

Coalescent Methods for Phylogenomic Inference Lab

Erin Molloy

Department of Computer Science
University of Maryland, College Park

Workshop on Phylogenomics @ Cesky Krumlov
January 25, 2024

Acknowledgements

Some slides have been adapted from **Laura Kubatko**'s coalescent lab in 2019; I encourage you to read her slides for more information.



Some slides include images by the individuals below; I encourage you to check out these papers (click on links).



James
Degnan (et al)



Frederik
Leliaert (et al)



Siavash
Mirarab



Luay
Nakhleh



Tandy
Warnow

Coalescent Lab — Day 1

Motivation for coalescent methods (**Activity A**)

Coalescent basics (**Activity B**)

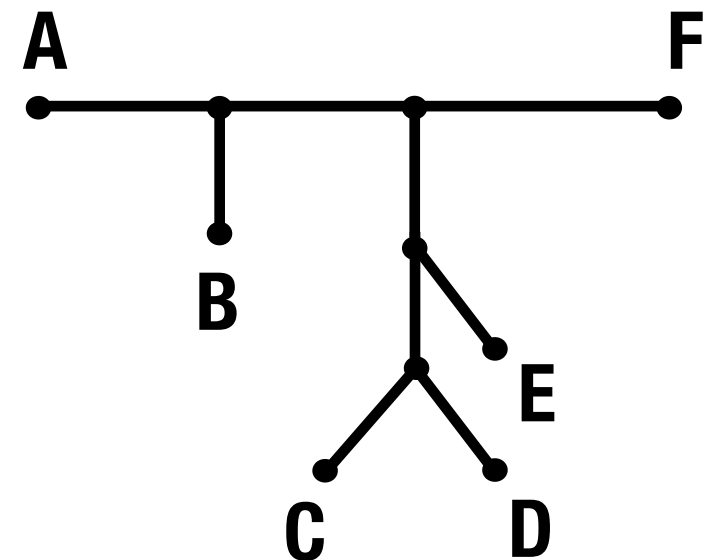
Species tree estimation with summary methods (**Activity C**)

Evaluation model fit (**Activity D** — optional / do tomorrow)



Input:

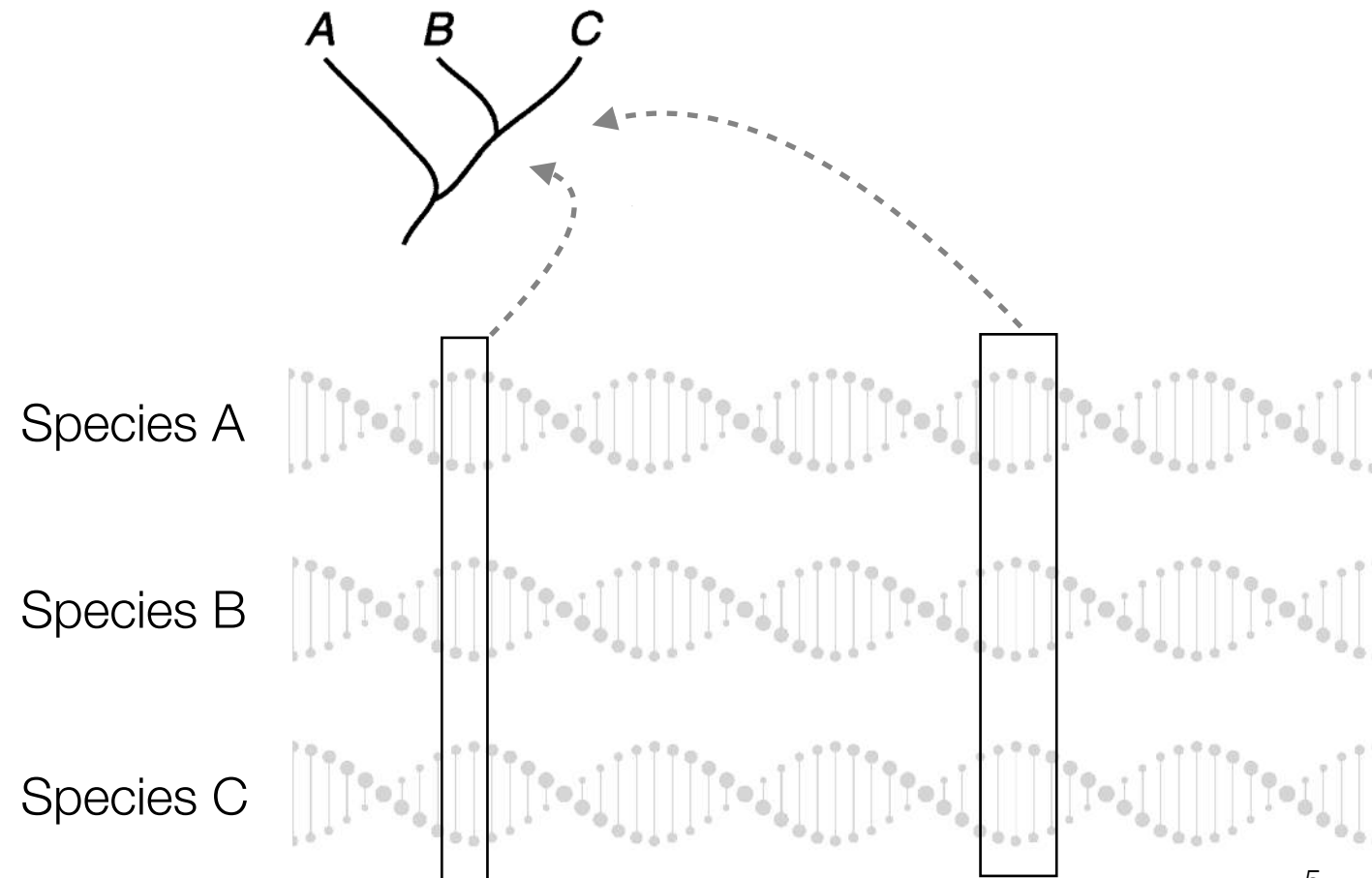
Data matrix with n species & m genes (alignments)



Output:

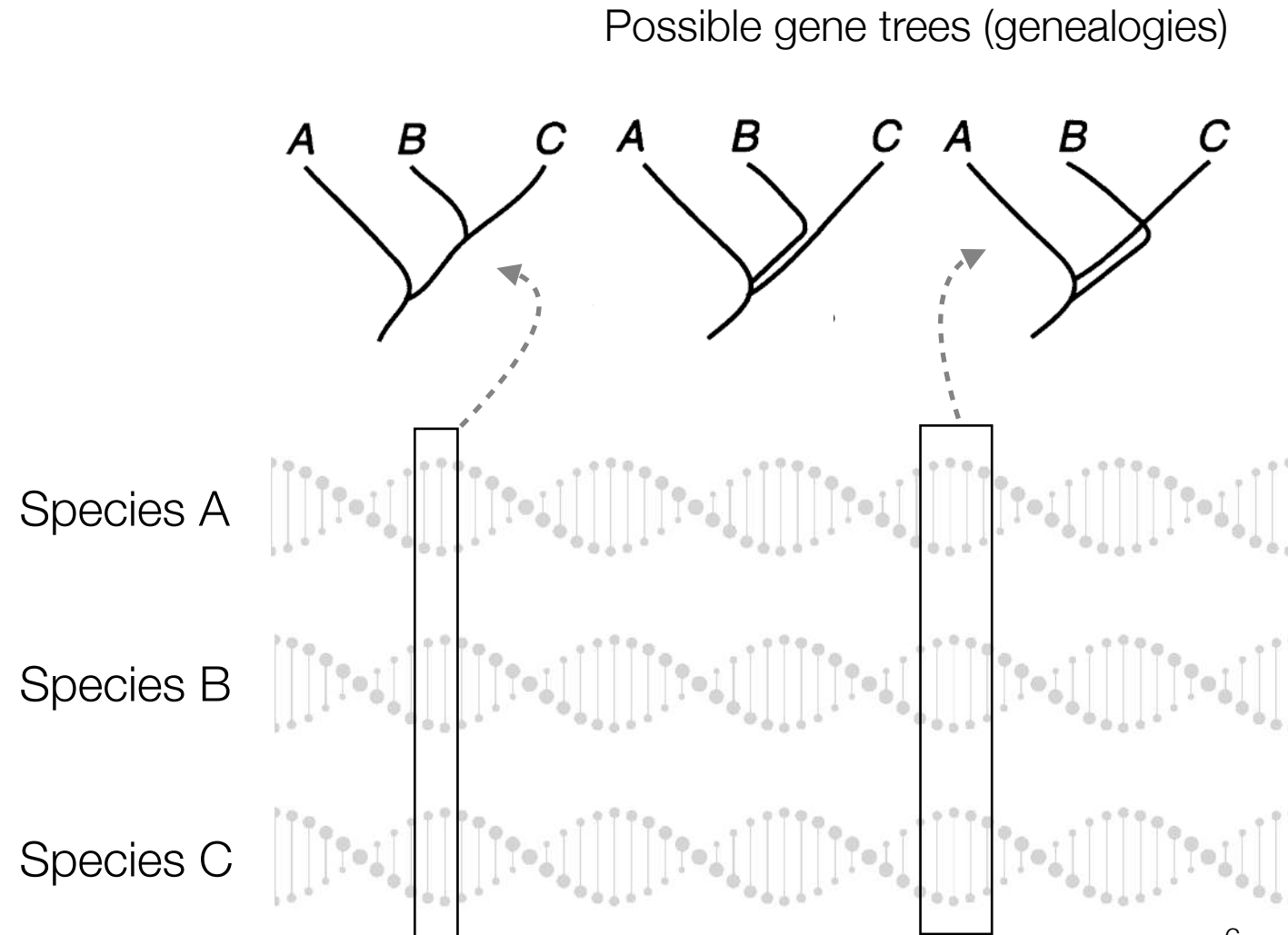
Unrooted, binary tree T with leaves labeled by n species

Standard molecular sequence evolution models assume all genomic regions evolve down same tree!



Standard molecular sequence evolution models assume all genomic regions evolve down same tree!

However, different regions of the genome can have different evolutionary histories!



GENE TREES IN SPECIES TREES

WAYNE P. MADDISON

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA

Abstract.—Exploration of the relationship between gene trees and their containing species trees leads to consideration of how to reconstruct species trees from gene trees and of the concept of phylogeny as a cloud of gene histories. When gene copies are sampled from various species, the gene tree relating these copies might disagree with the species phylogeny. This discord can arise from horizontal transfer (including hybridization), lineage sorting, and gene duplication and extinction. Lineage sorting could also be called *deep coalescence*, the failure of ancestral copies to coalesce (looking backwards in time) into a common ancestral copy until deeper than previous speciation events. These events depend on various factors; for instance, deep coalescence is more likely if the branches of the species tree are short (in generations) and wide (in population size). A similar dependence on process is found in historical biogeography and host–parasite relationships. Each of the processes of discord could yield a different parsimony criterion for reconstructing the species tree from a set of gene trees: with horizontal transfer, choose the species tree that minimizes the number of transfer events; with deep coalescence, choose the tree minimizing the number of extra gene lineages that had to coexist along species lineages; with gene duplication, choose the tree minimizing duplication and/or extinction events. Maximum likelihood methods for reconstructing the species tree are also possible because coalescence theory provides the probability that a particular gene tree would occur given a species tree (with branch lengths and widths specified). In considering these issues, one is provoked to reconsider precisely what is phylogeny. Perhaps it is misleading to view some gene trees as agreeing and other gene trees as disagreeing with the species tree; rather, all of the gene trees are part of the species tree, which can be visualized like a fuzzy statistical distribution, a cloud of gene histories. Alternatively, phylogeny might be (and has been) viewed not as a history of what happened, genetically, but as a history of what could have happened, i.e., a history of changes in the probabilities of interbreeding. [Biogeography; coalescence; coevolution; evolution; gene duplication; gene genealogy; gene trees; horizontal transfer; hybridization; lineage sorting; parsimony; phylogeny; species concepts; species trees; tree reconciliation.]

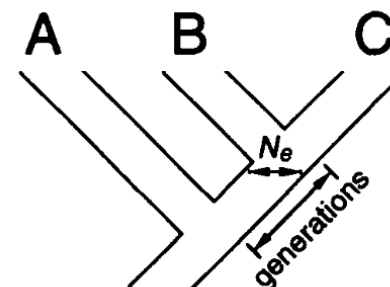
A phylogenetic tree of species contains smaller trees descending within its branches: the trees of genes. Recently, the relationship between gene trees and species trees has been the focus of some attention (e.g., Fitch, 1970; Goodman et al., 1979; Avise et al., 1983; Tajima, 1983; Pamilo and Nei, 1988; Takahata, 1989; Roth, 1991; Wu, 1991; Doyle, 1992; Hudson, 1992; Page, 1993; Baum and Shaw, 1995; Maddison, 1995, 1996). One aspect of this relationship is the congruence between the species tree and a tree of gene copies sampled from those species. Imagine that one gene copy was sampled from each species, and the gene tree relating these gene copies is examined. One might expect that two sister species would have sister copies in the gene tree and that other aspects of the gene tree would be congruent with the species tree, but this need not be the case (Fitch, 1970; Avise et al., 1983; Tajima, 1983; Pam-

ilo and Nei, 1988; Doyle, 1992). In this article, I review the processes by which discord can arise and then explore how a species tree can be reconstructed from gene trees by considering these processes of discord. However, discordant gene trees will also provoke me to reconsider precisely what species trees (i.e., phylogenies) are.

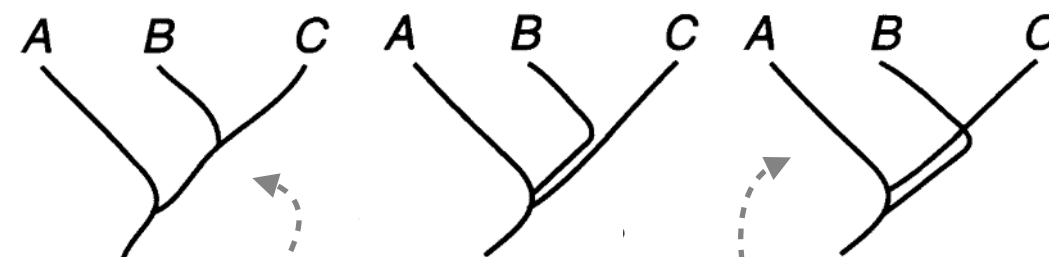
GENE TREES AND SPECIES TREES

Genes have gene trees because of gene replication. As a gene copy at a locus in the genome replicates and its copies are passed on to more than one offspring, branching points are generated in the gene tree. Because the gene copy has a single ancestral copy, barring recombination, the resulting history is a branching tree. (Point mutation can cause some of the copies to be imperfect representations of the original, but this process does not compromise the existence of the tree.) Sexual reproduc-

Species tree



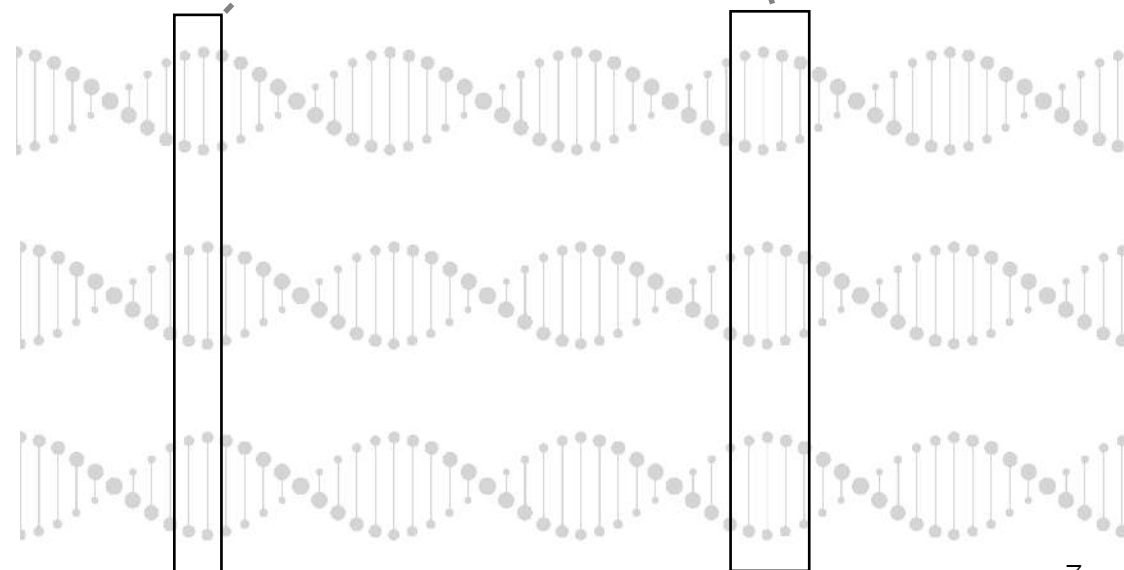
Possible gene trees (genealogies)



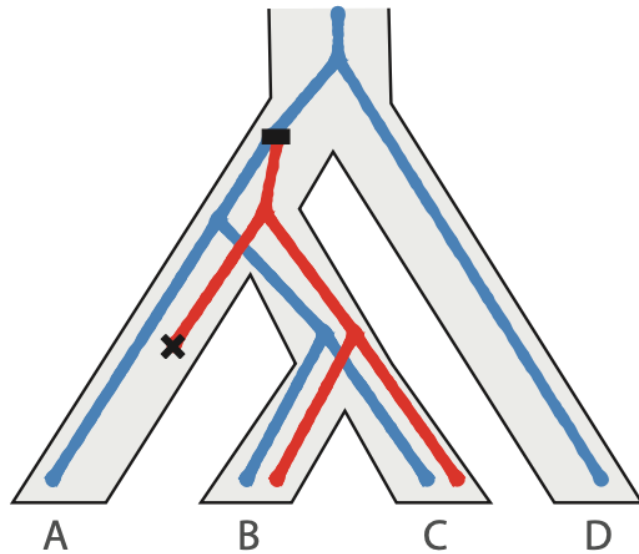
Species A

Species B

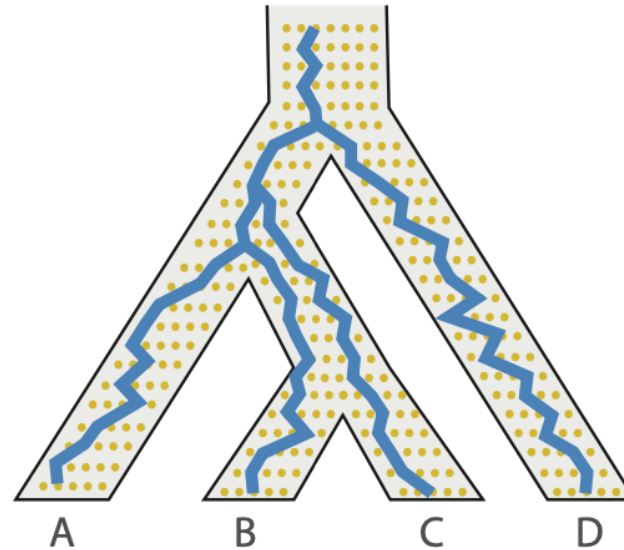
Species C



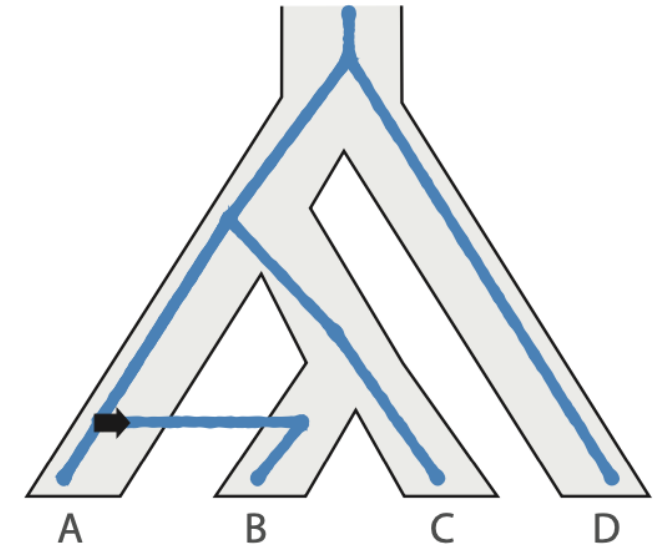
Biological Reasons Gene Trees differ from Species Tree



i Gene duplication and loss (GDL)

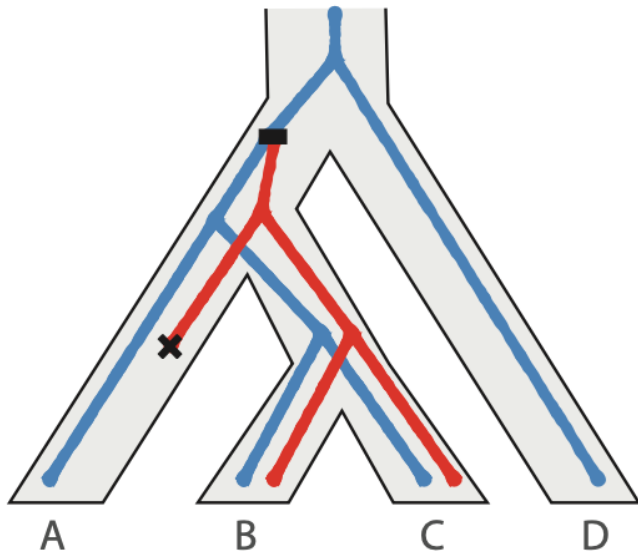


ii Incomplete lineage sorting (ILS)

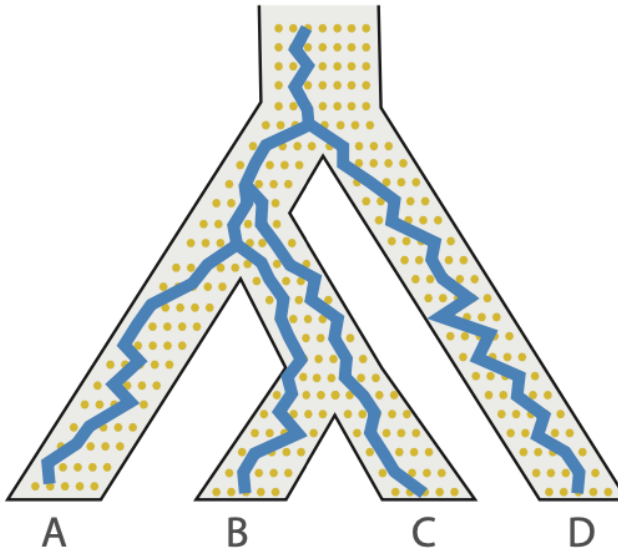


iii Horizontal gene transfer (HGT)

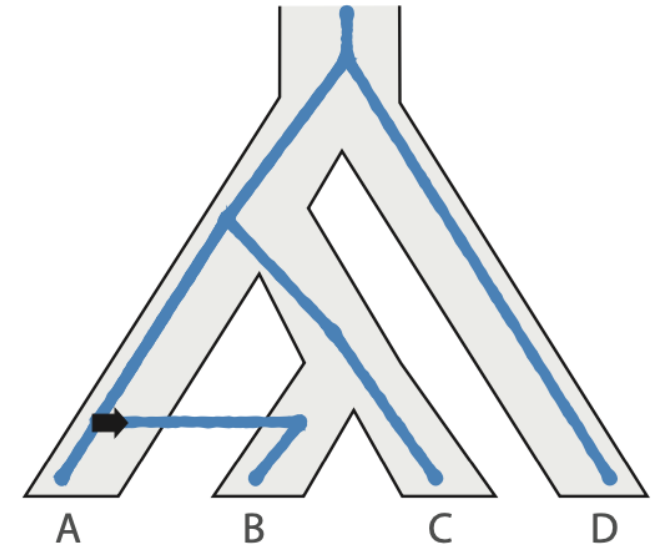
Biological Reasons Gene Trees differ from Species Tree



i Gene duplication and loss (GDL)



ii Incomplete lineage sorting (ILS)



iii Horizontal gene transfer (HGT)

species tree



time

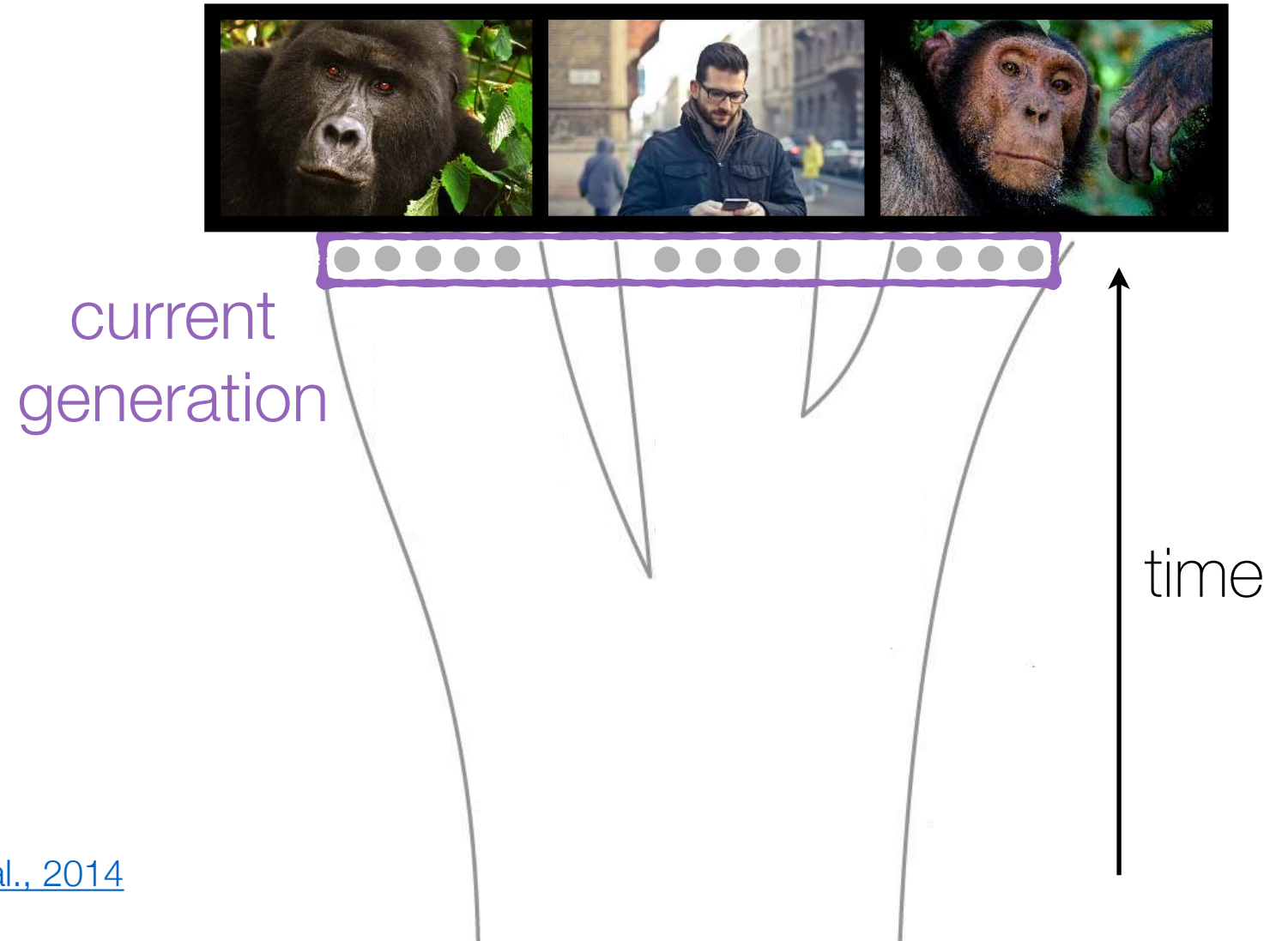
species tree

one individual
in population

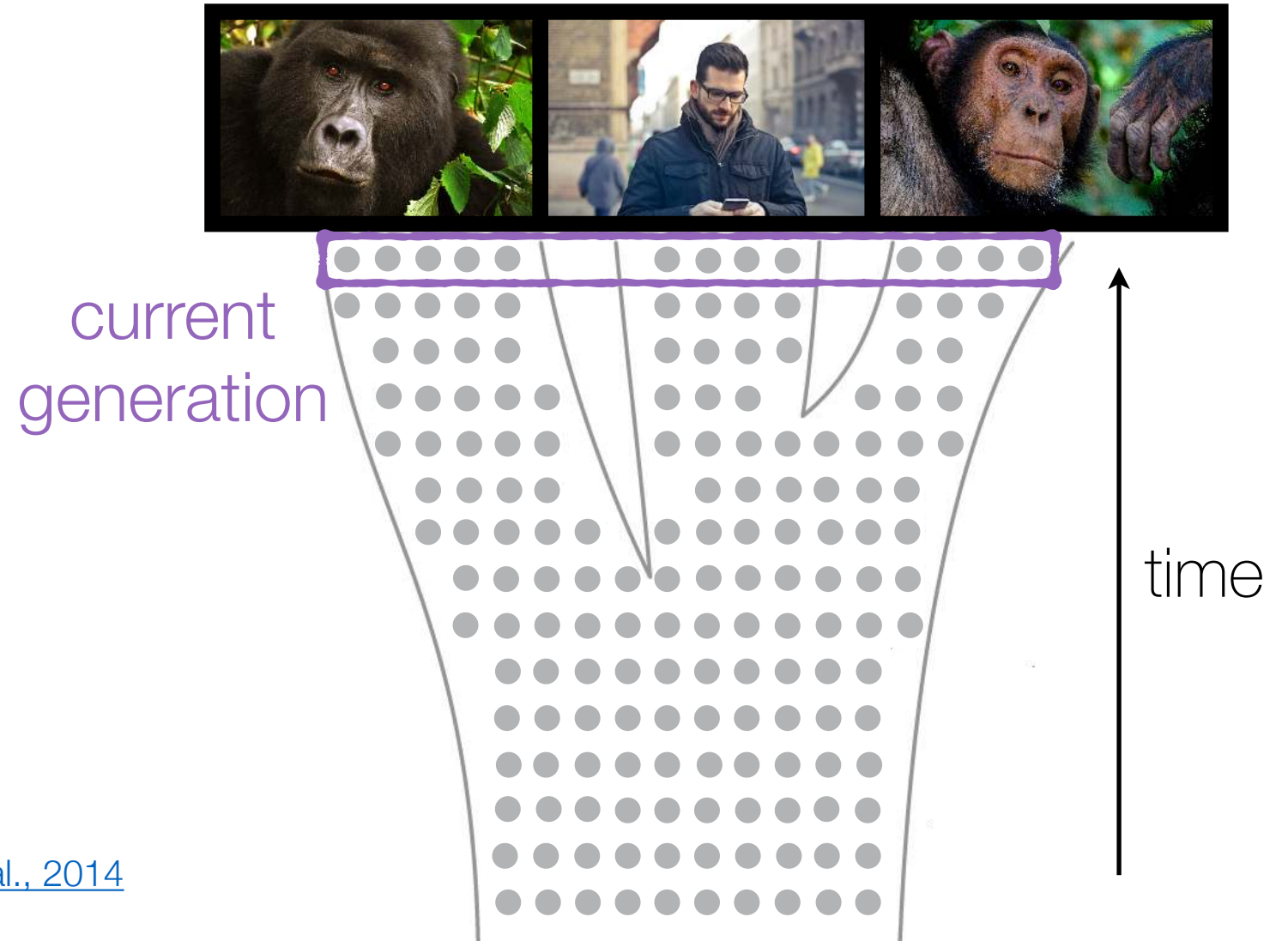


time

species tree

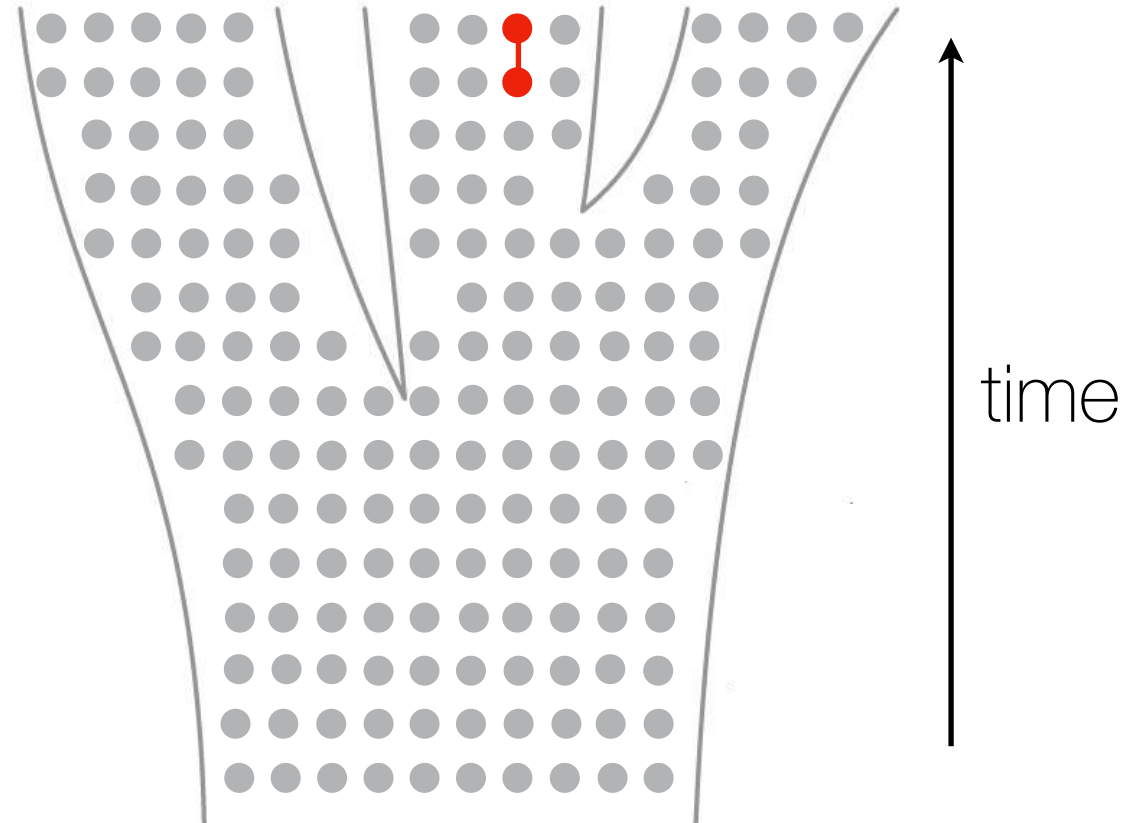


species tree



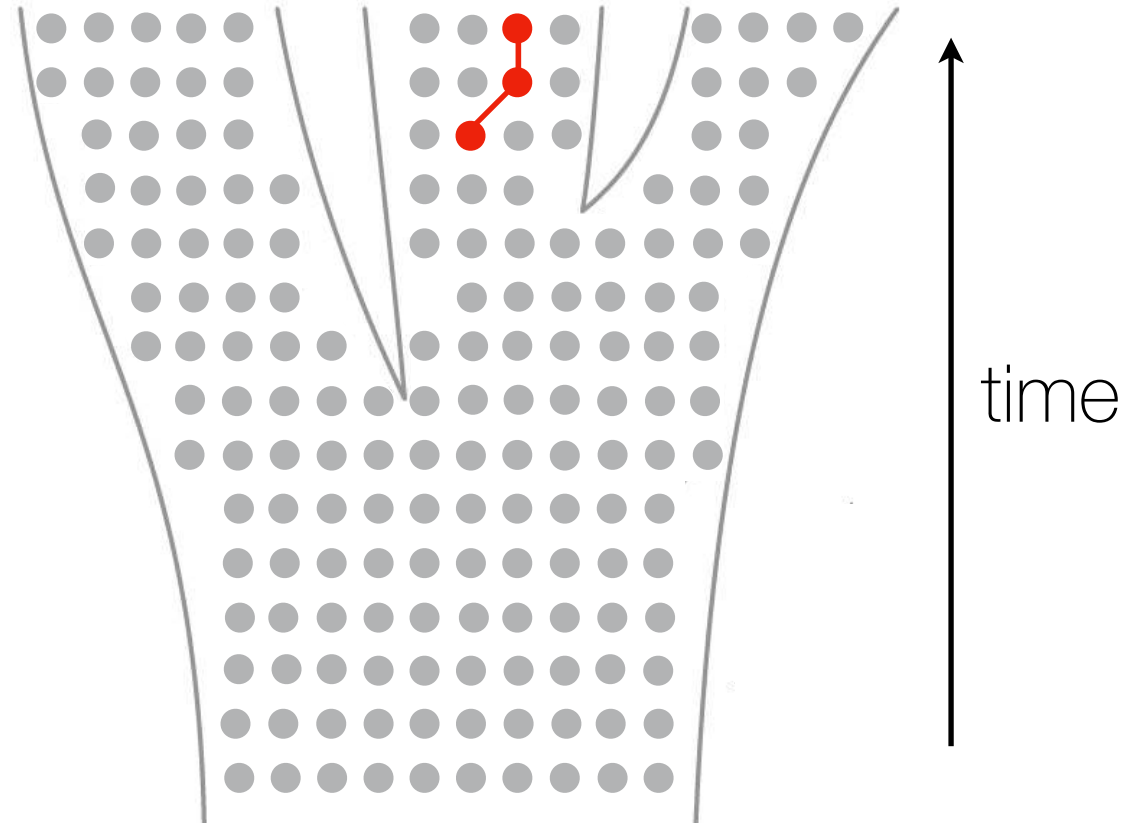
“gene”
inherited from
individual
in previous
generation

species tree



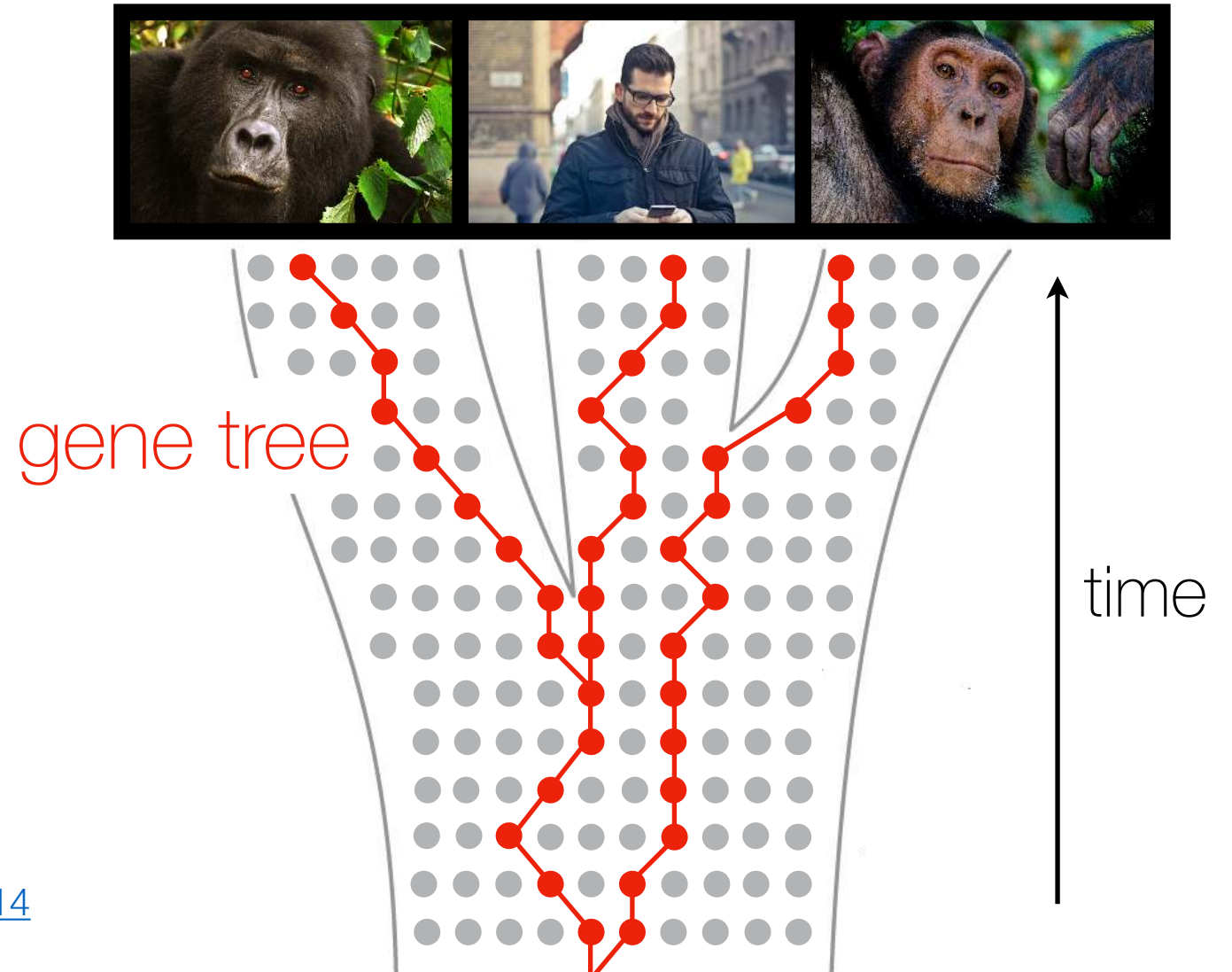
“gene”
inherited from
individual
in previous
generation

species tree



species tree

“gene”
inherited from
individual
in previous
generation



gene tree

differs

from

species tree

(incomplete lineage sorting)

species tree



gene tree

time

gene tree

differs

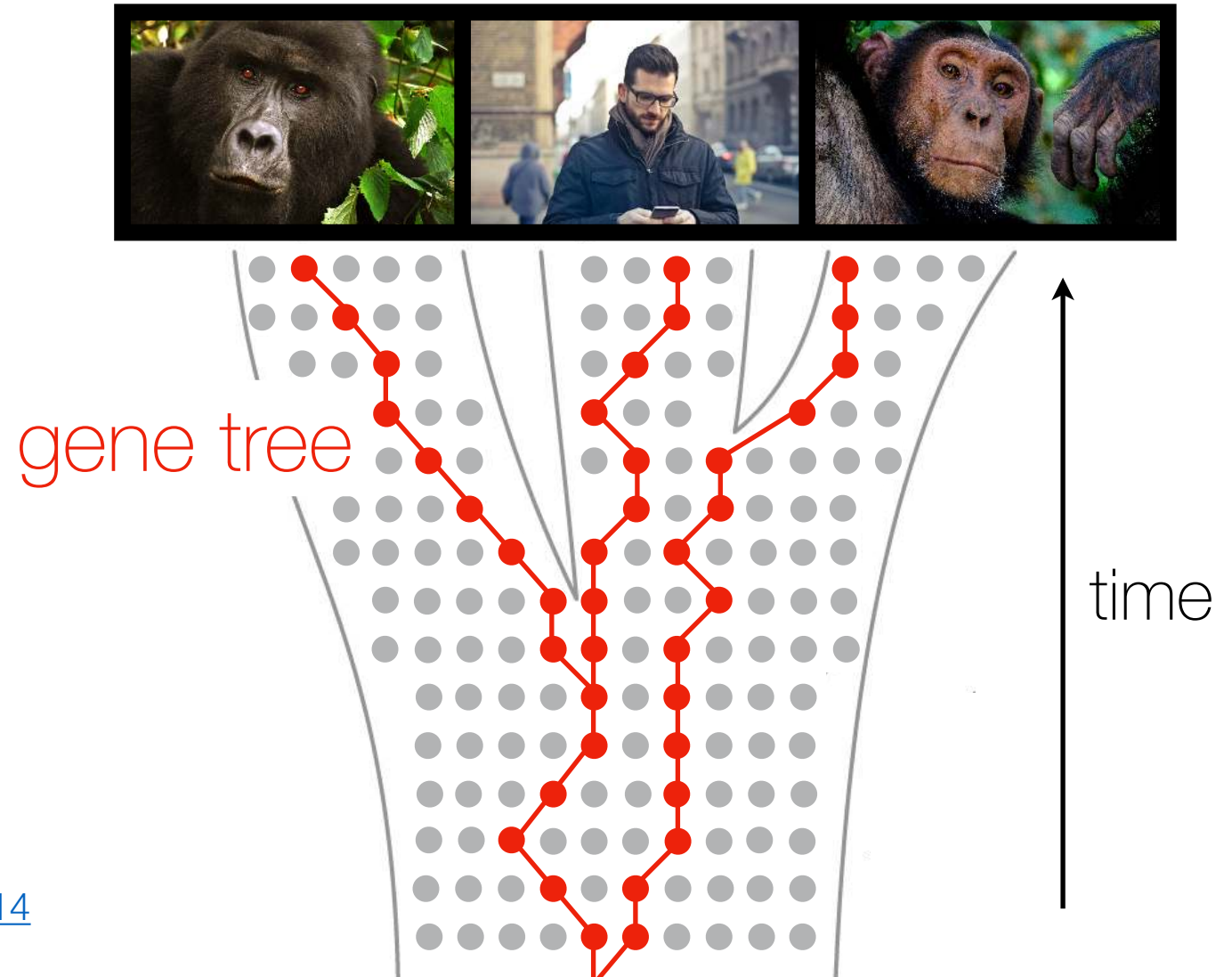
from

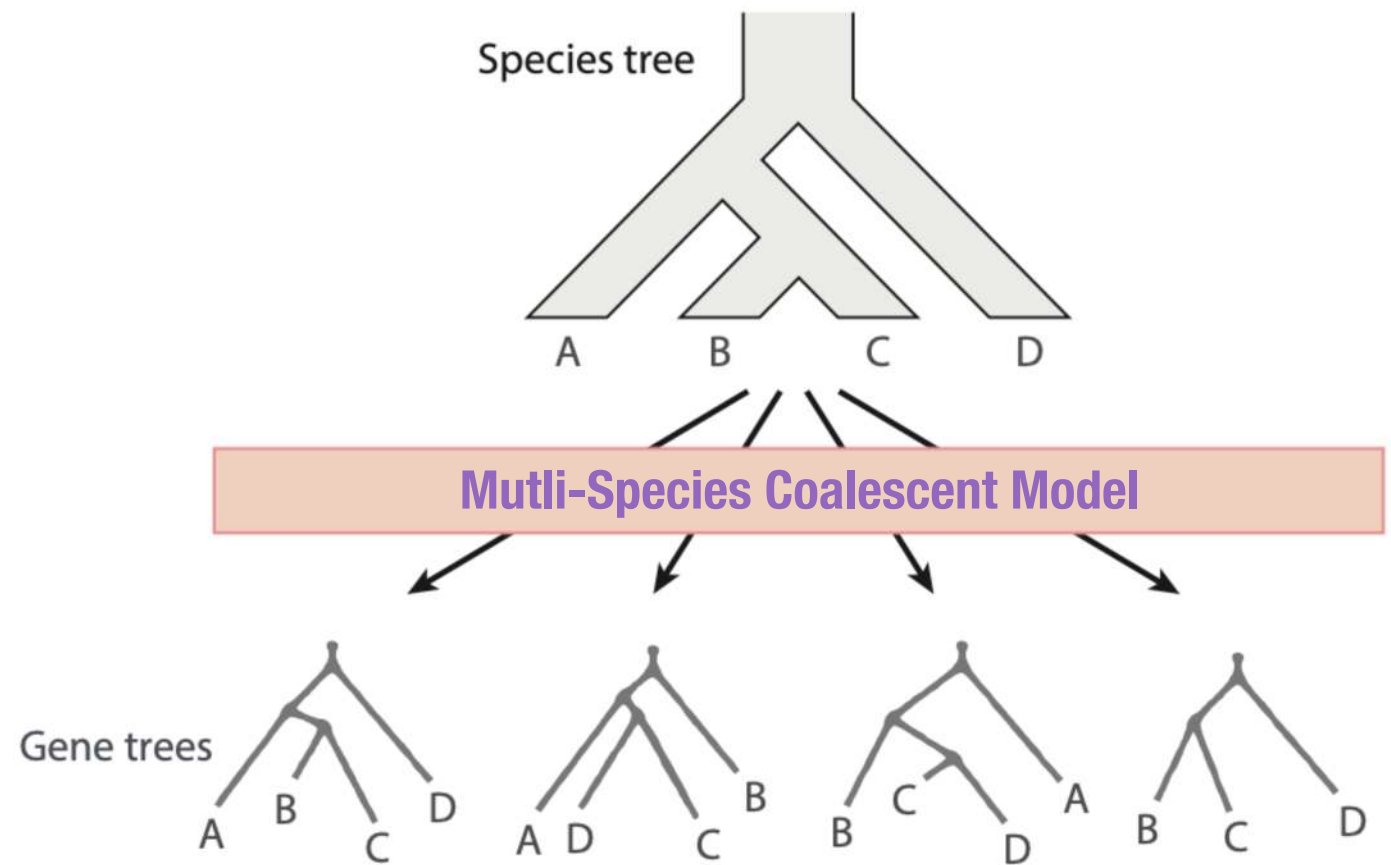
species tree

(incomplete lineage sorting)

ILS is modeled by **Multi-Species
Coalescent (MSC)**!

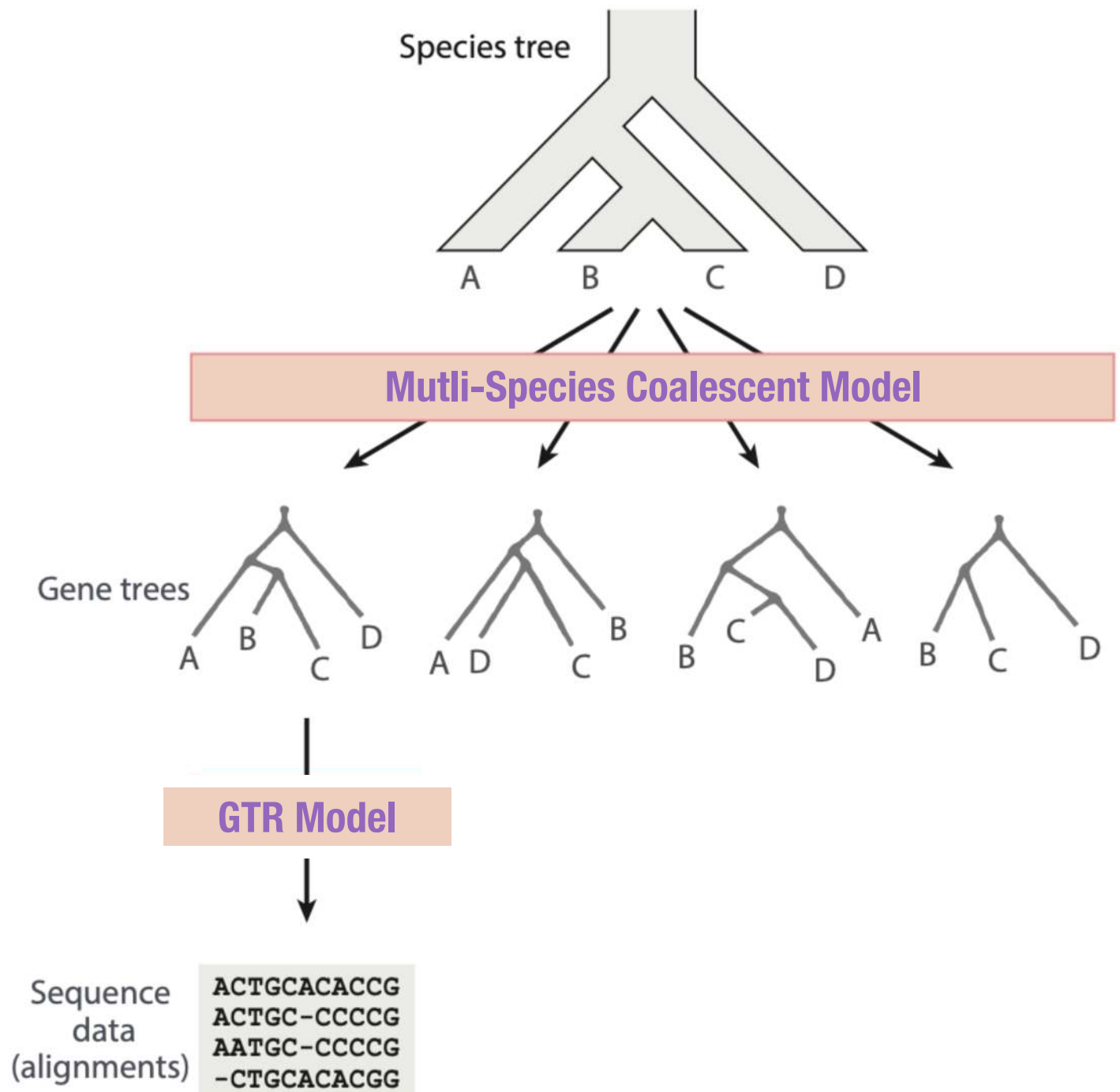
species tree



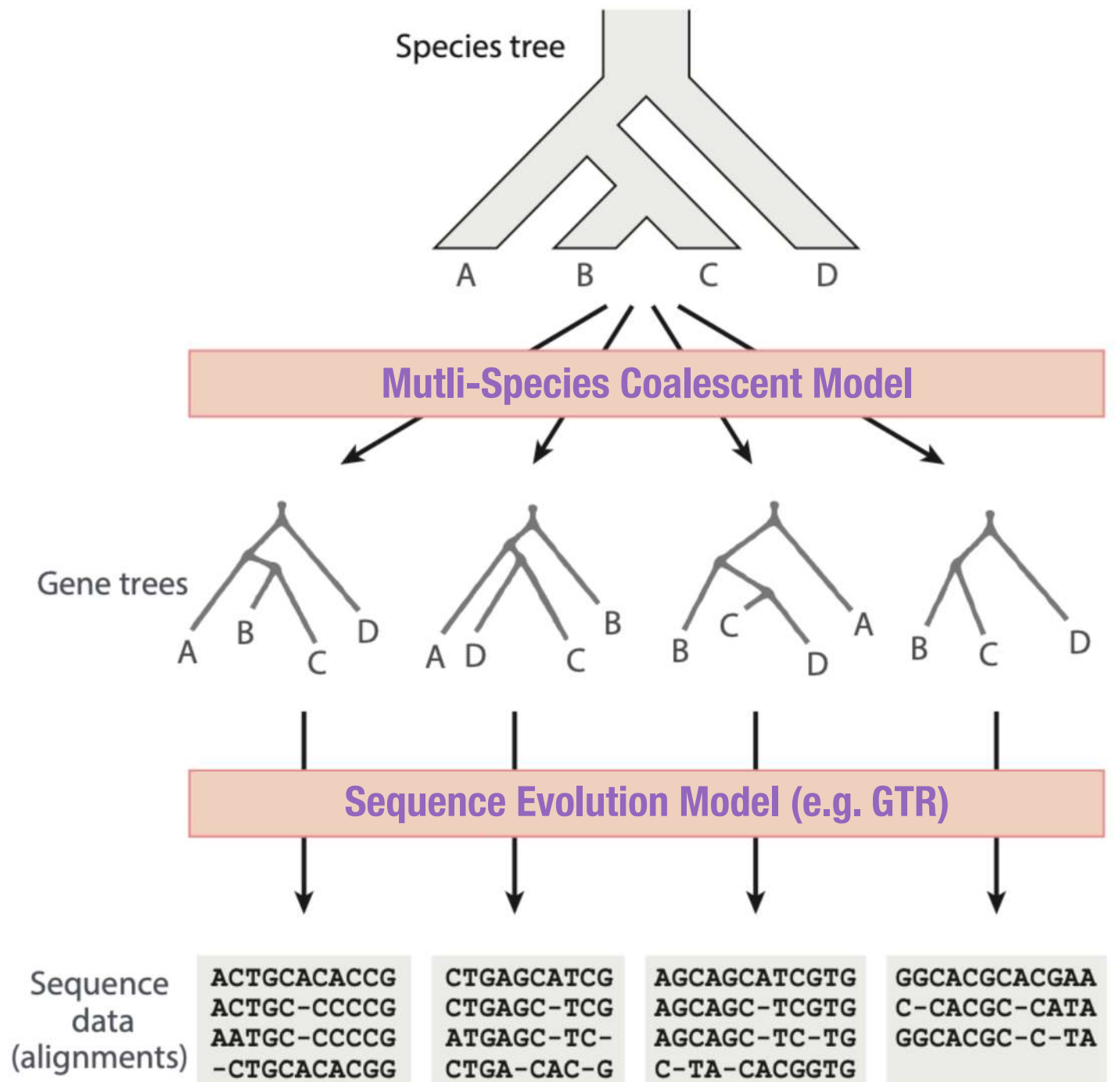


(to be discussed in detail)

But we don't
directly observe
gene trees...



But we don't
directly observe
gene trees...



The Concatenation Approach

Step 0.

Lots of data processing!!!!

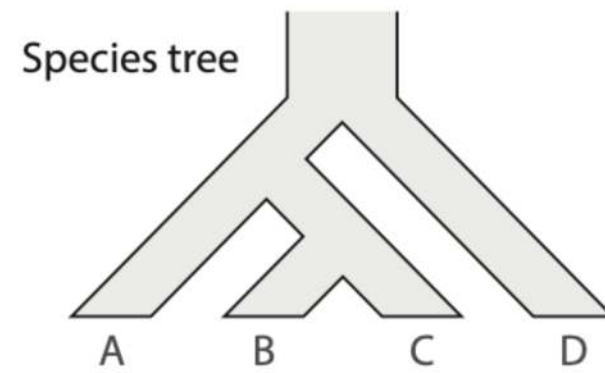
Step 1.

Concatenate alignments.

Sequence data (alignments)	ACTGCACACCG ACTGC-CCCCG AATGC-CCCCG -CTGCACACGG	CTGAGCATCG CTGAGC-TCG ATGAGC-TC- CTGA-CAC-G	AGCAGCATCGTG AGCAGC-TCGTG AGCAGC-TC-TG C-TA-CACGGTG	GGCACGCACGAA C-CACGC-CATA GGCACGC-C-TA
----------------------------	--	--	--	--

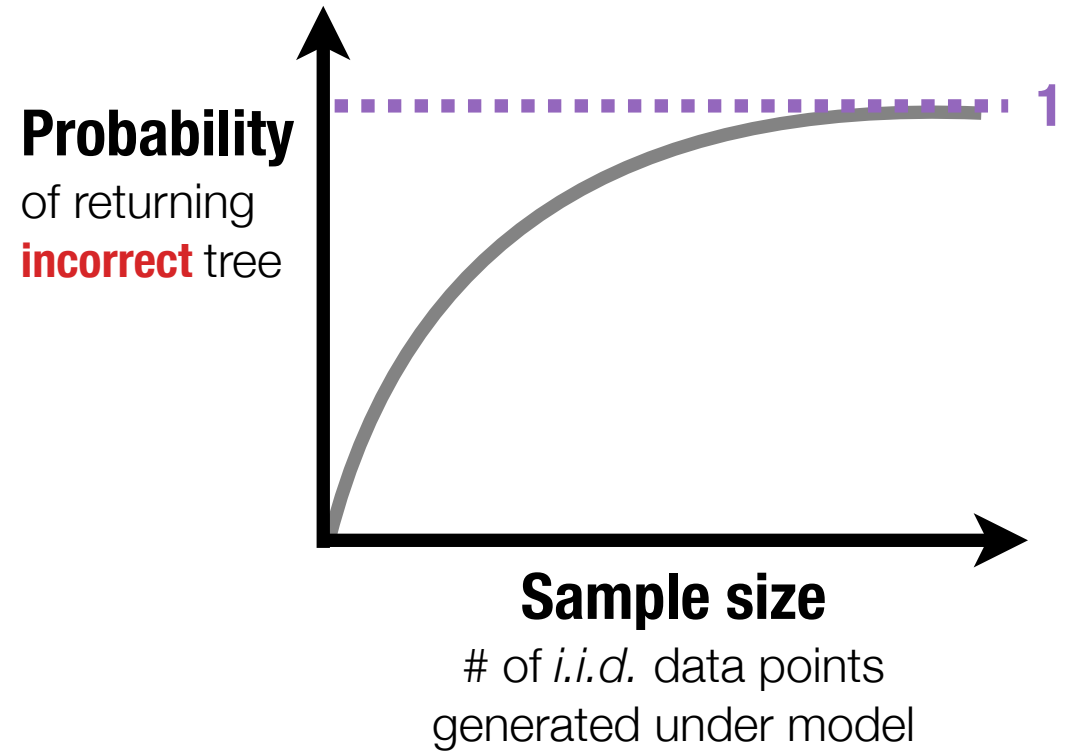
Step 2.

Estimate species tree under standard models that assume all regions evolved down the same tree i.e. they **ignore** gene tree heterogeneity



Why not just concatenate?

1. Can be **positively misleading** under the MSC model.
[\[Roch & Steel, 2015\]](#)
2. Mixed accuracy in simulations
[\[Kubatko & Degnan, 2007\]](#)
3. Model misspecification can cause impact branch length estimation & uncertainty quantification [\[Slides by Kubatko, 2019\]](#)
4. Can be computationally expensive

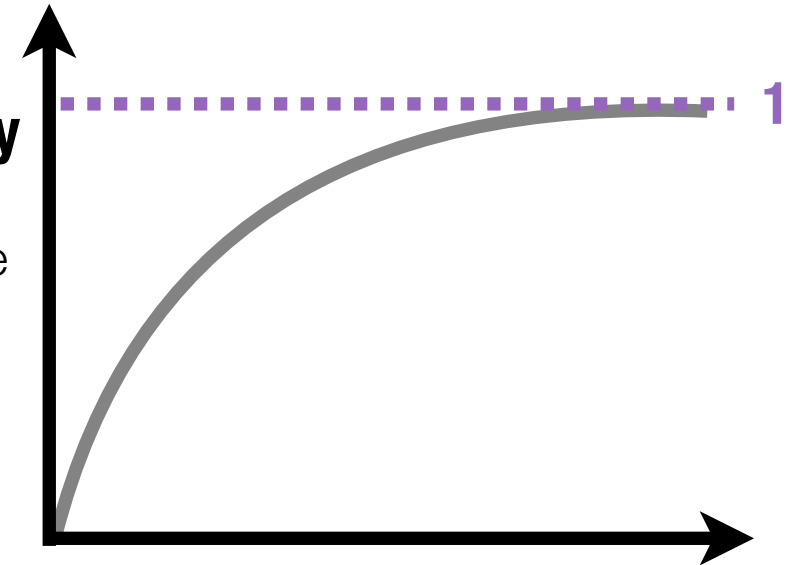


Why not just concatenate?

1. Can be **positively misleading** under the MSC model.

[\[Roch & Steel, 2015\]](#)

Probability
of returning
incorrect tree



Sample size

of *i.i.d.* data points
generated under model

Anomaly Zone

most probable gene tree
disagrees with species tree!!!!

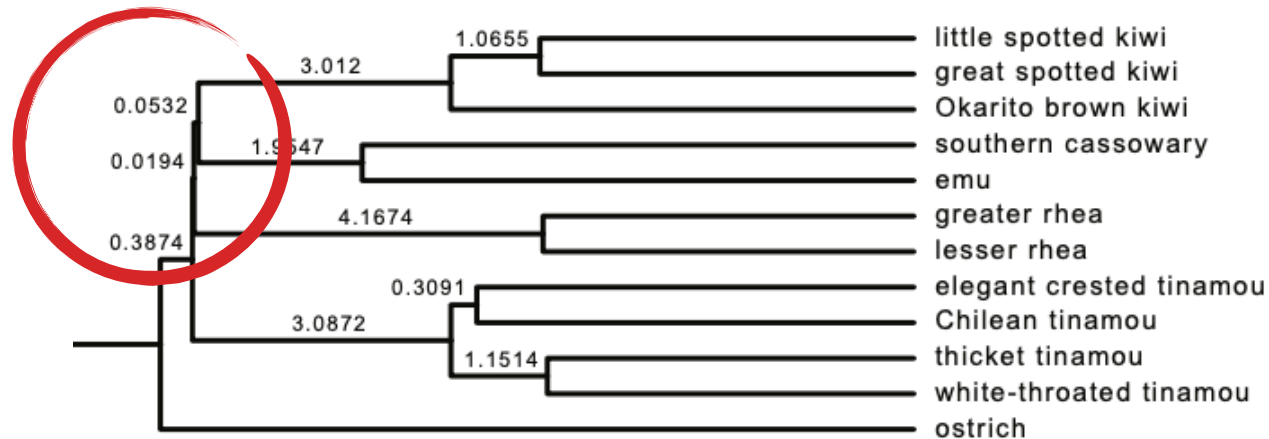


Image credit: Figure 1c in
[Molloy, Gatesy & Springer, 2021](#)

Evidence for rapid radiations and ILS in many major clades



Palaeognathae

[e.g., [Cloutier et al., Syst Biol, 2019](#)]



Neoaves

[e.g., [Jarvis et al., Science, 2014](#)]



Placental Mammals

[e.g., [McCormack et al., Genome Res, 2012](#)]

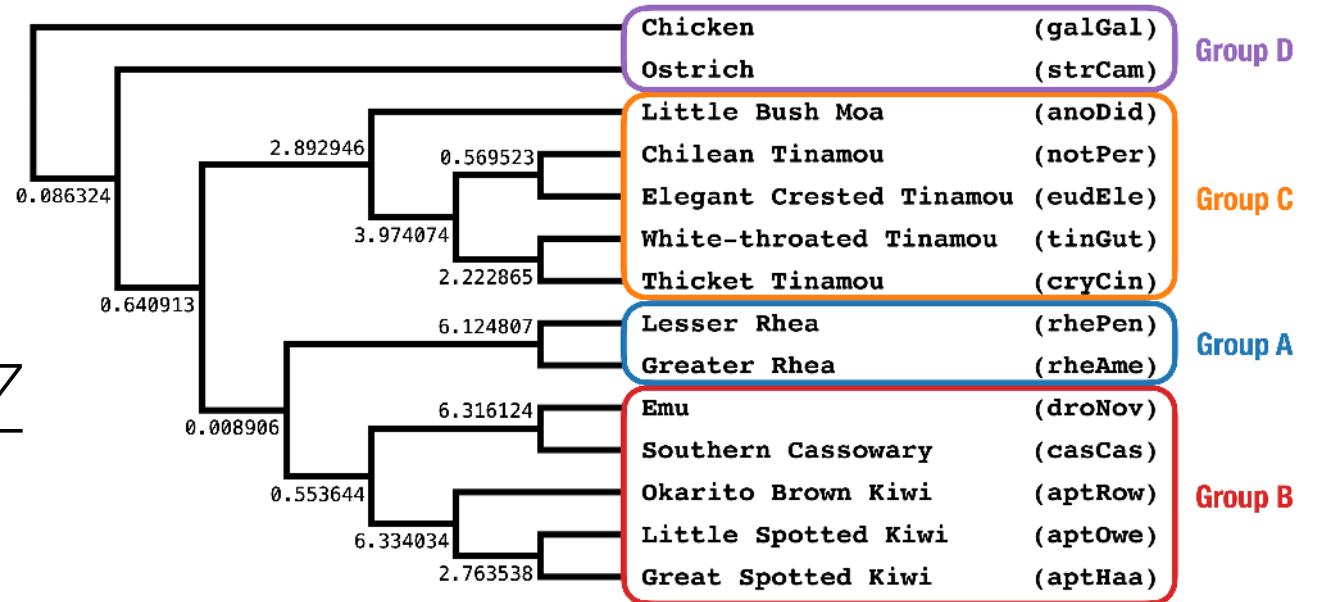


Green Plants

[e.g., [Leebens-Mack et al., Nature, 2019](#)]

Activity A

Check if this model
species tree is in the AZ



20 minutes

<https://github.com/molloy-lab/ck-phylo-workshop>

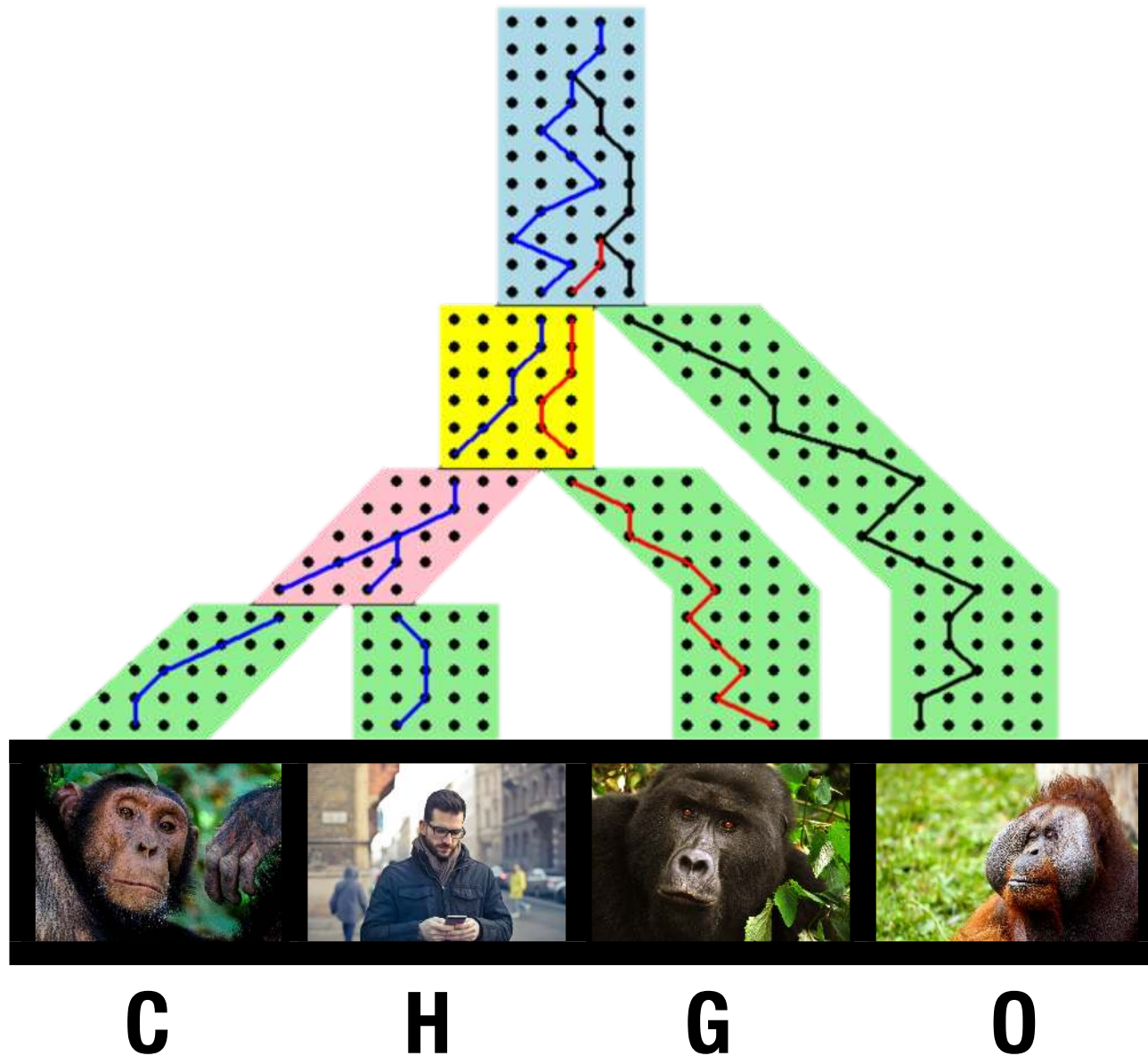
Coalescent Lab — Day 1

Motivation for coalescent methods (**Activity A**)

Coalescent basics (**Activity B**)

Species tree estimation with summary methods (**Activity C**)

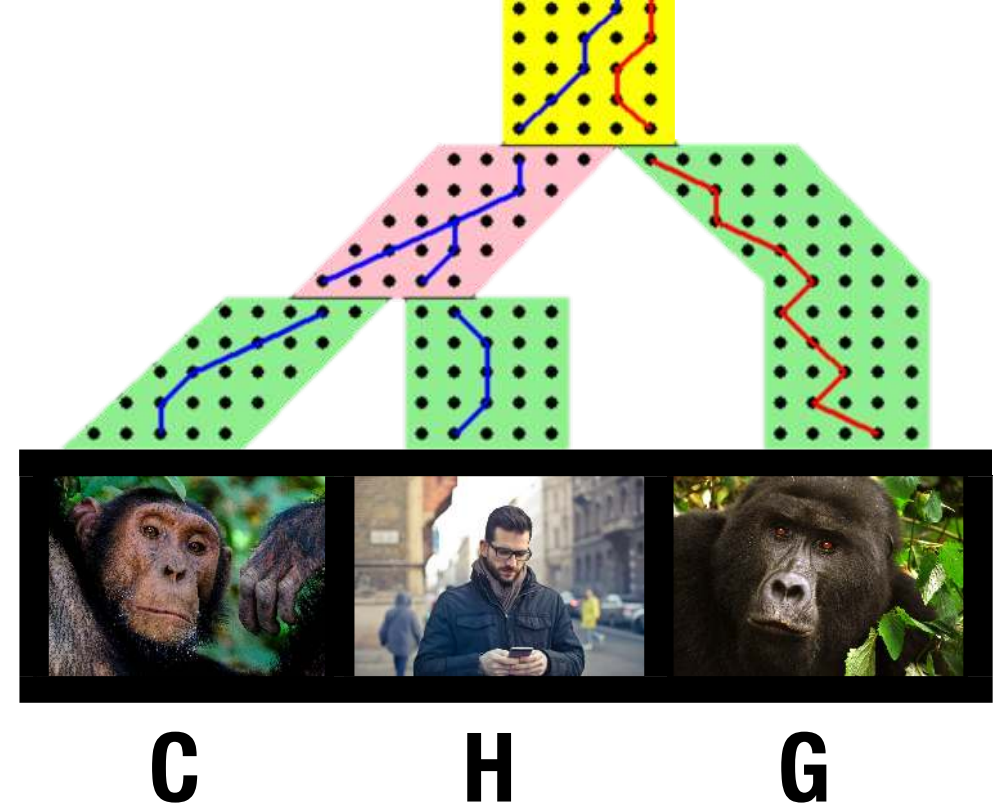
Evaluation model fit (**Activity D** — optional / do tomorrow)



MSC model has the following parameters:

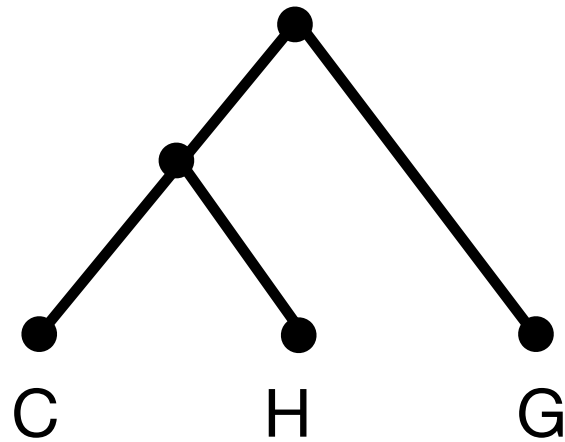
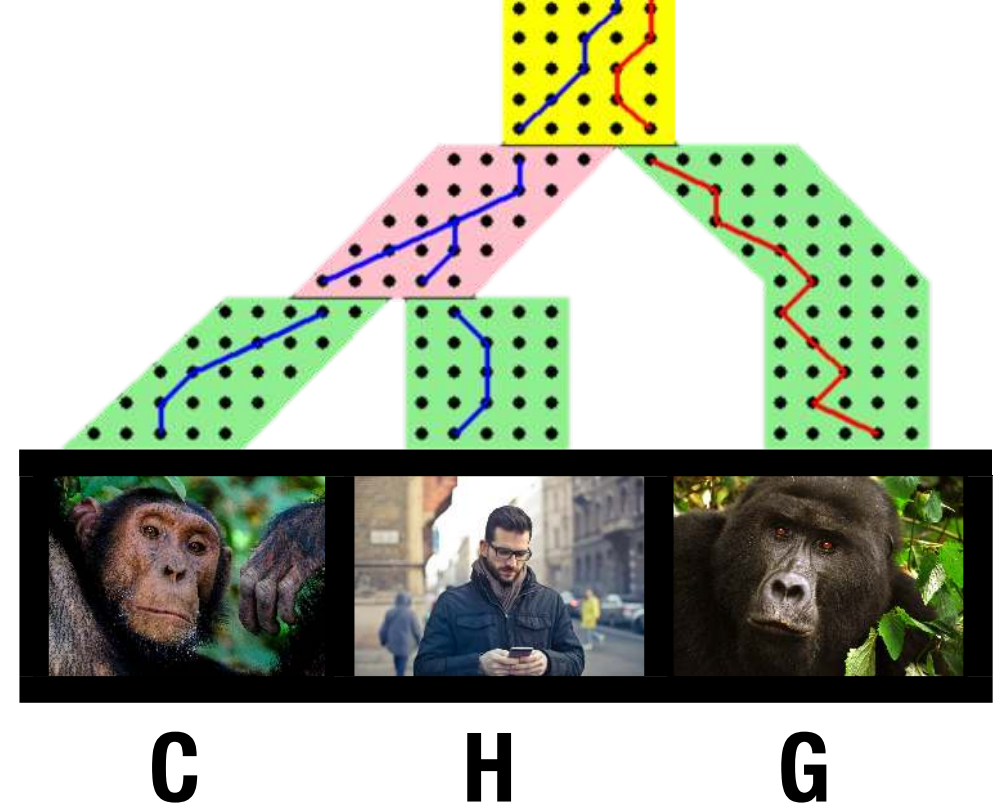
- **species tree topology** in this case “(((C,H),G),O);”
- **branch lengths**
(# of generations)
- **branch widths**
(effective population size)

Consider an MSC model species tree with 3 taxa.

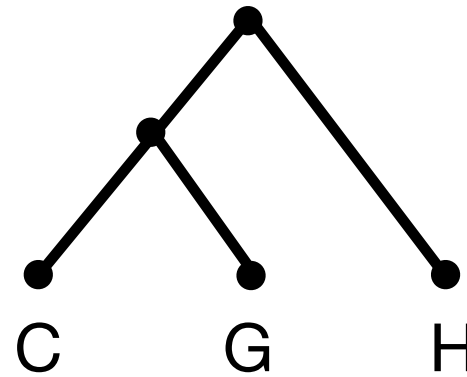


Consider an MSC model species tree with 3 taxa.

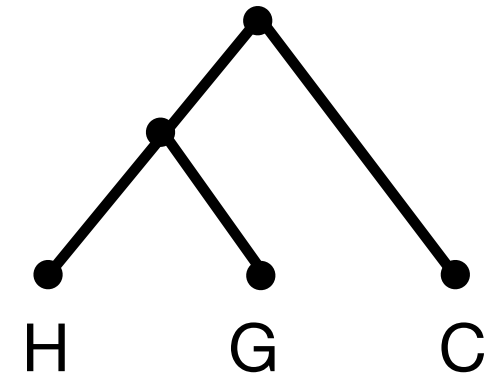
It can generate 3 gene tree topologies.



Same topology as species tree

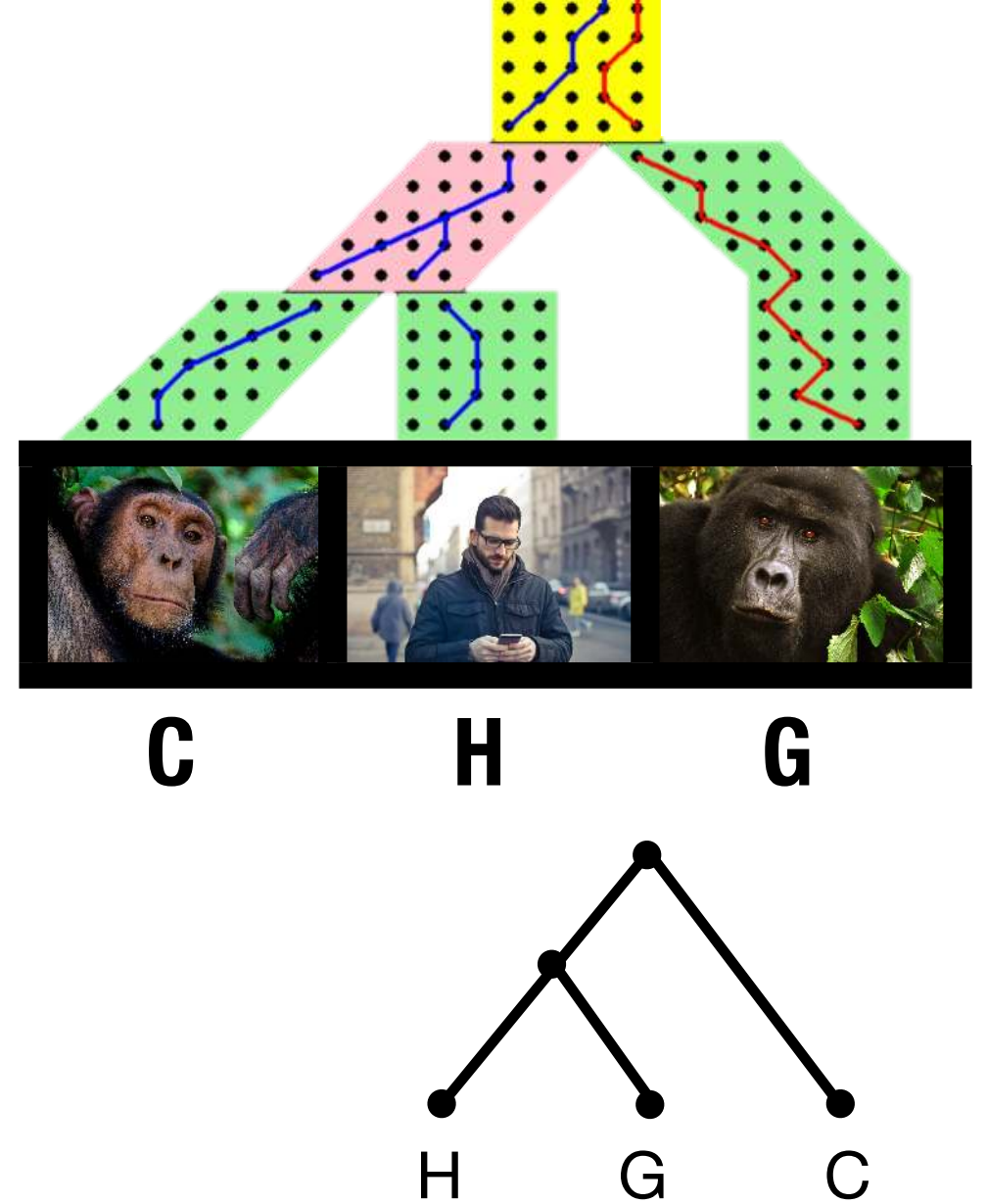


Different topologies than species tree



Consider an MSC model species tree with 3 taxa.

Q: What is probability of “(H,G),C);”?

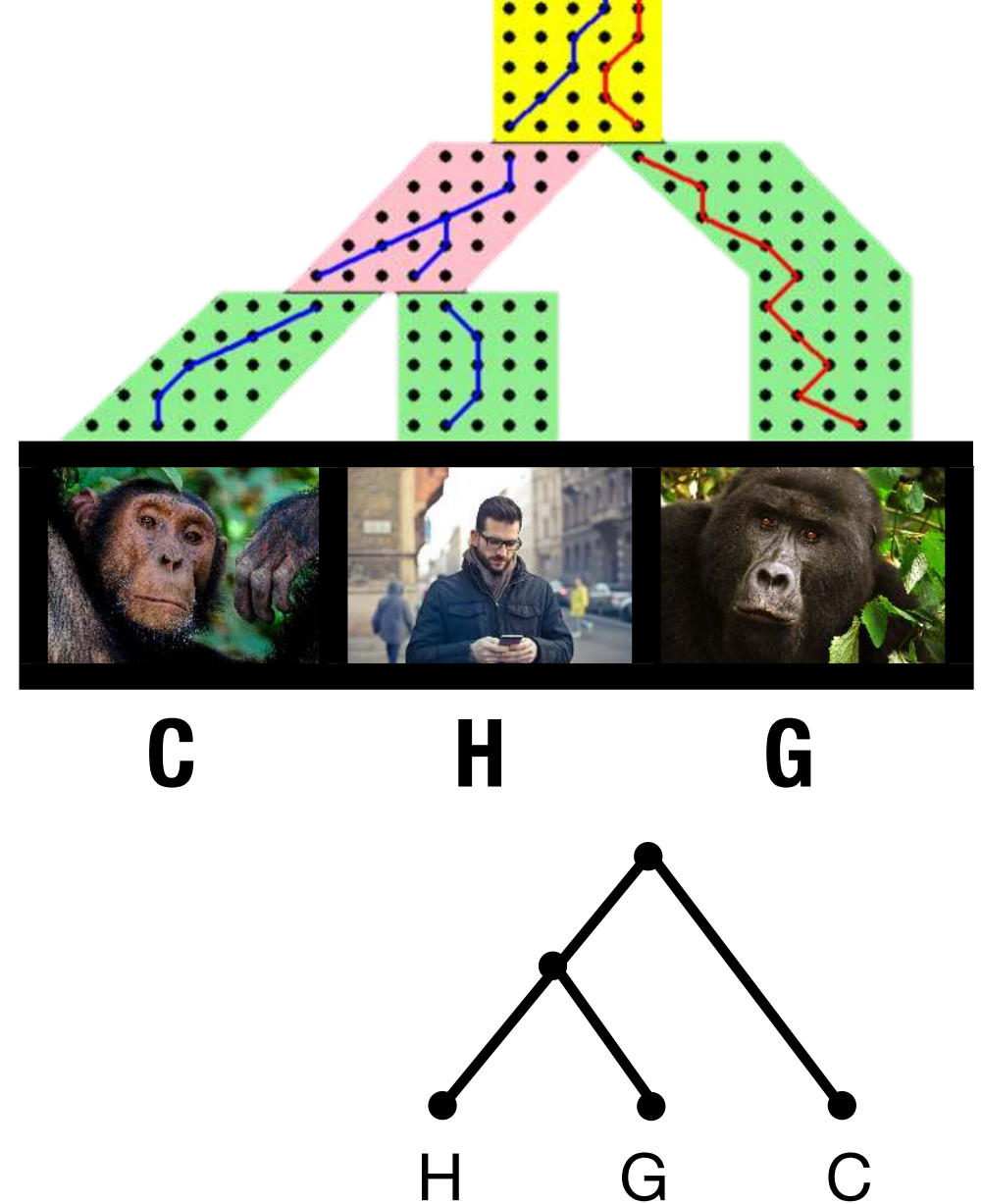


Consider an MSC model species tree with 3 taxa.

Q: What is probability of “(H,G),C);”?

This gene tree must be generated with the following events:

- (1) Lineages h & c enter **internal branch** and FAIL to coalesce on it
- (2) Lineages h, c, & g coalesce enter **above root “branch”** and h & g coalesce first



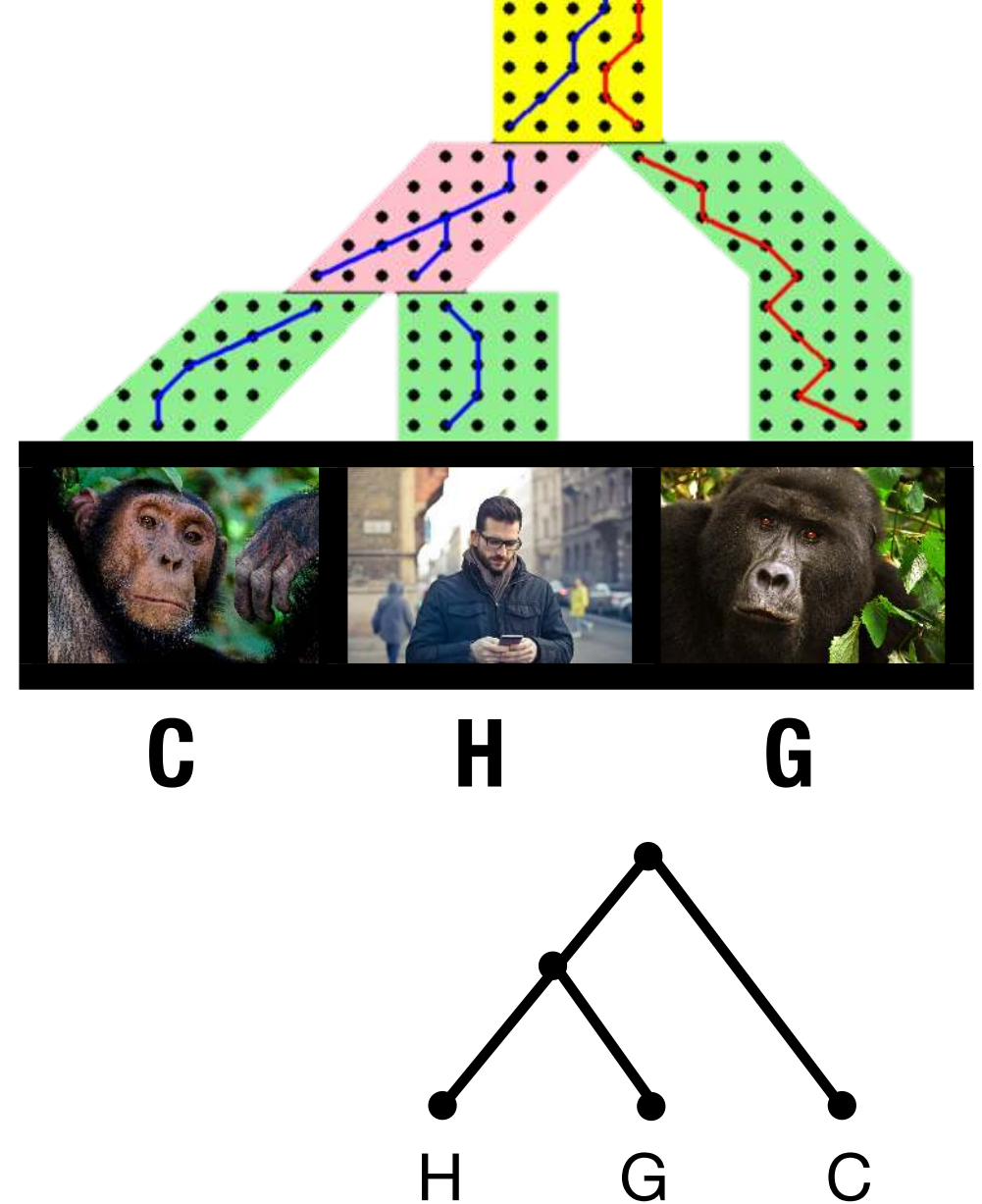
Consider an MSC model species tree with 3 taxa.

Q: What is probability of “(H,G),C);”?

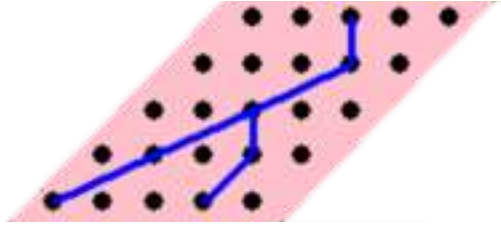
This gene tree must be generated with the following events:

- (1) Lineages h & c enter **internal branch** and FAIL to coalesce on it
- (2) Lineages h, c, & g coalesce enter **above root “branch”** and h & g coalesce first

What is the probability of (1) and (2)?



Coalescence

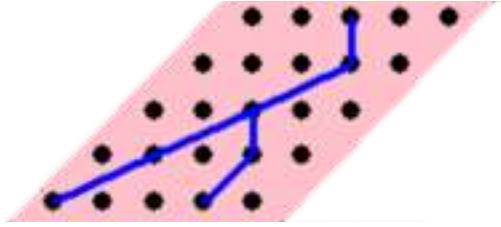


2 lineages enter branch & successfully coalesce

Consider population size of $2N_e$ per generation.

Q: What is the probability 2 lineages coalesce after 1 generation?

Coalescence



2 lineages enter branch & successfully coalesce

Consider population size of $2N_e$ per generation.

Q: What is the probability 2 lineages coalesce after 1 generation?

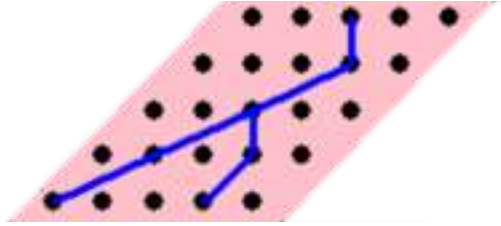
Imagine rolling a die labeled with $x = 2N_e$ possible ancestors.

The **first roll** produces an ancestor.

The **second roll** produces the same ancestor with probability $1/x$.

This is the probability of coalescence after 1 generation!

Coalescence



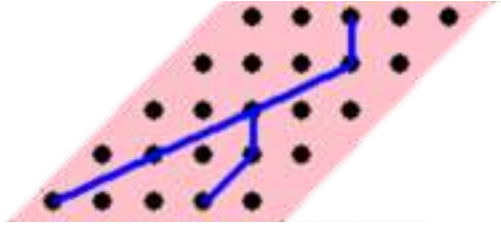
2 lineages enter branch & successfully coalesce

Consider population size of $2N_e$ per generation.

Q: What is the probability 2 lineages coalesce after 1 generation?

Q: What about after exactly 2 generations?

Coalescence



2 lineages enter branch & successfully coalesce

Consider population size of $2N_e$ per generation.

Q: What is the probability 2 lineages coalesce after 1 generation?

Q: What about after exactly 2 generations?

Again, imagine rolling a die labeled with $x = 2N_e$ possible ancestors.

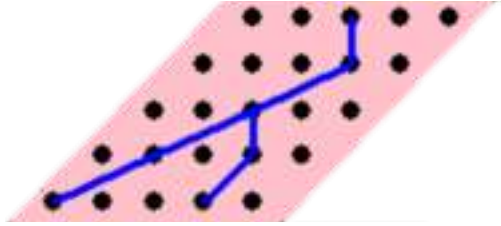
In order for this to occur, we need
(1) failure to coalesce after 1 gen &
(2) success after the next generation

First event has probability: $1 - 1/x$

Second event has probability: $1/x$

Total probability is $(1 - 1/x) \cdot 1/x$

Coalescence



2 lineages enter branch & successfully coalesce

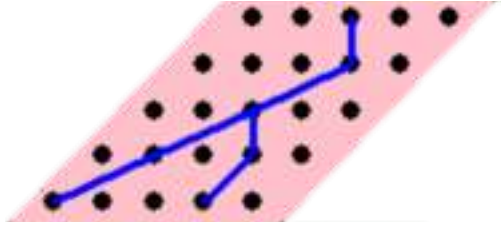
Consider population size of $2N_e$ per generation.

Q: What is the probability 2 lineages coalesce after 1 generation?

Q: What about after exactly 2 generations?

Q: What about after exactly t generations?

Coalescence



2 lineages enter branch & successfully coalesce

Consider population size of $2N_e$ per generation.

Q: What is the probability 2 lineages coalesce after 1 generation?

Q: What about after exactly 2 generations?

Q: What about after exactly t generations?

Again, imagine rolling a die labeled with $x = 2N_e$ possible ancestors.

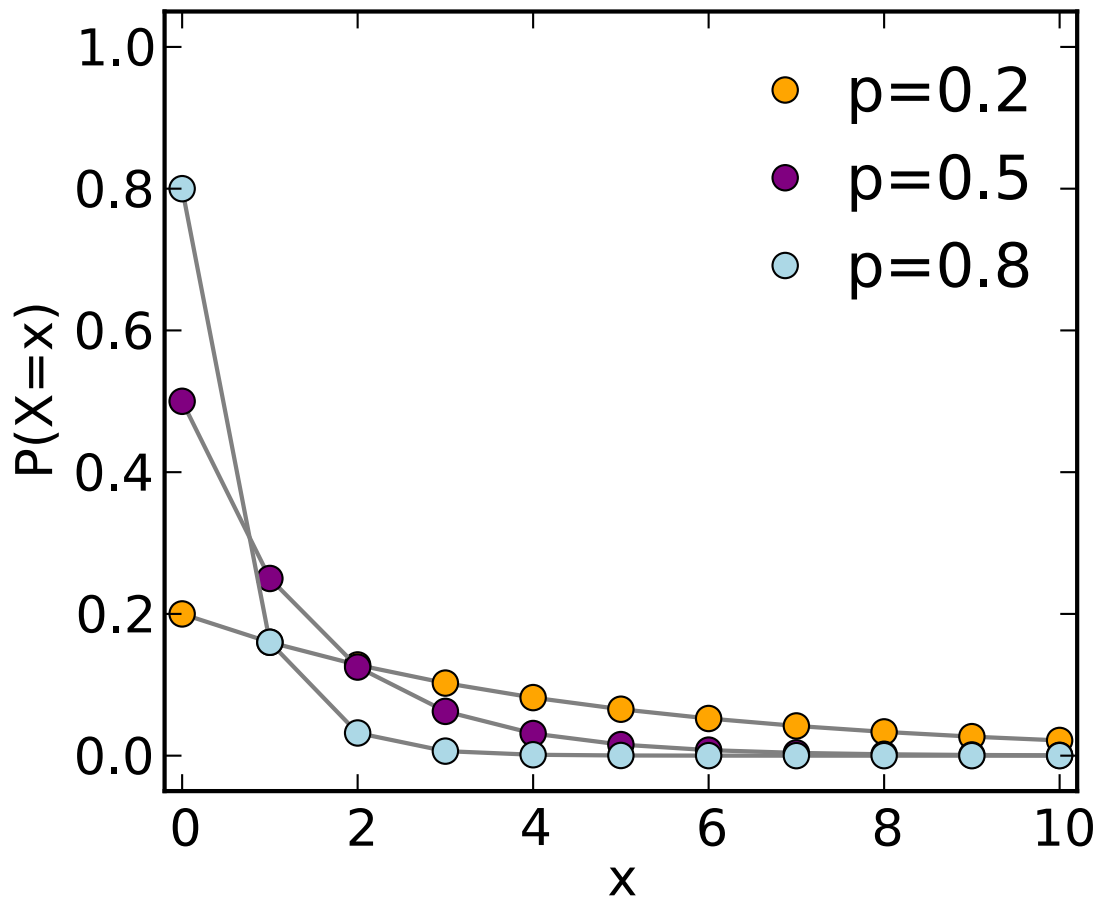
In order for this to occur, we need

- (1) failure to coalesce for $t - 1$ gens &
- (2) success after the next generation

Total probability is $(1 - 1/x)^{t-1} \cdot 1/x$

Seem familiar?

Geometric Distribution

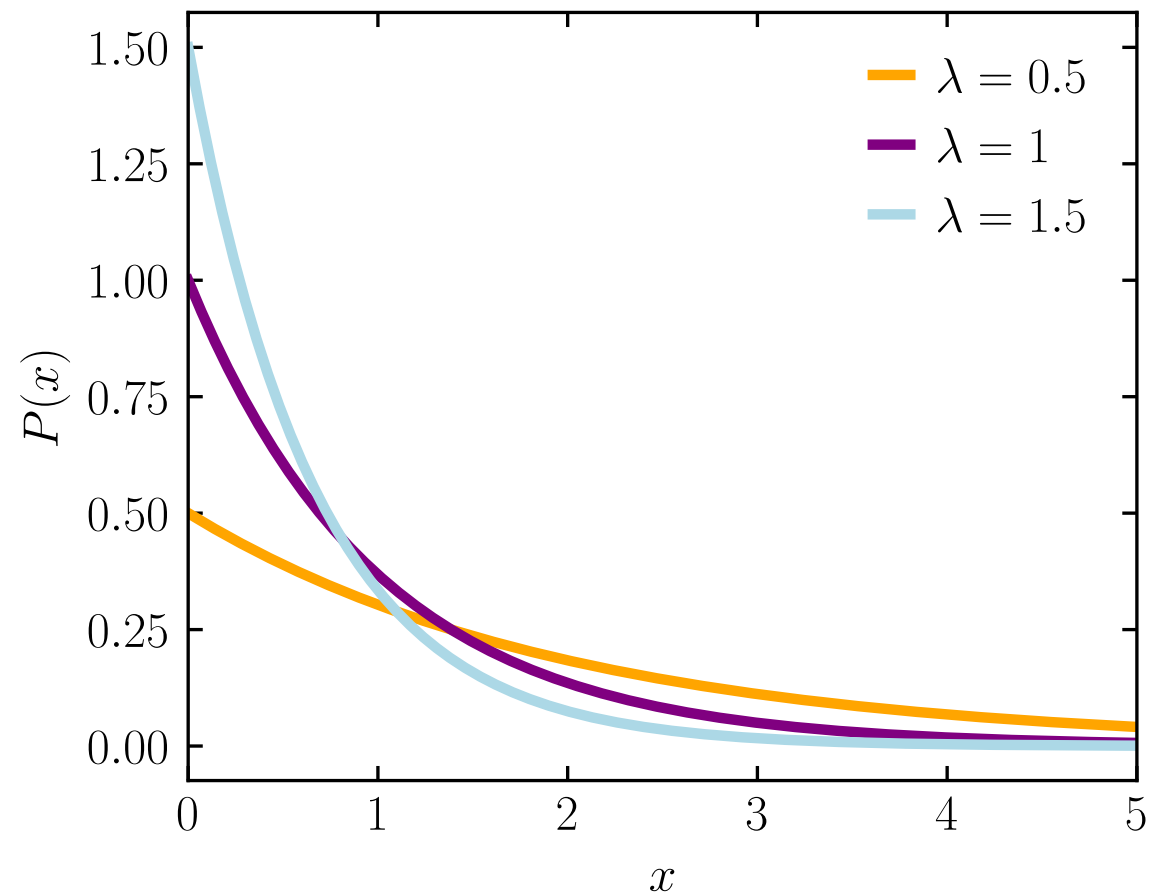
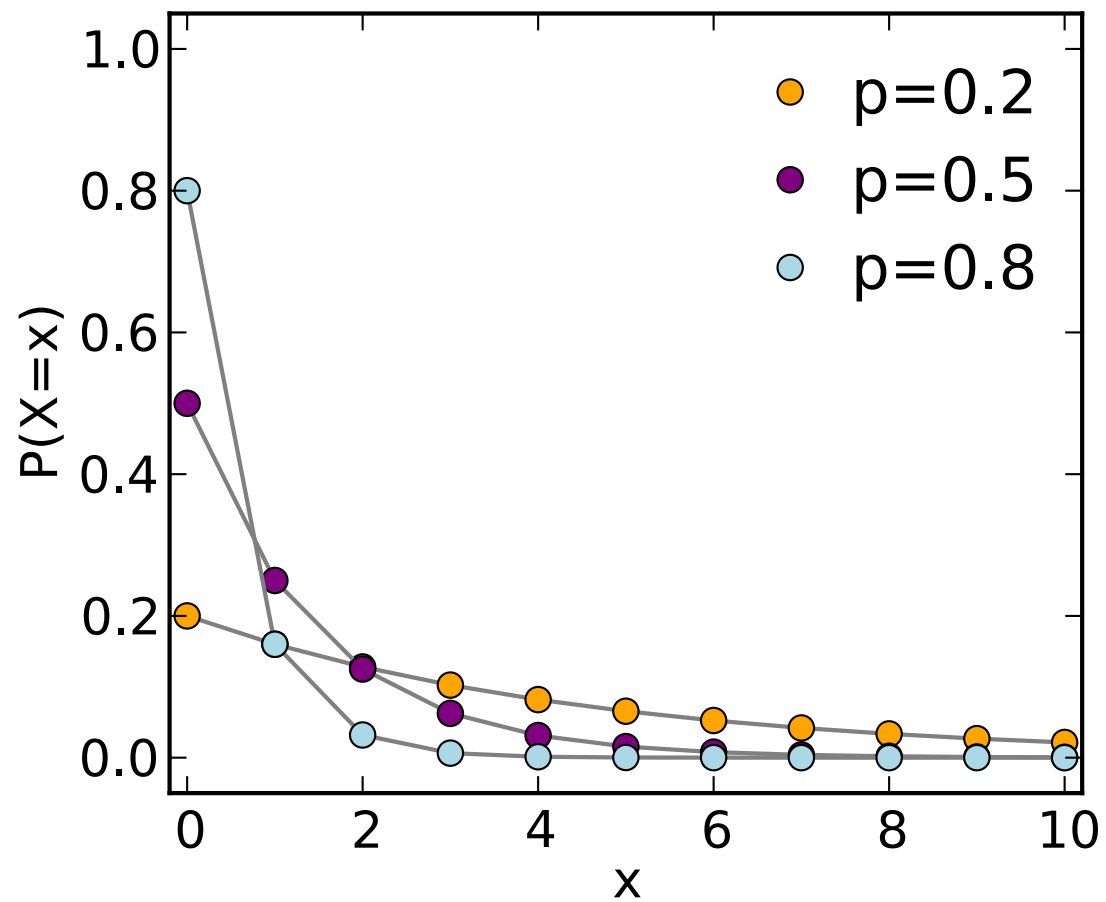


A discrete r.v. X is a **geometrically distributed**, denoted, $X \sim \text{Geom}(p)$, if its PMF

$$f_X(x; p) = \begin{cases} (1 - p)^x p & \text{for } x \in \mathbb{Z}_{\geq 0} \\ 0 & \text{otherwise} \end{cases}$$

models the number of failures until the first success, where $p \in \mathbb{R}_{>0, \leq 1}$ is probability of success.

Look similar?



Exponential Distribution

A continuous r.v. X is an **exponentially distributed**, denoted $X \sim \text{Exp}(\lambda)$, if its PDF

$$f_X(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \in \mathbb{R}_{\geq 0} \\ 0 & \text{for } x \in \mathbb{R}_{< 0} \end{cases}$$

models the **waiting time until next rare event**, where $\lambda \in \mathbb{R}_{> 0}$ is the expected number of rare events or mean rate.

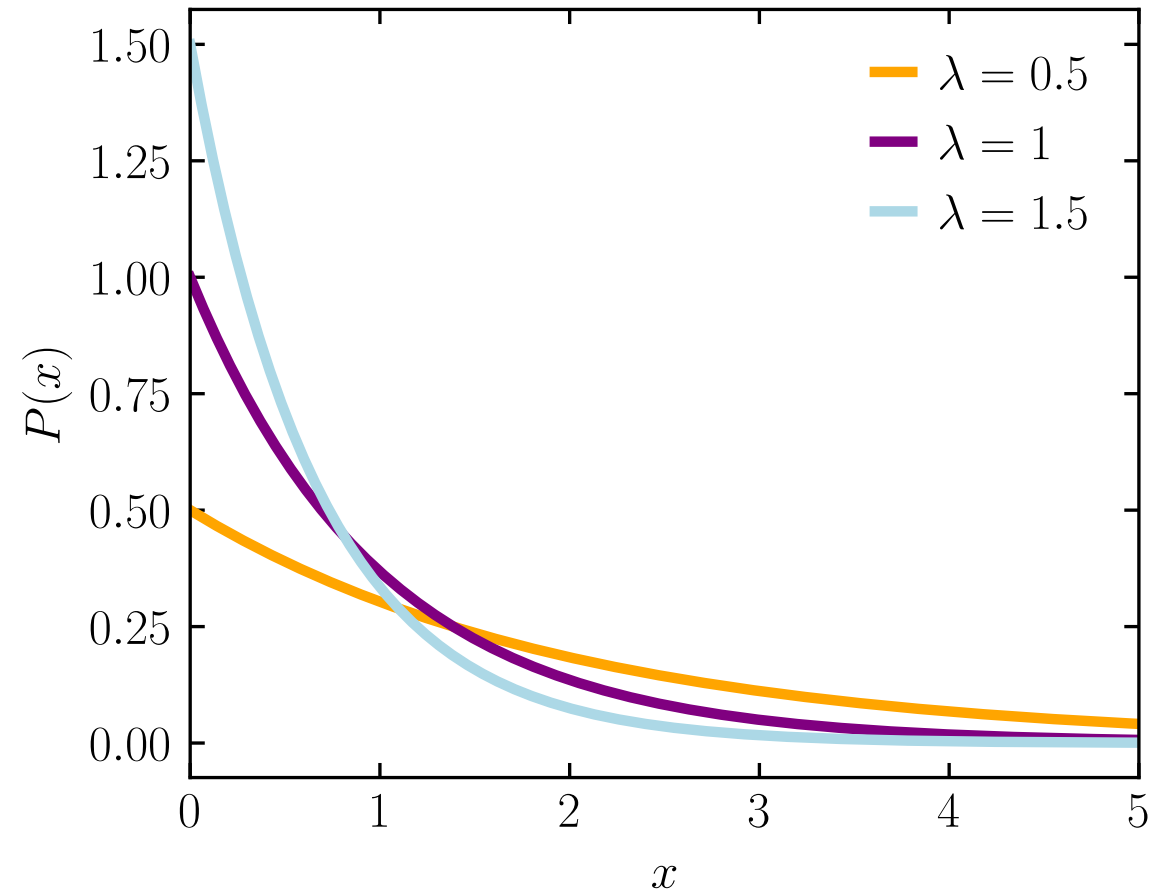


Image credit: WikipediaCC BY 4.0

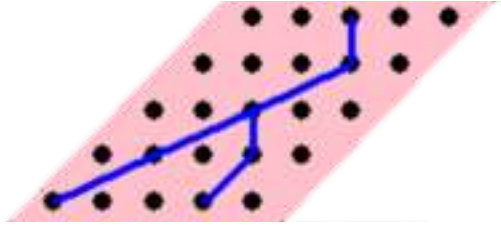
Take-Aways

The probability that 2 lineages coalesce after t generations is modeled as a geometrically distributed R.V., where the probability of success is $p = 1/2N_e$.

The expected # of generations until coalescence is thus $2N_e$ (so we wait longer for larger populations).

When p is small (i.e., $2N_e$ is sufficiently large), the geometric distribution can be approximated with an exponential distribution, which is done in **Kingman's Coalescent**.

Coalescence



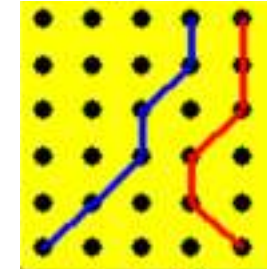
2 lineages enter branch & successfully coalesce

Probability 2 lineages coalescence on branch with t genes, each with effective population size of $2N_e$ (so $p = 1/2N_e$):

$$\sum_{i=1}^t (1-p)^{i-1} p = 1 - (1-p)^t \approx 1 - e^{-tp}$$

HINT: Look up CDF ;)

No coalescence

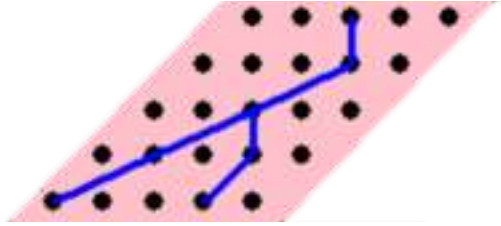


2 lineages enter branch & fail to coalesce

Probability that 2 lineages do NOT coalesce:

$$(1 - 1/2N_e)^t \approx e^{-tp}$$

Coalescence



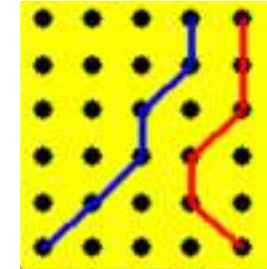
2 lineages enter branch & successfully coalesce

Probability 2 lineages coalescence on branch with t genes, each with effective population size of $2N_e$ (so $p = 1/2N_e$):

$$\sum_{i=1}^t (1-p)^{i-1} p = 1 - (1-p)^t \approx 1 - e^{-tp}$$

HINT: Look up CDF ;)

No coalescence

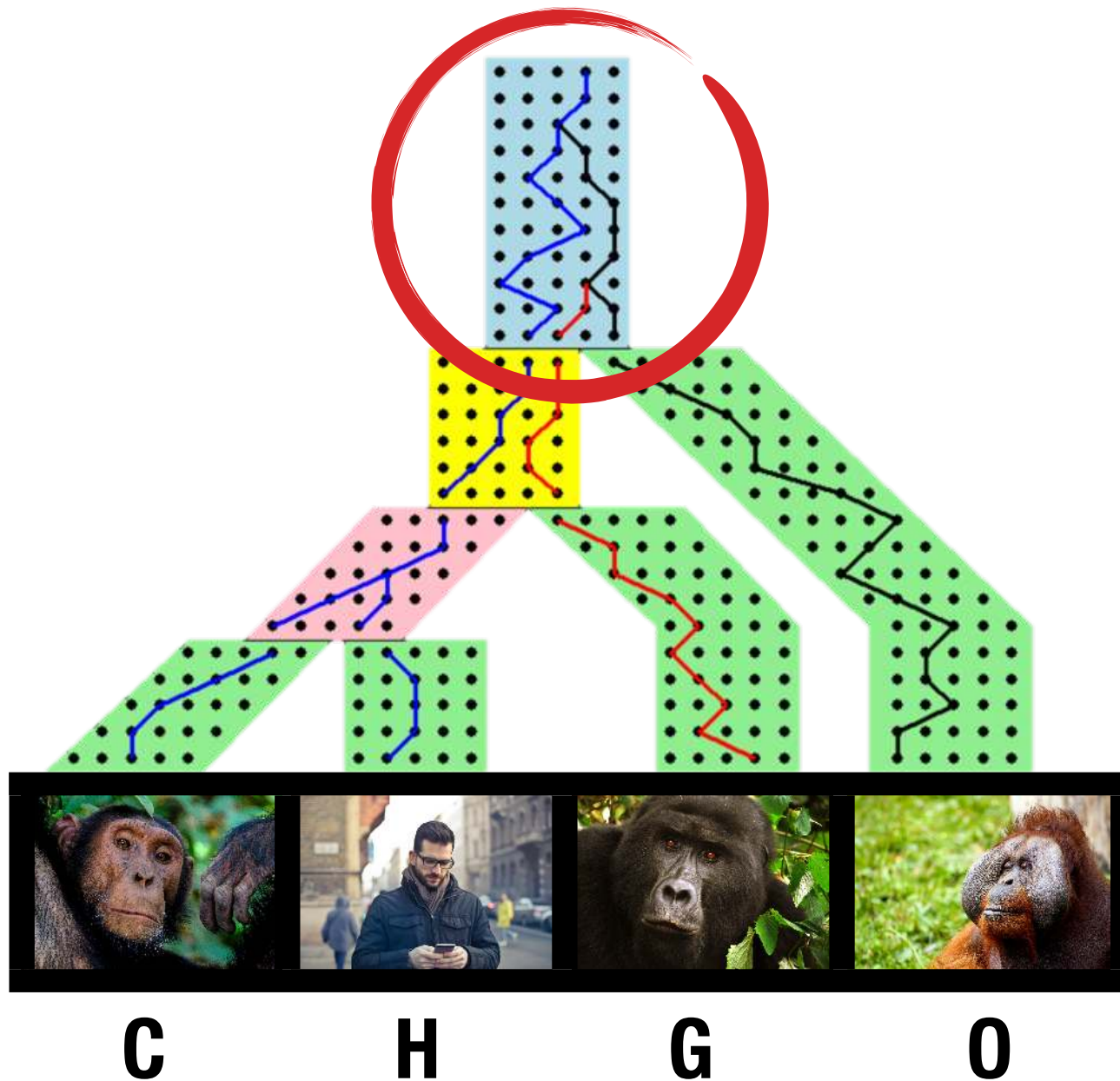


2 lineages enter branch & fail to coalesce

Probability that 2 lineages do NOT coalesce:

$$(1 - 1/2N_e)^t \approx e^{-tp}$$

Instead write as e^{-x} , where $x = t/2N_e$ is the length of branch in **coalescent units**!



Lastly, what happens when more than 2 lineages enter a branch?

Every pair coalesces with **equal** probability.

General Result

The **probability** that i lineages coalesce into j lineages on a branch of x coalescent units:

$$g_{i,j}(x) = \sum_{k=j}^i e^{-k(k-1)x} \frac{(2k-1)(-1)^{k-j}}{j!(k-j)!(j+k-1)} \prod_{m=0}^{k-1} \frac{(j+m)(i-m)}{i+m}$$

where $1 \leq j \leq i$ [[Tavare, 1984](#); [Rosenberg, 2002](#)].

This result allows us to compute the probability of gene trees given an MSC model species tree [[Rannala & Yang, 2003](#); [Degnan & Salter/Kubatko, 2005](#)]!

Also see book chapter by [Rannala, Edwards, Leache, and Yang](#).

Some useful equations

Let $\tau = \frac{t}{2N_e}$ be the length of a branch in the species tree in **coalescent units**. Then, the following will be useful for calculating gene tree probabilities:

2 lineages enter branch

$$g_{2,1}(\tau) = 1 - e^{-\tau}$$

$$g_{2,2}(\tau) = e^{-\tau}$$

3 lineages enter branch

$$g_{3,1}(\tau) = 1 - \frac{3}{2}e^{-\tau} + \frac{1}{2}e^{-3\tau}$$

$$g_{3,2}(\tau) = \frac{3}{2}e^{-\tau} - \frac{3}{2}e^{-3\tau}$$

$$g_{3,3}(\tau) = e^{-3\tau}$$

Now we are ready to compute the probability of a gene tree given an MSC model species tree.

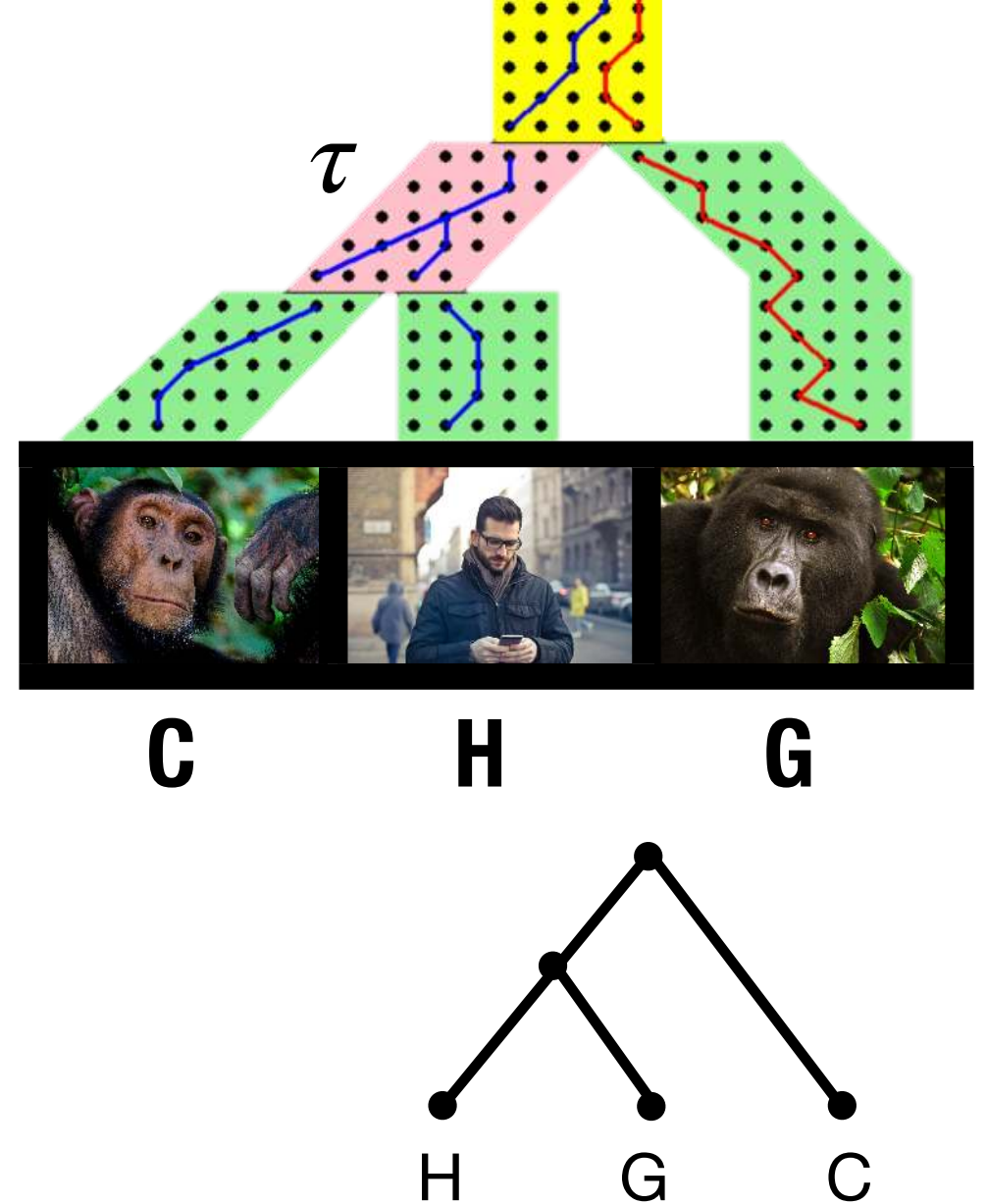
Consider an MSC model species tree with 3 taxa.

Q: What is probability of “(H,G),C);”?

This gene tree must be generated with the following events:

- (1) Lineages h & c enter **internal branch** and FAIL to coalesce on it
- (2) Lineages h, c, & g coalesce enter **above root “branch”** and h & g coalesce first

What is the probability of (1) and (2)?



Consider an MSC model species tree with 3 taxa.

Q: What is probability of “(H,G),C);”?

This gene tree must be generated with the following events:

(1) Lineages h & c enter **internal branch** and FAIL to coalesce on it

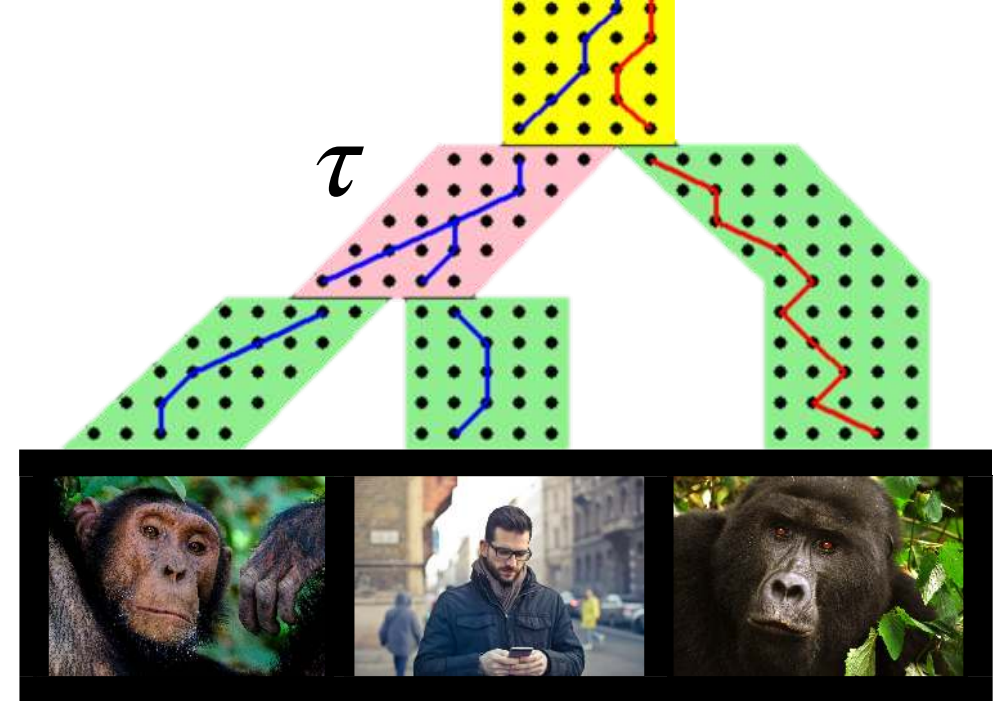
$$g_{2,2}(\tau) = e^{-\tau}$$

(2) Lineages h, c, & g coalesce enter **above root “branch”** and h & g coalesce first

$$\frac{1}{\binom{3}{2}} = \frac{1}{3}$$

Putting it together:

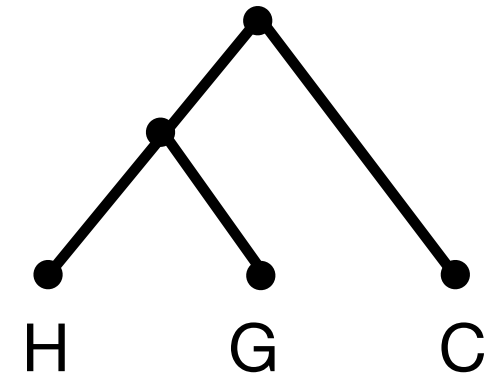
$$P(hg) = \frac{1}{3}e^{-\tau}$$



C

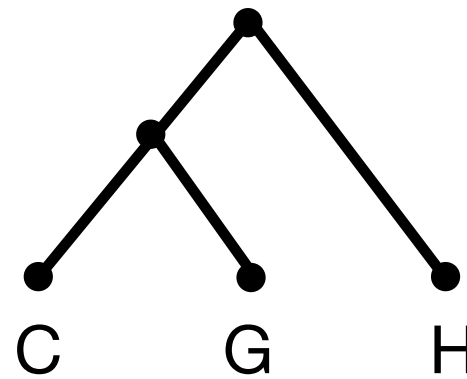
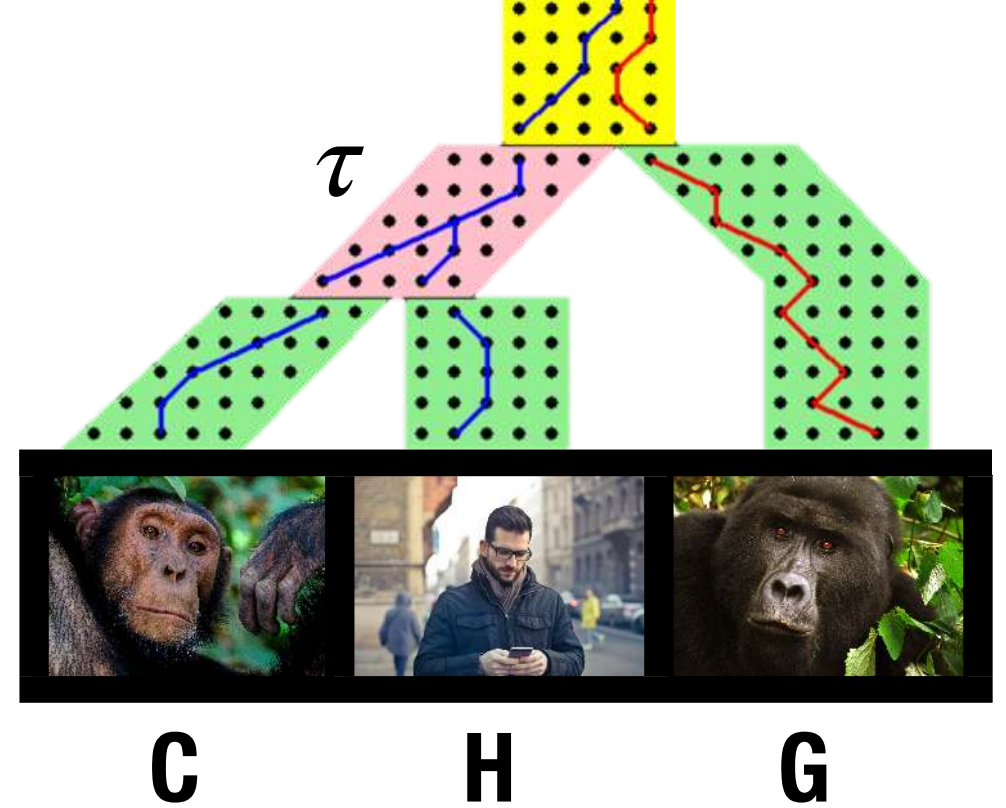
H

G

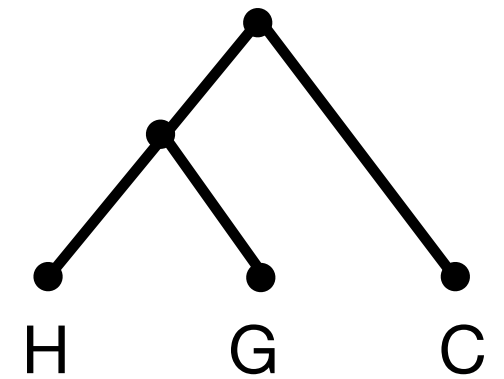


Consider an MSC model species tree with 3 taxa.

Continuing in this fashion, we get the probability distribution:



$$P(cg) = \frac{1}{3}e^{-\tau}$$

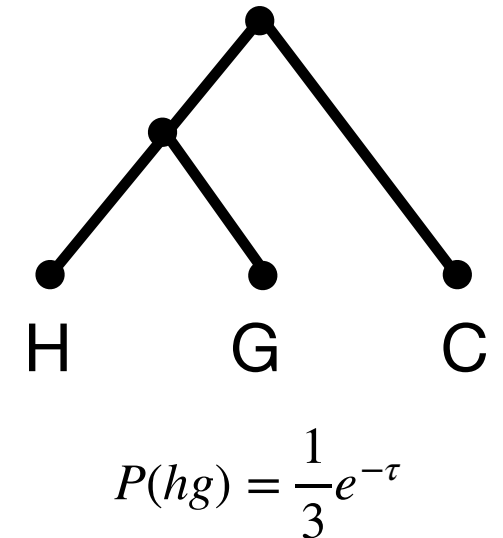
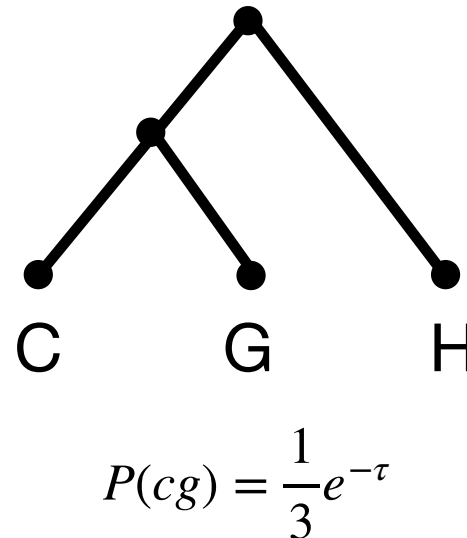
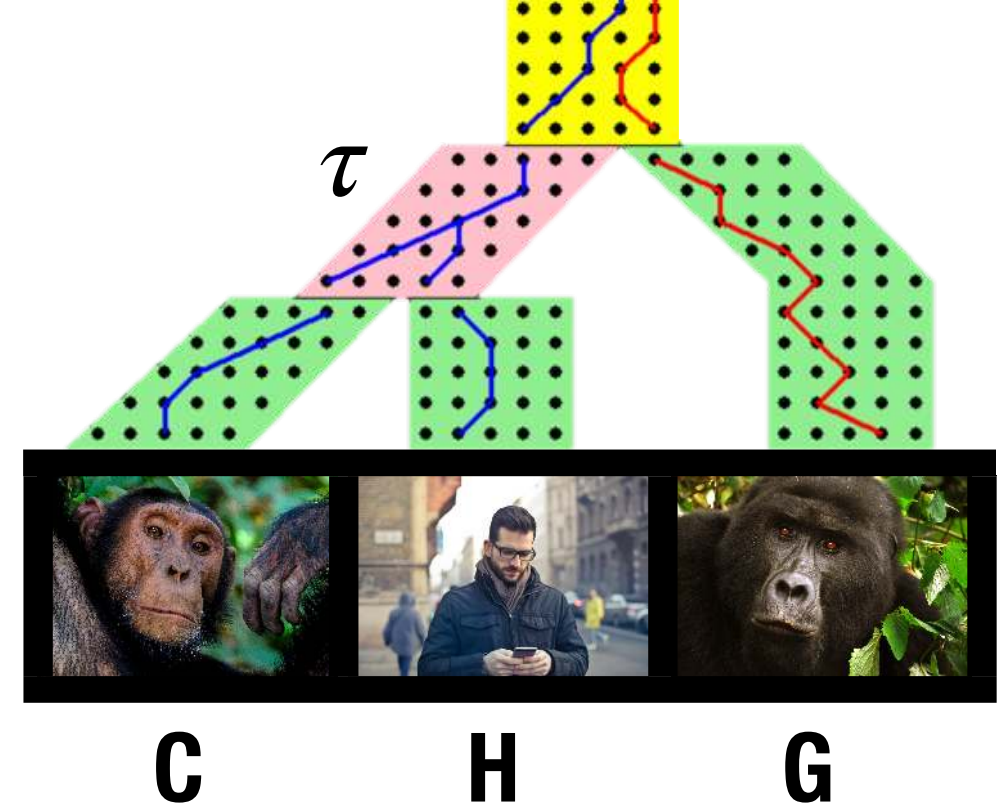
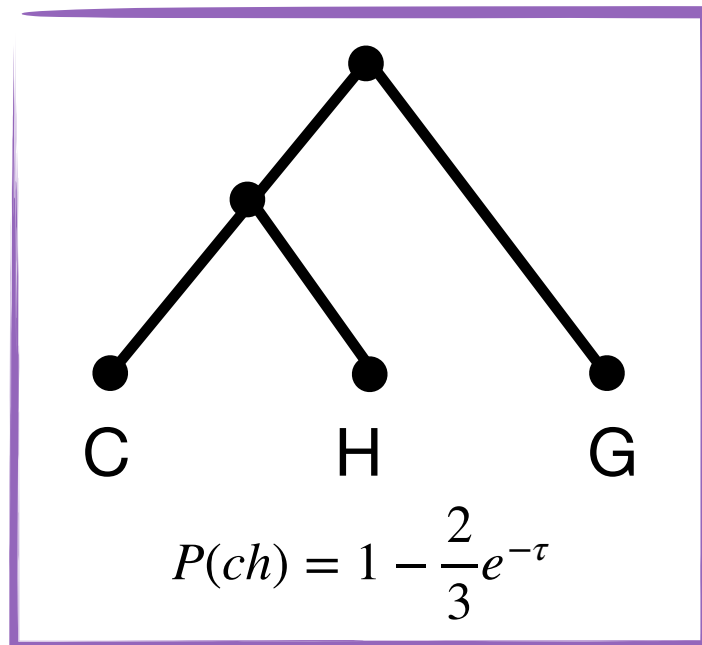


$$P(hg) = \frac{1}{3}e^{-\tau}$$

Consider an MSC model species tree with 3 taxa.

Continuing in this fashion, we get the probability distribution:

Same topology as species tree



Anomaly Zone

Definition. A gene tree is **anomalous** if has higher probability under the MSC than the gene tree with the same topology as the species tree.

Definition. A species tree is in the anomaly zone if it has an **anomalous** gene tree.

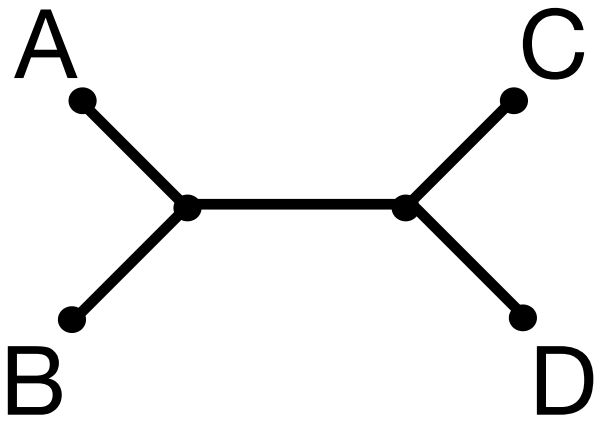
Result. No anomalous **triplets** (rooted 3-leaf trees) or **quartets** (unrooted 4-leaf trees).

[[Degnan & Rosenberg, 2006](#); [Allman, Degnan, Rhodes, 2011](#); [Degnan, 2013](#)]

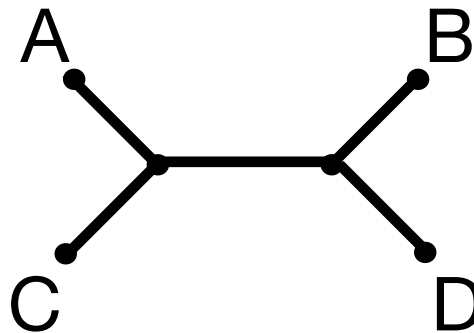
Triplets vs. Quartets

Result. No anomalous **triplets** (rooted 3-leaf trees) or **quartets** (unrooted 4-leaf trees).

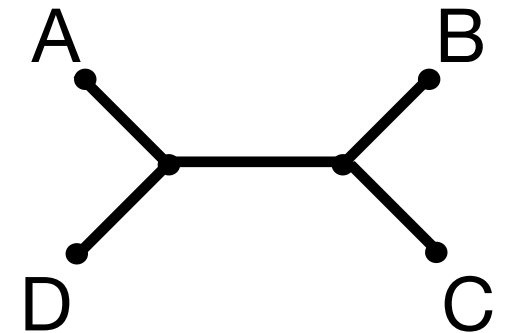
Same topology as unrooted
species tree



$$P(ch) = 1 - \frac{2}{3}e^{-\tau}$$

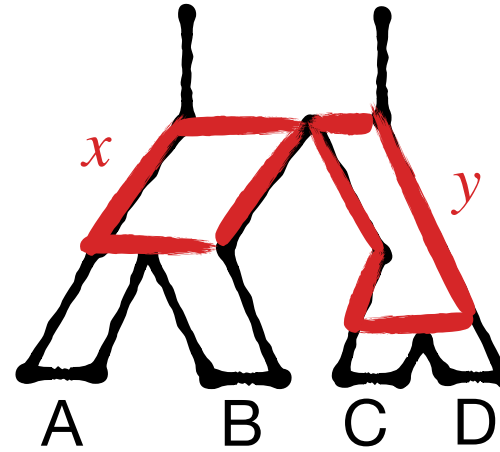


$$P(cg) = \frac{1}{3}e^{-\tau}$$



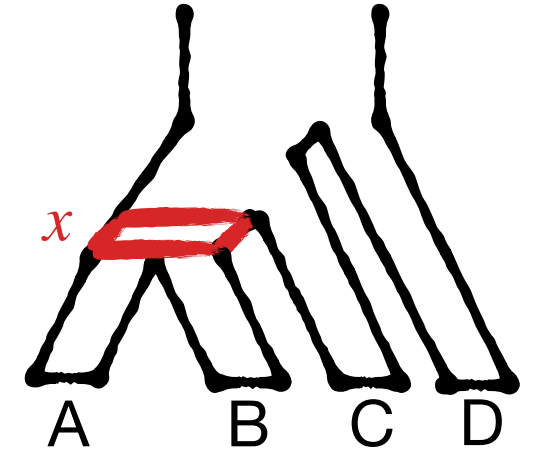
$$P(co) = \frac{1}{3}e^{-\tau}$$

Probability
distribution for
both **balanced** &
pectinate
species trees



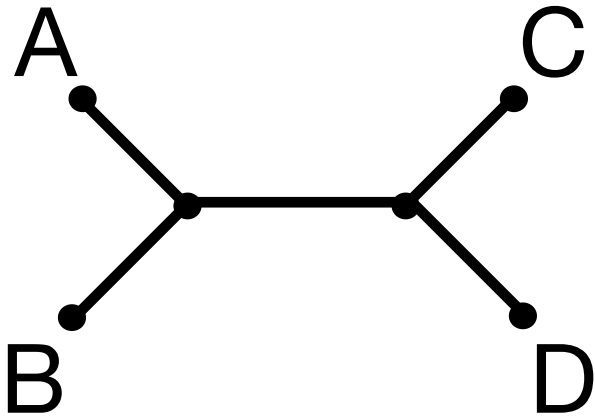
Balanced species tree

$$\tau = x + y$$

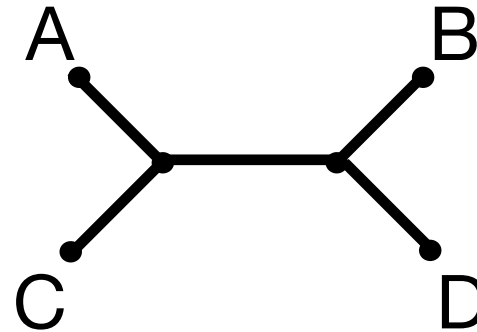


Pectinate species tree

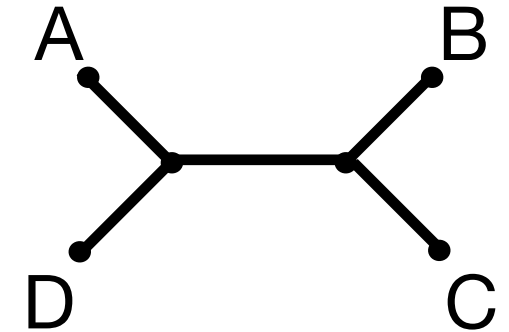
$$\tau = x$$



$$P(ch) = 1 - \frac{2}{3}e^{-\tau}$$



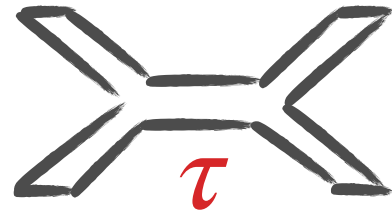
$$P(cg) = \frac{1}{3}e^{-\tau}$$



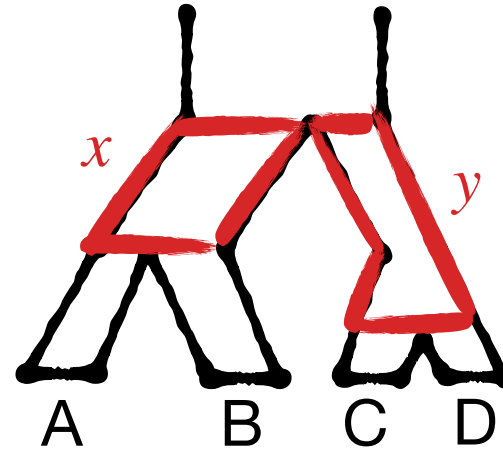
$$P(co) = \frac{1}{3}e^{-\tau}$$

Probability
distribution for
both **balanced** &
pectinate
species trees

τ is length of internal branch
in CUs

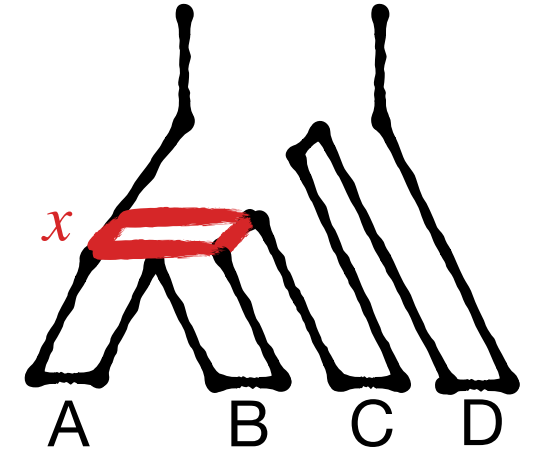


Unrooted
species tree



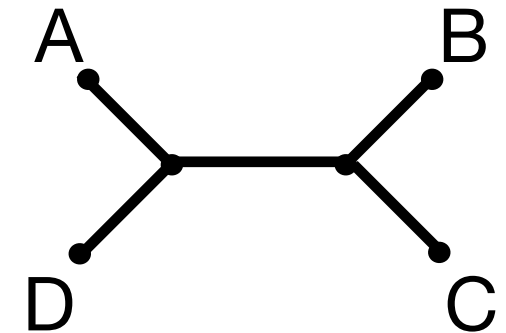
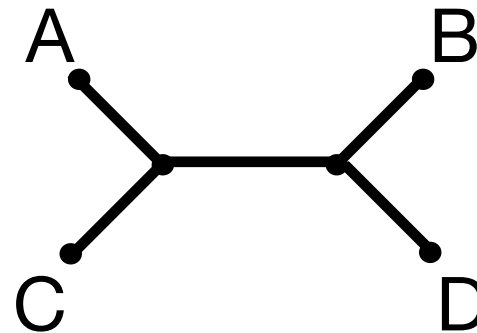
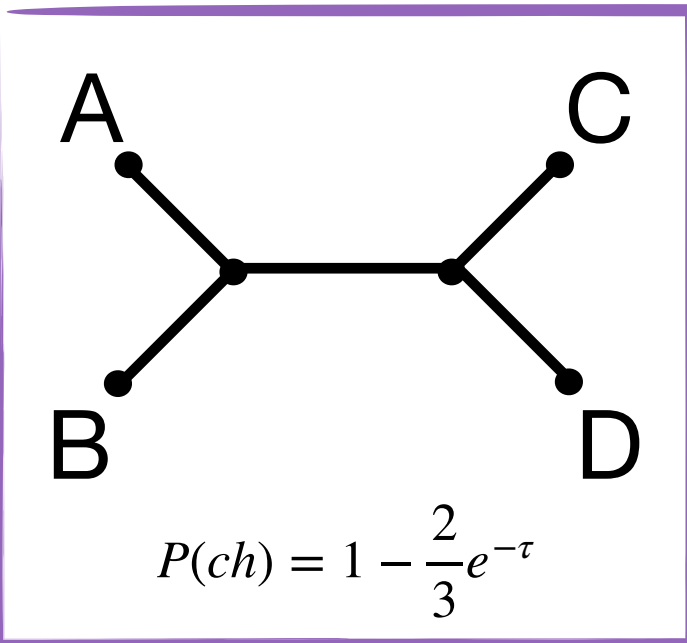
Balanced species tree

$$\tau = x + y$$



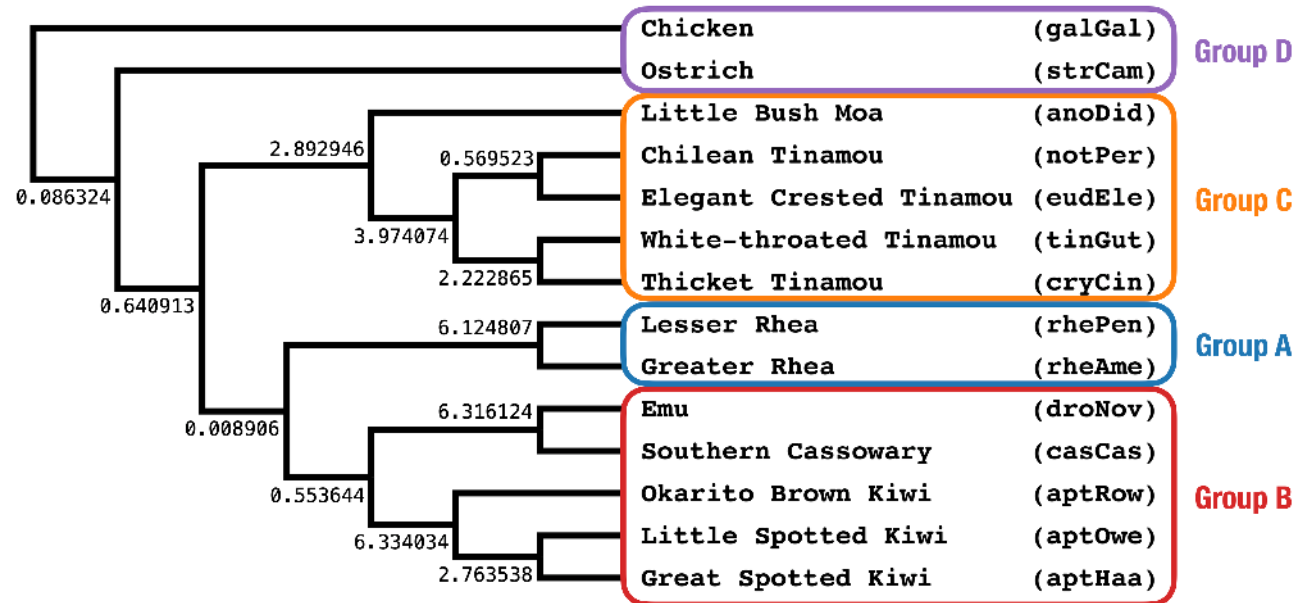
Pectinate species tree

$$\tau = x$$



Activity B

How does
branch
length (in CUs)
impact gene
tree discordance?



20 minutes

<https://github.com/molloy-lab/ck-phylo-workshop>

Coalescent Lab — Day 1

Motivation for coalescent methods (**Activity A**)

Coalescent basics (**Activity B**)

Species tree estimation with summary methods (**Activity C**)

Evaluation model fit (**Activity D** — optional / do tomorrow)

IS A NEW AND GENERAL THEORY OF MOLECULAR SYSTEMATICS EMERGING?

Scott V. Edwards^{1,2}

¹Museum of Comparative Zoology and Department of Organismic & Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, Massachusetts 02138

²E-mail: sedwards@fas.harvard.edu

Received September 30, 2008

Accepted October 1, 2008

The advent and maturation of algorithms for estimating species trees—phylogenetic trees that allow gene tree heterogeneity and whose tips represent lineages, populations and species, as opposed to genes—represent an exciting confluence of phylogenetics, phylogeography, and population genetics, and ushers in a new generation of concepts and challenges for the molecular systematist. In this essay I argue that to better deal with the large multilocus datasets brought on by phylogenomics, and to better align the fields of phylogeography and phylogenetics, we should embrace the primacy of species trees, not only as a new and useful practical tool for systematics, but also as a long-standing conceptual goal of systematics that, largely due to the lack of appropriate computational tools, has been eclipsed in the past few decades. I suggest that phylogenies as gene trees are a “local optimum” for systematics, and review recent advances that will bring us to the broader optimum inherent in species trees. In addition to adopting new methods of phylogenetic analysis (and ideally reserving the term “phylogeny” for species trees rather than gene trees), the new paradigm suggests shifts in a number of practices, such as sampling data to maximize not only the number of accumulated sites but also the number of independently segregating genes; routinely using coalescent or other models in computer simulations to allow gene tree heterogeneity; and understanding better the role of concatenation in influencing topologies and confidence in phylogenies. By building on the foundation laid by concepts of gene trees and coalescent theory, and by taking cues from recent trends in multilocus phylogeography, molecular systematics stands to be enriched. Many of the challenges and lessons learned for estimating gene trees will carry over to the challenge of estimating species trees, although adopting the species tree paradigm will clarify many issues (such as the nature of polytomies and the star tree paradox), raise conceptually new challenges, or provide new answers to old questions.

KEY WORDS: Fossil, genome, macroevolution, Neanderthal, phylogeography, polytomy.

The title of this essay is borrowed from one of the famous essays written by Stephen Jay Gould, “Is a new and general theory of evolution emerging?”, published in *Paleobiology* in 1980 (Gould 1980). Gould was speculating as to whether the constellation of observations and trends from the fossil record and developmental biology, collectively known as “macroevolution,” might constitute a genuinely new set of phenomena, a set that had not been covered adequately by the reigning paradigm of Darwinian microevolution. Of course whether one answers Gould’s question in the positive or negative depends on one’s perspective; although

Gould and others would not have raised the question unless one could answer “yes,” many evolutionary biologists have argued that the quantitative framework provided by microevolution can adequately account for the observations of punctuation, stasis, and apparent saltation that had suggested a new paradigm to some (Charlesworth et al. 1982; Smith 1983; Estes and Arnold 2007). Yet there is a pervasive feeling that the paradigms laid down by the Modern Synthesis still may not adequately capture the plethora of phenomena ushered in by modern evolutionary biology (Erwin 2000; Pigliucci 2007). Although the paradigm that I question is

Scott Edwards
championed the
coalescent in
systematics **15 years**
ago...



...and now coalescent
methods are super
popular!

**Many
coalescent
methods
have been
developed
in last
decade!**

Gene tree summary methods for unrooted trees

e.g. [BUCKy](#) (-pop), [NJst](#) / [USTAR](#) / [ASTRID](#), [ASTRAL](#) / [ASTER](#), [TREE-QMC](#), [wQFM](#)

Gene tree summary methods for rooted trees

e.g. [MDC](#), [STEM](#), [MP-EST](#)

Site-based methods

e.g. [SNAPP](#), [SVDQuartets](#), [CASTER](#)

Bayesian co-estimation methods (co-estimate gene trees & species tree)

e.g. [*BEAST](#), [StarBEAST2](#)

Also see methods based on population allele frequencies

e.g. [PoMo](#), implemented in [RevBayes](#)

**Many
coalescent
methods
have been
developed
in last
decade!**

Gene tree summary methods for unrooted trees

e.g. [BUCKy](#) (-pop), [NJst](#) / [USTAR](#) / [ASTRID](#), [ASTRAL](#) / [ASTER](#), [TREE-QMC](#), [wQFM](#)

Gene tree summary methods for rooted trees

e.g. [MDC](#), [STEM](#), [MP-EST](#)

Site-based methods

e.g. [SNAPP](#), [SVDQuartets](#), [CASTER](#)

Bayesian co-estimation methods (co-estimate gene trees & species tree)

e.g. [*BEAST](#), [StarBEAST2](#)

Also see methods based on population allele frequencies

e.g. [PoMo](#), implemented in [RevBayes](#)

The Summary Method Approach

Step 0.

Lots of data processing!!!!

Step 1.

Estimate gene trees.

Sequence data
(alignments)

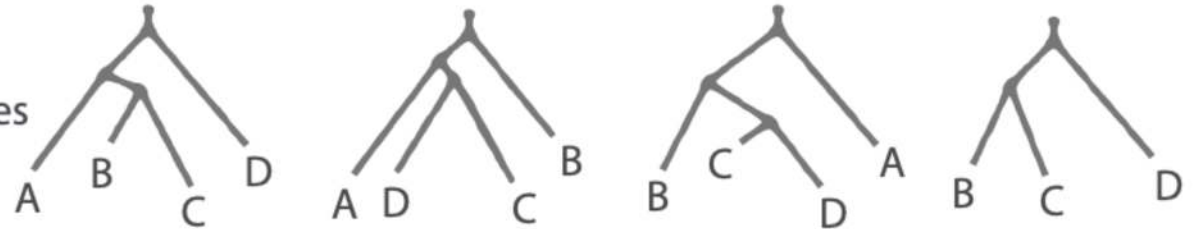
```
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG
```

```
CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G
```

```
AGCAGCATCGTG
AGCAGC-TCGTG
AGCAGC-TC-TG
C-TA-CACGGTG
```

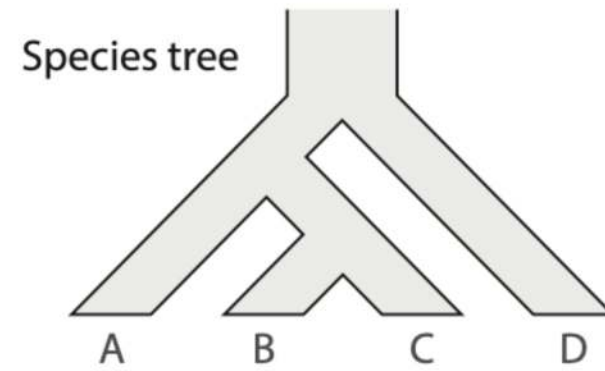
```
GGCACGCACGAA
C-CACGC-CATA
GGCACGC-C-TA
```

Gene trees



Step 2.

Estimate species tree from gene trees under MSC.



Many popular coalescent methods are based on triplets or quartets

1. Very fast to compute likelihood for 3 or 4 taxa (unlike larger #'s of taxa) + species tree that maximizes **pseudo-likelihood** is consistent estimator*

See [MP-EST](#) for triplets & [PhyloNetworks](#) for quartets!

2. Species tree that maximizes **triplet score** or **quartet score** is consistent estimator* + fast & accurate heuristics for these optimization problems

See [STELAR](#) for triplets & [ASTRAL](#) / [ASTER](#) or [TREE-QMC](#) for quartets.

*Assumes error-free gene trees (or sequence length is unbounded); see [Roch, Nute & Warnow, 2018](#)

Many popular coalescent methods are based on triplets or quartets

1. Very fast to compute likelihood for 3 or 4 taxa (unlike larger #'s of taxa) + species tree that maximizes **pseudo-likelihood** is consistent estimator*

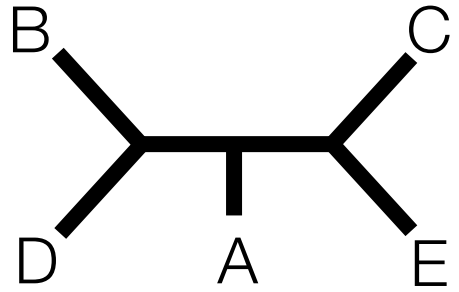
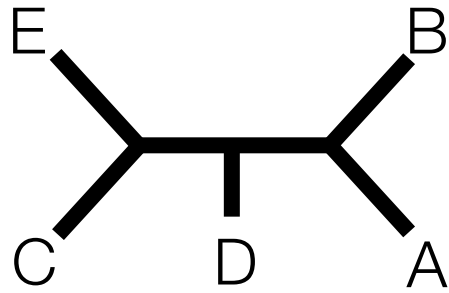
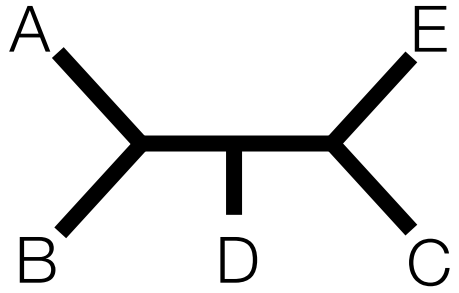
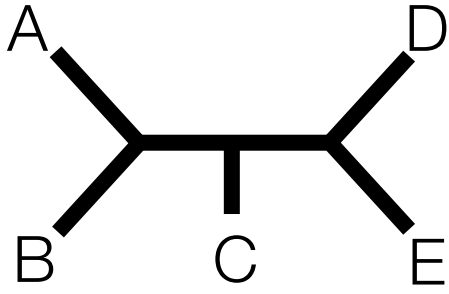
See [MP-EST](#) for triplets & [PhyloNetworks](#) for quartets!

2. Species tree that maximizes **triplet score** or **quartet score** is consistent estimator* + fast & accurate heuristics for these optimization problems

See [STELAR](#) for triplets & [ASTRAL](#) / [ASTER](#) or [TREE-QMC](#) for quartets.

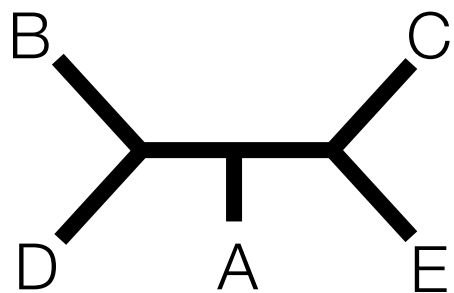
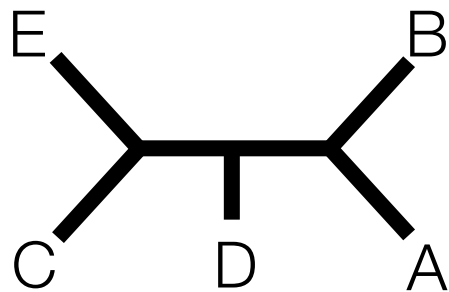
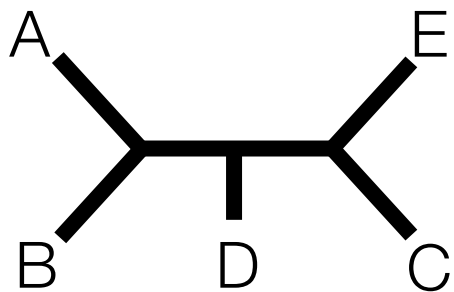
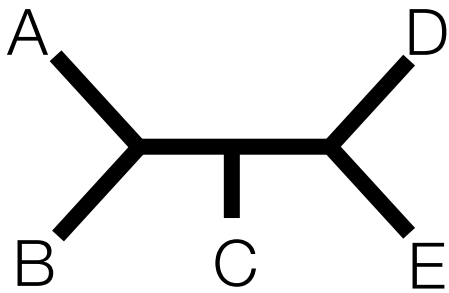
*Assumes error-free gene trees (or sequence length is unbounded); see [Roch, Nute & Warnow, 2018](#)

Input Data:



...

Input Data:



■ ■ ■

X,Y,Z,W

$$X, Y | Z, W$$
$$X, Z | Y, W$$

X,W|Y,Z

A,B,C,D

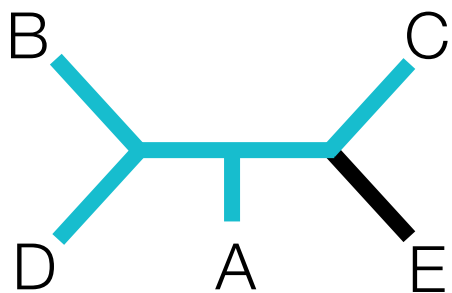
A,B,C,E

A,B,D,E

A,C,D,E

B,C,D,E

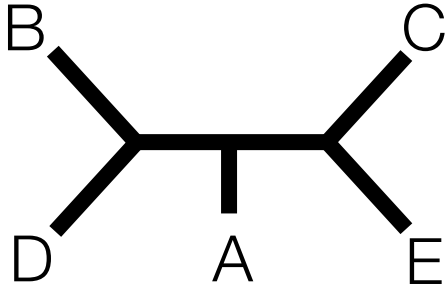
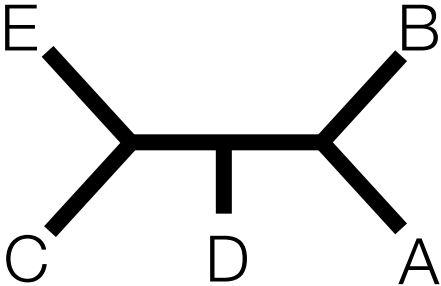
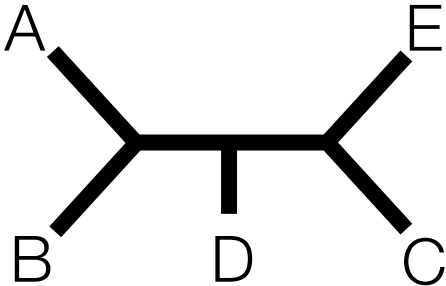
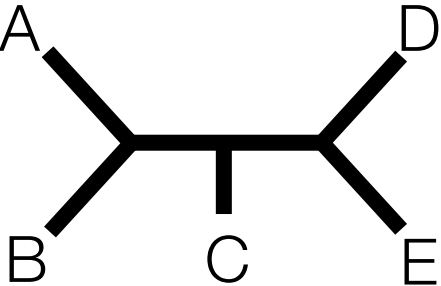
Input Data:



■ ■ ■

X,Y,Z,W	X,Y Z,W	X,Z Y,W	X,W Y,Z
A,B,C,D	3	1	0
A,B,C,E			
A,B,D,E			
A,C,D,E			
B,C,D,E			

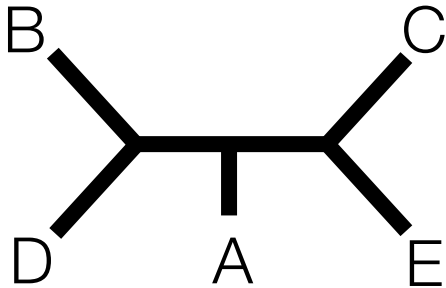
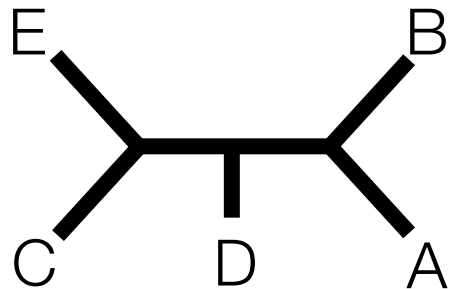
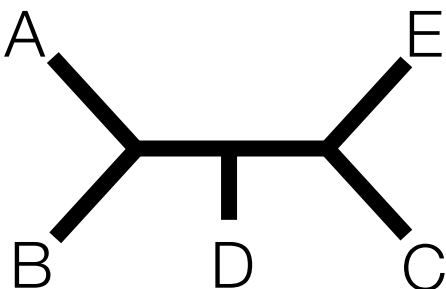
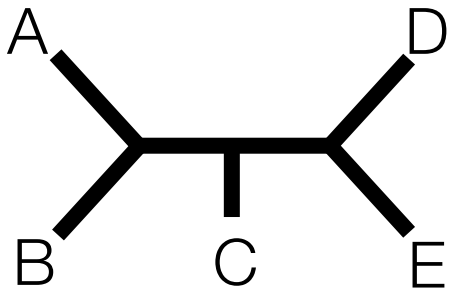
Input Data:



...

X,Y,Z,W	X,Y Z,W	X,Z Y,W	X,W Y,Z
A,B,C,D	3	1	0
A,B,C,E	4	0	0
A,B,D,E	3	0	1
A,C,D,E	1	3	0
B,C,D,E	1	3	0

Input Data:



...

X,Y,Z,W

X,Y|Z,W

X,Z|Y,W

X,W|Y,Z

A,B,C,D

3

1

0

A,B,C,E

4

0

0

A,B,D,E

3

0

1

A,C,D,E

1

3

0

B,C,D,E

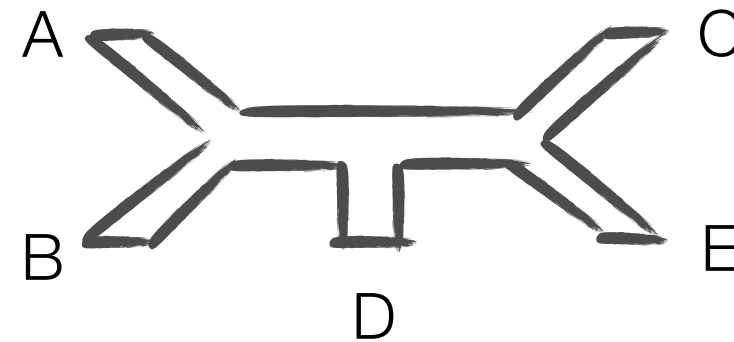
1

3

0

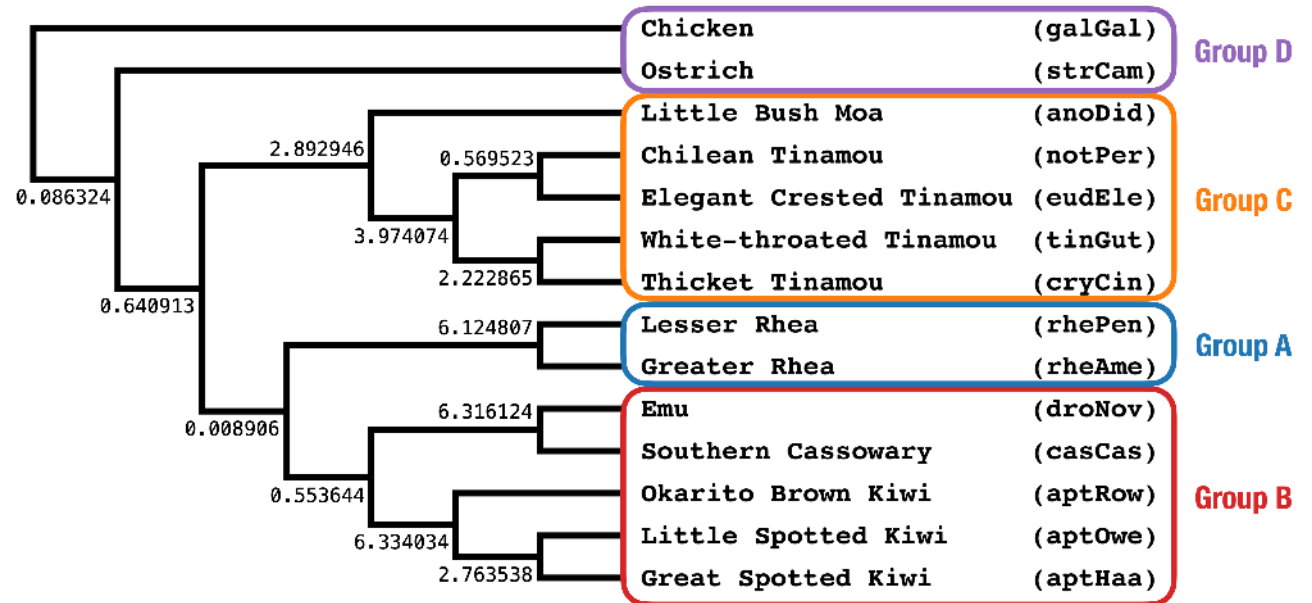
Solution:

Quartet score = 14



Activity C

At last, let's
estimate
species trees
from (simulated)
gene trees!!!



20 minutes

<https://github.com/molloy-lab/ck-phylo-workshop>

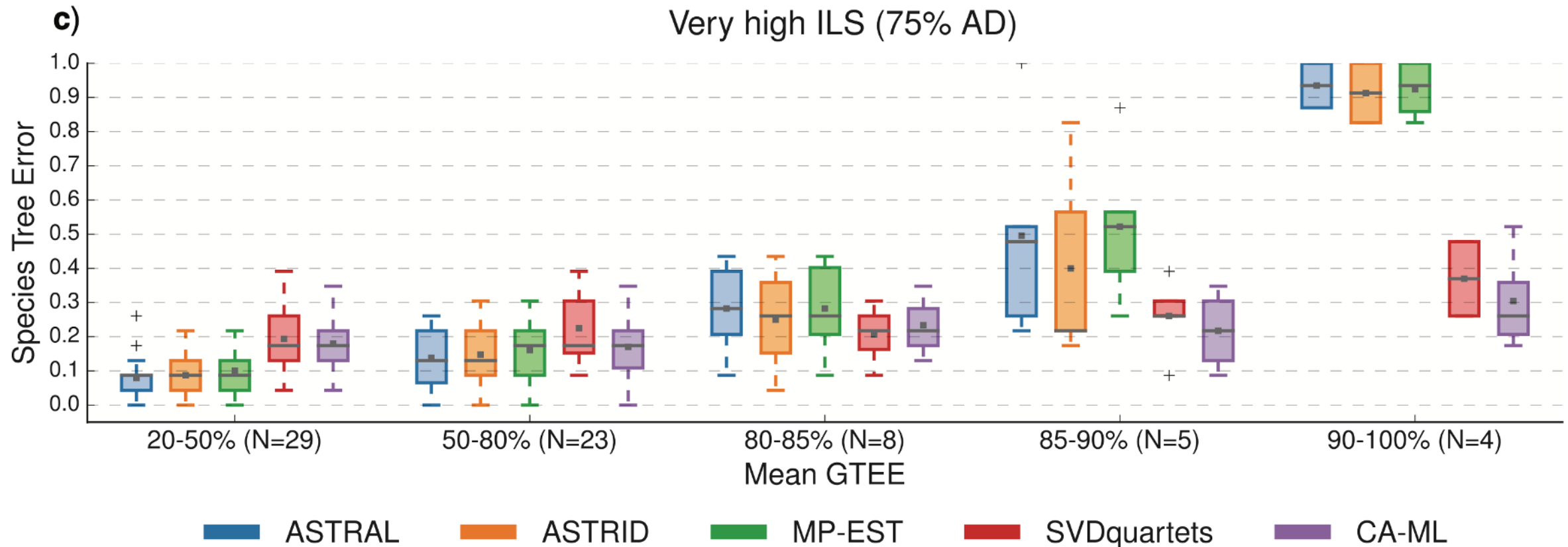
Discussion Questions

1. Are you concerned about incomplete lineage sorting in your system?
2. How would you evaluate whether ILS was a potential problem?

Don't forget — we made some assumptions:

- Gene trees evolve independently within the same model species tree (no linkage!!)
- Coalescent events in different populations are independent
- All pairs of lineages in a population are equally likely to coalesce
- Assumptions of Kingman's coalescent, e.g., no population structure (within a branch), no selection, etc.
- No intra-locus recombination (otherwise evolutionary history for gene is NOT a tree)
- No gene flow (otherwise evolutionary history of species is NOT a tree)
- No genome / gene duplication or gene loss
- **Error-free, complete gene trees — perfect data!**

So what's better concatenation or summary method?



Coalescent Lab — Day 1

Motivation for coalescent methods (**Activity A**)

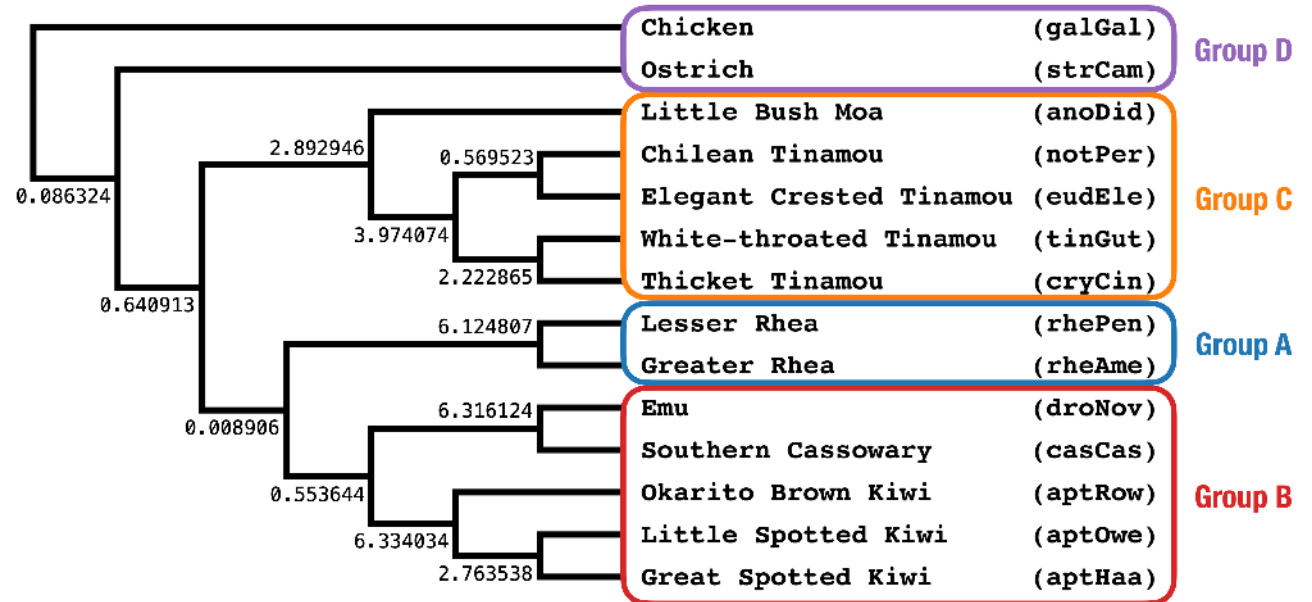
Coalescent basics (**Activity B**)

Species tree estimation with summary methods (**Activity C**)

Evaluation model fit (**Activity D** — optional / do tomorrow)

Activity D (Optional)

Let's use
pseudo-
likelihood to
evaluate
model fit!



remaining time

<https://github.com/molloy-lab/ck-phylo-workshop>

Many popular coalescent methods are based on triplets or quartets

1. Very fast to compute likelihood for 3 or 4 taxa (unlike larger #'s of taxa) + species tree that maximizes **pseudo-likelihood** is consistent estimator*

See [MP-EST](#) for triplets & [PhyloNetworks](#) for quartets!

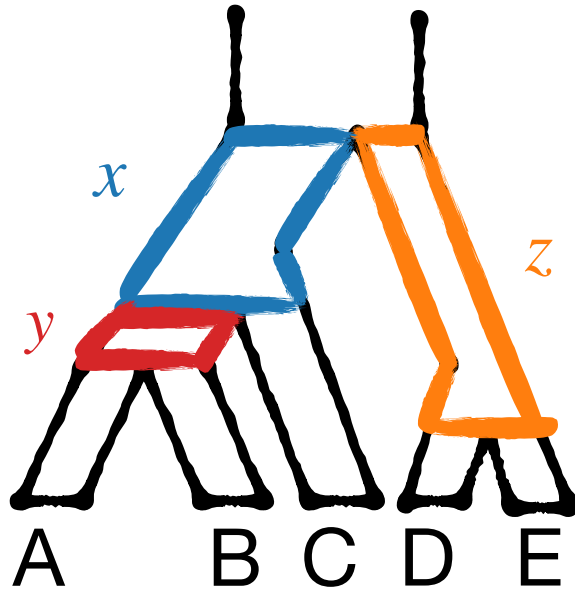
2. Species tree that maximizes **triplet score** or **quartet score** is consistent estimator* + fast & accurate heuristics for these optimization problems

See [STELAR](#) for triplets & [ASTRAL](#) / [ASTER](#) or [TREE-QMC](#) for quartets.

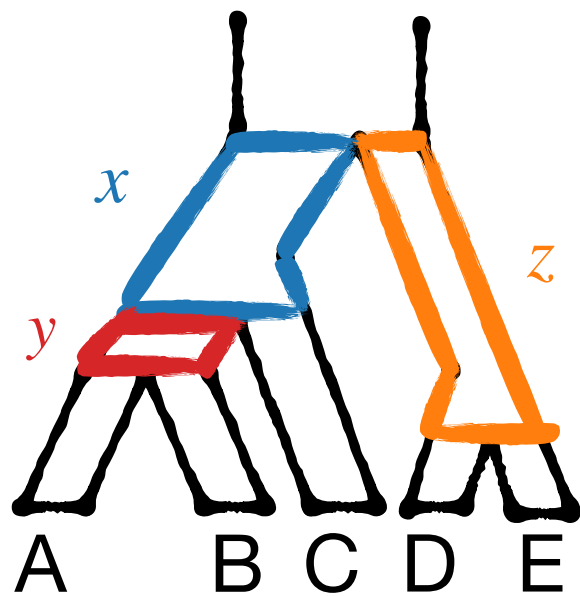
*Assumes error-free gene trees (or sequence length is unbounded); see [Roch, Nute & Warnow, 2018](#)

Model

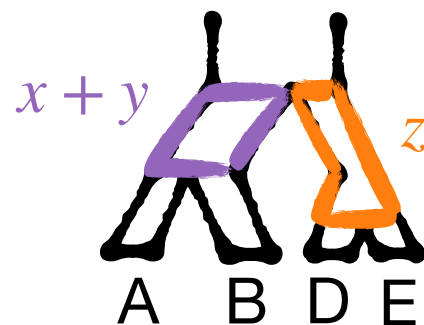
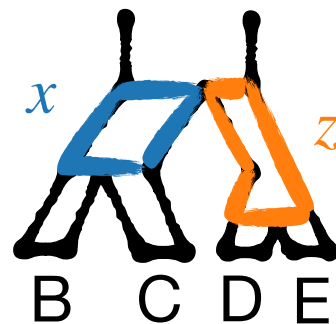
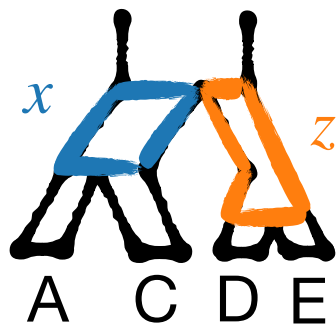
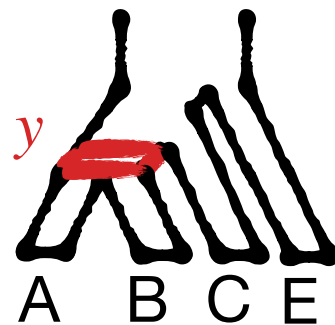
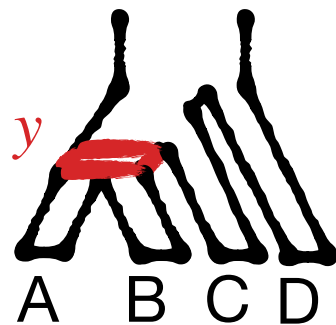
Image Credit: Rooted species trees adapted from
[Allman, Degnan & Rhodes, 2011](#)



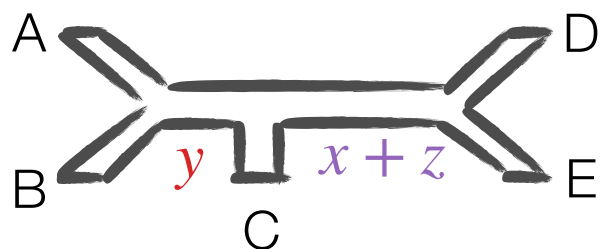
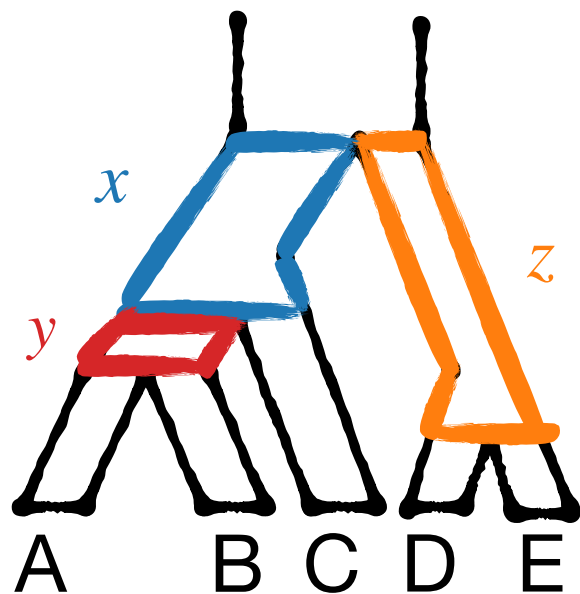
Model



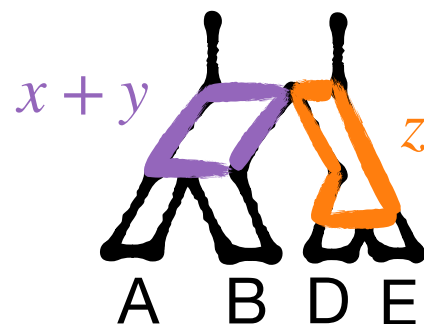
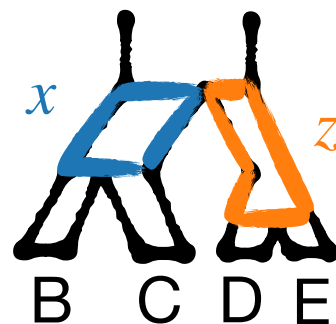
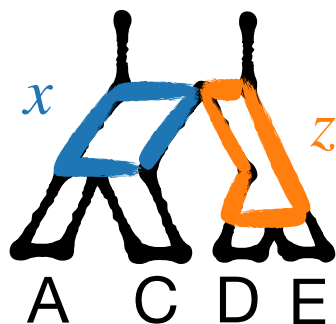
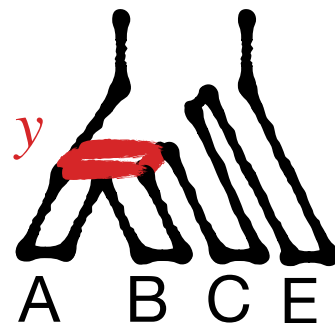
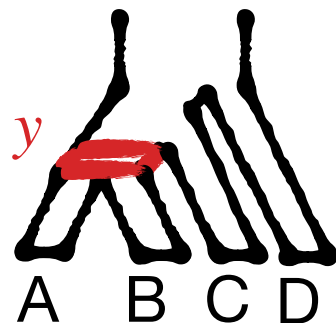
Rooted sub-models



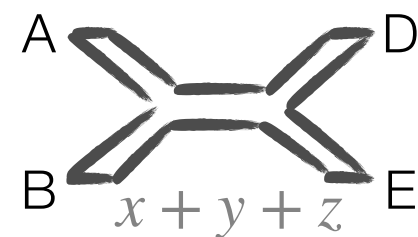
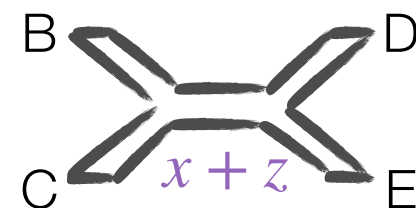
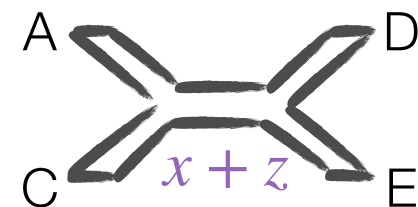
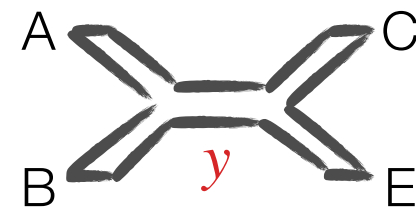
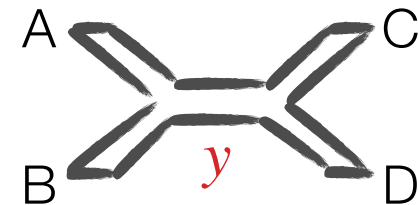
Model



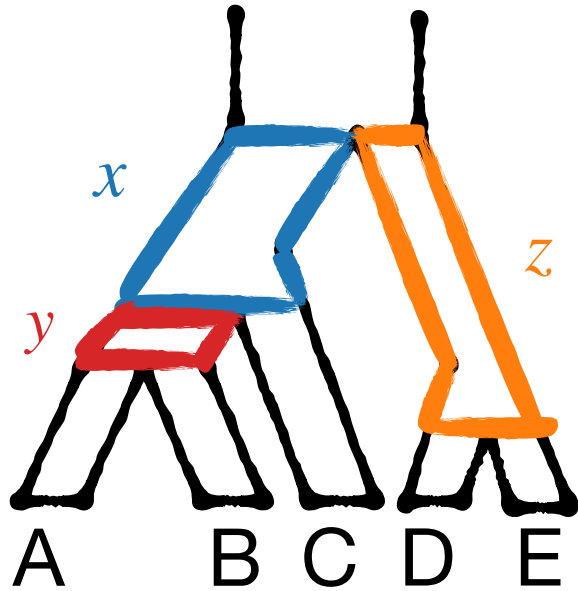
Rooted sub-models



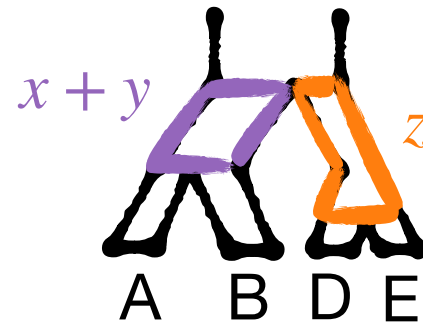
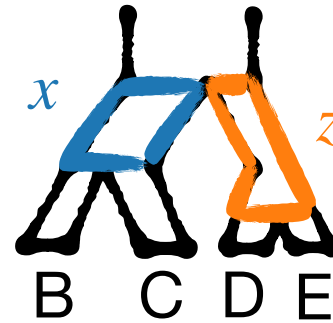
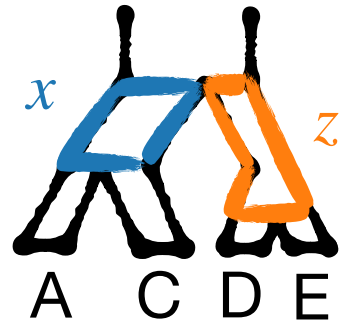
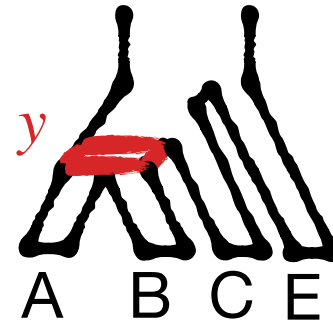
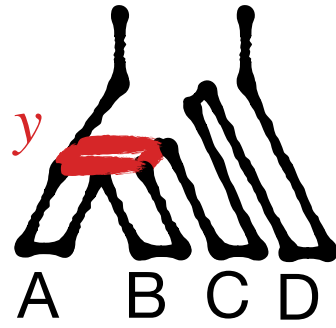
Unrooted sub-models



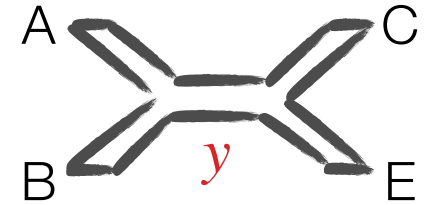
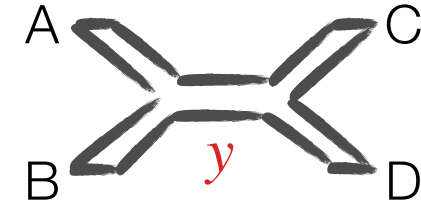
Model



Rooted sub-models

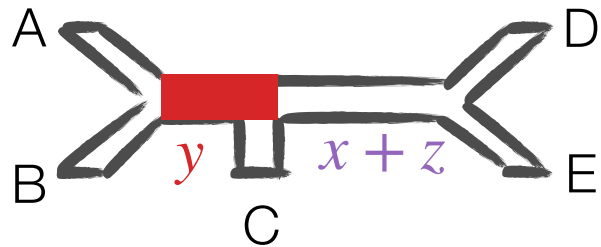


Unrooted sub-models

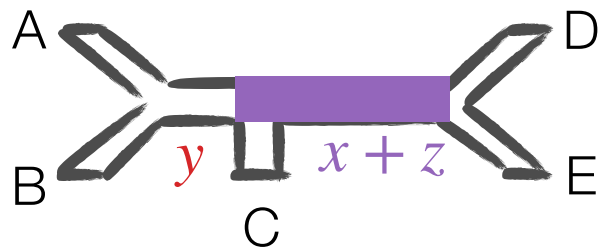
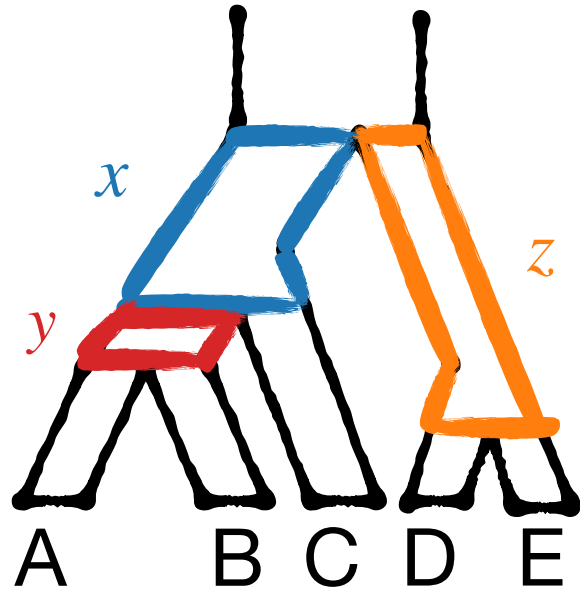


IMPORTANT:

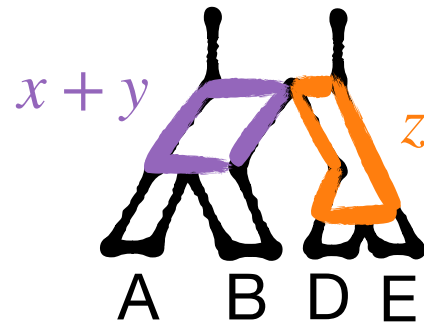
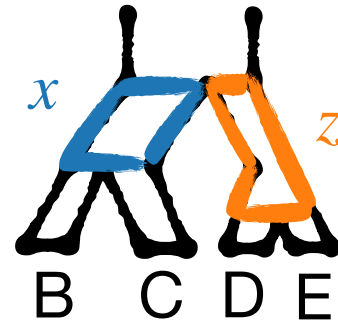
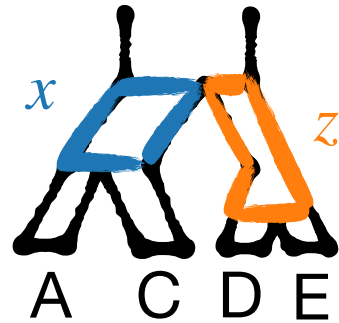
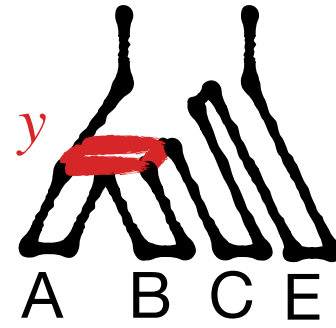
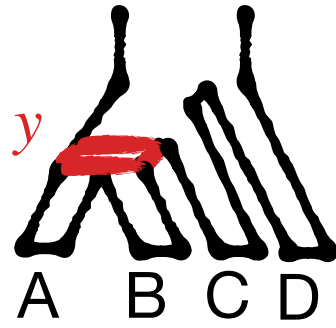
First 2 quartets are
“**around**” same branch in
unrooted model tree



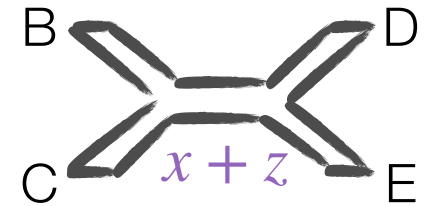
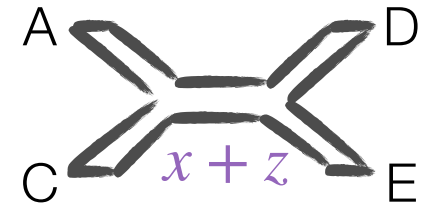
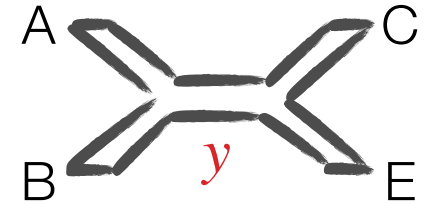
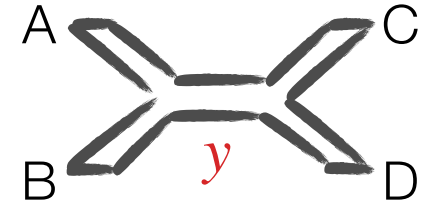
Model



Rooted sub-models



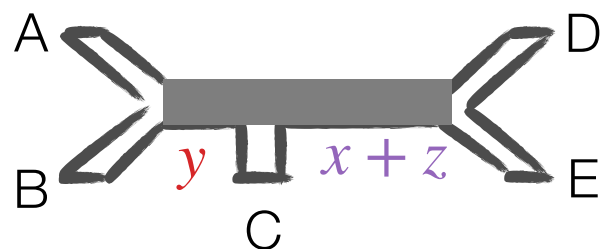
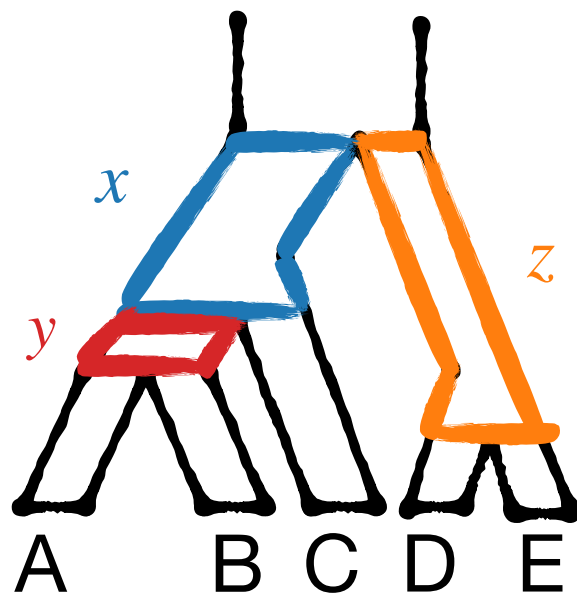
Unrooted sub-models



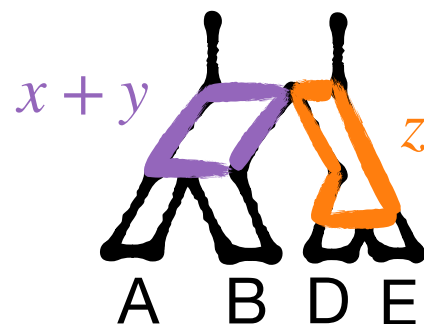
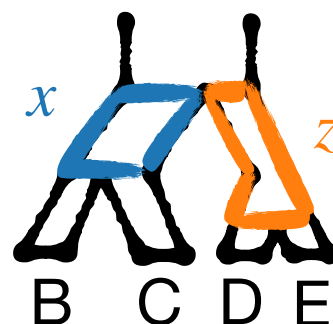
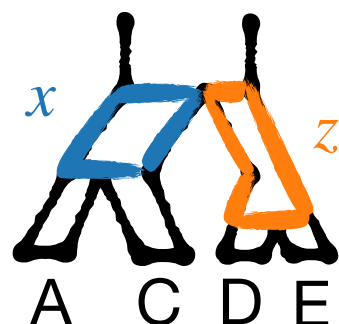
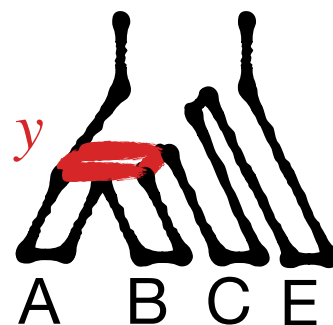
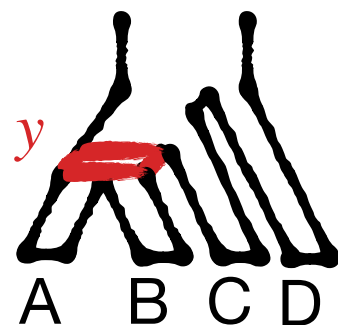
IMPORTANT:

Second 2 quartets are
“around” same branch

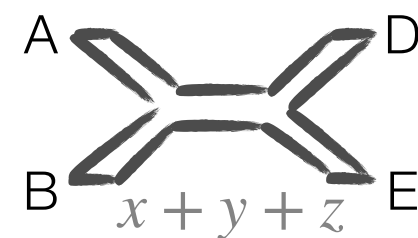
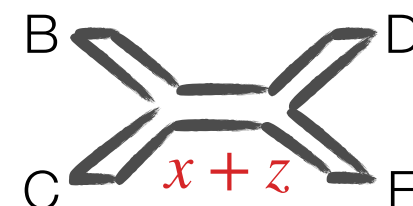
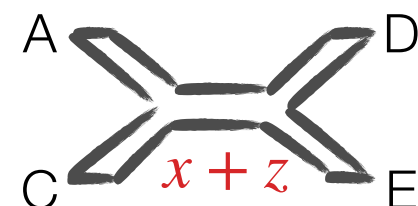
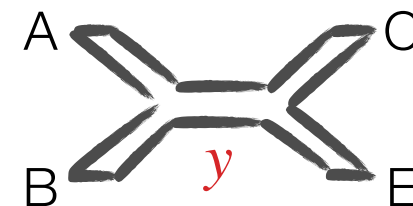
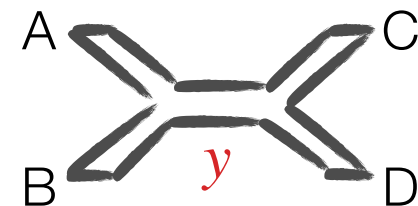
Model



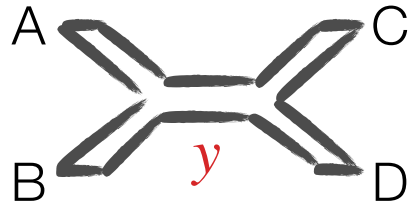
Rooted sub-models



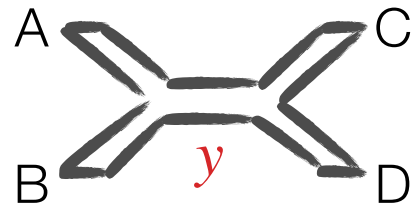
Unrooted sub-models



To compute
the likelihood
of each sub-
model



To compute
the likelihood
of each sub-
model



Exp. probs.

A,B|C,D

$$p_1 = 1 - \frac{2}{3}e^{-y}$$

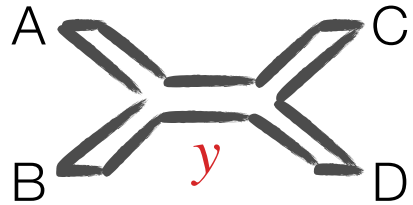
A,C|B,D

$$p_2 = \frac{1}{3}e^{-y}$$

A,D|B,C

$$p_3 = \frac{1}{3}e^{-y}$$

To compute
the likelihood
of each sub-
model



Exp. probs.

A,B|C,D

$$p_1 = 1 - \frac{2}{3}e^{-y}$$

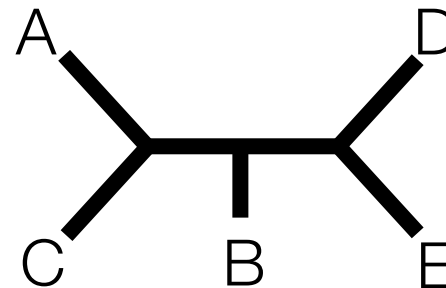
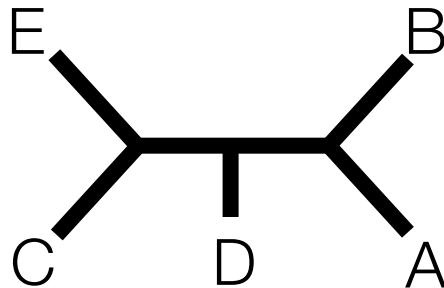
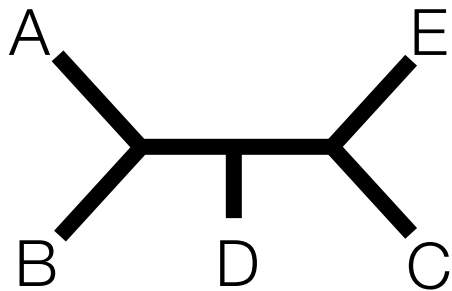
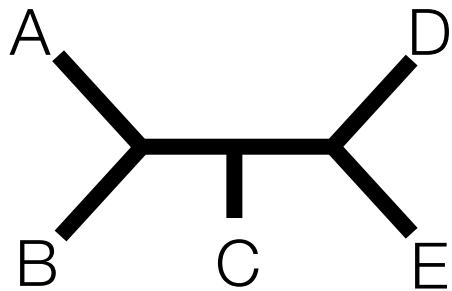
A,C|B,D

$$p_2 = \frac{1}{3}e^{-y}$$

A,D|B,C

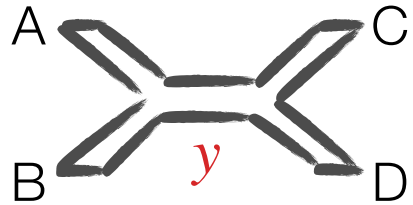
$$p_3 = \frac{1}{3}e^{-y}$$

Input Data:



...

To compute
the likelihood
of each sub-
model



Exp. probs.

$$p_1 = 1 - \frac{2}{3}e^{-y}$$

$$p_2 = \frac{1}{3}e^{-y}$$

$$p_3 = \frac{1}{3}e^{-y}$$

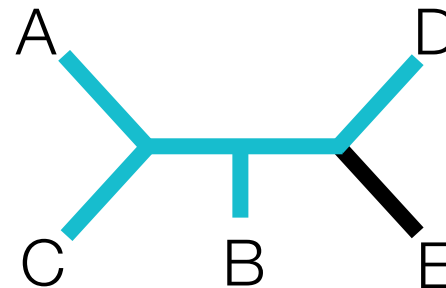
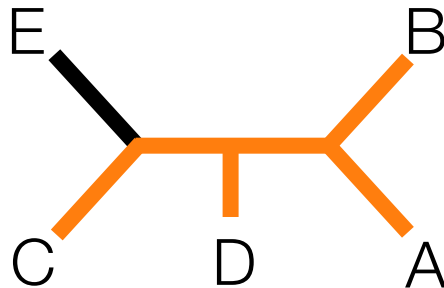
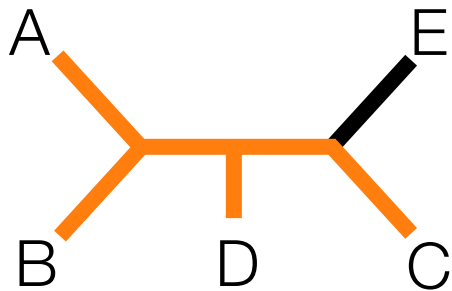
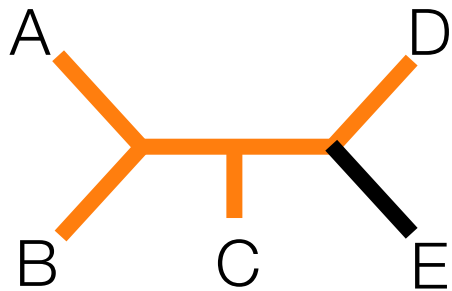
Obs. freqs.

3

1

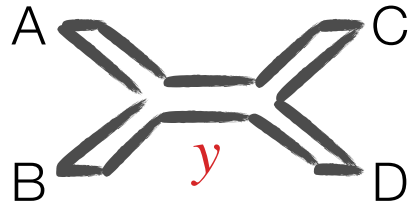
0

Input Data:



...

To compute
the likelihood
of each sub-
model



Exp. probs.

$$p_1 = 1 - \frac{2}{3}e^{-y}$$

$$p_2 = \frac{1}{3}e^{-y}$$

$$p_3 = \frac{1}{3}e^{-y}$$

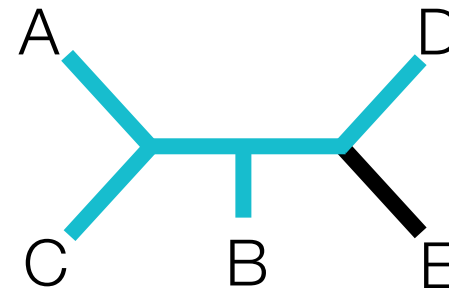
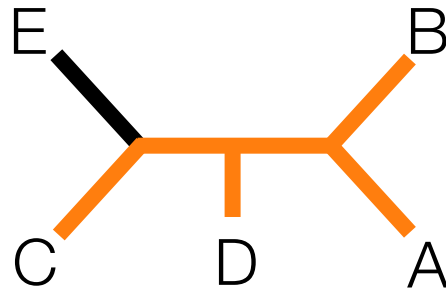
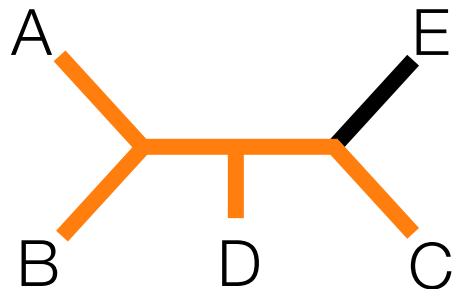
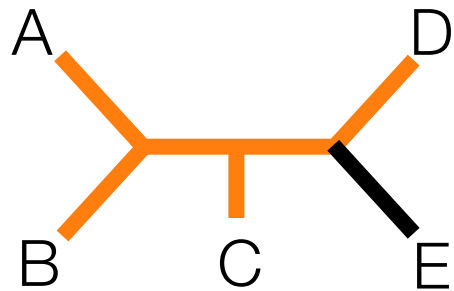
Obs. freqs.

3

1

0

Input Data:

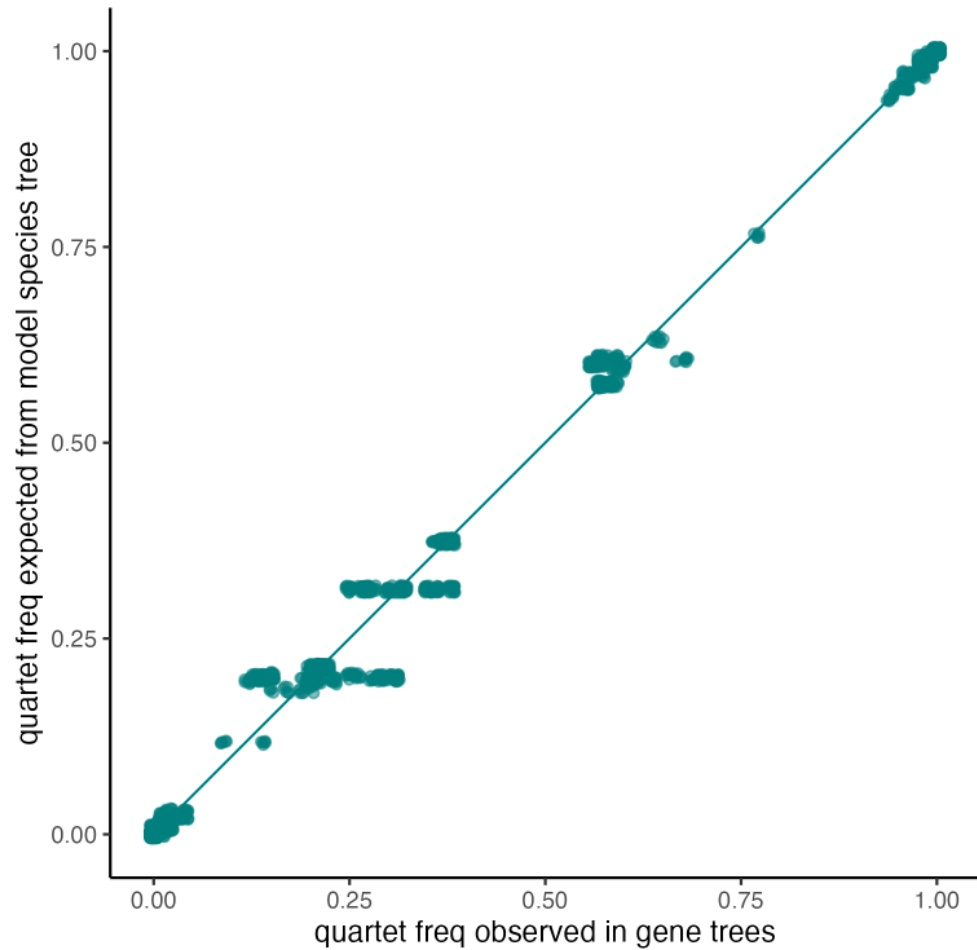


...

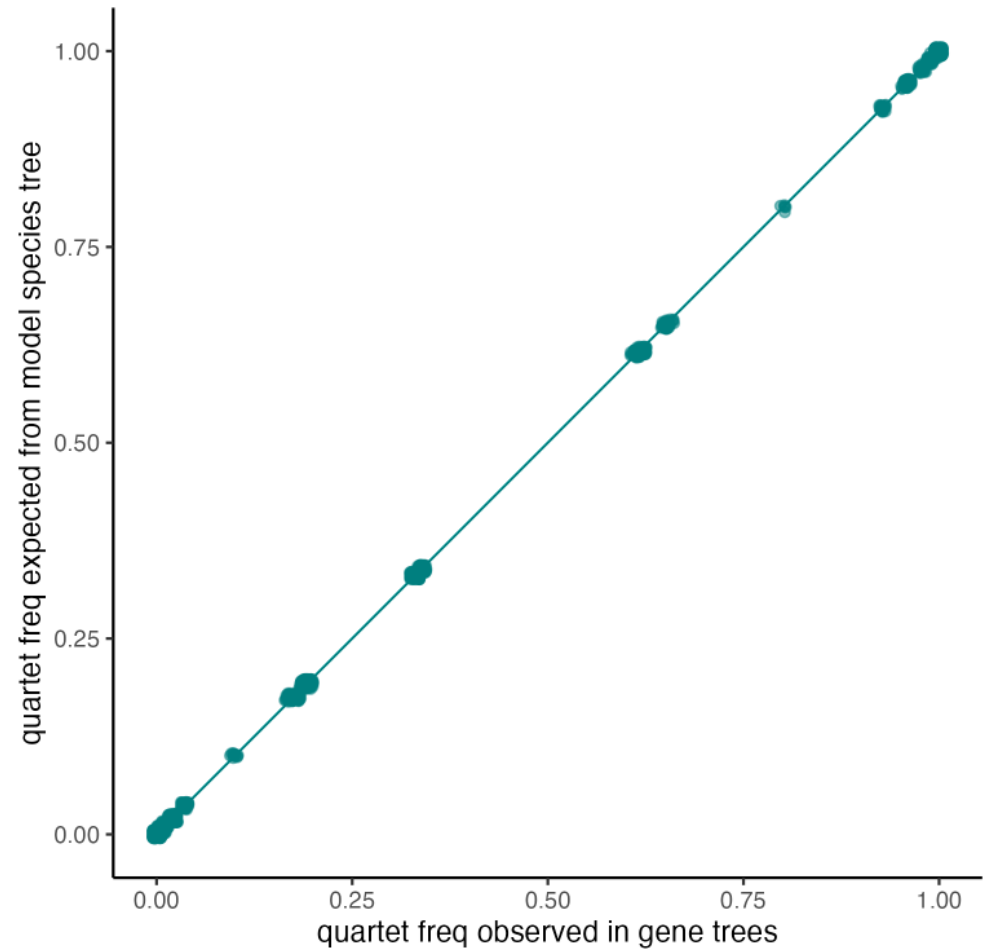
Repeat for other sub-models

Plot fit!

Observed vs expected quartet CFs

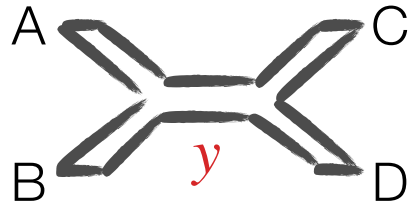


100 gene trees



10,000 gene trees

To compute
the likelihood
of each sub-
model



Exp. probs.

$$p_1 = 1 - \frac{2}{3}e^{-y}$$

$$p_2 = \frac{1}{3}e^{-y}$$

$$p_3 = \frac{1}{3}e^{-y}$$

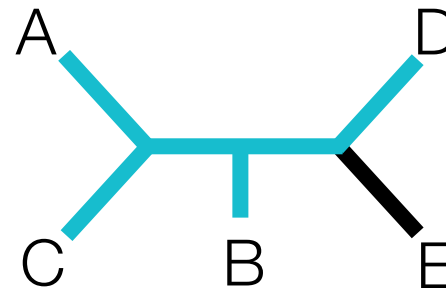
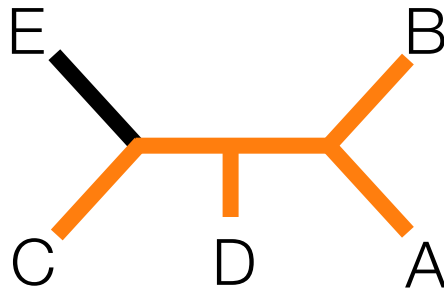
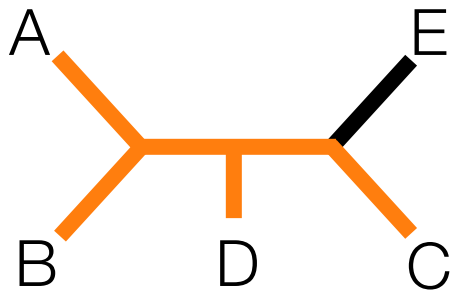
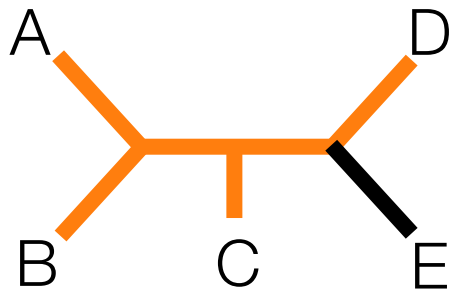
Obs. freqs.

3

1

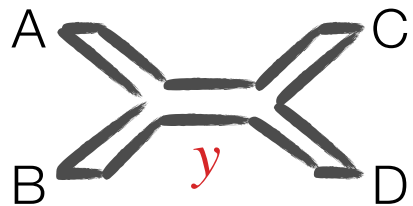
0

Input Data:



...

To compute
the likelihood
of each sub-
model



Exp. probs.

$$p_1 = 1 - \frac{2}{3}e^{-y}$$

$$p_2 = \frac{1}{3}e^{-y}$$

$$p_3 = \frac{1}{3}e^{-y}$$

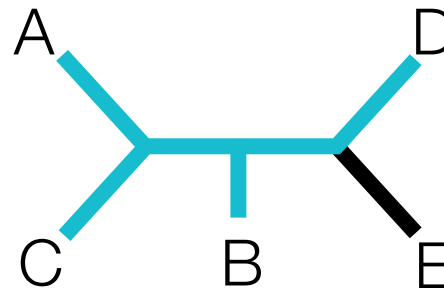
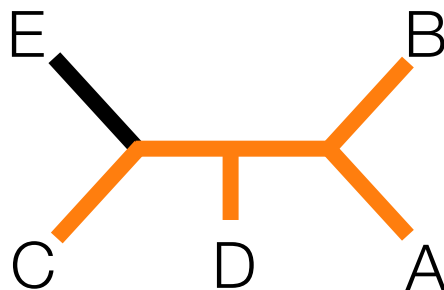
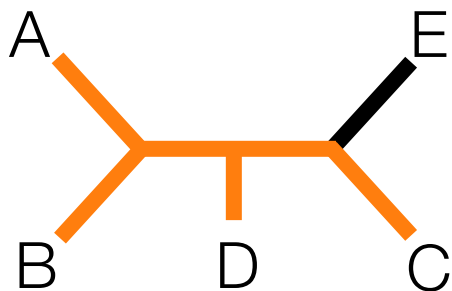
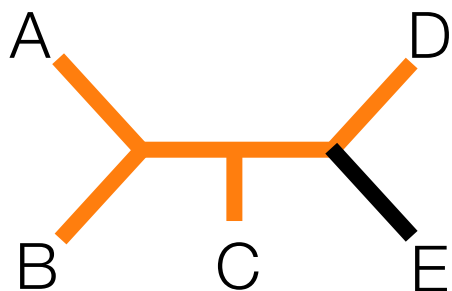
Obs. freqs.

3

1

0

Input Data:

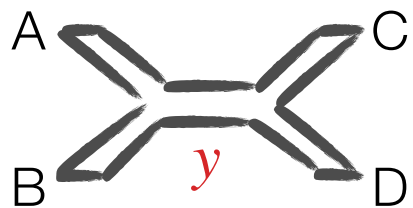


...

Repeat for other sub-models

Plot fit!

To compute
the likelihood
of each sub-
model



Exp. probs.

$$p_1 = 1 - \frac{2}{3}e^{-y}$$

$$p_2 = \frac{1}{3}e^{-y}$$

$$p_3 = \frac{1}{3}e^{-y}$$

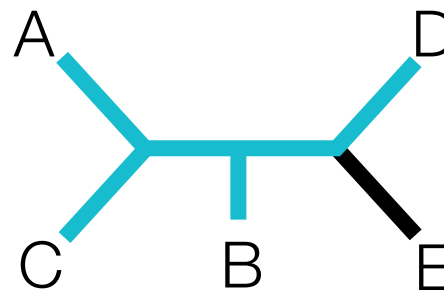
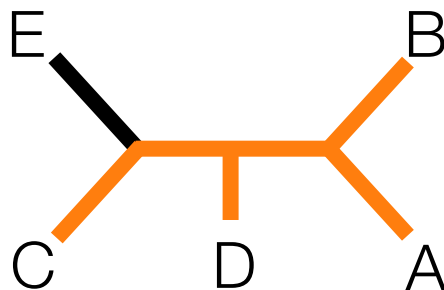
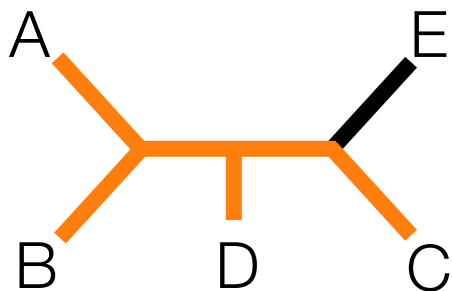
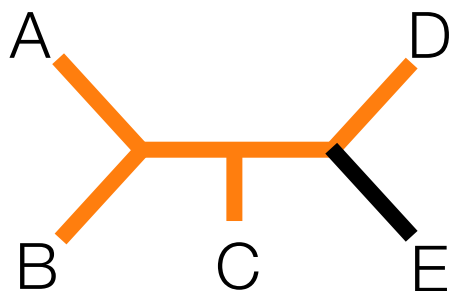
Obs. freqs.

3

1

0

Input Data:



...

Repeat for other sub-models

Plot fit!

Compute
pseudo-lk

Coalescent Lab — Day 1

Motivation for coalescent methods (**Activity A**)

Coalescent basics (**Activity B**)

Species tree estimation with summary methods (**Activity C**)

Evaluation model fit (**Activity D** — optional / do tomorrow)