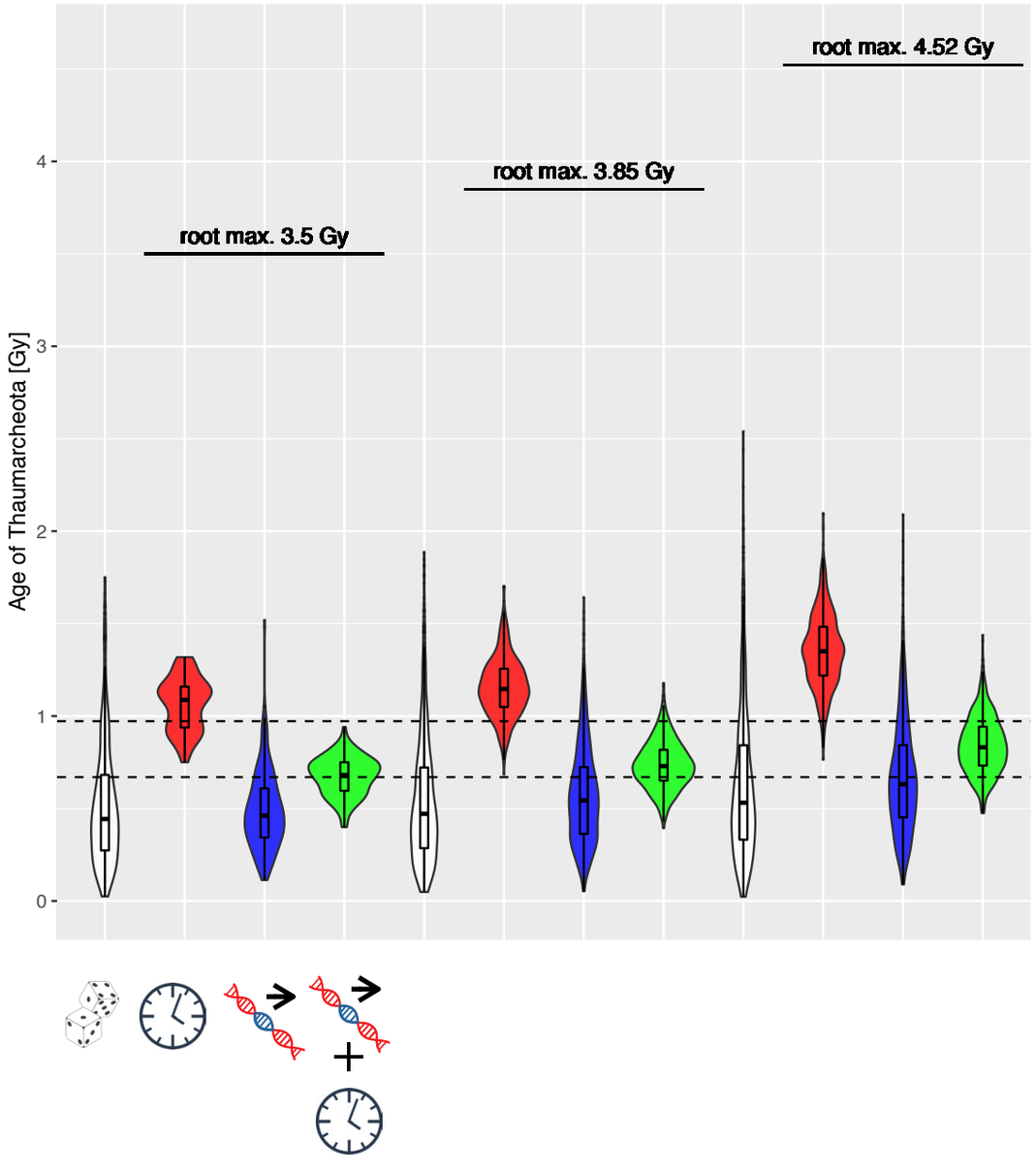
ssolo@elte.hu

# Do clocks & transfers predict rocks?

Combing **RMCs with relative constraints and fossil calibrations** we can infer dated trees in RevBayes




Thaumarchaeota are the dominant archaea in most soil systems where they constitute up to 5% of all prokaryotes



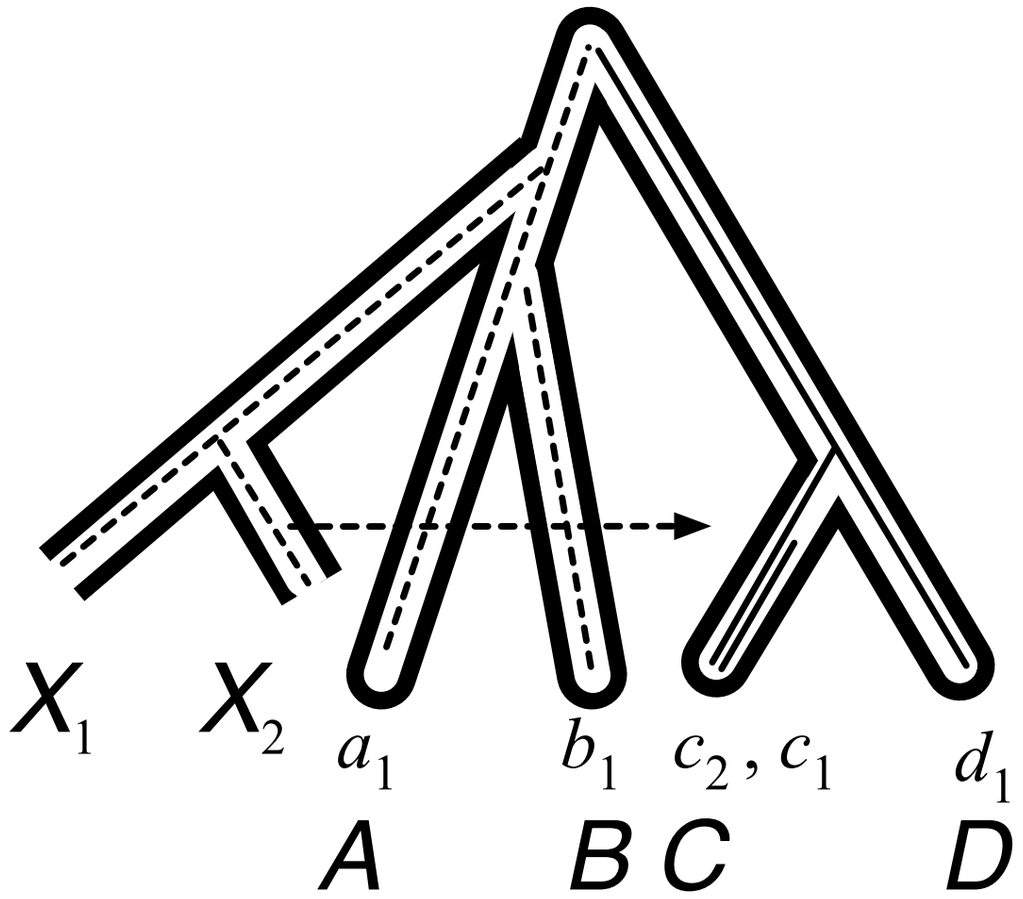

972.4–669.9 My
Morris PNAS 2018

implemented in **RevBayes**

**https://revbayes.github.io**

# Lateral gene transfer from the dead

.. but the species lineage from which a gene was transferred **may have gone extinct** or not have been sampled.



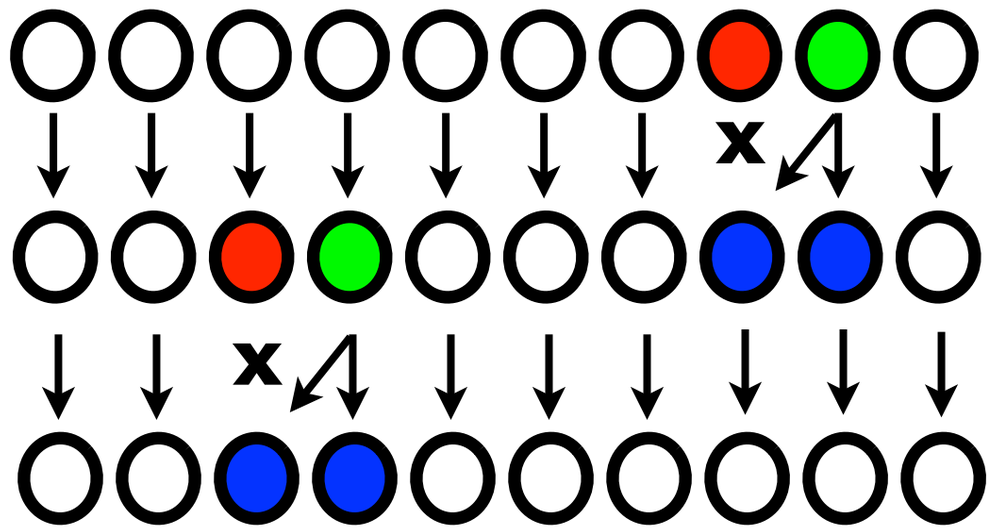

Nicolas Lartillot

**Szöllősi**, Tannier, Lartillot & Daubin *Systematic Biology* (2013)
*Lateral Gene Transfer from the Dead*

# A minimal model of speciation dynamics

## the Moran process

$N$ species

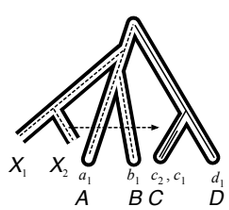


rate per species :

$\sigma$

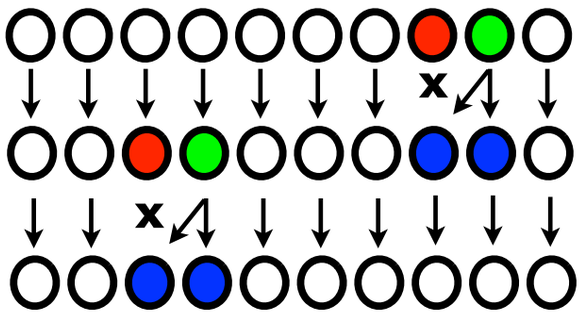
total rate :

$N\sigma$

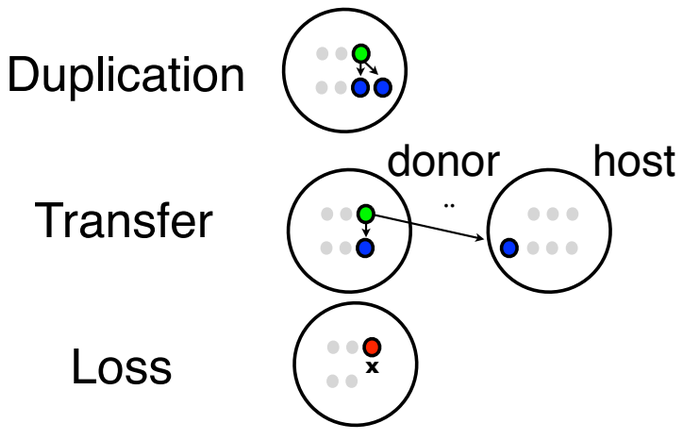


# A minimal model of speciation dynamics

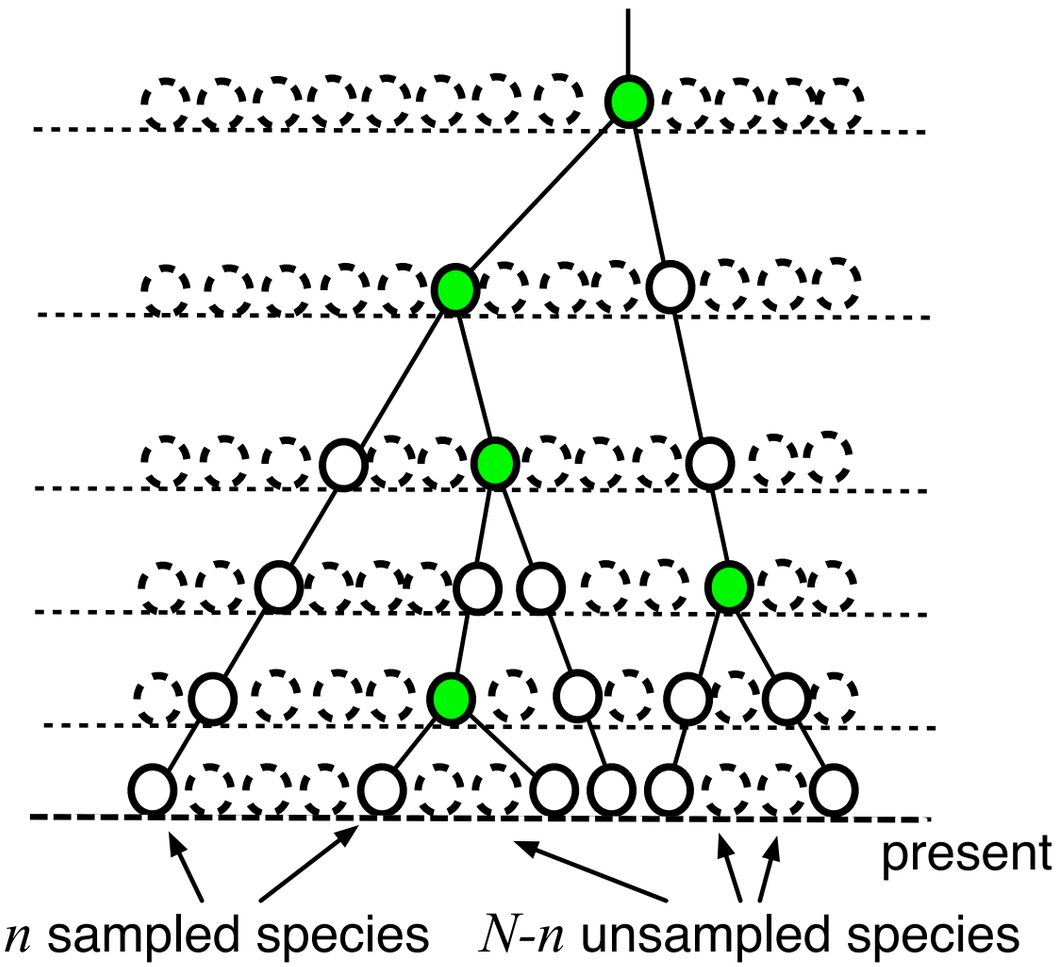
a) speciation dynamics

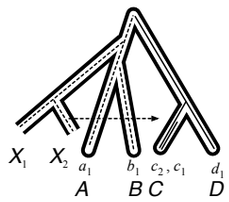


b) gene birth and death



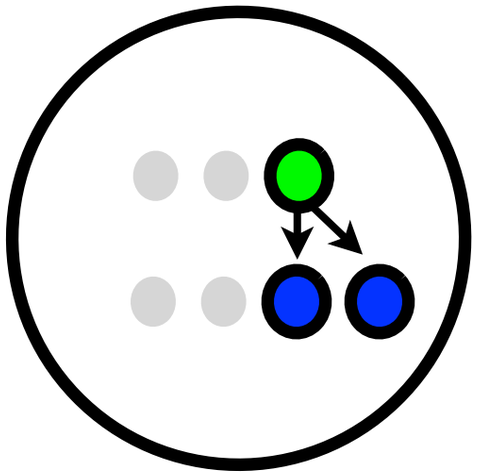

c) represented species phylogeny



# Gene birth-death by **D**uplication, **T**ransfer and **L**oss

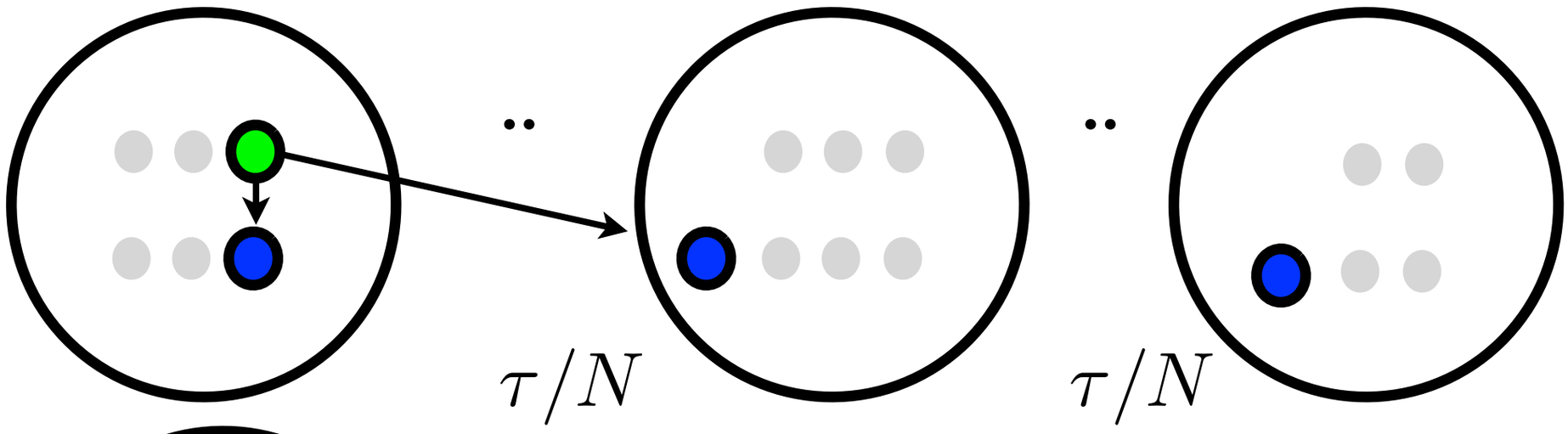
**D**uplication



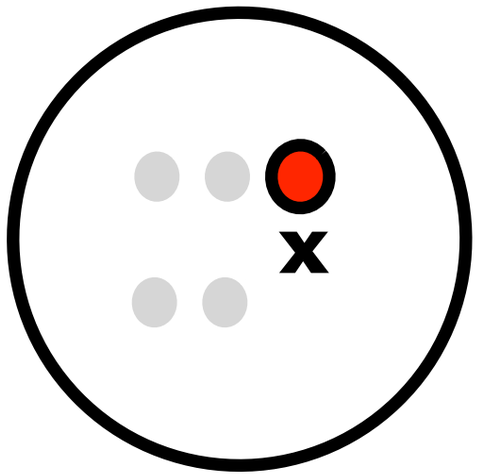
$\delta$ duplication rate

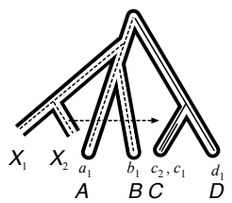
$\tau$ transfer rate

**T**ransfer




**L**oss

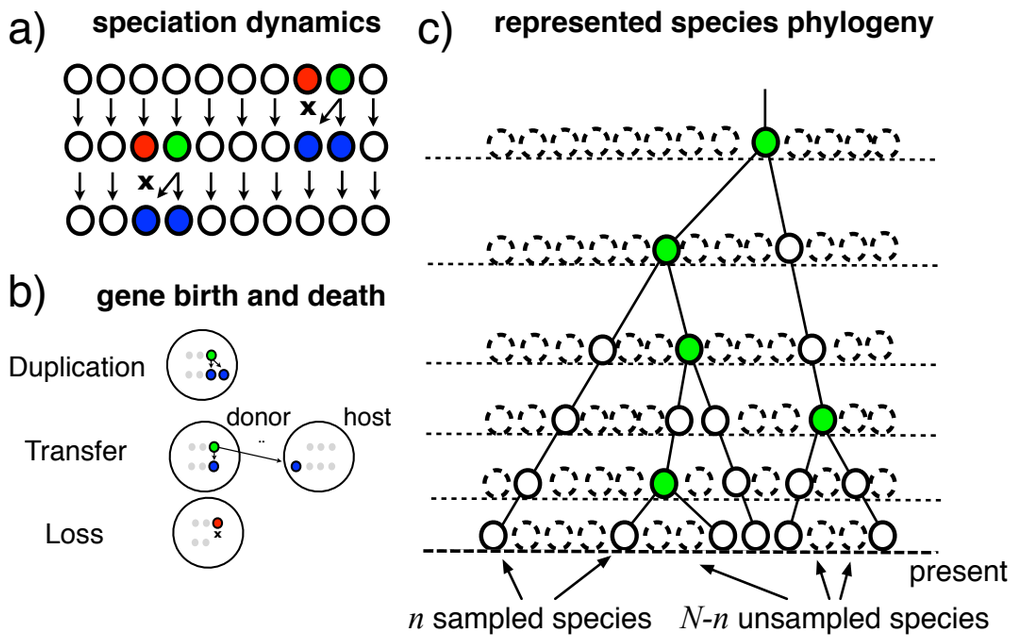


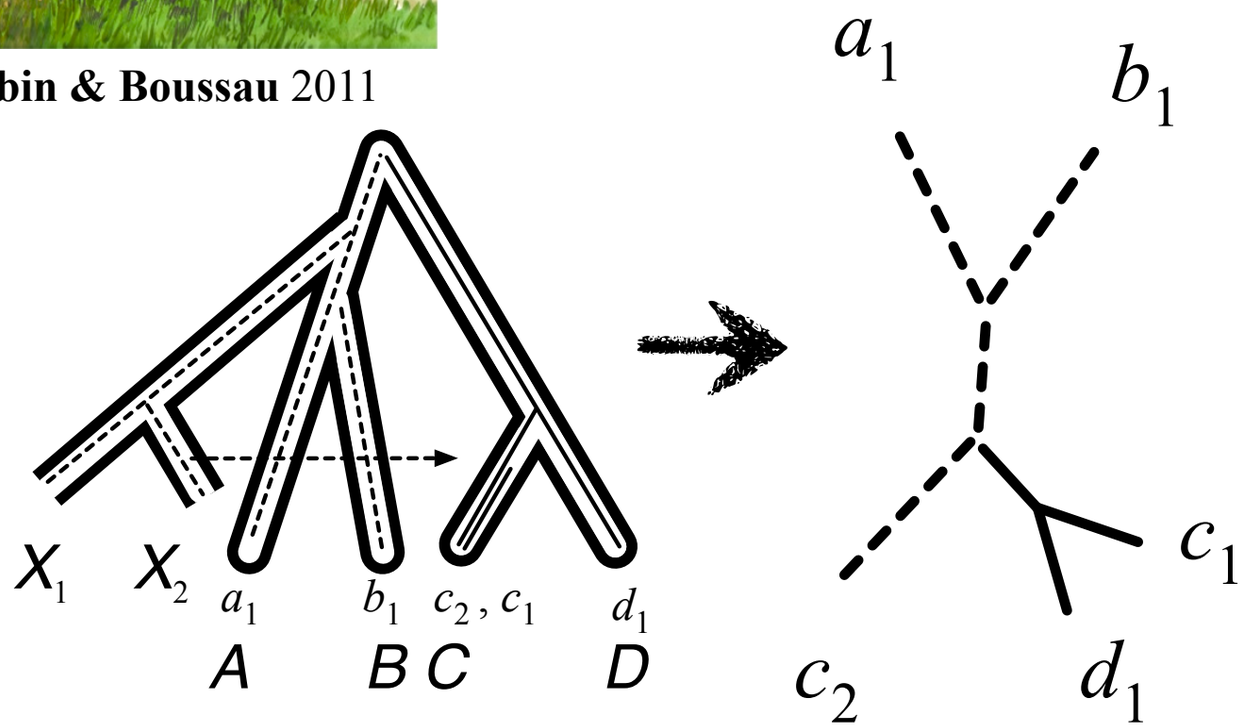


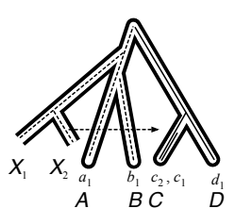
# The combination of speciation dynamics and gene birth and death generates gene trees

$\sigma$ speciation/extinction rate
$\delta$ gene duplication rate
$\tau$ gene transfer rate
$\lambda$ gene loss rate

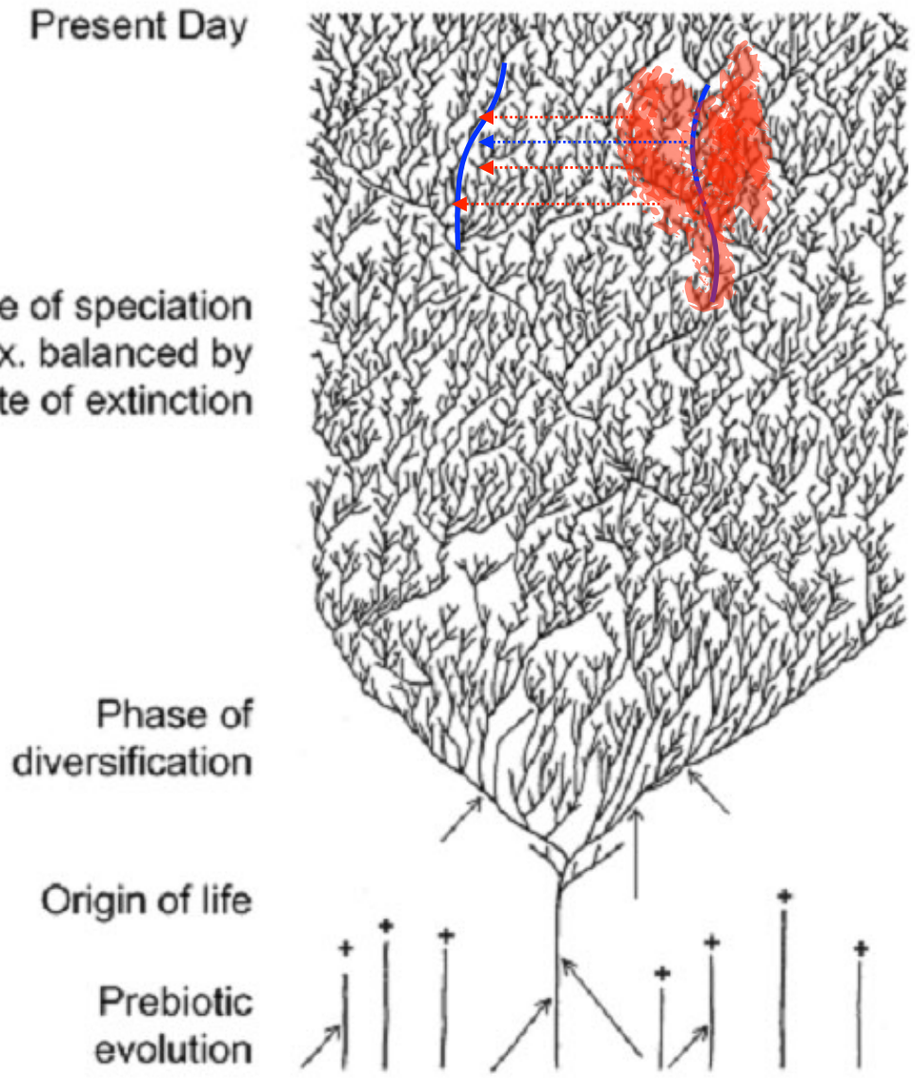
**Daubin & Boussau** 2011

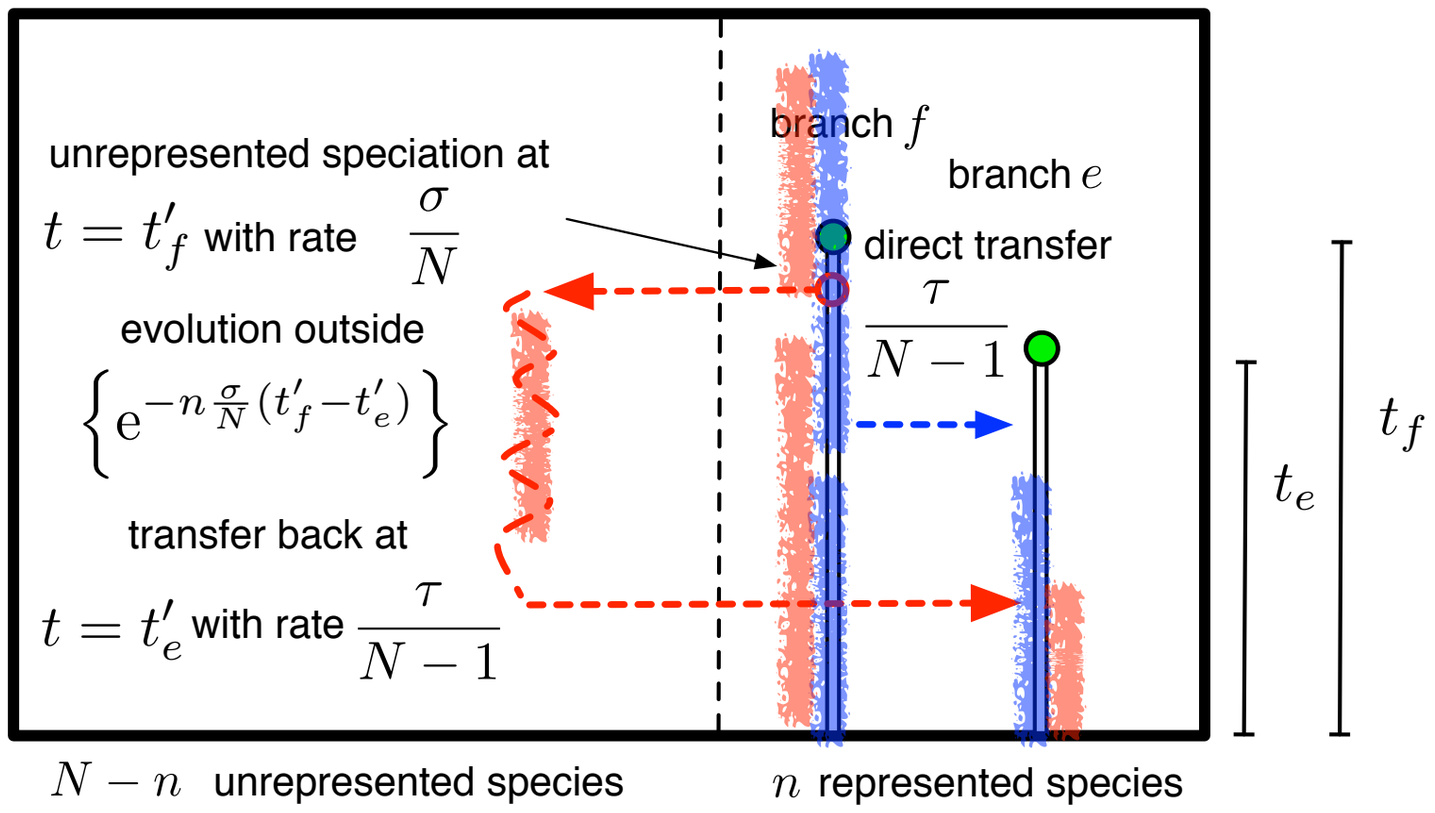


# ..(almost) all transfers are from the dead..




**Gogarten lab**

$$T_{\text{direct}} \approx \int_0^{\frac{N}{n\sigma}} \frac{\tau}{N} \mathrm{d}t'_e = \frac{1}{N}\frac{\tau}{2n}\left[\frac{2N}{\sigma}\right],$$

$$T_{\text{indirect}} \approx \int_0^{\frac{N}{n\sigma}} \int_{t'_e}^{\frac{N}{n\sigma}} \tau \left\{ \mathrm{e}^{-n\frac{\sigma}{N}(t'_f - t'_e)} \right\} \frac{\sigma}{N} \mathrm{d}t'_f \mathrm{d}t'_e = \frac{1}{\mathrm{e}n}\frac{\tau}{2n}\left[\frac{2N}{\sigma}\right],$$

**Szöllősi**, Tannier, Lartillot & Daubin *Systematic Biology* (2013)
*Lateral Gene Transfer from the Dead*

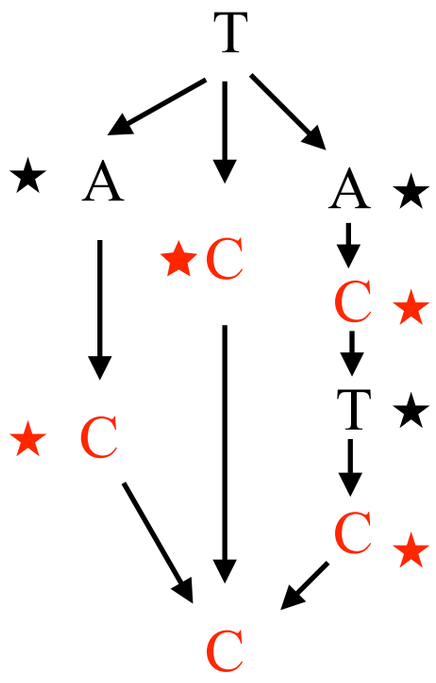
# The story of homologs genes can be reconstructed

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

$$p(A|G)$$

*requires summing over all possible substitution paths.*

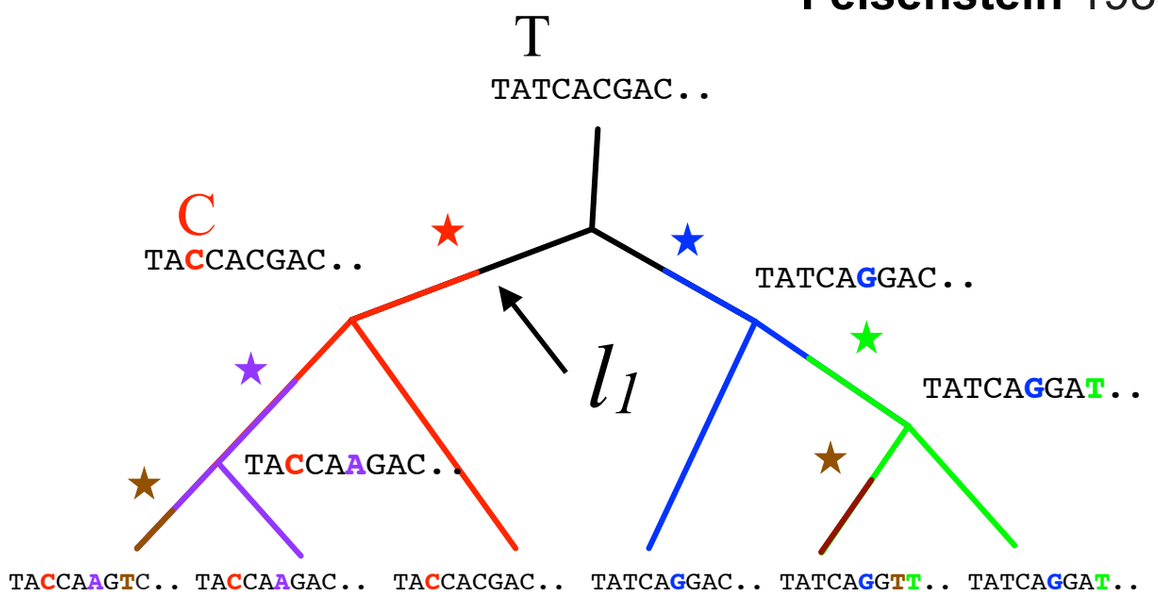
**sum over subs. along branch conditional on states on top and bottom**

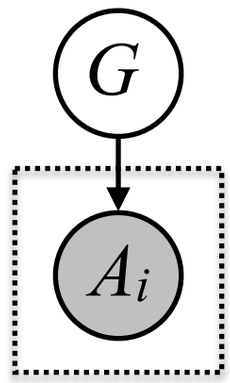



$$P_{\mathrm{TC}}(l_1)$$

**sum over ancestral states**

**Felsenstein** 1981



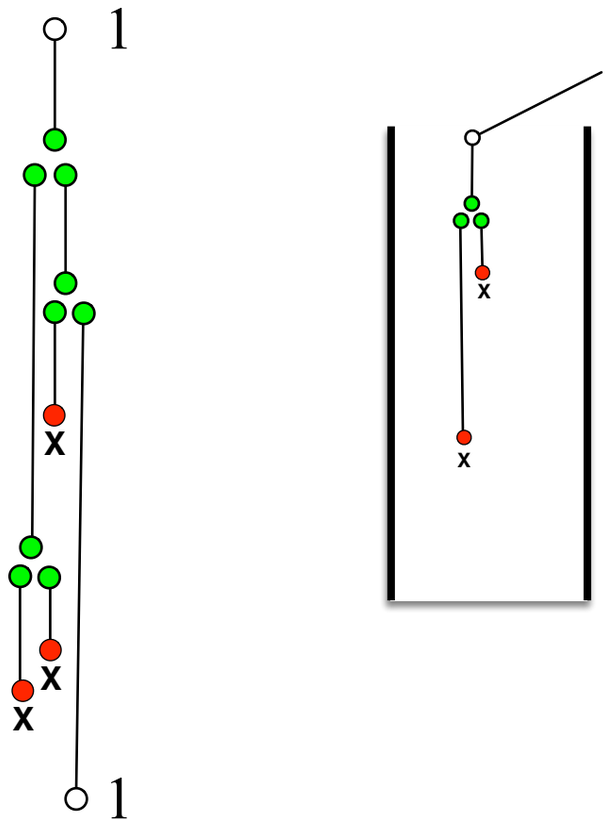

$$= .. \times P_{\mathrm{TC}}(l_1) \times P_{\mathrm{CC}}(l_2) \times P_{\mathrm{CC}}(l_3) \times P_{\mathrm{CC}}(l_4) \times ..$$

# .. but 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. For the DTL process to calculate the likelihood of a ***gene tree*** we sum over all possible ***gene birth and death events*** along a given ***species tree.***

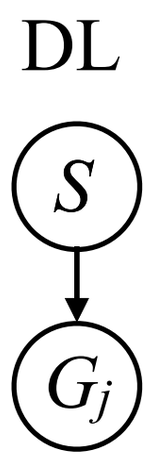
**sum over gene birth and death events along a branch conditional on reconciliation**

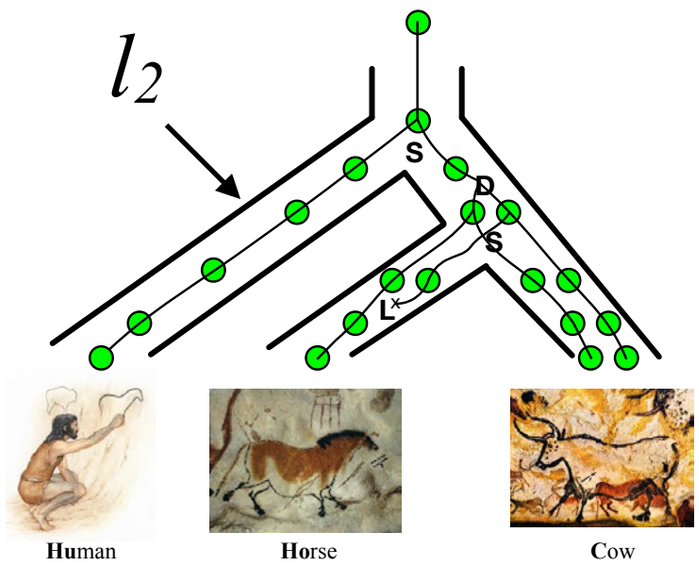
$P_{11}(l_{2,}\mathbf{Hu})$ $P_{10}(\mathbf{Ho})$




**sum over all reconciliations with species tree conditional on gene tree**

**Arvestad** 2010

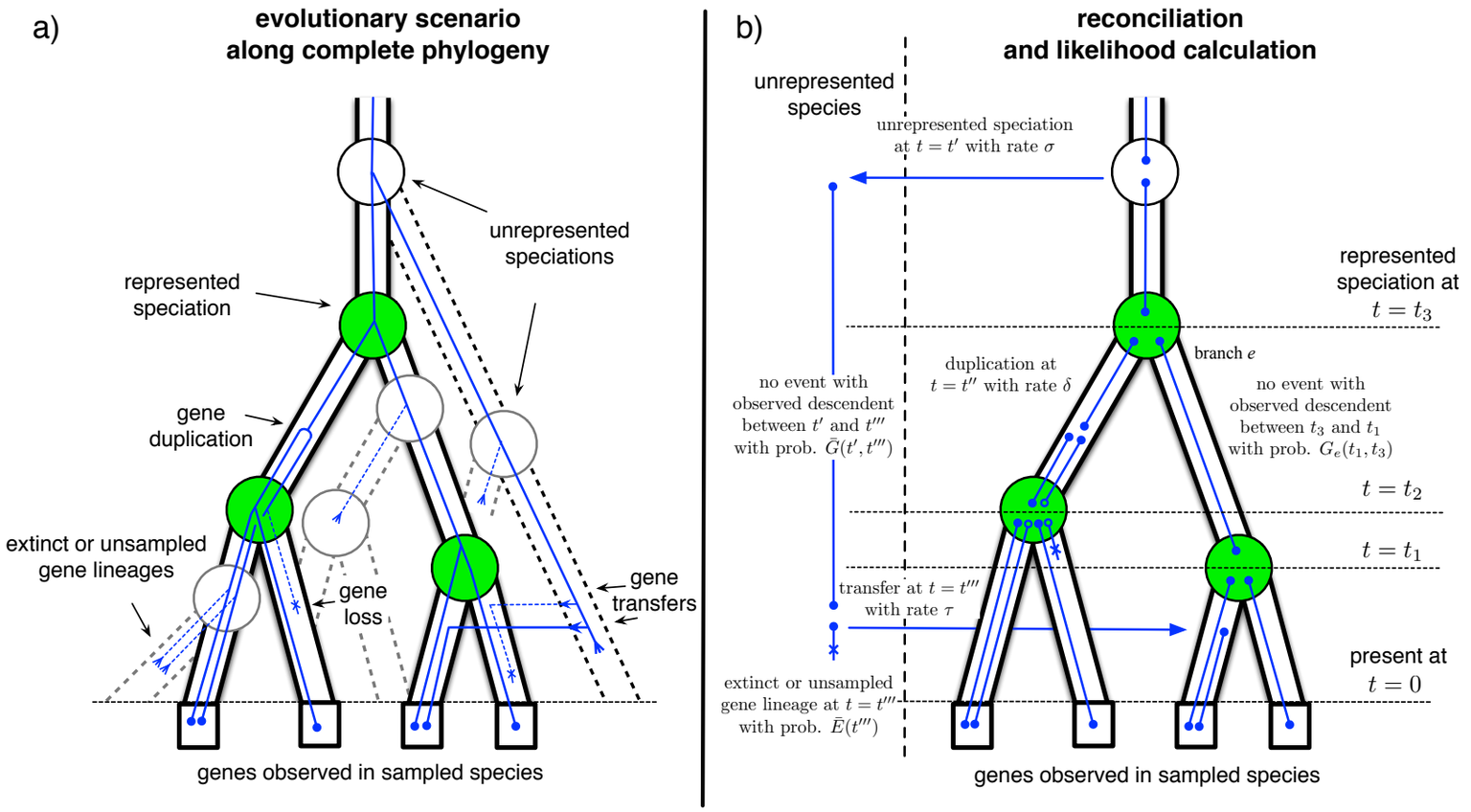


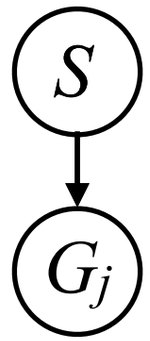
# .. but gene trees are generated along the species tree

Given a model of gene family evolution a species tree induces a probability distribution over gene trees. For the DTL process to calculate the likelihood of a ***gene tree*** we sum over all possible ***gene birth and death events*** along a given ***species tree.***

**calculating** $p(G_i|S)$ **is possible if** $n \ll N$




DTL




implemented in ALE:

**http://github.com/ssolo/ALE**


**Szöllősi**, Tannier, Lartillot & Daubin *Systematic Biology* (2013)
*Lateral Gene Transfer from the Dead*

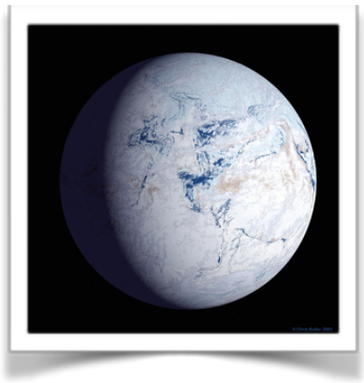
**Szöllősi**, Rosikiewicz, Boussau, Tannier & Daubin *Systematic Biology* (2013)
*Efficient exploration of the space of reconciled gene trees*

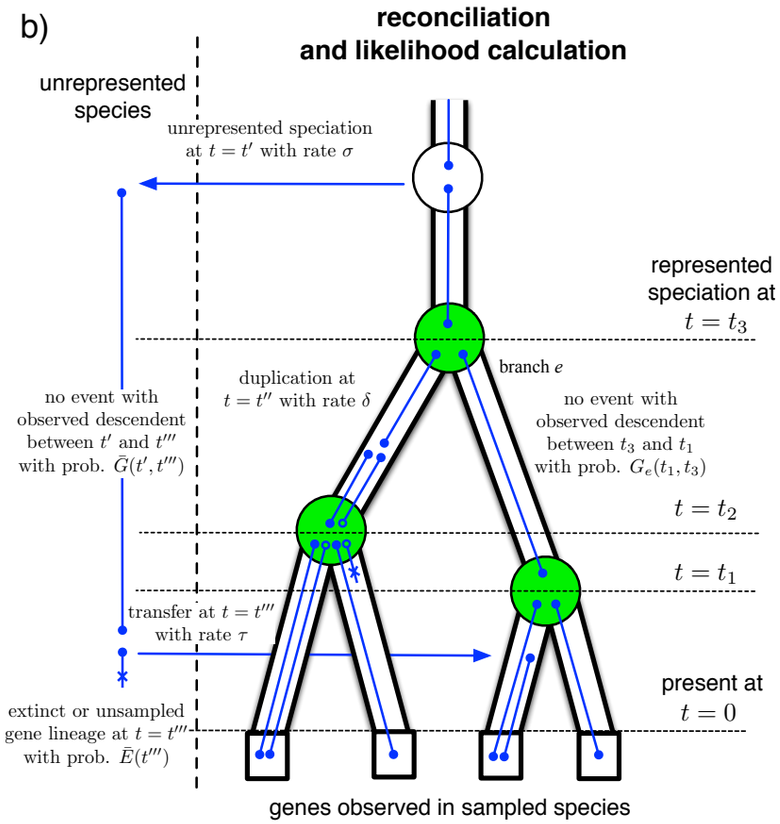
# .. but gene trees are generated along the species tree

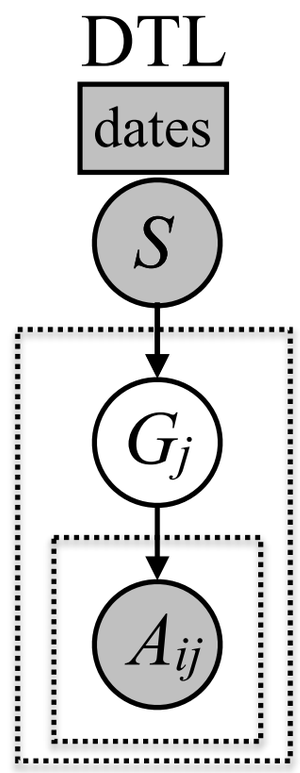
Modeling variation in N, the total number of species, over geological times, could be of particular interest. Indeed, a corollary of the observation that LGT events record evolutionary paths along the complete species tree is that the phylogenies of genes from a limited sample of extant species carry information about extinct lineages, and therefore about the size and dynamics of ancient biodiversity.

**calculating** $p(G_i|S)$ **is possible if** $n \ll N$

**Can we detect mass extinction events?**






implemented in ALE:

**http://github.com/ssolo/ALE**

**Szöllősi**, Tannier, Lartillot & Daubin *Systematic Biology* (2013)
*Lateral Gene Transfer from the Dead*

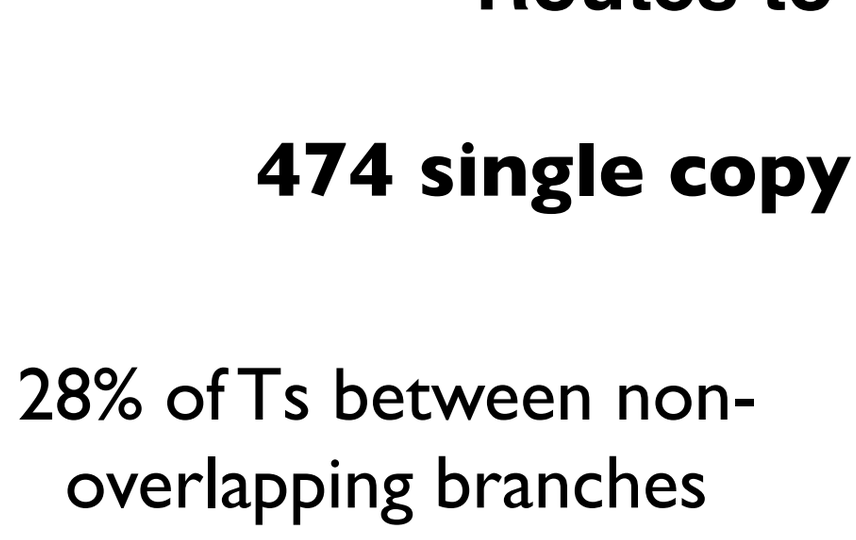
**Szöllősi**, Rosikiewicz, Boussau, Tannier & Daubin *Systematic Biology* (2013)
*Efficient exploration of the space of reconciled gene trees*

# Routes to cyanobacterial genomes

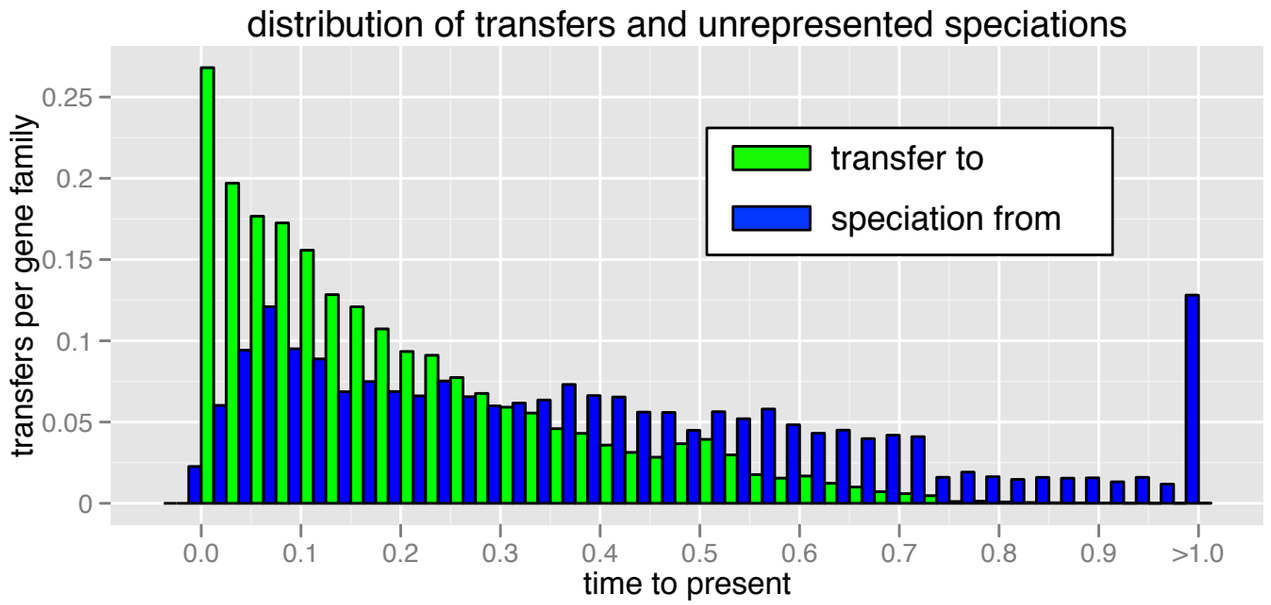
## 474 single copy families from 36 cyanobacteria

28% of Ts between non-overlapping branches

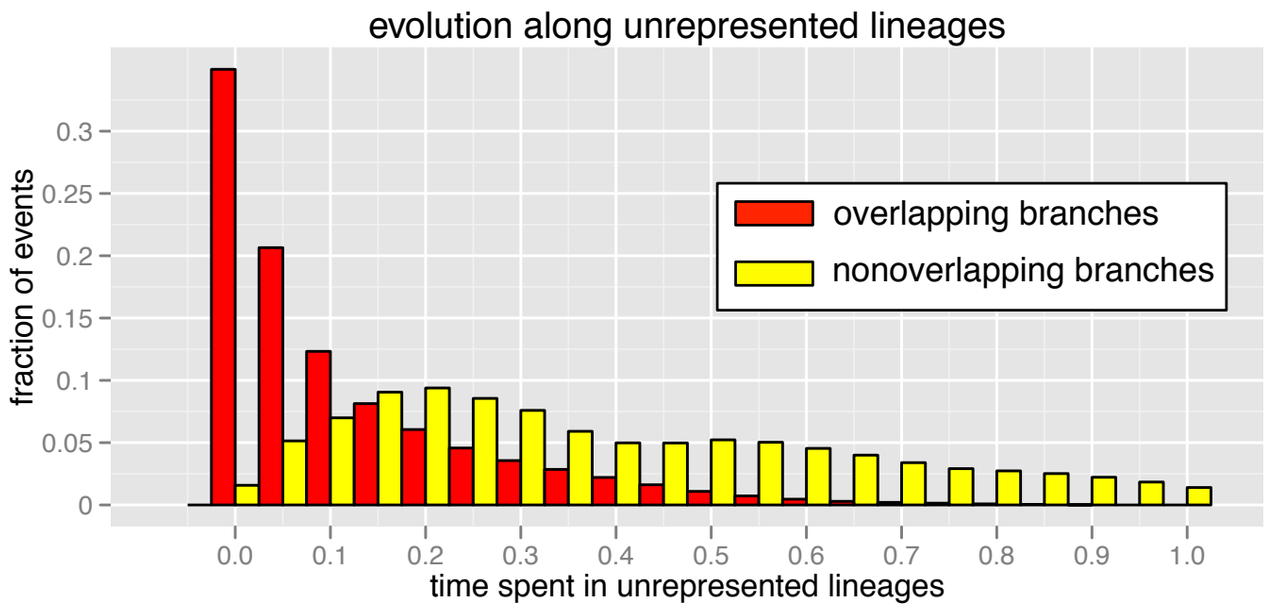
6% from above the root of Cyanobacteria




a)



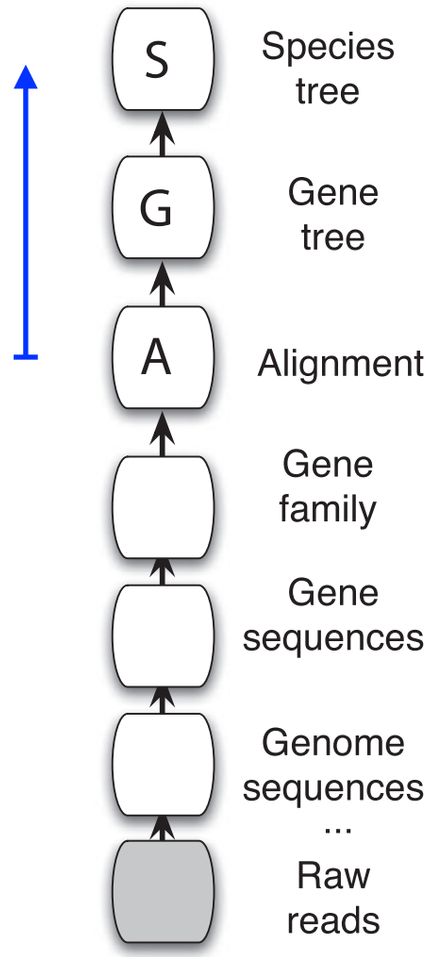

b)



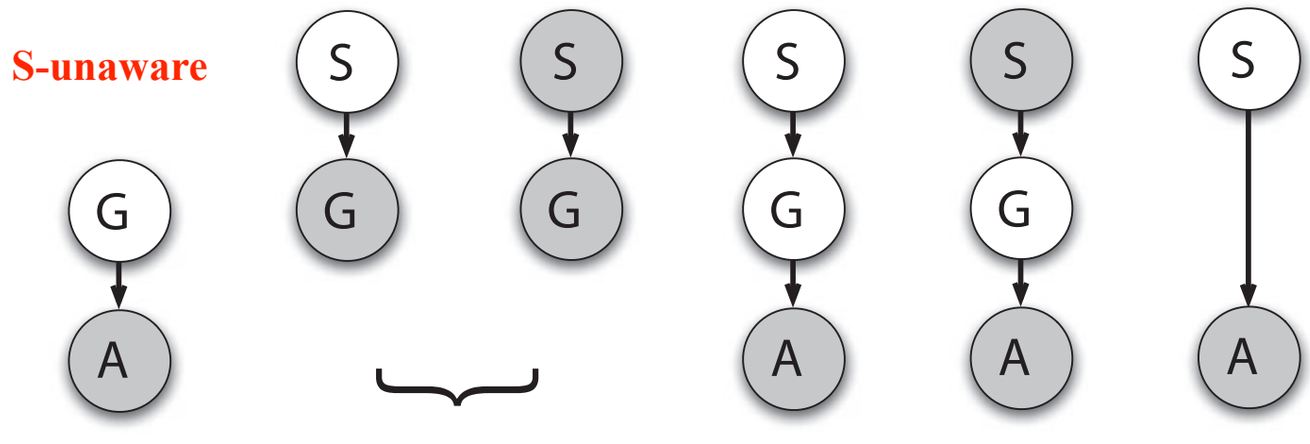

**Szöllősi**, Tannier, Lartillot & Daubin *Systematic Biology* (2013)
*Lateral Gene Transfer from the Dead*

# Species tree-awareness

**Phylogenomics inference pipeline**




**Gene tree-species tree models published in the literature**




P(A|G)

PhyML [Guindon & Gascuel, 2003];
RAxML [Stamatakis et al., 2005];
MrBayes [Ronquist et al., 2012];
BEAST [Drummond et al., 2012];

P(G|S)

BEST [Edwards et al., 2007];
DLCoal [Rasmussen & Kellis, 2012];
MP-EST [Liu et al., 2010];
ODT [Szöllősi et al., 2012];
ASTRAL [Mirarab et al. 2014]

P(A|G,S)

*BEAST [Heled & Drummond, 2008];
PHYLDOG [Boussau et al., 2013];
Prime-GSR [Akerborg et al., 2009];
IMa2 [Hey, 2011];
Prime-DLRS [Sjőstrand et al., 2012];
ALE [Szöllősi et al., 2013b];

P(A|S)

SNAPP [Bryant et al., 2012];
ALE [Szöllősi et al., 2013b];

**Szöllősi,.., Boussau** 2015 Syst. Biol.

# Efficiently integrating over the space of reconciled gene trees

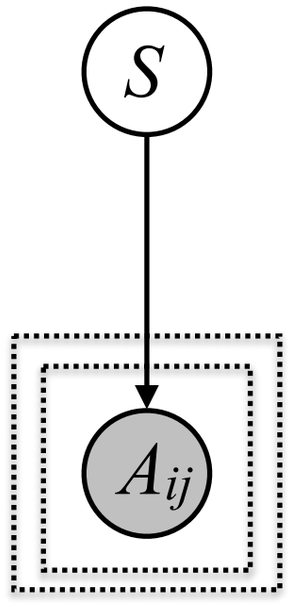
using ALE we can approximate the integral in the DTL version of the Felsenstein equation:

$$P(A|S, \text{rates}) = \int_G p(A|G)p(G|S, \text{rates})$$

**Felsenstein** 1988

This opens the possibility for efficient MCMC or ML exploration of $\mathcal{L}(S, \text{rates}|A)$

DTL




**Szöllősi**, Rosikiewicz, Boussau, Tannier & Daubin *Systematic Biology* (2013)
*Efficient exploration of the space of reconciled gene trees*

**Szöllősi**, Tannier, Daubin & Boussau *Systematic Biology* (2015)
*The inference of gene trees with species trees*

# Efficiently exploring the space of reconciled gene trees

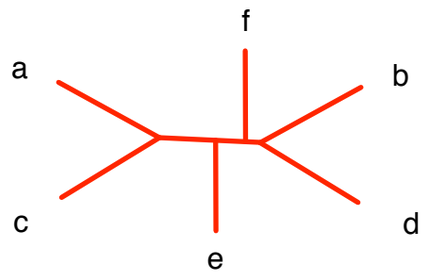
Based on a sample of trees conditional clade probabilities can be used to estimate posterior probability of any gene tree that can be amalgamated. This is usually a very large number of trees (e.g. for $10^4$ samples $10^{12}$ trees, but up to $10^{40}$).

**Felsenstein** 1981

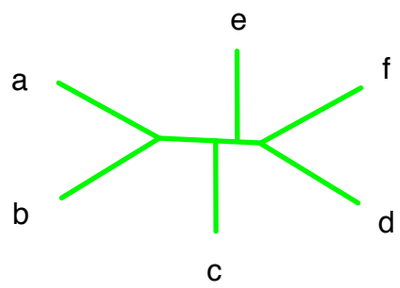
$p(A|G)$

**Gene tree sample**

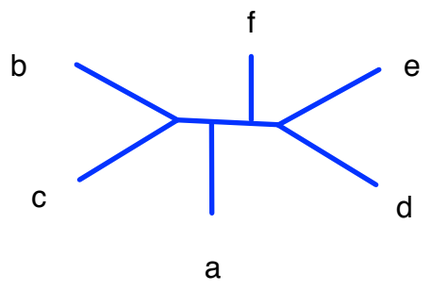
$\approx$ 2/10 **2X**




$\approx$ 3/10 **3X**



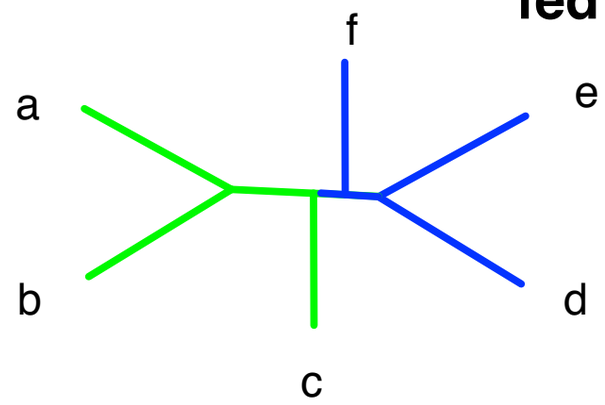

$\approx$ 5/10 **5X**




$$p(A|G) \approx 3/8 \times 5/10$$

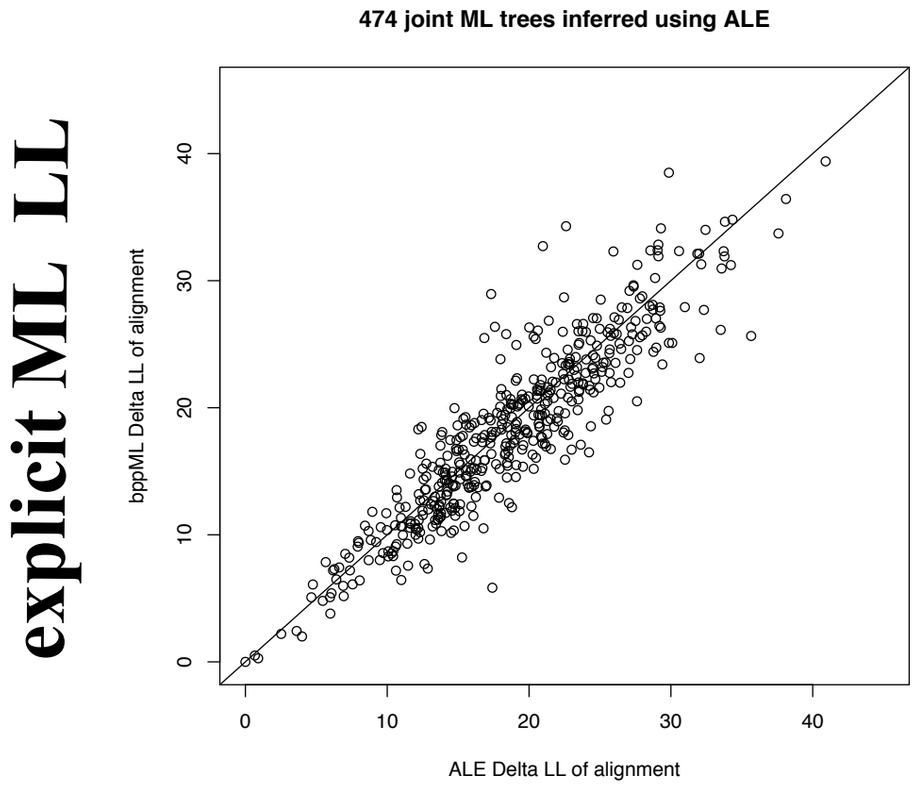
$$\frac{\textbf{ab-c}}{\textbf{abc}} = 3/8$$

$$\frac{\textbf{ed-f}}{\textbf{fed}} = 5/10$$



# Efficiently exploring the space of reconciled gene trees

Based on a sample of trees conditional clade probabilities can be used to estimate posterior probability of any gene tree that can be amalgamated. This is usually a very large number of trees (e.g. for $10^4$ samples $10^{12}$ trees, but up to $10^{40}$).

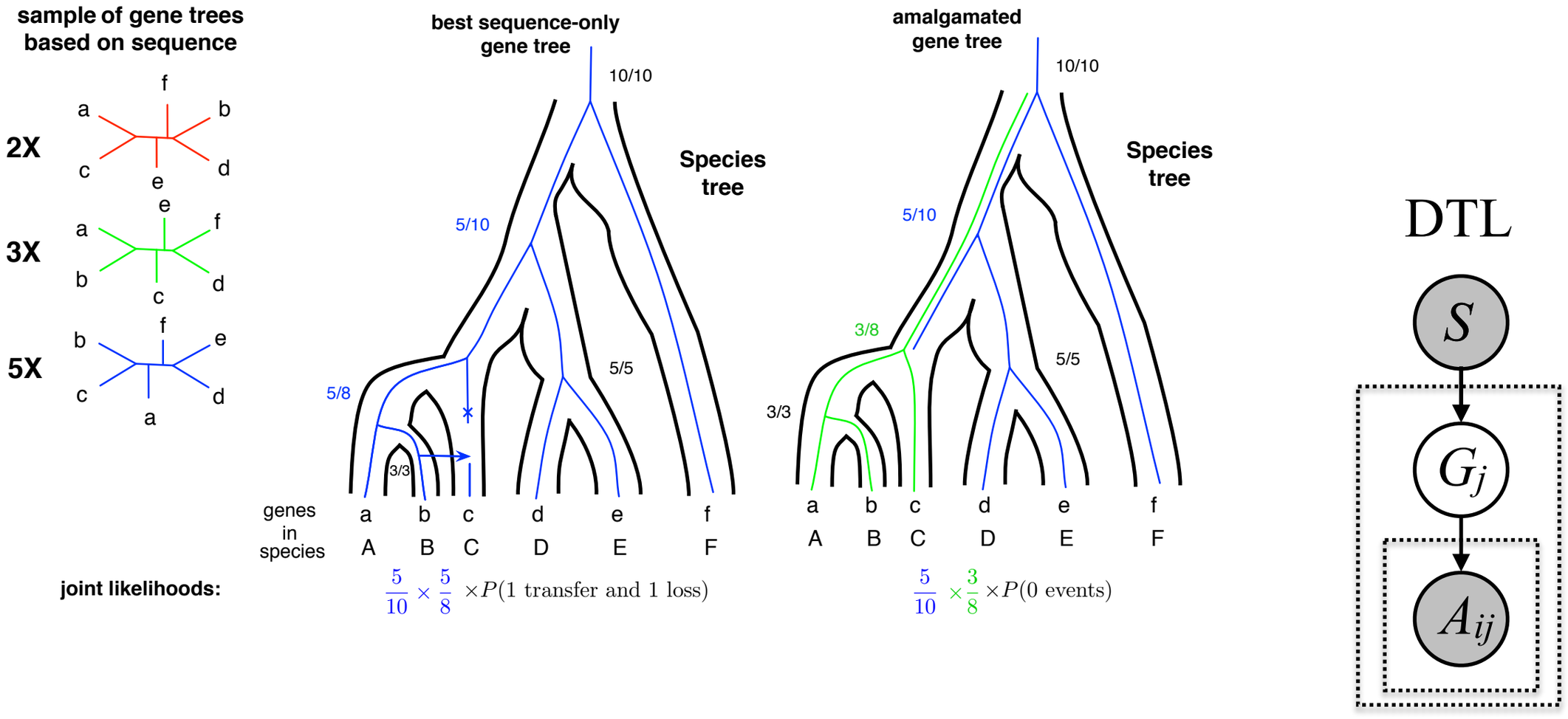



explicit ML LL

ALE LL

# Efficiently exploring the space of reconciled gene trees

Based on a sample of trees conditional clade probabilities can be used to estimate posterior probability of any gene tree that can be amalgamated. This is usually a very large number of trees (e.g. for $10^4$ samples $10^{12}$ trees, but up to $10^{40}$). *The dynamic programming used in gene tree-species tree reconciliation can be extended to approximate the joint likelihood efficiently for a very large set of gene trees.*




implemented in ALE:

**http://github.com/ssolo/ALE**


**Szöllősi**, Tannier, Lartillot & Daubin *Systematic Biology* (2013)
*Lateral Gene Transfer from the Dead*

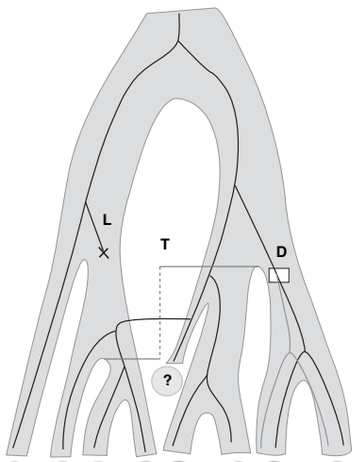
**Szöllősi**, Rosikiewicz, Boussau, Tannier & Daubin *Systematic Biology* (2013)
*Efficient exploration of the space of reconciled gene trees*

**Szöllősi**, Tannier, Daubin & Boussau *Systematic Biology* (2015)
*The inference of gene trees with species trees*

# Using phylogenetic incongruence to reconstruct a dated ToL

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.

**Rooted dated species tree**

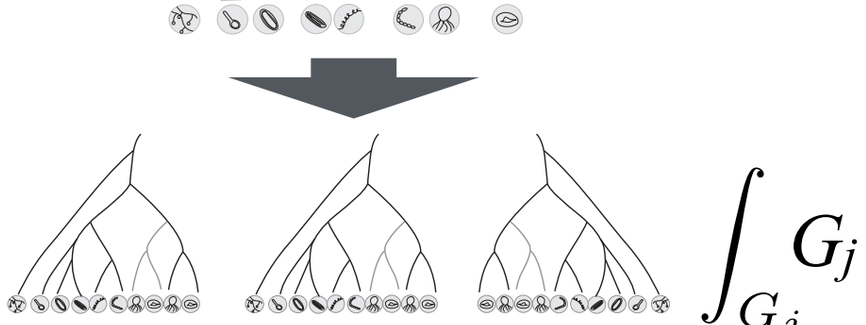



**species tree**

$S$

**Daubin & Boussau** 2011

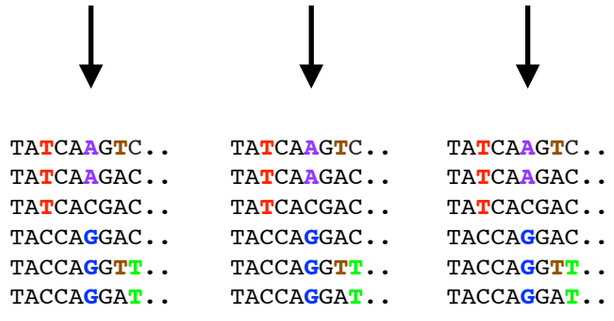
**Fossils & HGT**



$\int_{G_j} G_j$

**gene trees**

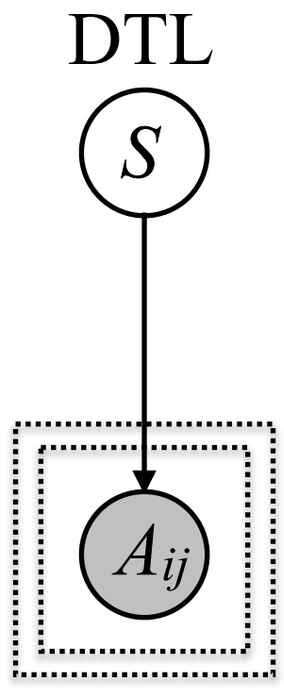
**Alignment of homologous genes from complete genomes**




$A_{ij}$

**alignments**

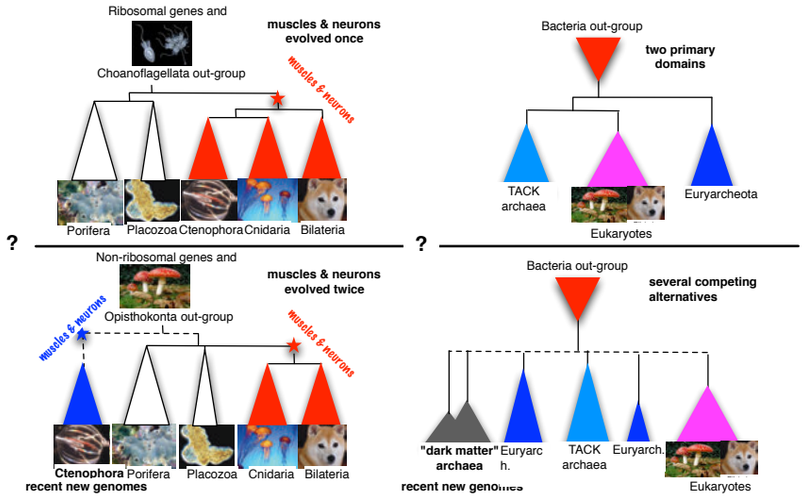
DTL



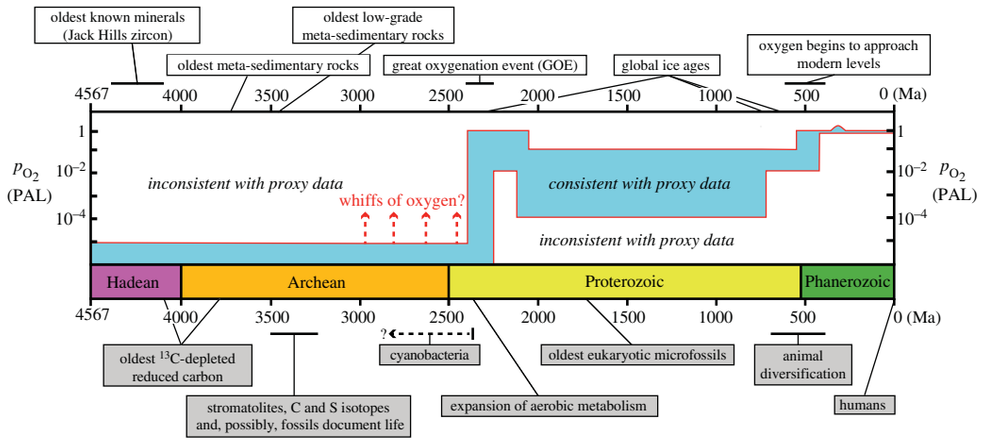
# Using phylogenetic incongruence to reconstruct a dated ToL

Using a hierarchical model wherein gene trees are generated along the species tree and sequences are generated along gene trees we can *jointly* infer gene trees and species trees from sequences.

## Rooted dated species tree




## Fossils & HGT





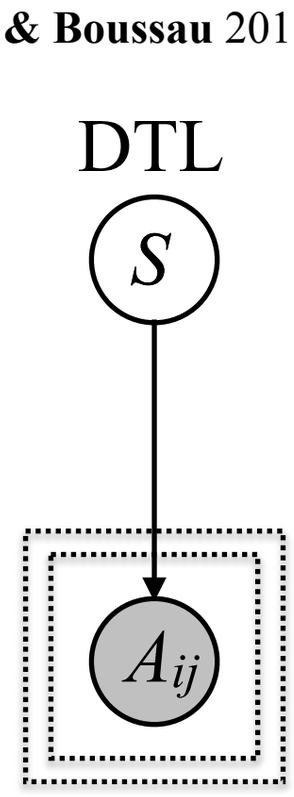


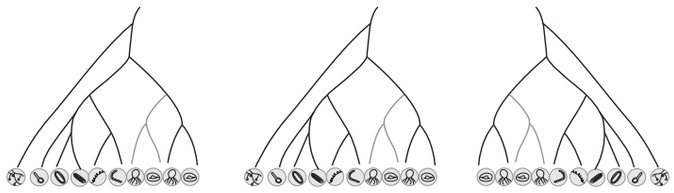
**species tree**

$S$

**Daubin & Boussau** 2011

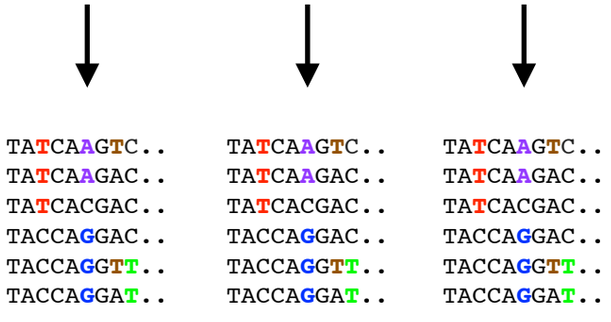
DTL



$\int_{G_j} G_j$

**gene trees**




$A_{ij}$

**alignments**

**GENECLOCKS**