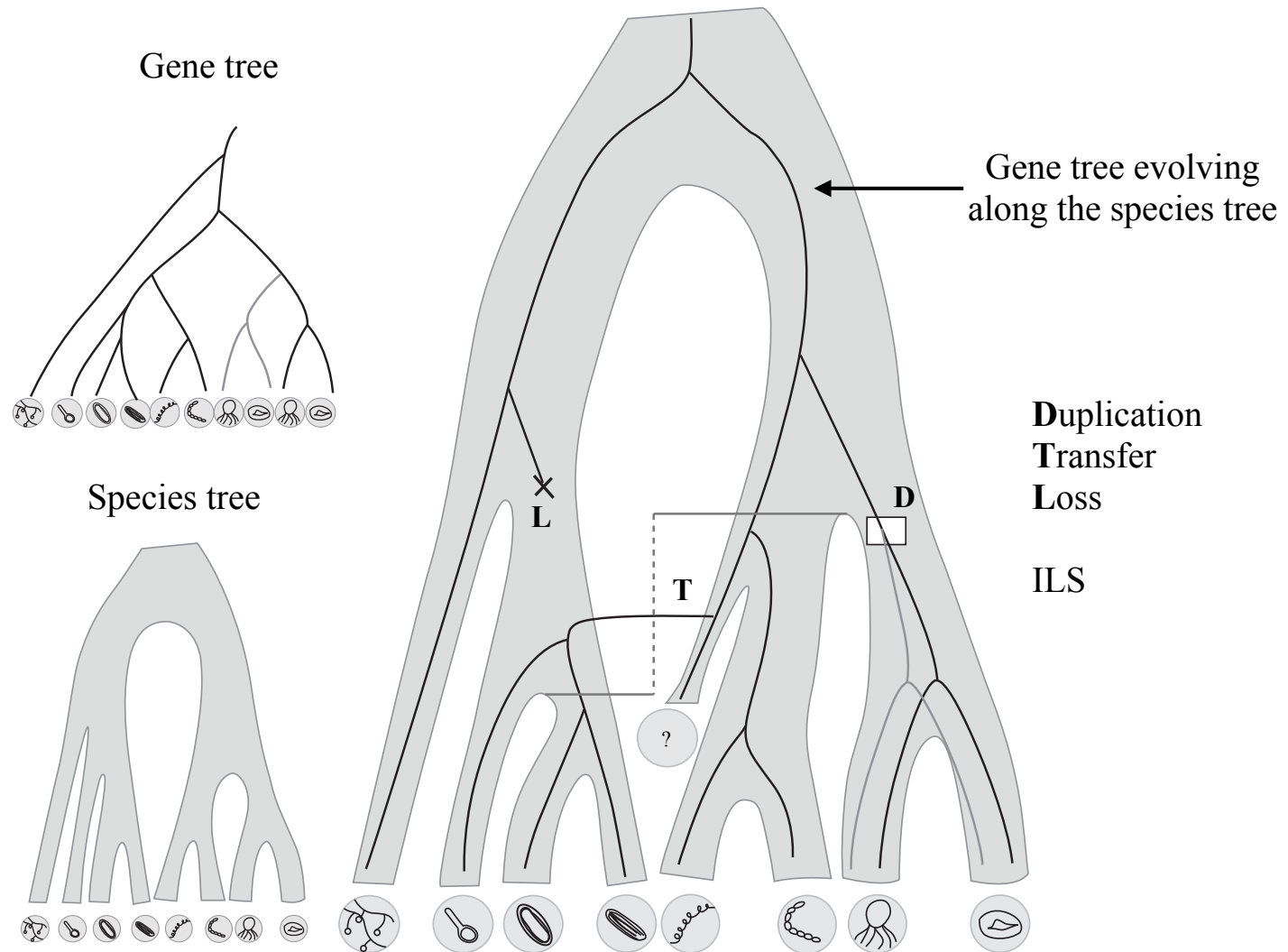


The problem is gene trees are not species trees

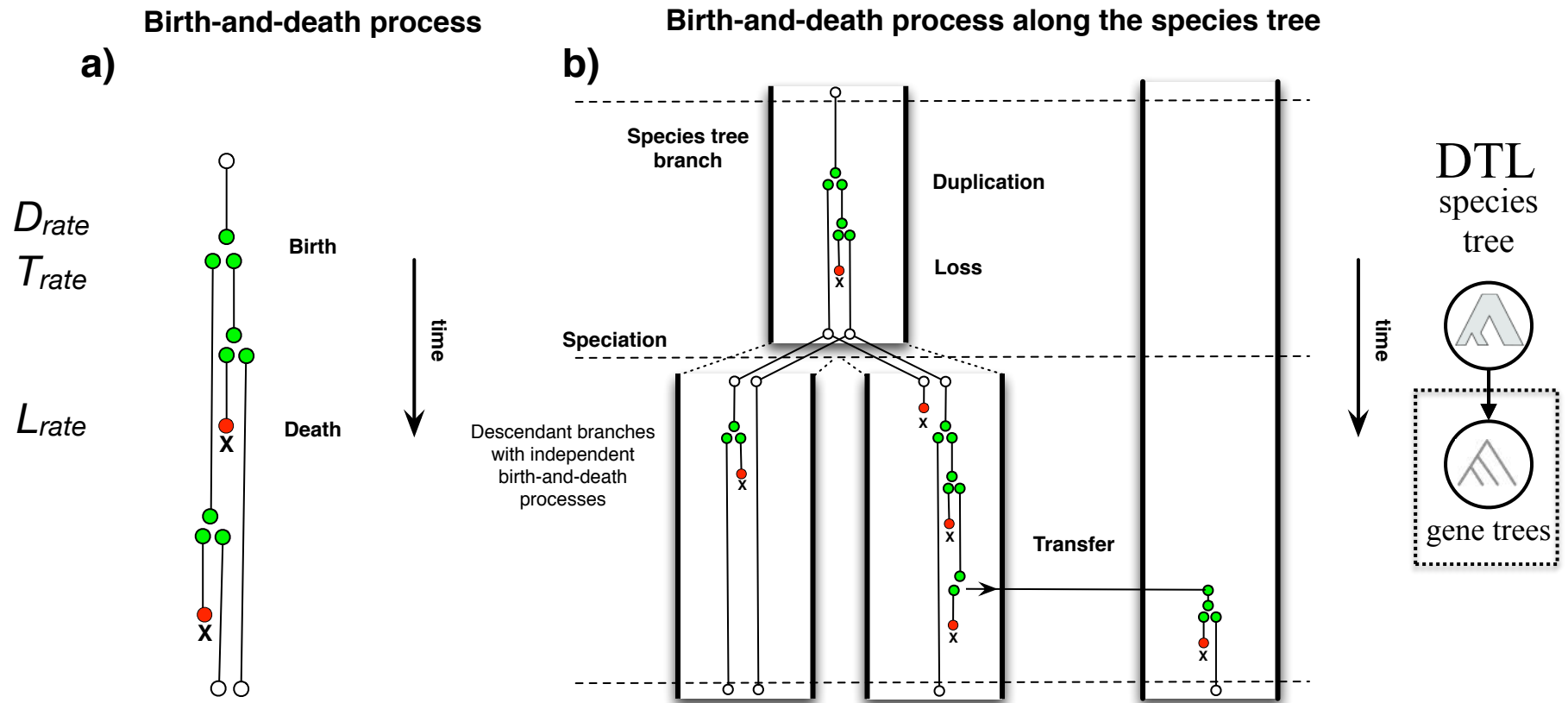
A gene tree is a deformation of the species tree through the prism of genome evolution and population genetics processes.



.. gene trees are generated along the species tree

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

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

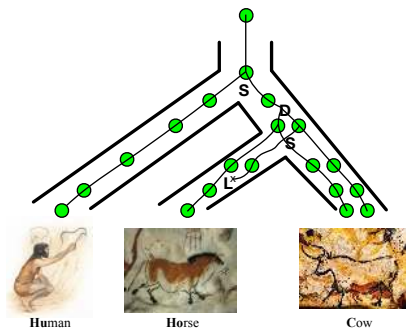


Tofigh PhD thesis 2009

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

.. 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

$$\sim 10 \times \log(\text{\#species}) \times \text{\#genes}$$

DTL

$$\sim 10 \times \text{\#species}^2 \times \text{\#genes}$$

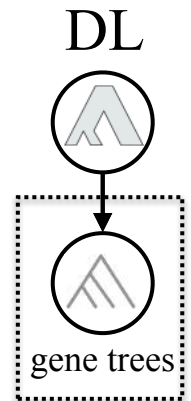
parameters
(ML or Bayes)

DL

D&L rates
branch lengths, root

DTL

D,T&L rates
dated tree



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

Given the **species tree**, which **gene tree** produced my **sequences**? ..
and in **what evolutionary context**?

Joint likelihood: $P(\text{alignment} | \text{species tree}) P(\text{species tree} | \text{alignment})$

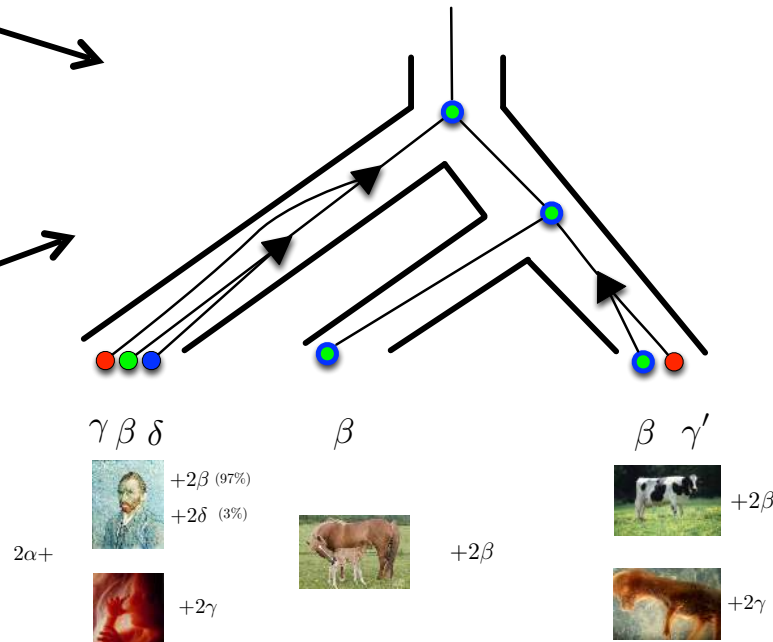


Gene family alignment

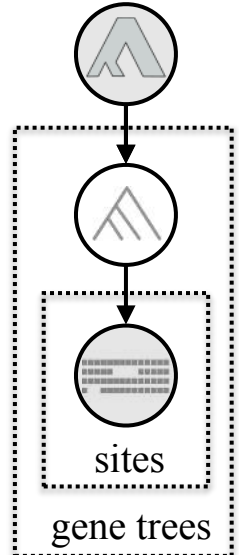


Rooted binary species tree

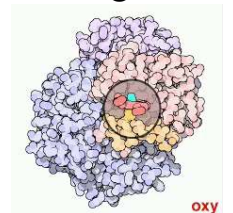
Most likely **species tree-aware** gene tree



species tree



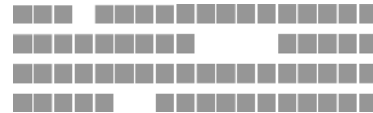
hemoglobin



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

Given the **species tree**, which **gene tree** produced my **sequences**? ..
and in **what evolutionary context**?

Joint likelihood: $P(\text{alignment} | \text{gene tree}) P(\text{gene tree} | \text{species tree})$

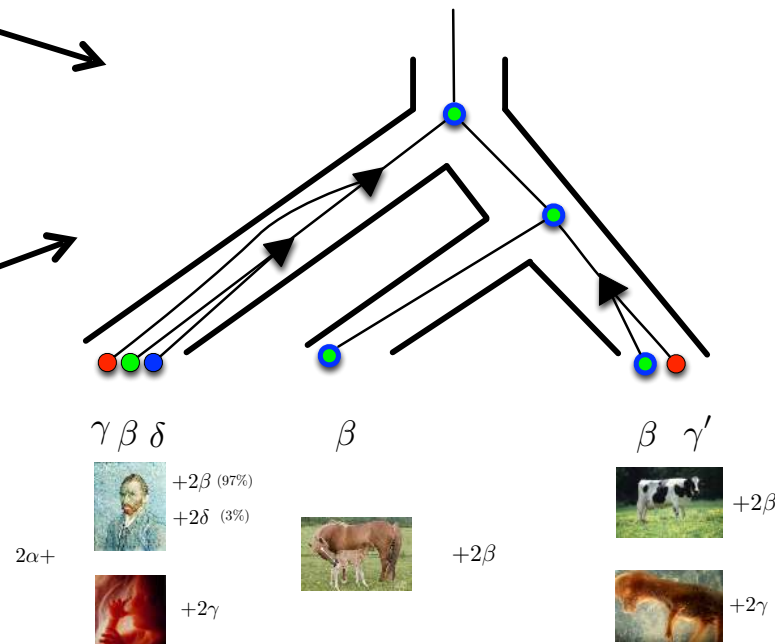


Gene family
alignment

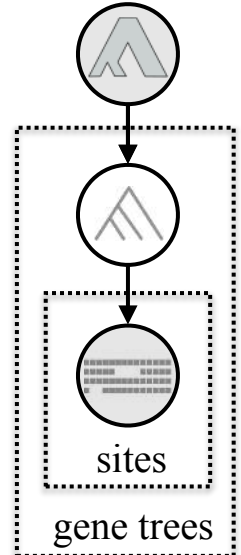


Rooted binary
species tree

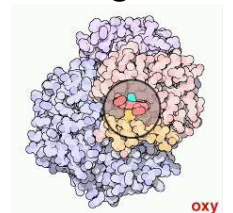
Most likely **species tree-aware** gene tree



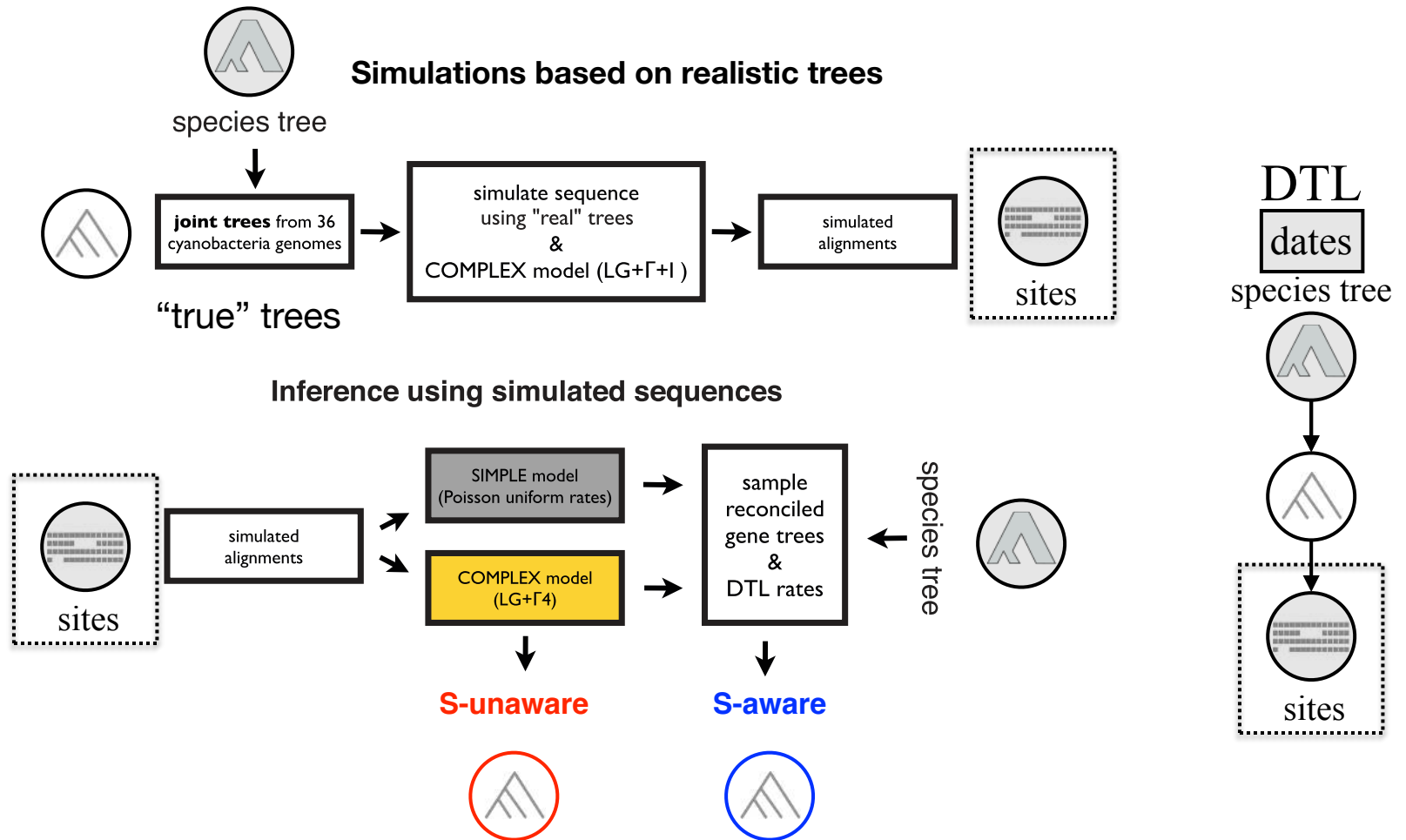
species tree



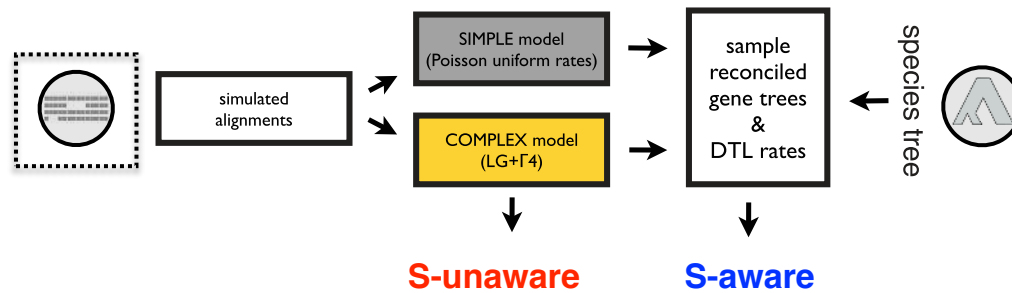
hemoglobin



“Realistic simulations” suggest S-aware methods are important

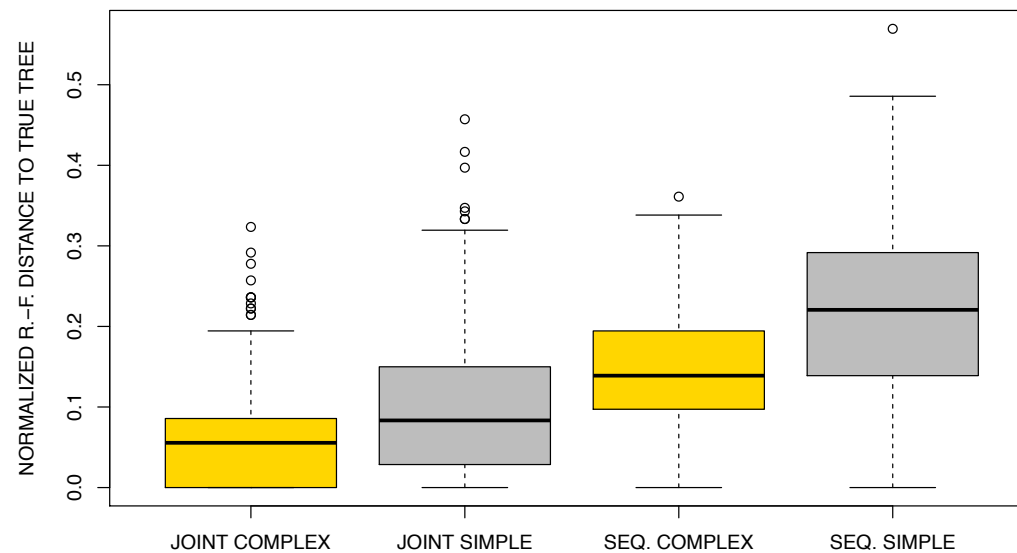


“Realistic simulations” suggest S-aware methods are important



Distance to true tree

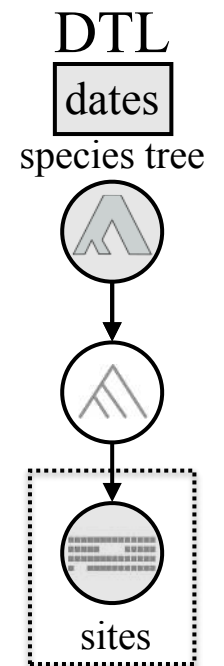
(based on 474 near universal single copy fams.)



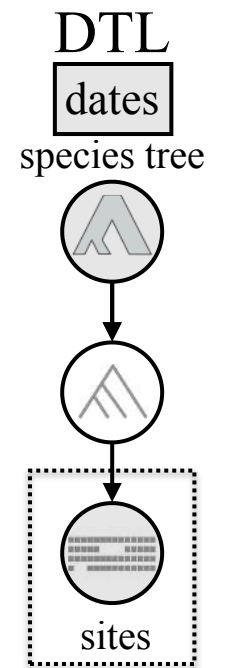
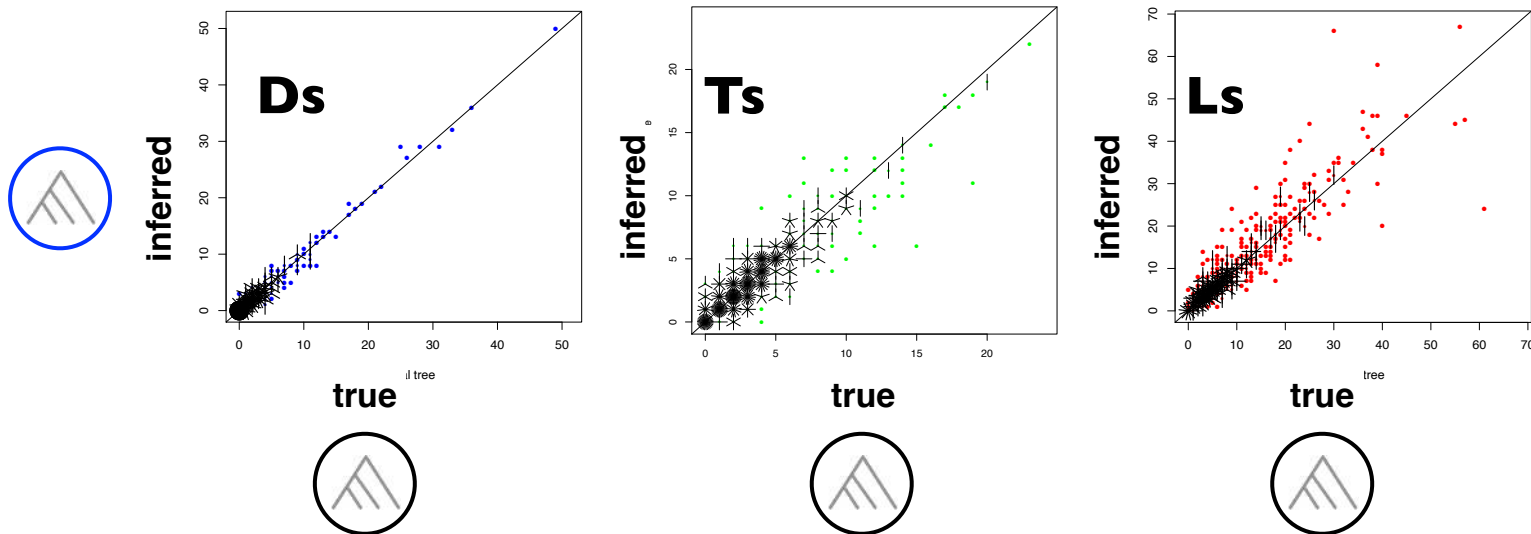
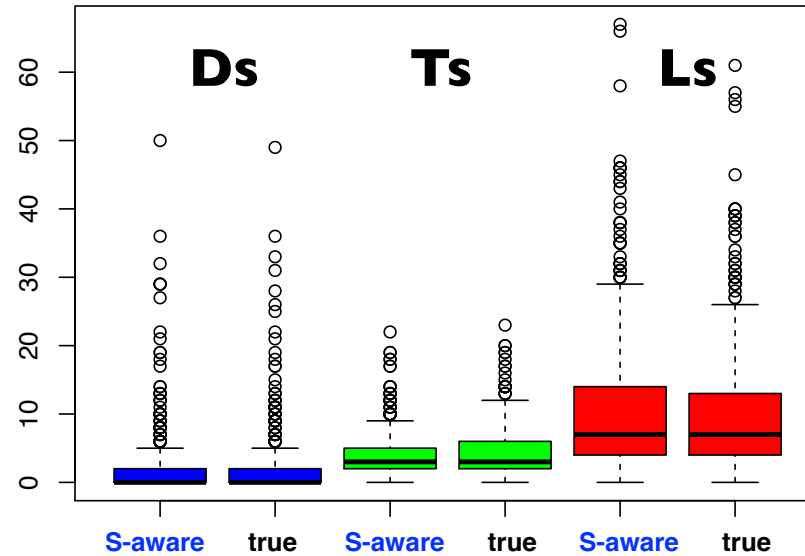
S-aware



S-unaware

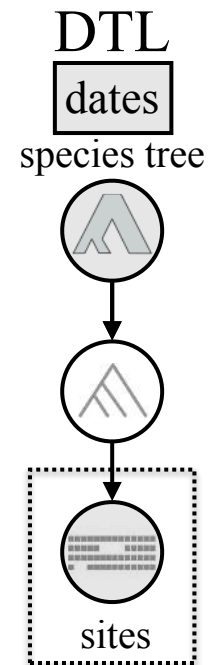
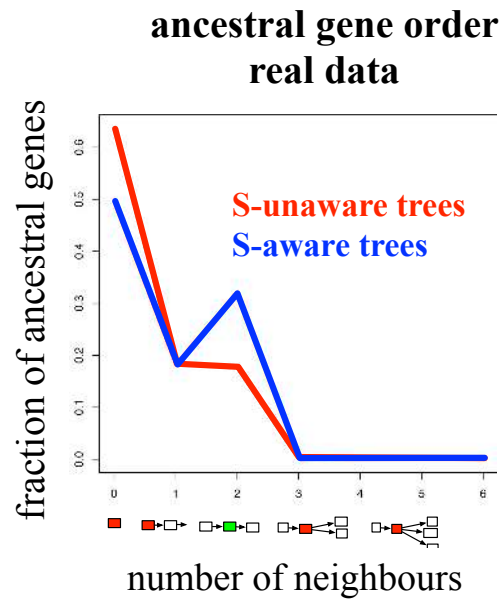


S-aware methods accurately recover number of DTL events



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>

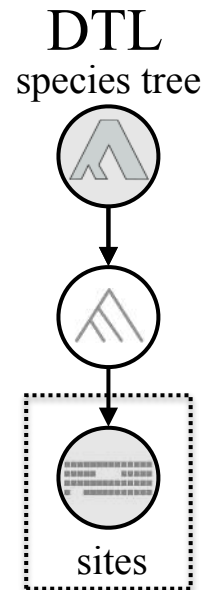
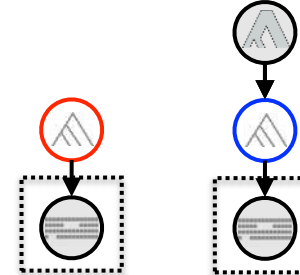
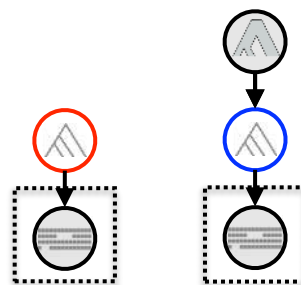
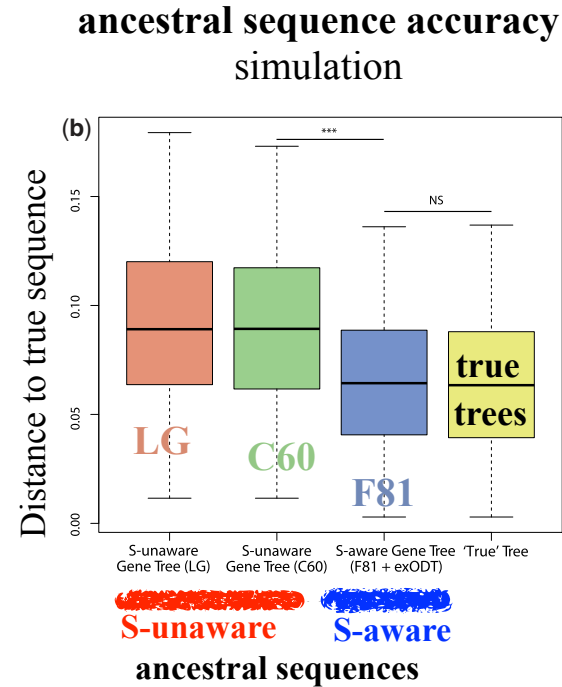
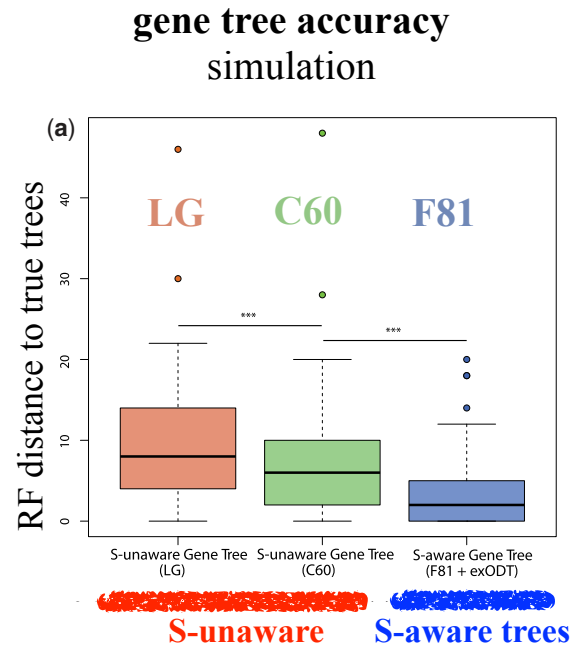
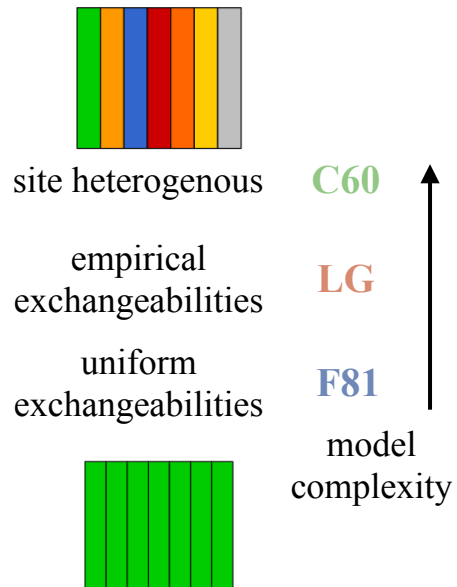
Groussin, Hobbs, Szöllősi, Gribaldo, Arcus & Gouy *Mol. Biol. Evol.* (2015)
Toward More Accurate Ancestral Protein Genotype–Phenotype Reconstructions with the Use of Species Tree-Aware Gene Trees

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

S-aware methods are better at reconstructing gene trees..

S-aware gene trees and ancestral sequences reconstructed along them are significantly more accurate according to simulations



S-aware method are better at reconstructing gene trees..

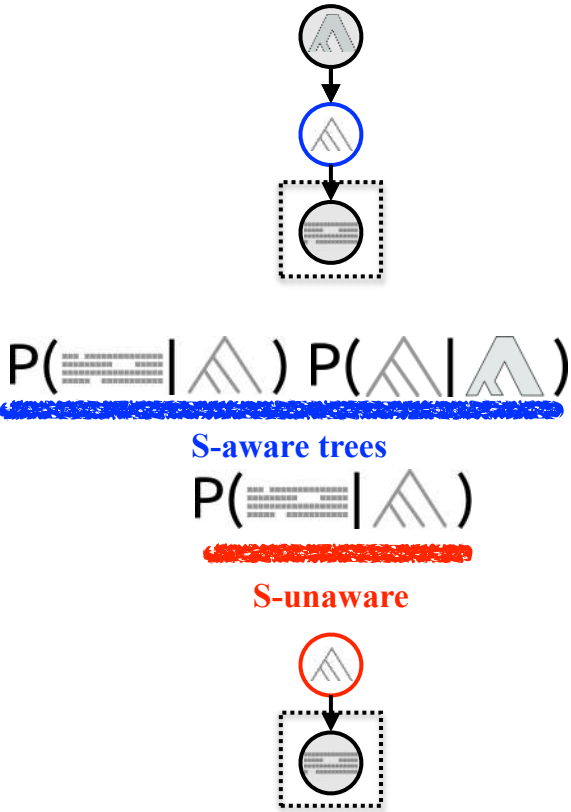
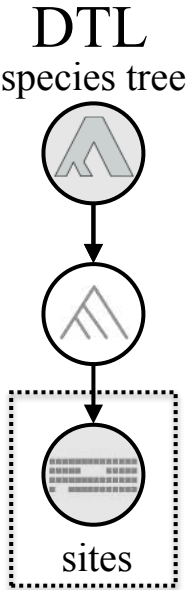
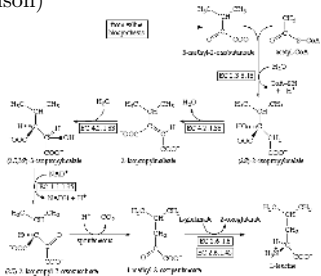
Ancestral sequences reconstructed along S-aware gene trees produce a biochemically more realistic and thermodynamically more stable ancestral protein.

in vitro biochemical essay
on reconstructed ancestral sequences

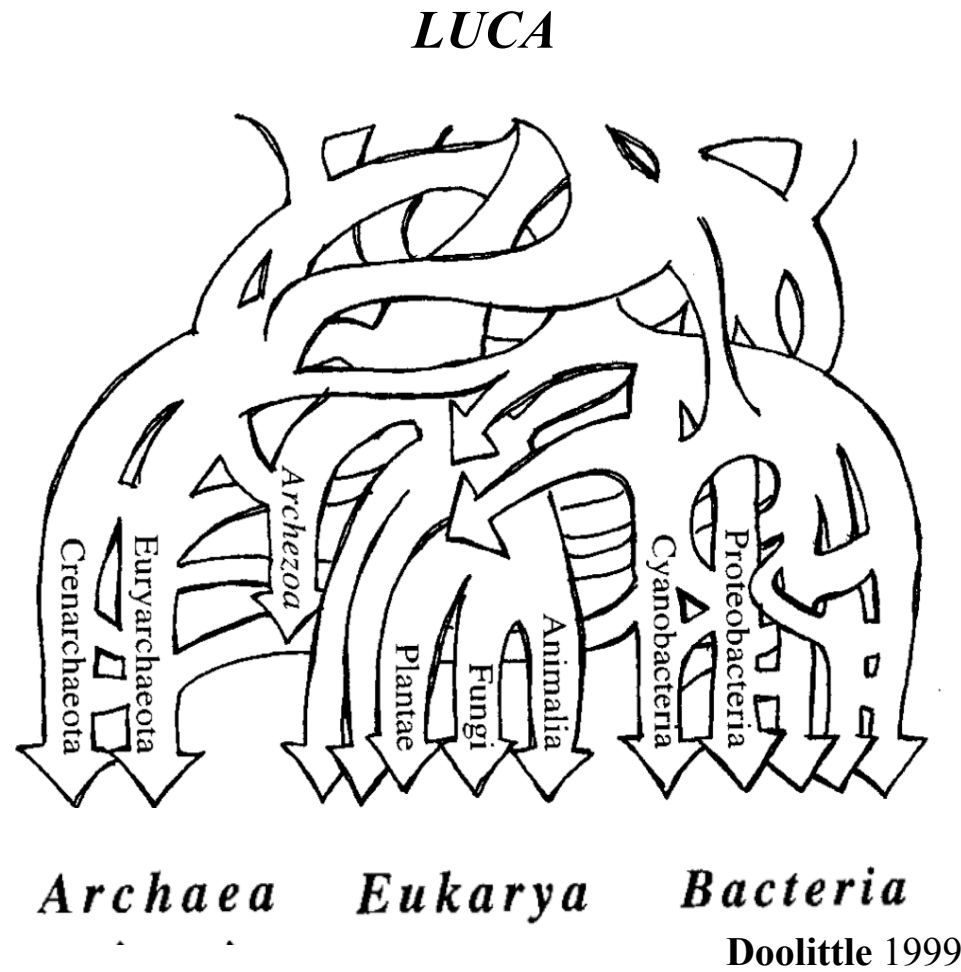
Enzyme	K_M^{app} (mM)	T_{opt} (°C)	$\Delta G_{N-U}^\ddagger$ (kJmol ⁻¹)	resurrected LeuB of Firmicutes LCA
BPSYC	0.2	47	94.9	
BSUB	0.7	53	95.9	
BCVX	1.1	69	100.7	
S-aware Tree + LG	1.5	85	114.4	
S-aware Tree + EX_EHO	1.6	85	110.9	
S-unaware Tree + EX_EHO	6.8	78	91.4	

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:



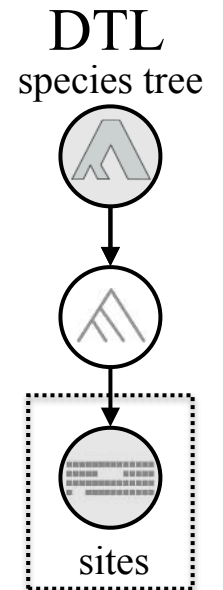
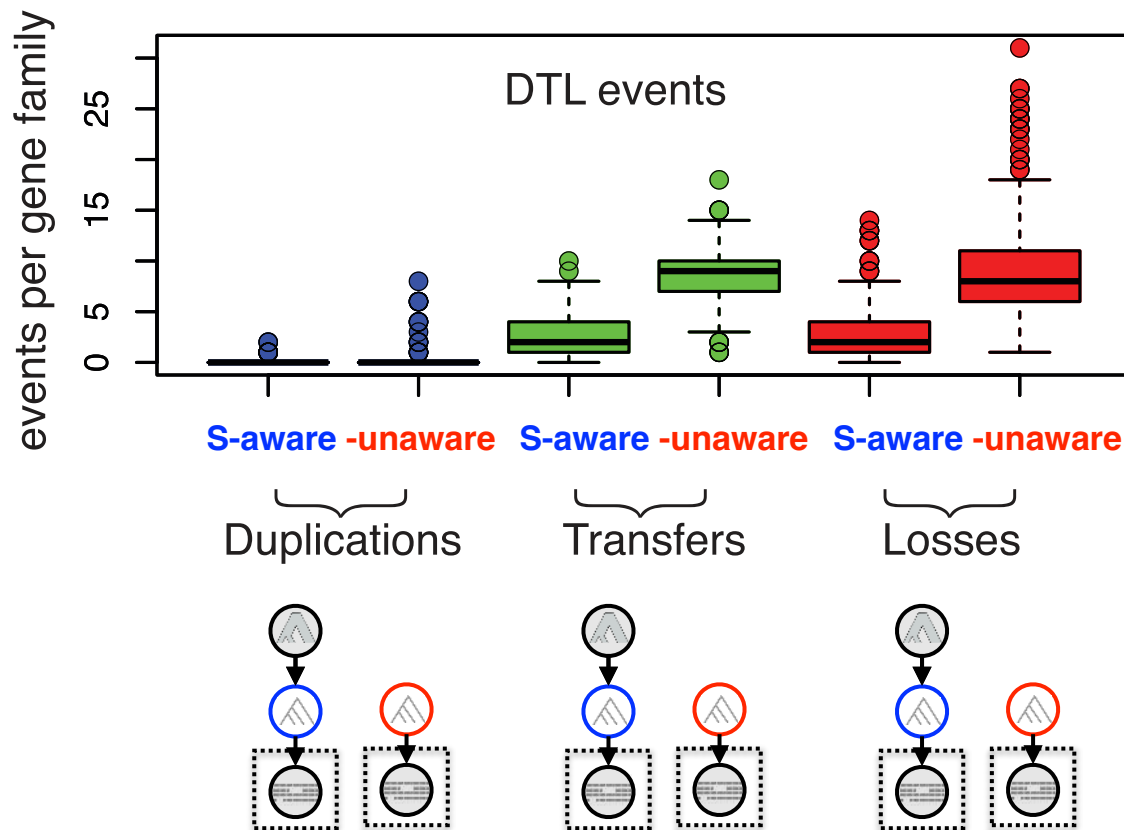
just how much HGT is there?



just how much HGT is there?

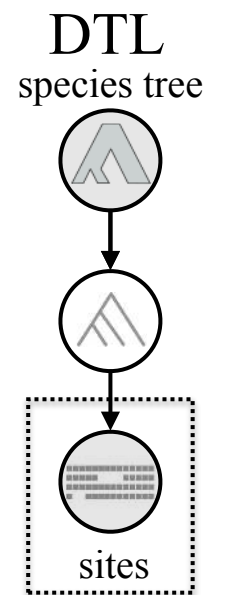
Two out of three transfers inferred based S-unaware gene trees are the result phylogenetic errors.

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



just how much HGT is there?

what about ancestral gene content?

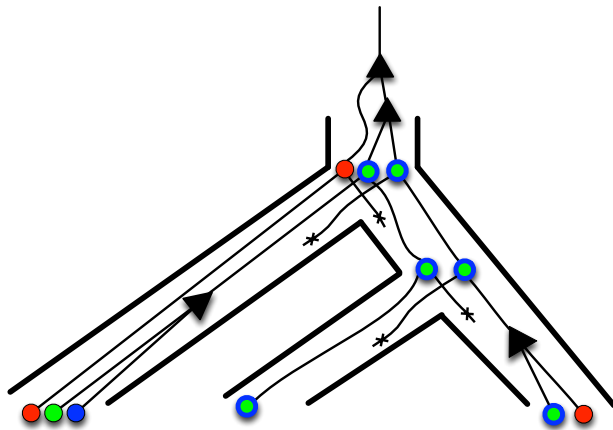


just how much HGT is there?

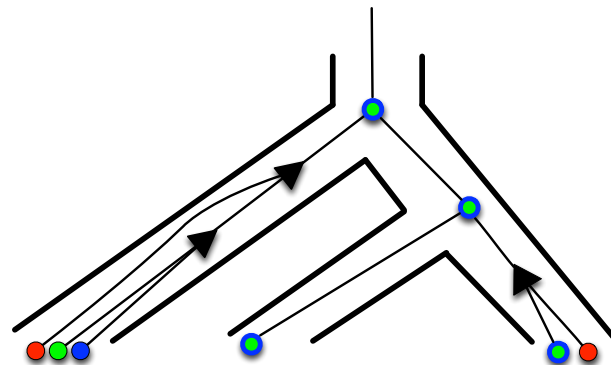
what about ancestral gene content?

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

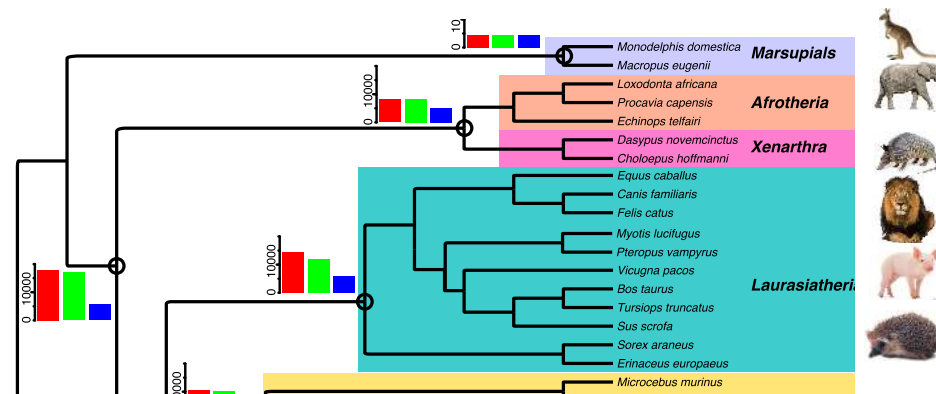
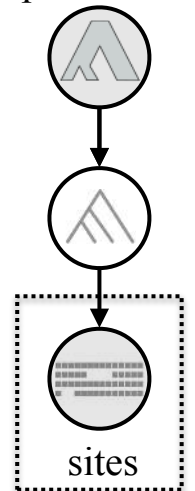
gene tree with errors



correct gene tree



DTL
species tree

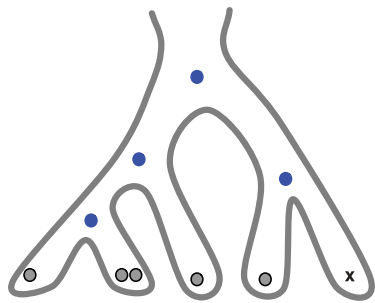


just how much HGT is there?

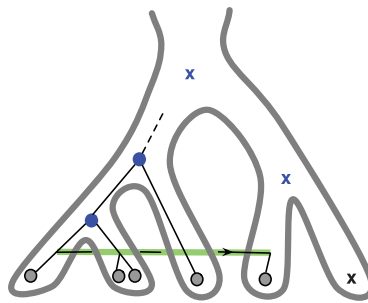
what about ancestral gene content?

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

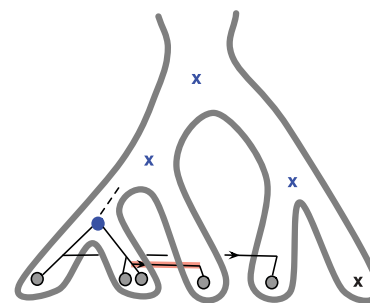
A only phylogenetic profile



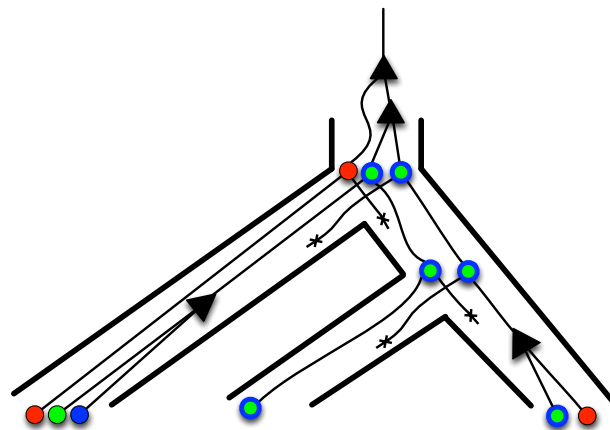
B including gene phylogeny



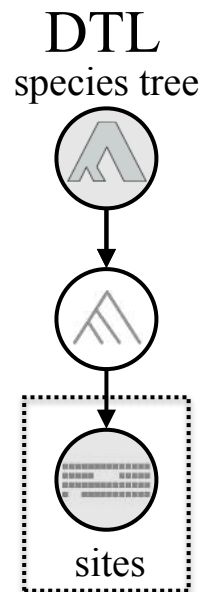
C gene tree errors



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

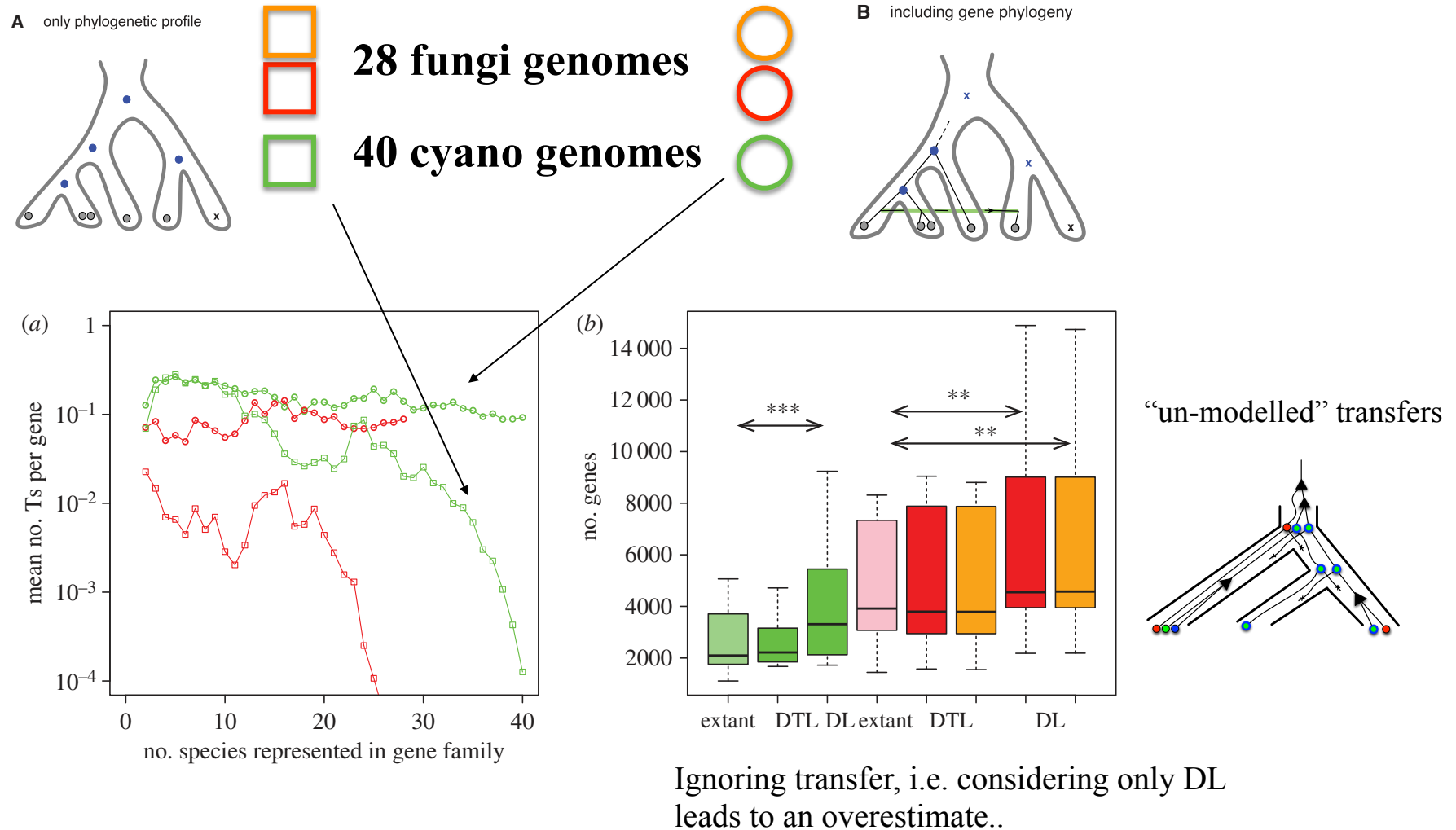


gene tree with “unmodelled” transfers



just how much HGT is there?

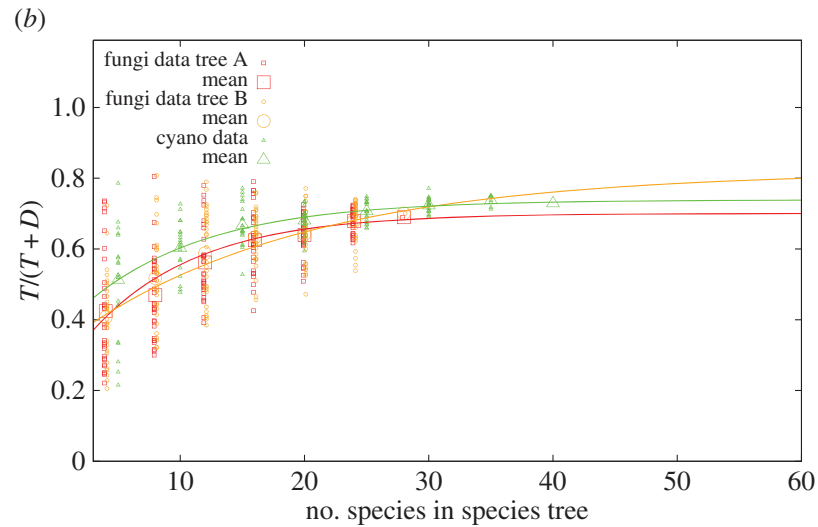
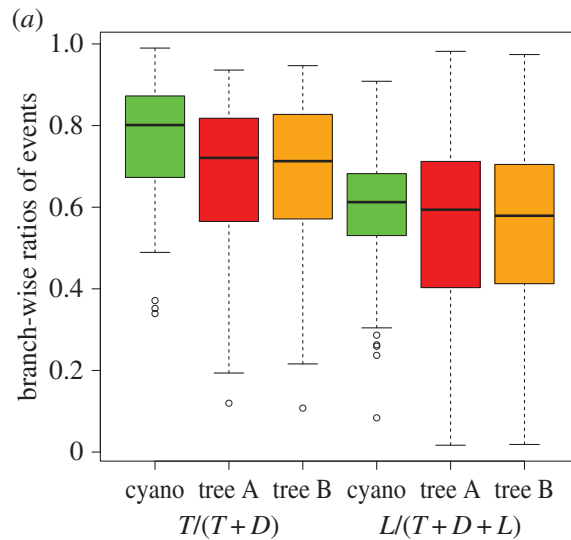
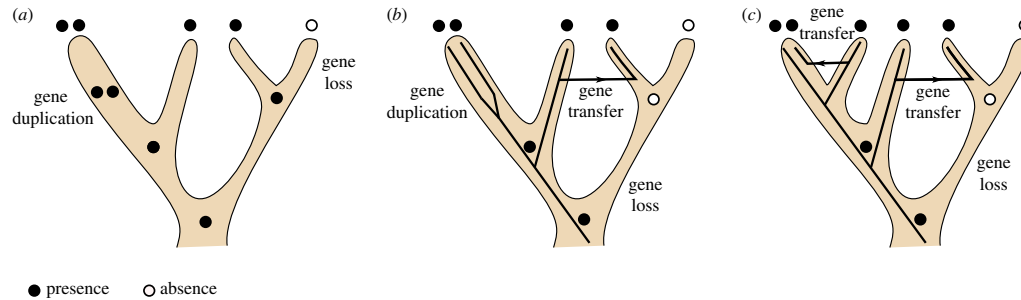
Is HGT frequent only in prokaryotes?



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

just how much HGT is there?

There is extensive gene transfer among fungi

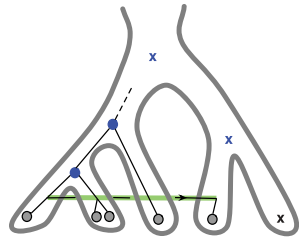


Szöllősi, Davin, Tannier, Daubin & Boussau *Phil.Trans.Roy.Soc.B* (2015)
Genome-scale phylogenetic analysis finds extensive gene transfer among fungi.

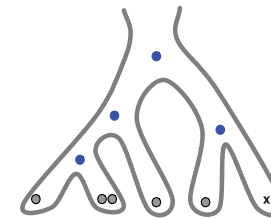
just how much HGT is there?

There is extensive gene transfer among fungi

B including gene phylogeny



A only phylogenetic profile



28 fungi genomes

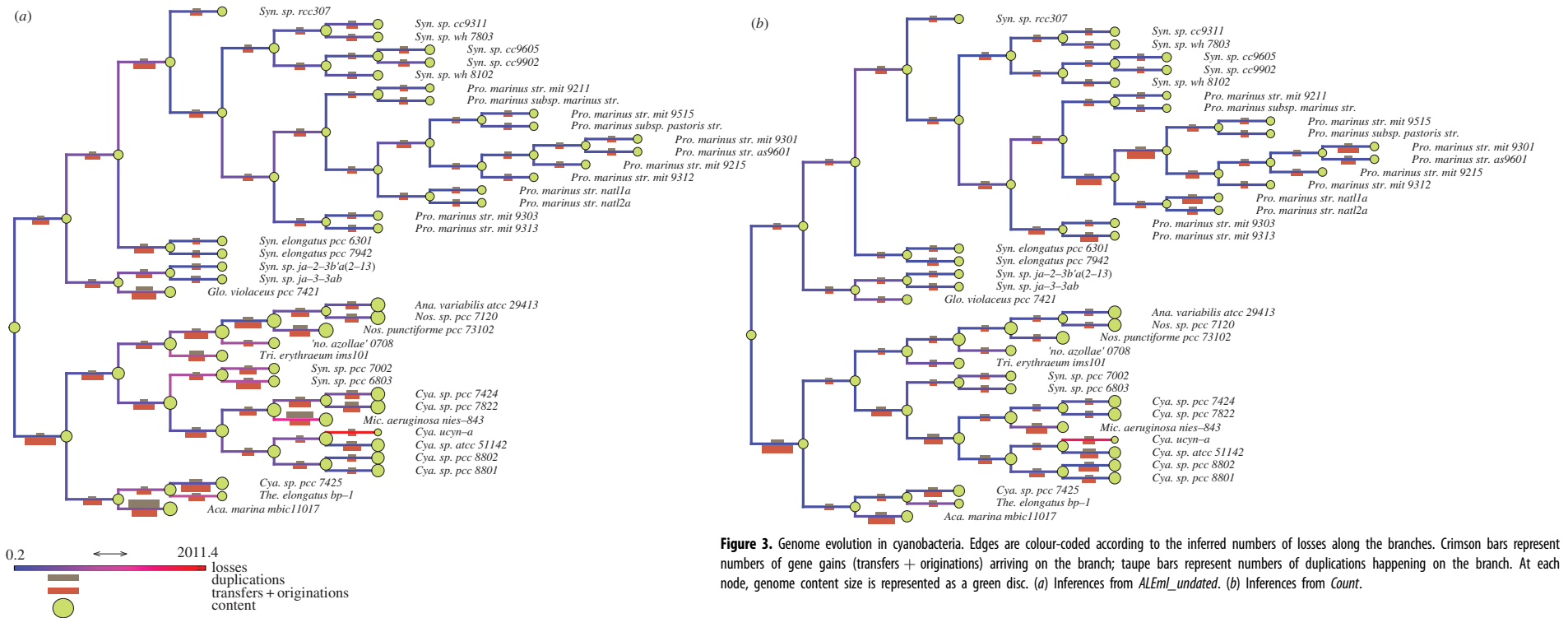


Figure 3. Genome evolution in cyanobacteria. Edges are colour-coded according to the inferred numbers of losses along the branches. Crimson bars represent numbers of gene gains (transfers + originations) arriving on the branch; taupe bars represent numbers of duplications happening on the branch. At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

just how much HGT is there?

Stronger transfer highways in fungi

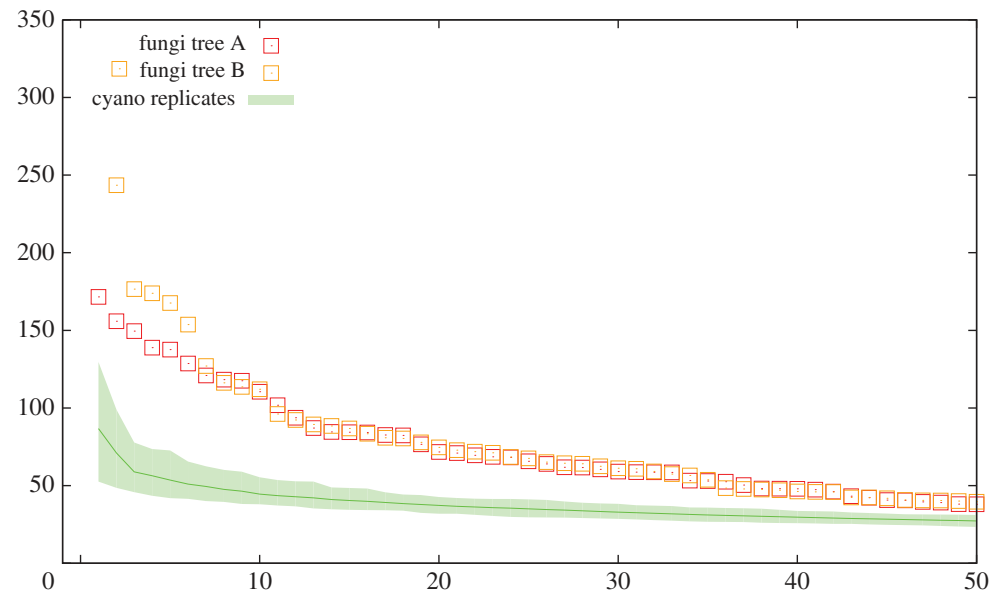
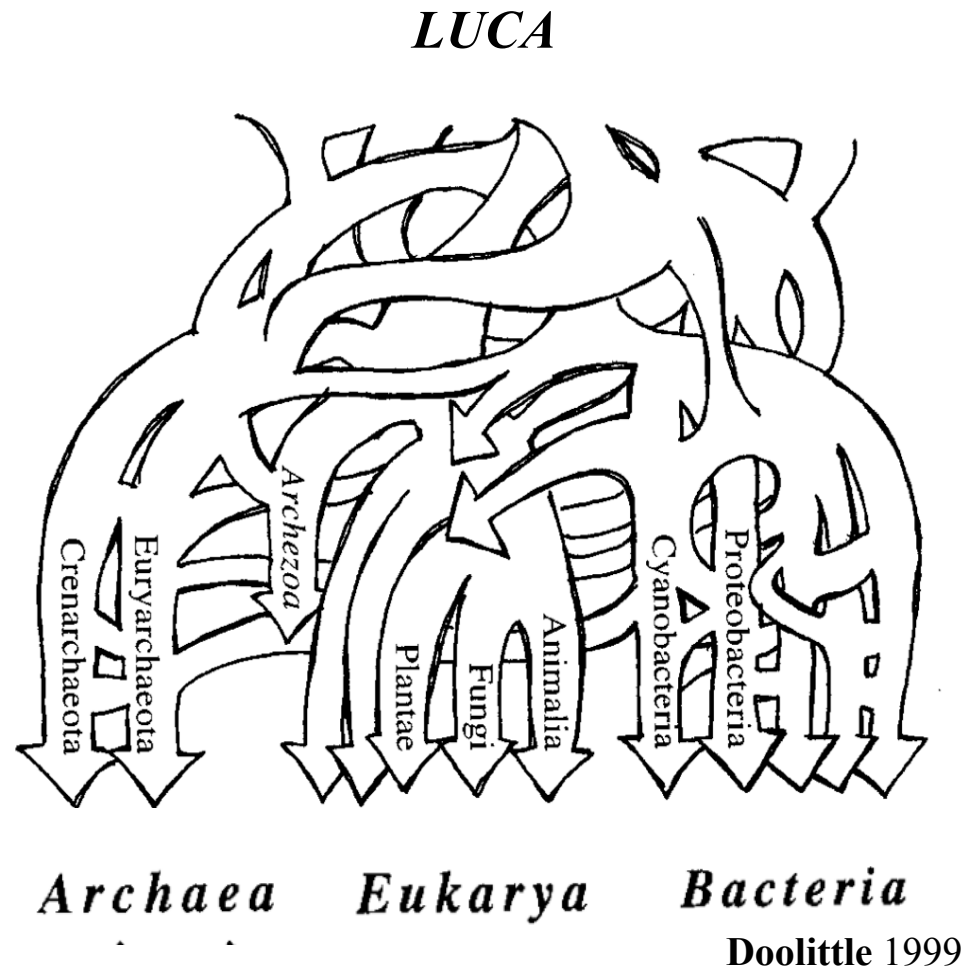


Figure 6. Stronger highways in fungi. Data points, red for tree A and orange for tree B, correspond to numbers of transfers between pairs of branches ('highways of transfers') in either fungi phylogeny plotted in decreasing order. The continuous line (green online) shows the mean number of transfers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

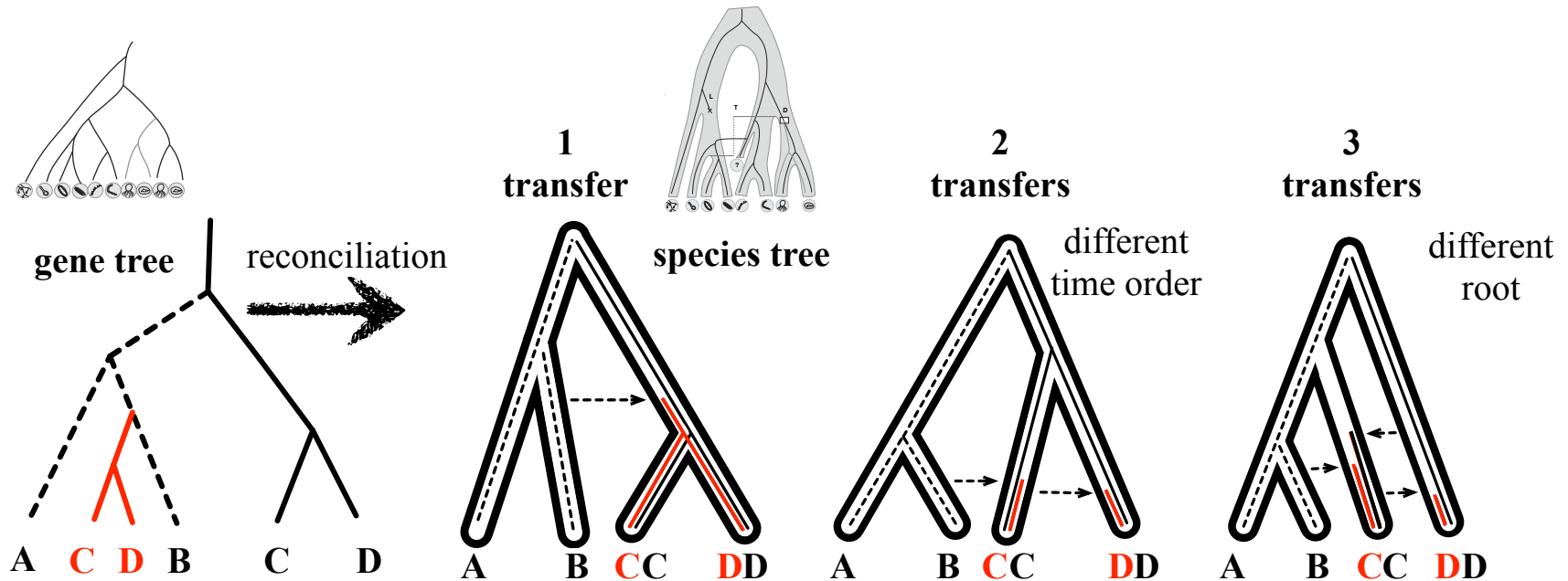
Horizontal gene transfer as noise

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.



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.



Gene trees and species trees can be jointly reconstructed

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

parallel computation scheme

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

server:
calculate

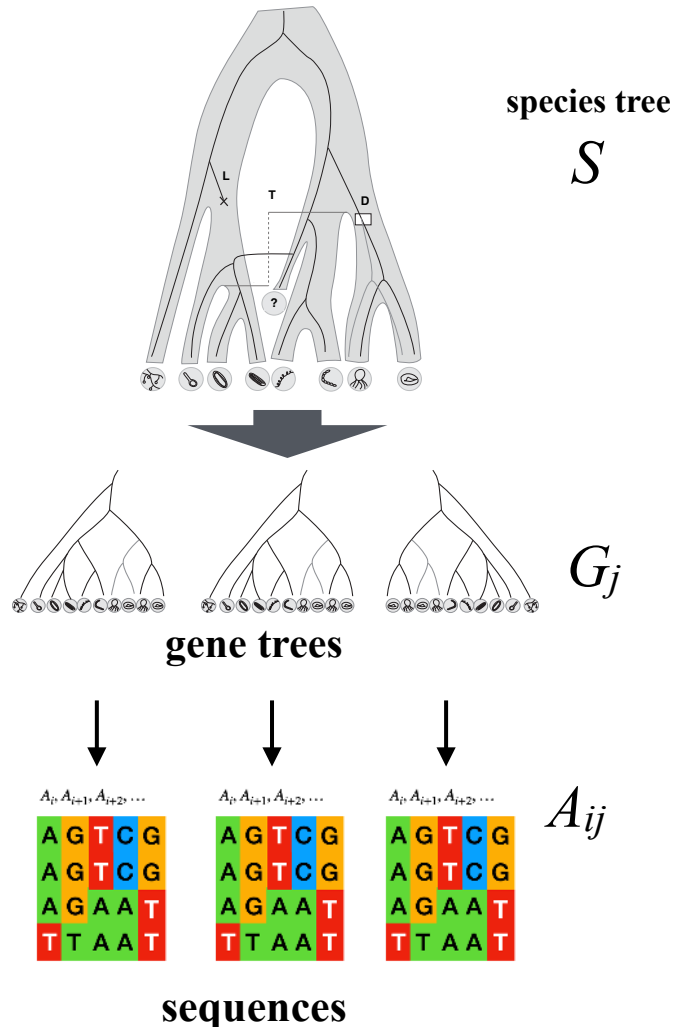
$$\prod_j$$

optimise S
and estimate rates

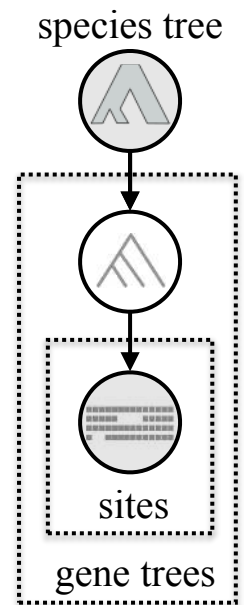
clients:
calculate

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

optimise (or integrate over) G_j



Daubin & Boussau 2011



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

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

Given the **species tree**, which **gene tree** produced my **sequences**? ..
and in **what evolutionary context**?

Joint likelihood: $P(\text{alignment} | \text{gene tree}) P(\text{gene tree} | \text{species tree})$

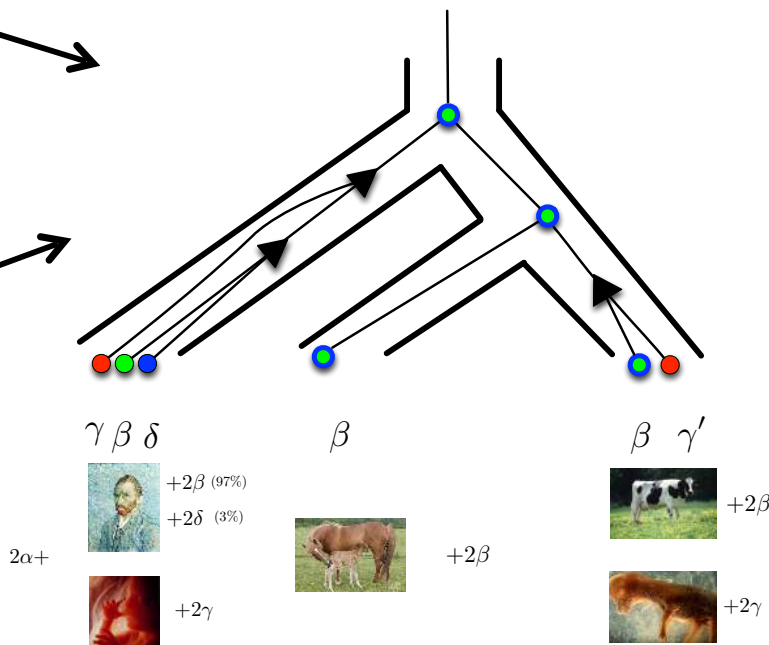


Gene family alignment

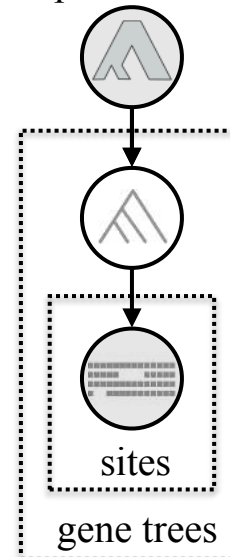


Rooted binary species tree

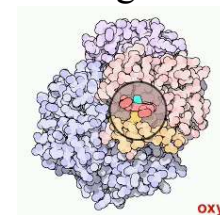
Most likely **species tree-aware** gene tree



species tree



hemoglobin



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

Given the **species tree**, which **gene tree** produced my **sequences**? ..
and in **what evolutionary context**?

Joint likelihood: $P(\text{alignment} | \text{gene tree}) P(\text{gene tree} | \text{species tree})$

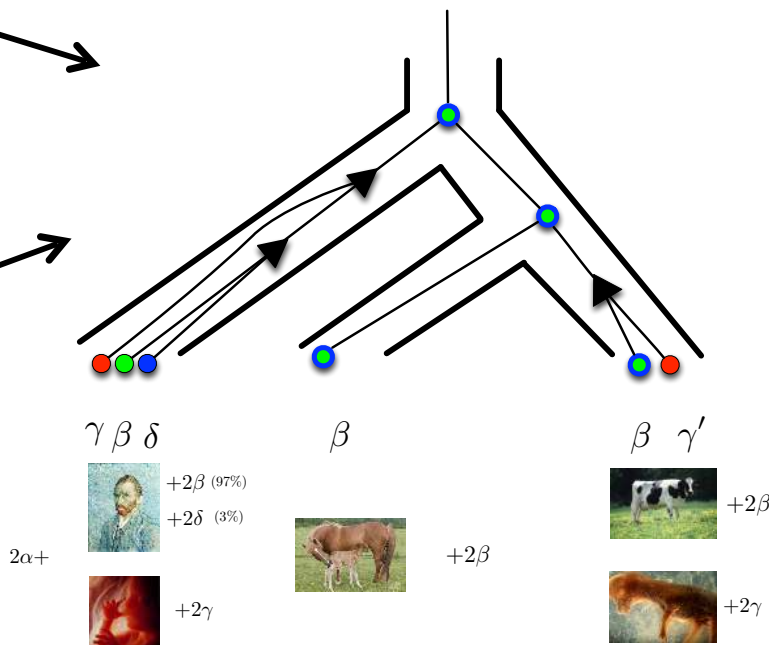


Gene family alignment

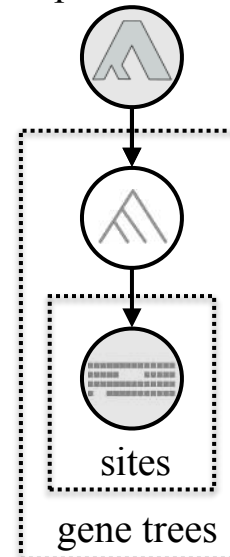


Rooted binary species tree

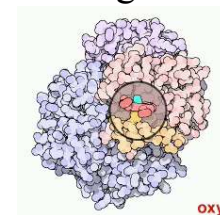
Most likely **species tree-aware** gene tree



species tree



hemoglobin



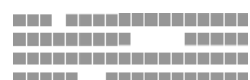
Efficiently exploring the space of reconciled gene trees

Based on a sample of trees conditional clade probabilities can be used to efficiently sum over gene tree uncertainty and sample **species tree-aware** gene trees along with their reconciliations.

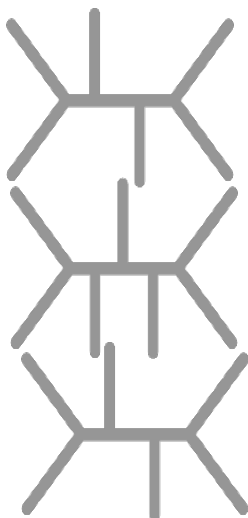
Given the **species tree**, which **gene tree** produced my **sequences**? ..
and in **what evolutionary context**?

Sequence likelihood:

$$P(\text{sequences} | \text{gene tree})$$



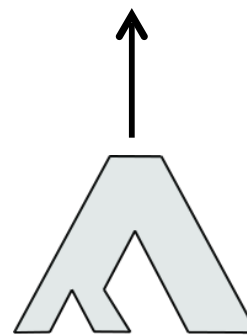
Gene family
alignment



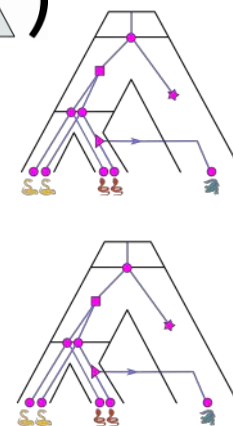
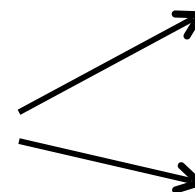
Gene tree
distribution

$$\sum_{\text{gene tree}} P(\text{sequences} | \text{gene tree}) P(\text{gene tree} | \text{species tree})$$

Joint likelihood:



Rooted
species tree



Reconciled
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



<http://github.com/ssolo/ALE>

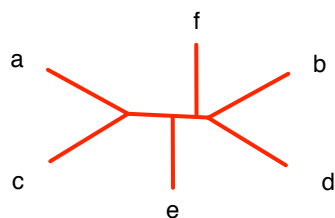
Efficiently exploring the space of reconciled gene trees

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}).

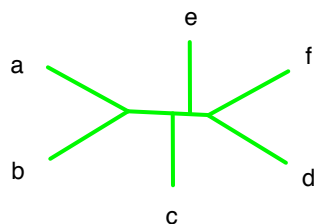
Felsenstein 1981

$$P(\text{[tree]} | \text{[sample]}) \text{ Gene tree sample}$$

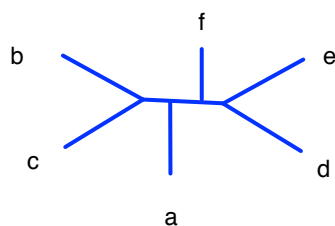
$$\approx 2/10 \quad 2X$$



$$\approx 3/10 \quad 3X$$



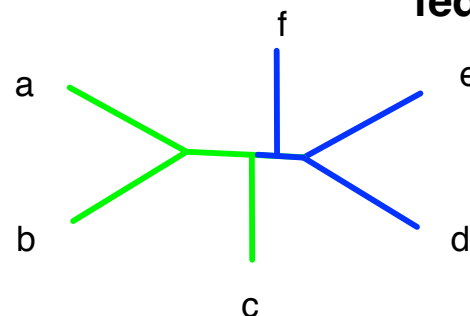
$$\approx 5/10 \quad 5X$$



$$P(\text{[tree]} | \text{[sample]}) \approx 3/8 \times 5/10$$

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

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



more precisely:

$$P(\text{[tree]} | \text{[sample]}) \propto P(\text{[tree]} | \text{[sample]}) P(\text{[tree]})$$

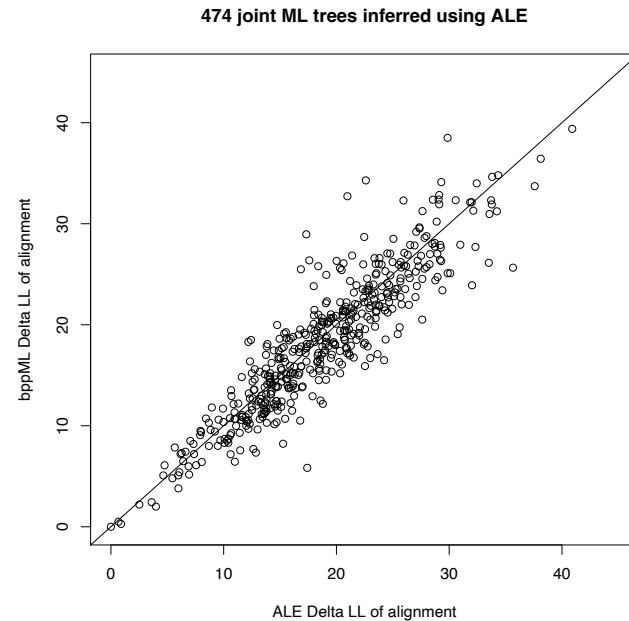
posterior likelihood prior
(uniform)

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}).

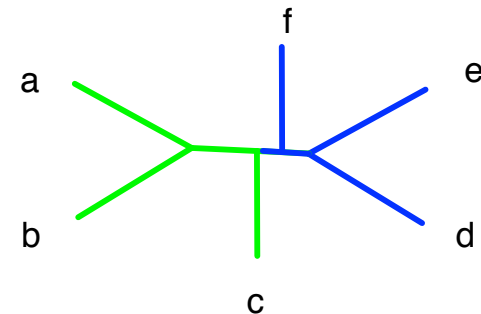
$$P(\text{alignment} \mid \text{tree})$$

explicit ML LL



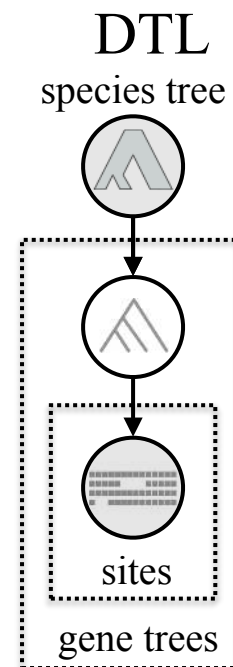
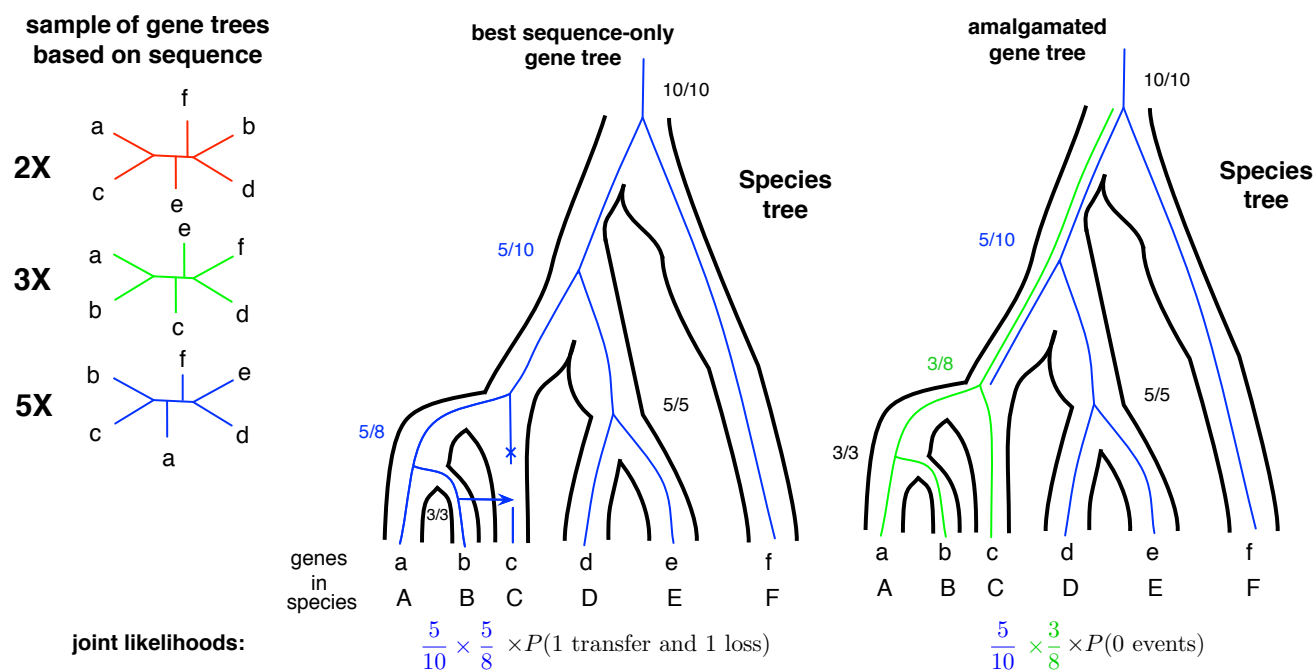
ALE LL

$$\approx 3/8 \times 5/10$$



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}). *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

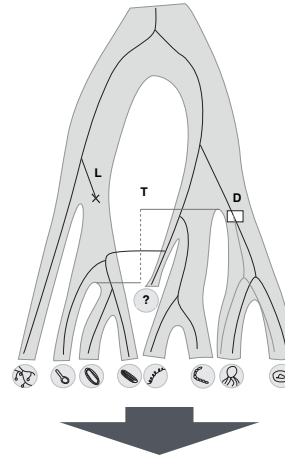
Szöllősi, Rosikiewicz, Boussau, Tannier & Daubin *Systematic Biology* (2013)
Efficient exploration of the space of reconciled gene trees

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**

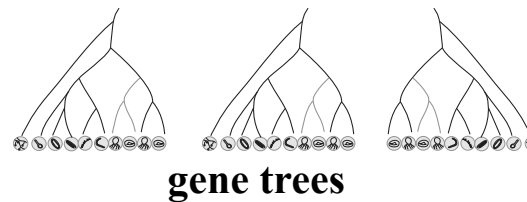


S



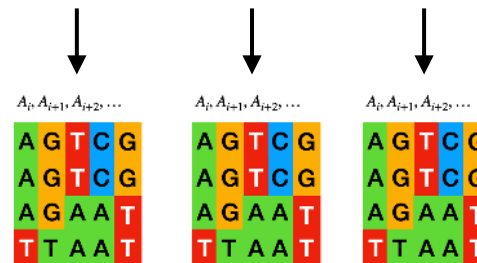
Daubin & Boussau 2011

Fossils & HGT



$$\int_{G_j} G_j$$

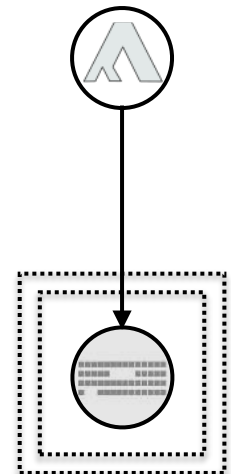
**Alignment of
homologous genes
from complete genomes**



alignments

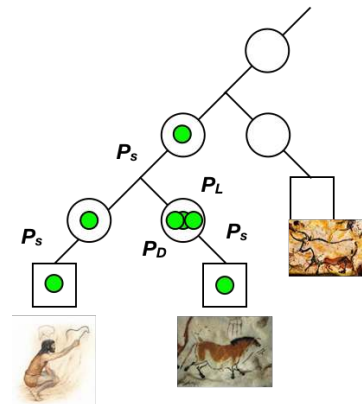
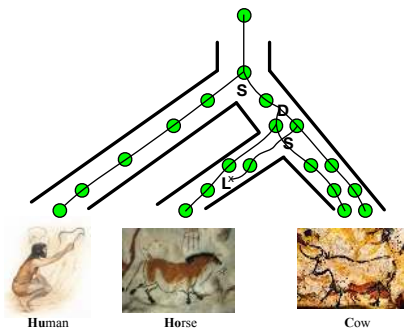
A_{ij}

**DTL
species tree**



.. 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
 $\sim 10 \times \log(\text{\#species}) \times \text{\#genes}$

DTL

$$\sim 10 \times \text{\#species}^2 \times \text{\#genes}$$

DL

$$\log(\# \text{species}) \times \# \text{genes}$$

“undated”

DTL

$$\# \text{species} \times \log(\# \text{species}) \times \# \text{genes}$$

**parameters
(ML or Bayes)**

DL

D&L rates
branch lengths, root

DTL

D,T&L rates
dated tree

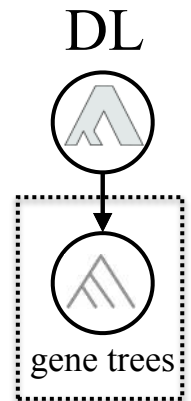
DL

D&L rates, root

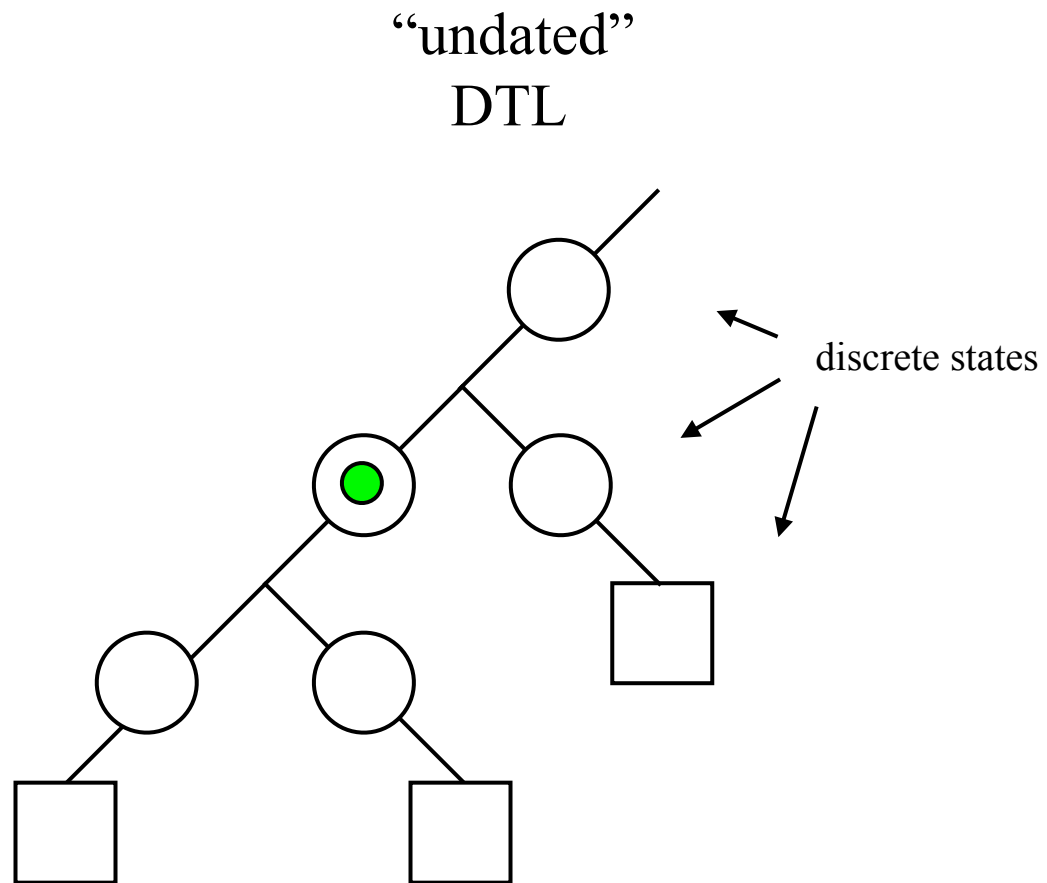
“undated”

DTL

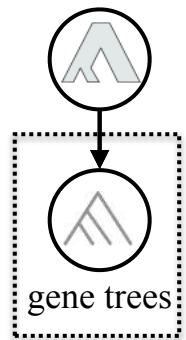
D,T&L rates, root



faster models of gene family evolution with DT&L



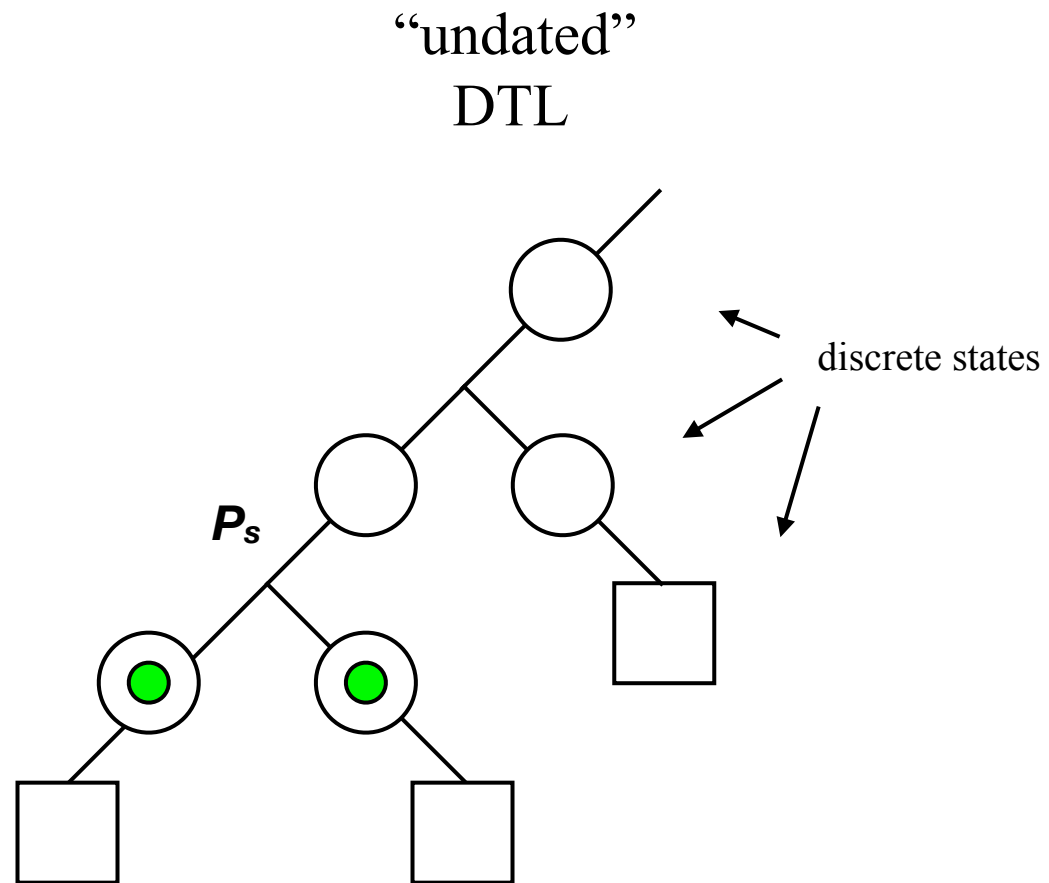
**undated
DTL**



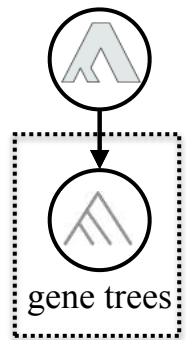
implemented in ALE:

<http://github.com/ssolo/ALE>

faster models of gene family evolution with DT&L



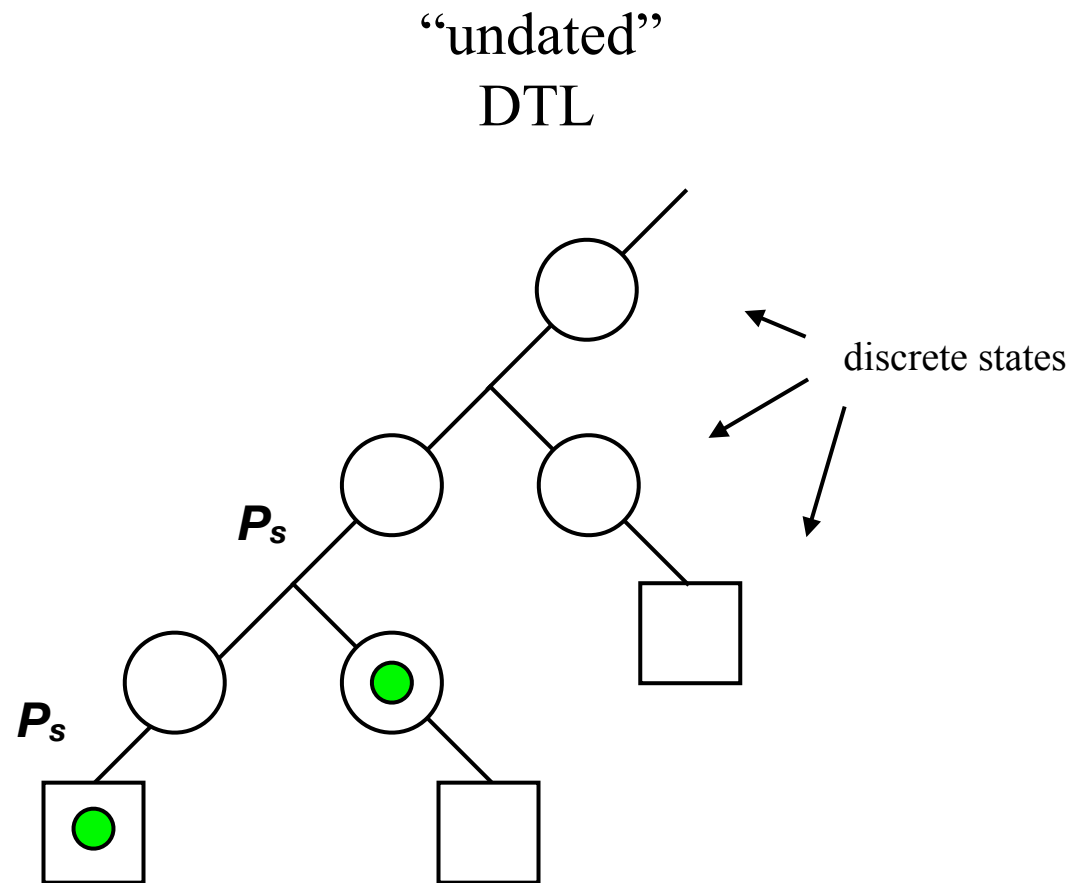
undated
DTL



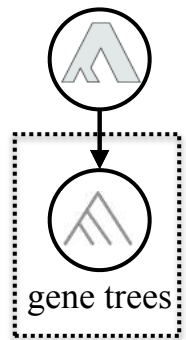
implemented in ALE:

<http://github.com/ssolo/ALE>

faster models of gene family evolution with DT&L



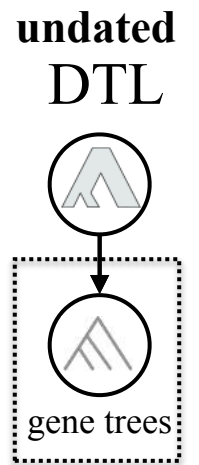
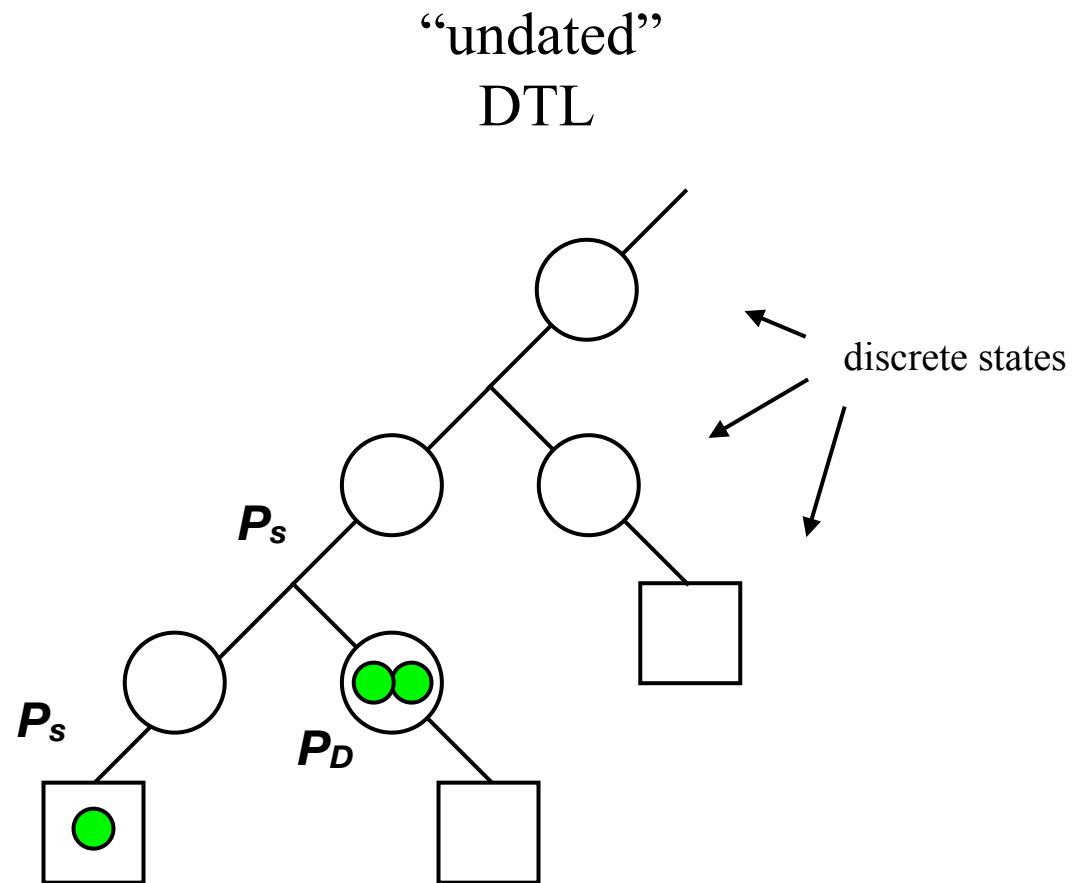
undated
DTL



implemented in ALE:

<http://github.com/ssolo/ALE>

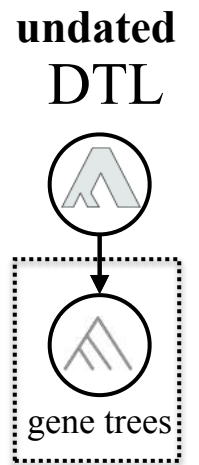
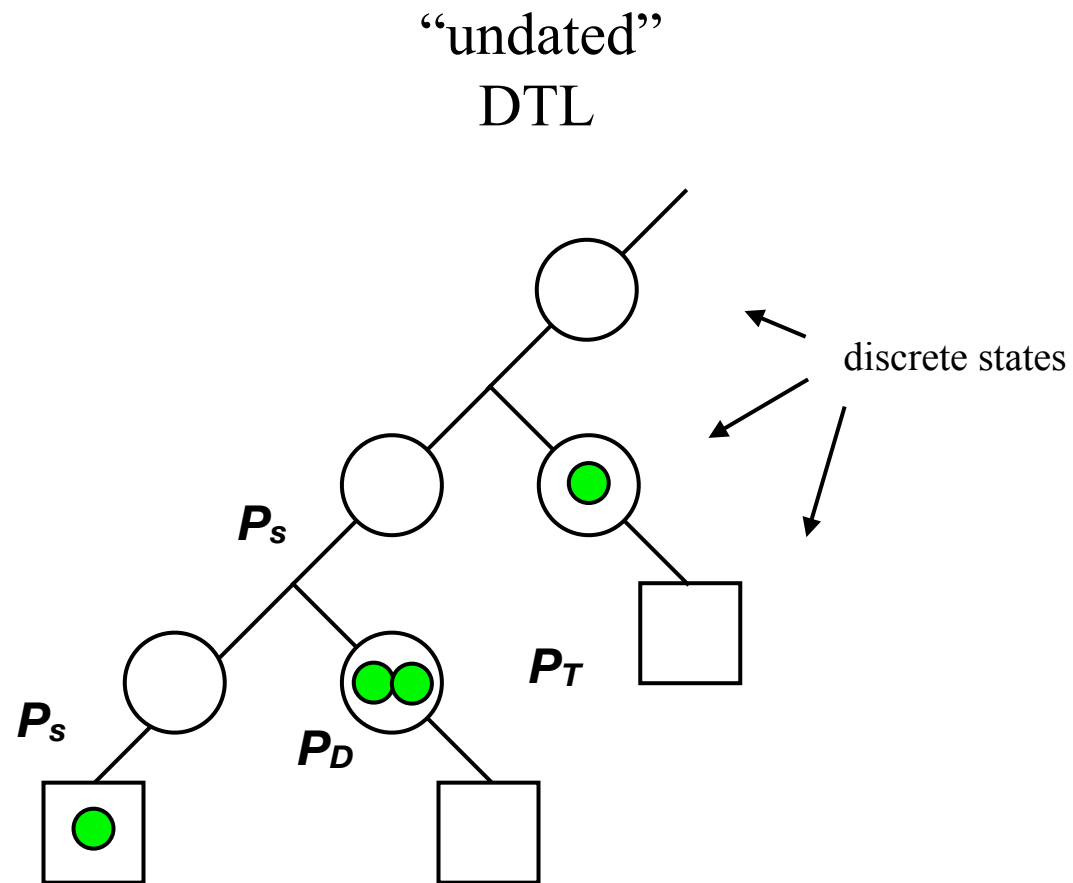
faster models of gene family evolution with DT&L



implemented in ALE:

<http://github.com/ssolo/ALE>

faster models of gene family evolution with DT&L

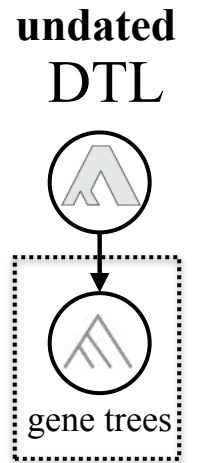
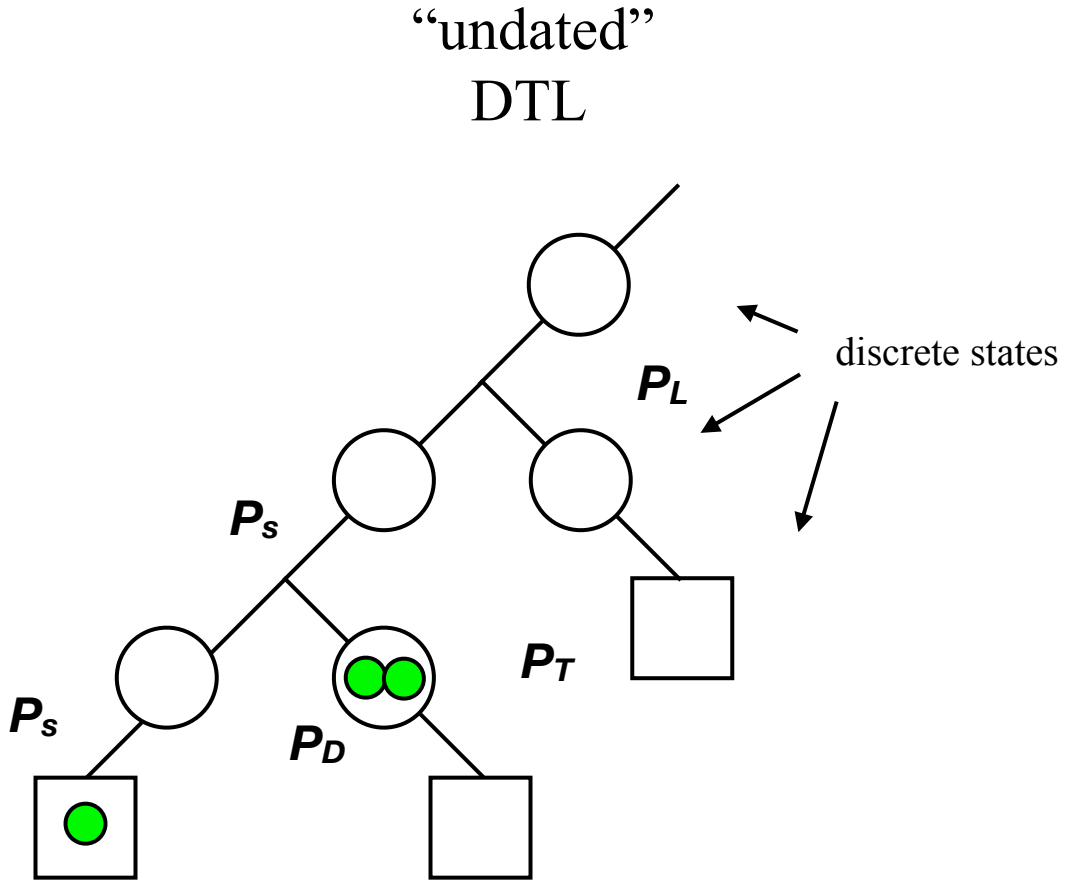


implemented in ALE:

<http://github.com/ssolo/ALE>

faster models of gene family evolution with DT&L

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

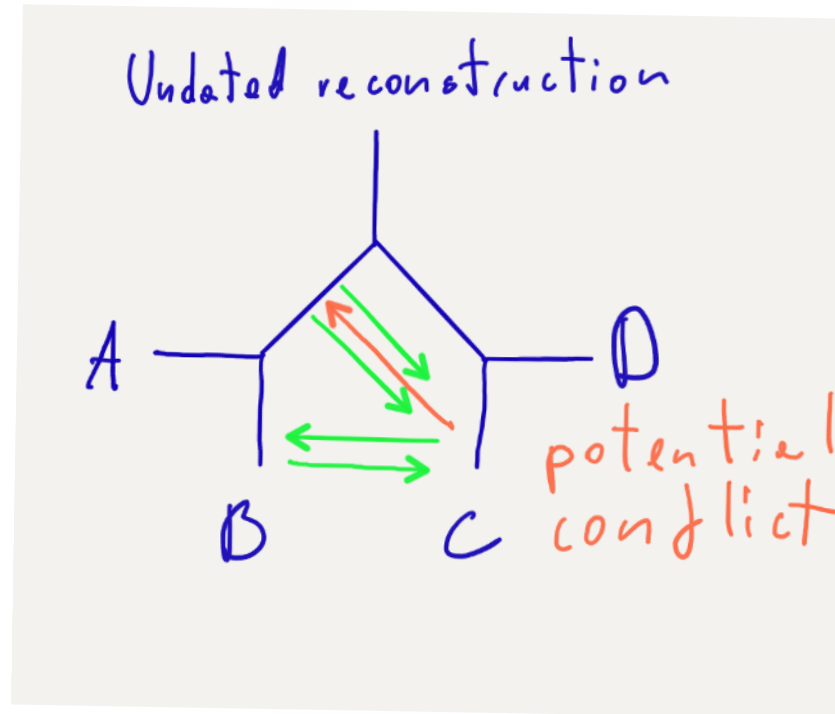


implemented in ALE:

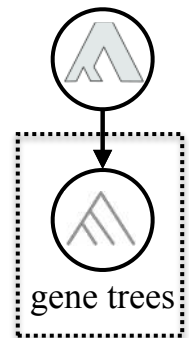
<http://github.com/ssolo/ALE>

faster models of gene family evolution with DT&L

“undated”
DTL



undated
DTL



DTL



“undated” DTL

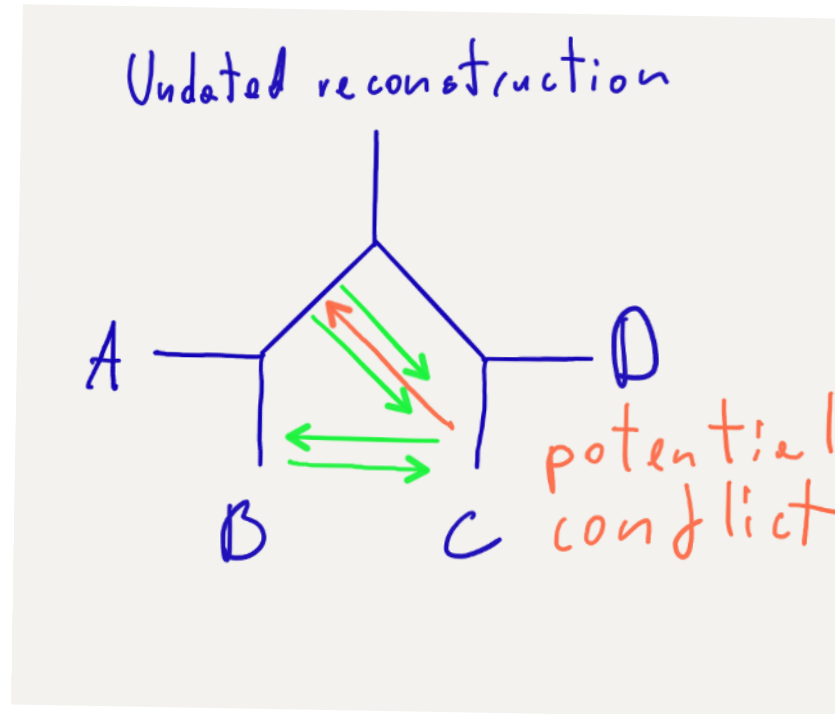


implemented in ALE:

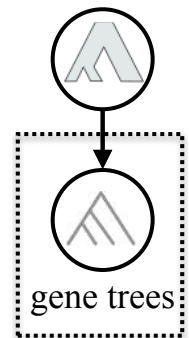
<http://github.com/ssolo/ALE>

faster models of gene family evolution with DT&L

“undated”
DTL



undated
DTL



DTL



“undated” DTL



implemented in ALE:

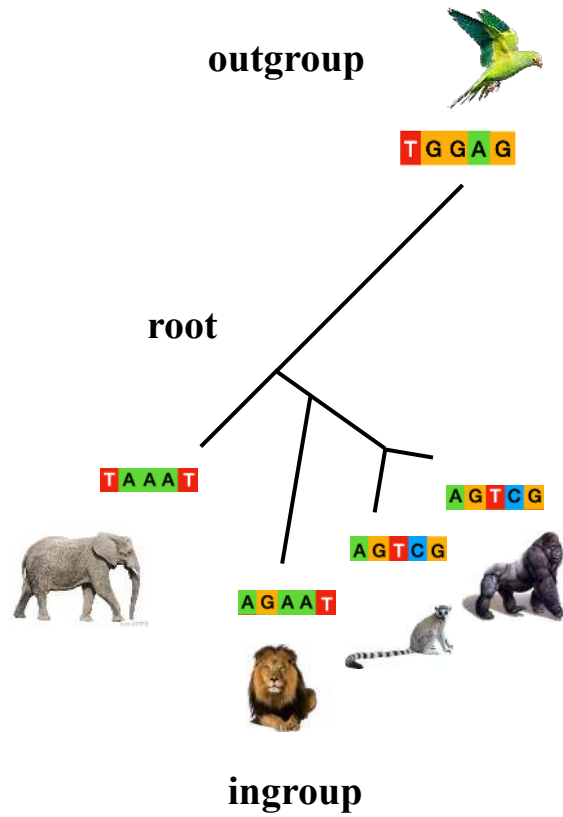
<http://github.com/ssolo/ALE>



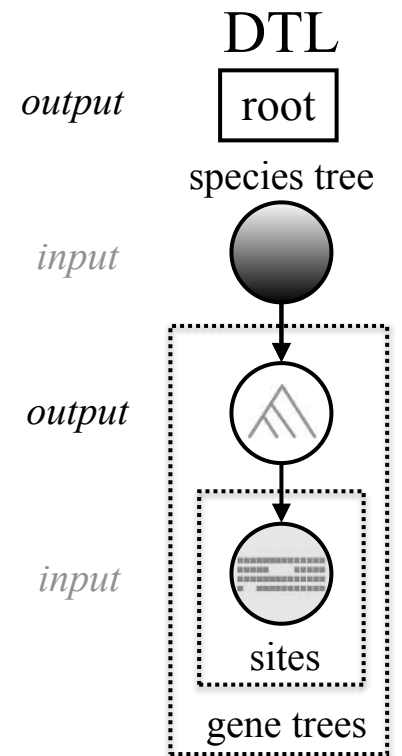
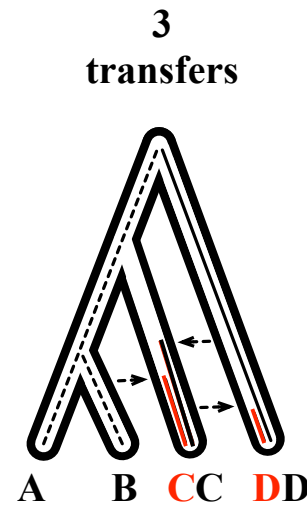
GENECLOCKS

RECONSTRUCTING A DATED TREE OF LIFE USING PHYLOGENETIC INCONGRUENCE

OUTGROUP ROOTING



OUTGROUP-FREE ROOTING



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 **using the history of 1% as a scaffold**.

62 genomes / **31,236** gene families

Integrative modeling of gene and genome evolution roots the archaeal tree of life

Tom A. Williams^{a,b,1}, Gergely J. Szöllősi^{c,2}, Anja Spang^{d,2}, Peter G. Foster^e, Sarah E. Heaps^{b,f}, Bastien Boussau^g, Thijs J. G. Ettema^d, and T. Martin Embley^b

^aSchool of Earth Sciences, University of Bristol, Bristol BS8 1TQ, United Kingdom; ^bInstitute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne NE2 4HH, United Kingdom; ^cMTA-ELTE Lendület Evolutionary Genomics Research Group, 1117 Budapest, Hungary; ^dDepartment of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, SE-75123 Uppsala, Sweden; ^eDepartment of Life Sciences, Natural History Museum, London SW7 5BD, United Kingdom; ^fSchool of Mathematics & Statistics, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom; and ^gUniv Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR5558, F-69622 Villeurbanne, France

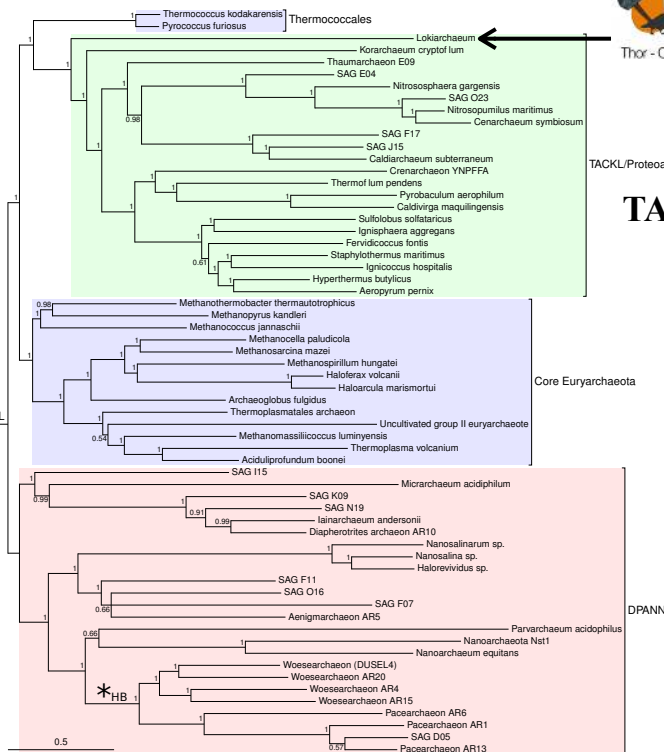
Edited by W. Ford Doolittle, Dalhousie University, Halifax, Canada, and approved April 24, 2017 (received for review November 7, 2016)



U.Bristol



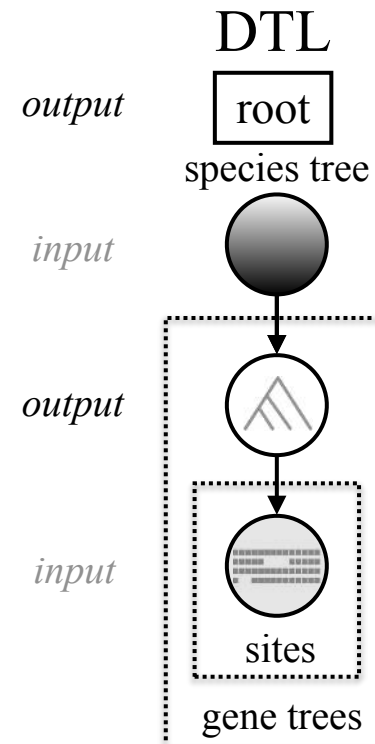
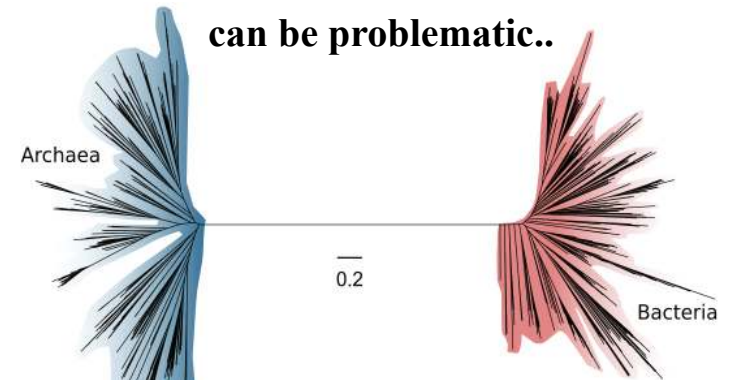
SZÖLLŐSI



TACKL

core
Euryarchaeota

DPANN



outgroup-free rooting

Resolving LACA

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 using the history of 1% as a scaffold.

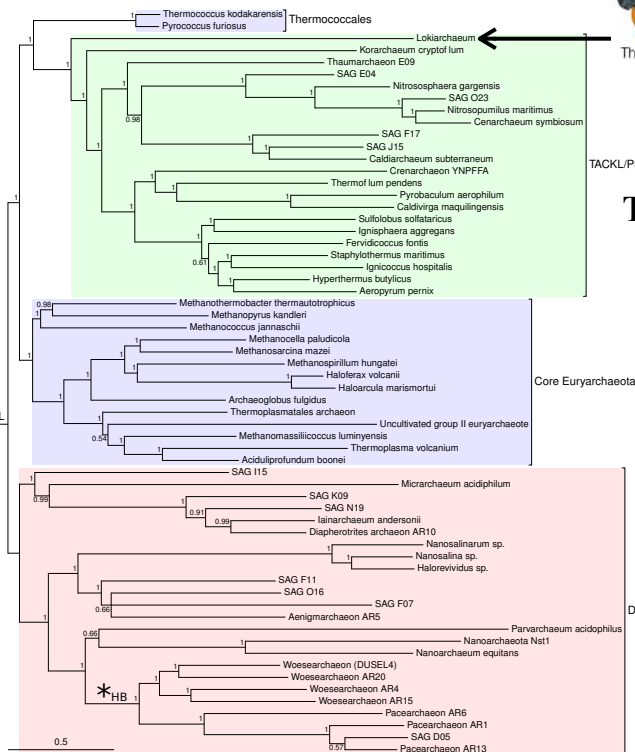
62 genomes / 31,236 gene families

Integrative modeling of gene and genome evolution roots the archaeal tree of life

Tom A. Williams^{a,b,1}, Gergely J. Szöllősi^{c,2}, Anja Spang^{d,2}, Peter G. Foster^e, Sarah E. Heaps^{b,f}, Bastien Boussau^g, Thijs J. G. Ettema^d, and T. Martin Embley^b

^aSchool of Earth Sciences, University of Bristol, Bristol BS8 1TQ, United Kingdom; ^bInstitute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne NE2 4HH, United Kingdom; ^cMTA-ELTE Lendület Evolutionary Genomics Research Group, 1117 Budapest, Hungary; ^dDepartment of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, SE-75123 Uppsala, Sweden; ^eDepartment of Life Sciences, Natural History Museum, London SW7 5BD, United Kingdom; ^fSchool of Mathematics & Statistics, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom; and ^gUniv Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR5558, F-69622 Villeurbanne, France

Edited by W. Ford Doolittle, Dalhousie University, Halifax, Canada, and approved April 24, 2017 (received for review November 7, 2016)



Thor - Odin - Loki - Heimdall

TACKL

core Euryarchaeota

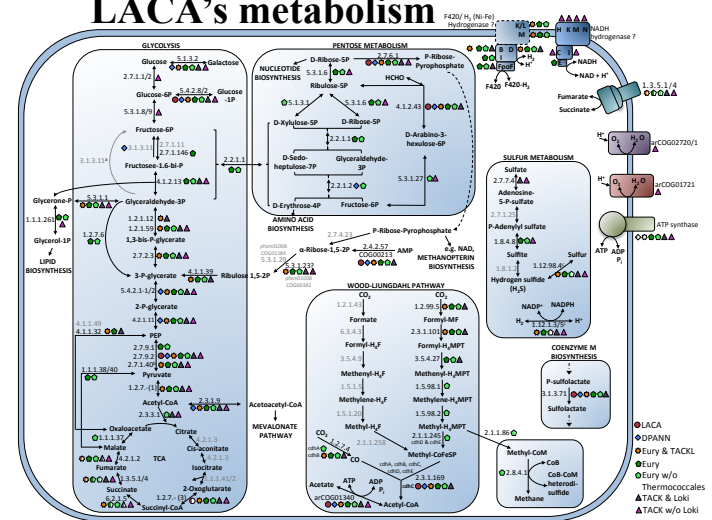
DPANN

NIOZ



SZÖLLŐSI

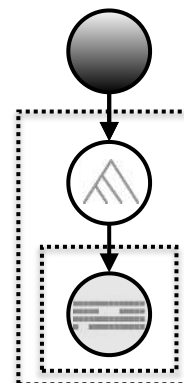
LACA's metabolism



An anaerobe that could fix CO₂ to acetyl-CoA and generate acetate and ATP from it.

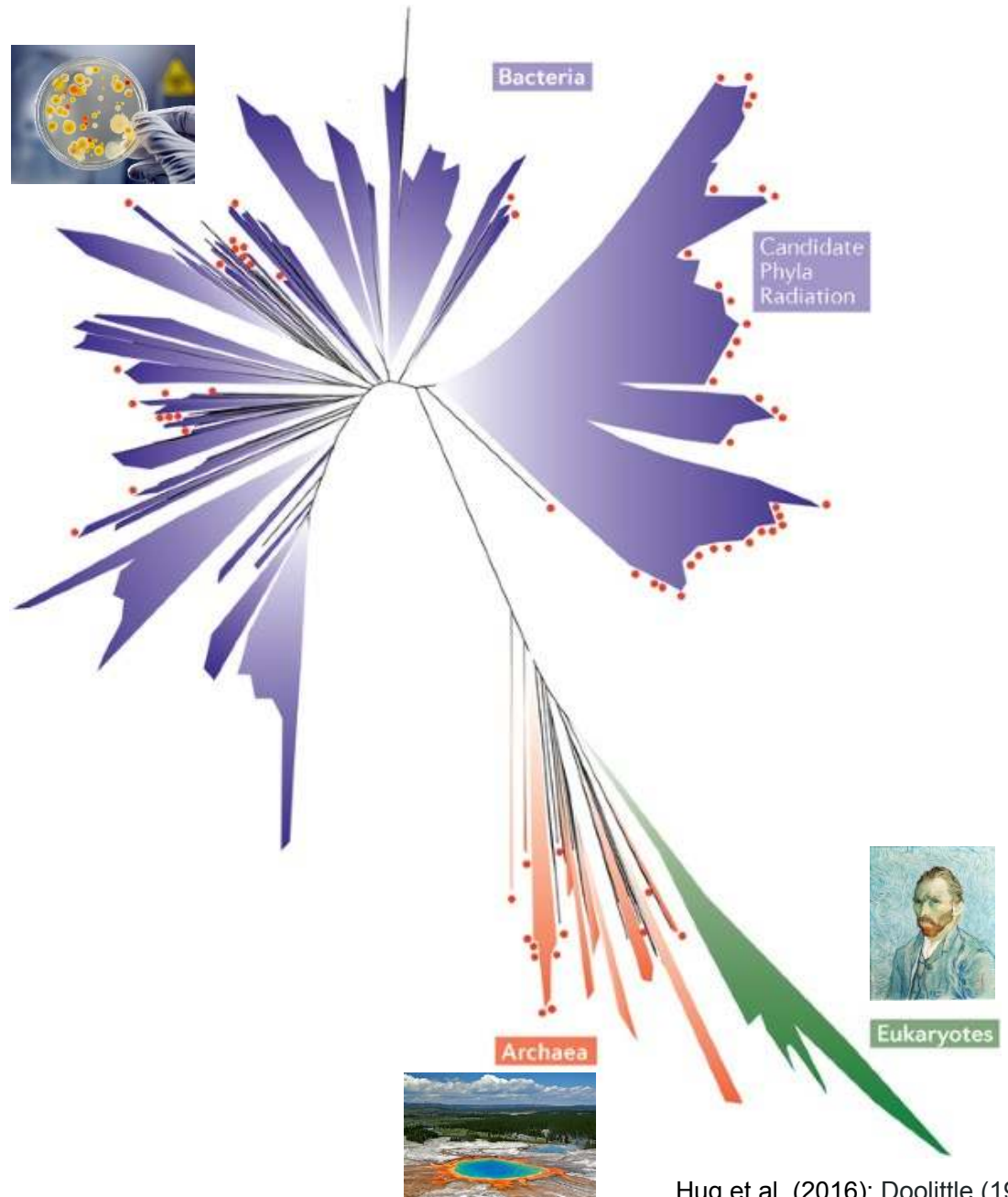
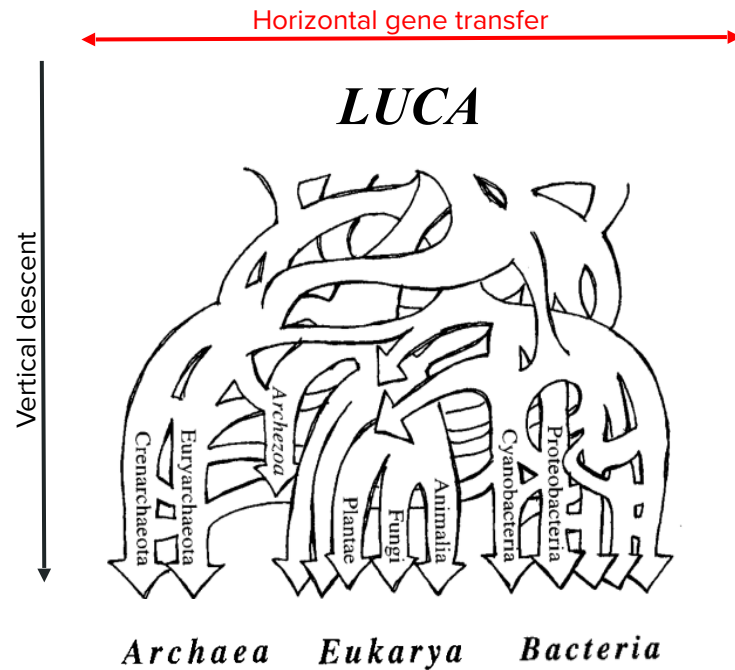
DTL

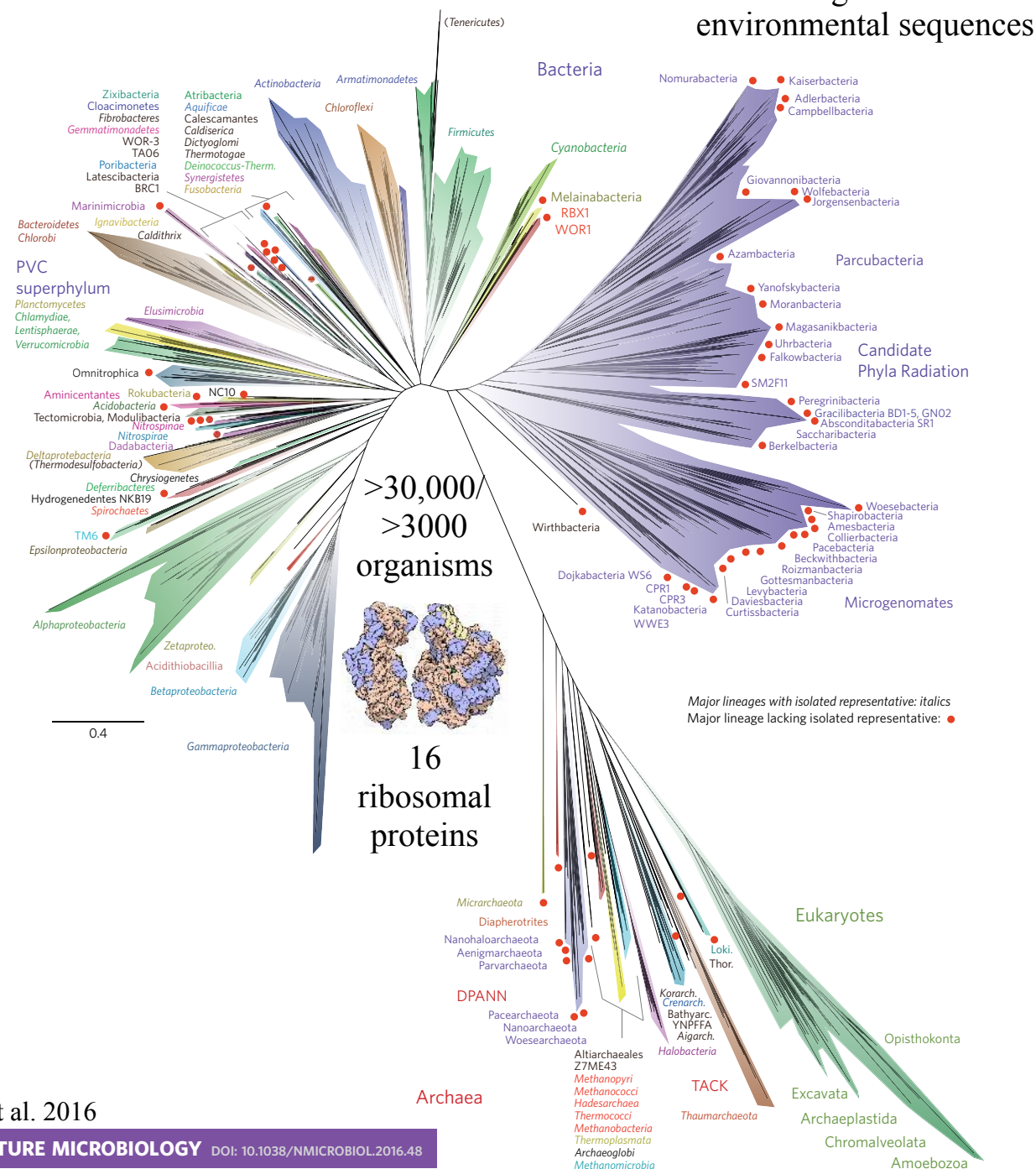
ancestral gene contents



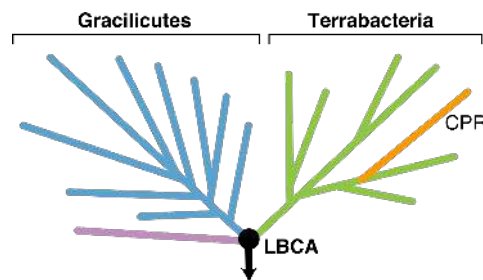
Early bacterial evolution is a playground full of open fundamental questions

- Among the most genetically and metabolically diverse group of cellular lifeforms
 - Shaped evolution of the planet
 - Enormous “new” diversity
 - Interesting open questions about their early evolution
1. Is there a tree? Where's the root?
 2. What was the last bacterial common ancestor like?
 3. How did the rise of oxygen (re)shape bacterial evolution?

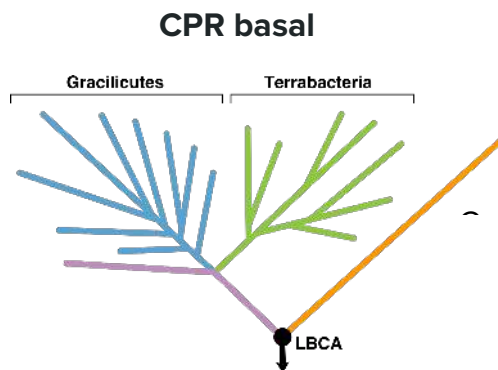




Hug et al. 2016

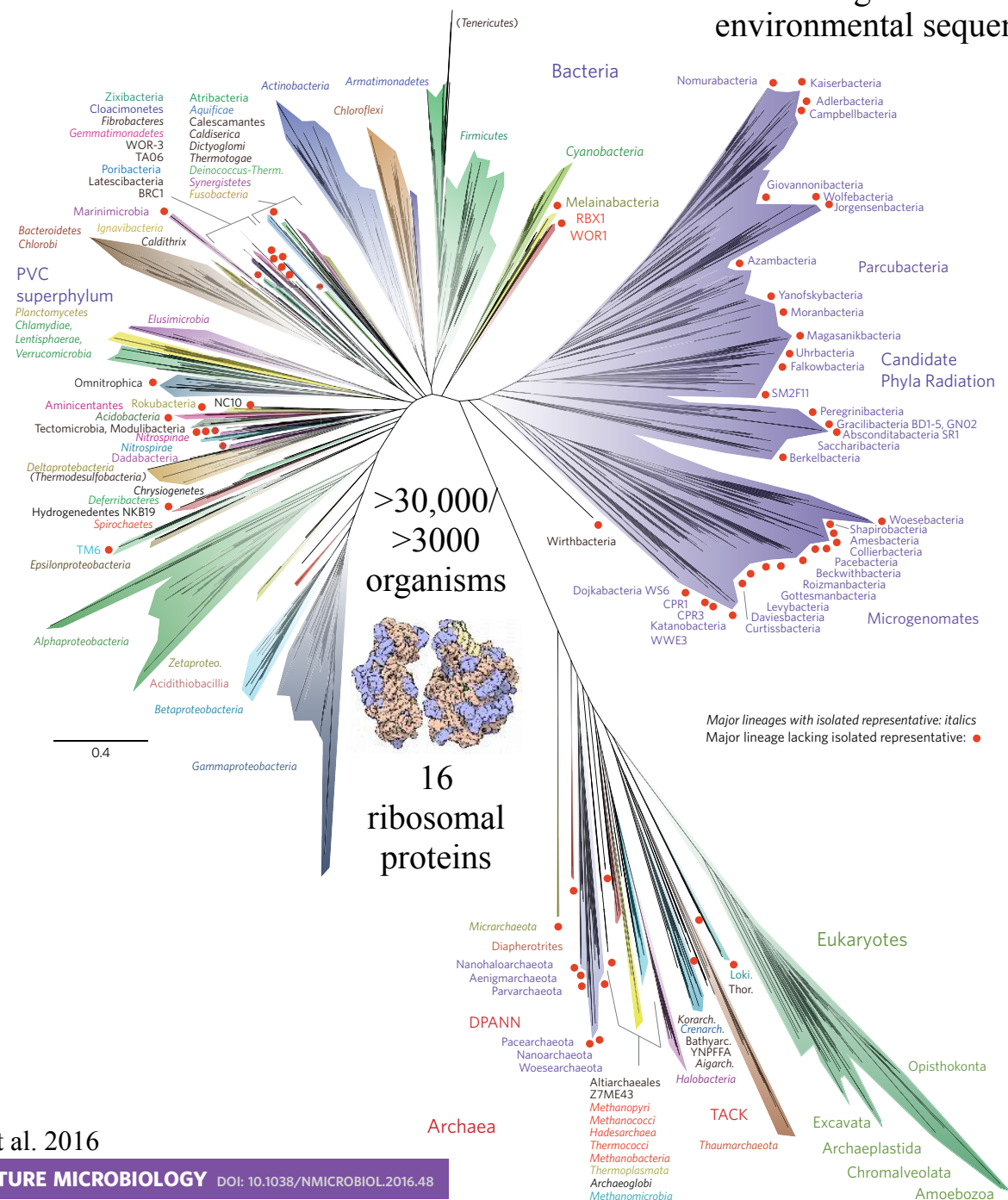


- More complex, free-living ancestor
- Diderm (double-membraned) ancestor?



- Small-genomed, host-associated ancestor?
- Monoderm (single-membraned ancestor?)

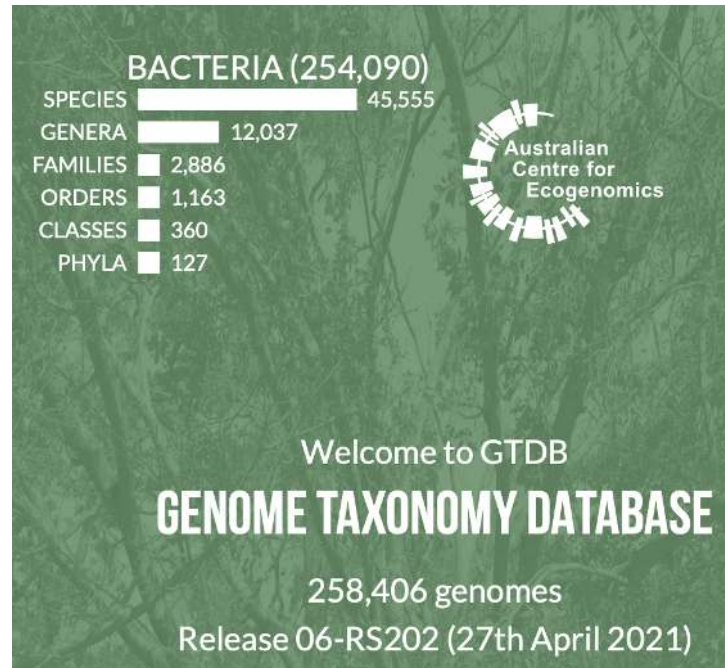
>1000 new genomes from
environmental sequences



Hug et al. 2016

Fewer genomes, but better models, more signal?

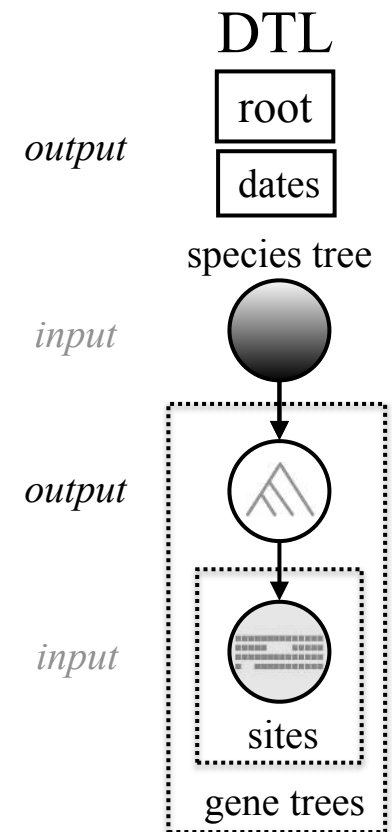
What we need is a good sample..



..to access alternative rooting signal.

2x

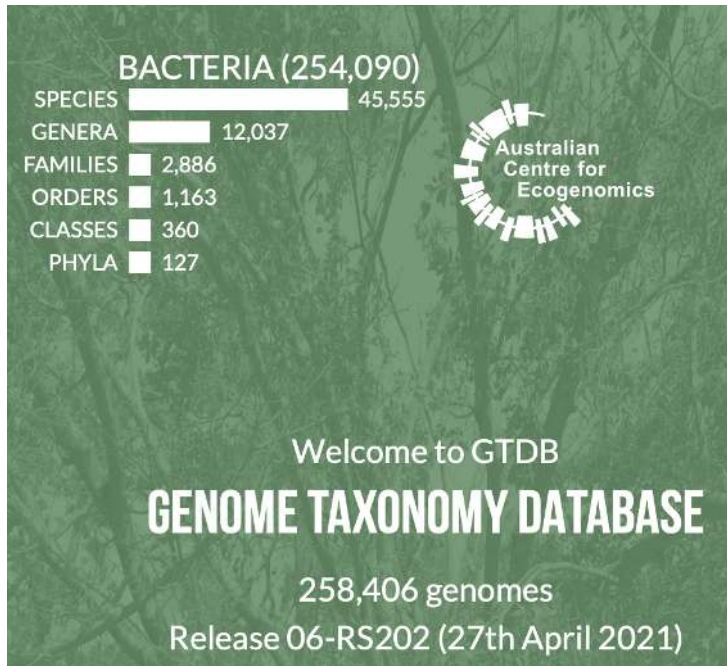
265 genomes / 11,272 gene families
62 marker genes



outgroup-free rooting

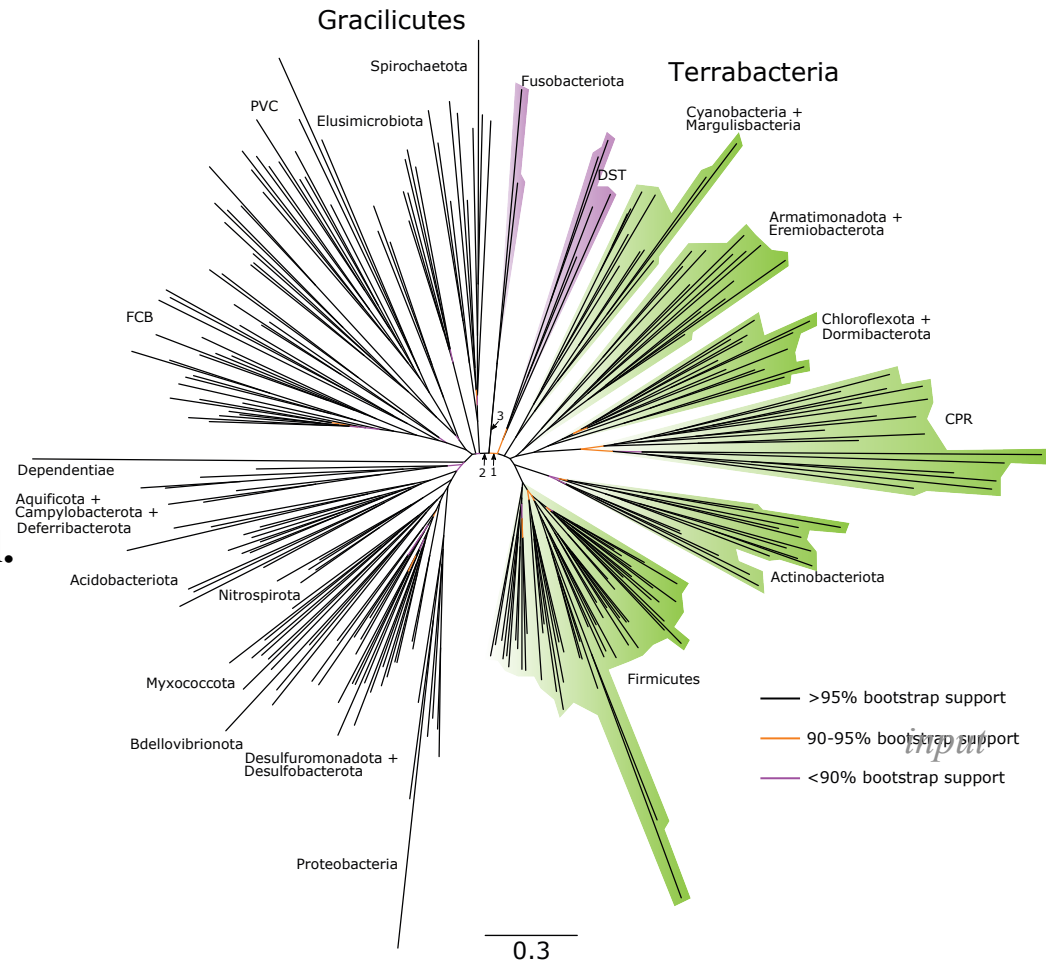
Fewer genomes, but better models, more signal?

What we need is a good sample..



..to access alternative rooting signal.

265 genomes / 11,272 gene families
62 marker genes
using the history of 1% as a scaffold



SZÖLLŐSI

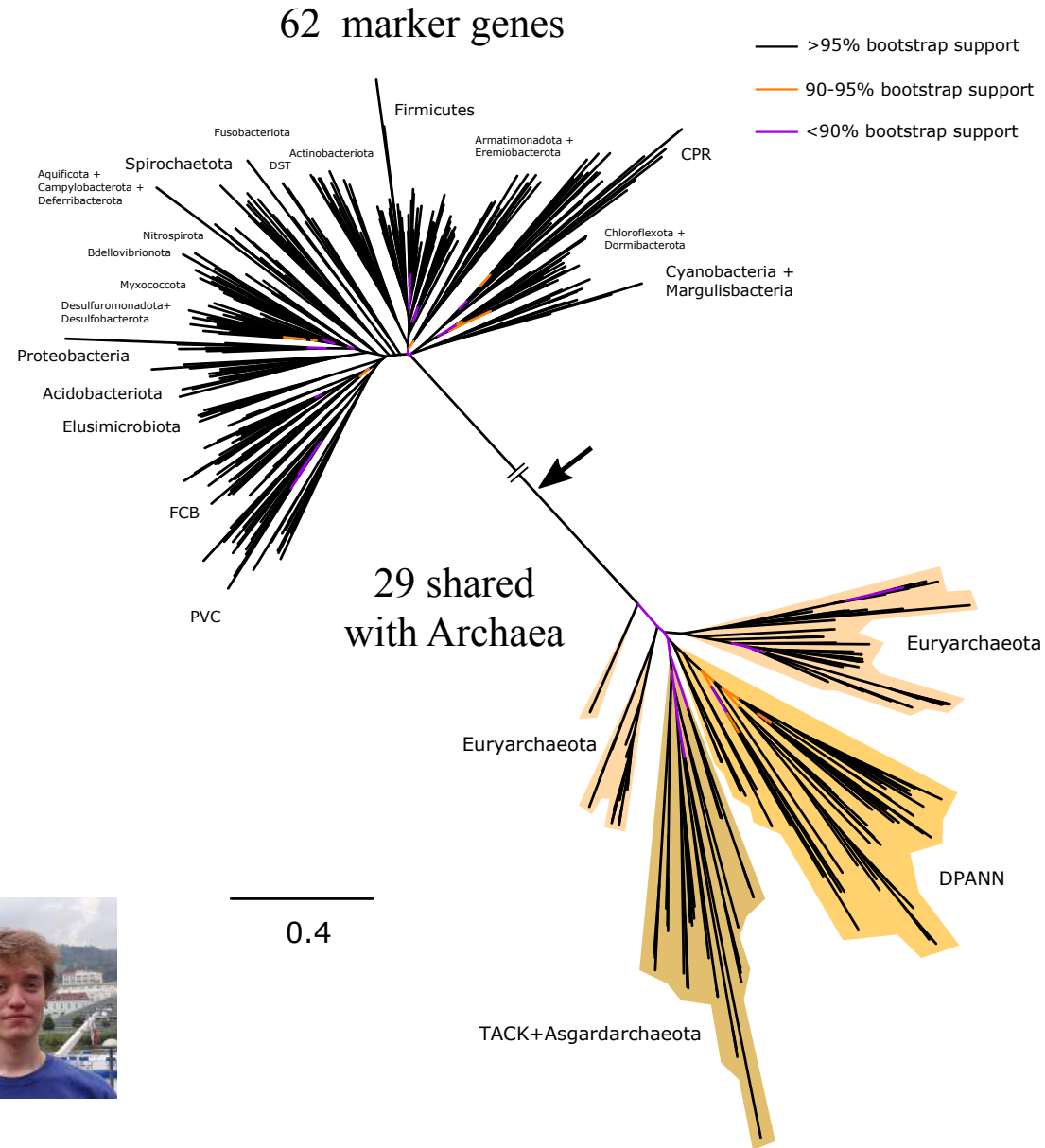


Coleman *et al.*, *Science* **372**, eabe0511 (2021) 7 May 2021

ssolo@elte.hu

Using the history of 1% as a scaffold

Can an archaeal outgroup resolve the root?



Coleman *et al.*, *Science* **372**, eabe0511 (2021) 7 May 2021

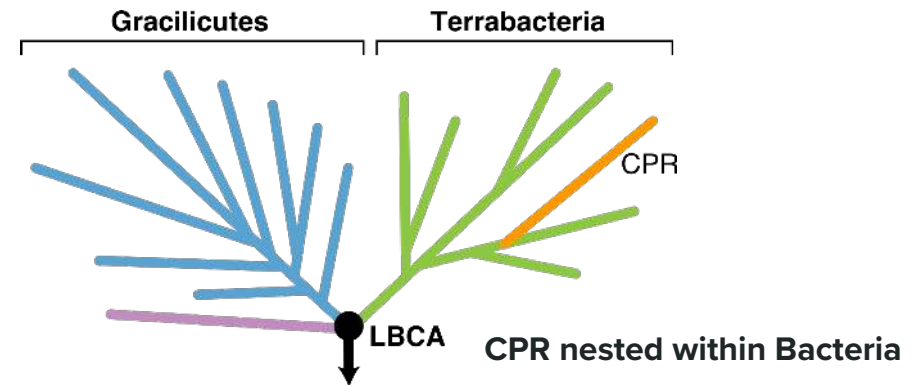


Using the history of 1% as a scaffold

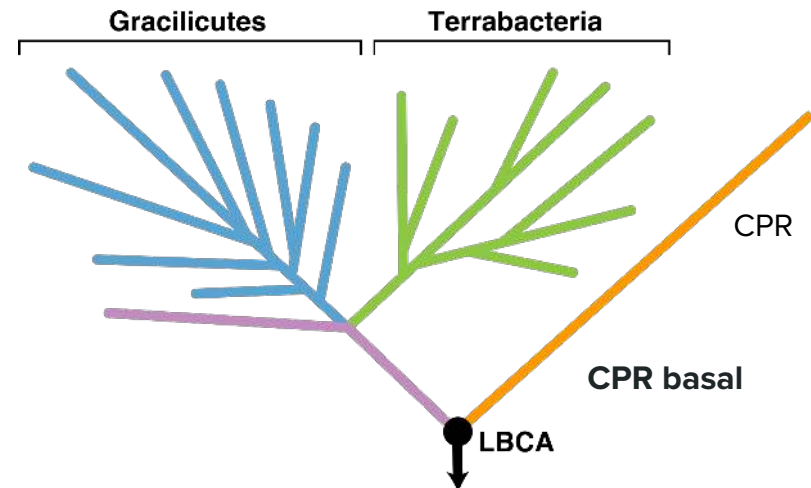
The archaeal outgroup *cannot* resolve the root

Root hypothesis	log-likelihood difference to ML	p-value	Study
Observed outgroup root (Fig. S1)	0	0.58	This study (ML tree)
Between Firmicutes and Actinobacteriota	-3.3	0.50	(26)
Deinococcota, Synergistota and Thermotogota basal*	-3.7	0.52	
Planctomycetota basal	-4.9	0.49	(8)
Chloroflexota basal	-6.2	0.52	(9)
CPR basal	-16.7	0.37	(11, 16)
DPANN basal within archaeal outgroup	-19.4	0.37	(1, 20)
Fusobacteriota basal	-24.2	0.33	
Between Gracilicutes and Terrabacteria	-24.5	0.33	This study (ALE root, see below)

Table S2: Support for published hypotheses using outgroup rooting. *Our unrooted topology was incompatible with some published hypotheses, including a clade of Thermotogales and Aquificales at the root (6, 7).



- More complex, free-living ancestor
- Diderm (double-membraned) ancestor?

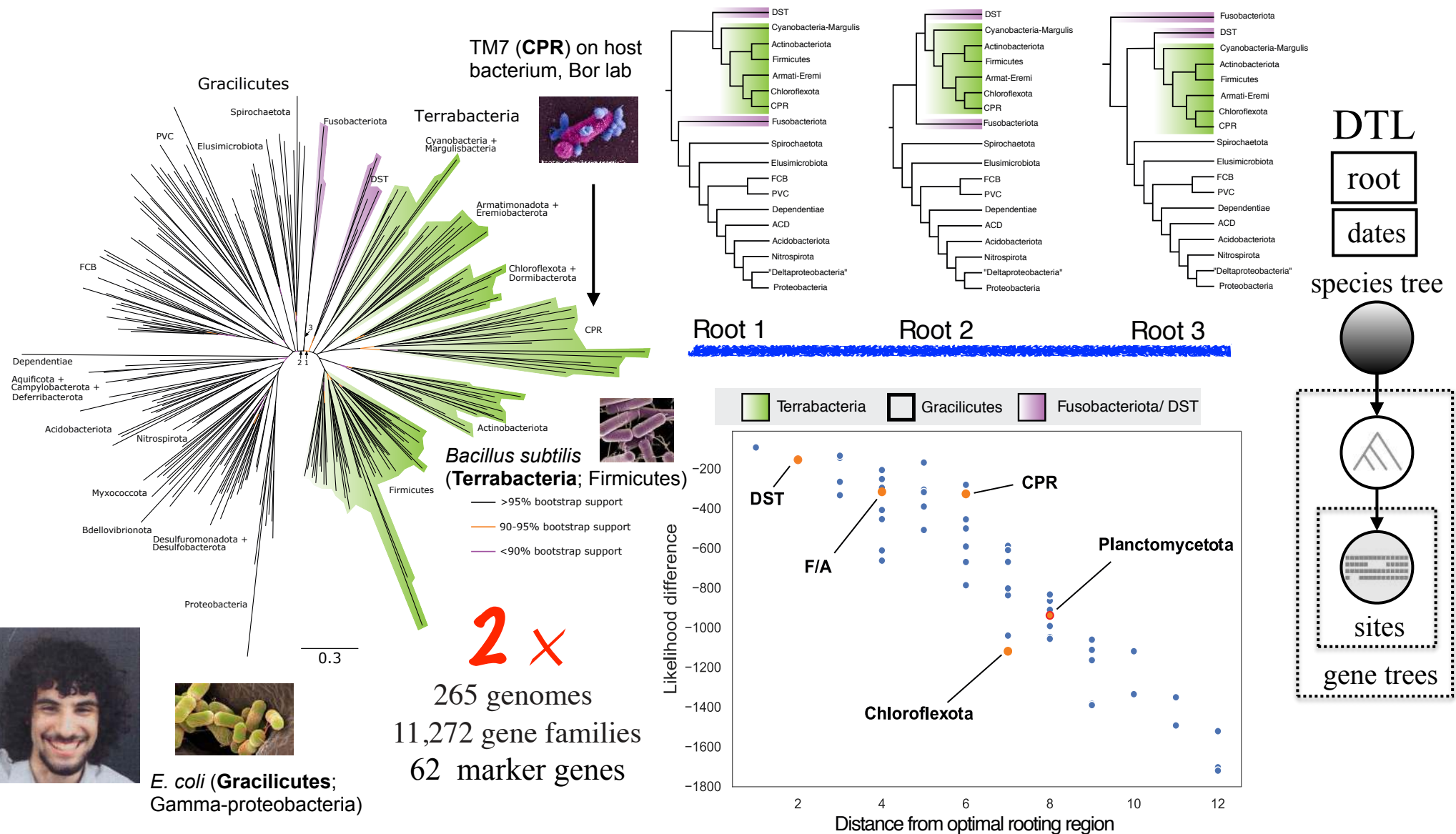


- Small-genomed, host-associated ancestor?
- Monoderm (single-membraned ancestor?)



Transfers can root the bacterial tree of life

The root falls between two major clades of Bacteria, the Gracilicutes and the Terrabacteria, on one of three statistically equivalent adjacent branches

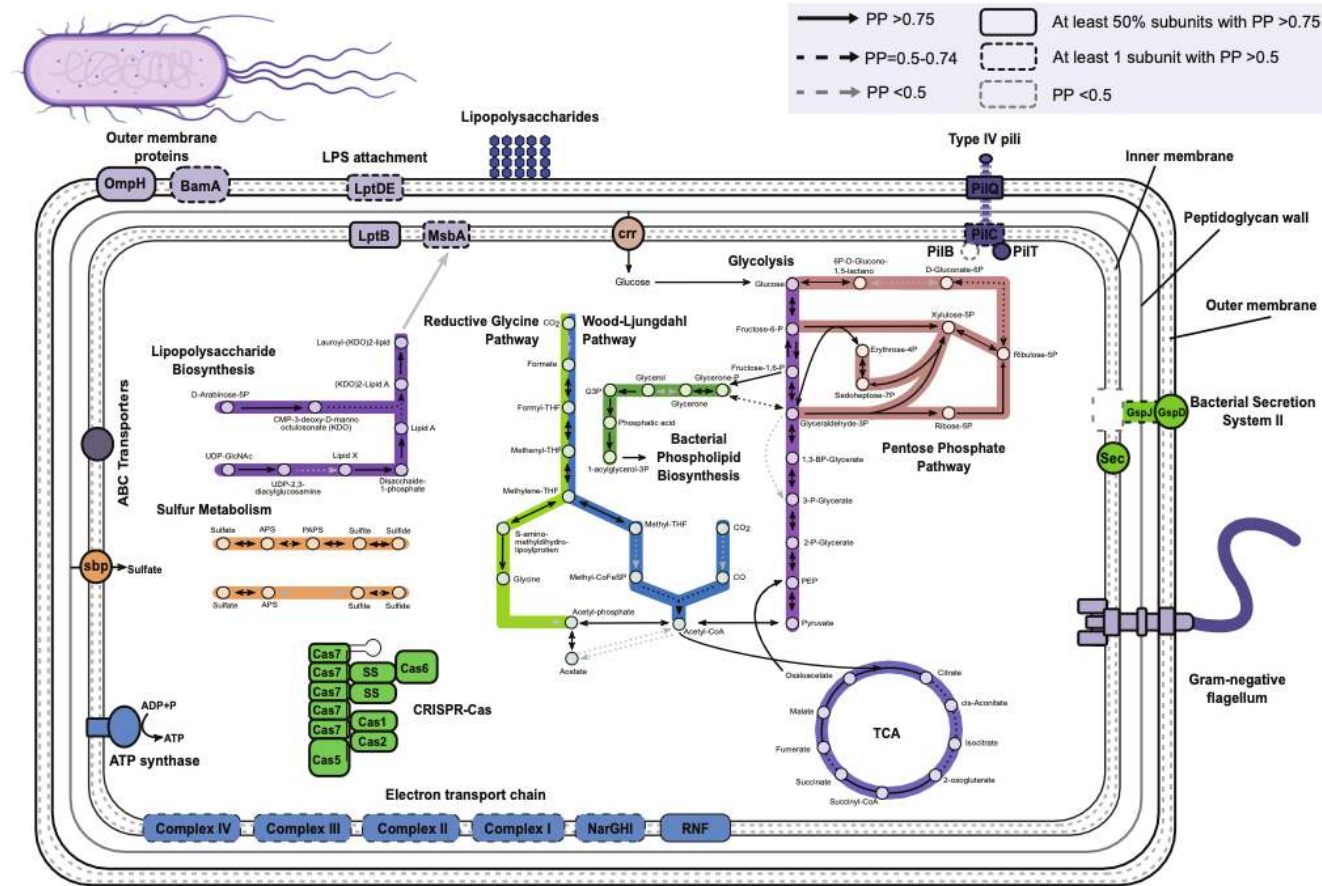


Resolving LBCA

2x

265 genomes / 11,272 gene families

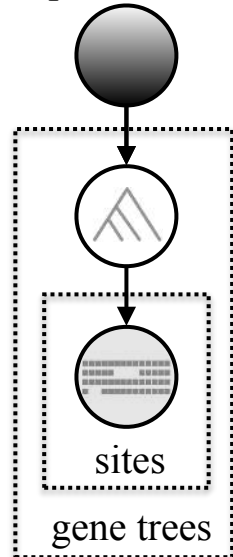
62 marker genes



DTL

ancestral
gene
contents

species tree

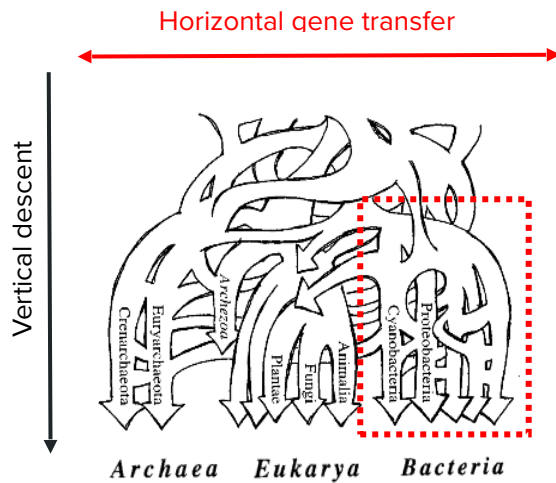


- ➡ Map genes to the root of the inferred tree using reconciliations.
- ➡ Genome size ~1.6-2.7Mb.
- ➡ Anaerobic acetogen.
- ➡ Rod-shaped; flagella and chemotaxis; CRISPR-Cas; **double-membraned**.



just how much HGT is there?

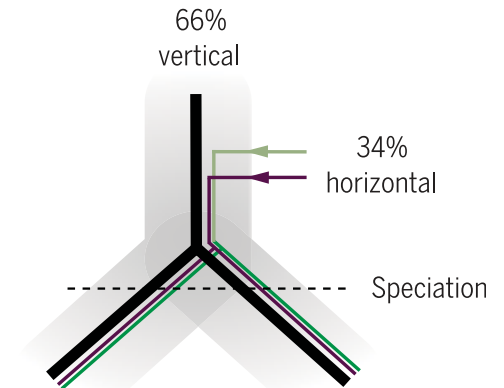
Is there a tree of bacteria?



2x

265 genomes / 11,272 gene families
62 marker genes

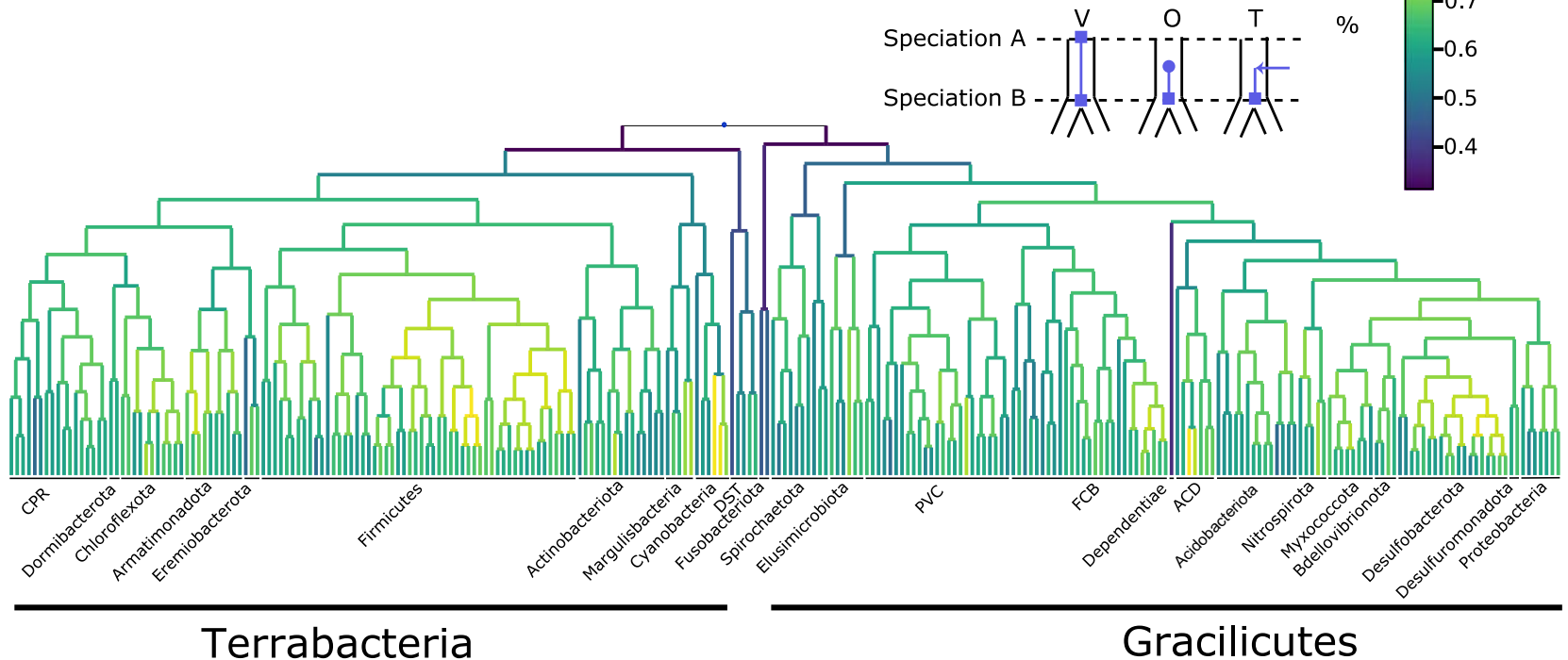
Average gene transmission



% Vertical Evolution

A

Branchwise verticality
 $V/(V+O+T)$



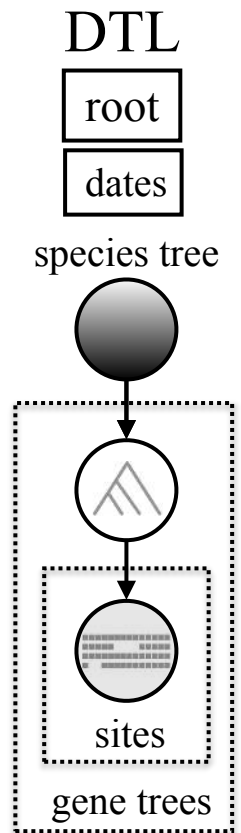
Applying species tree-aware inference



GENECLOCKS

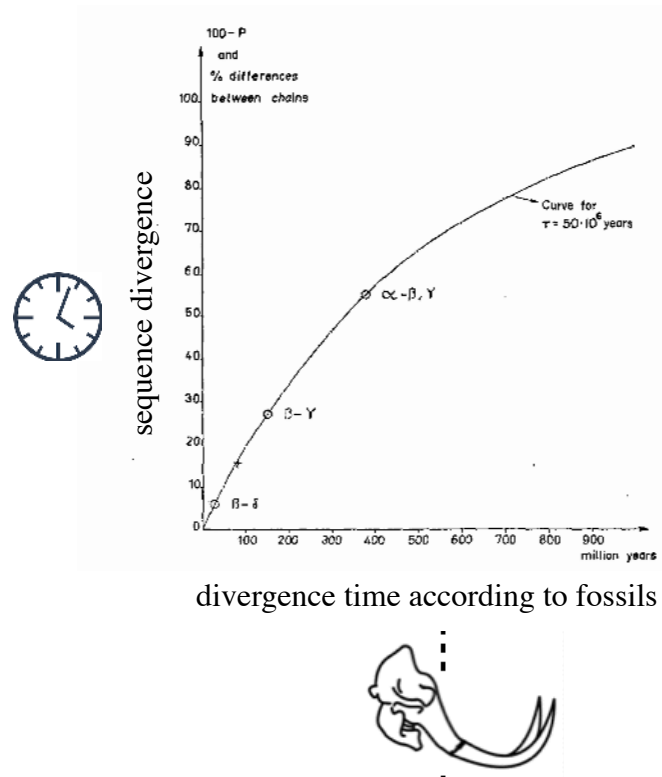
RECONSTRUCTING A DATED TREE OF LIFE USING PHYLOGENETIC INCONGRUENCE

WHERE IS THE DATED TREE?



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**.

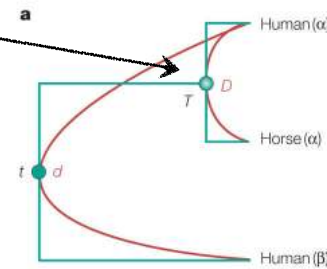


divergence time according to fossils

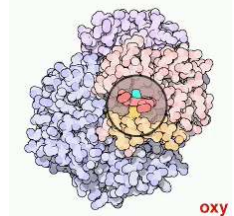
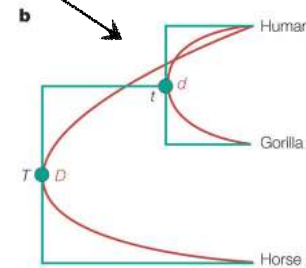


Zukerkandl and Pauling 1965

~130 million years
18 aa substitutions



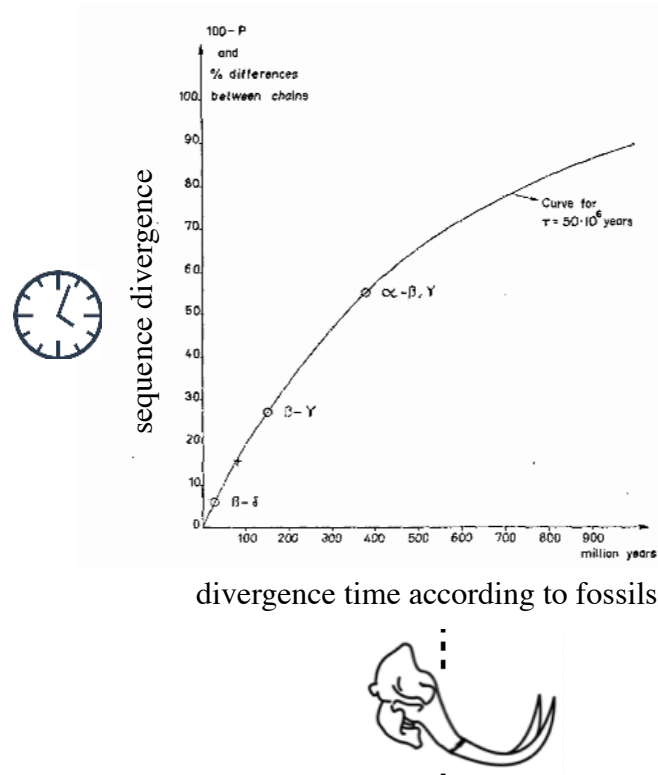
1 & 2 substitutions
~11 million years



$$2\alpha + 2\beta$$

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**.

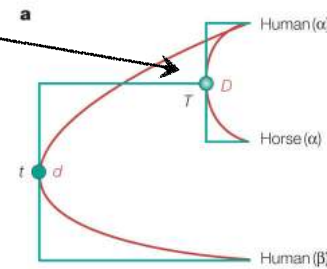


divergence time according to fossils

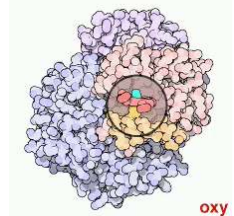
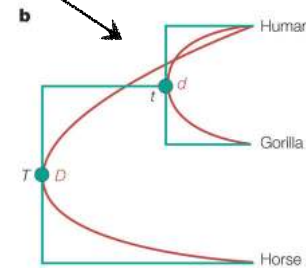


Zukerkandl and Pauling 1965

~130 million years
18 aa substitutions



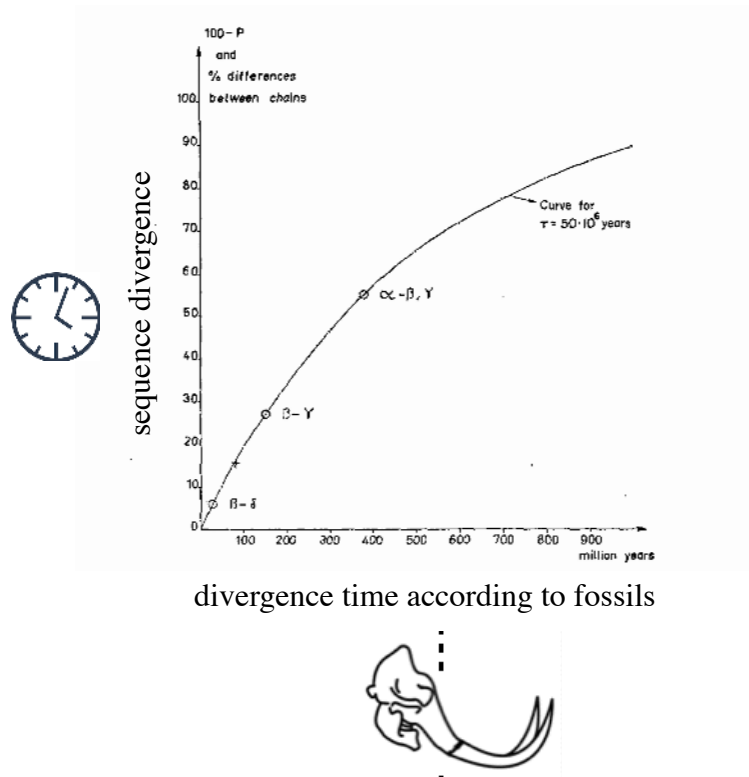
1 & 2 substitutions
~11 million years



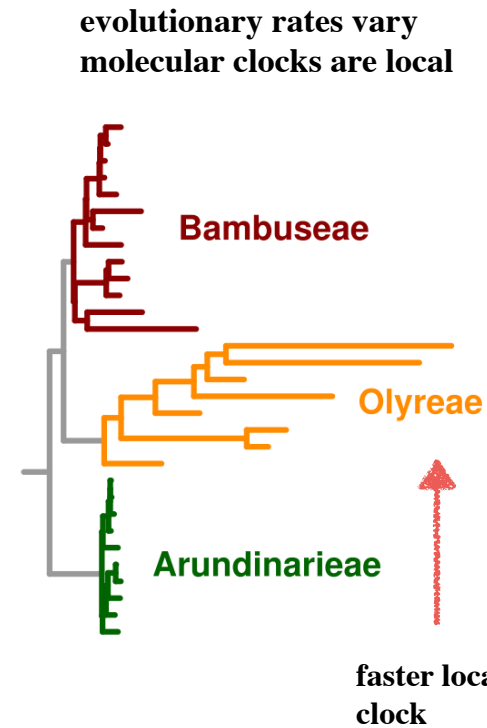
$$2\alpha + 2\beta$$

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**.



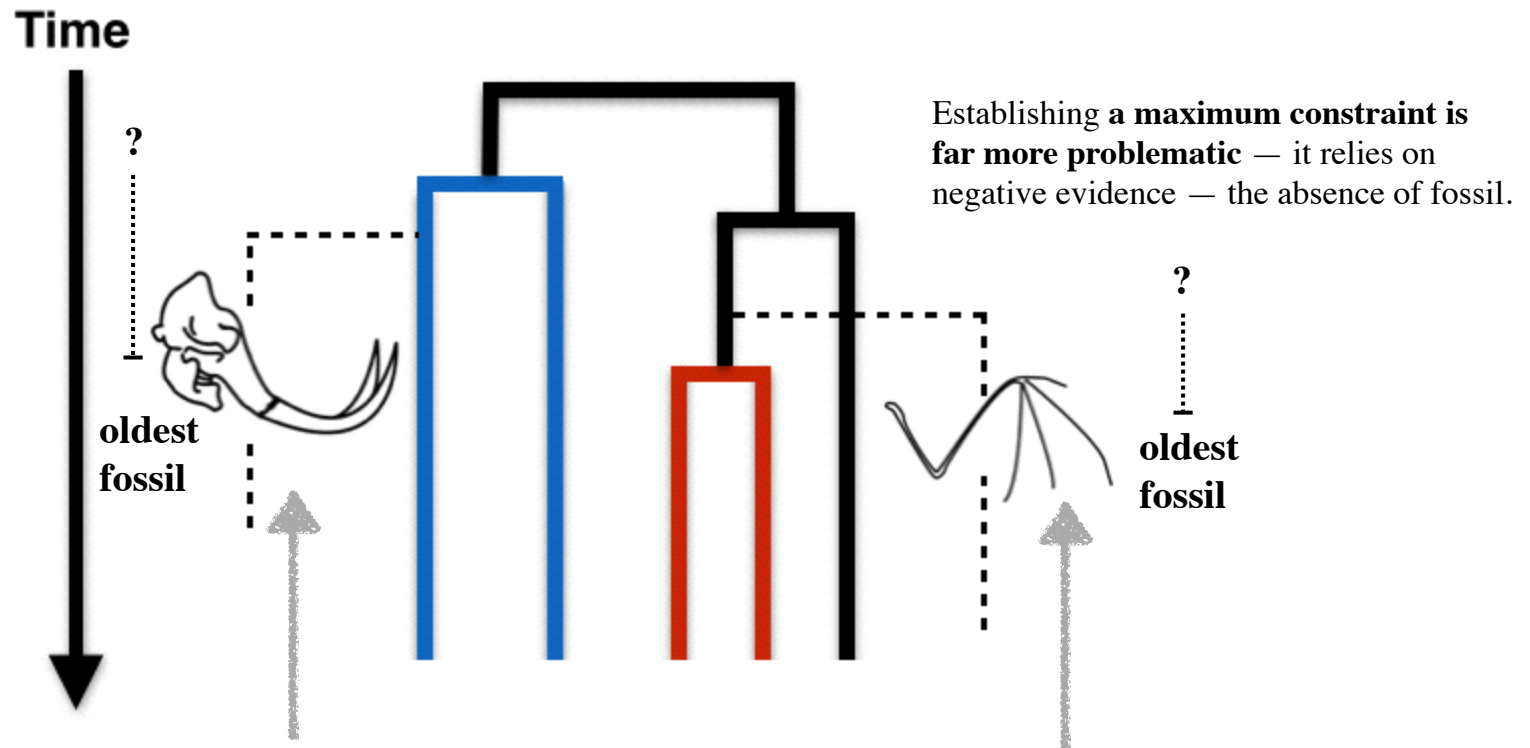
Zukerkandl and Pauling 1965



Wysocki et al. 2014 & Wikipedia

Rocks

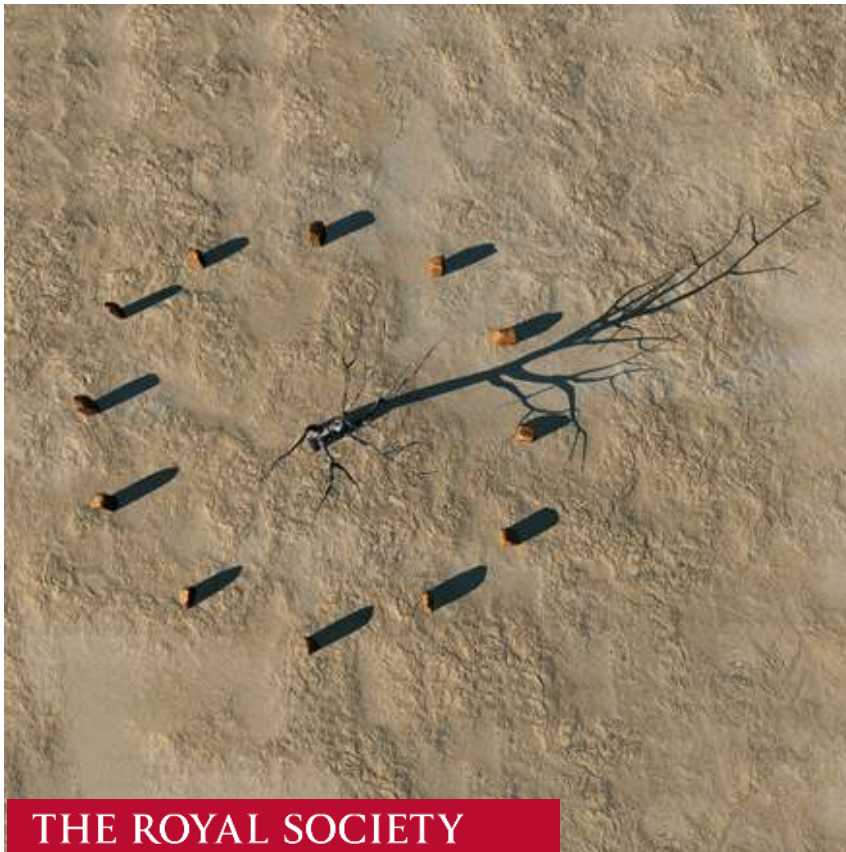
The geological record is the only source of information concerning absolute time



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

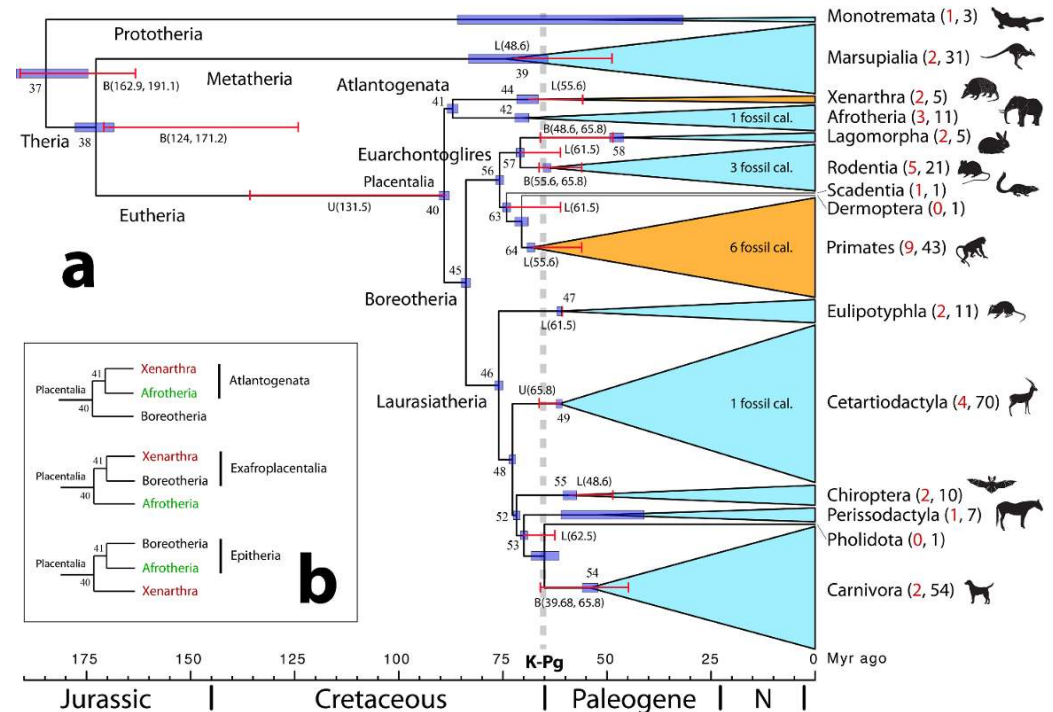
Rocks & Clocks

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



Donoghue and Yang 2015

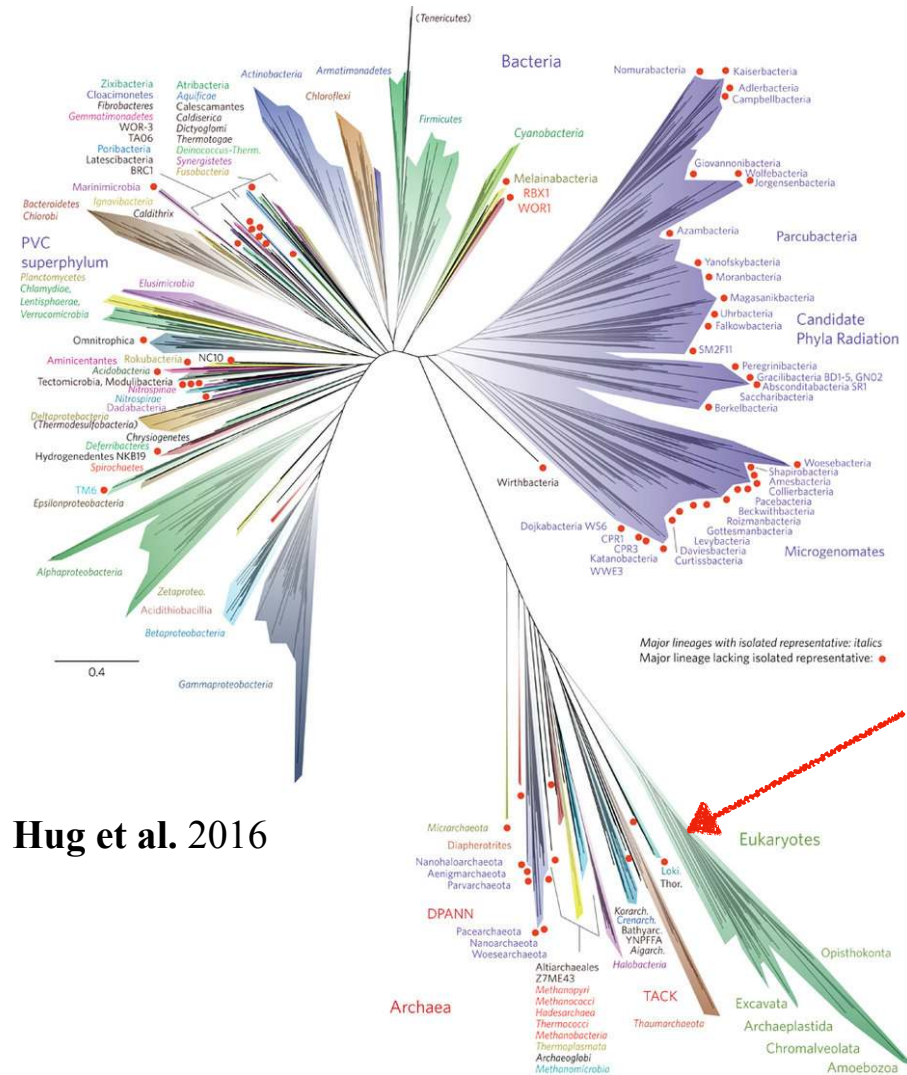
... today, Bayesian RMC methods have resolved most, but not all controversies, using **sequence based local molecular clocks anchored by multiple fossil calibrations**.



dos Reis et al. 2012



Rocks & Clocks

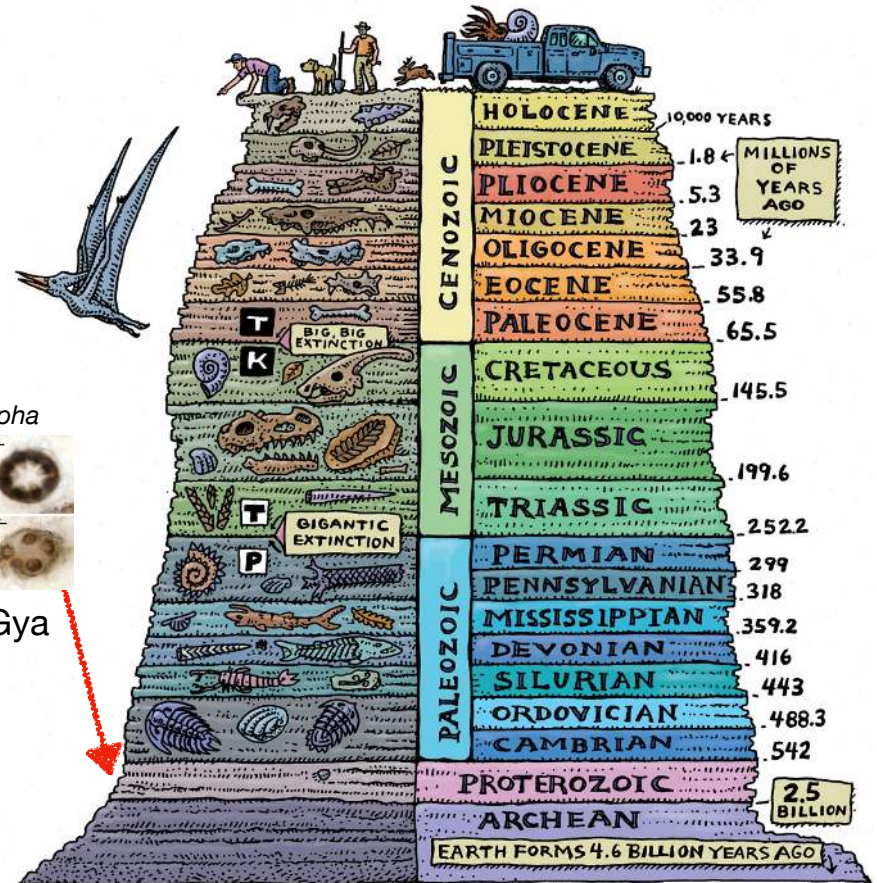


Hug et al. 2016

Bangiomorpha

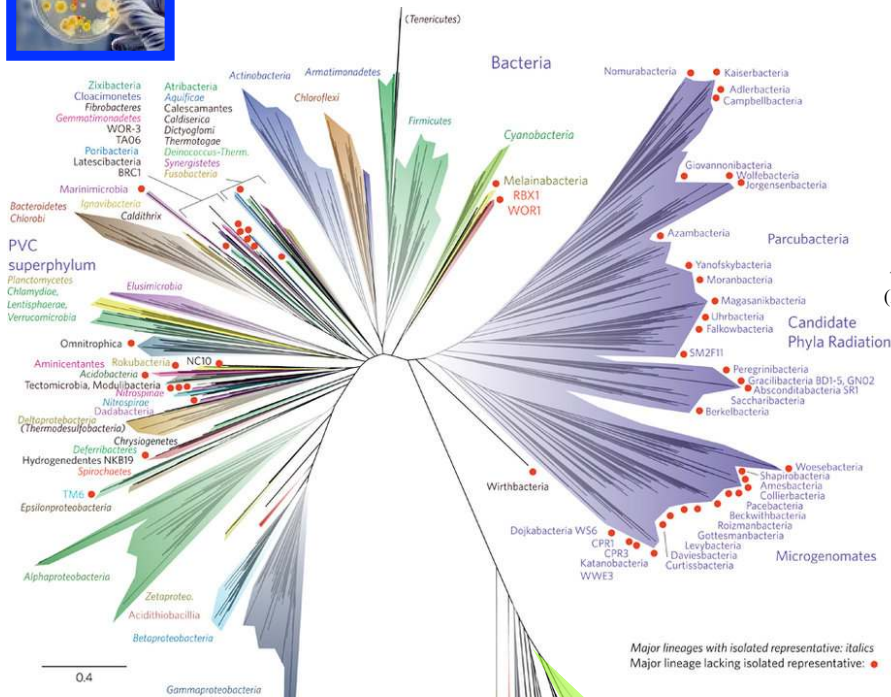


~ 1-1.5 Gya



Rocks & Clocks

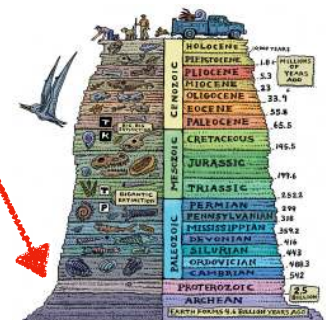
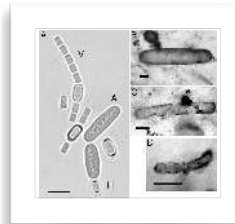
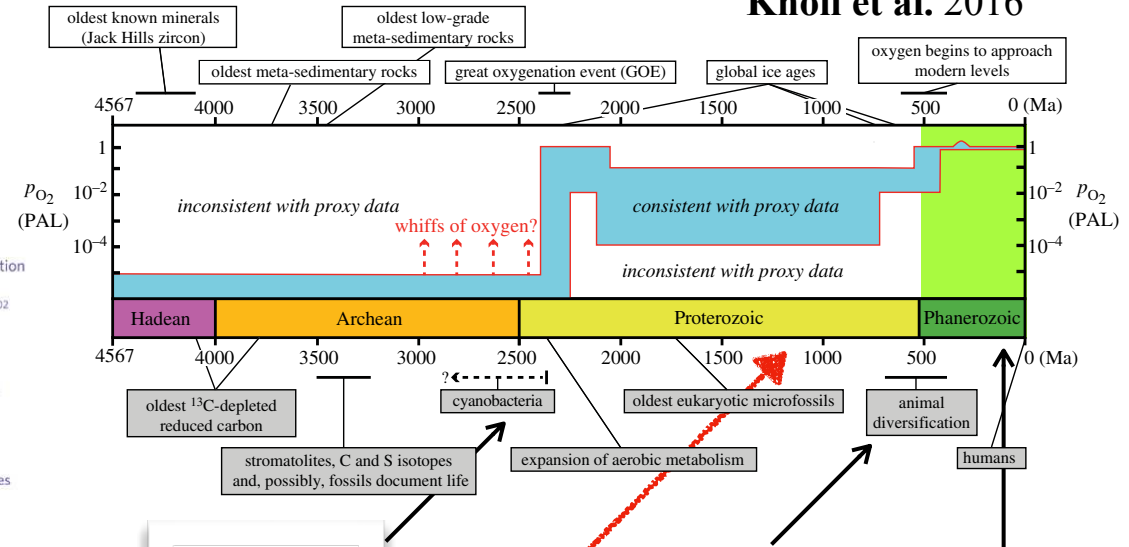
For majority of life and most it's history we lack sufficient fossils to anchor local clocks.



Hug et al. 2016

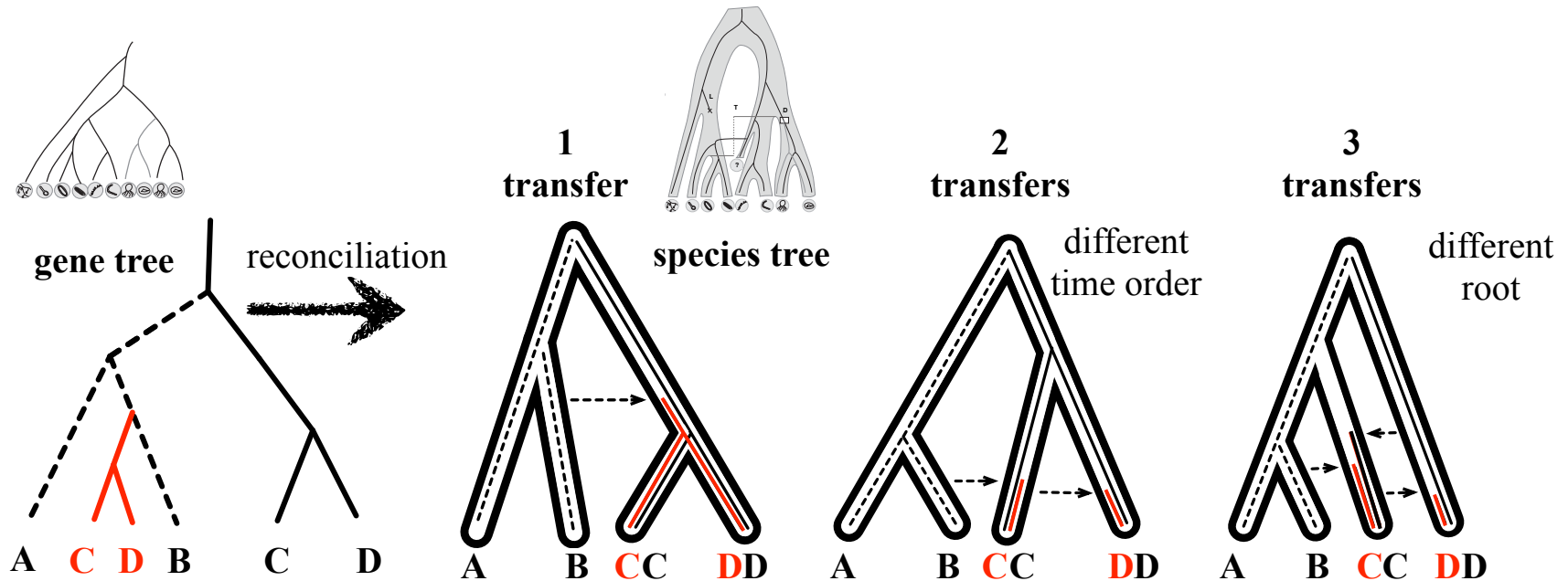


Knoll et al. 2016



Horizontal gene transfer as information

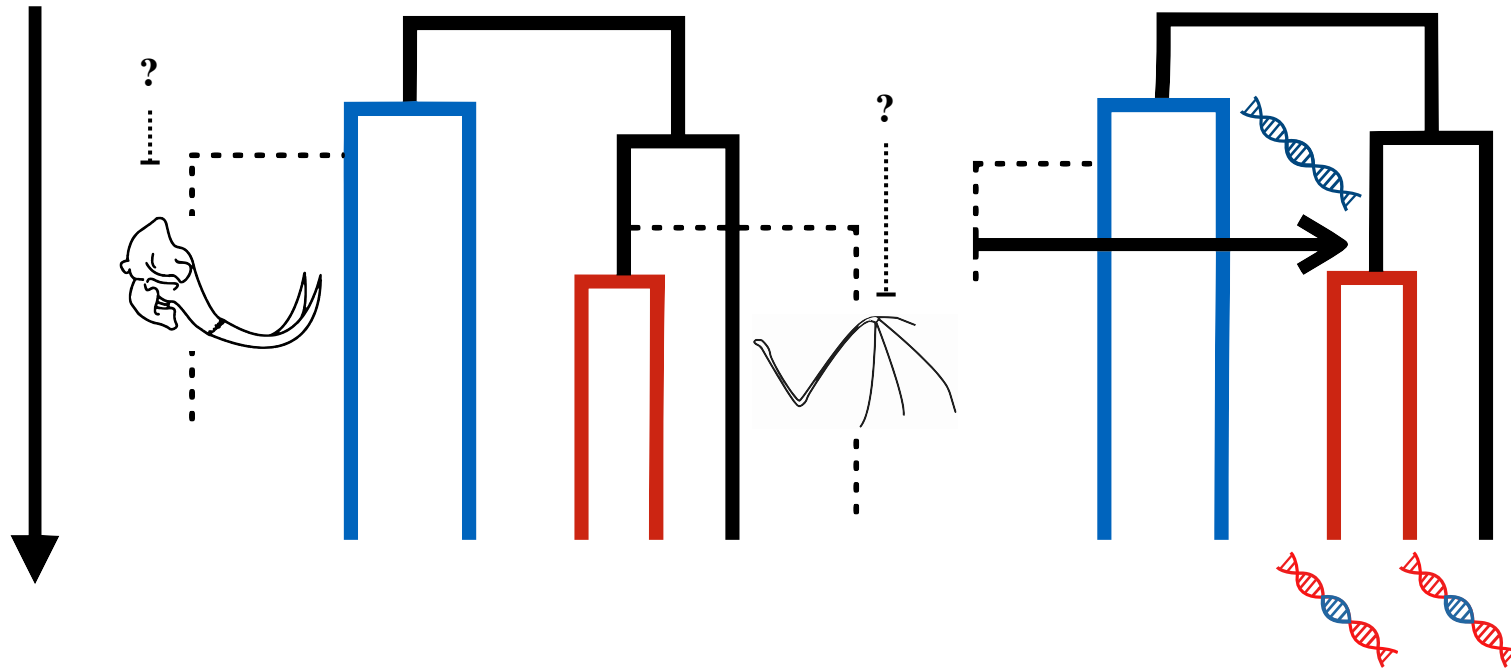
Transfer events, encoded in the topologies of gene trees can be thought of as “*molecular fossils*” that record the order of speciation events.



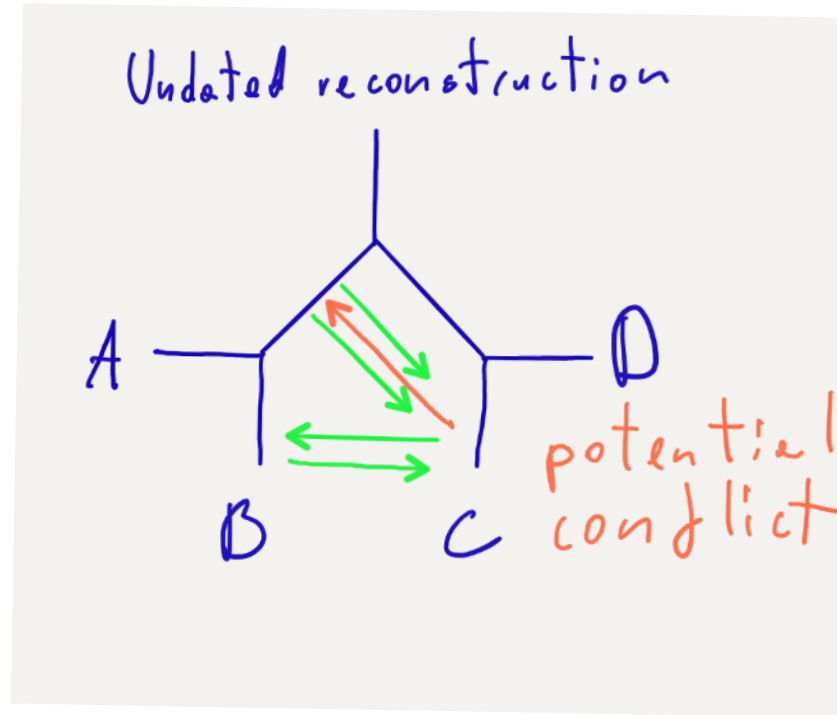
... and genes from other species!

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

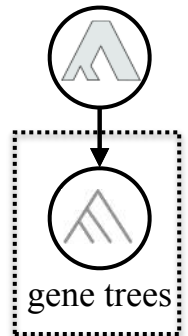
Transfers are not informative on absolute age, but do provide **direct evidence on relative ages**.



“undated” DTL



undated DTL



DTL



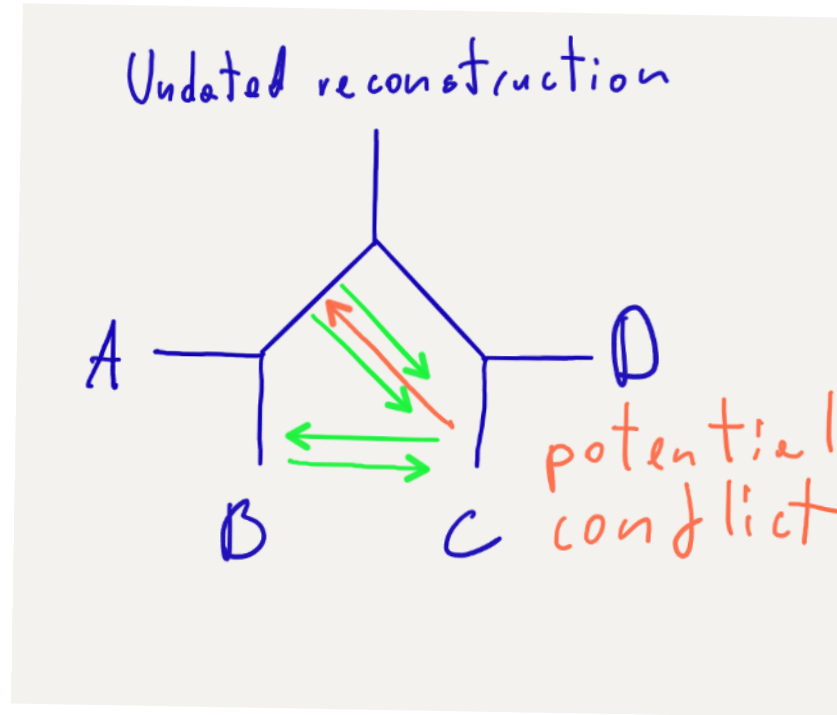
“undated” DTL



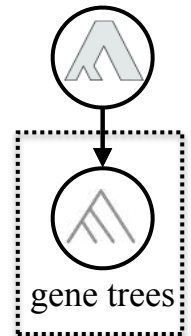
implemented in ALE:

<http://github.com/ssolo/ALE>

“undated” DTL



undated DTL



DTL



“undated” DTL

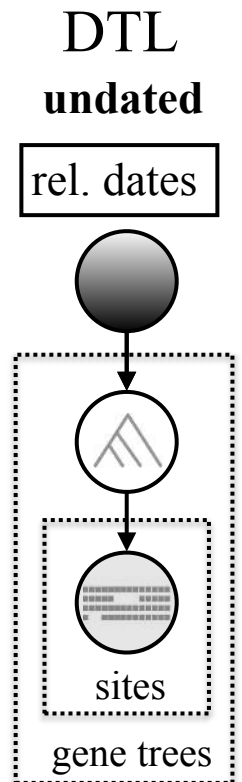
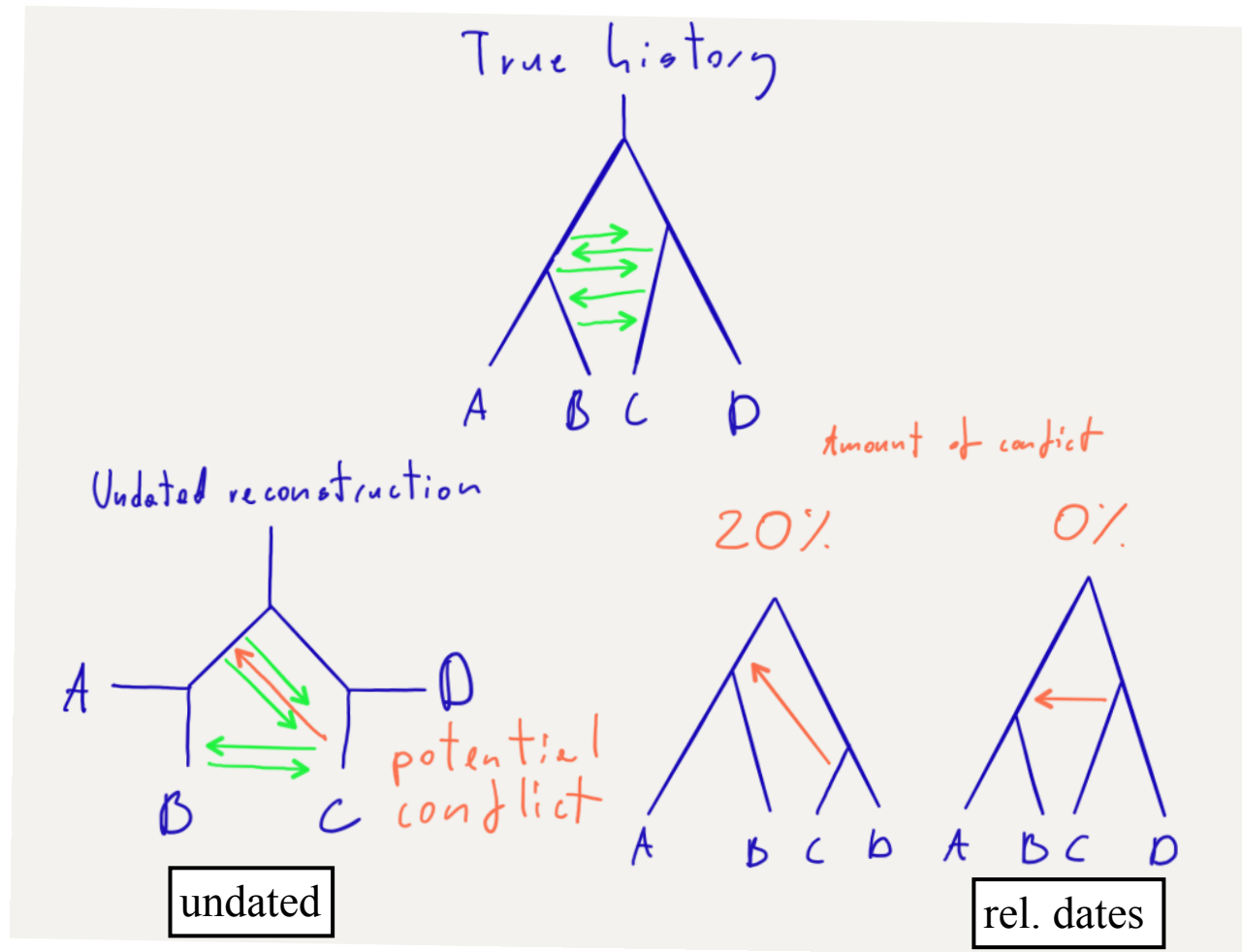


implemented in ALE:

<http://github.com/ssolo/ALE>

Relative age constrains from transfers

HGT events inferred by an “undated” version of the species tree-aware method ALE were input into the MaxTiC (**m**aximal **t**ime consistency) optimisation method to obtain relative age constrains.



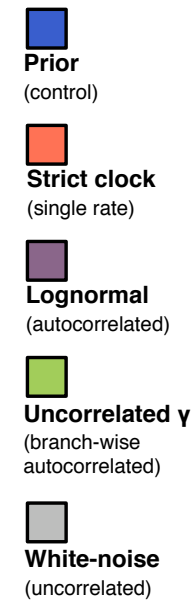
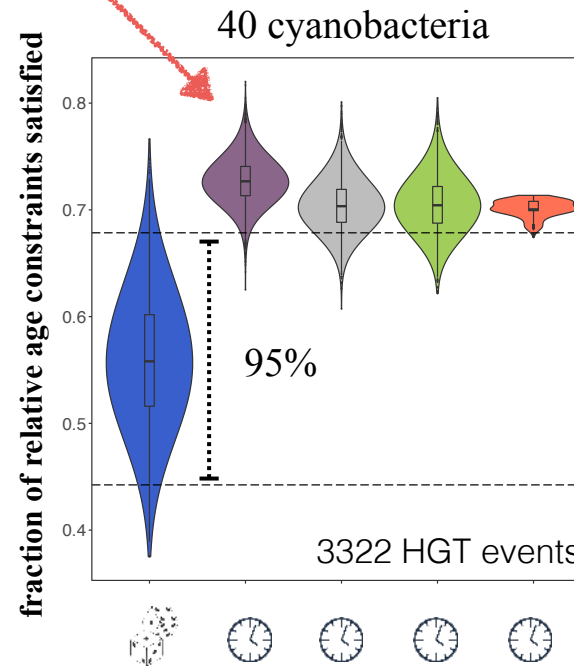
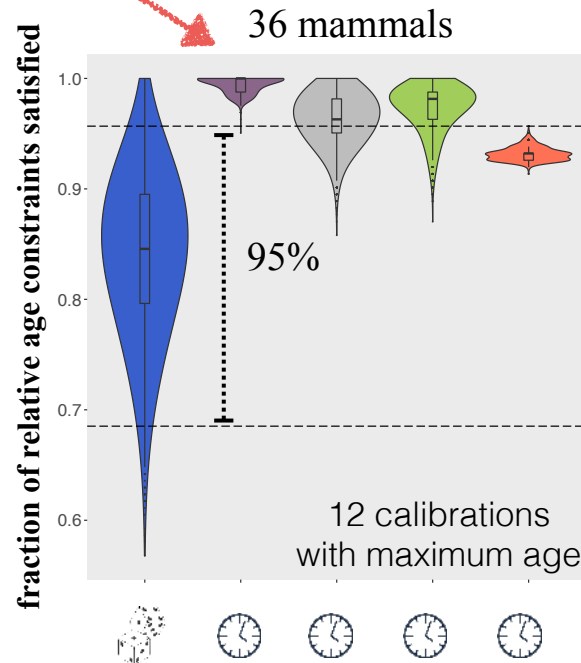
Eric
Tannier

Rocks, clocks and genes from other species

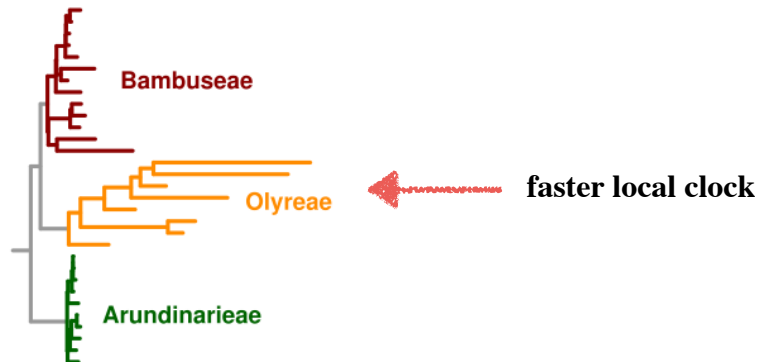
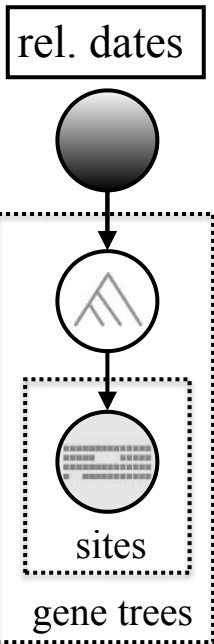
To directly compare relative age constraints, we measured how different relaxed molecular clock models, without fossil calibrations, are able to predict the relative timing of speciations implied by fossils and by transfers.

faster local clock
in rodents

faster local clock in
prochlorococcus



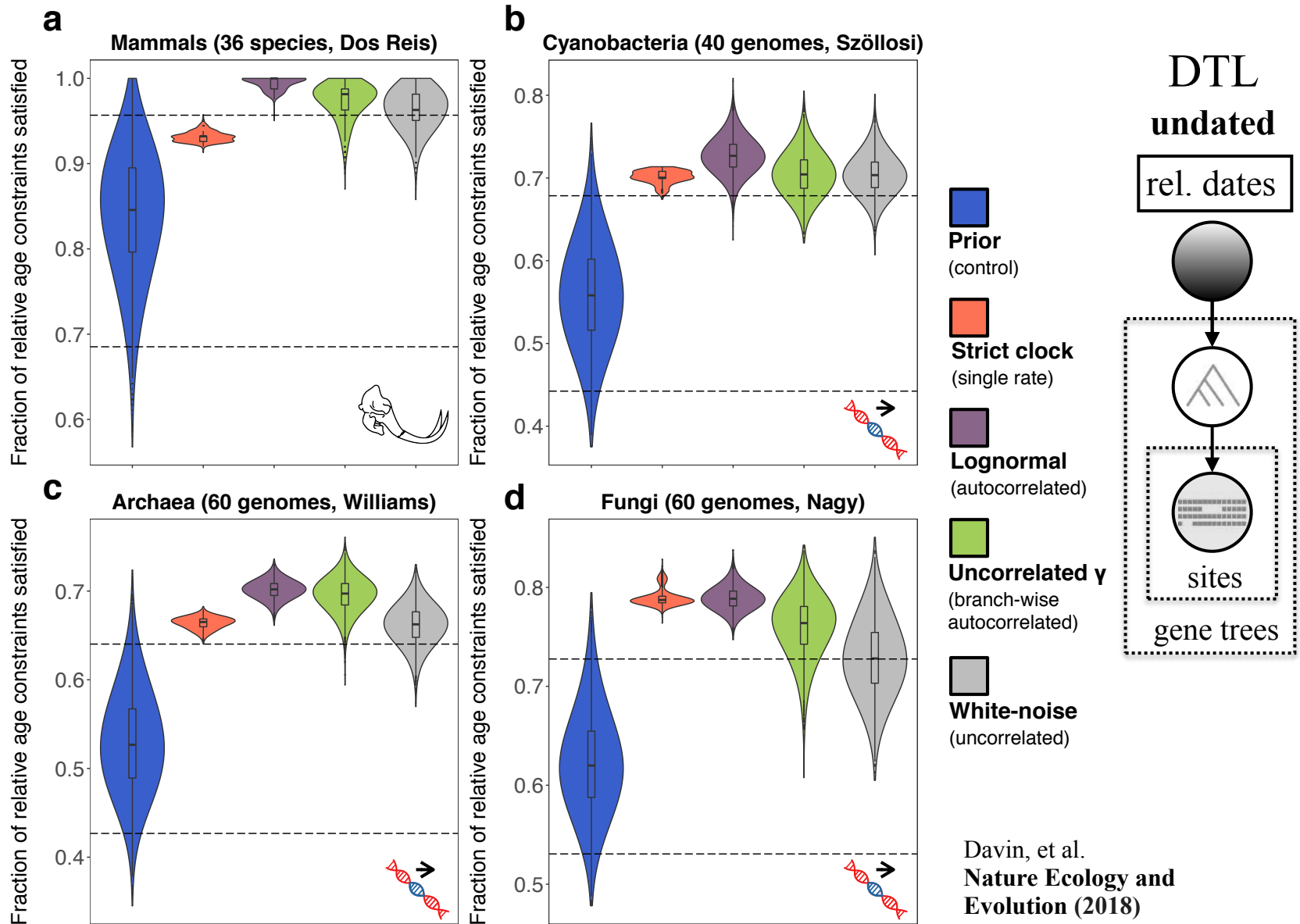
DTL
undated



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

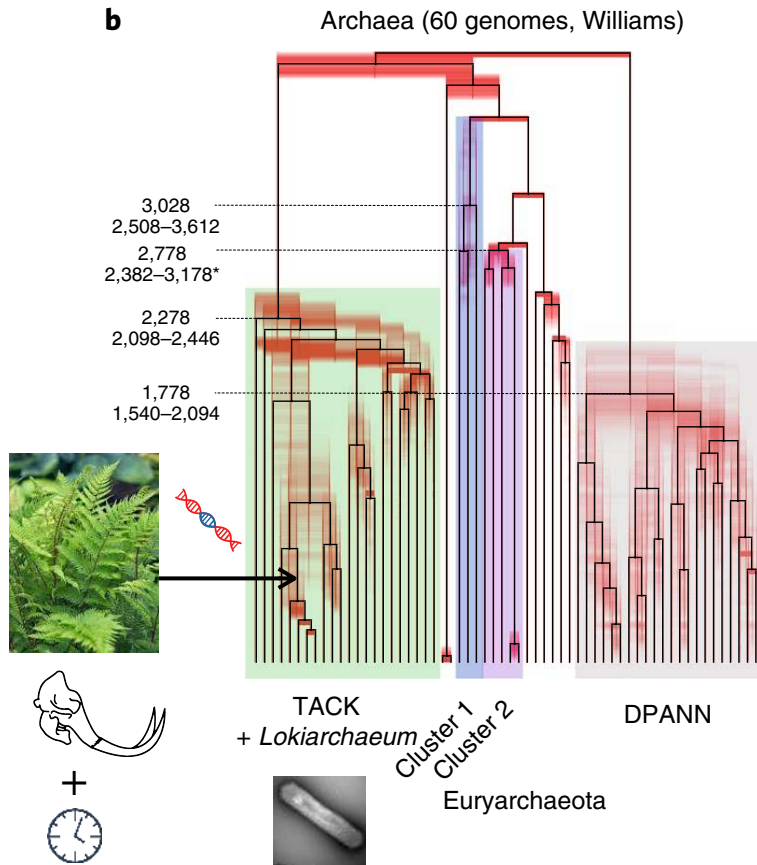
Transfer based relative dates correlate with molecular clocks

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.

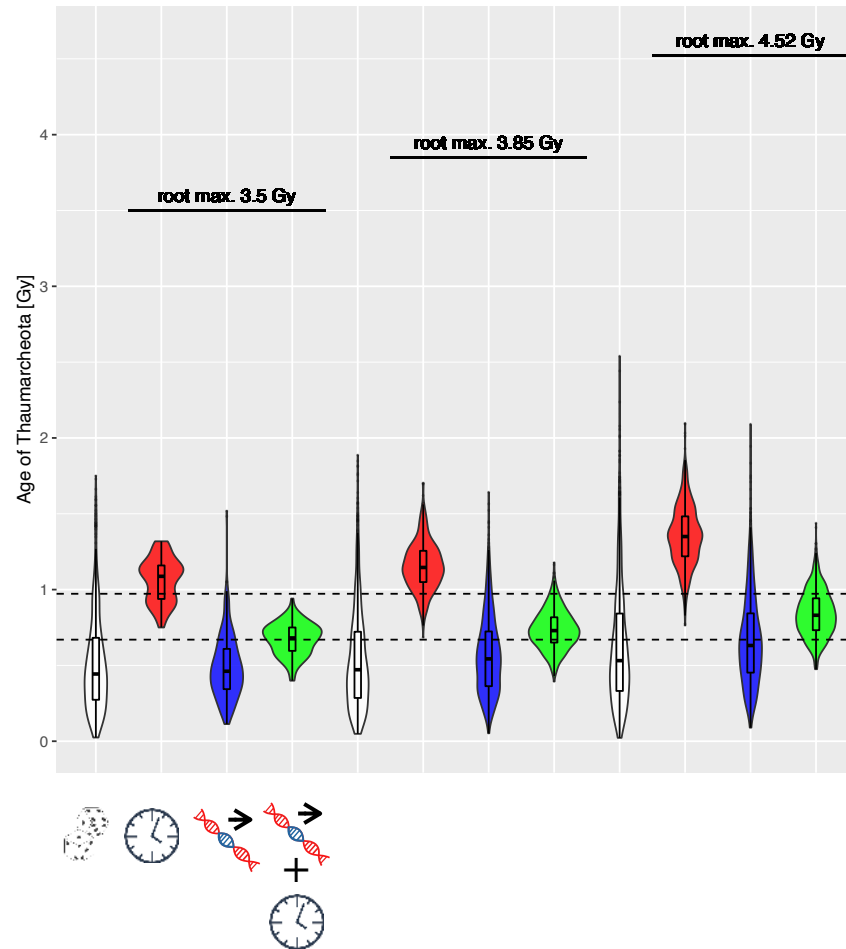


Do clocks & transfers predict rocks?

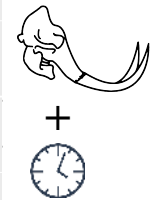
Combining **RMCs with relative constraints and fossil calibrations** we can infer dated trees in RevBayes



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

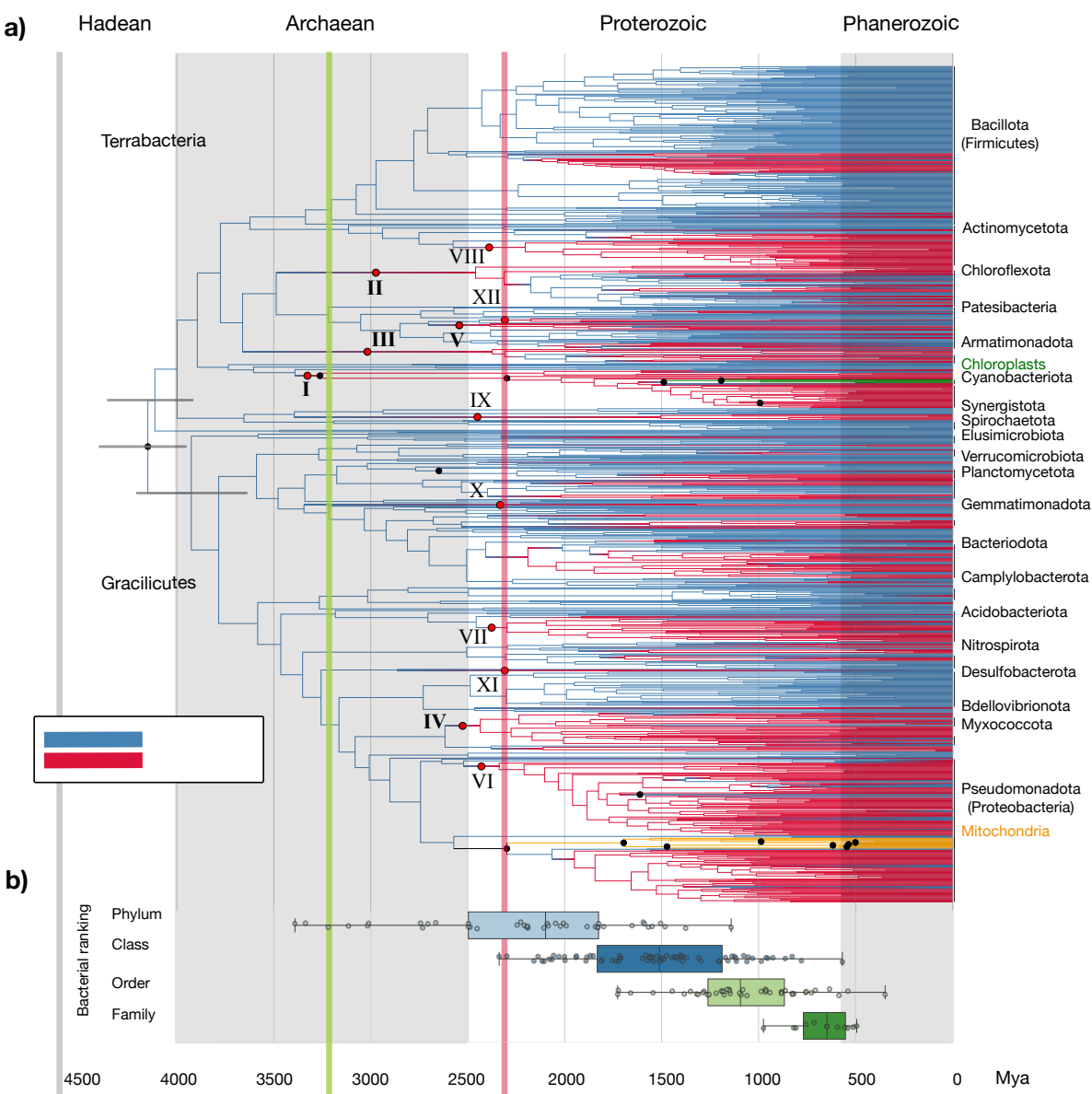


implemented in **MCMCdate**

<https://github.com/dschrempf/mcmc-date>

Work in progress..

A geological timescale for bacterial evolution and oxygen adaptation



P(A | A)

Condition	log-likelihood	improvement over MFI	AU test p-value
Moon-forming impact (MFI)	-4736205	0	1×10^{-14}
MFI + Fossils	-4735968	237	8×10^{-17}
MFI + Fossils + GOE (Concatenate)	-4735871	334	2×10^{-8}
MFI + Fossils + GOE (ExtraTrees)	-4735539	666	9×10^{-6}
MFI + Fossils + GOE (XGBoost)	-4735472	733	1

new “reldate” model



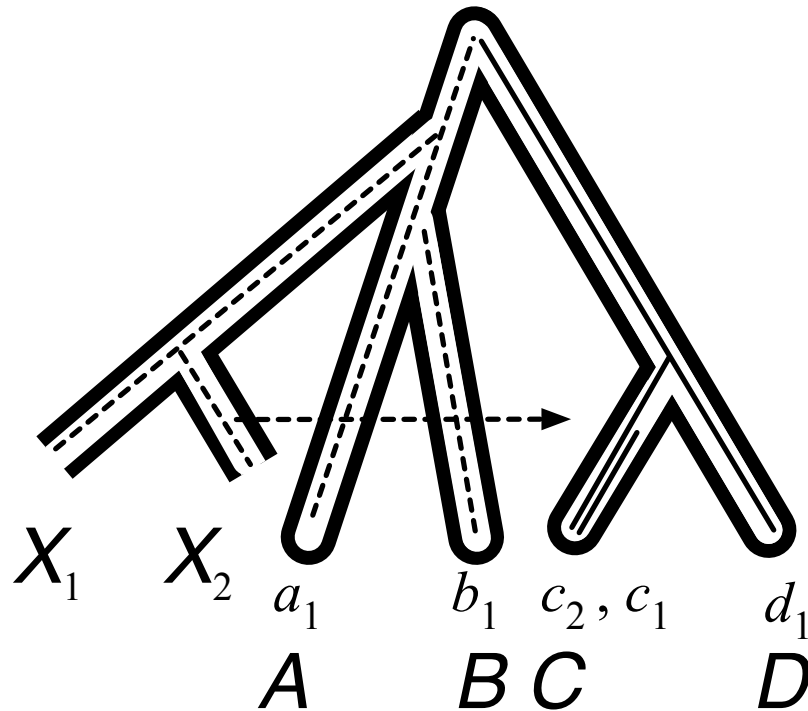
<http://github.com/ssolo/ALE>

Moon-forming impact
Oxygenic photosynthesis
Great Oxidation Event

<https://github.com/dschrempf/mcmc-date>

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.

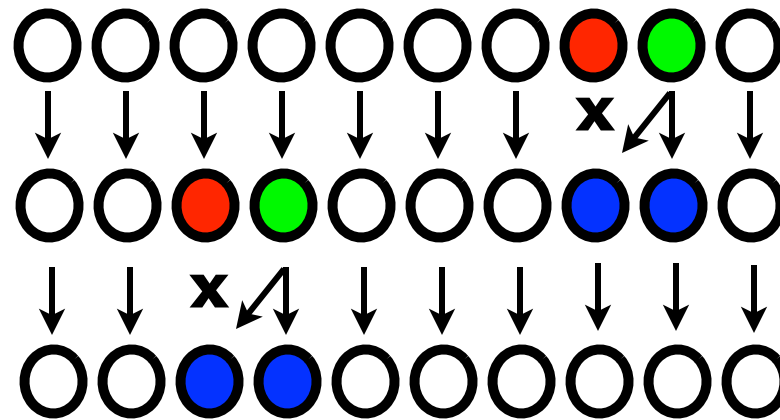


Nicolas
Lartillot

A minimal model of speciation dynamics

the Moran process

N species

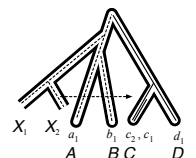


rate per species :

σ

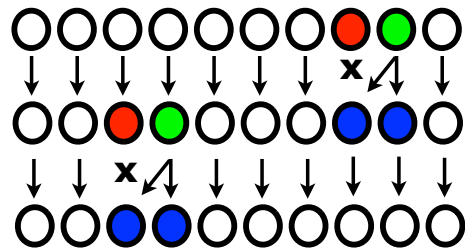
total rate :

$N\sigma$

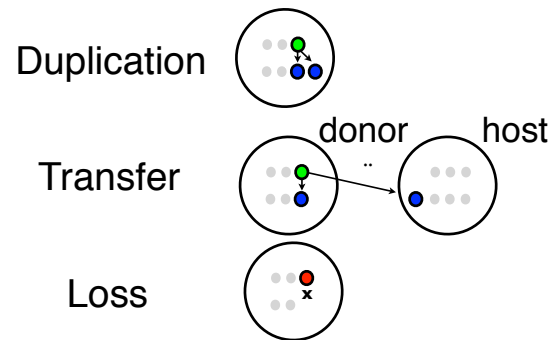


A minimal model of speciation dynamics

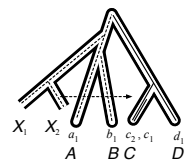
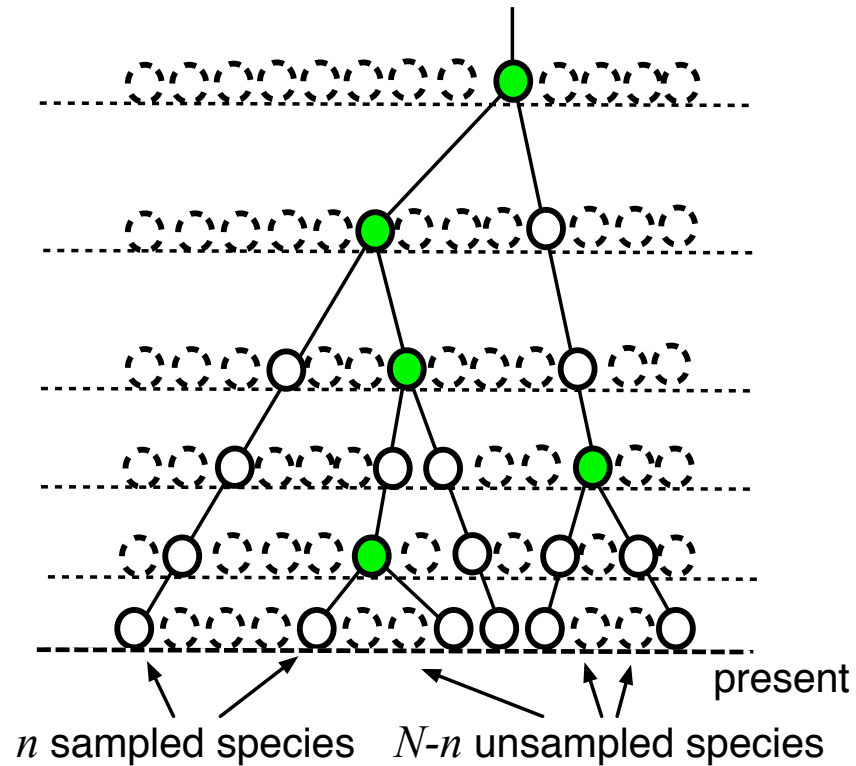
a) speciation dynamics



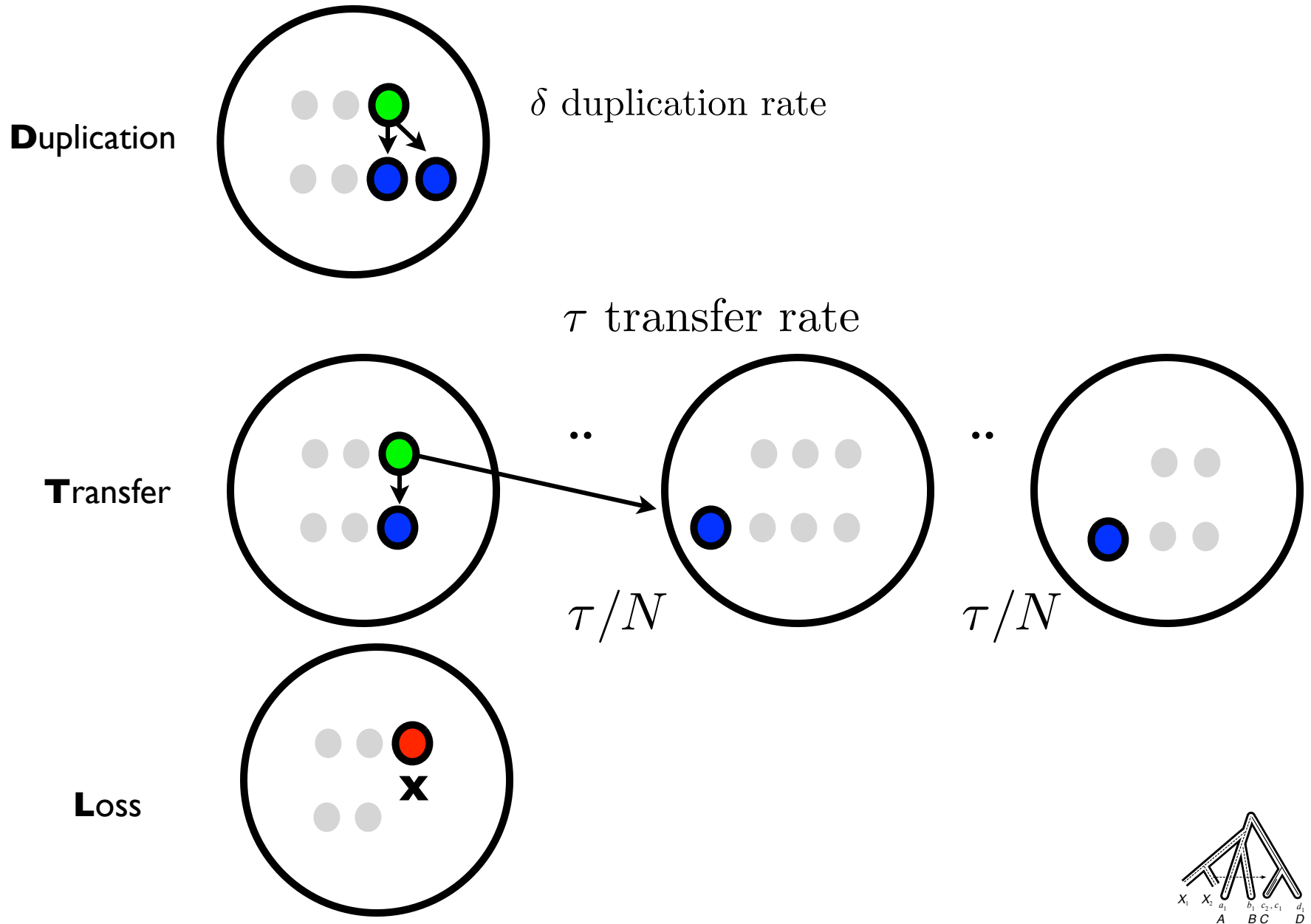
b) gene birth and death



c) represented species phylogeny



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



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

σ speciation/extinction rate

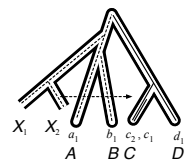
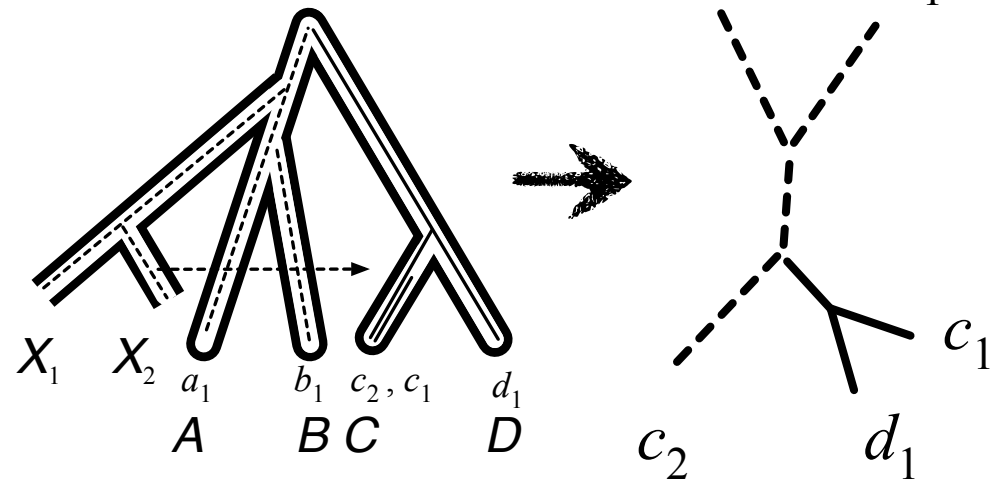
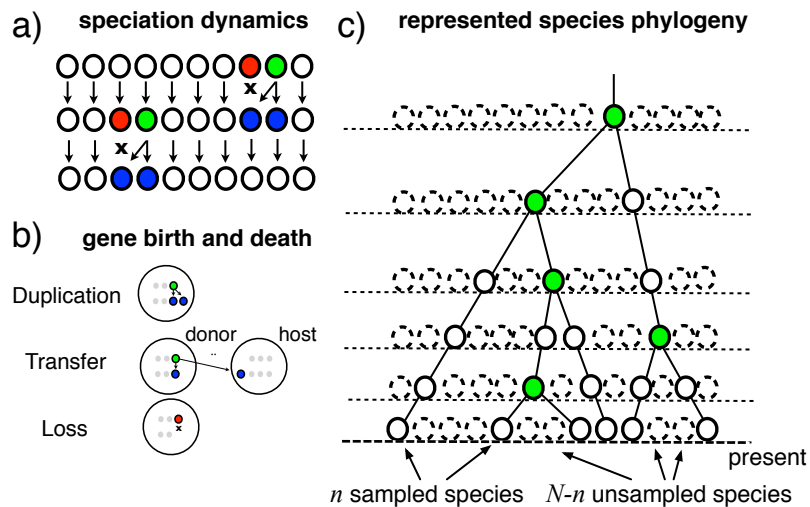
δ gene duplication rate

τ gene transfer rate

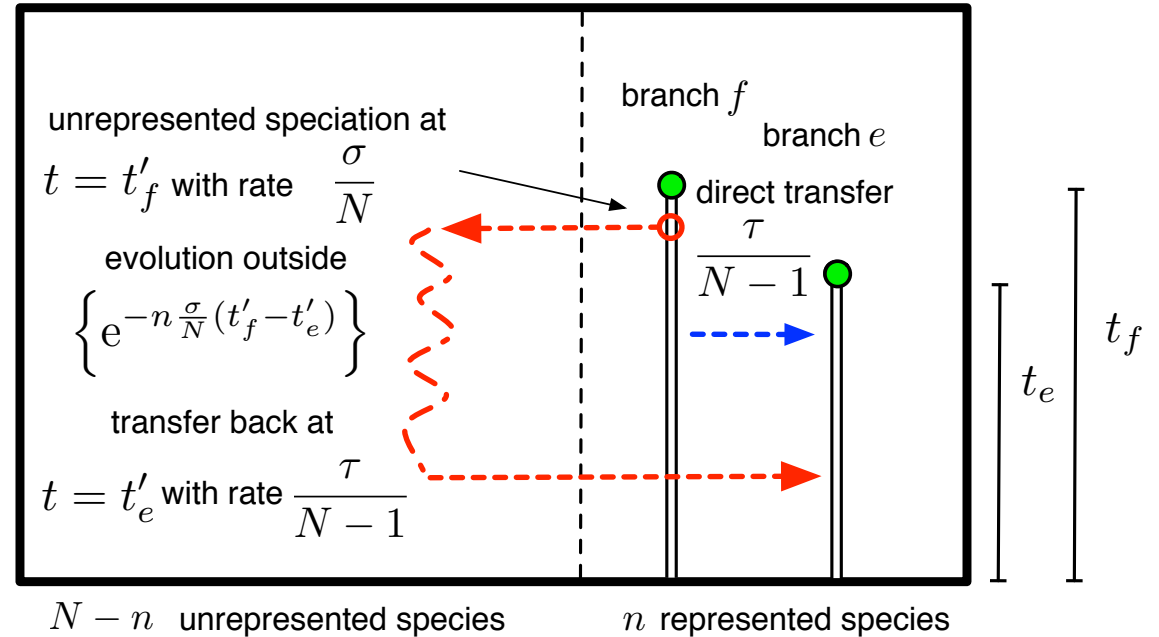
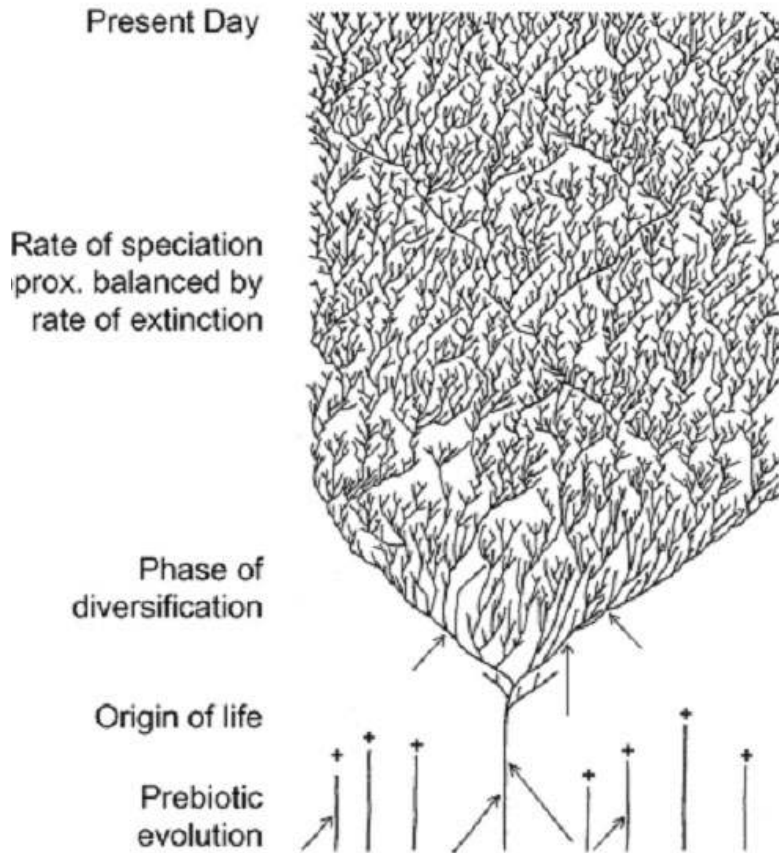
λ gene loss rate



Daubin & Boussau 2011



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

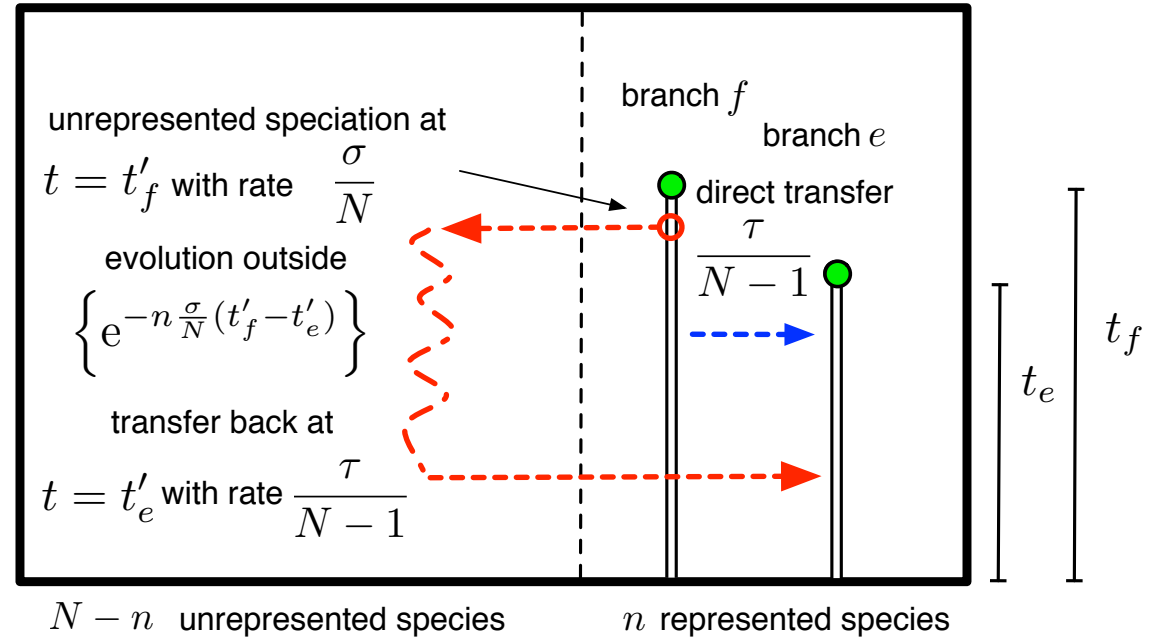
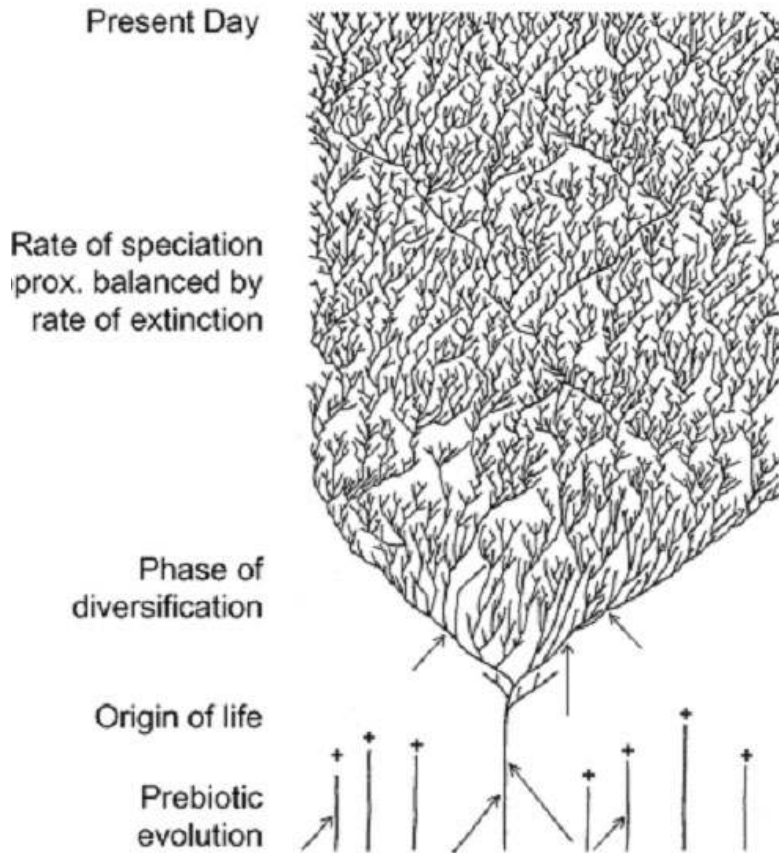


Gogarten lab

$$T_{\text{direct}} \approx \int_0^{\frac{N}{n\sigma}} \frac{\tau}{N} dt'_e = \frac{1}{N} \frac{\tau}{2n} \left[\frac{2N}{\sigma} \right],$$

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

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

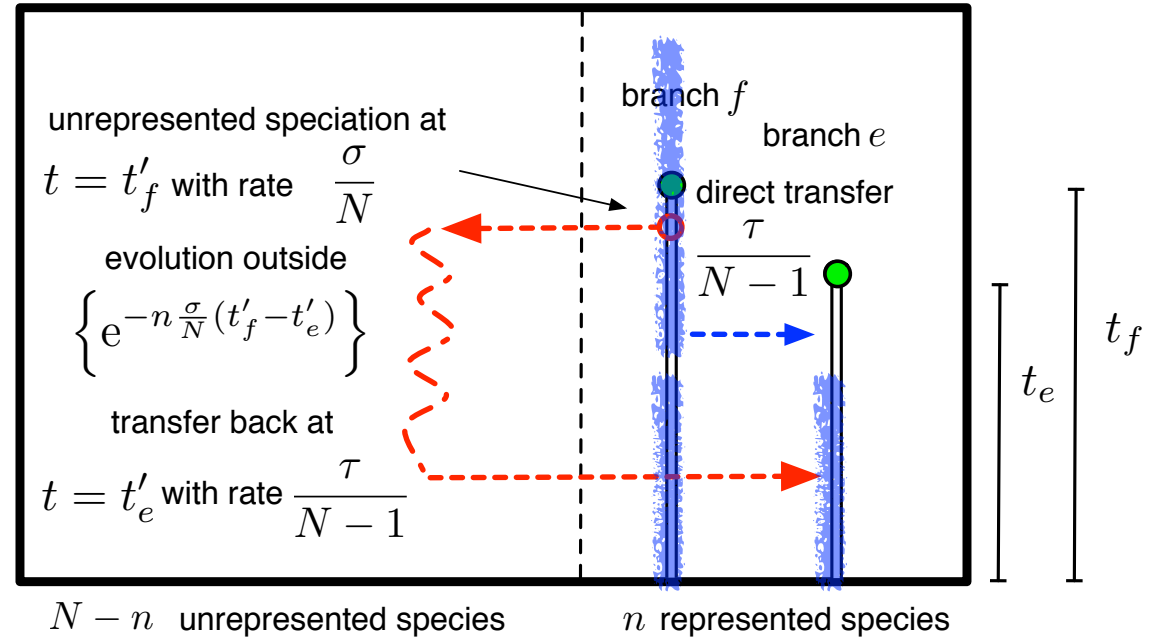
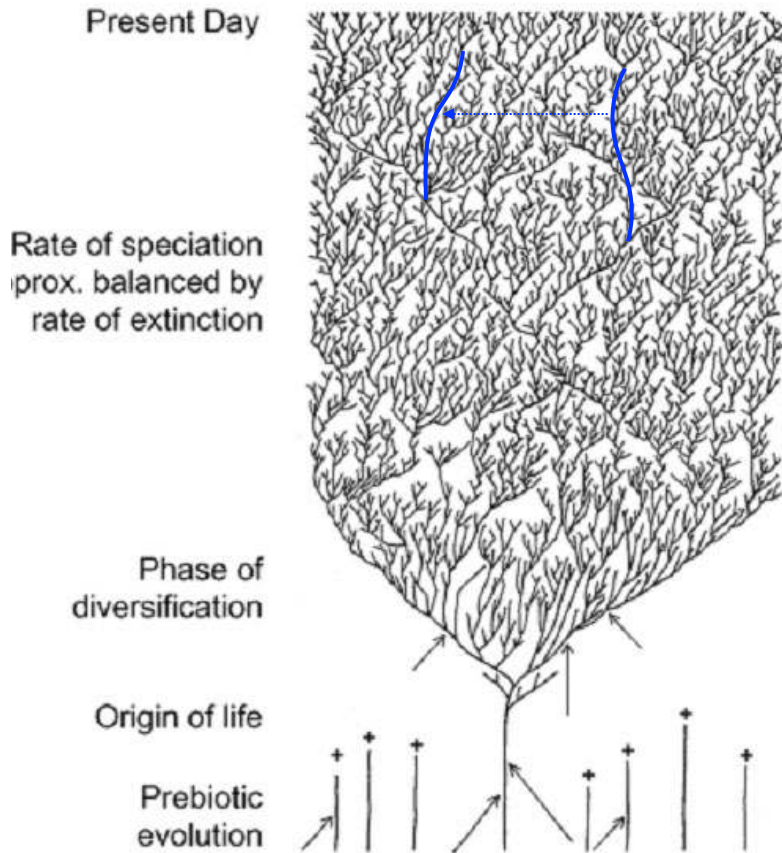


Gogarten lab

$$T_{\text{direct}} \approx \int_0^{\frac{N}{n\sigma}} \frac{\tau}{N} dt'_e = \frac{1}{N} \frac{\tau}{2n} \left[\frac{2N}{\sigma} \right],$$

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

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

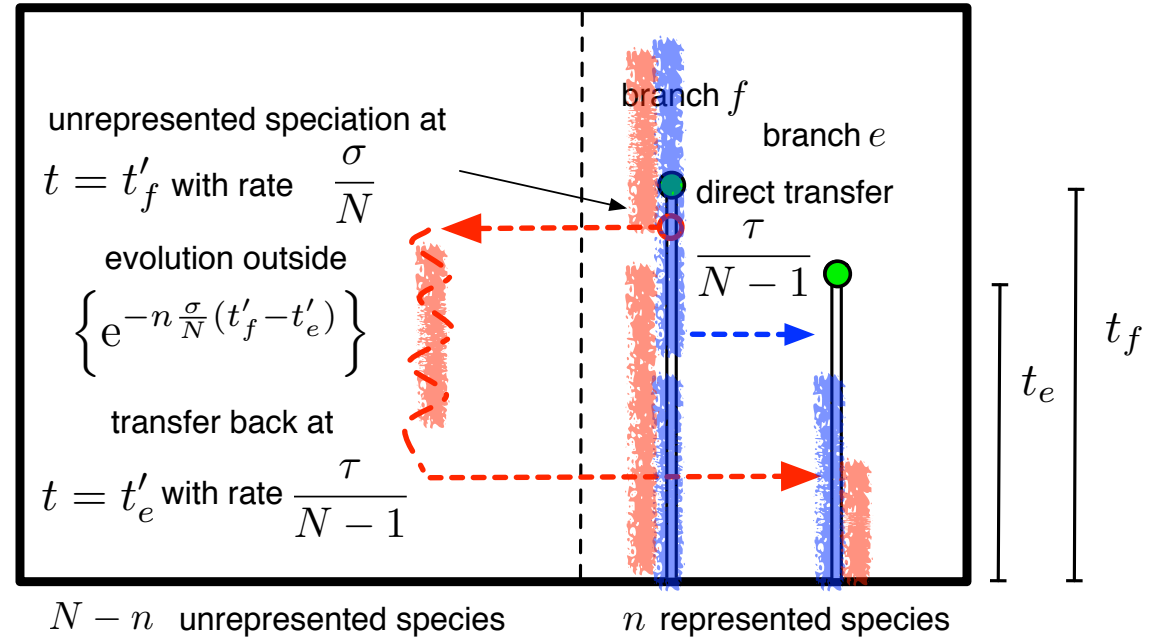
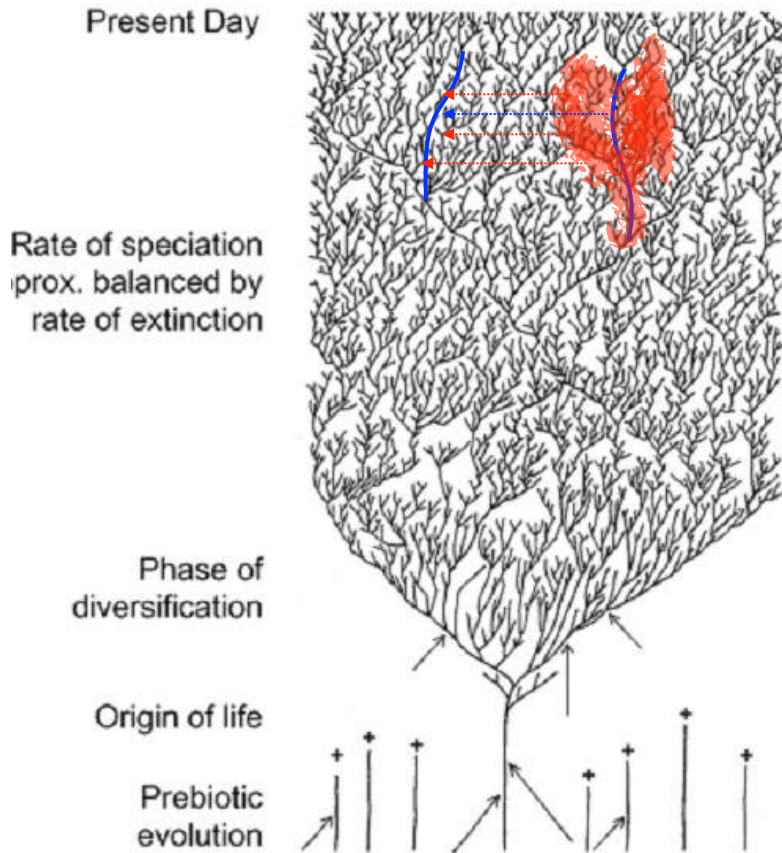


Gogarten lab

$$T_{\text{direct}} \approx \int_0^{\frac{N}{n\sigma}} \frac{\tau}{N} dt'_e = \frac{1}{N} \frac{\tau}{2n} \left[\frac{2N}{\sigma} \right],$$

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

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



Gogarten lab

$$T_{\text{direct}} \approx \int_0^{\frac{N}{n\sigma}} \frac{\tau}{N} dt'_e = \frac{1}{N} \frac{\tau}{2n} \left[\frac{2N}{\sigma} \right],$$

$$T_{\text{indirect}} \approx \int_0^{\frac{N}{n\sigma}} \int_{t'_e}^{\frac{N}{n\sigma}} \tau \left\{ e^{-n \frac{\sigma}{N} (t'_f - t'_e)} \right\} \frac{\sigma}{N} dt'_f dt'_e$$

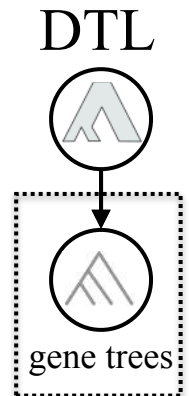
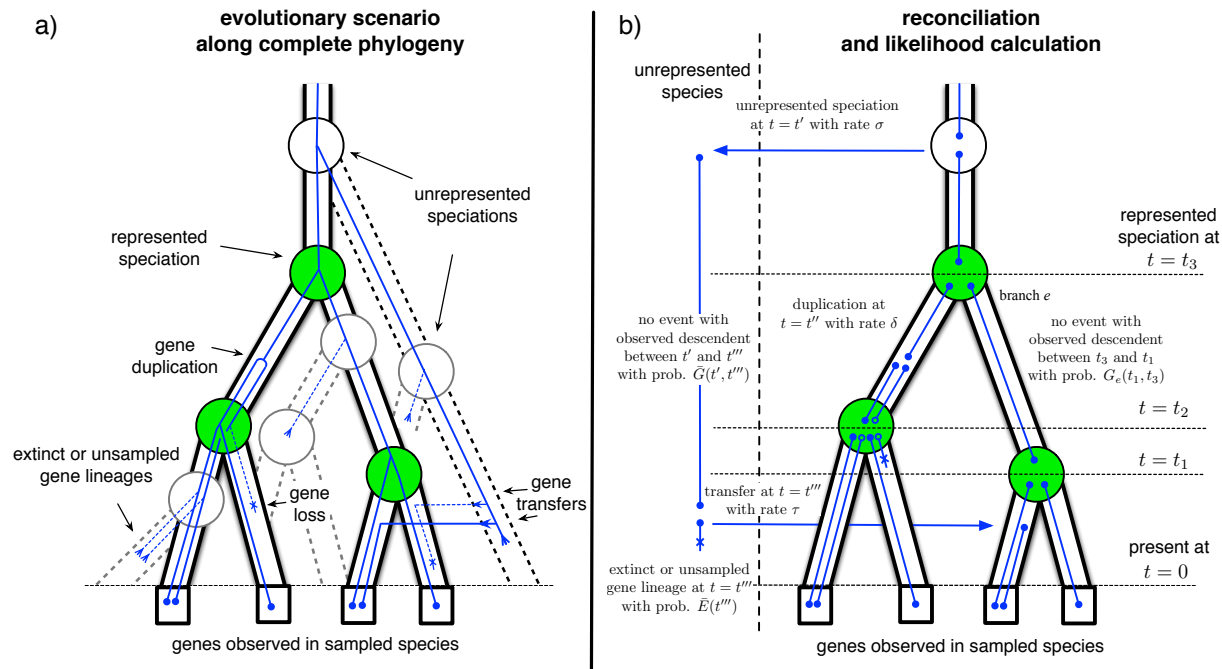
$$= \frac{1}{en} \frac{\tau}{2n} \left[\frac{2N}{\sigma} \right],$$

.. but gene trees are generated along the species tree

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

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

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



implemented in ALE:

<http://github.com/ssolo/ALE>

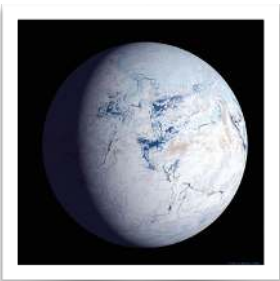
Szöllősi, Tannier, Lartillot & Daubin *Systematic Biology* (2013)
Lateral Gene Transfer from the Dead

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

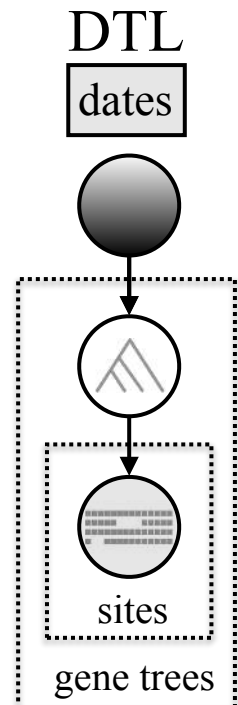
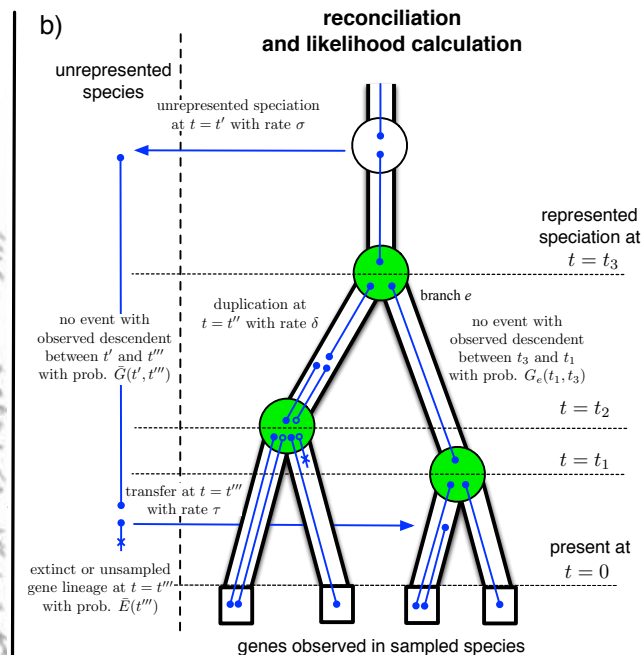
.. but gene trees are generated along the species tree

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.**

Can we detect mass extinction events?



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



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

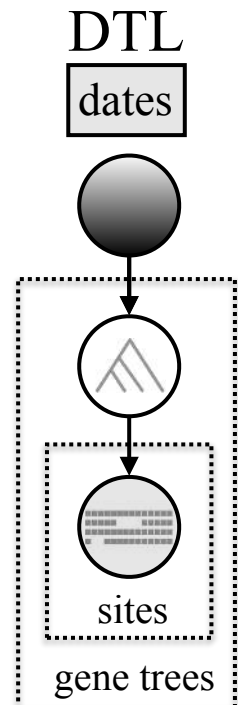
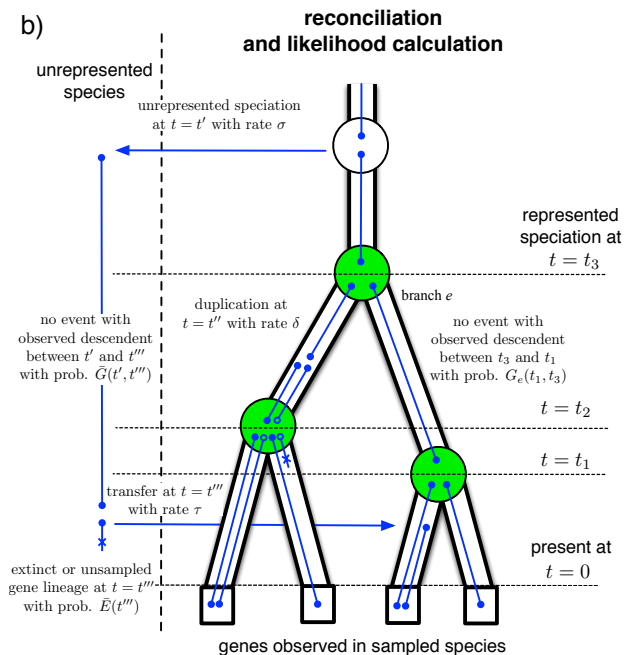
.. but gene trees are generated along the species tree

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.**

Can we detect mass extinction events?



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



implemented in ALE:

<http://github.com/ssolo/ALE>

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

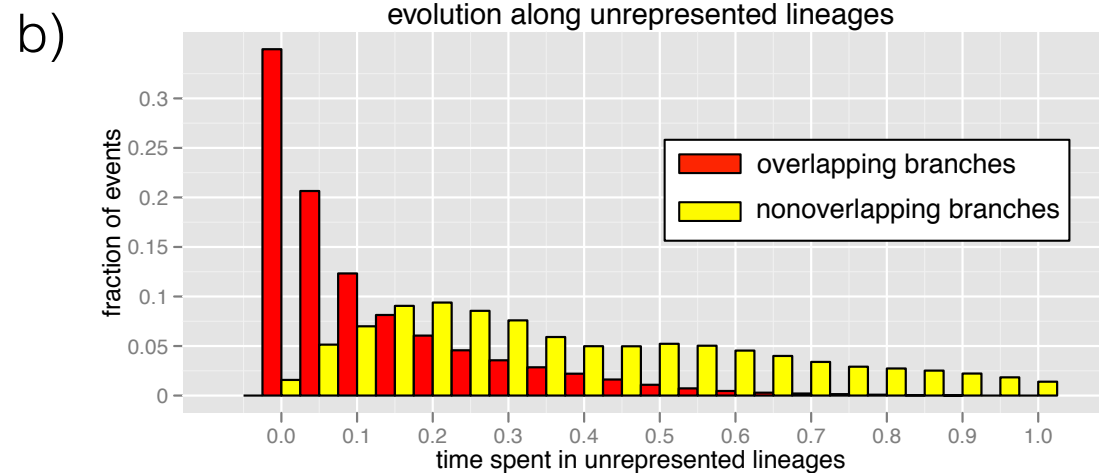
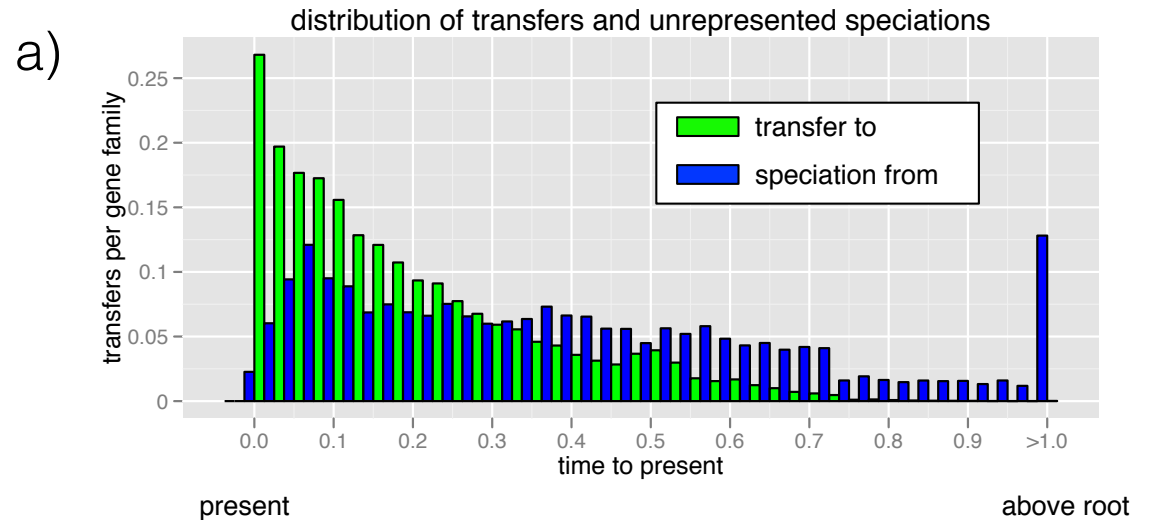
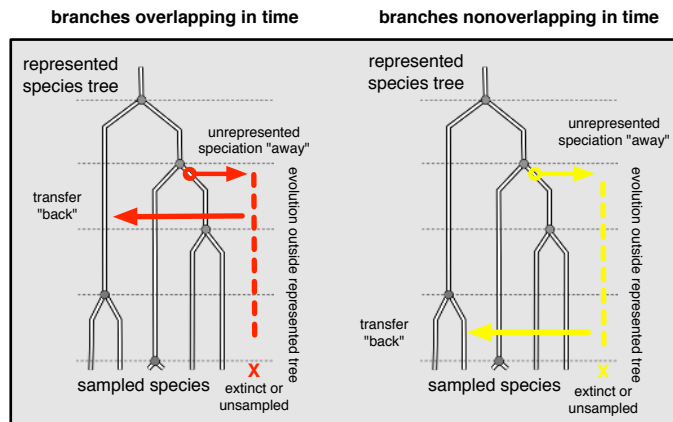
Szöllősi, Rosikiewicz, Boussau, Tannier & Daubin *Systematic Biology* (2013)
Efficient exploration of the space of reconciled gene trees

Routes to cyanobacterial genomes

474 single copy families from 36 cyanobacteria

28% of Ts between non-overlapping branches

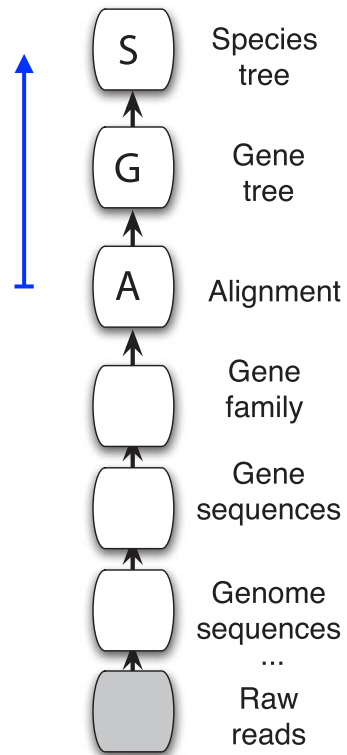
6% from above the root of Cyanobacteria



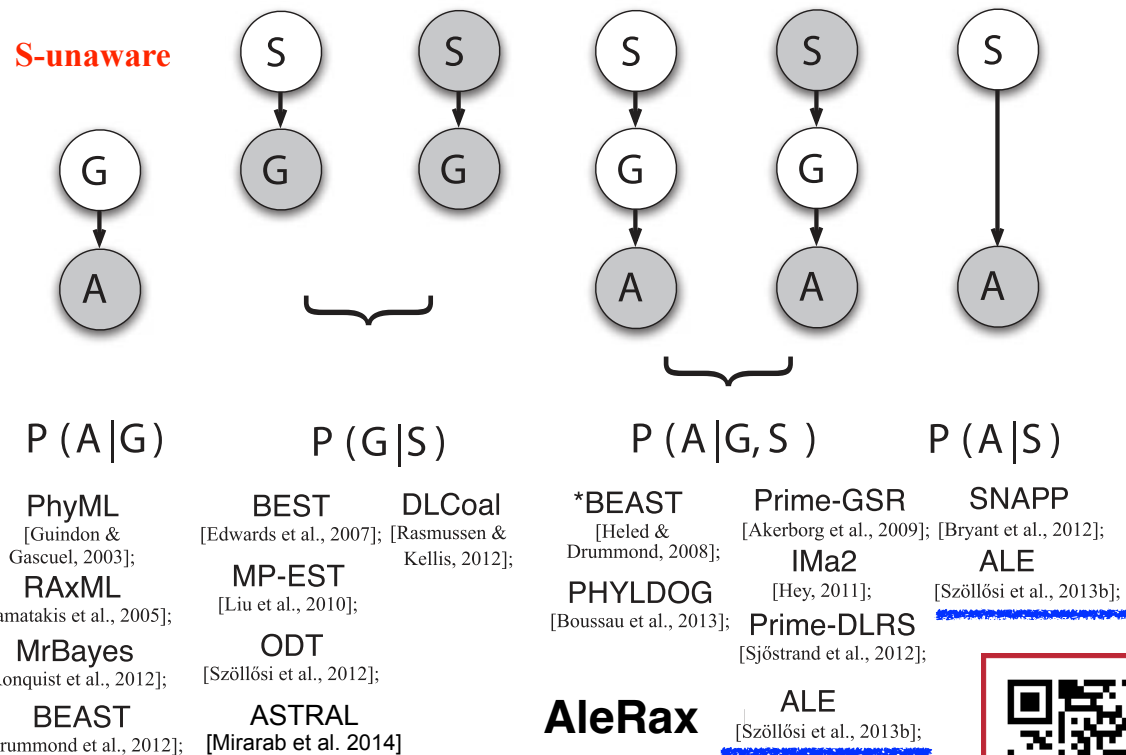
Species tree-awareness

S-aware

Phylogenomics inference pipeline



Gene tree-species tree models published in the literature



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



10.1101/2023.10.06.56109

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

Efficiently integrating over the space of reconciled gene trees

AleRax

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

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

Benoit Morel,^{1,2} Tom A. Williams,³ Alexandros Stamatakis,^{4,1,2} and Gergely J. Szöllősi,^{5,6,7}

¹Computational Molecular Evolution group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

²Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany

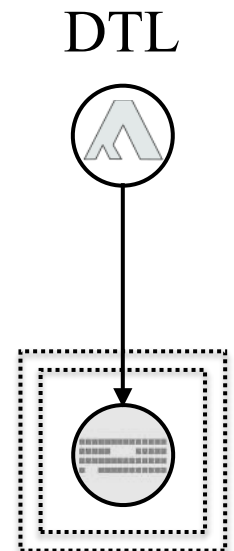
³School of Biological Sciences, University of Bristol, Bristol, UK

⁴Biodiversity Computing Group, Institute of Computer Science, Foundation for Research and Technology - Hellas

⁵ELTE-MTA "Lendület" Evolutionary Genomics Research Group, Pázmány P. stny. 1A., H-1117 Budapest, Hungary

⁶Institute of Evolution, Centre for Ecological Research, Konkoly-Thege M. út 29-33. H-1121 Budapest, Hungary

⁷Model-Based Evolutionary Genomics Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan.



HGT as a postdoc!



Join us for a postdoc at the interface of computational & evolutionary biology! Use probabilistic models & machine learning to model coevolution, reconstruct the Tree of Life, understand somatic evolution or pursue your own project..

Model-Based Evolutionary Genomics Unit
モデルベース進化ゲノミクスユニット
<https://www.oist.jp/research/research-units/modevolgenom>

Okinawa Institute of Science and Technology

`gergely.szollosi@oist.jp`