Gene tree-species tree methods in RevBayes

Bastien Boussau

LBBE, CNRS, Université de Lyon

What I spend my time doing

- Developing methods for:
  - molecular evolution
  - phylogenetic reconstruction
  - genomics, genome scale phylogenetic reconstruction
- Using these methods to investigate:
  - Genome evolution (gene family evolution, drift)
  - Evolution of life on Earth (species trees, ancestral trait reconstruction)

Why the tautological title?

"Phylogenomic" has been used to describe attempts at reconstructing species trees based on 10-100 genes

We are interested in the species tree, but also in genome evolution

A genome can contains >20,000 genes

Genome-scale: thousands of genes, dozens of species

Genomes, Genes, gene families, gene trees and species trees

Genome

Genomes, Genes, gene families, gene trees and species trees

Genome

Gene

Homo sapiens; GeneA: `ACTGGTGATGACATGAC…`

Genomes, Genes, gene families, gene trees and species trees

Genome

Gene

Homo sapiens; GeneA: ACTGGTGATGACATGAC…

Gene family alignment

Homo sapiens; GeneA: ACTGGTGATGACATAAC…
Homo sapiens; GeneB: ACTGTTGATGACATGAC…
Mus musculus; GeneC: ACTGATGATGACAAGAC…
Mus musculus; GeneD: ACTGGTGA--CCATGAC…
Bison bison; GeneE: ACTGGTGATGACACGAC…
Canis lupus; GeneF: ACT--TCATGAAACGAC…

Genomes, Genes, gene families, gene trees and species trees

Genome

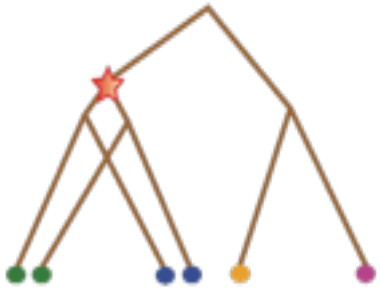
Gene

Homo sapiens; GeneA: ACTGGTGATGACATGAC…

Gene family alignment

Homo sapiens; GeneA: ACTGGTGATGACATAAC…
Homo sapiens; GeneB: ACTGTTGATGACATGAC…
Mus musculus; GeneC: ACTGATGATGACAAGAC…
Mus musculus; GeneD: ACTGGTGA--CCATGAC…
Bison bison; GeneE: ACTGGTGATGACACGAC…
Canis lupus; GeneF: ACT--TCATGAAACGAC…

Gene tree



Genomes, Genes, gene families, gene trees and species trees

Genome

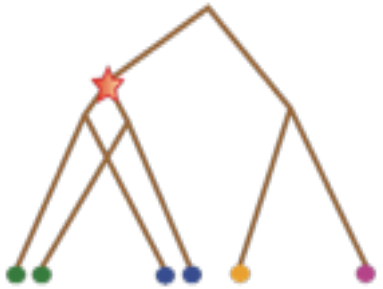
Gene

Homo sapiens; GeneA: ACTGGTGATGACATGAC…

Gene family alignment

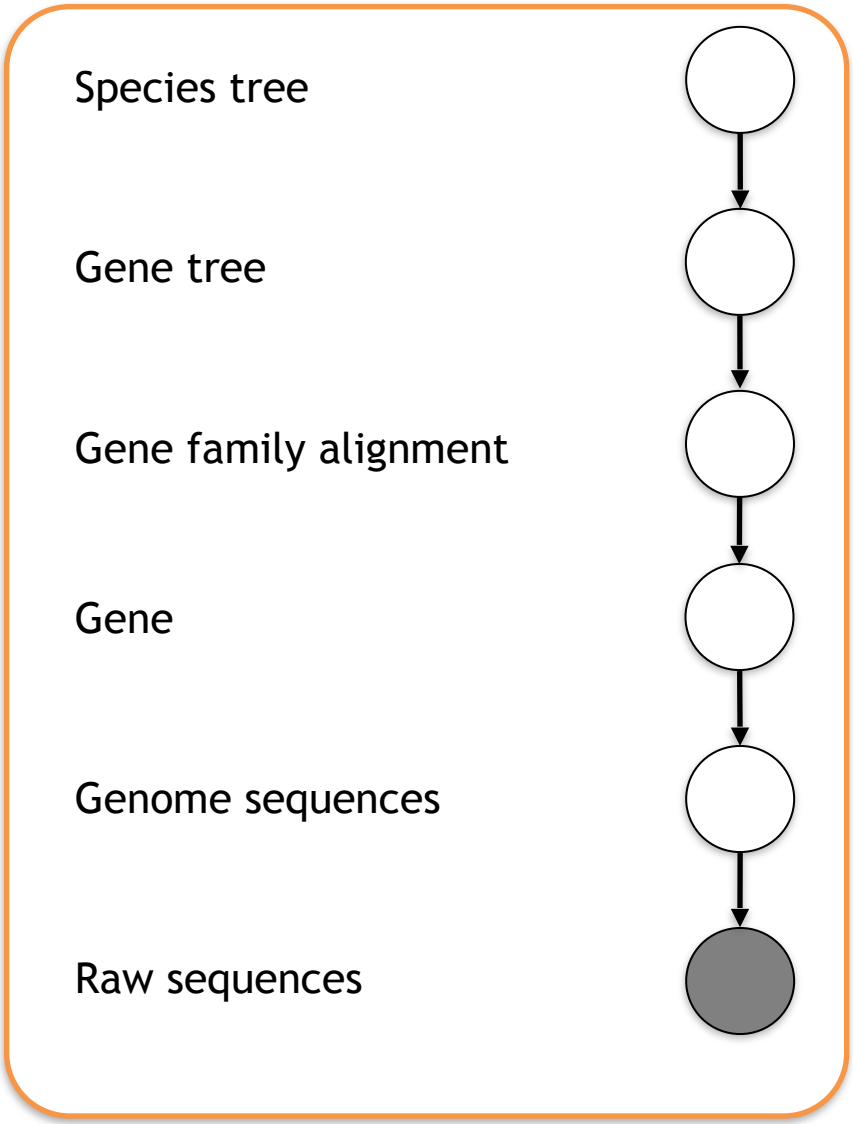
Homo sapiens; GeneA: ACTGGTGATGACATAAC…
Homo sapiens; GeneB: ACTGTTGATGACATGAC…
Mus musculus; GeneC: ACTGATGATGACAAGAC…
Mus musculus; GeneD: ACTGGTGA--CCATGAC…
Bison bison; GeneE: ACTGGTGATGACACGAC…
Canis lupus; GeneF: ACT--TCATGAAACGAC…

Gene tree

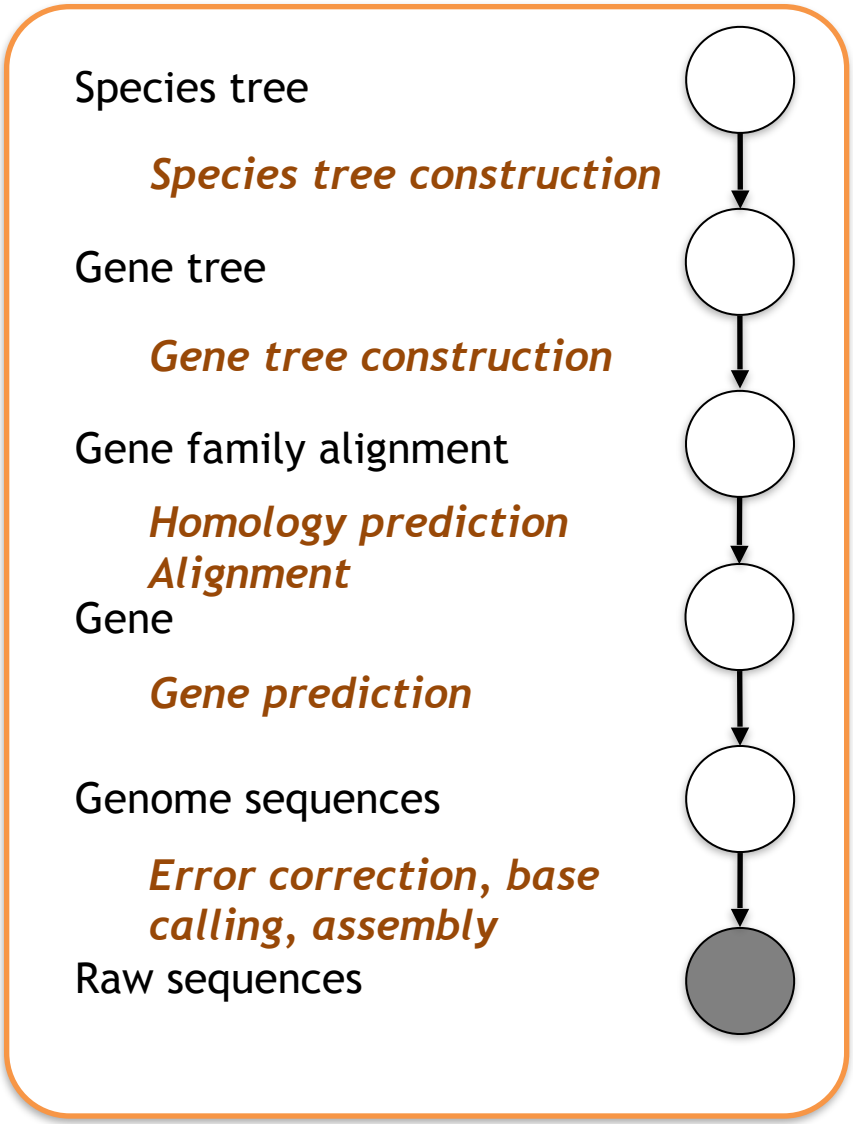


Species trees

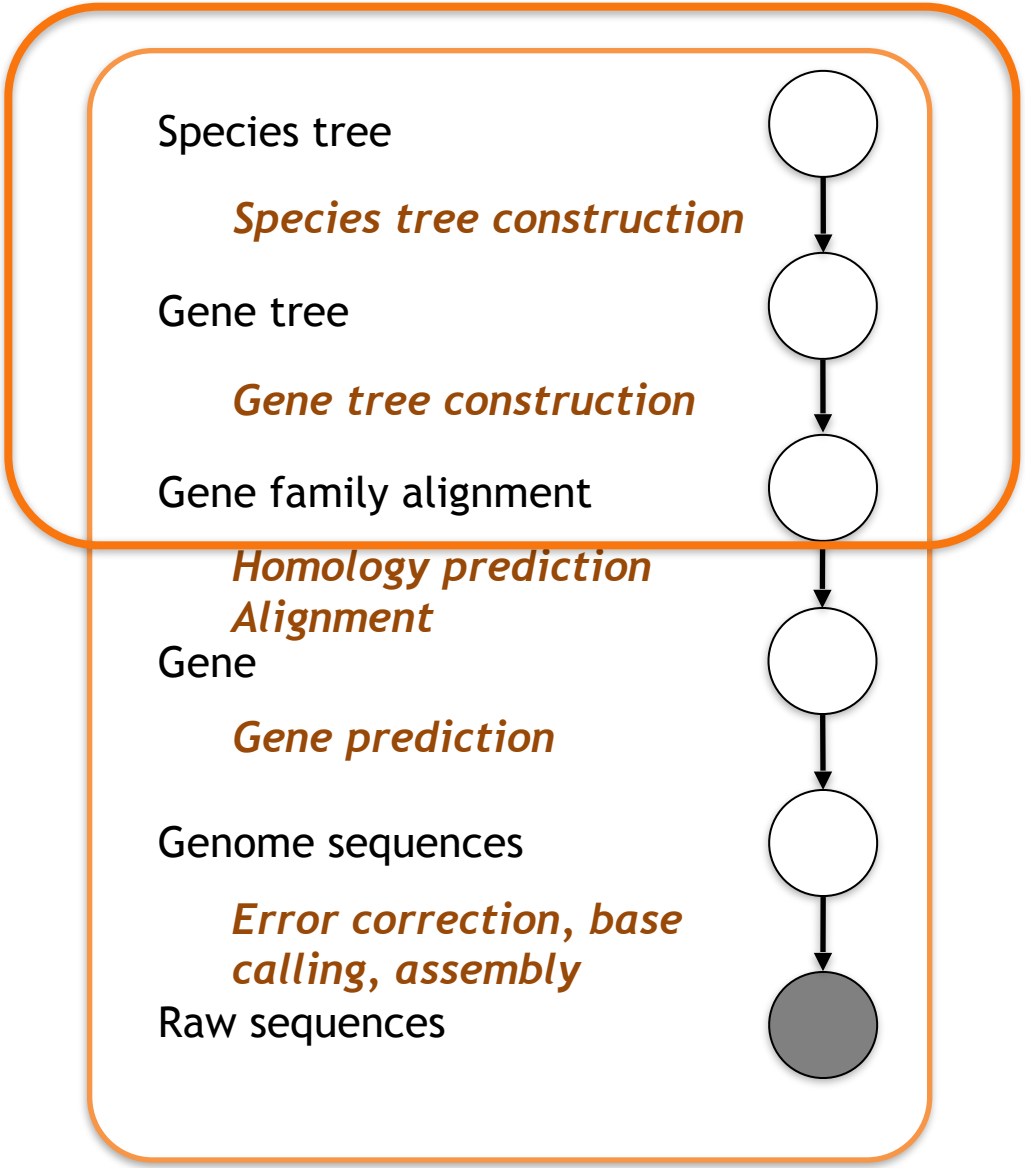
A graphical model for phylogenomics



A graphical model for phylogenomics

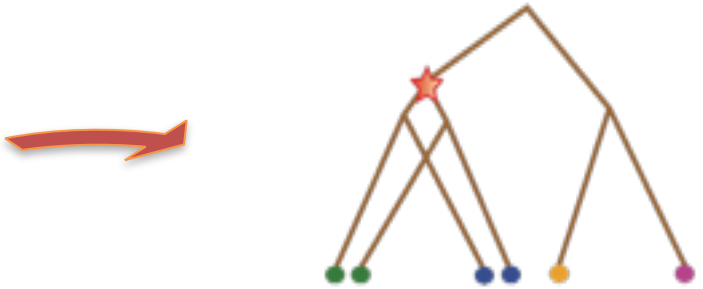


A graphical model for phylogenomics



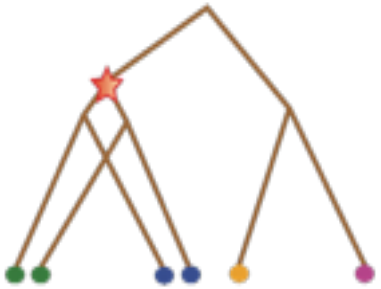
The usual approach to reconstructing phylogenetic trees

Homo sapiens; GeneA: ACTGGTGATGACATAAC…
Homo sapiens; GeneB: ACTGTTGATGACATGAC…
Mus musculus; GeneC: ACTGATGATGACAAGAC…
Mus musculus; GeneD: ACTGGTGA--CCATGAC…
Bison bison; GeneE: ACTGGTGATGACACGAC…
Canis lupus; GeneF: ACT--TCATGAAACGAC…



The usual approach to reconstructing phylogenetic trees

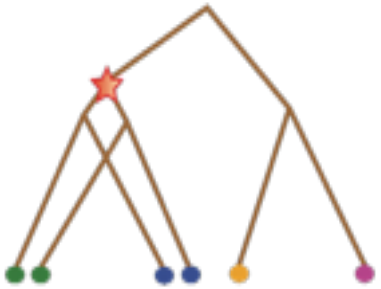
Homo sapiens; GeneA: ACTGGTGATGACATAAC…
Homo sapiens; GeneB: ACTGTTGATGACATGAC…
Mus musculus; GeneC: ACTGATGATGACAAGAC…
Mus musculus; GeneD: ACTGGTGA--CCATGAC…
Bison bison; GeneE: ACTGGTGATGACACGAC…
Canis lupus; GeneF: ACT--TCATGAAACGAC…



- Parsimony
- Model-based approaches: e.g. Felsenstein pruning algorithm (1981) to compute P(alignment|Gene tree)

The usual approach to reconstructing phylogenetic trees

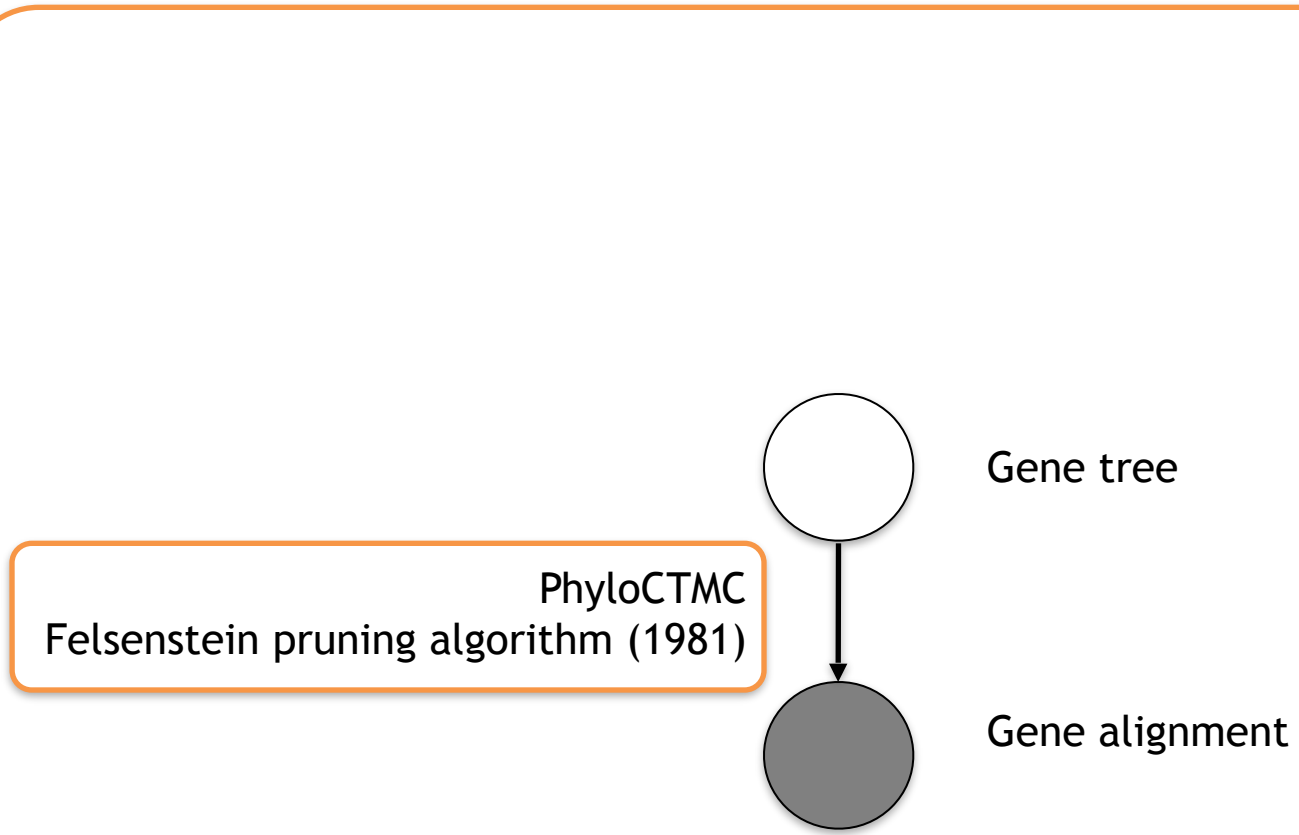
Homo sapiens; GeneA: ACTGGTGATGACATAAC…
Homo sapiens; GeneB: ACTGTTGATGACATGAC…
Mus musculus; GeneC: ACTGATGATGACAAGAC…
Mus musculus; GeneD: ACTGGTGA--CCATGAC…
Bison bison; GeneE: ACTGGTGATGACACGAC…
Canis lupus; GeneF: ACT--TCATGAAACGAC…



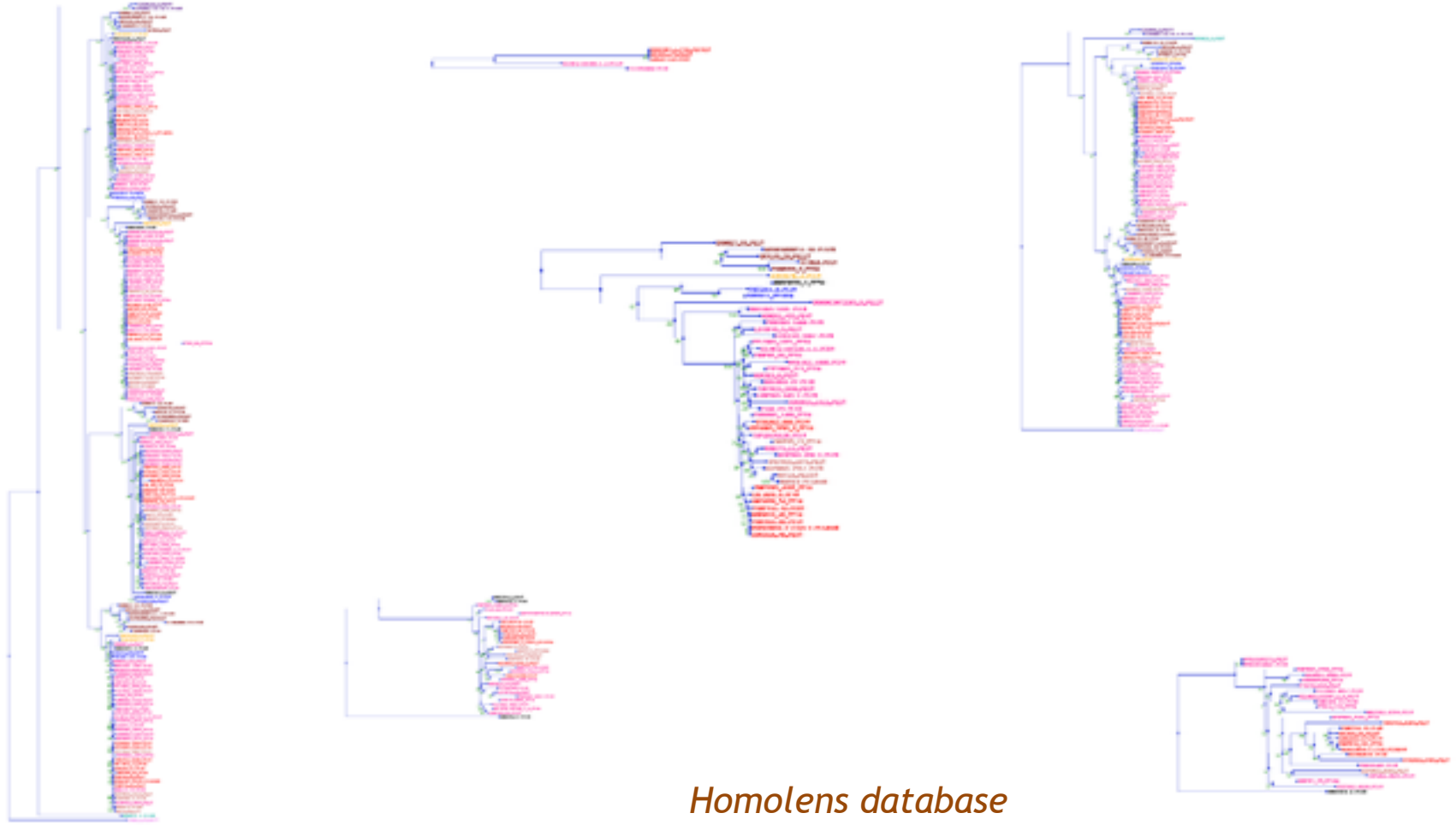
- Parsimony
- Model-based approaches: e.g. Felsenstein pruning algorithm (1981) to compute P(alignment|Gene tree)

No species tree object in the usual approach

The gene tree graphical model

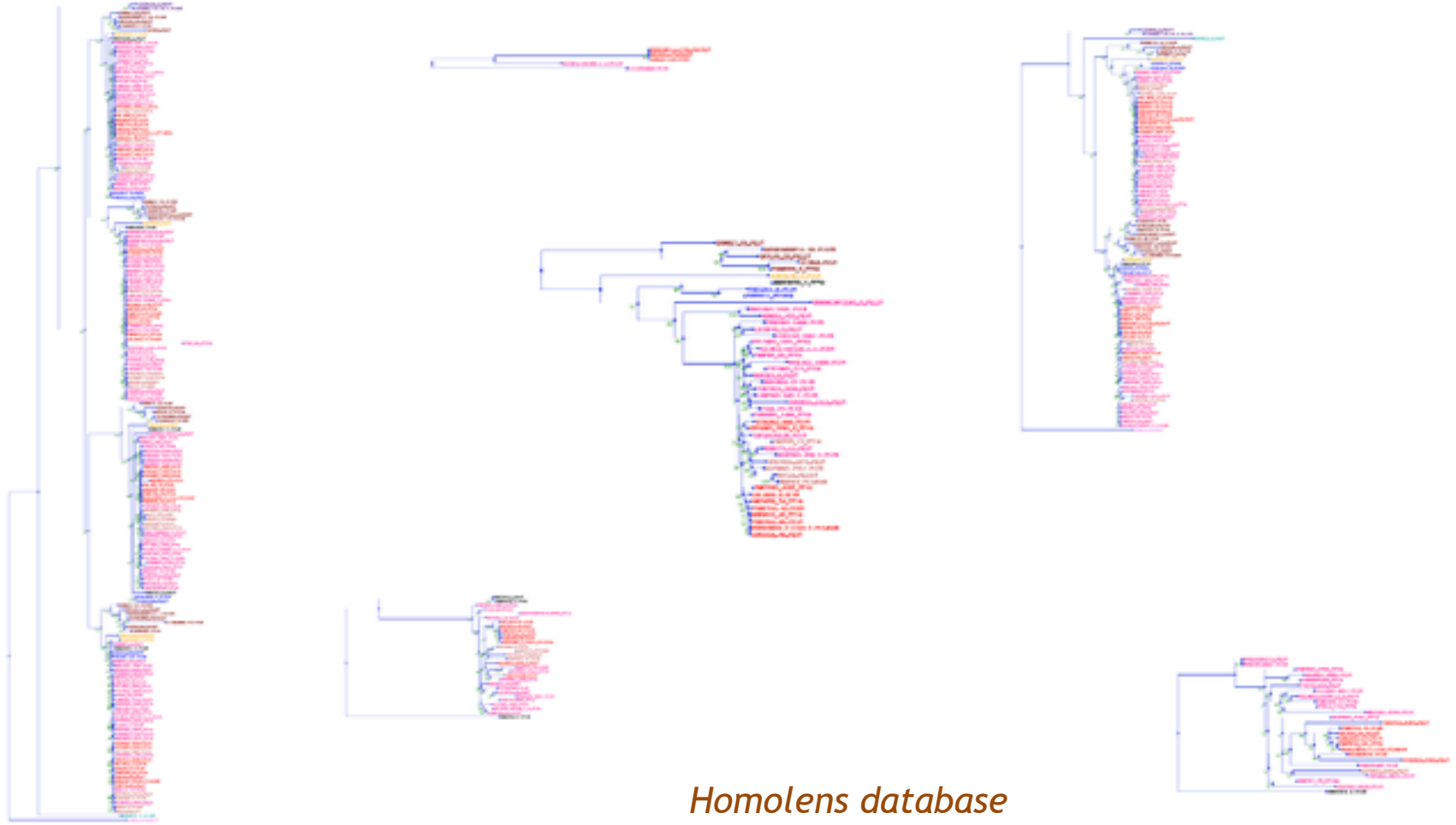


Gene trees inferred from sequences alone



Homolens database

Gene trees inferred from sequences alone



Homolens database

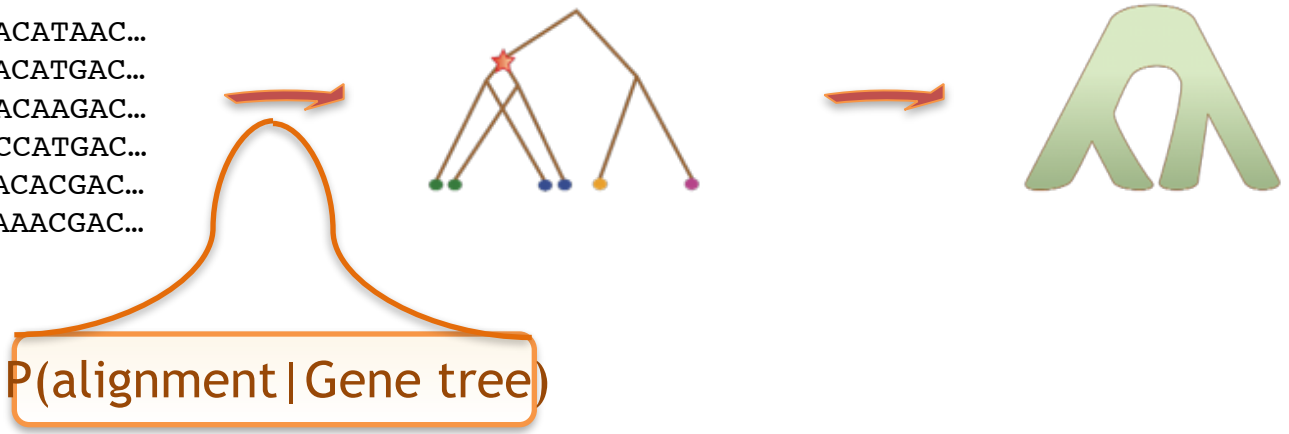
Complex, messy, not to be trusted

The species tree-gene tree approach to reconstructing phylogenetic trees

Homo sapiens GeneA: ACTGGTGATGACATAAC…
Homo sapiens GeneB: ACTGTTGATGACATGAC…
Mus musculus GeneC: ACTGATGATGACAAGAC…
Mus musculus GeneD: ACTGGTGA--CCATGAC…
Bison bison; GeneE: ACTGGTGATGACACGAC…
Canis lupus; GeneF: ACT--TCATGAAACGAC…

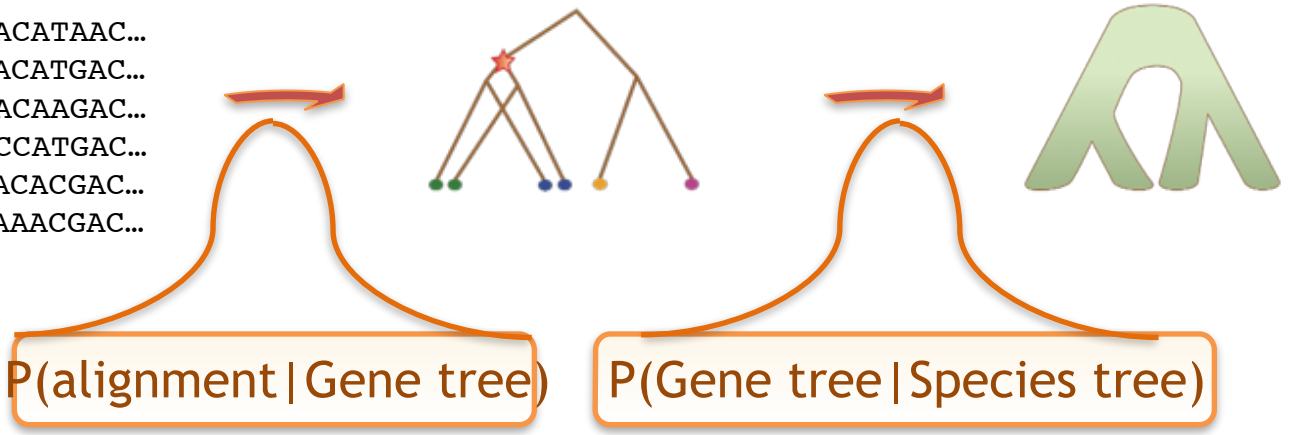
The species tree-gene tree approach to reconstructing phylogenetic trees

Homo sapiens GeneA: ACTGGTGATGACATAAC…
Homo sapiens GeneB: ACTGTTGATGACATGAC…
Mus musculus GeneC: ACTGATGATGACAAGAC…
Mus musculus GeneD: ACTGGTGA--CCATGAC…
Bison bison; GeneE: ACTGGTGATGACACGAC…
Canis lupus; GeneF: ACT--TCATGAAACGAC…



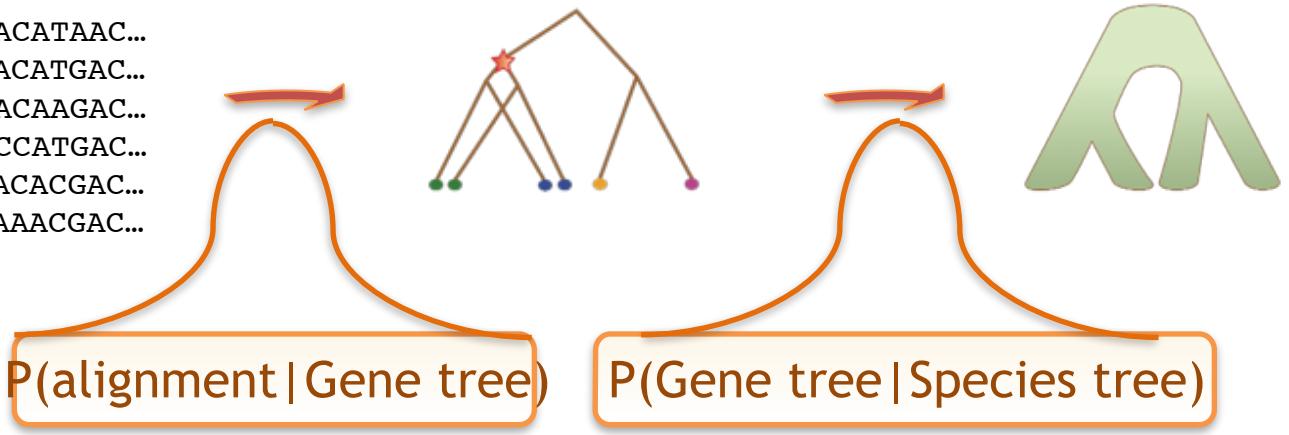
The species tree-gene tree approach to reconstructing phylogenetic trees

Homo sapiens GeneA: ACTGGTGATGACATAAC…
Homo sapiens GeneB: ACTGTTGATGACATGAC…
Mus musculus GeneC: ACTGATGATGACAAGAC…
Mus musculus GeneD: ACTGGTGA--CCATGAC…
Bison bison; GeneE: ACTGGTGATGACACGAC…
Canis lupus; GeneF: ACT--TCATGAAACGAC…



The species tree-gene tree approach to reconstructing phylogenetic trees

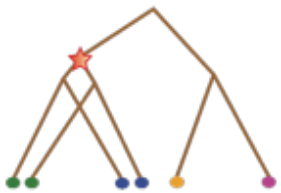
Homo sapiens GeneA: ACTGGTGATGACATAAC…
Homo sapiens GeneB: ACTGTTGATGACATGAC…
Mus musculus GeneC: ACTGATGATGACAAGAC…
Mus musculus GeneD: ACTGGTGA--CCATGAC…
Bison bison; GeneE: ACTGGTGATGACACGAC…
Canis lupus; GeneF: ACT--TCATGAAACGAC…



P(alignment,Gene tree|Species tree)=P(alignment|Gene tree)xP(Gene tree|Species tree)

The species tree-gene tree approach to reconstructing phylogenetic trees

Homo sapiens GeneA: ACTGGTGATGACATAAC…
Homo sapiens GeneB: ACTGTTGATGACATGAC…
Mus musculus GeneC: ACTGATGATGACAAGAC…
Mus musculus GeneD: ACTGGTGA--CCATGAC…
Bison bison; GeneE: ACTGGTGATGACACGAC…
Canis lupus; GeneF: ACT--TCATGAAACGAC…



P(alignment|Gene tree)

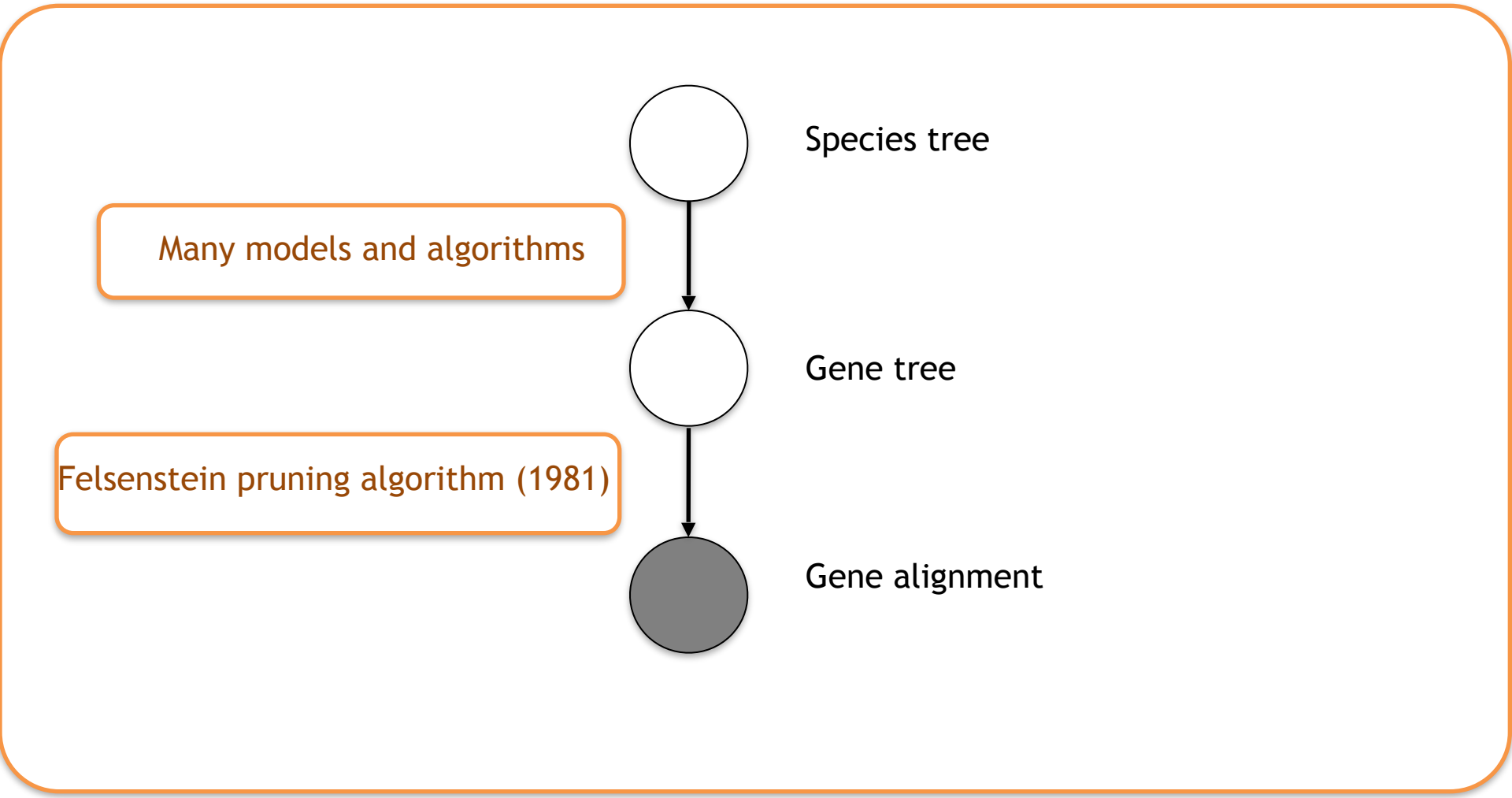
P(Gene tree|Species tree)

Felsenstein pruning algorithm (1981)

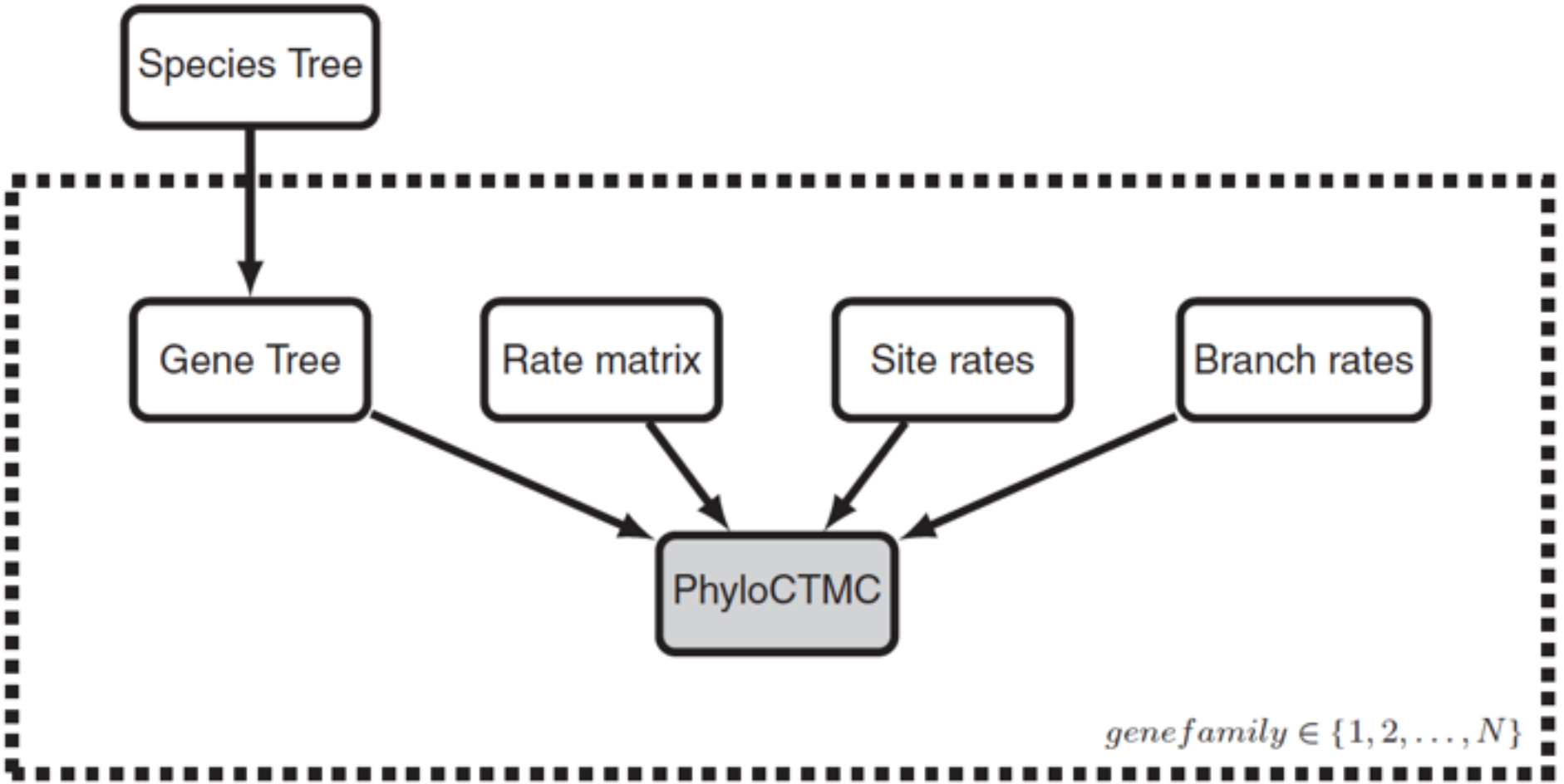
Many models and algorithms

P(alignment,Gene tree|Species tree)=P(alignment|Gene tree)xP(Gene tree|Species tree)

The species tree-gene tree graphical model

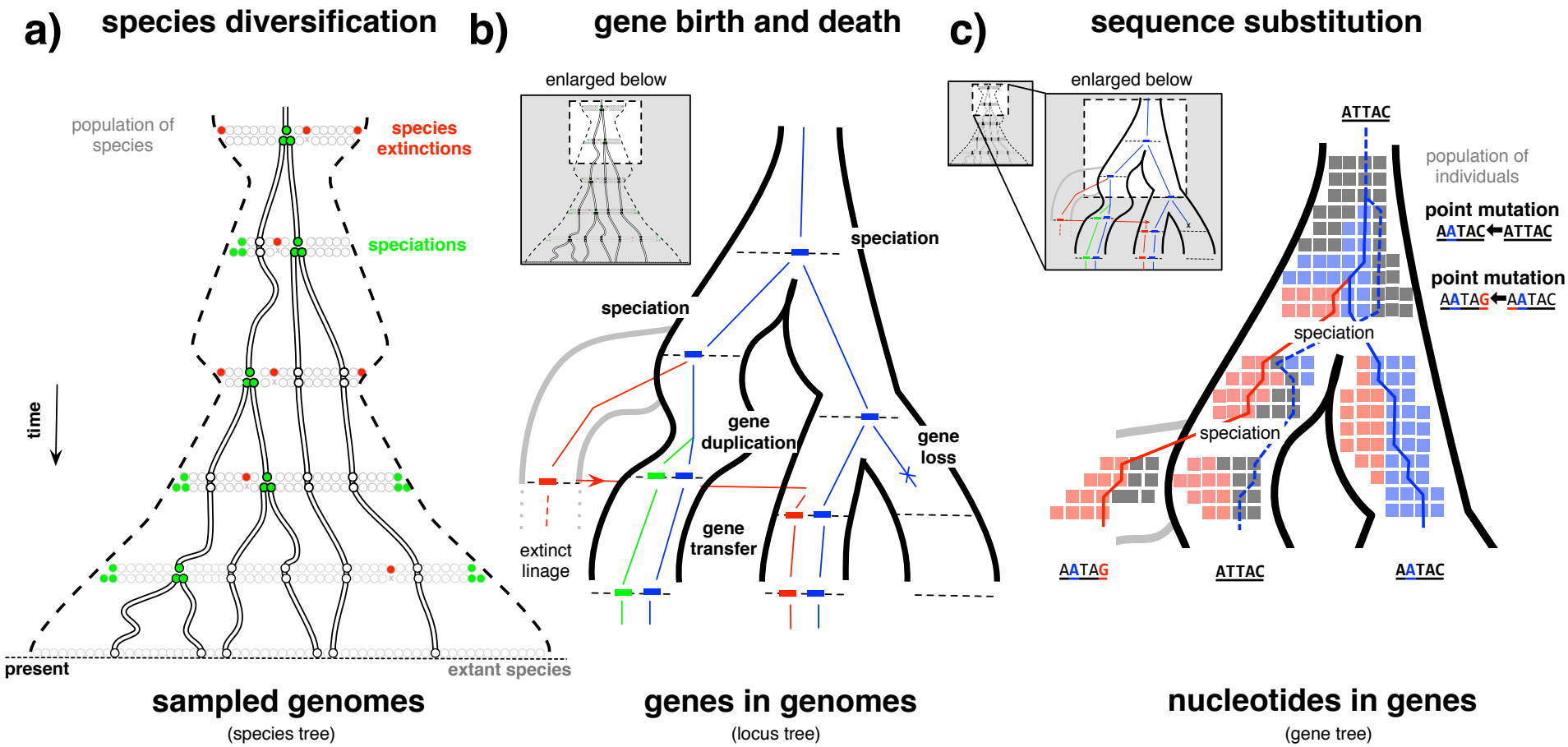


The species tree-gene tree graphical model



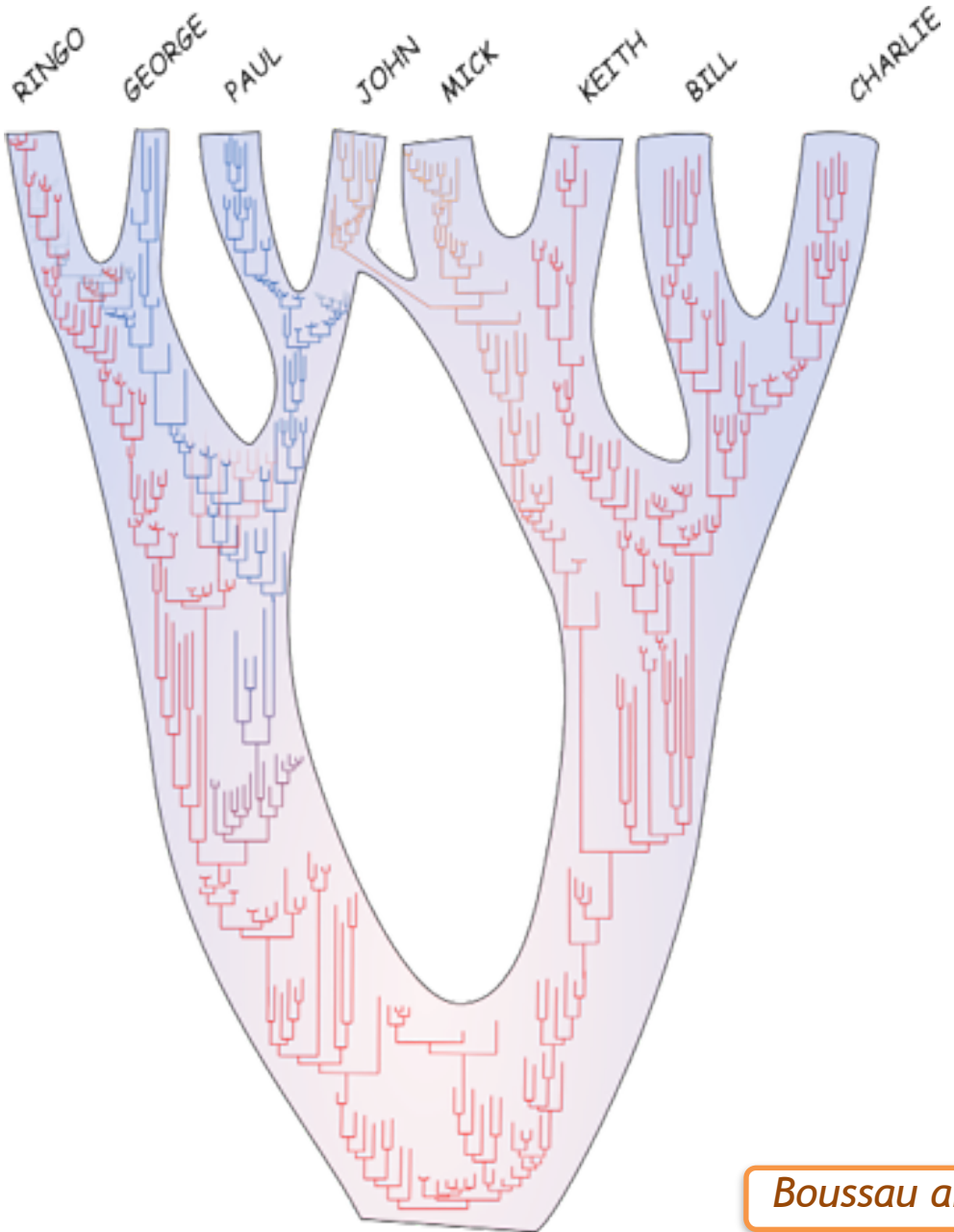
Hoehna *et al., Syst Biol 2014*

A hierarchy of processes



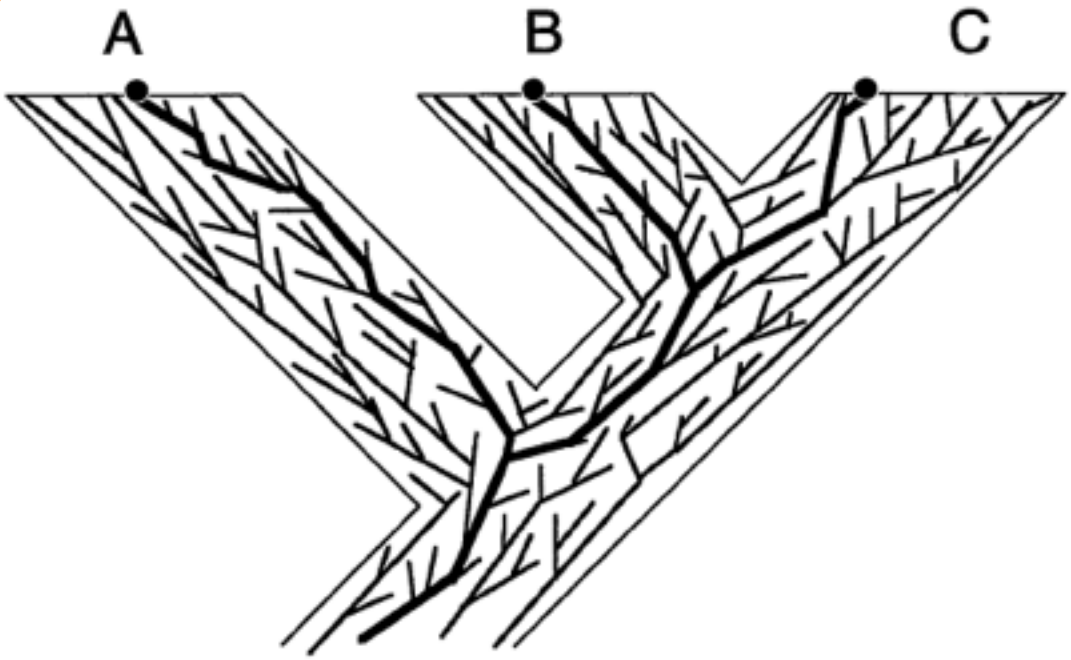
Szöllősi *et al., Syst Biol 2014*

Gene tree and species tree are linked



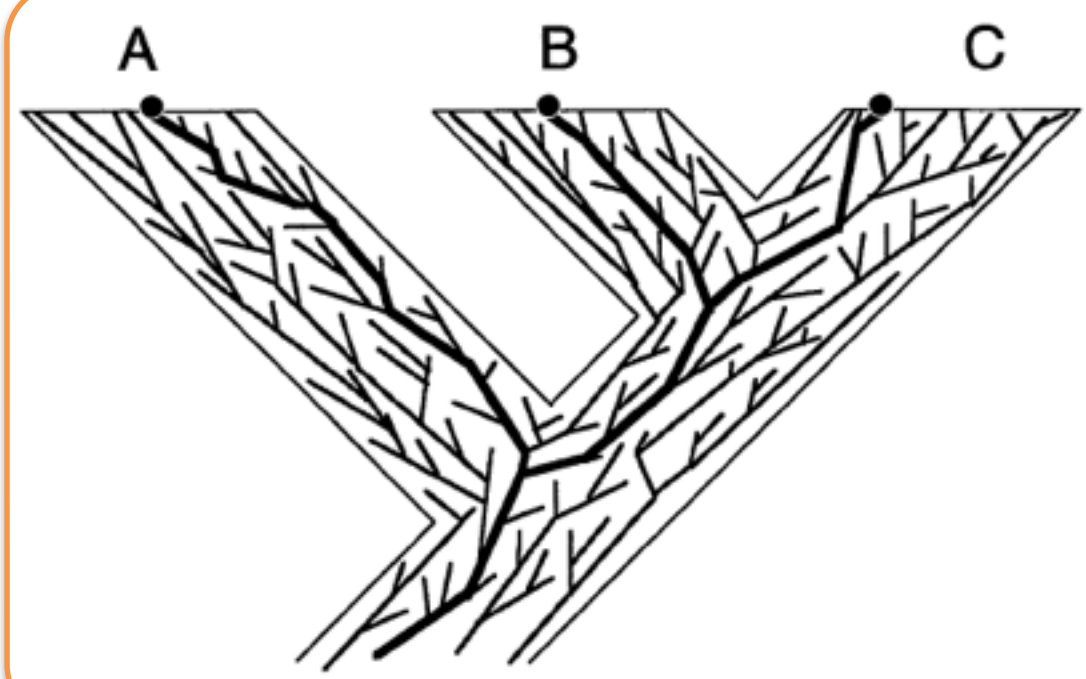
Boussau and Daubin, Tree 2010

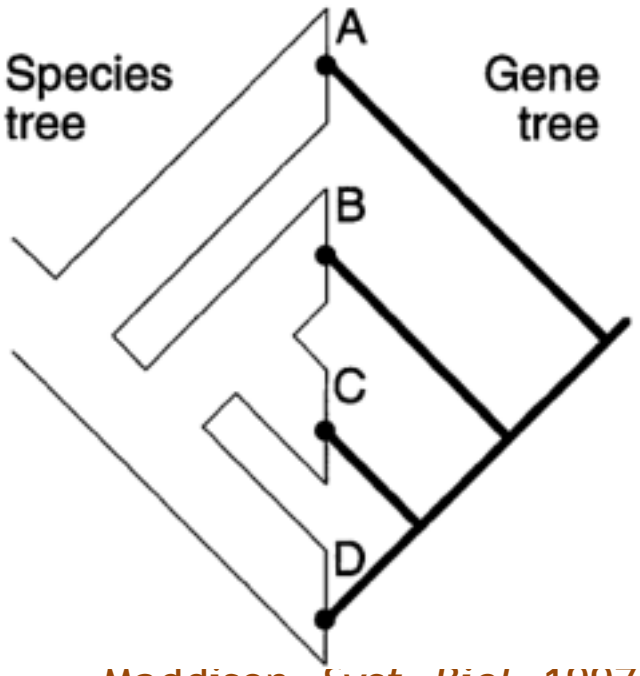
Gene trees in species trees



Maddison, *Syst. Biol.* 1997

Gene trees in species trees





Maddison, *Syst. Biol.* 1997

Challenges

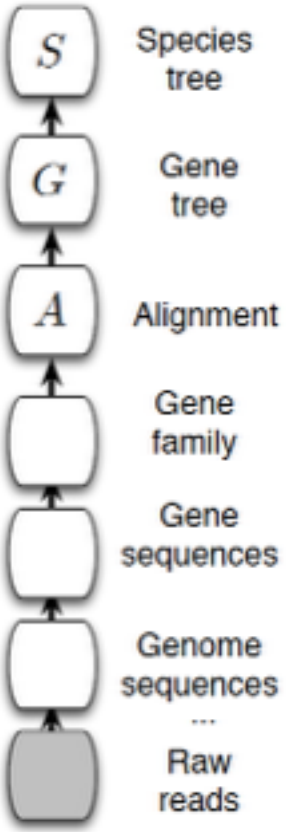
Given gene trees: reconstruct species tree

Given species tree and alignment: reconstruct gene tree

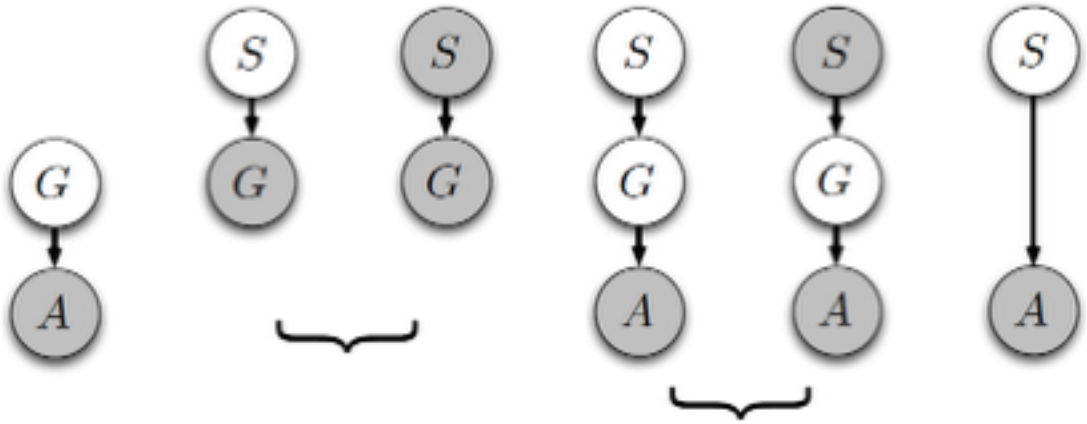
Given alignments: reconstruct genes and species trees

The landscape of gene tree-species tree methods

Phylogenomics inference pipeline



Gene tree-species tree models published in the literature



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

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

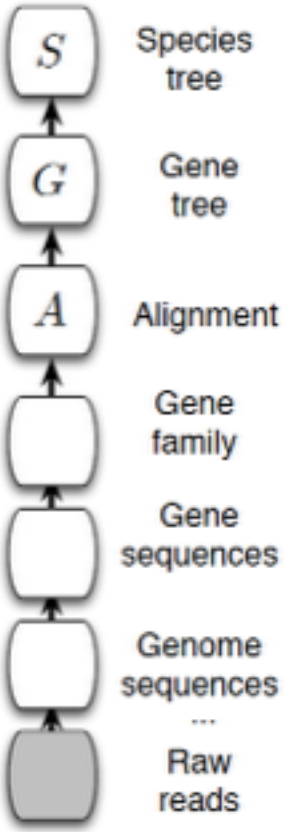
*BEAST [Heled & Drummond, 2008];
PHYLDOG [Boussau et al., 2013];
Prime-GSR [Akerborg et al., 2009];
IMa2 [Hey, 2011];
Prime-DLRS [Sjostrand et al., 2012];
exODT [Szöllősi et al., 2013a].

SNAP [Bryant et al., 2012];
ALE [Szöllősi et al., 2013b];
POMO (De Maio et al., 2013)

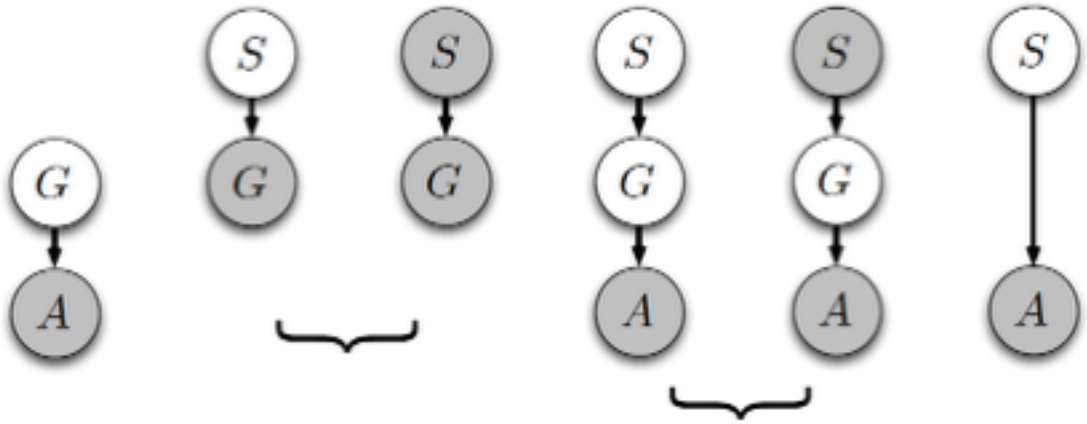
Szöllősi *et al., Syst Biol 2014*

The landscape of gene tree-species tree methods

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];

$P(A|G,S)$

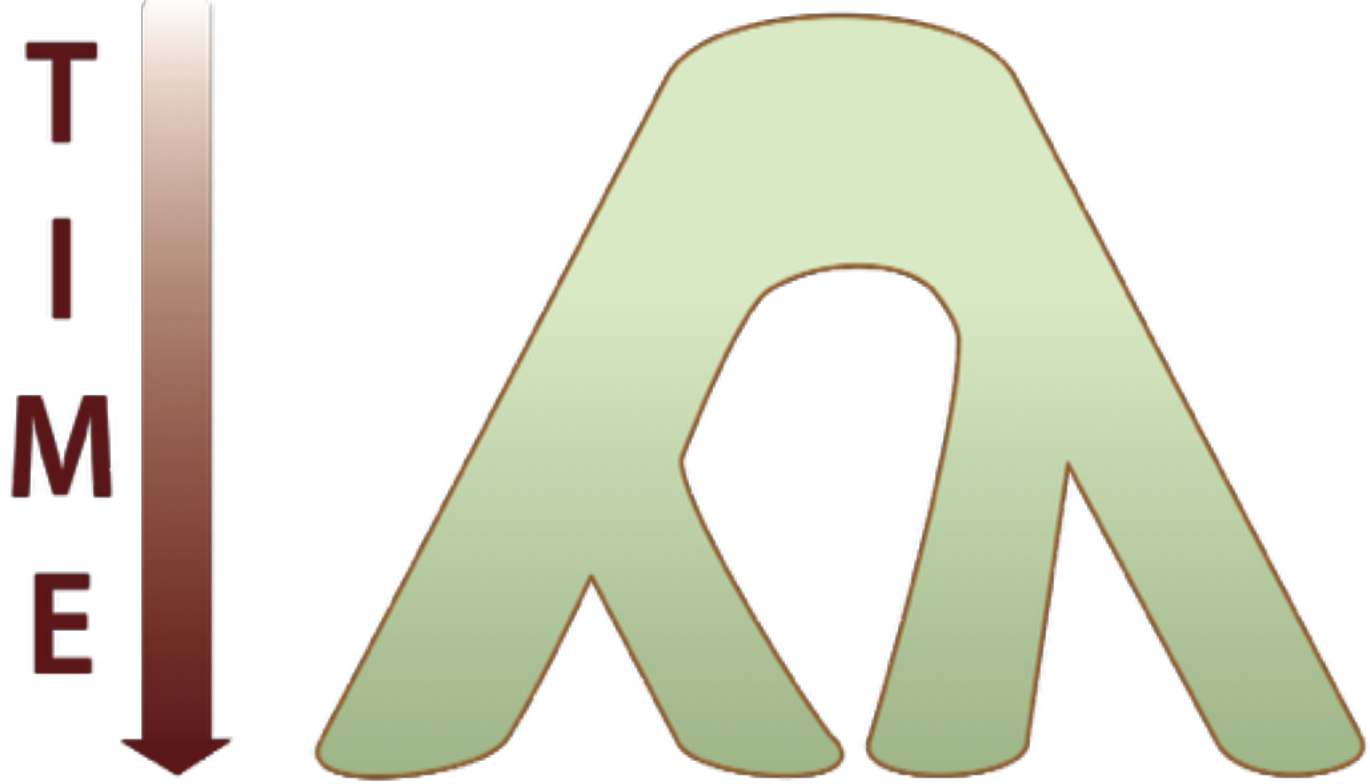
*BEAST [Heled & Drummond, 2008];
PHYLDOG [Boussau et al., 2013];
Prime-GSR [Akerborg et al., 2009];
IMa2 [Hey, 2011];
Prime-DLRS [Sjostrand et al., 2012];
exODT [Szöllősi et al., 2013a].

$P(A|S)$

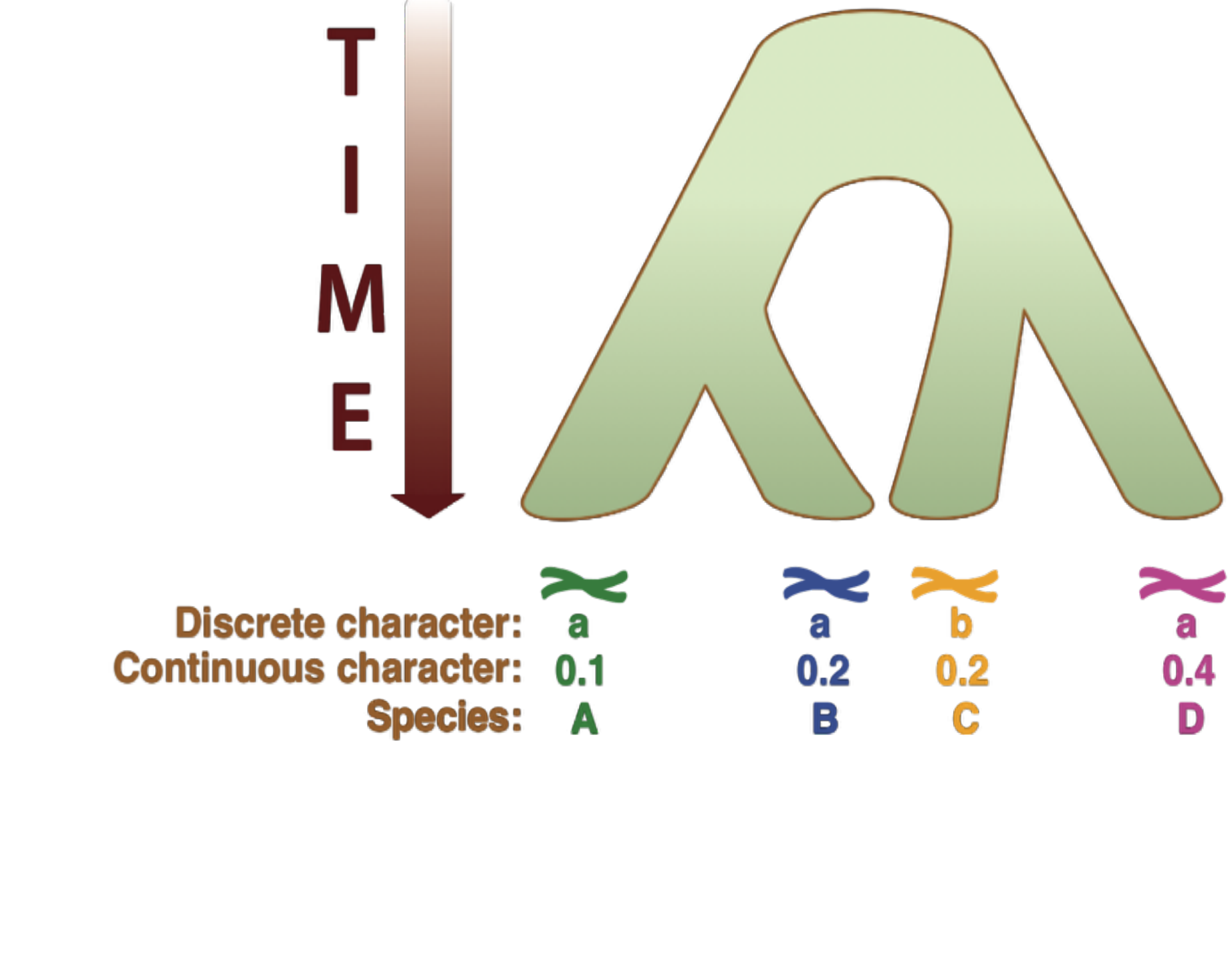
SNAP [Bryant et al., 2012];
ALE [Szöllősi et al., 2013b];
POMO (De Maio et al., 2013)

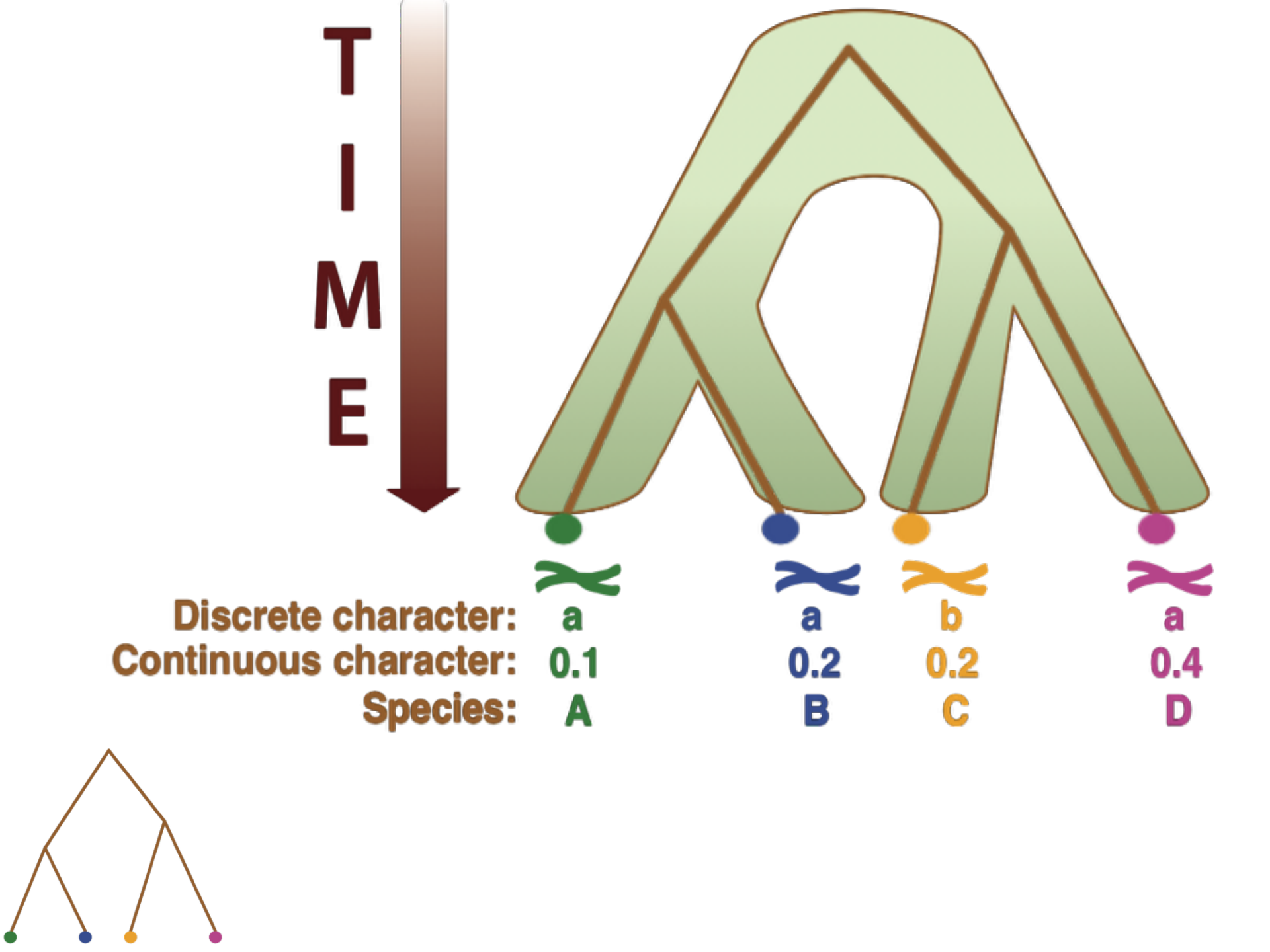
RevBayes

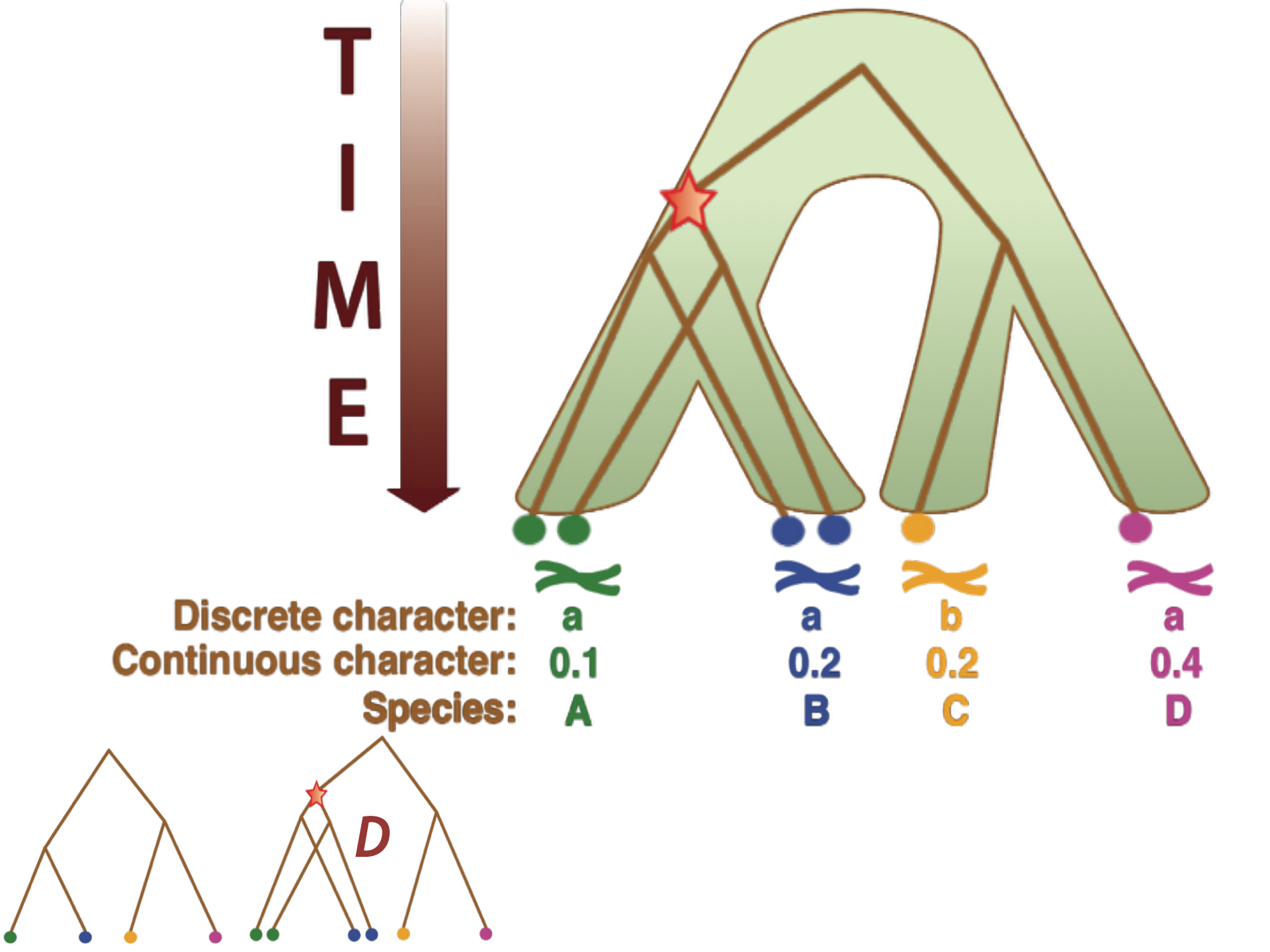
Szöllősi *et al., Syst Biol 2014*

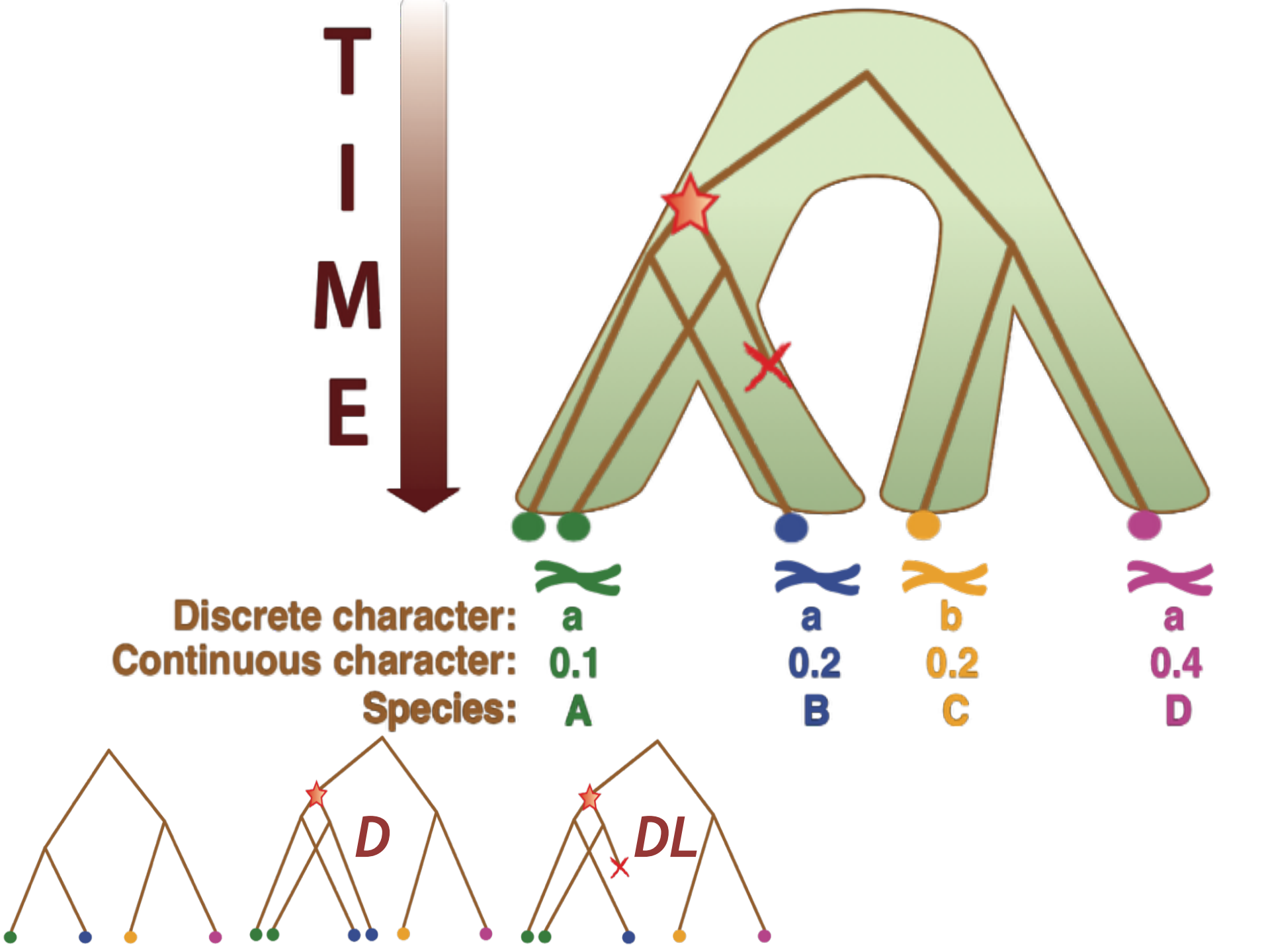


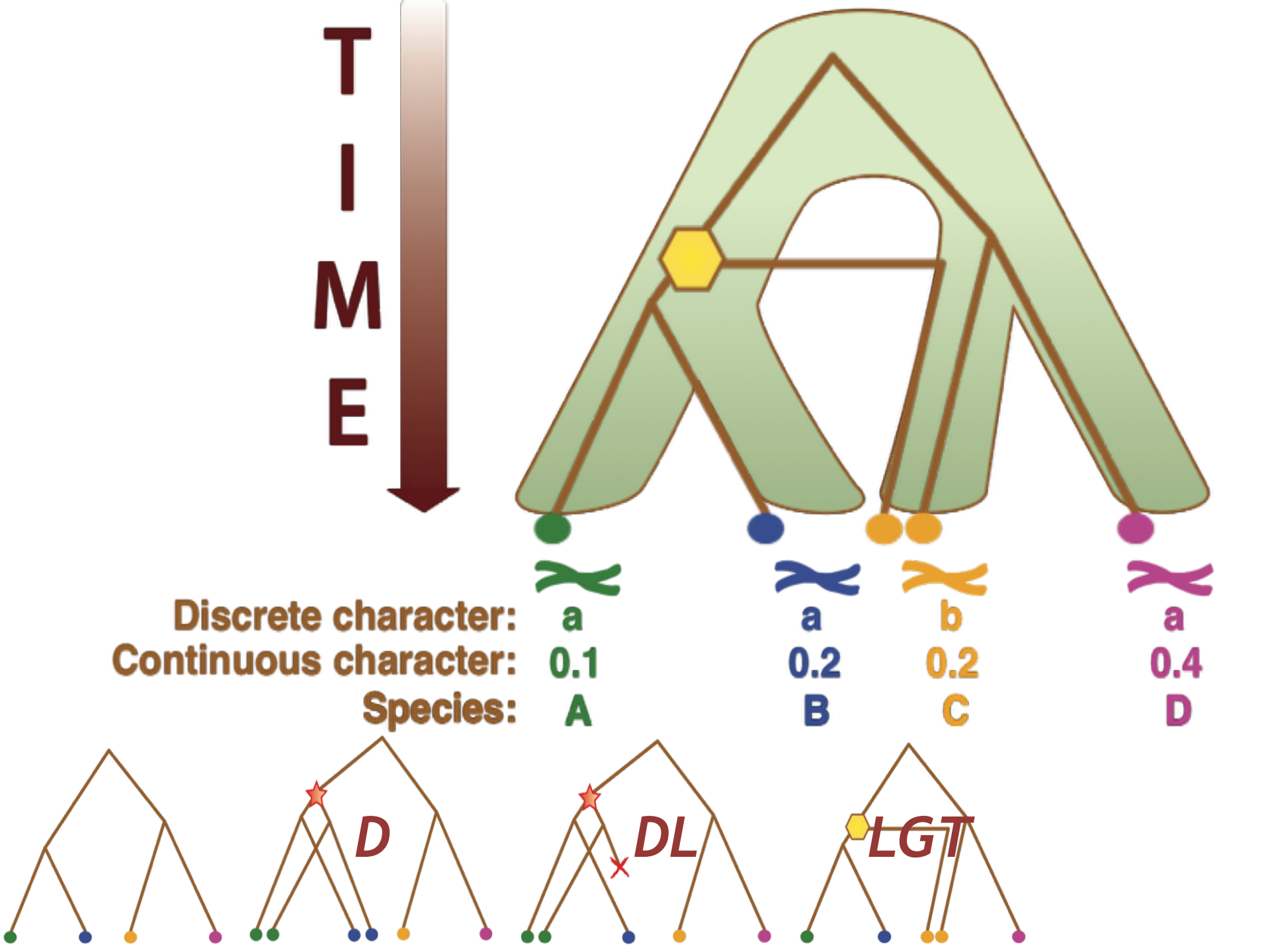
Discrete character:	a	a	b	a
Continuous character:	0.1	0.2	0.2	0.4
Species:	A	B	C	D

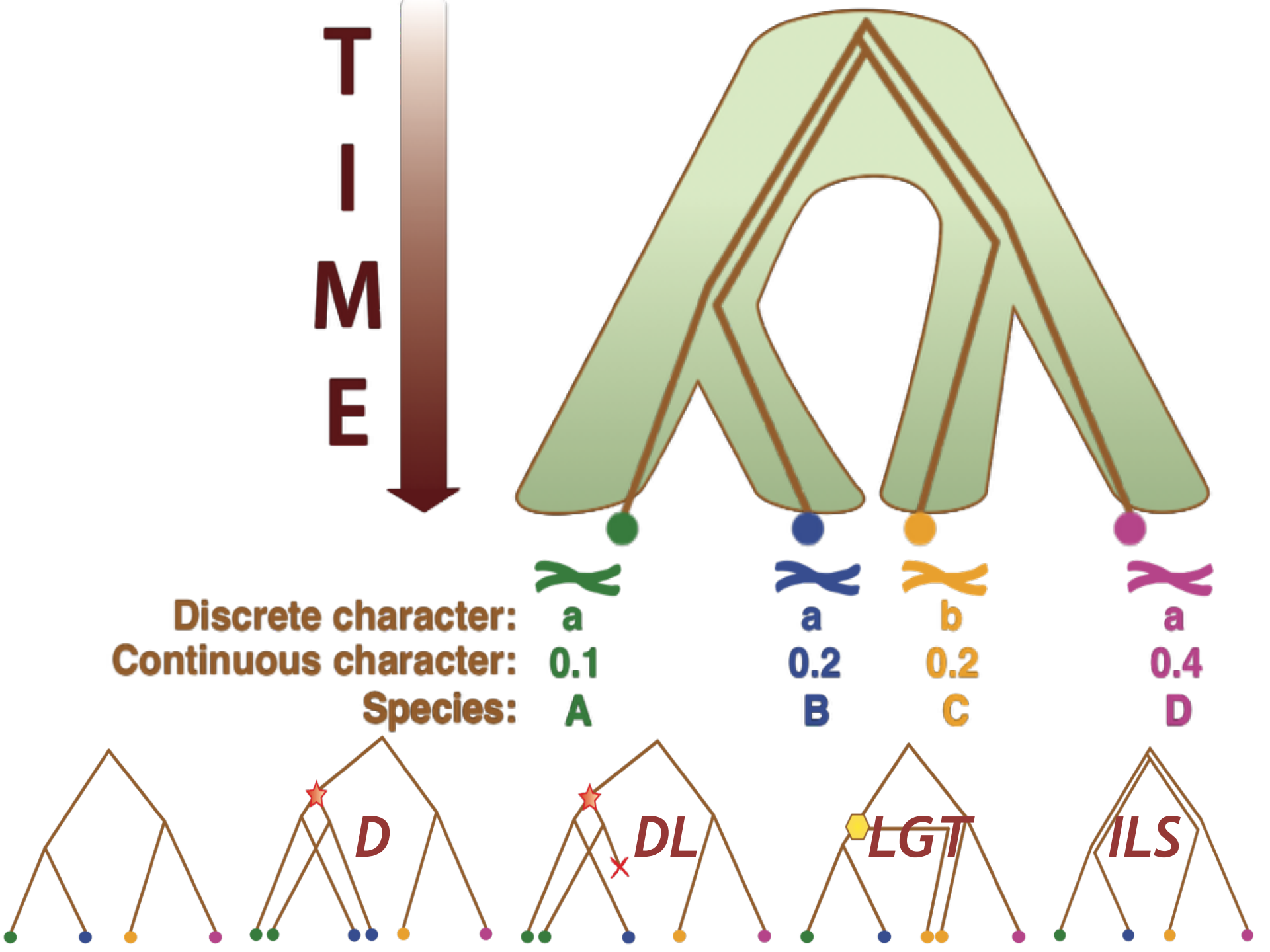

TIME
Discrete character:
Continuous character:
Species:
a
0.1
A
a
0.2
B
b
0.2
C
a
0.4
D

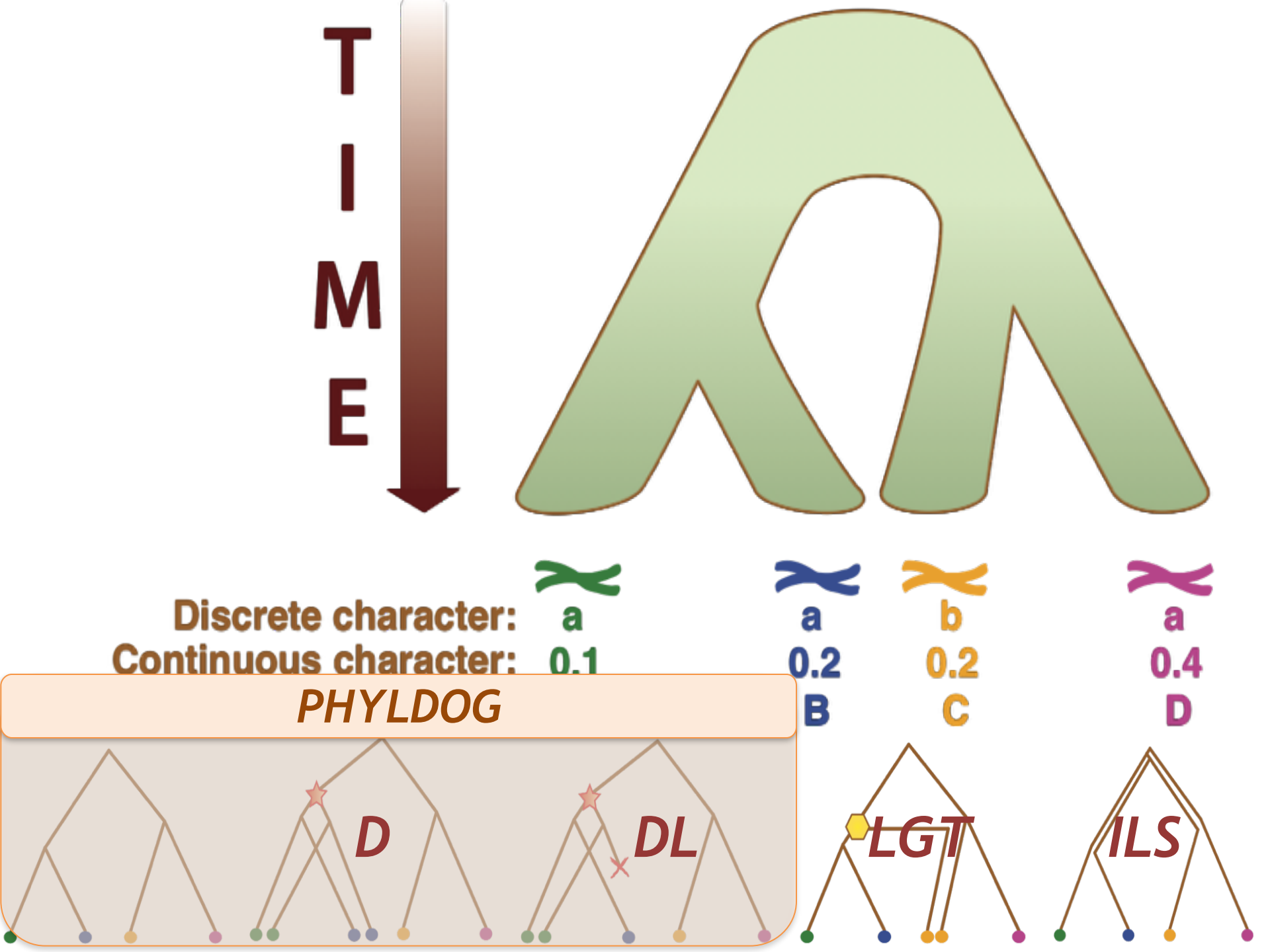

TIME
Discrete character:
Continuous character:
Species:
a
0.1
A
a
0.2
B
b
0.2
C
a
0.4
D

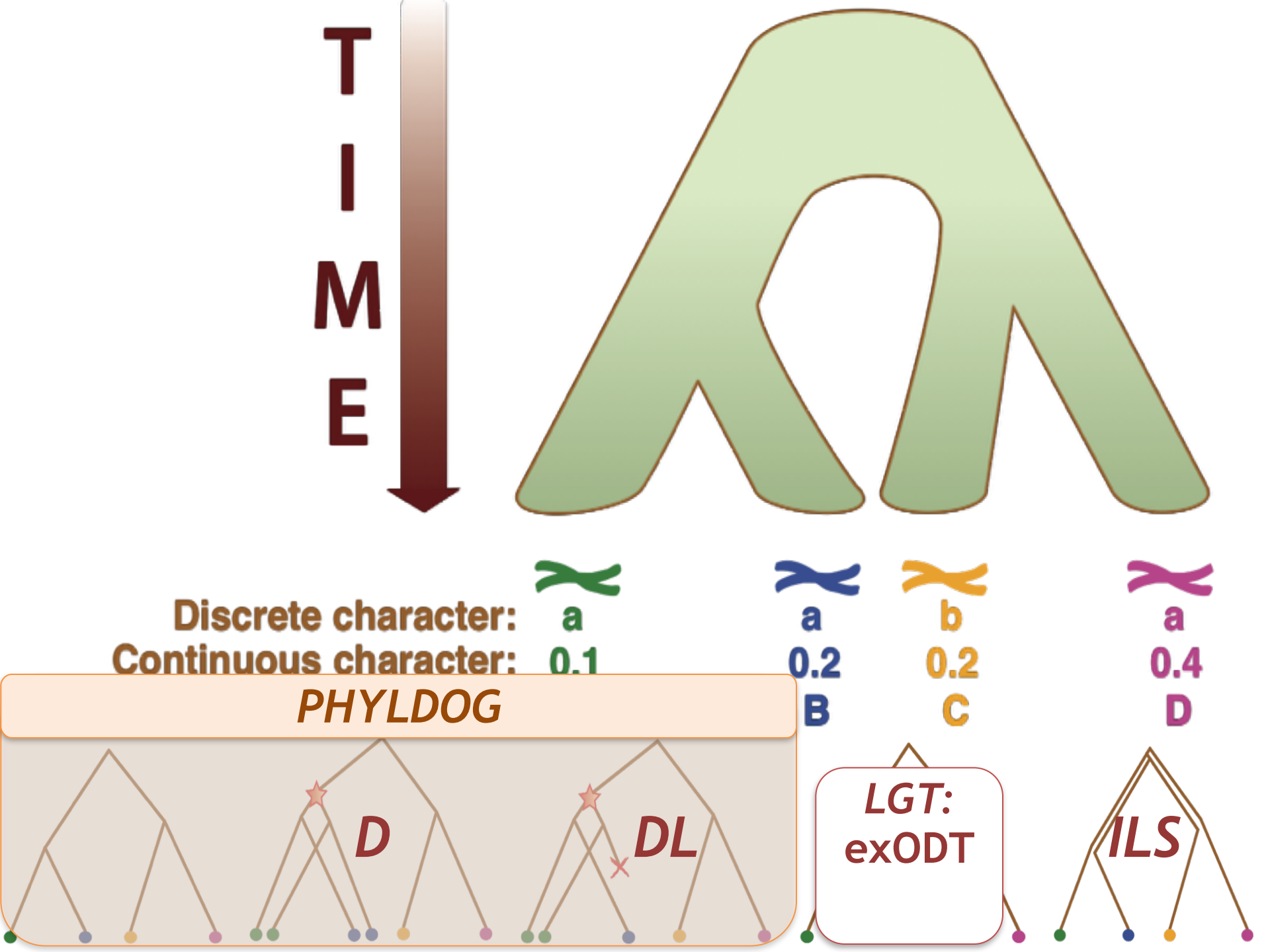

TIME
Discrete character:
Continuous character:
Species:
a
0.1
A
a
0.2
B
b
0.2
C
a
0.4
D
D

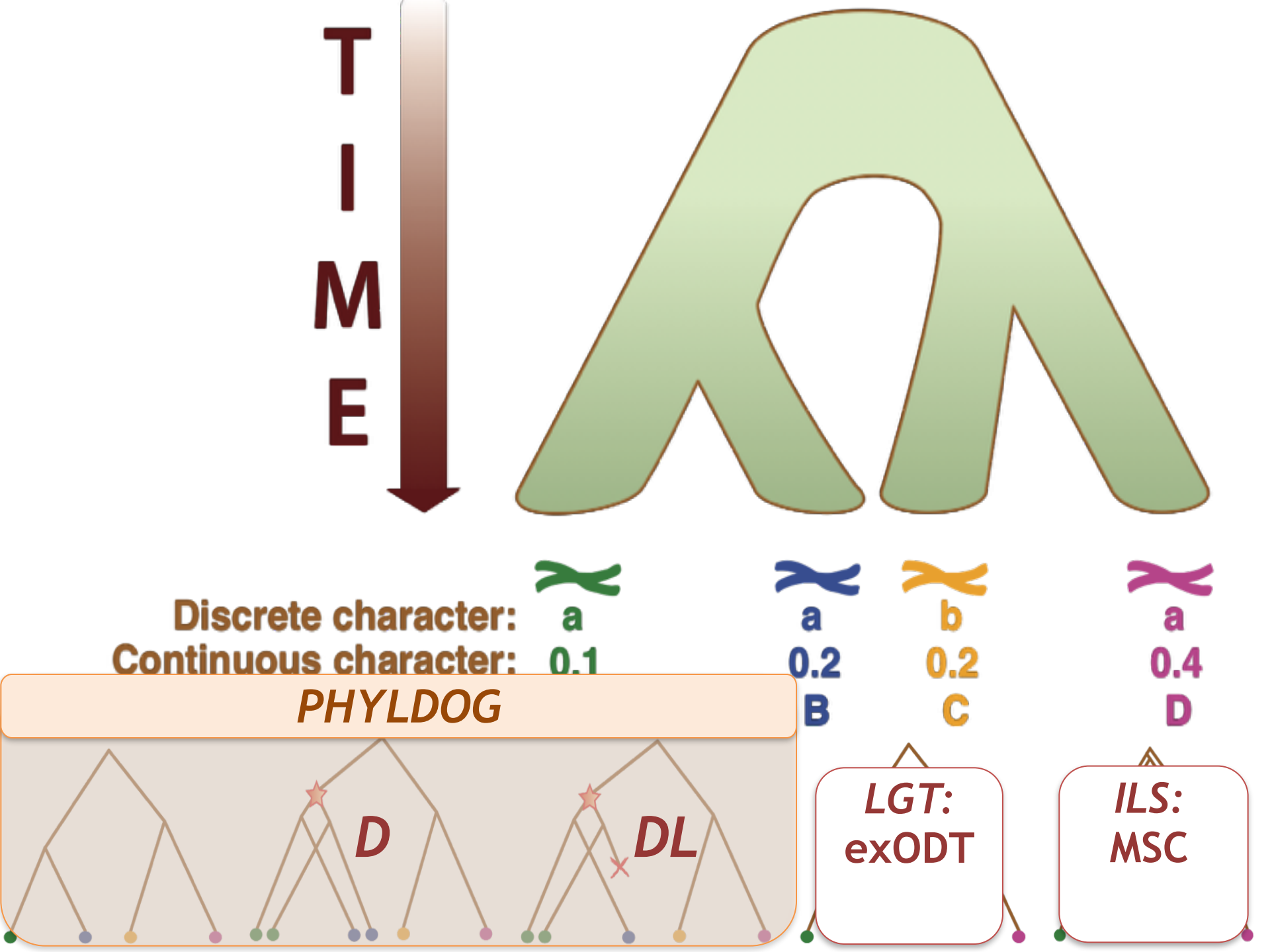

TIME
Discrete character:
Continuous character:
Species:
a
0.1
A
a
0.2
B
b
0.2
C
a
0.4
D
D
DL


TIME
Discrete character:
Continuous character:
Species:
a
0.1
A
a
0.2
B
b
0.2
C
a
0.4
D
D
DL
LGT


TIME
Discrete character:
Continuous character:
Species:
a
0.1
A
a
0.2
B
b
0.2
C
a
0.4
D
D
DL
LGT
ILS

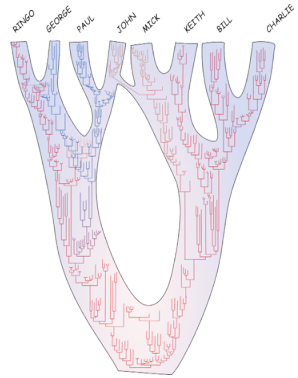

TIME
Discrete character:
Continuous character:
a
0.1
a
0.2
B
b
0.2
C
a
0.4
D
PHYLDOG
D
DL
LGT
ILS

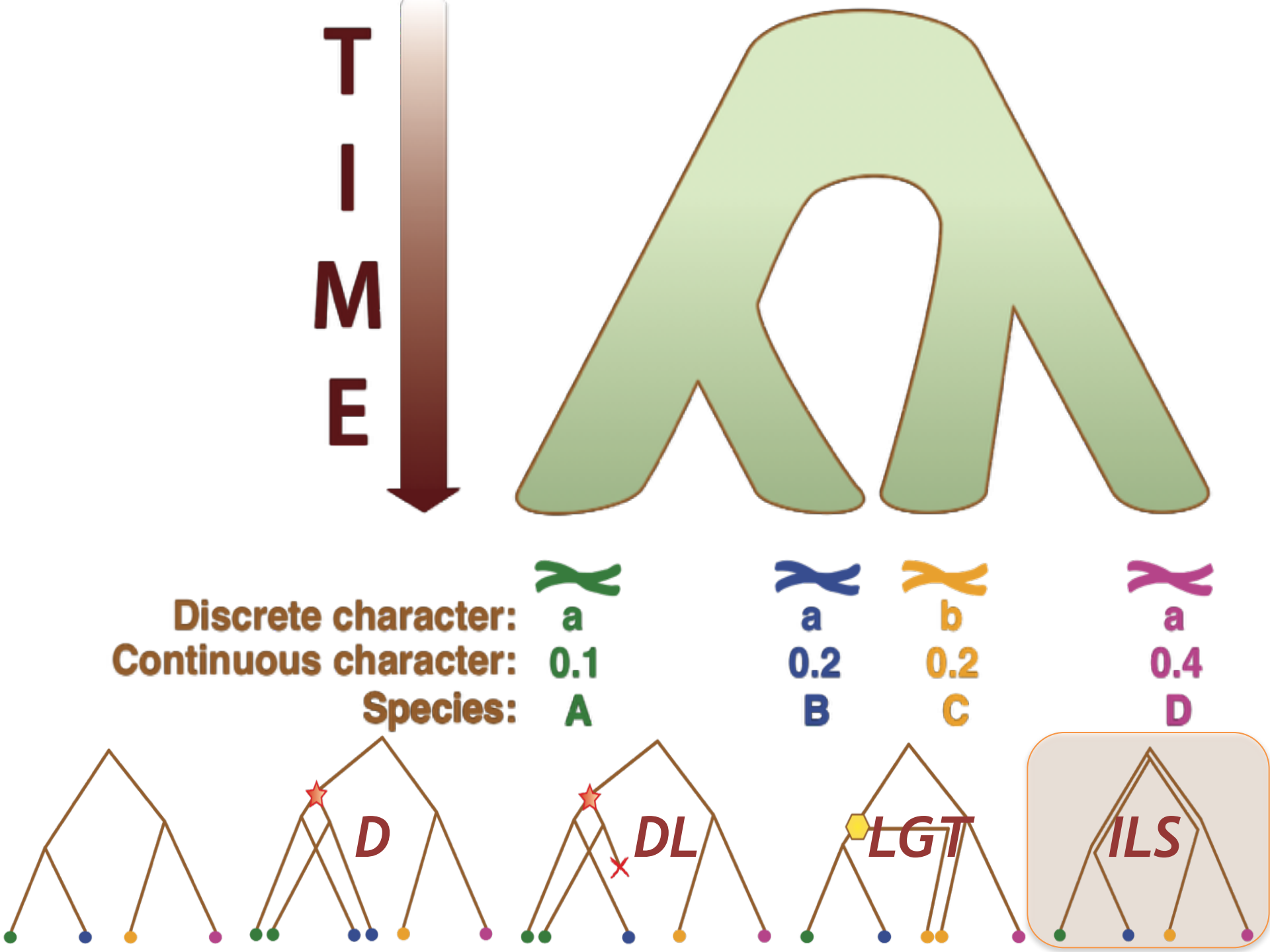

TIME
Discrete character:
Continuous character:
a
0.1
a
0.2
B
b
0.2
C
a
0.4
D
PHYLDOG
D
DL
LGT:
exODT
ILS


TIME
Discrete character:
Continuous character:
a
0.1
a
0.2
B
b
0.2
C
a
0.4
D
PHYLDOG
D
DL
LGT:
exODT
ILS:
MSC

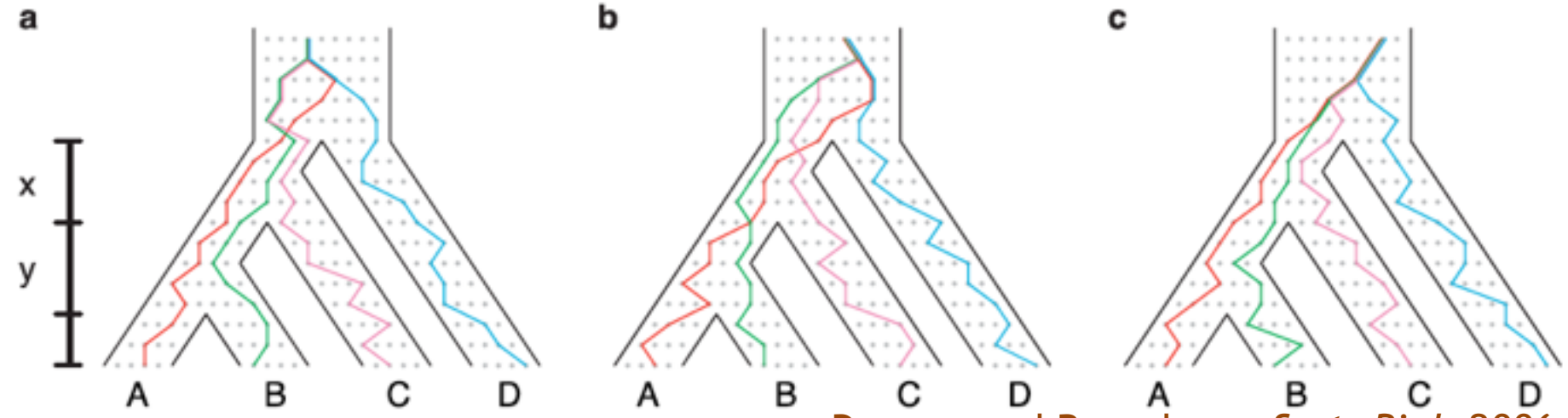
Plan

1. Modeling the relationship between species tree and gene tree
   - coalescent models
   - models of gene duplication and loss
   - models of gene transfer
   - models that combine the above

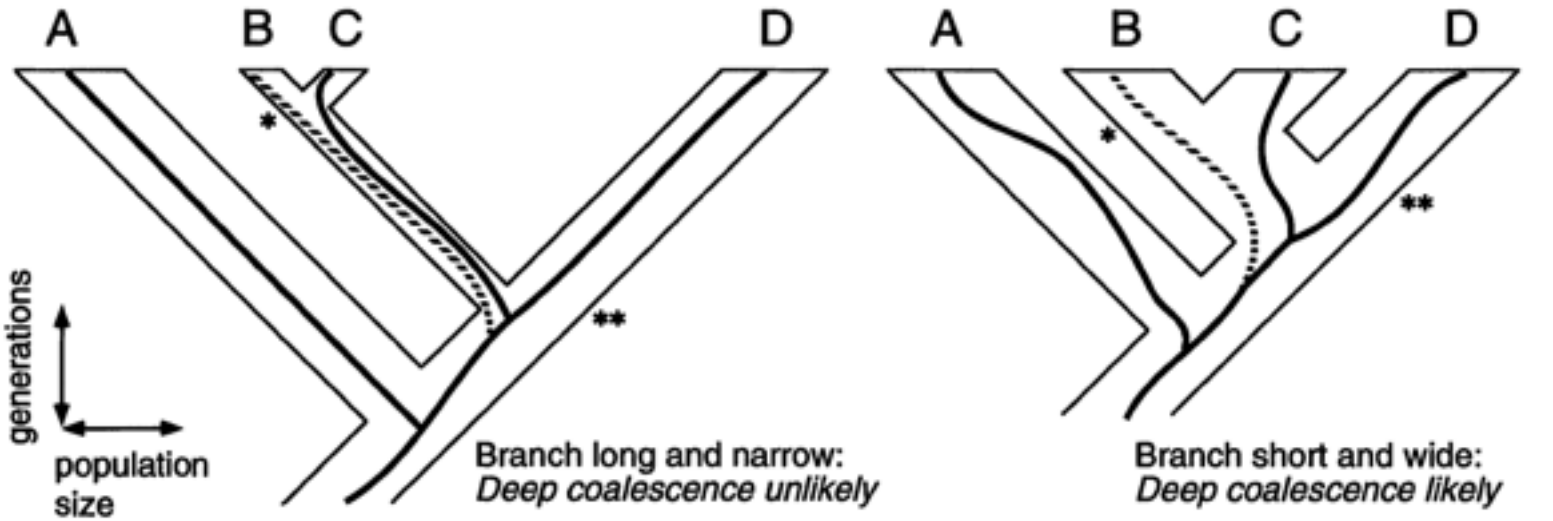



TIME
Discrete character:
Continuous character:
Species:
a
0.1
A
a
0.2
B
b
0.2
C
a
0.4
D
D
DL
LGT
ILS

Incomplete lineage sorting



Degnan and Rosenberg, *Syst. Biol.* 2006



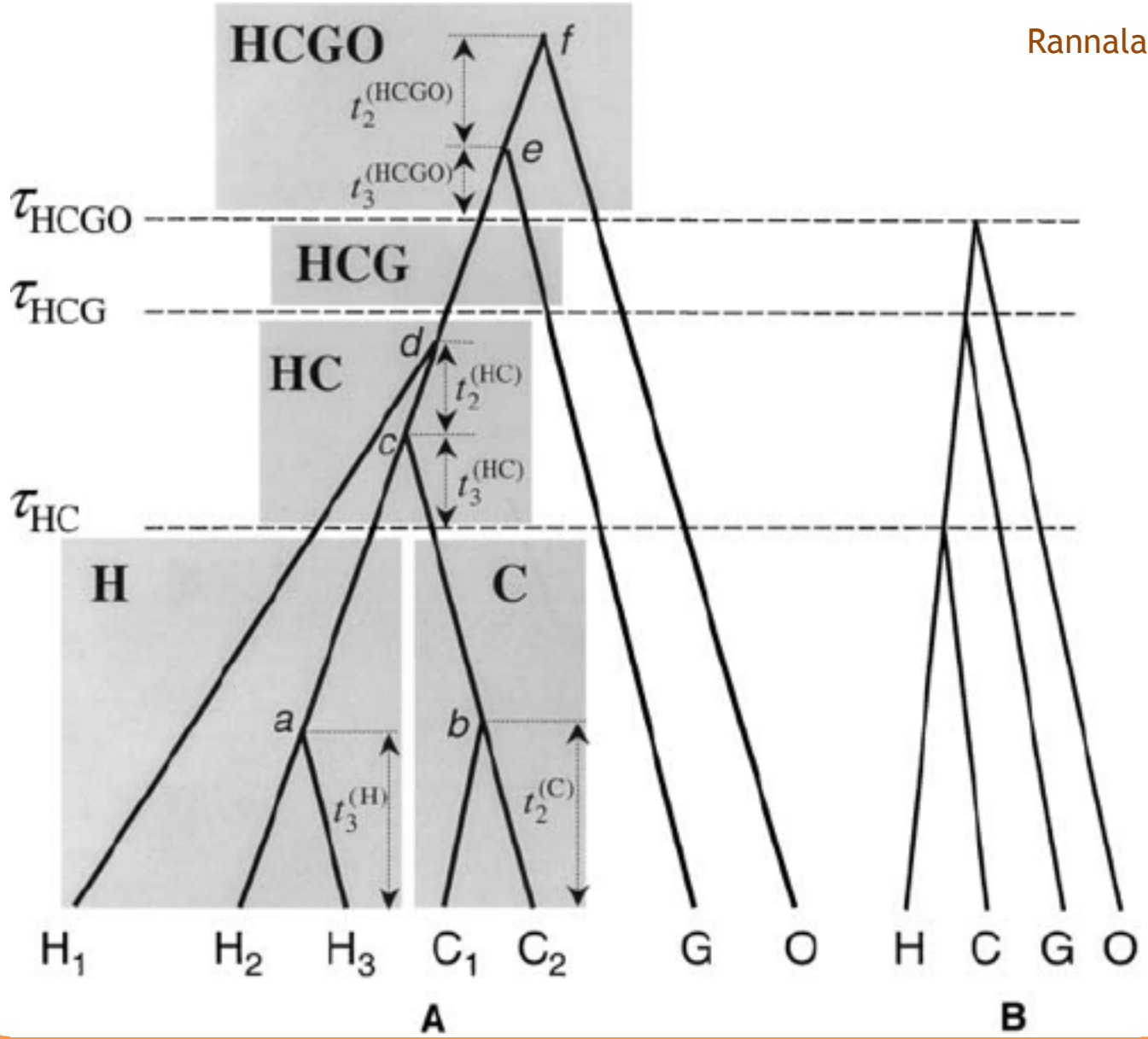
Maddison, *Syst. Biol.* 1997

Reconstructing species trees given ILS

- Parsimony method: minimizing deep coalescences (Maddison 1997)
- Distance method: NJst (Liu and Yu 2011)
- Summary statistic methods: STAR (Liu et al, 2009), GLASS (Mossel and Roch 2010), iGLASS (Jewett and Rosenberg 2012), ASTRAL (Mirarab et al., 2014)
- Maximum Likelihood methods: STEM (Kubatko et al. 2009), MP-EST (Liu et al., 2010)
- Bayesian methods: BeST (2007), *BEAST (Heled and Drummond, 2009)
- Bypassing the gene trees with math tricks: SNAPP (Bryant et al., 2012), POMO (DeMaio et al., 2013), SVDQuartets (Chifman and Kubatko, 2014)

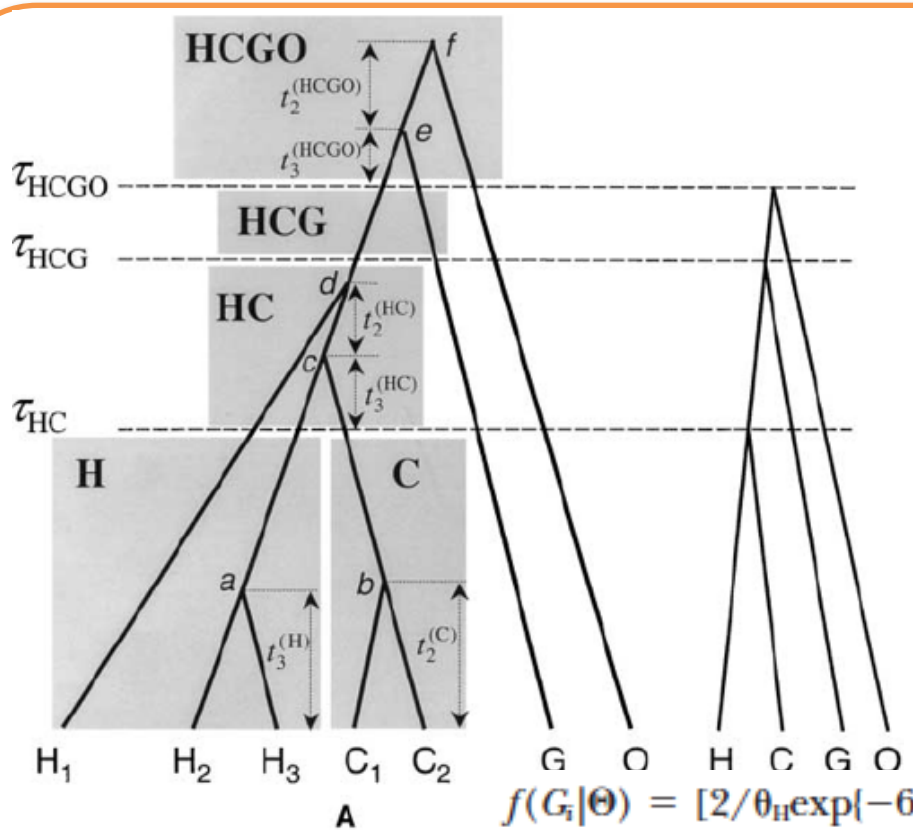
Likelihood of a gene tree given a species tree under the **multi-species coalescent**

Rannala and Yang, *Genetics* 2003



Likelihood of a gene tree given a species tree under the **multi-species coalescent**

Rannala and Yang, *Genetics* 2003



A

$$f(G_i|\Theta) = [2/\theta_H \exp\{-6t_3^{(H)}/\theta_H\}\exp\{-2(\tau_{HC} - t_3^{(H)})/\theta_H\}]$$

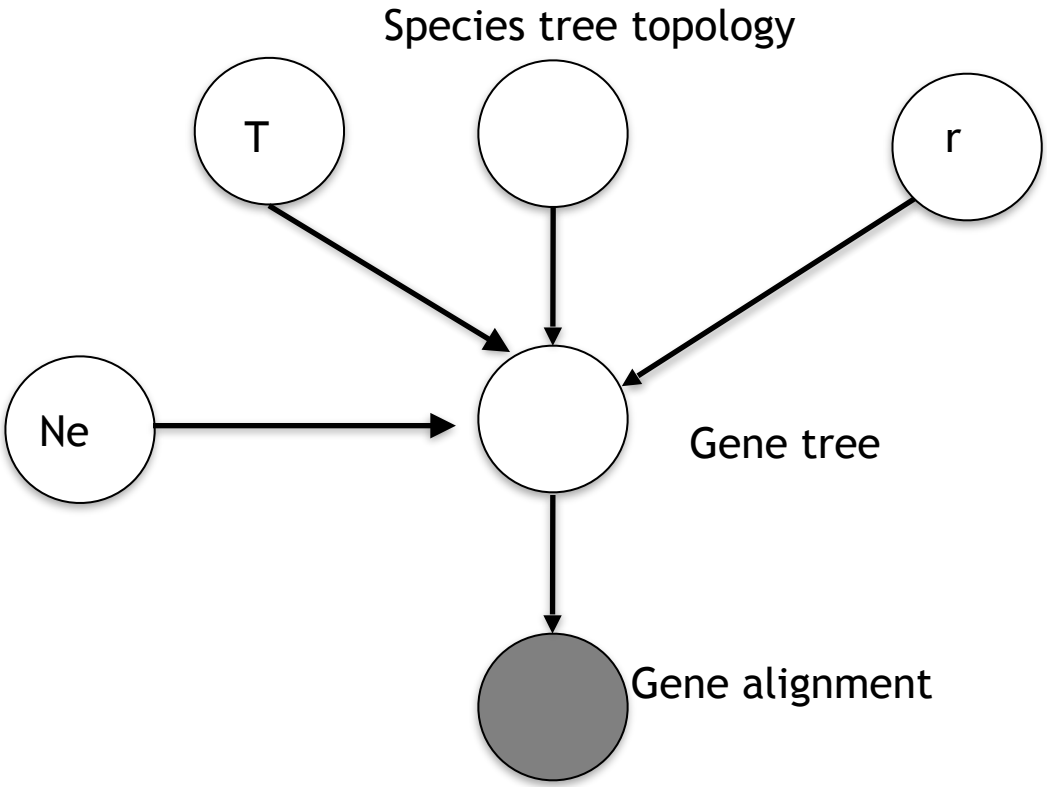
$$\times [2/\theta_C \exp\{-2t_2^{(C)}/\theta_C\}]$$

$$\times [2/\theta_{HC}\exp\{-6t_3^{(HC)}/\theta_{HC}\} \times 2/\theta_{HC}\exp\{-2t_2^{(HC)}/\theta_{HC}\}]$$

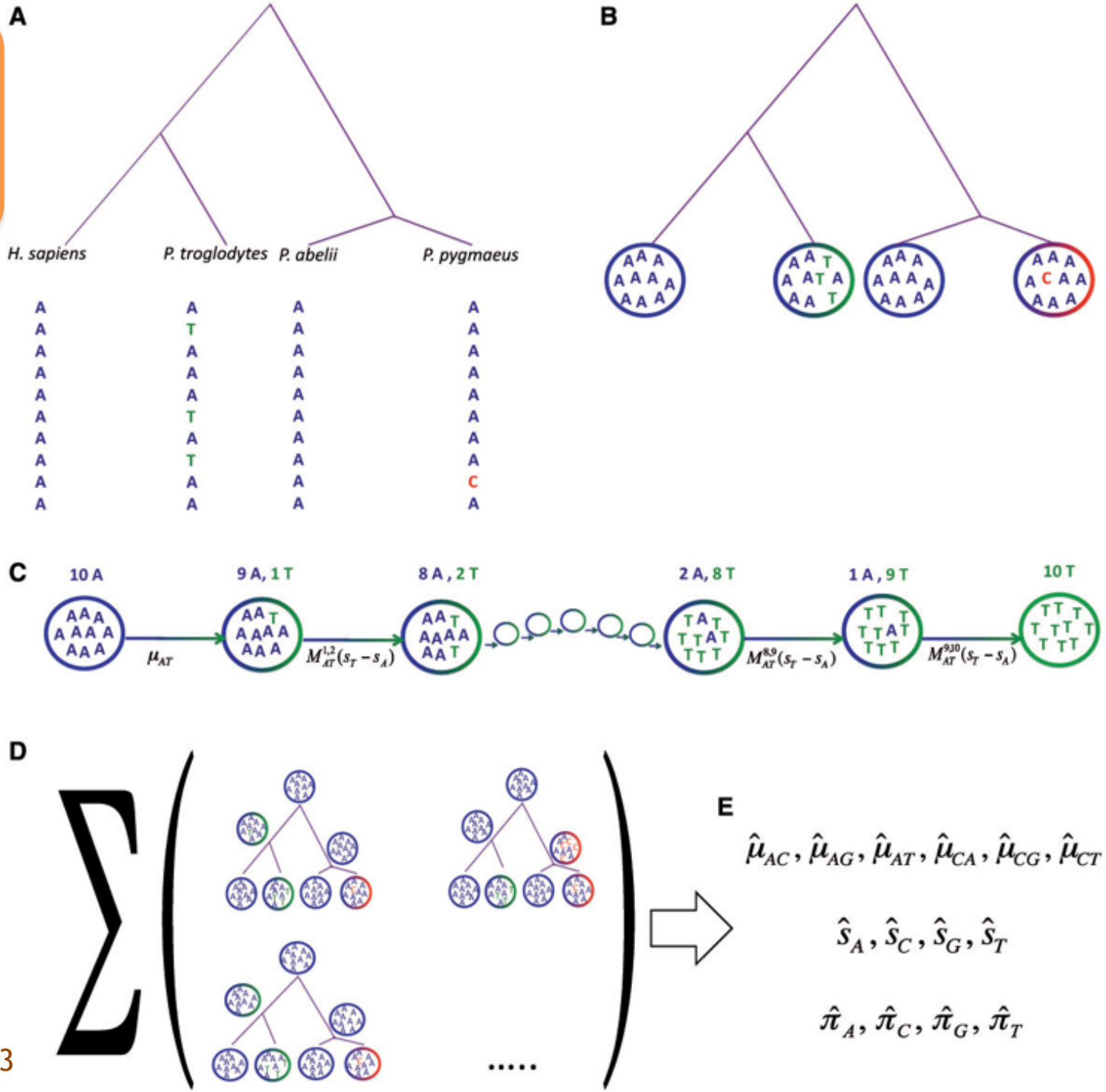
$$\times [\exp\{-2(\tau_{HCG} - \tau_{HC} - (t_3^{(HC)} + t_2^{(HC)}))/\theta_{HCG}\}]$$

$$\times [2/\theta_{HCGO}\exp\{-6t_3^{(HCGO)}/\theta_{HCGO}\} \times 2/\theta_{HCGO}\exp\{-2t_2^{(HCGO)}/\theta_{HCGO}\}].$$

The species tree-gene tree graphical model with ILS

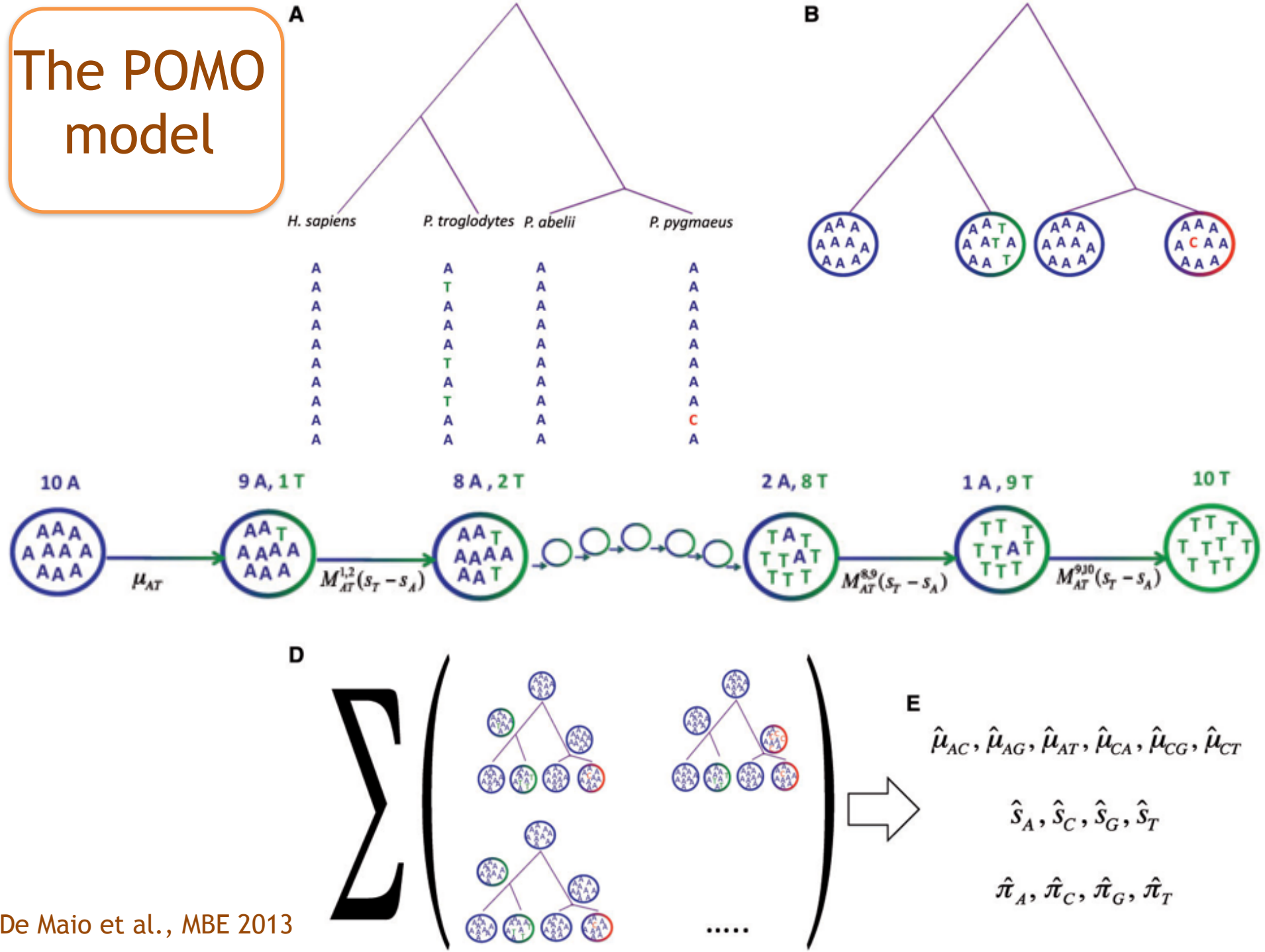


The POMO model



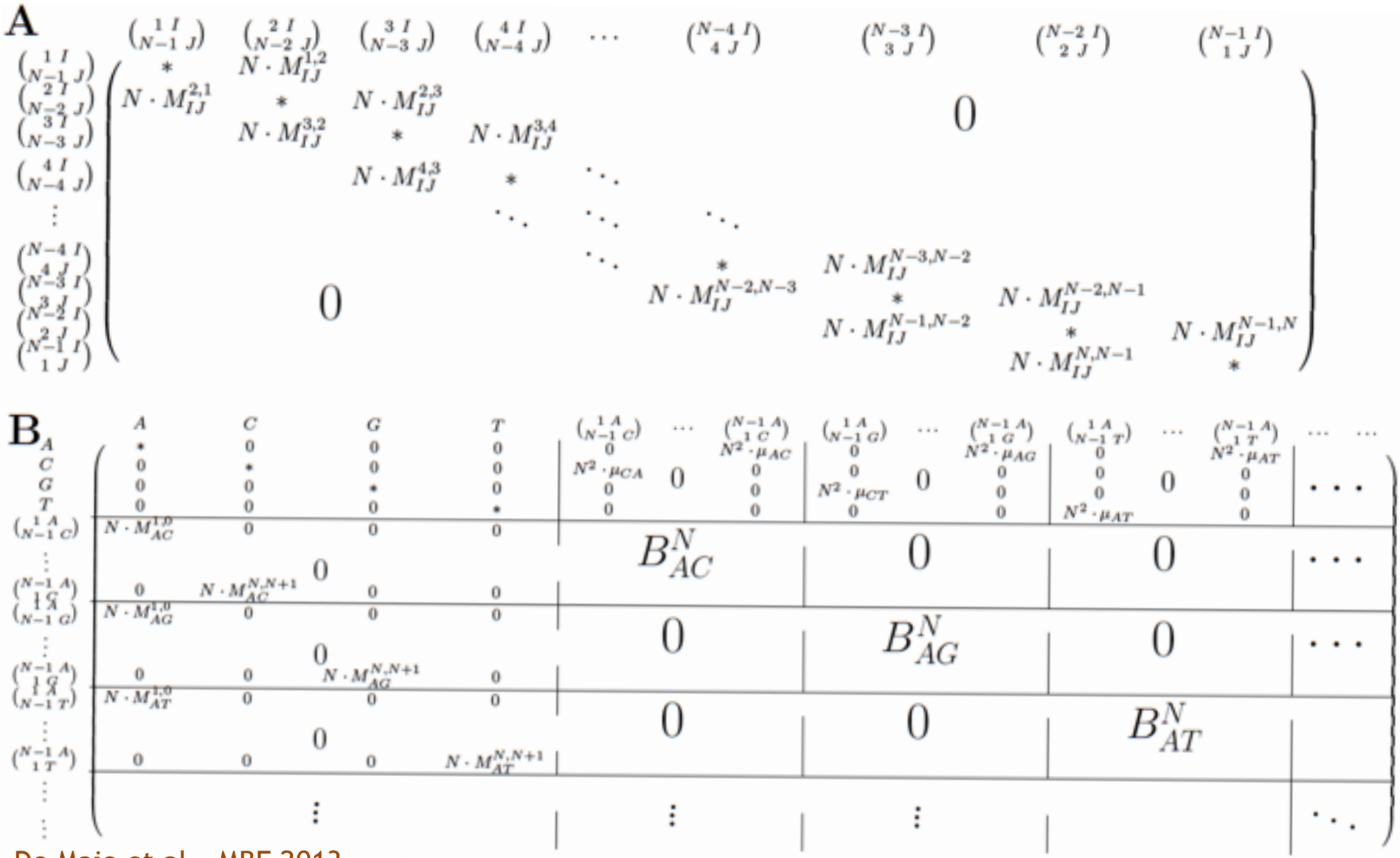
De Maio et al., MBE 2013

The POMO model



De Maio et al., MBE 2013

The POMO model



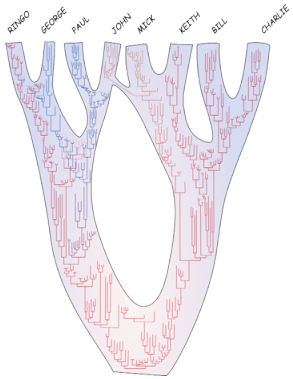
De Maio et al., MBE 2013

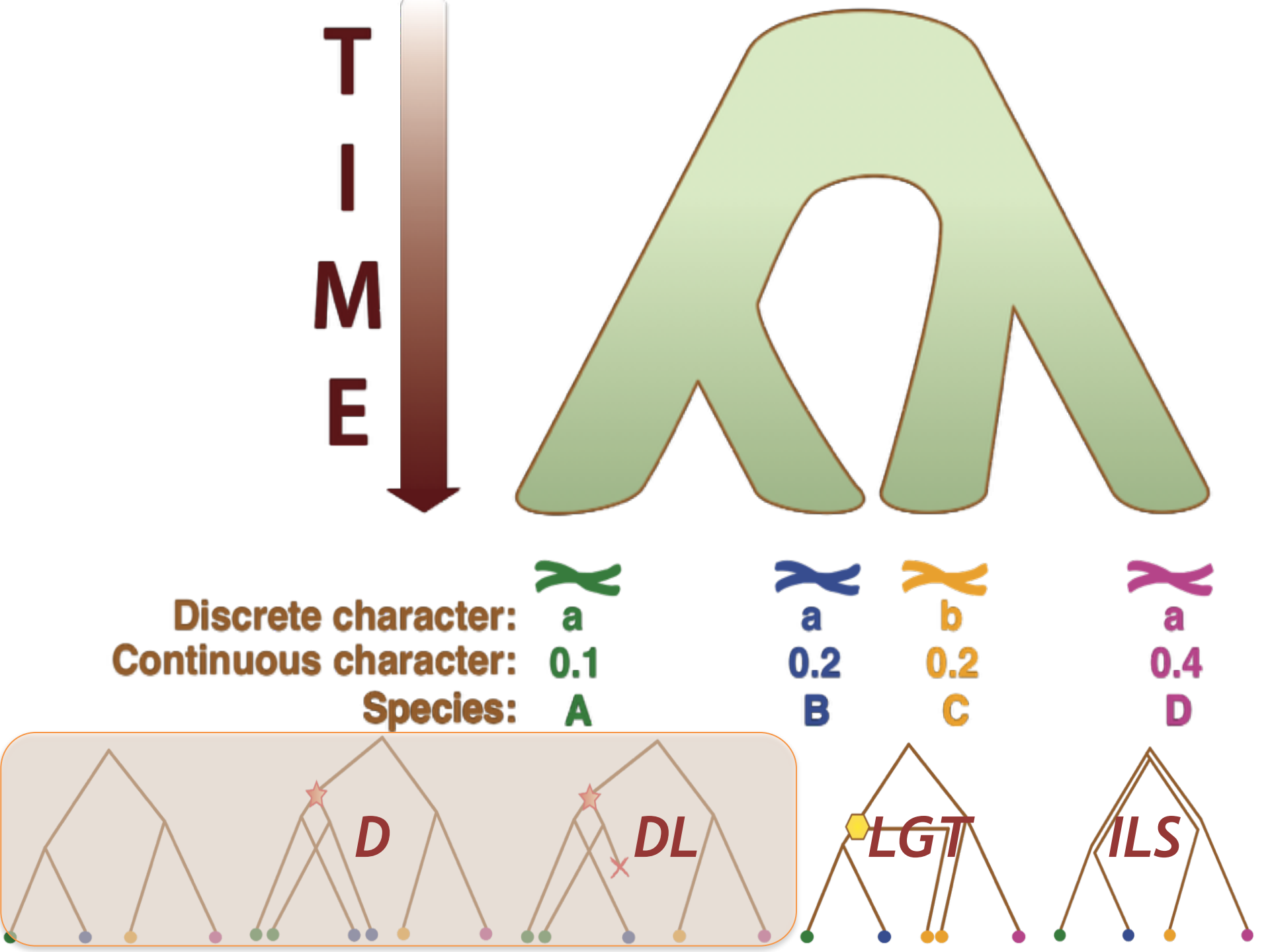
Pros and cons of the Pomo model

1. Pros:
    1. bypasses gene tree estimation (easy to combine with DL or DTL models…)
    2. does not assume linkage between sites
2. Cons:
    1. You don’t get gene trees
    2. It’s not sure what branch lengths mean anymore
    3. matrix can become big
    4. Only bimorphic sites are considered: 3- or 4-morphic sites need to be handled differently
    5. Not sure how it behaves in practice

Plan

1. Modeling the relationship between species tree and gene tree
   - coalescent models
   - models of gene duplication and loss
   - models of gene transfer
   - models that combine the above

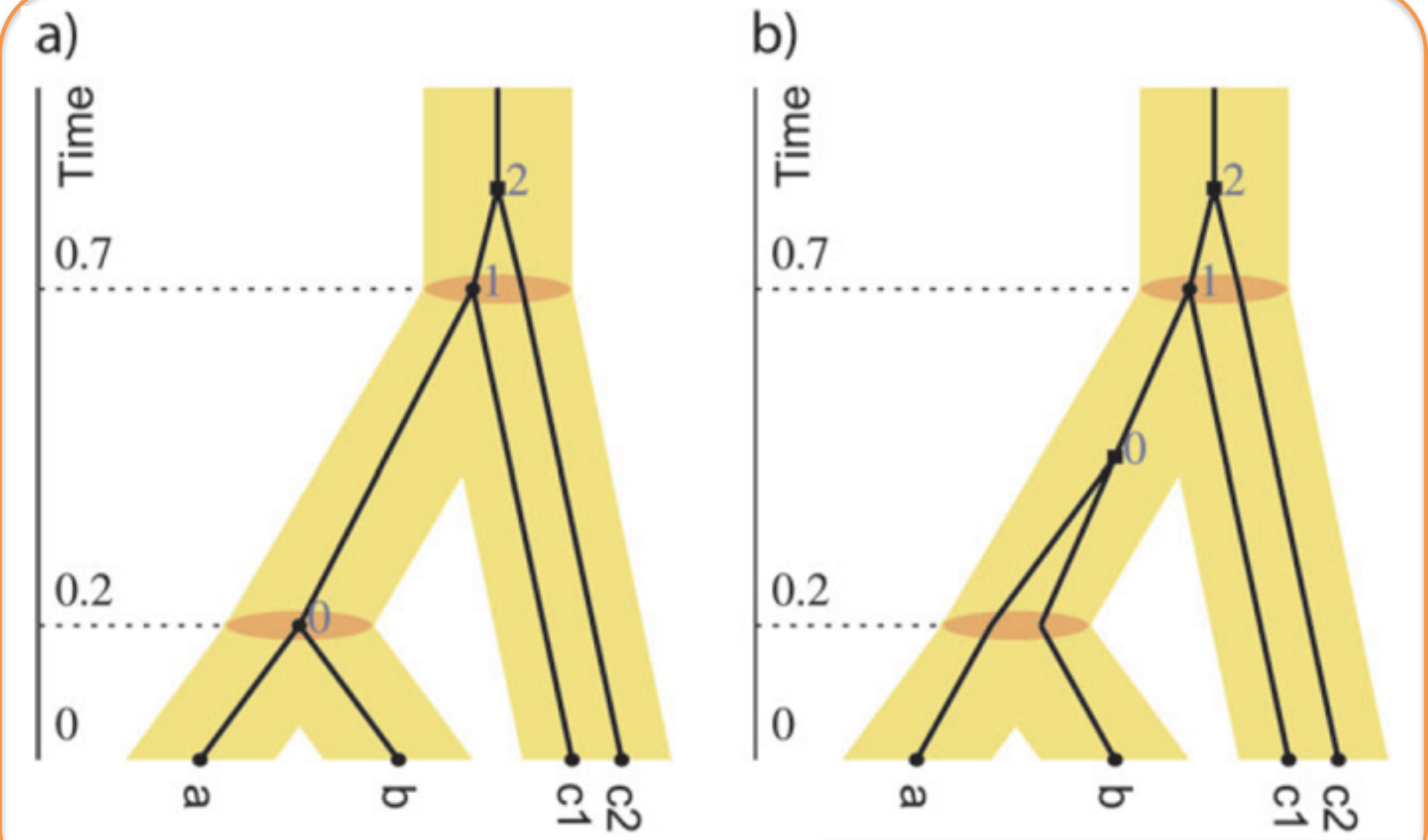



TIME
Discrete character:
Continuous character:
Species:
a
0.1
A
a
0.2
B
b
0.2
C
a
0.4
D
D
DL
LGT
ILS

DL models

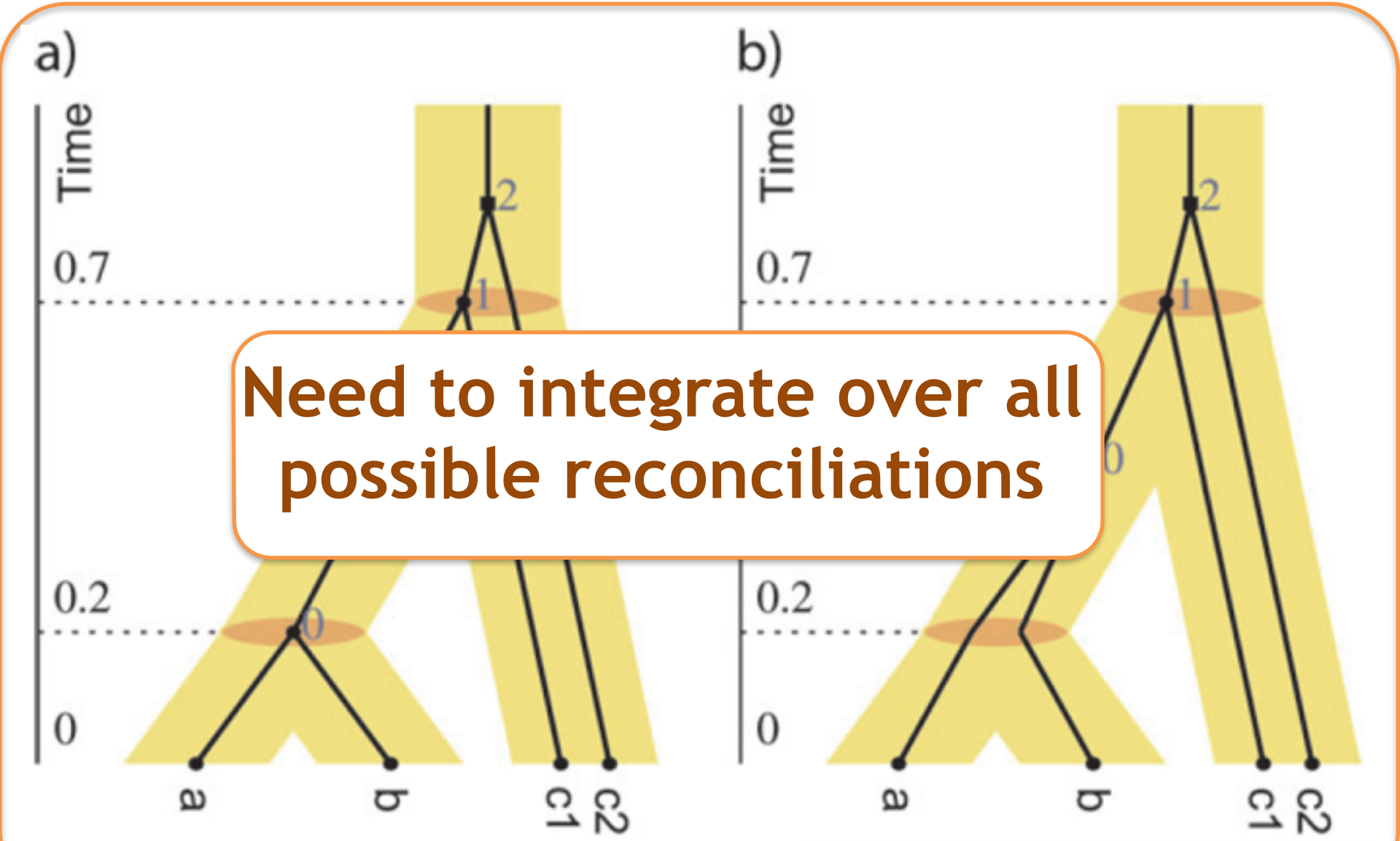
- **Reconciliation between a gene tree and a species tree:**
  - Parsimony method: minimizing numbers of duplications (Goodman et al. 1979; Page 1994; Zmasek and Eddy 2001)
  - Model-based method: Arvestad et al. 2003, 2004, 2009; Górecki et al. 2011
- **Reconstruction of a gene tree given a species tree and alignment:**
  - Parsimony method: Chen et al. 2000; Durand et al. 2006
  - Model-based method: Rasmussen and Kellis 2007, 2011; Akerborg et al. 2009; Sjöstrand et al. 2012
- **Reconstruction of a species tree given gene trees:**
  - Parsimony methods: Page and Charleston 1997, Bansal et al. 2007; Wehe et al. 2008; Bansal et al. 2010; Chang et al. 2011
- **Joint reconstruction of gene and species trees given alignments:**
  - Model-based method: Boussau et al., 2013

Many possible reconciliations



Sennblad and Lagergren, *Syst. Biol.* 2009

Many possible reconciliations

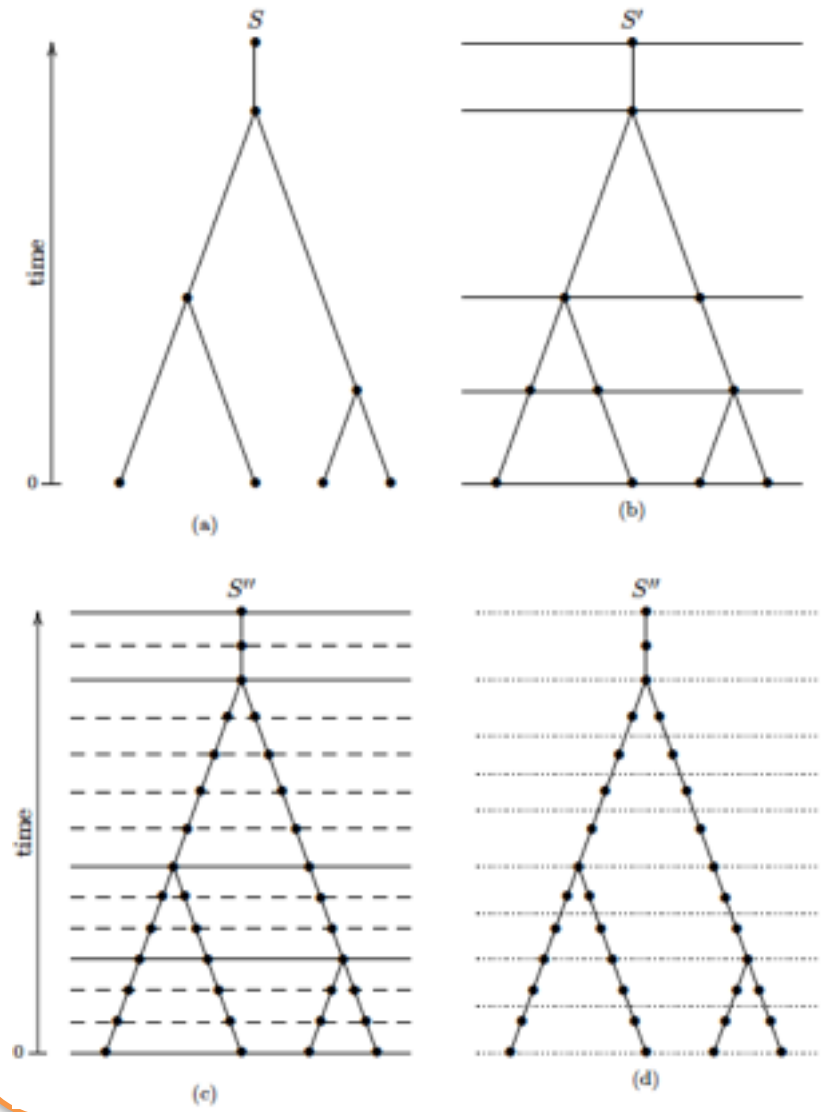


Need to integrate over all possible reconciliations

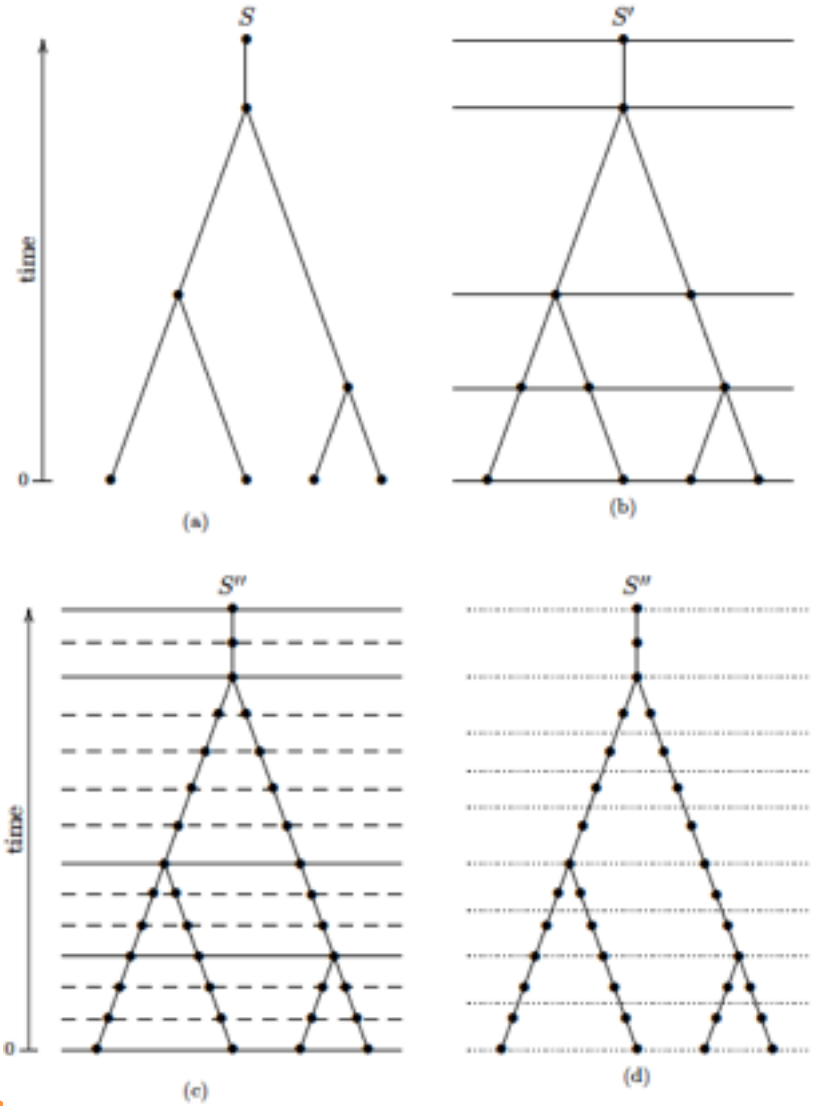
Sennblad and Lagergren, *Syst. Biol.* 2009

Considering all mappings: The species tree discretization approximation

Tofigh, *PhD thesis*



Considering all mappings: The species tree discretization approximation



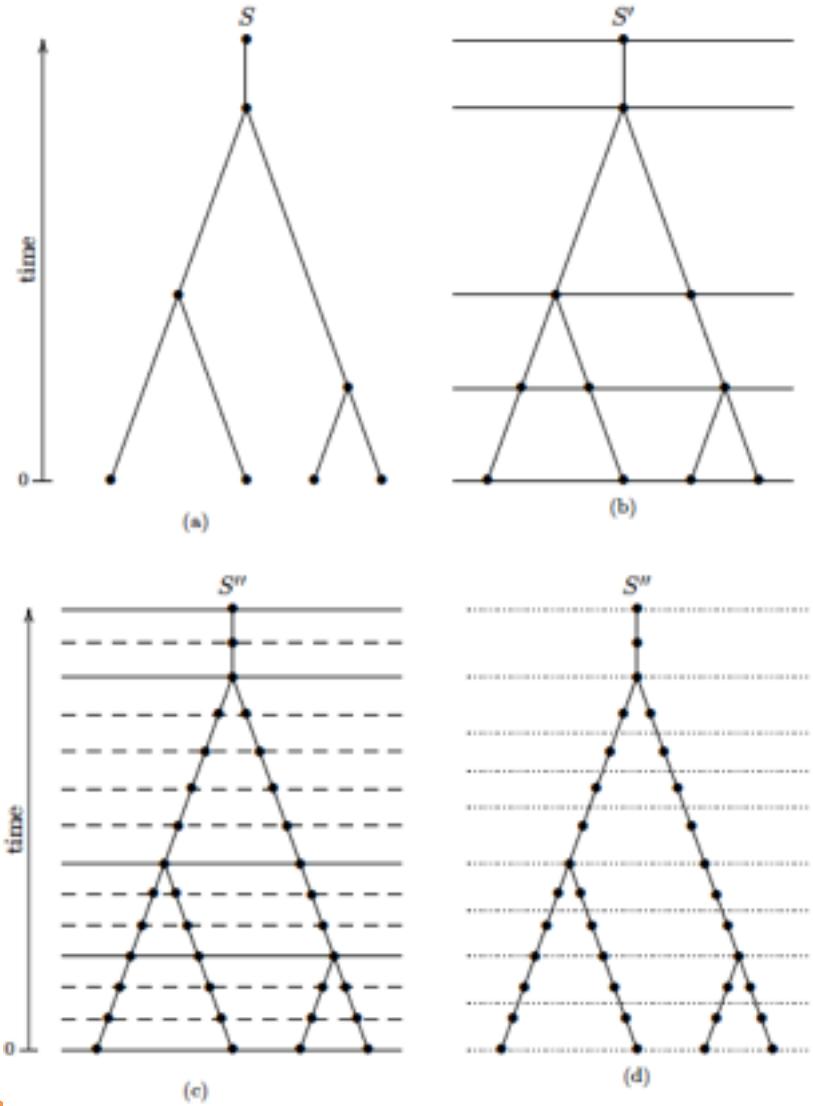
Tofigh, *PhD thesis*

Mapping between a gene tree and species tree

⇔

Mapping between nodes of the gene tree and nodes of the augmented species tree

Considering all mappings: The species tree discretization approximation



Tofigh, *PhD thesis*

Mapping between a gene tree and species tree

⇔

Mapping between nodes of the gene tree and nodes of the augmented species tree

Integrating over all mappings: done through dynamic programming from the leaves of the gene tree to its root

Birth-death models for DL

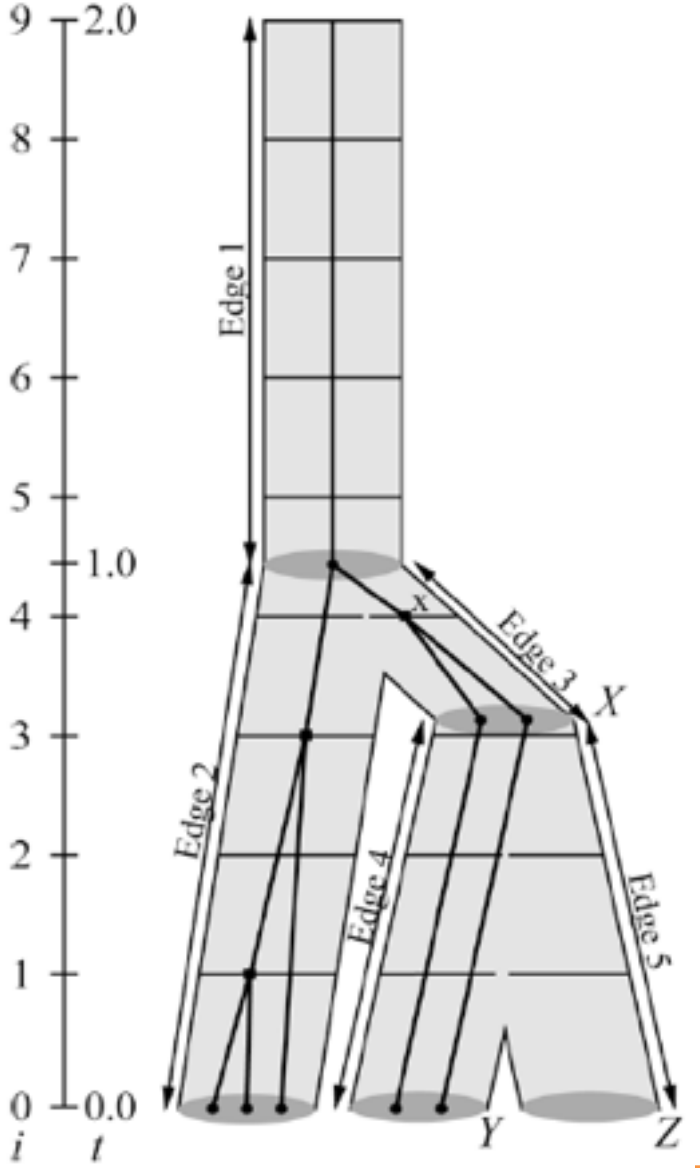
Model-based approaches:

often use a **birth-death process** (Arvestad et al., 2003, 2004, 2009; Akerborg et al., 2009; Sjöstrand et al., 2012; Boussau et al., 2013)

Birth=Duplication: λ

Death=Loss: μ

Akerborg et al., *PNAS* 2009



The Akerborg et al. 2009 model

- Input: Species tree with time, rooted gene tree, λ and μ assumed to be known
- Output: reconciliation between species tree and gene tree
- Relaxed-clock model
- Gene trees generated according to a birth-death model

$$\Pr[D, G|S] = \int \Pr[D|G, l = rt]\, p[r|G] p[G, t|S] dt dr$$

The Akerborg et al. 2009 model

- Input: Species tree with time, rooted gene tree, λ and μ assumed to be known
- Output: reconciliation between species tree and gene tree
- Relaxed-clock model
- Gene trees generated according to a birth-death model

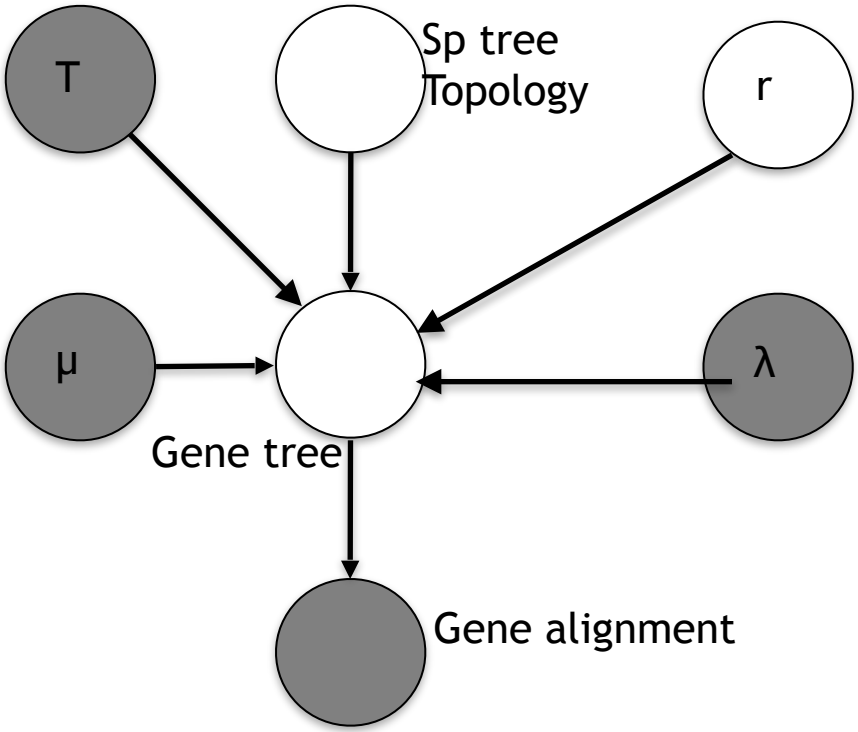
$$\Pr[D, G|S] = \int \Pr[D|G, l = rt]\, p[r|G] p[G, t|S] dt dr$$

Felsenstein 1981

Gamma prior

Dynamic programming to integrate over all mappings with BD model

The species tree-gene tree graphical model with DL (Akerborg 2009)



The Akerborg et al. 2009 model vs PHYLDOG

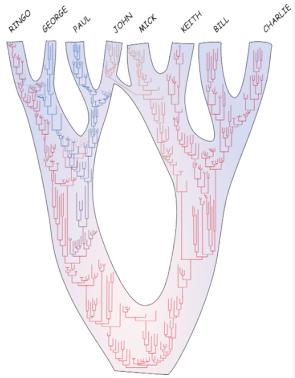
- Akerborg et al. 2009:
  - Generative model
  - Integrates over a large space of mappings
  - Integrates over all possible scenarios of duplications and losses

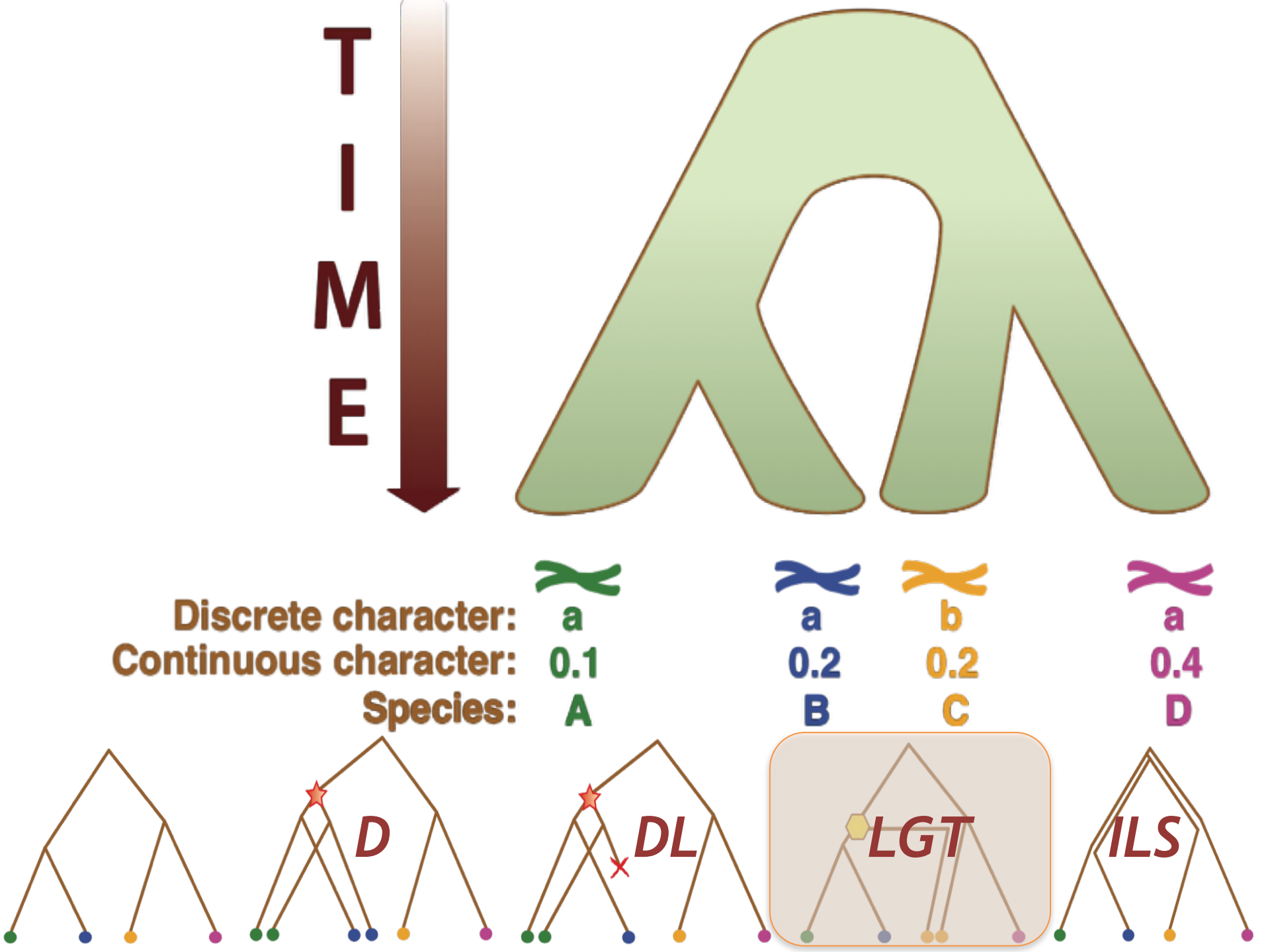
PHYLDOG

- use unrooted, non-clock gene trees
- Use the most parsimonious mapping
- Use a double-recursive tree traversal
- Integrates over a large subset of possible scenarios of duplications and losses (Boussau et al., Genome Research 2013)

Plan

1. Modeling the relationship between species tree and gene tree
   - coalescent models
   - models of gene duplication and loss
   - models of gene transfer
   - models that combine the above




TIME
Discrete character:
Continuous character:
Species:
a
0.1
A
a
0.2
B
b
0.2
C
a
0.4
D
D
DL
LGT
ILS

Pure transfer models

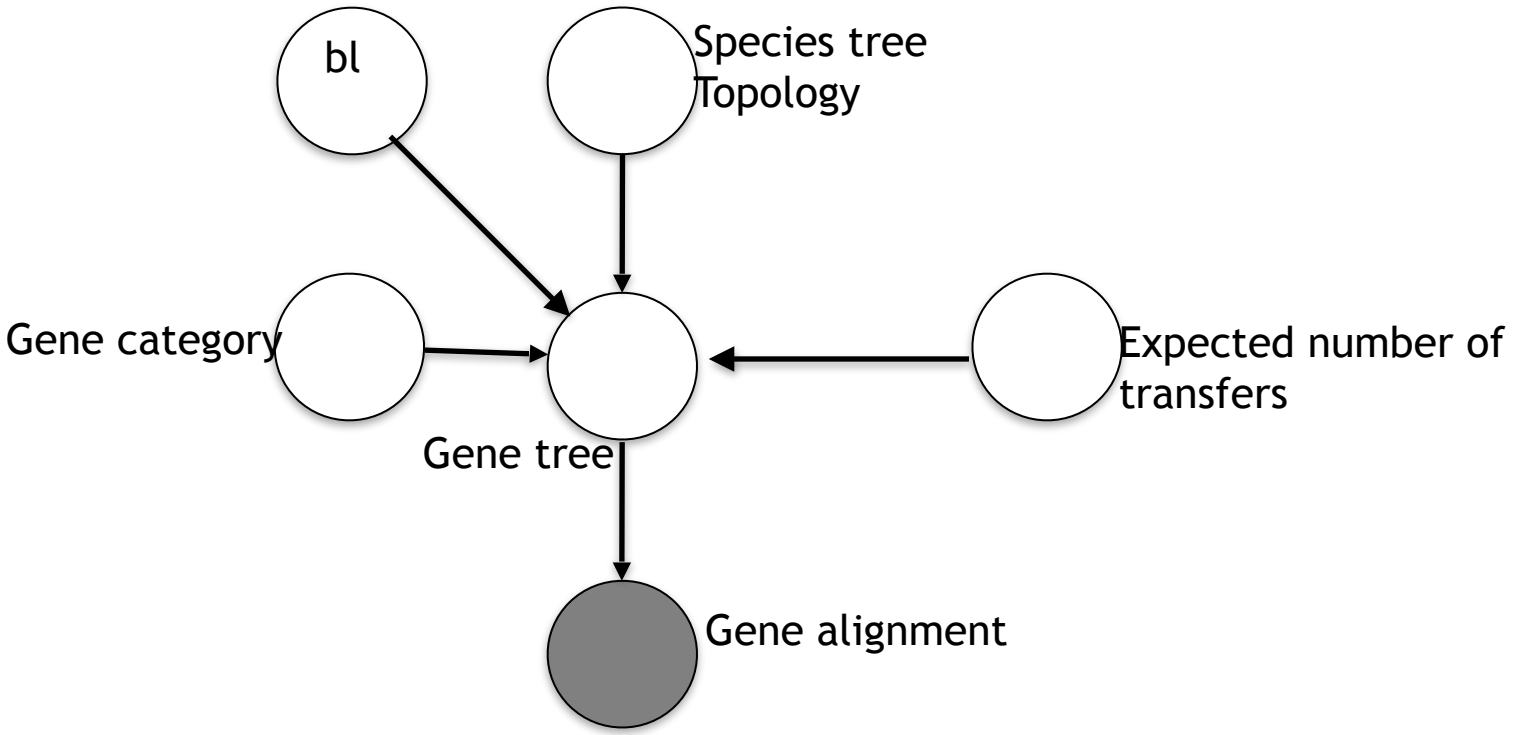
- **Identification of transfers (not probabilistic):**
  - Nakleh et al., 2005; Beiko and Hamilton 2006; Abby et al., 2010; Hill et al. 2010
- **Joint reconstruction of gene and species trees given alignments:**
  - Model-based method: Suchard, *Genetics* 2005

Suchard "random walk through tree space" model

- Input: gene alignments
- Output: species tree, gene trees
- Each gene tree is a Poisson number of transfers away from the species tree
- Random walk through gene trees
- C classes of genes, with different propensities to be transferred
- Has been used with at most 6 species, 144 orthologous gene alignments, but could be used with 8 species or even more

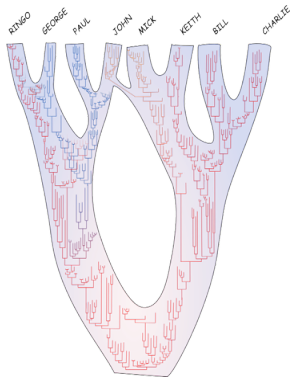
Suchard, *Genetics* 2005

The species tree-gene tree graphical model with transfers (Suchard 2005)



Plan

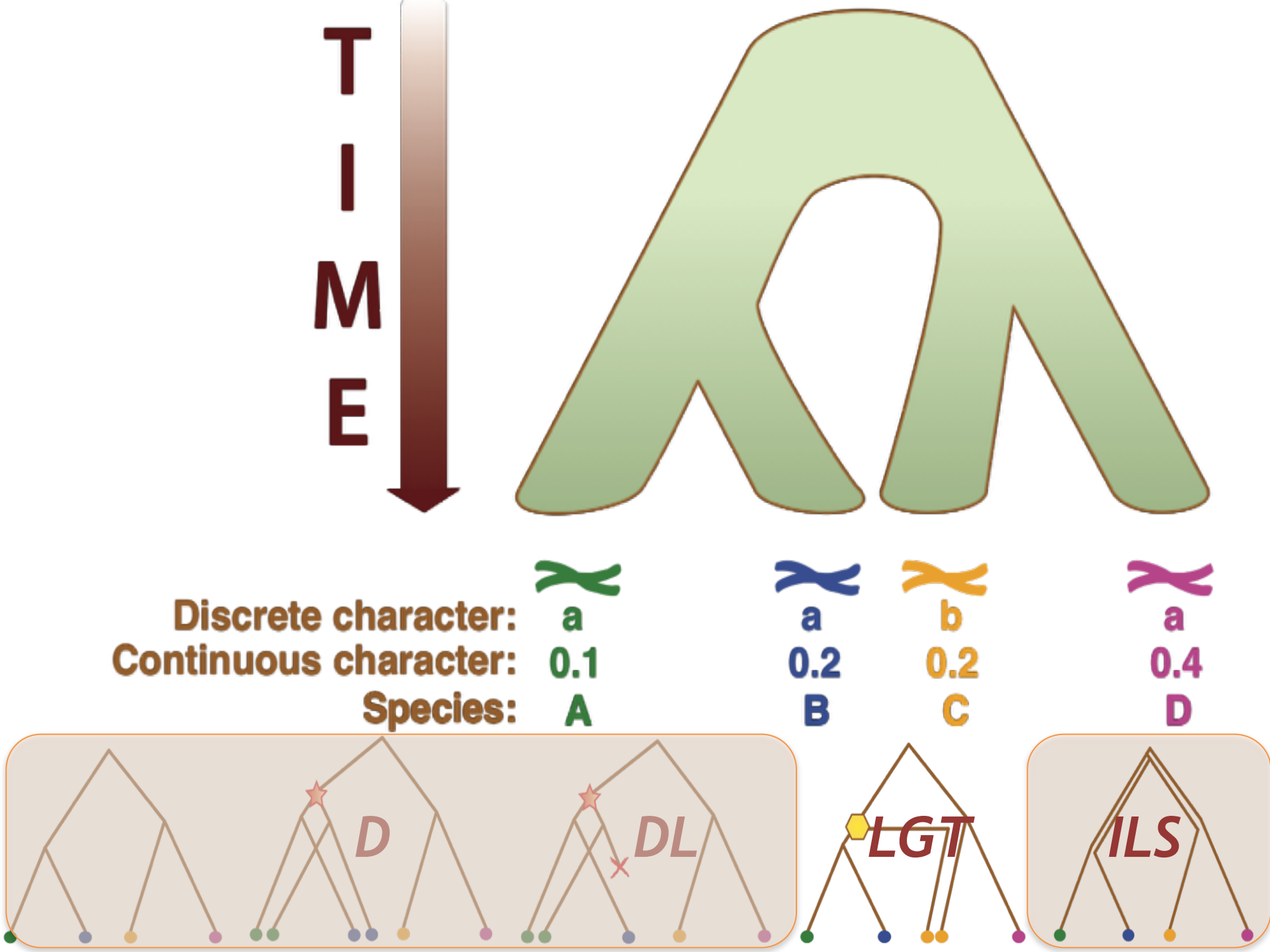
1. Modeling the relationship between species tree and gene tree
   - coalescent models
   - models of gene duplication and loss
   - models of gene transfer
   - models that combine the above



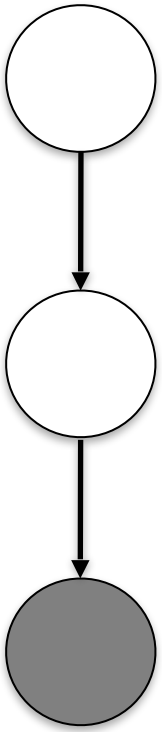
Model that combines DL and ILS

- **Reconciliation between gene tree and species tree:**
  - Model-based method: Rasmussen and Kellis, *Genome Research* 2012

- Input: dated and rooted species tree, DL rates, Ne
- Output: Locus tree and DL+ILS scenario


TIME
Discrete character:
Continuous character:
Species:
a
0.1
A
a
0.2
B
b
0.2
C
a
0.4
D
D
DL
LGT
ILS

The Rasmussen and Kellis model (2012)

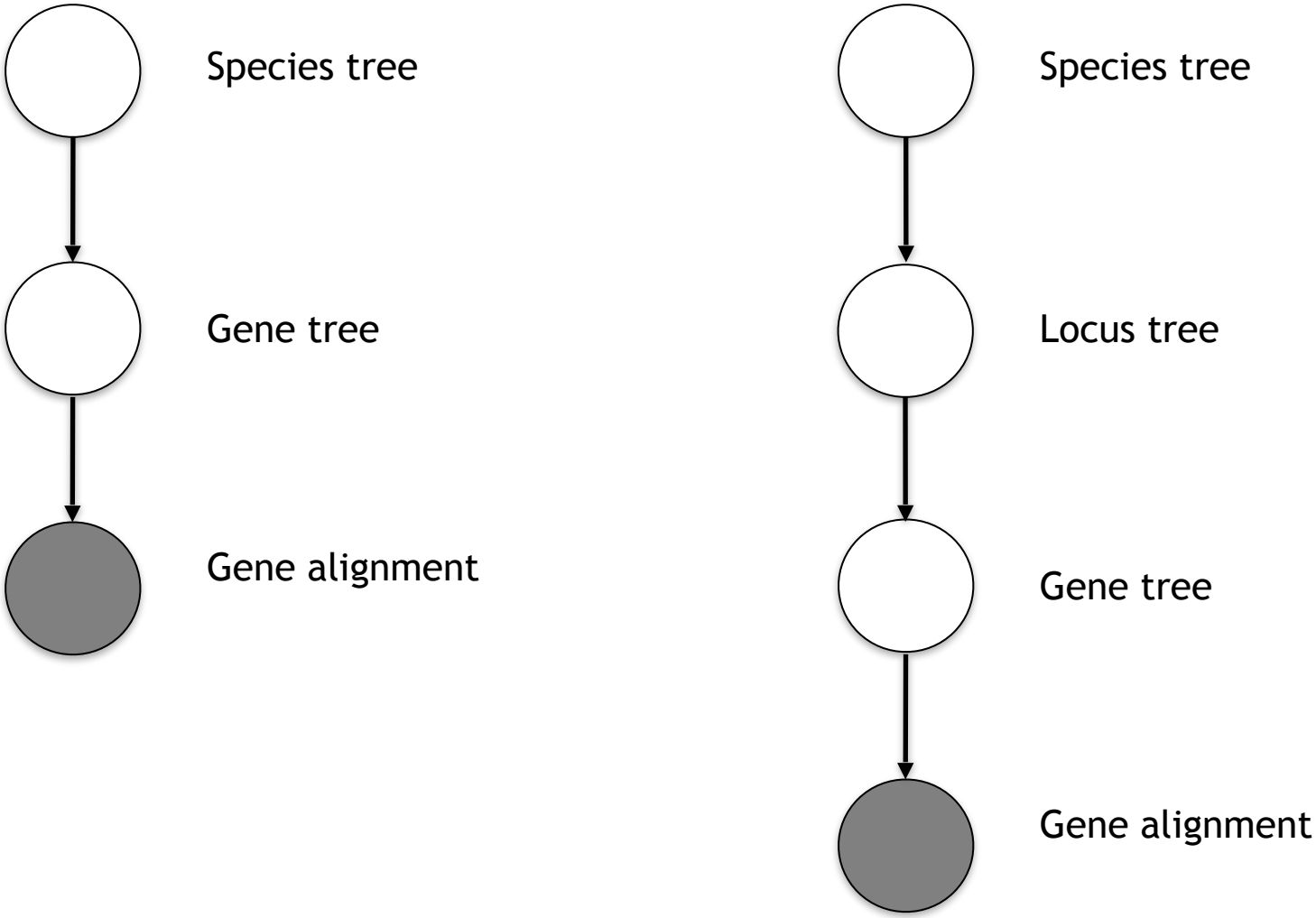


Species tree

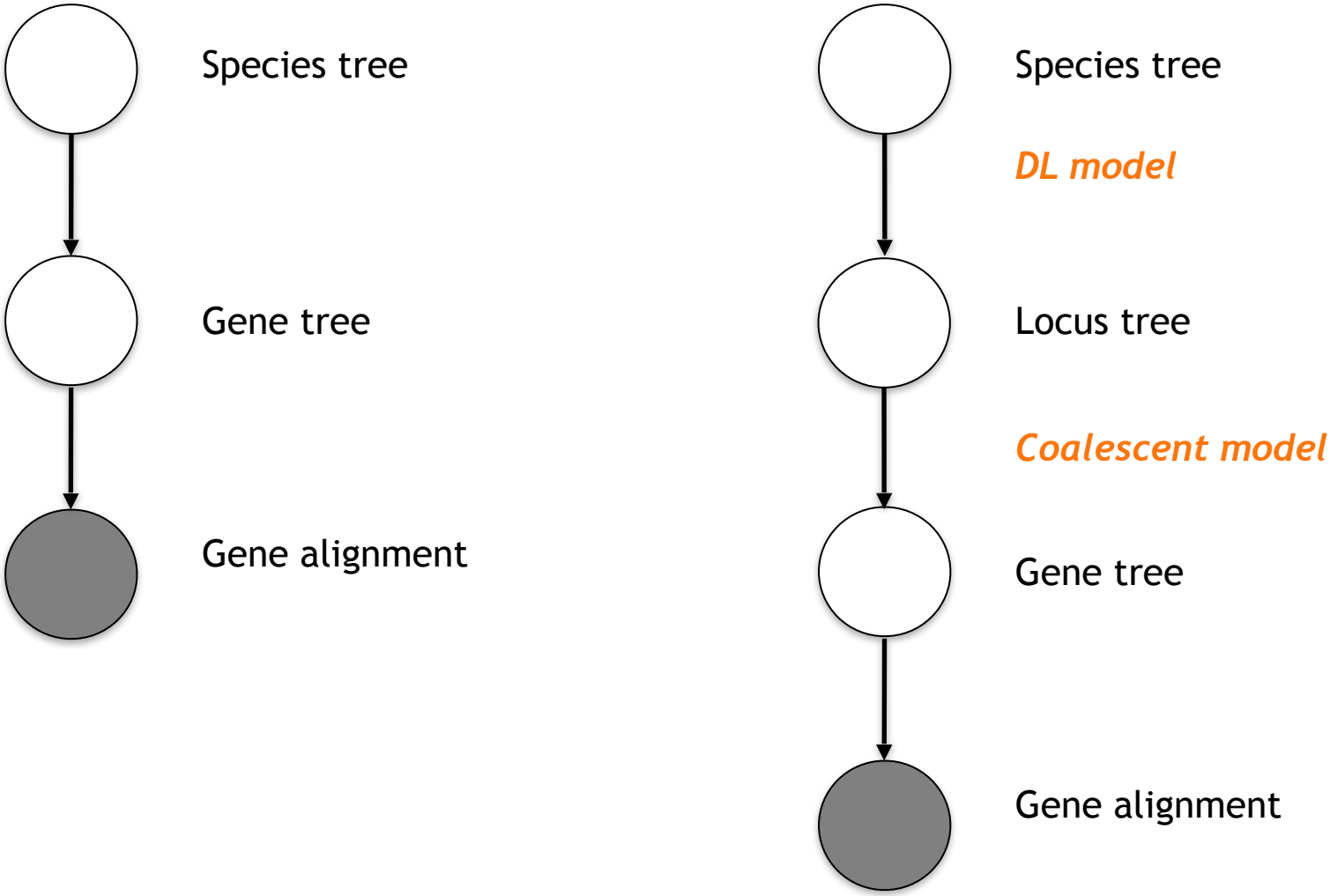
Gene tree

Gene alignment

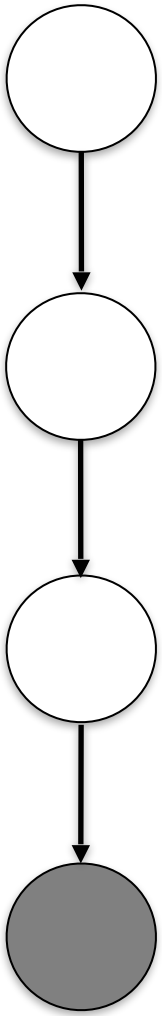
The Rasmussen and Kellis model (2012)



The Rasmussen and Kellis model (2012)



The Rasmussen and Kellis model (2012)



Species tree

DL model

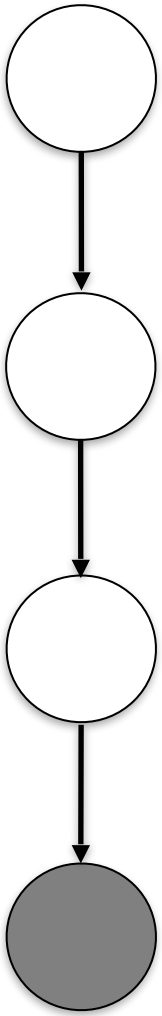
Locus tree

Coalescent model

Gene tree

Gene alignment

The Rasmussen and Kellis model (2012)



Species tree

DL model

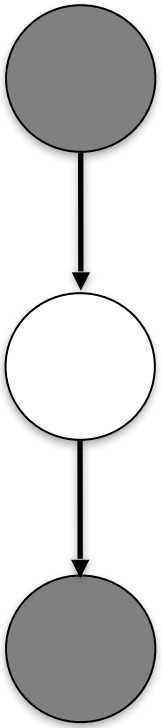
Locus tree

Coalescent model

Gene tree

Gene alignment

In practice:



Species tree

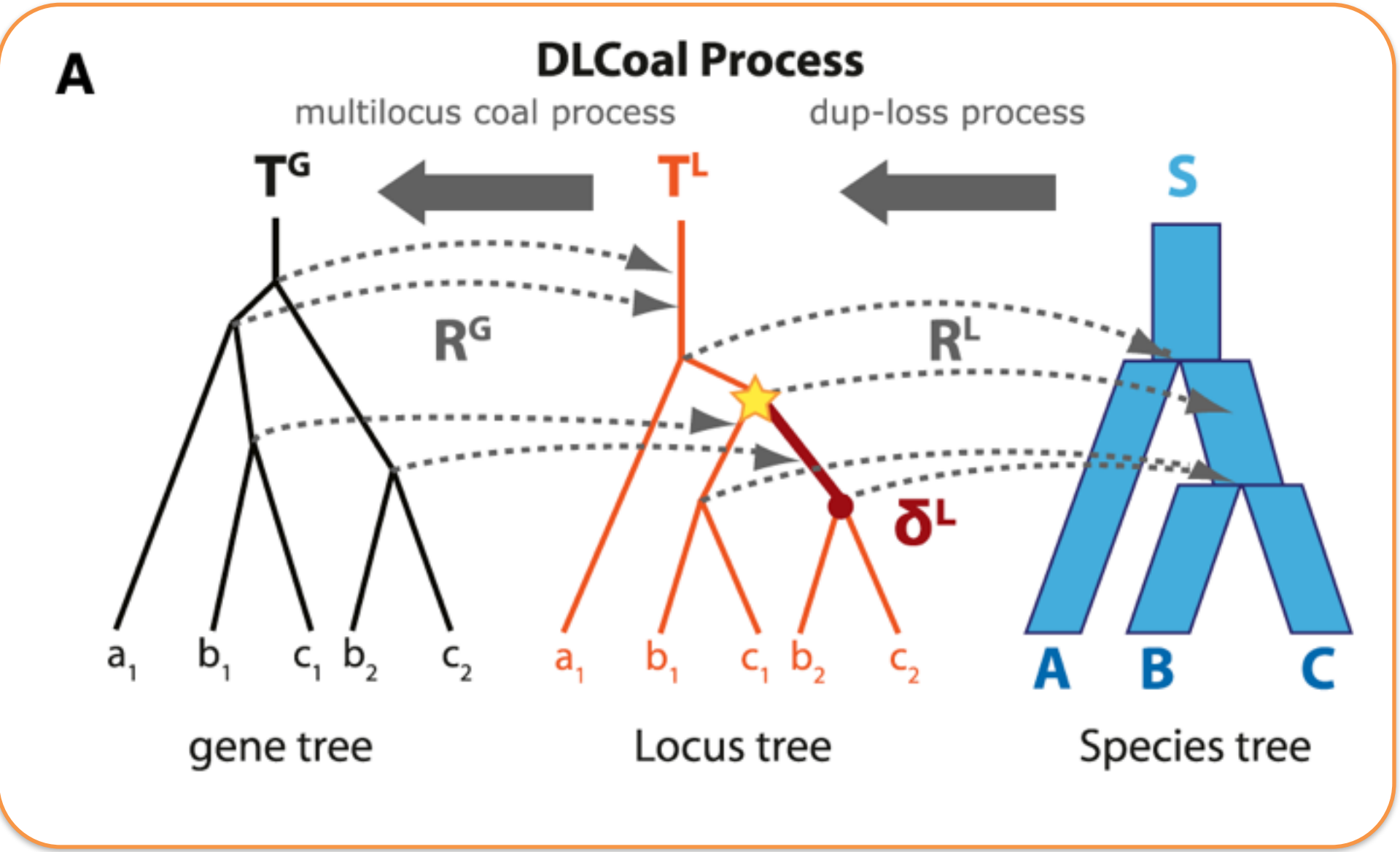
DL model

Locus tree

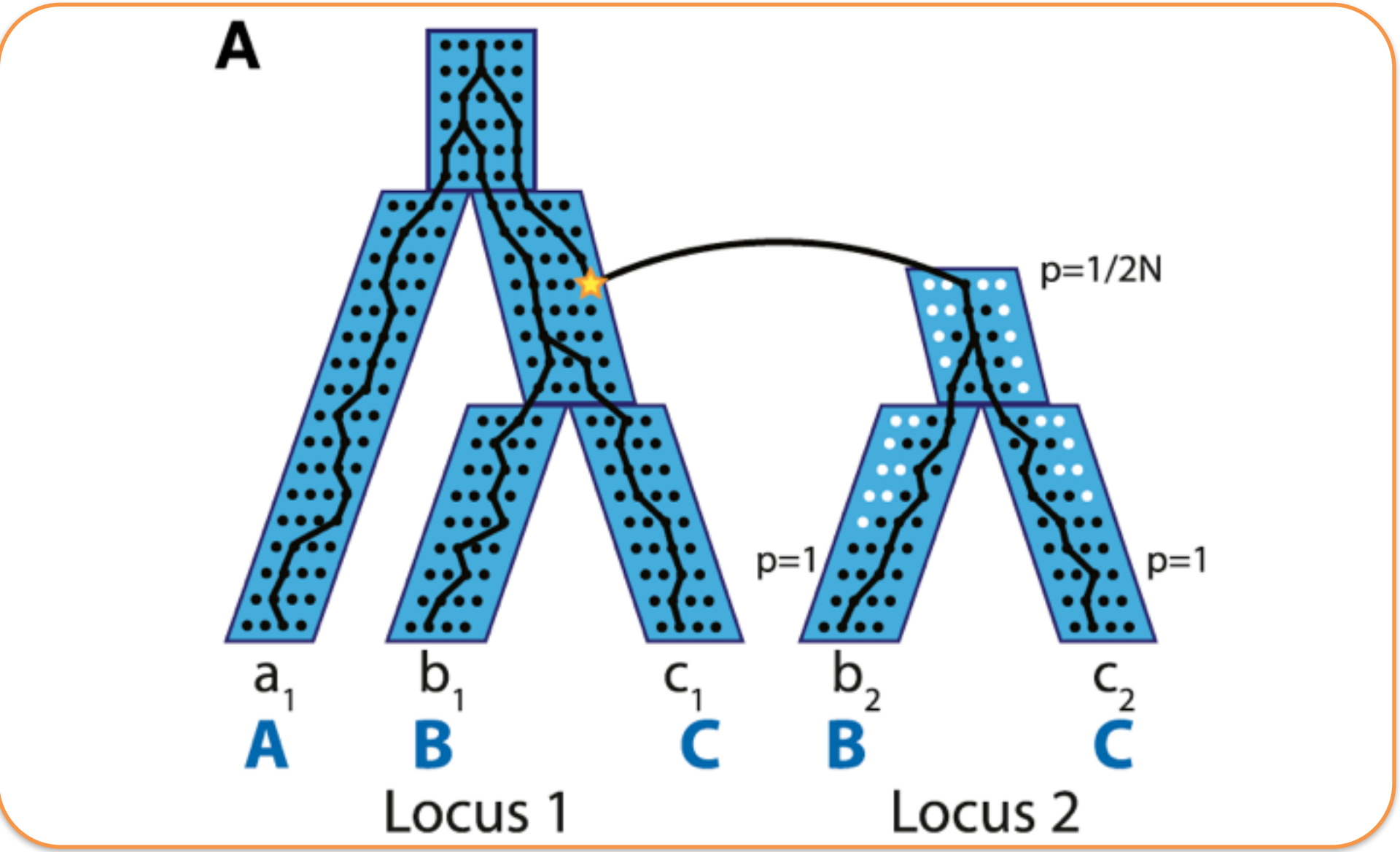
Coalescent model

Gene tree

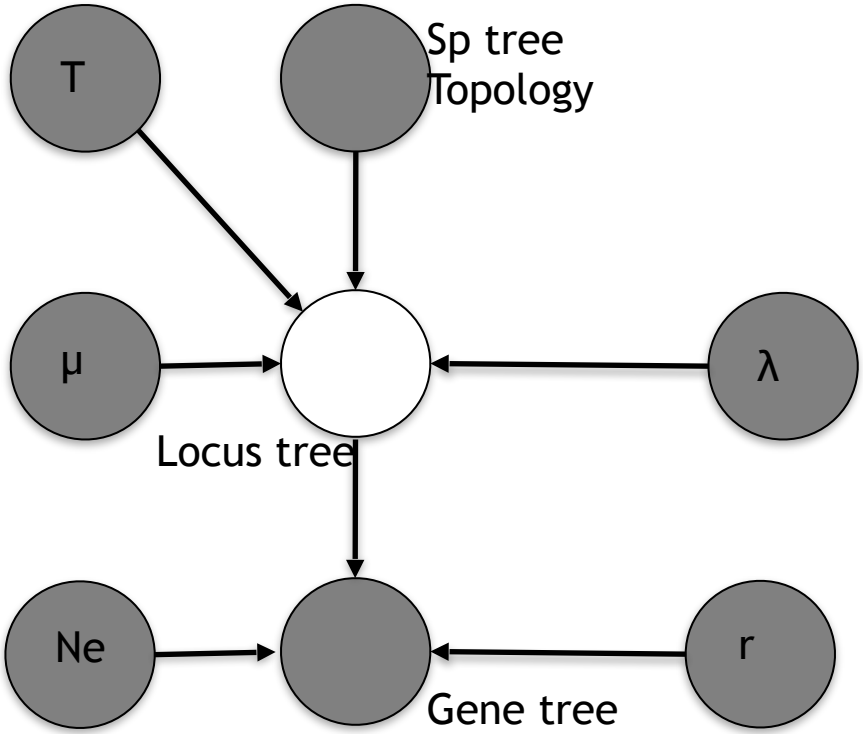
The Rasmussen and Kellis model (2012)

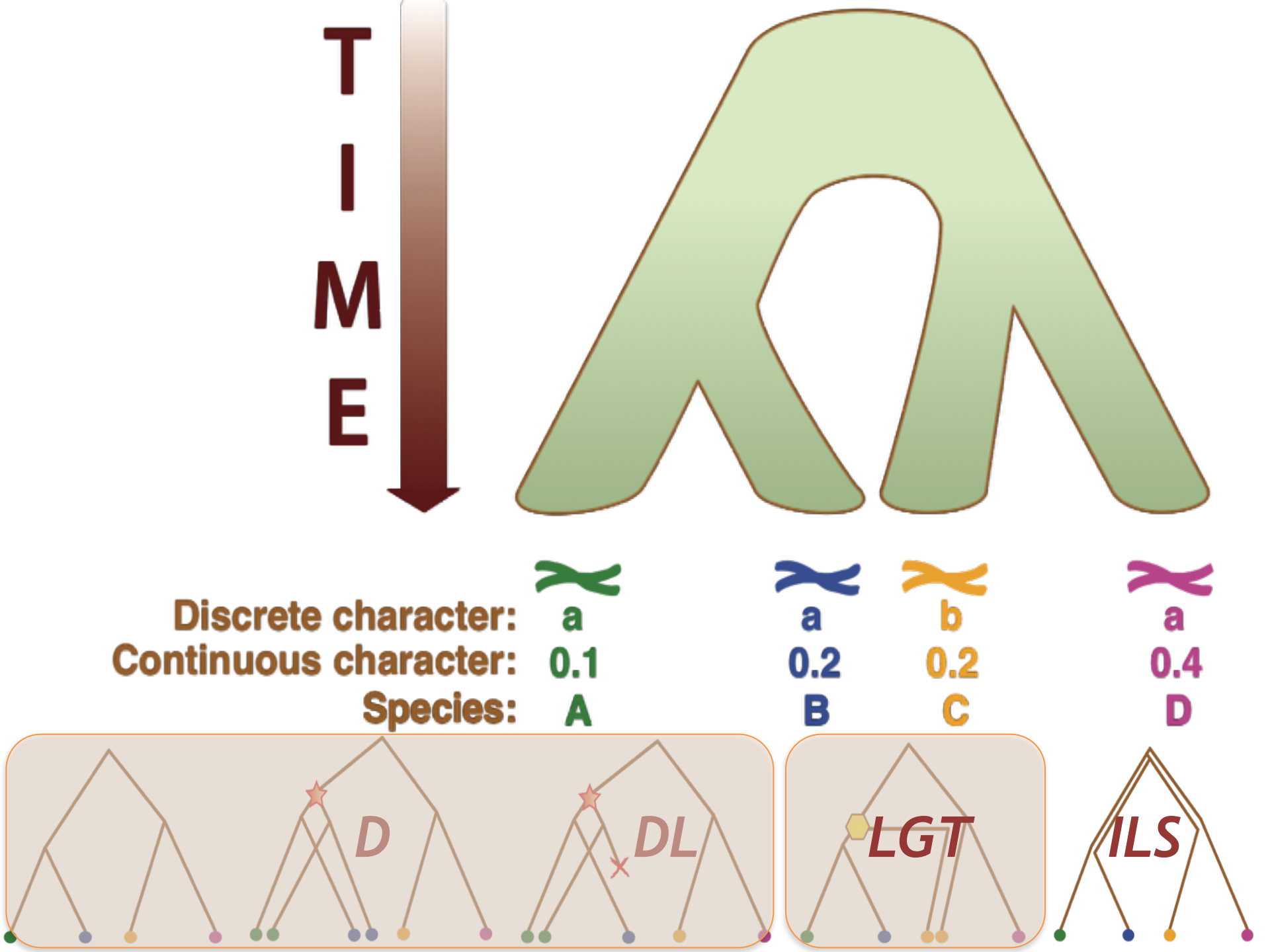


The Rasmussen and Kellis model (2012)



The species tree-gene tree graphical model with DL + ILS (Rasmussen and Kellis 2012)



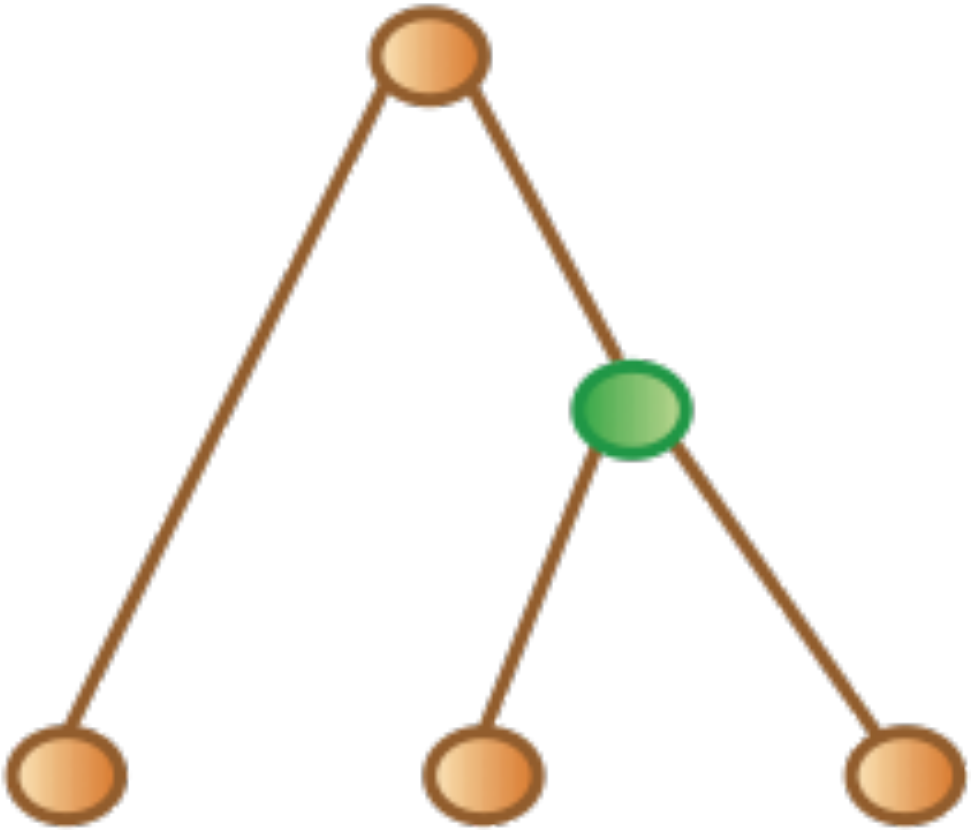

TIME
Discrete character:
Continuous character:
Species:
a
0.1
A
a
0.2
B
b
0.2
C
a
0.4
D
D
DL
LGT
ILS

Model that combines DL and T

- **Reconciliation between a species tree and a gene tree:**
  Parsimony setting: Gorecki, RECOMB 2004; Gorbunov and Lyubetski Mol. Biol. (Mosk), 2009; Libeskind-Hadas and Charleston, JCB 2009; Tofigh et al., IEEE ACM 2011; Doyon et al., Comparative Genomics 2011

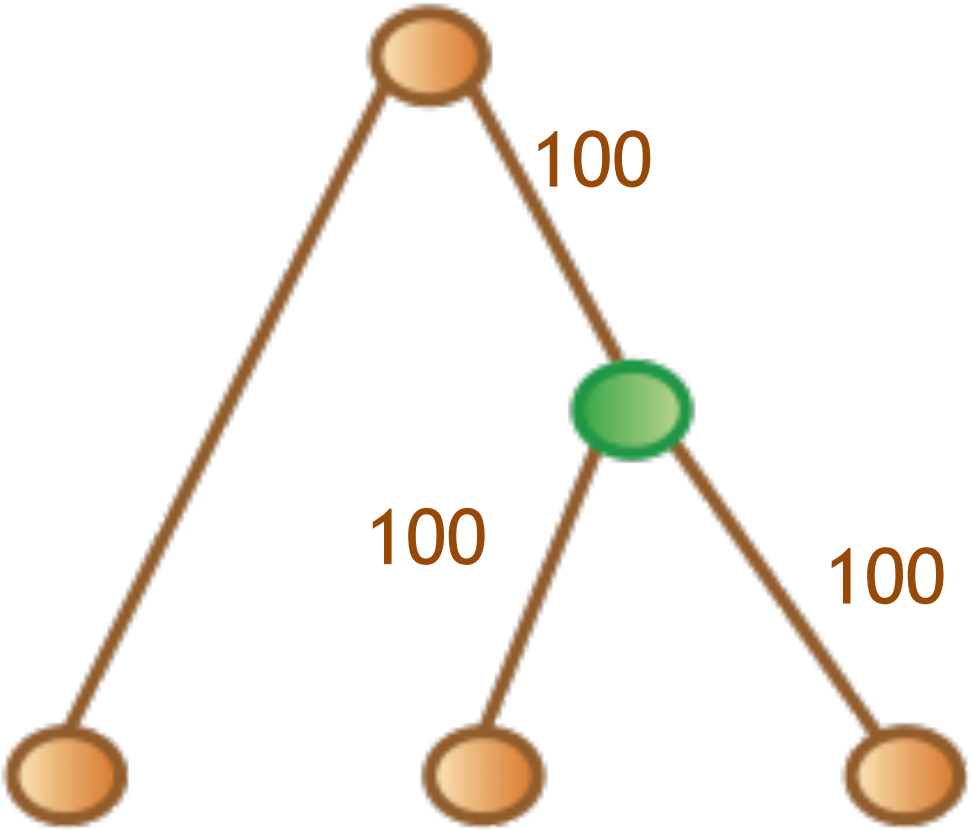
- **Reconstruction of a species tree given gene trees:**
  - Model-based method: Szöllősi et al., *PNAS* 2012; Szöllősi et al., Systematic Biology 2013a, 2013b; Sjöstrand et al., Syst Biol 2014

Dating the tree of life is difficult



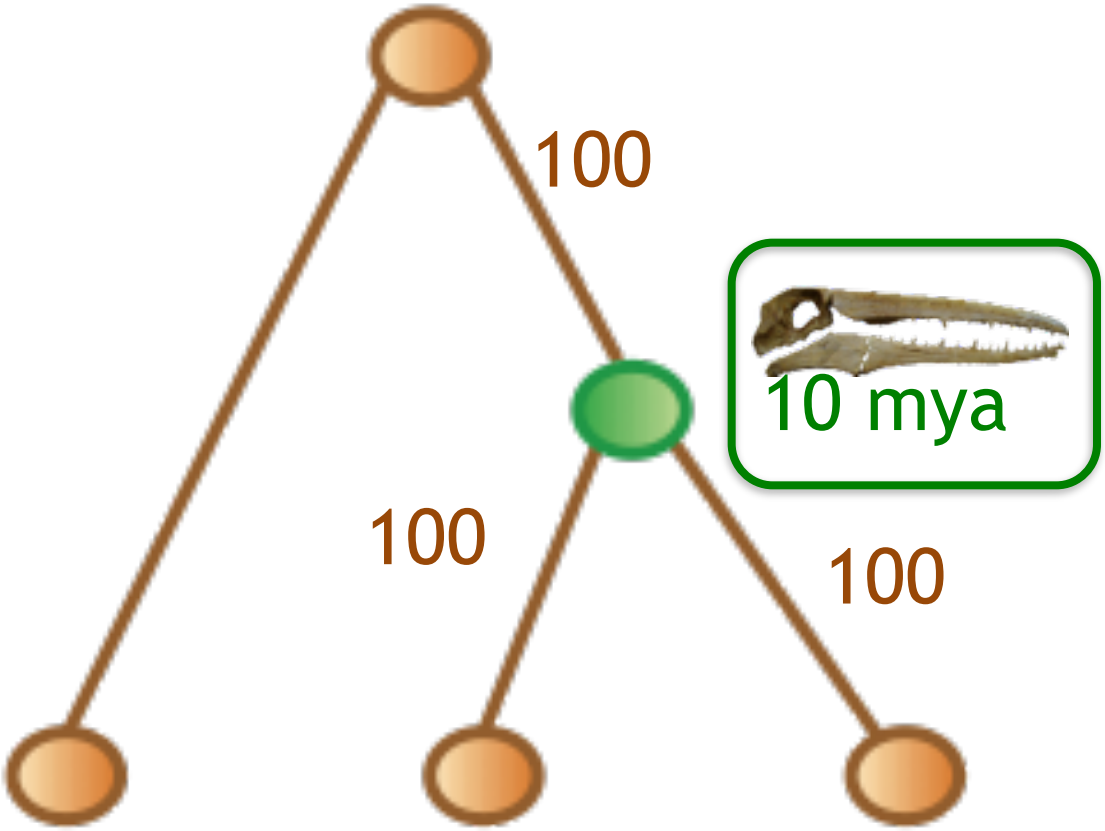
Species:

Dating the tree of life is difficult



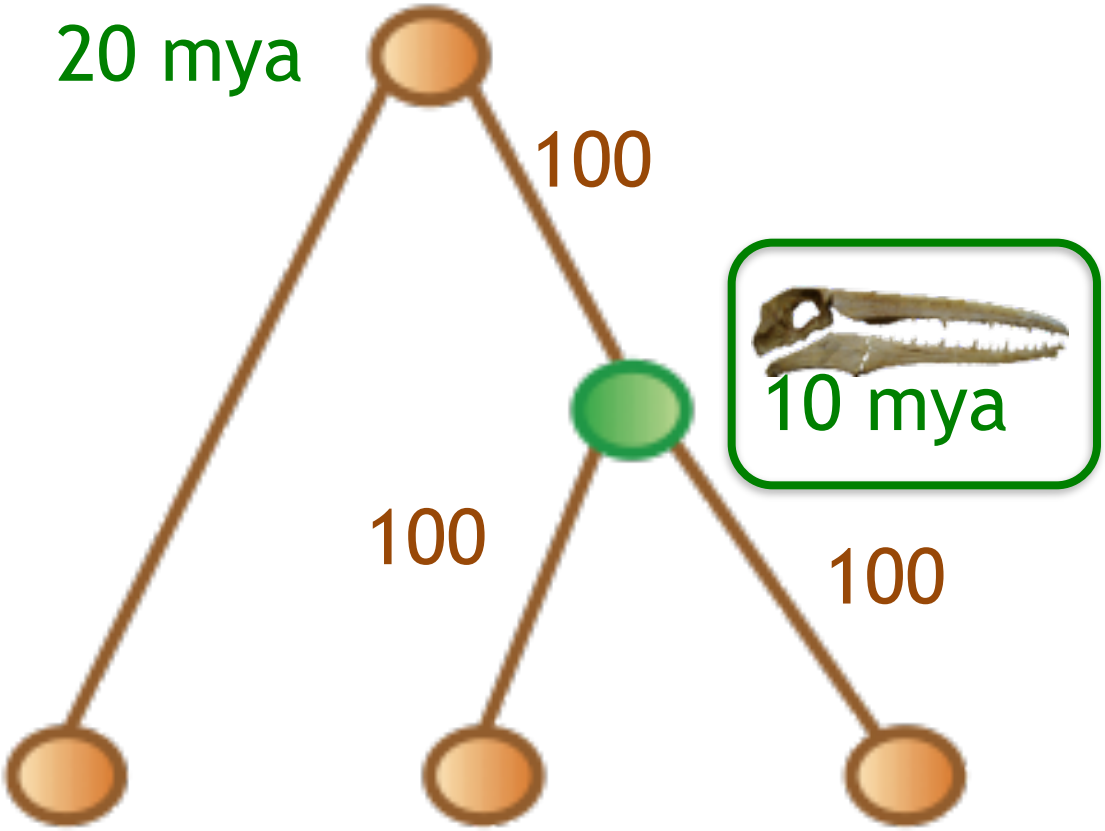
Species:

Dating the tree of life is difficult



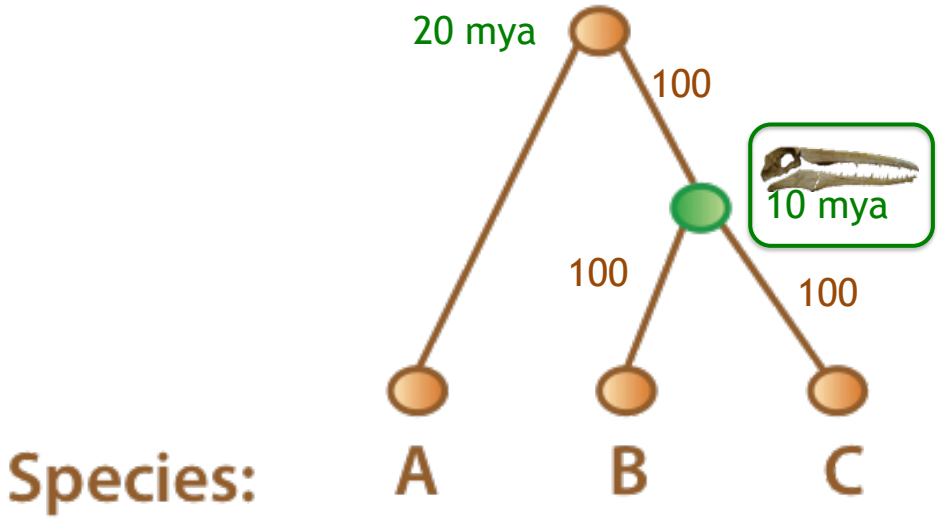
Species:

Dating the tree of life is difficult



Species:

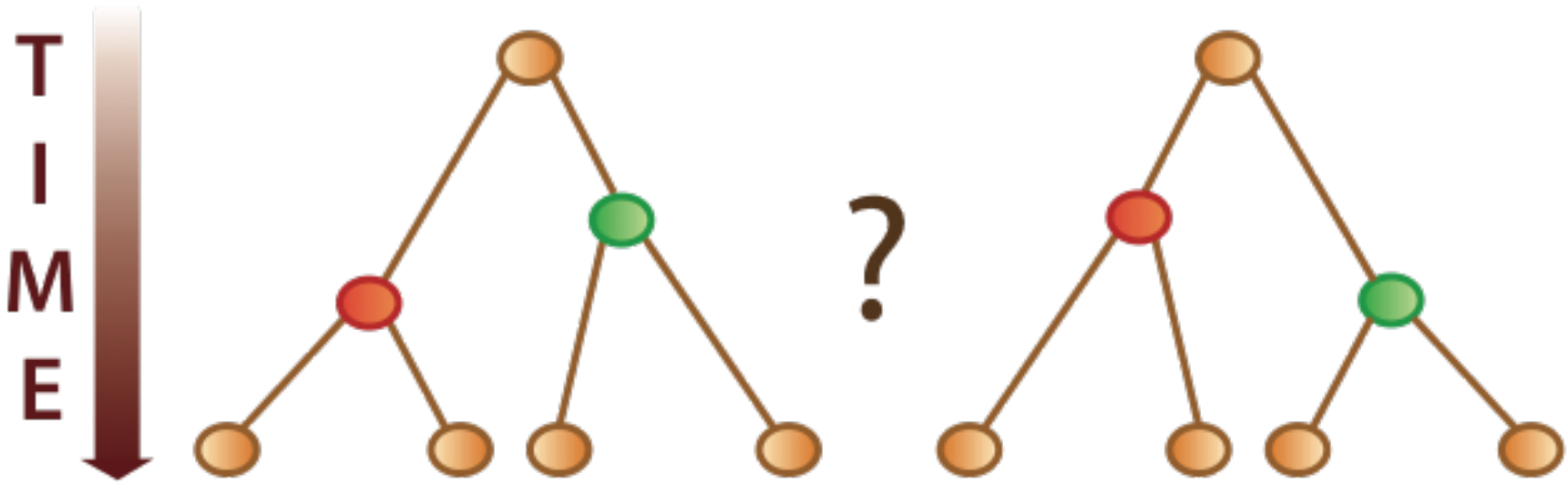
Dating the tree of life is difficult



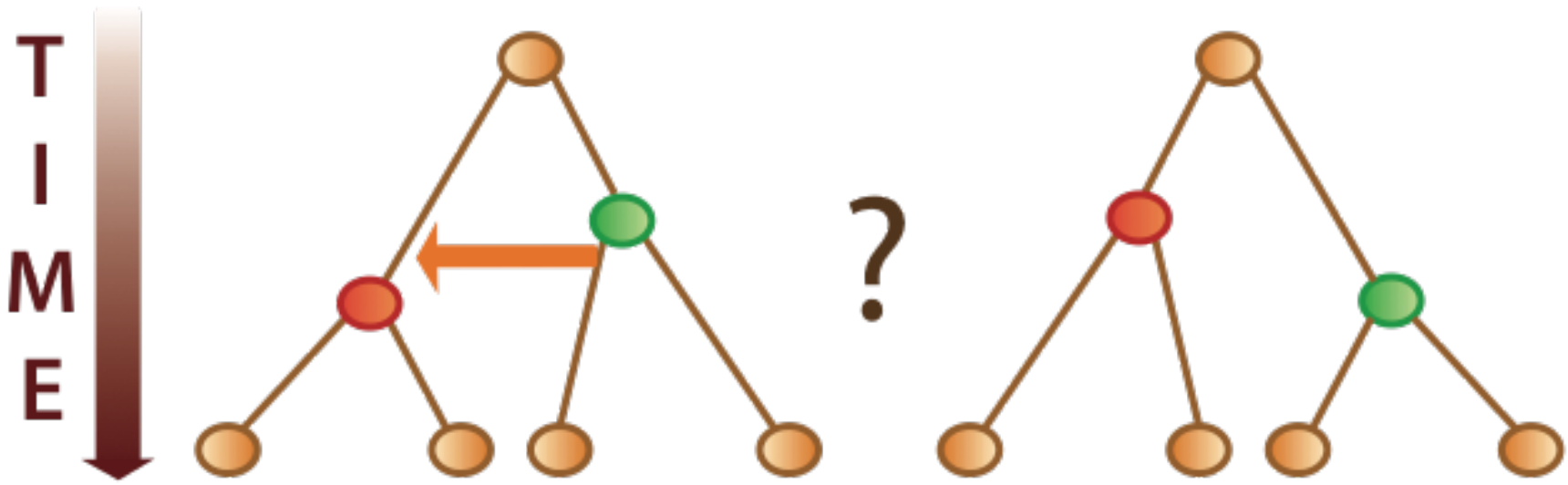
Very few fossils for the first 3 billion years of evolution on Earth

Other constraints are needed to date the tree of life

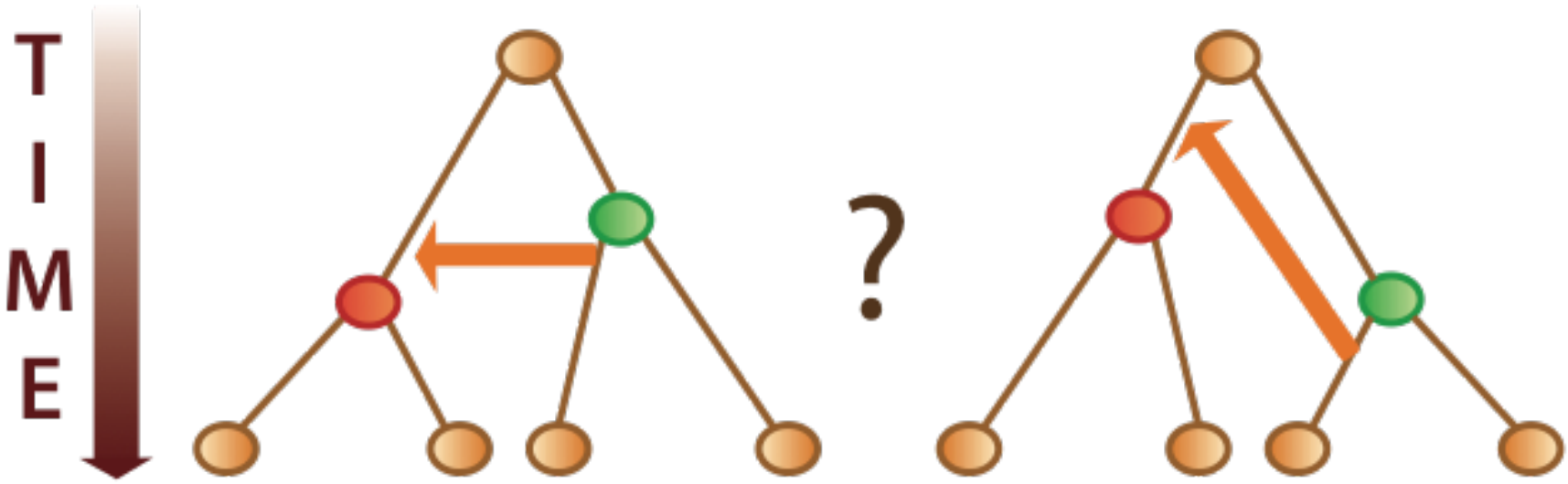
Using transfers to date clades



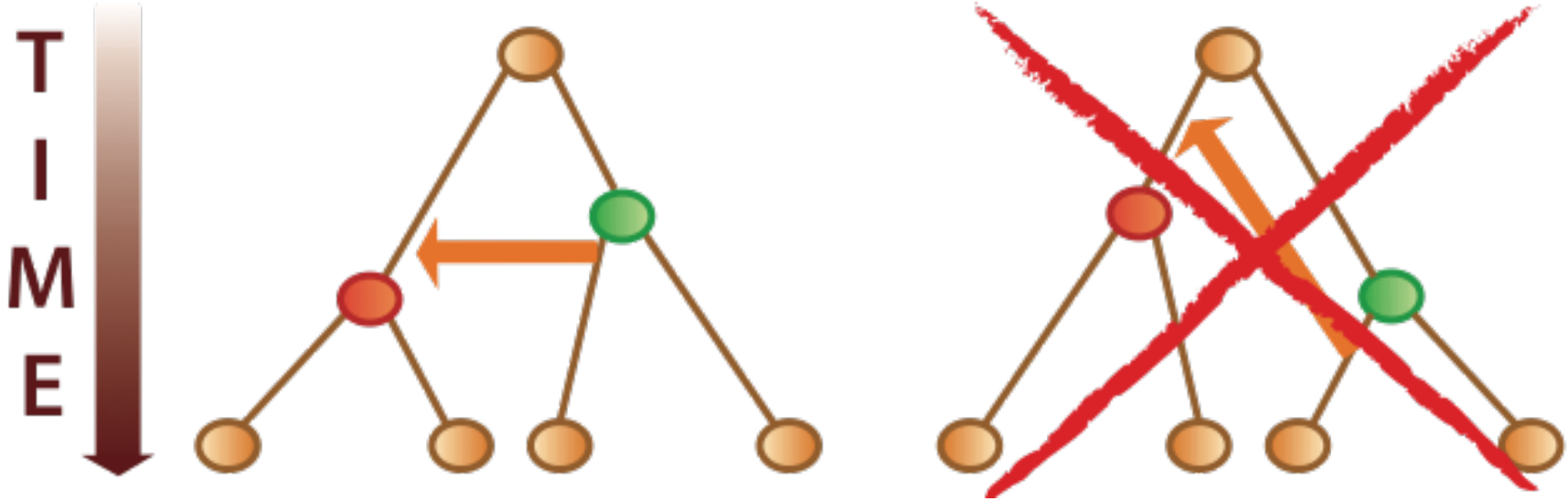
Using transfers to date clades



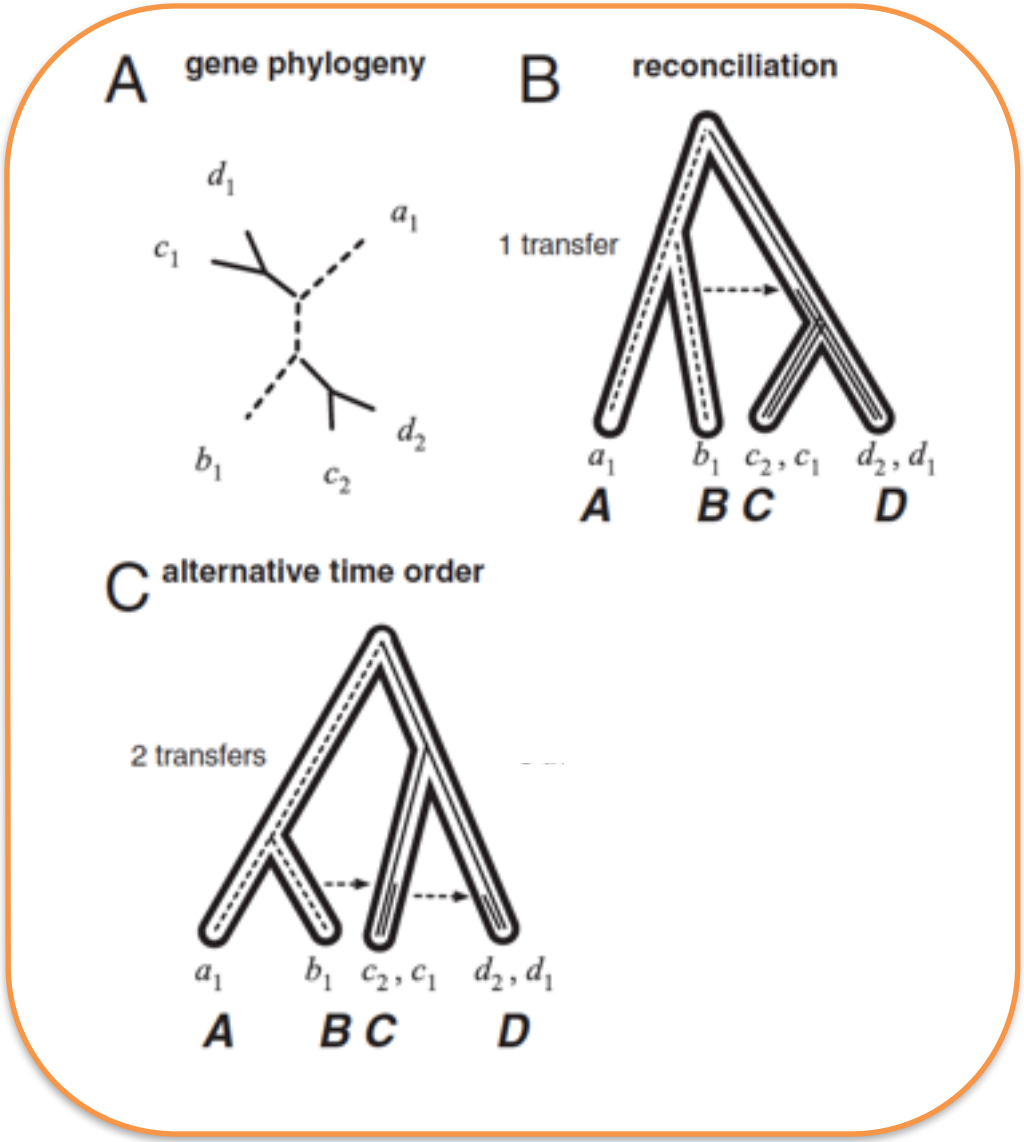
Using transfers to date clades



Using transfers to date clades

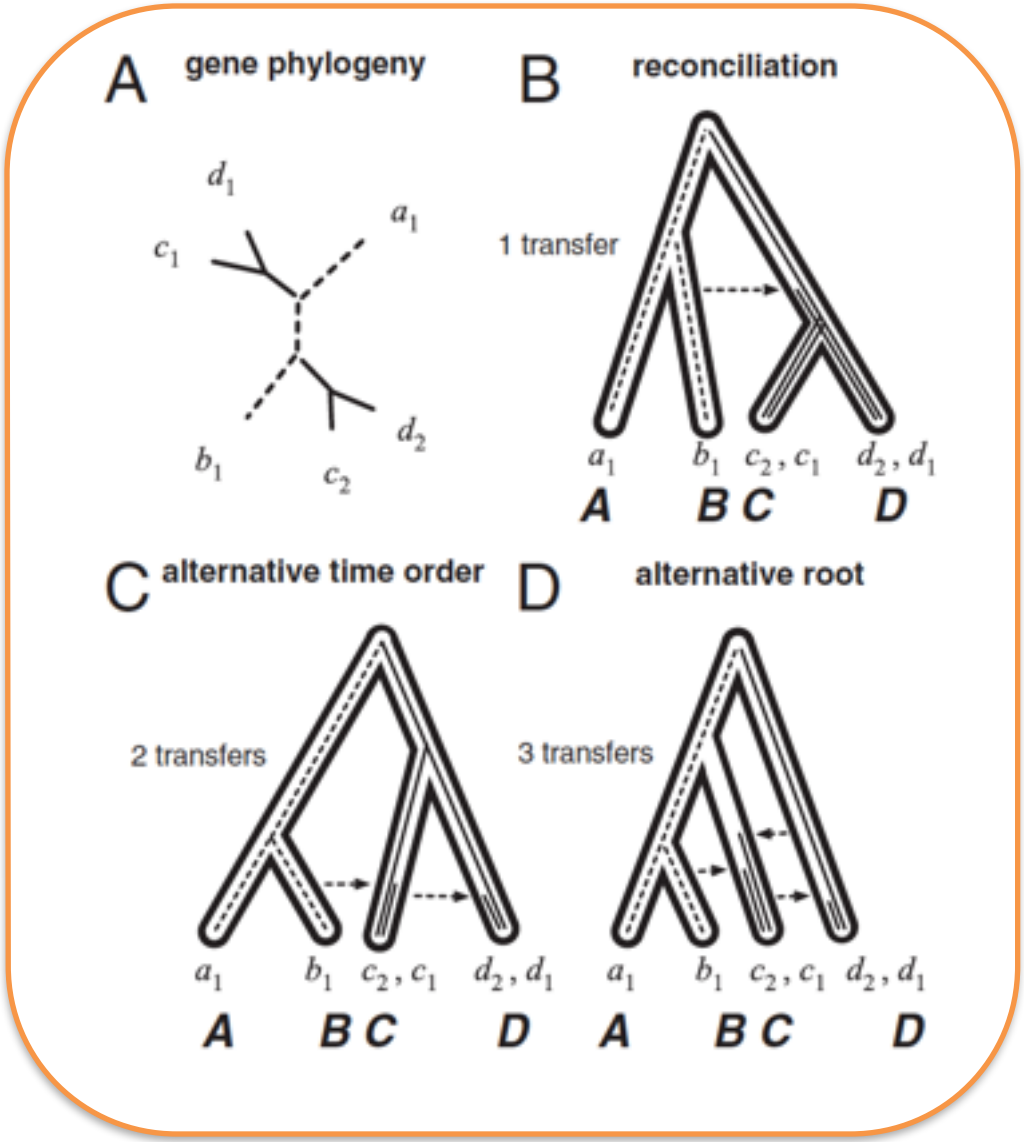


Gene transfers provide means to root and date species trees



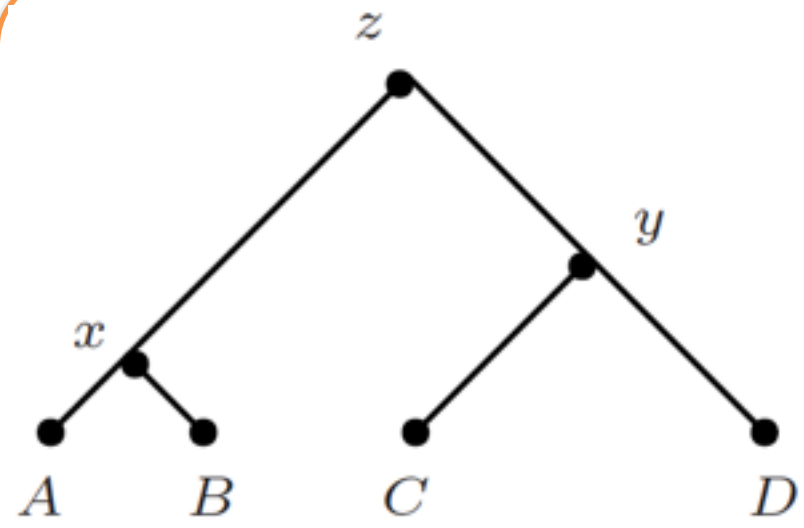
Szöllősi et al., PNAS 2012

Gene transfers provide means to root and date species trees

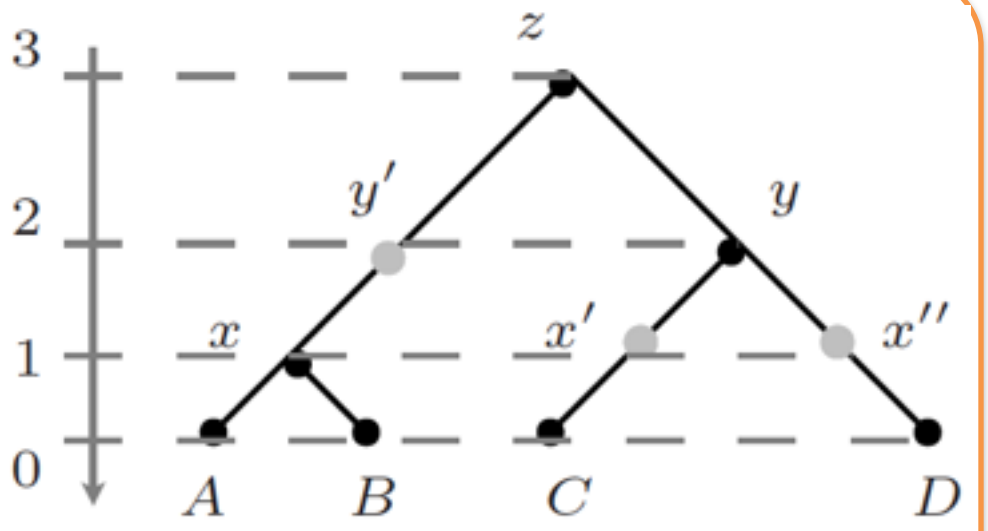


Szöllősi et al., PNAS 2012

Avoiding back to the future scenarios



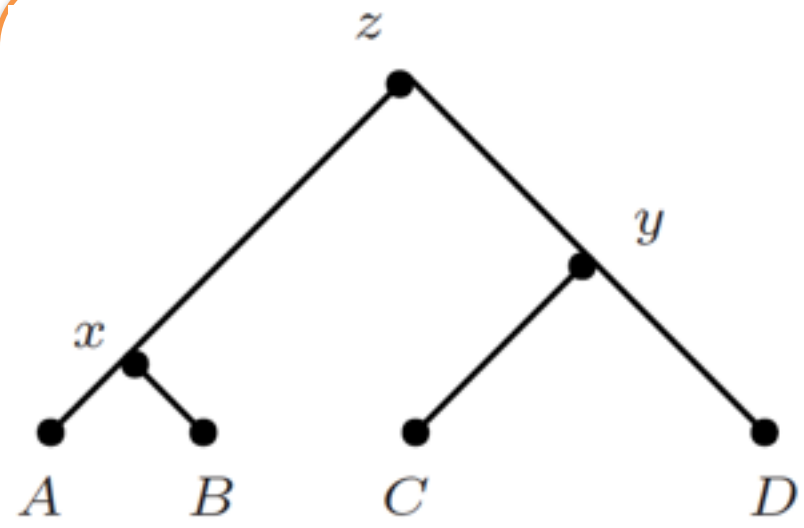
(a) A species tree S



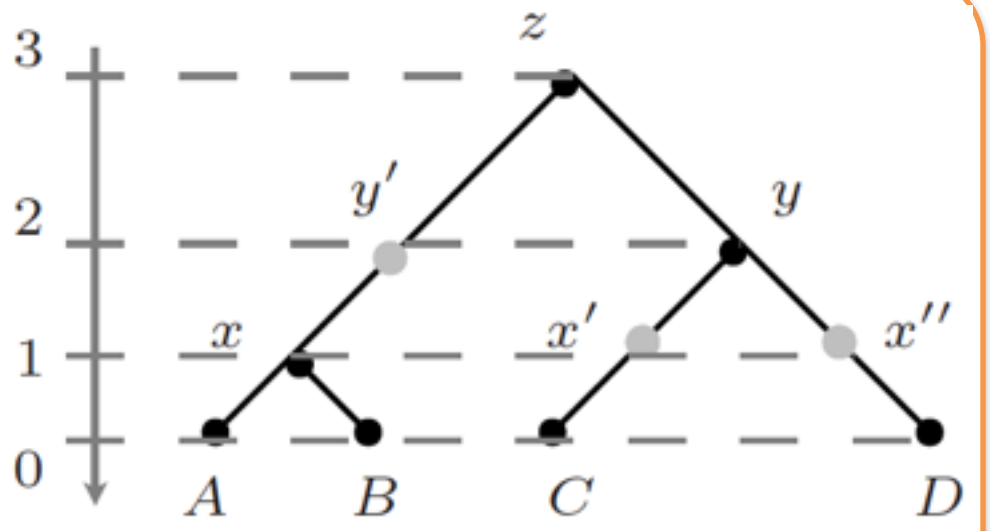
(b) The subdivision S' of S

Doyon et al., Comparative Genomics 2011

Avoiding back to the future scenarios



(a) A species tree S

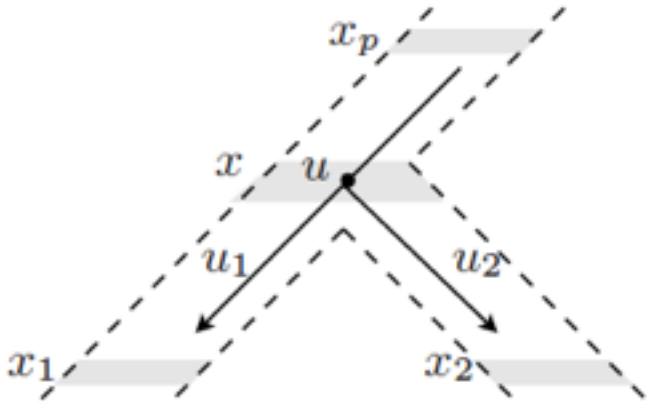


(b) The subdivision S' of S

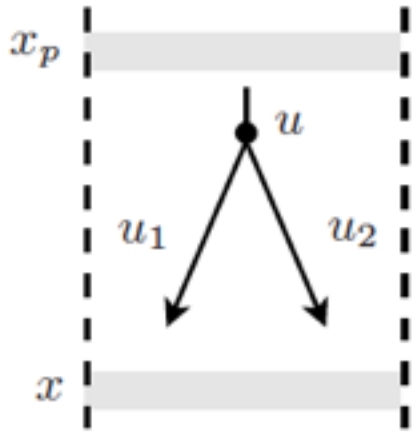
➔ *Algorithm in O (#NodesInS' . #NodesInG)*

Doyon et al., Comparative Genomics 2011

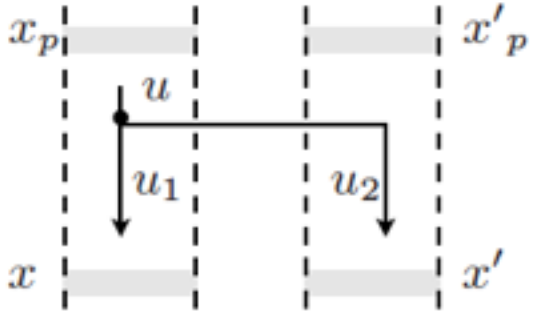
6 possible events are enough



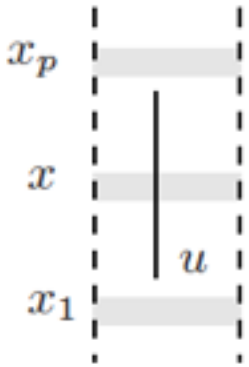
(a) Speciation ($\mathbb{S}$)



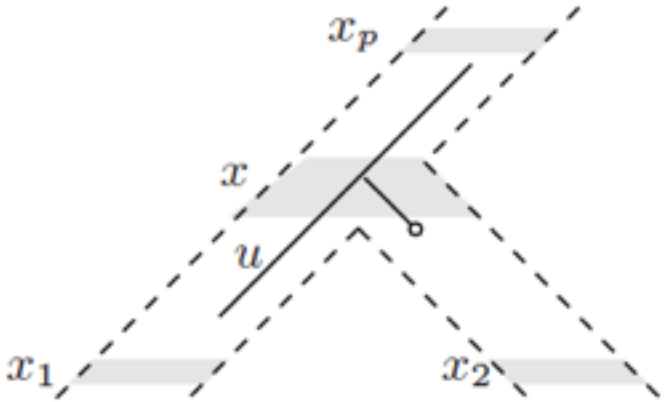
(b) Duplication ($\mathbb{D}$)



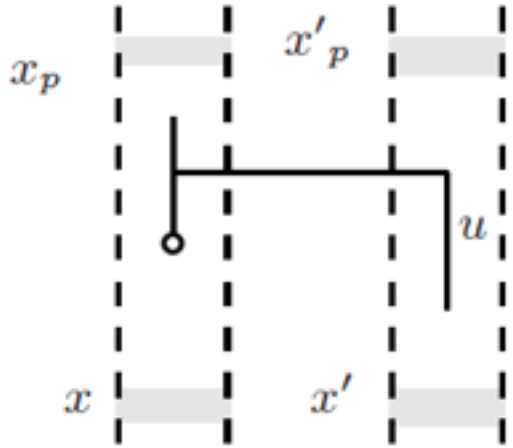
(c) Transfer ($\mathbb{T}$)



(d) $\varnothing$ event



(e) Speciation + Loss ($\mathbb{SL}$)



(f) Transfer+Loss ($\mathbb{TL}$)

Doyon et al., Comparative Genomics 2011

Dynamic programming in the Doyon et al. 2011 approach

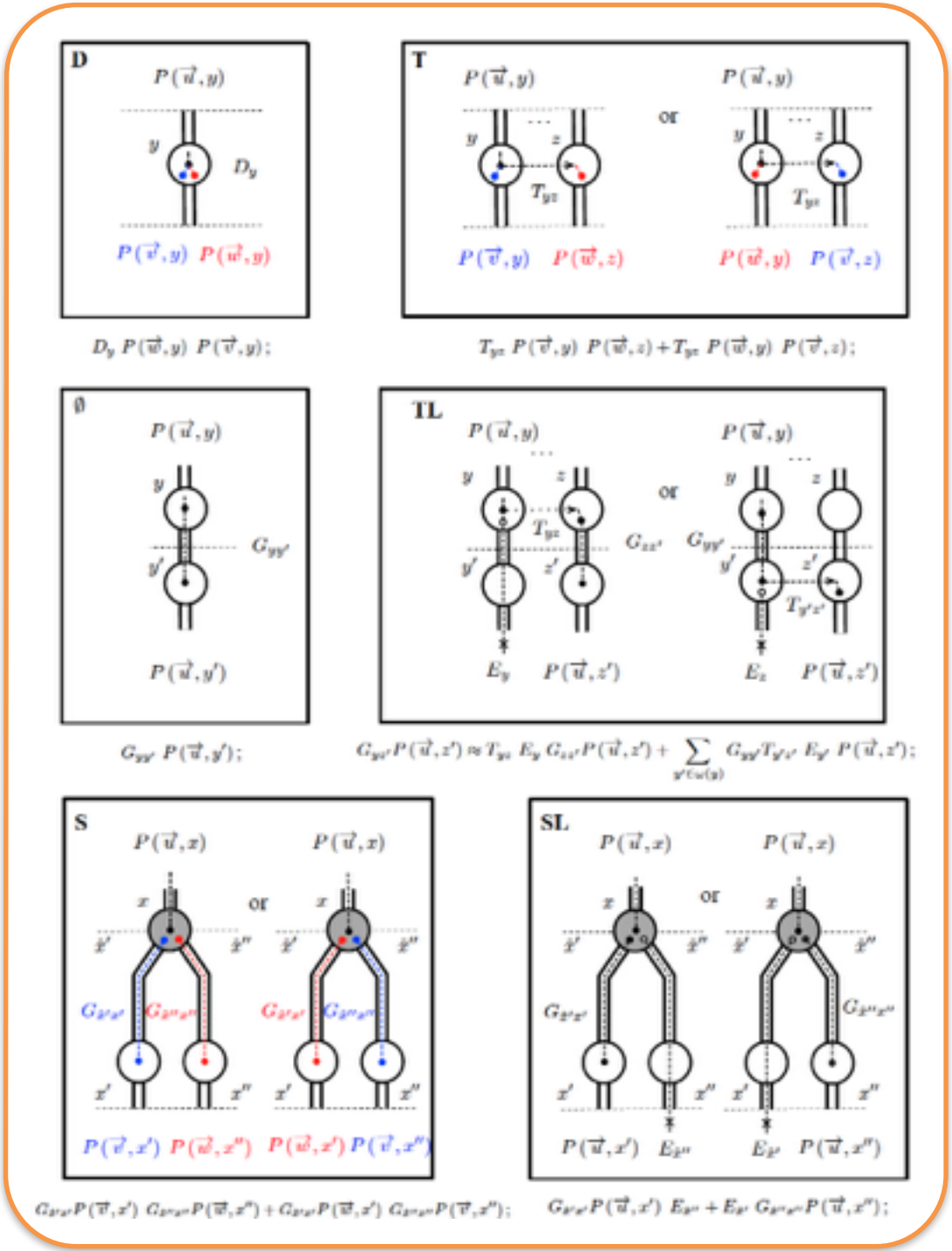
- Input: rooted augmented species tree S', rooted gene tree G
- Output: reconciliation scenario between S' and G
- Finding scenarios can be done in polynomial time using DP:
  - Fill the cost matrix c (#NodesInG, #NodesInS')
  - Post-order traversal of the gene tree
  - For each node i of G:
    - Map it to each node j of S'
    - Consider all of 6 events, and set c(i, j)=min(cost(possible events))
  - At the root, return the minimum cost found

From parsimony to model-based

- Input: gene trees
- Output: reconciled rooted gene trees, rooted species tree with ordered nodes
- Birth-death process to model gene evolution:
  - Birth parameters:
    - D
    - T: rate of receiving a gene through transfer
  - Death parameter:
    - L
- One or more sets of D, T, L parameters
- Double recursive DP for computing P(gene tree|species tree)
- Integrating over all possible scenarios
- Sliced species tree

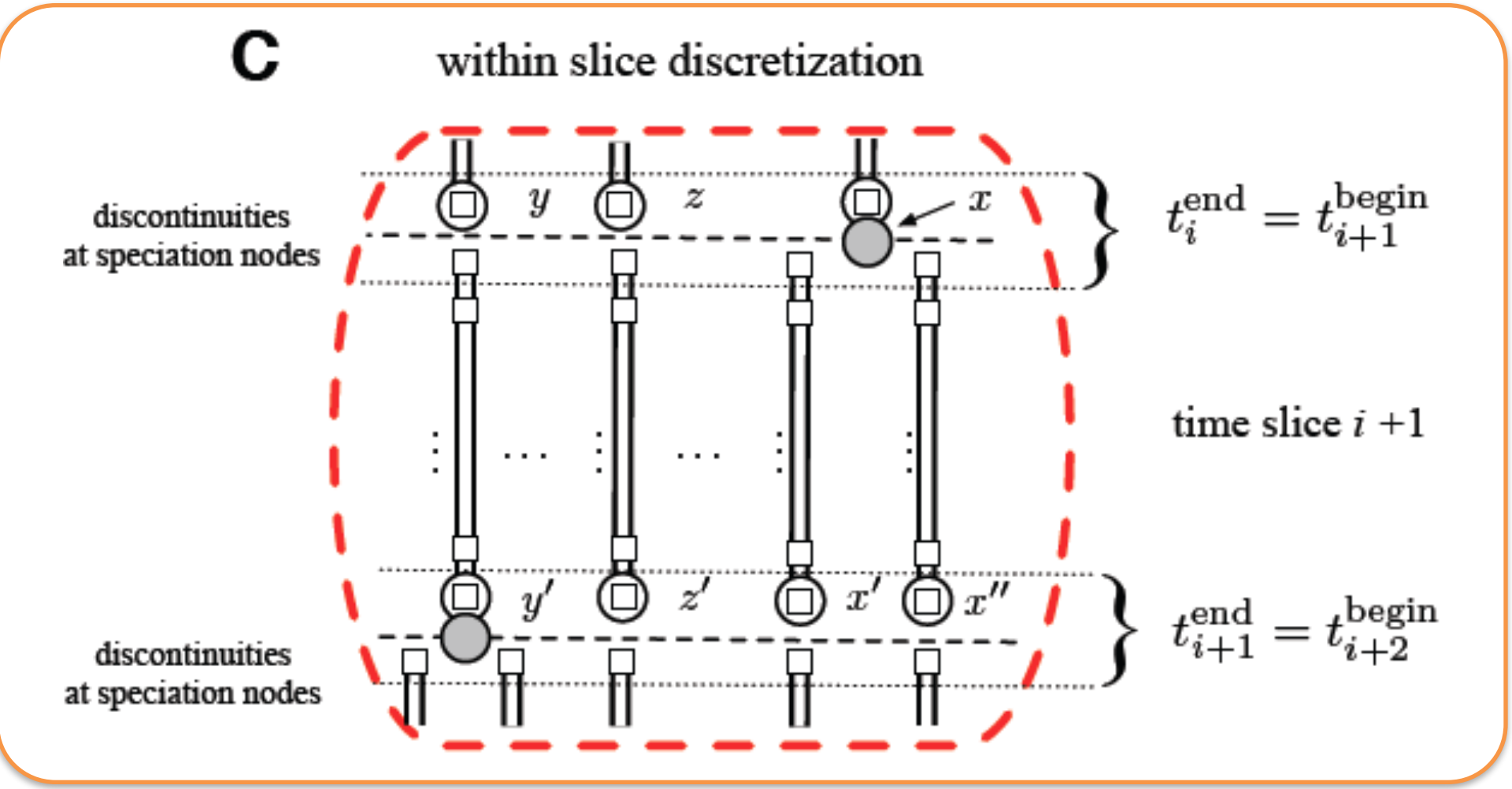
Szöllősi et al., PNAS 2012

Use of the 6 events from Doyon et al. 2011



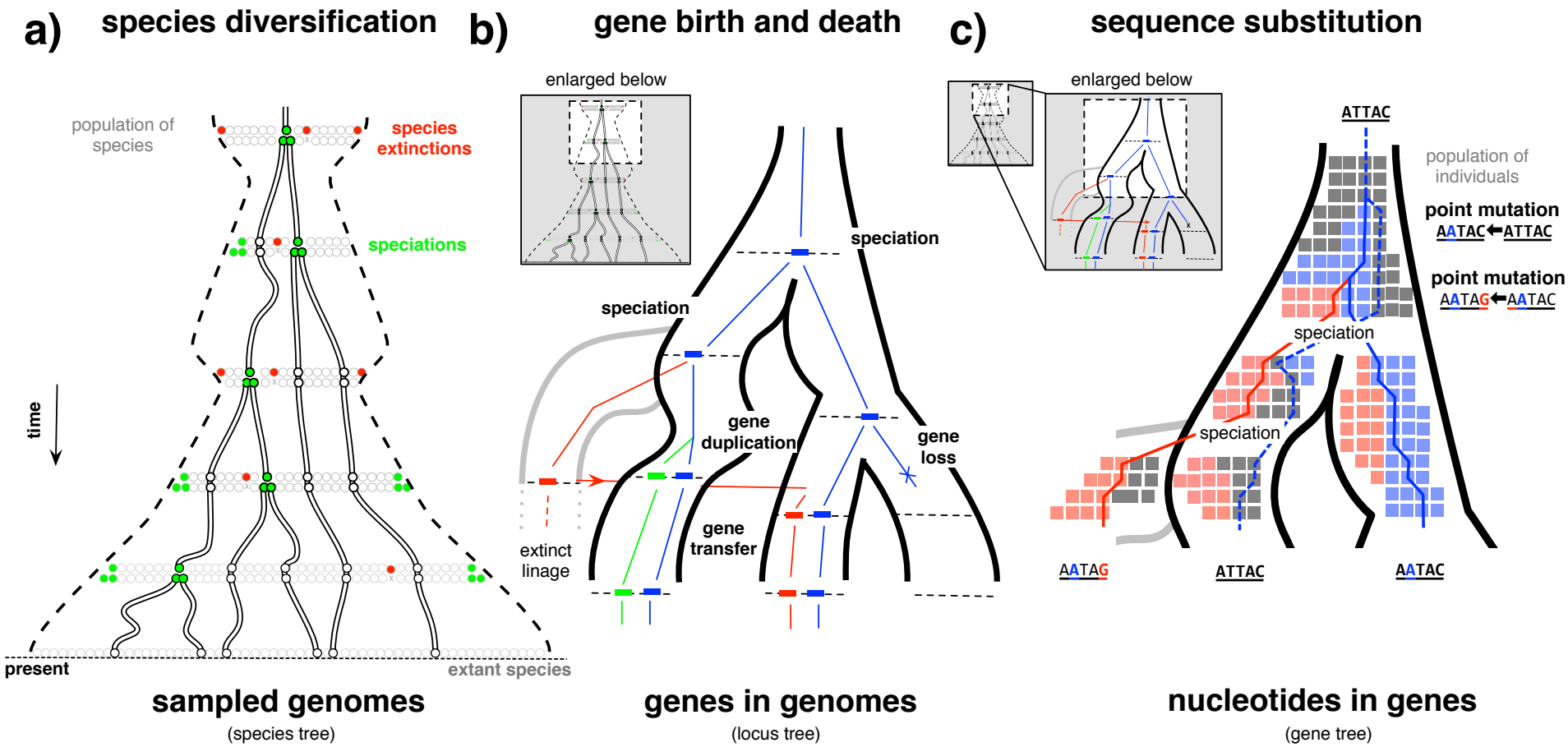
Szöllősi et al., PNAS 2012

Within-slice discretization for solving ordinary differential equations



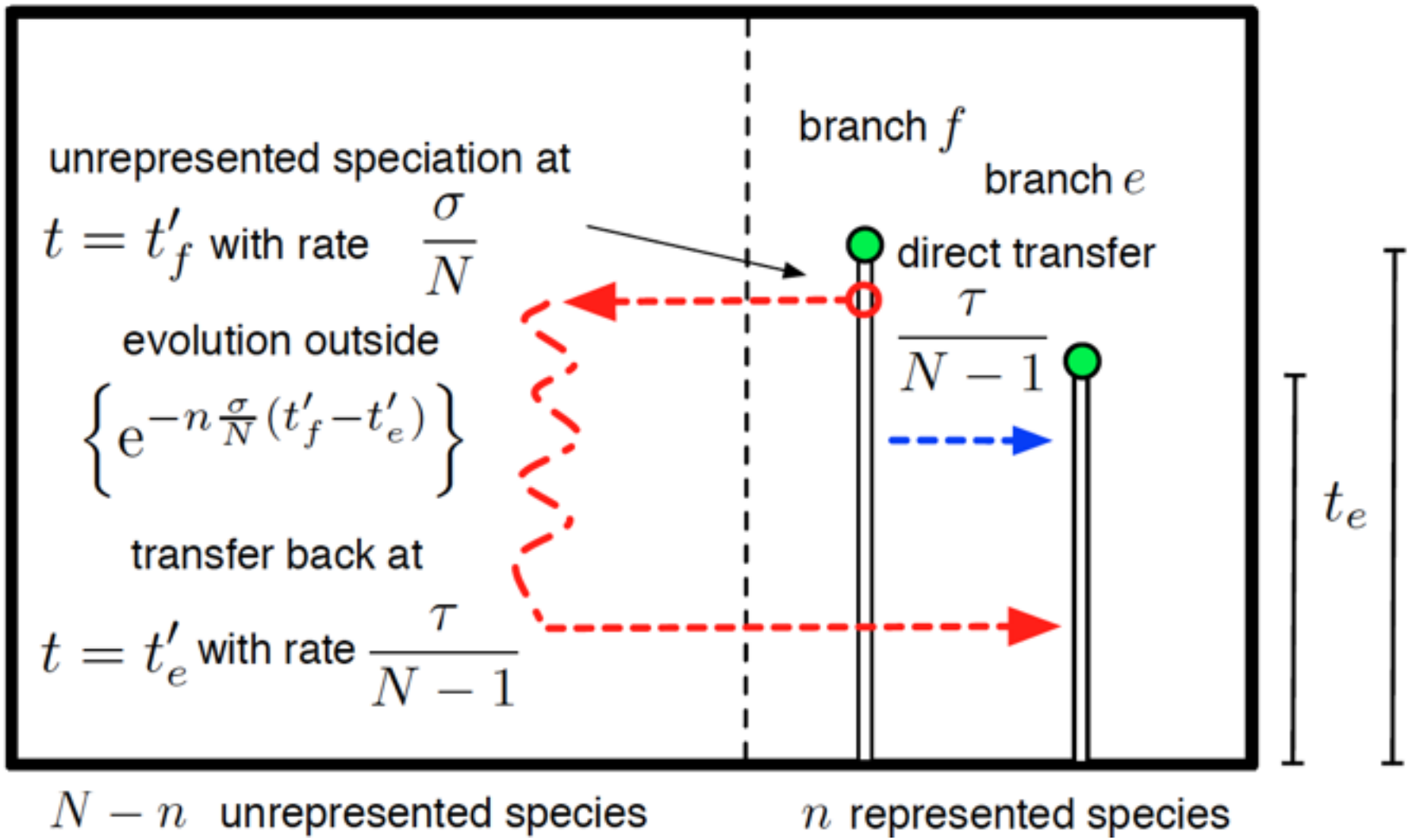
Szöllősi et al., PNAS 2012

A hierarchy of processes



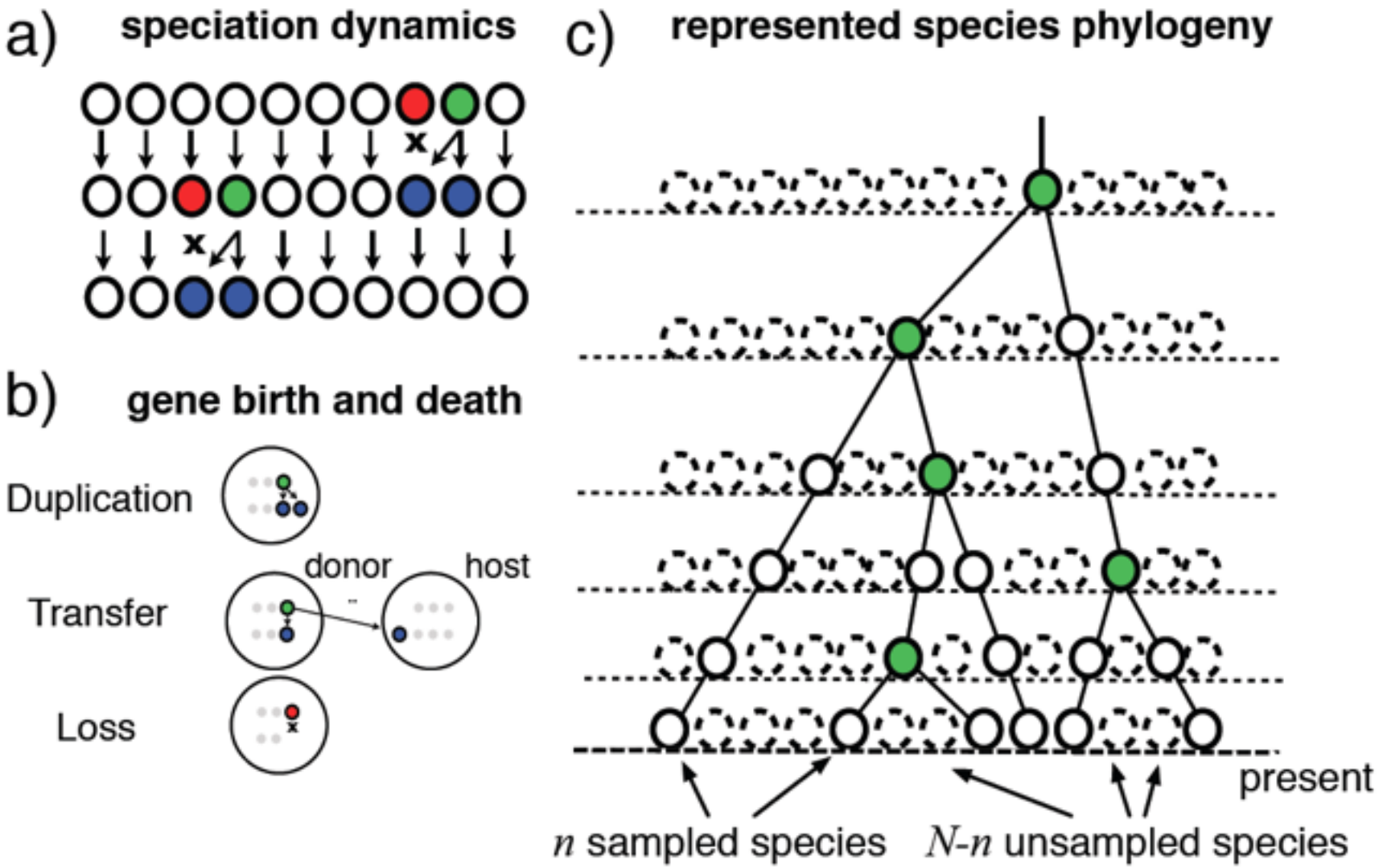
Szöllősi *et al., Syst Biol 2014*

Further refinement: seeing dead lineages



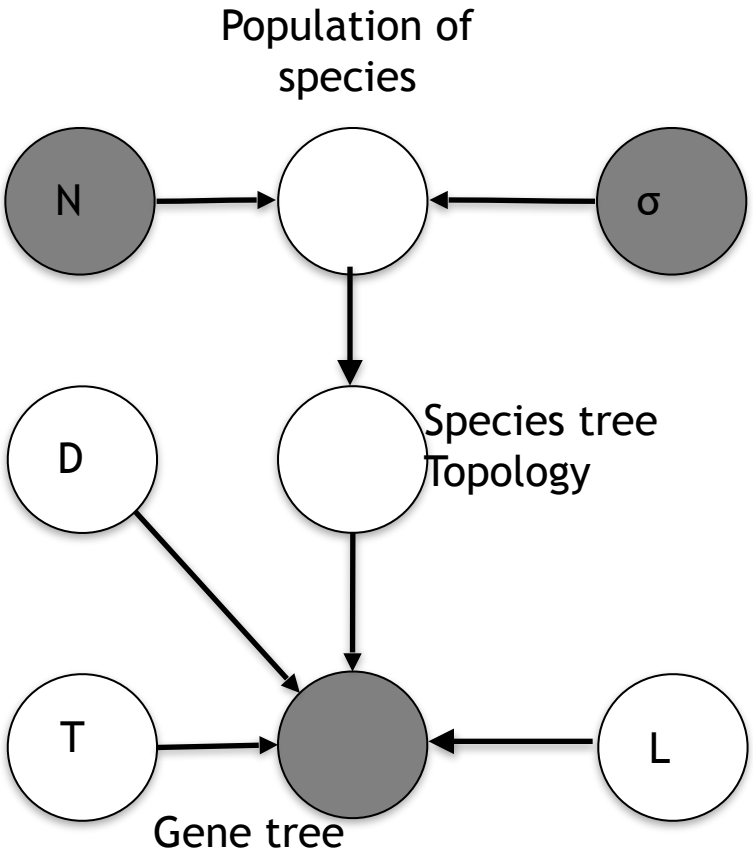
Szöllősi et al., Syst. Biol 2013

Further refinement: seeing dead lineages



Szöllősi et al., Syst. Biol 2013

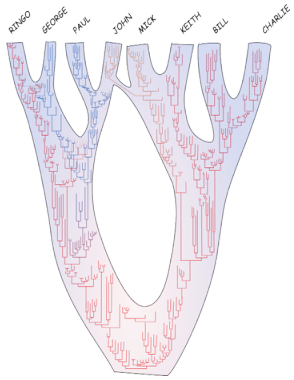
The exODT graphical model



Plan

Modeling the relationship between species tree and gene tree

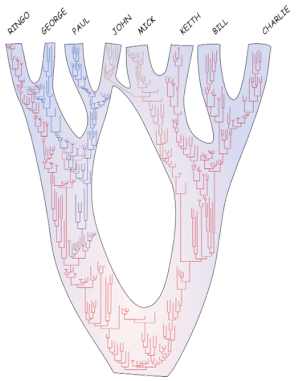
- coalescent models
- models of gene duplication and loss
- models of gene transfer
- models that combine the above



Gene tree species tree models in RevBayes, and tutorial

Gene tree species tree models in RevBayes

- Multispecies coalescent
- Pomo model
- In progress:
  - exODT (DTL model)
- Soon:
  - Akerborg et al. 2009 (DL model)
  - MP-EST



The tutorial

- A walk around the multispecies coalescent:

1. Simulation with the multispecies coalescent (play with the parameters!)
2. Then:
   - Inference with the multispecies coalescent from gene trees
   - Inference with the multispecies coalescent from gene alignments
   - Inference with the POMO model from gene alignments
   - Inference by concatenating the gene alignments

The tutorial

Files are in:

tutorials/RB_GeneTreeSpeciesTree_Tutorial/RB_tutorial_files/RevBayes_scripts/

in five folders. **It is highly recommended to start RevBayes from each of these folders when doing the exercises to avoid path problems.**

There is one simulation-only folder, used for better understanding the multispecies coalescent:

UnderstandingMultiSpeciesCoalescent

Then there are 4 inference folders.

In each inference folder, there are 3 scripts. The file XXXMCMC.Rev is the one that can launch the entire analysis, which consists of three parts:

Simulating the data

Constructing the model for inference

Preparing and running the MCMC