Estimation of Species Trees from DNA Sequence Data

mb, Sebastien Roch

February 17, 2022

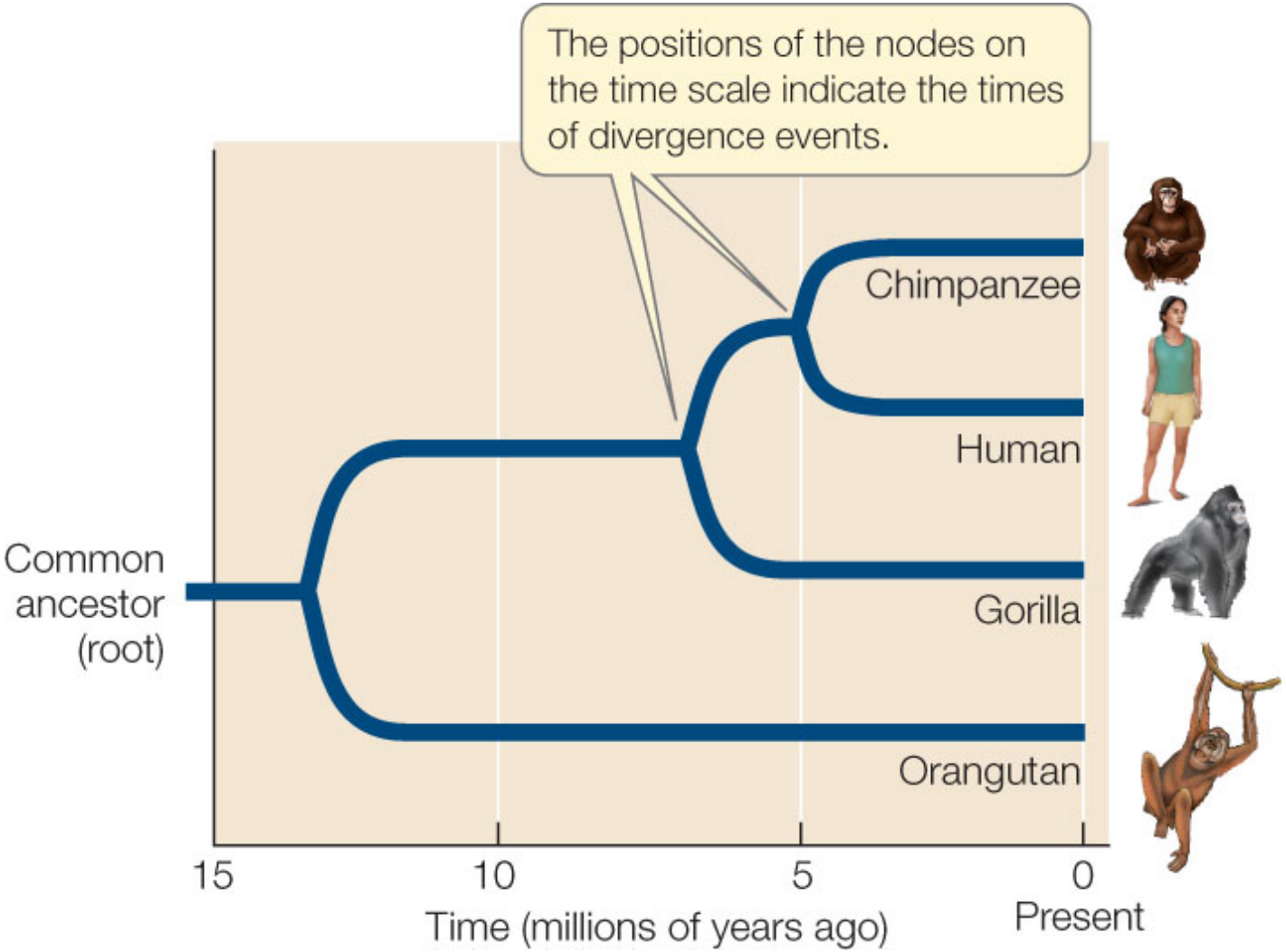
Global Roadmap

1. Part 1: Key Definitions
2. Part 2: Model of Evolution
3. Part 3: Inference Problem & Analytic Results
4. Part 4: Simulation Study

Global Roadmap

1. Part 1: Key Definitions
2. Part 2: Model of Evolution
3. Part 3: Inference Problem & Analytic Results
4. Part 4: Simulation Study

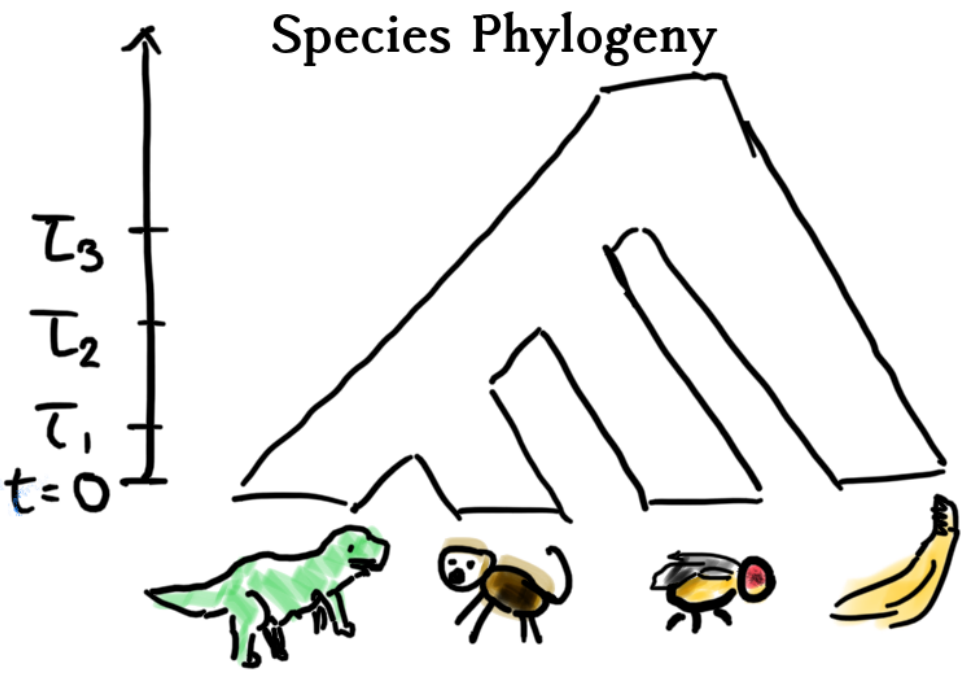
Phylogeny: the evolutionary history of a set of organisms



(Image Source: Savada 2014 [7])

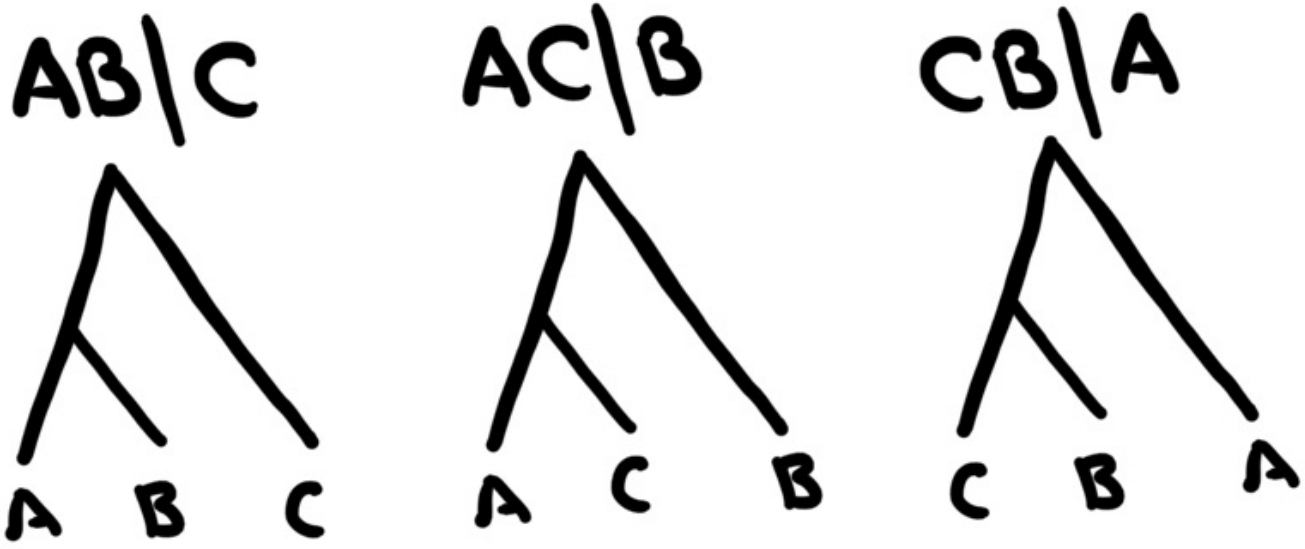
Species Phylogeny (Formal Definition)

A **species phylogeny** $S = (V, E; r, \bar{\rho}, \bar{\tau}, \bar{\theta})$ is a directed binary tree with root r, and n labeled leaves $L = [n]$, such that each edge $e \in E$ has an edge length $\tau_e \in (0, \infty)$, as well as mutation and recombination rate parameters $\theta_e \in [0, \infty)$ and $\rho_e \in [0, \infty)$.



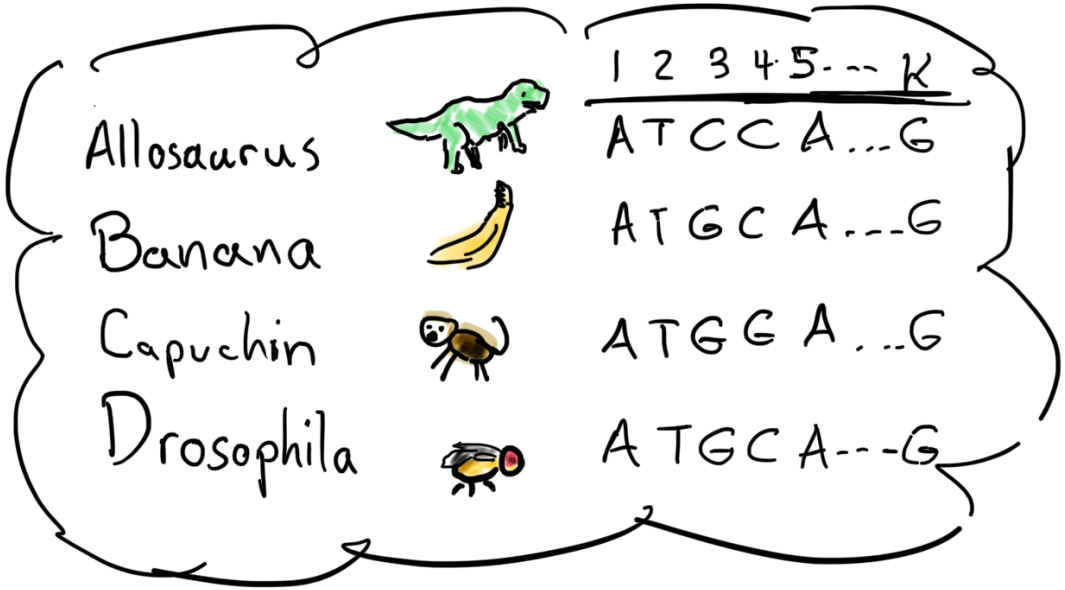
Definitions: Topology & Rooted Triple

- The **topology** of a tree is the branching structure, without regard to branch lengths.
- For a tree with 3 taxa, there are 3 possible topologies, referred to as **rooted triples**:



Phylogenetics: Inferring phylogeny from sequence data

A **multiple sequence alignment (MSA)** M is an $n \times k$ matrix whose entries are letters in the nucleotide alphabet $\{A, T, C, G\}$ such that entries in the same column are assumed to share a common ancestor.

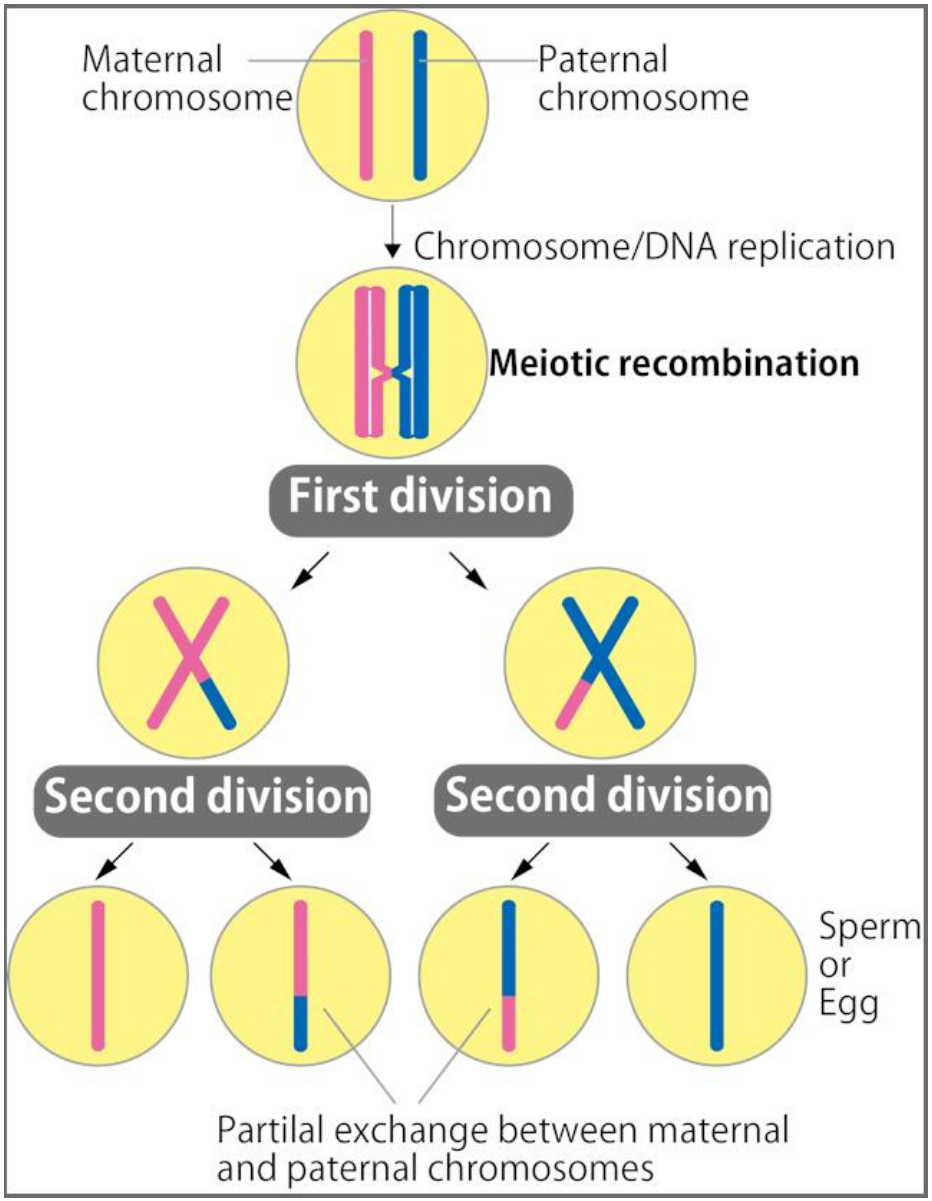


- $n =$ number of species
- $k =$ length of the *locus* of DNA sampled from each species

Interesting Questions

1. Which features of the species phylogeny can be estimated?
2. How to use sequence data sampled from the leaves to reconstruct a species phylogeny?
3. How accurate is a given inference method? Statistically consistent?
4. How much data is needed to have confidence in an estimate?
5. Which biological phenomena are major sources of estimation error?

Biological Phenomenon of Interest: Recombination



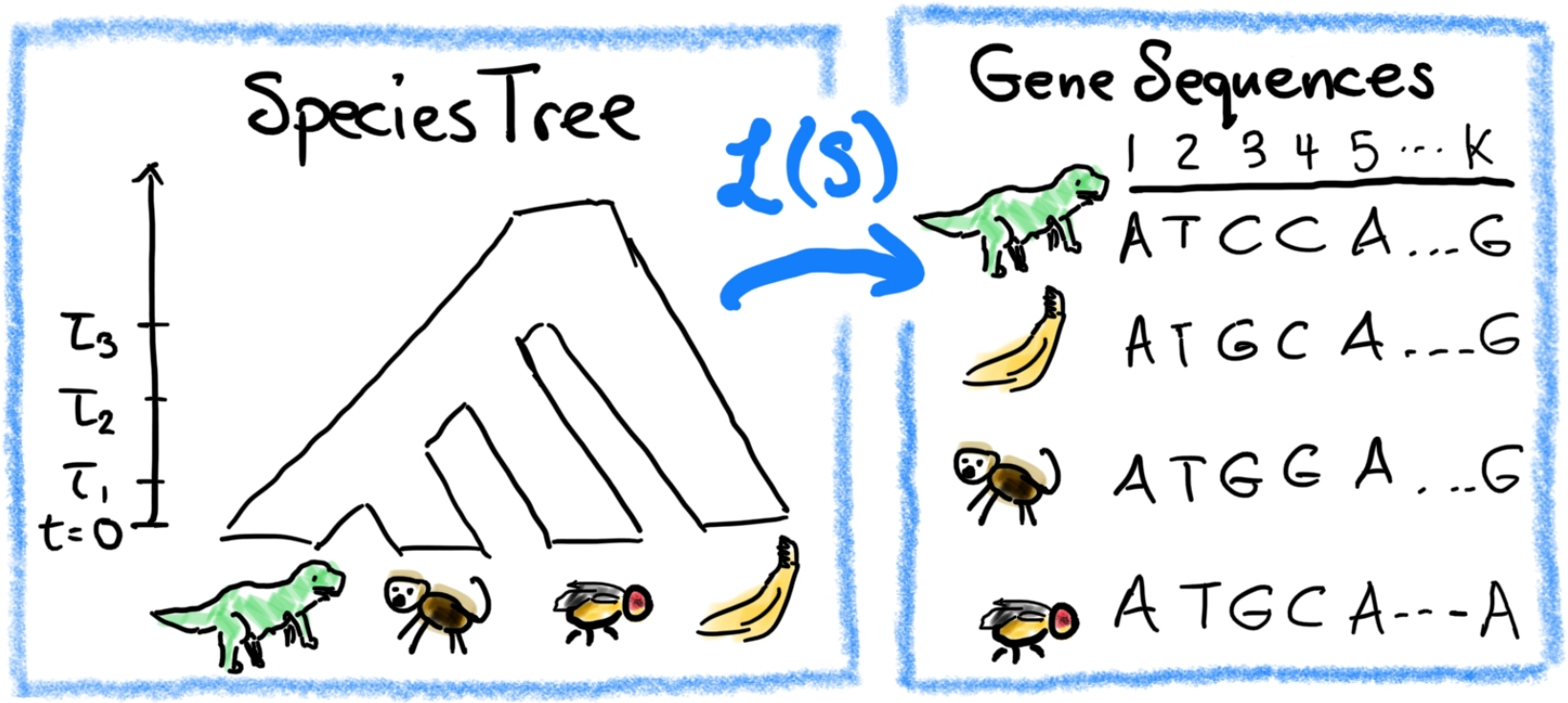
- Diploid organisms possess a maternal and paternal set of chromosomes
- Occurs during production of haploid sex cells
- Reduces linkage between genes on the same chromomosome

(Image source: [10])

Global Roadmap

1. Part 1: Key Definitions
2. Part 2: Model of Evolution
3. Part 3: Inference Problem & Analytic Results
4. Part 4: Simulation Study

Model of Evolution: Main Idea



Model of Evolution: Two Parts

Two-part model:

1. **Gene Tree Process:** We use a generalization of the multi-species coalescent (MSC) to allow for intralocus recombination.
2. **Sequence Evolution Process:** We run the Jukes-Cantor substitution process independently on each gene tree.

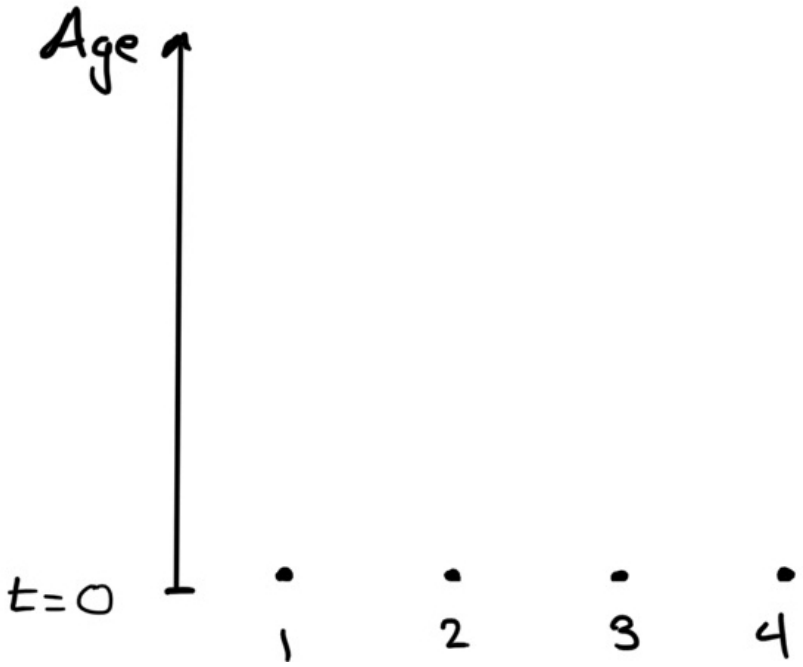
Model of Evolution: Two Parts

Two-part model:

1. **Gene Tree Process:** We use a generalization of the multi-species coalescent (MSC) to allow for intralocus recombination.
2. **Sequence Evolution Process:** We run the Jukes-Cantor substitution process independently on each gene tree.

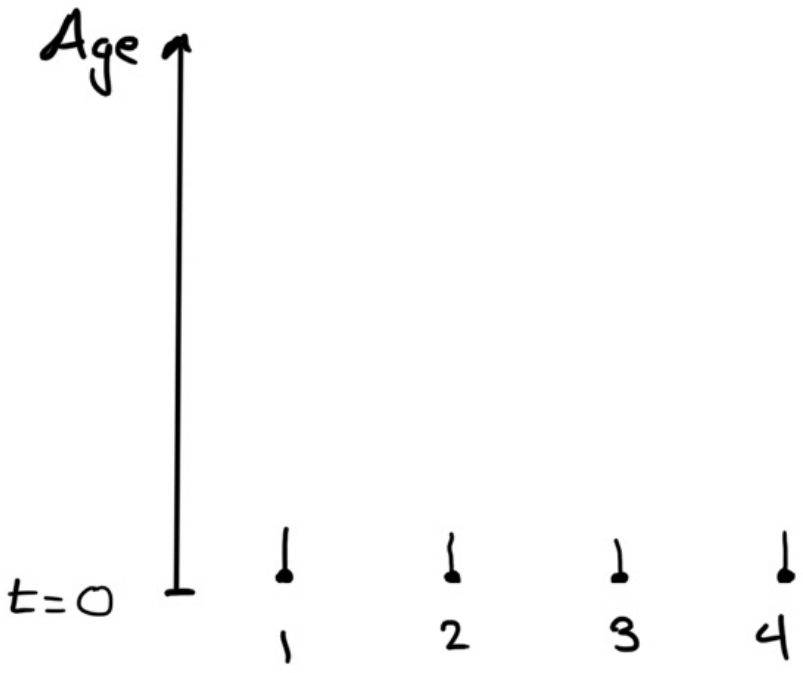
The Coalescent

- Widely-used model of how genes evolved from common ancestors
- "Backwards in time"
- Start by sampling n individuals at time $t = 0$. (Here $n = 4$)



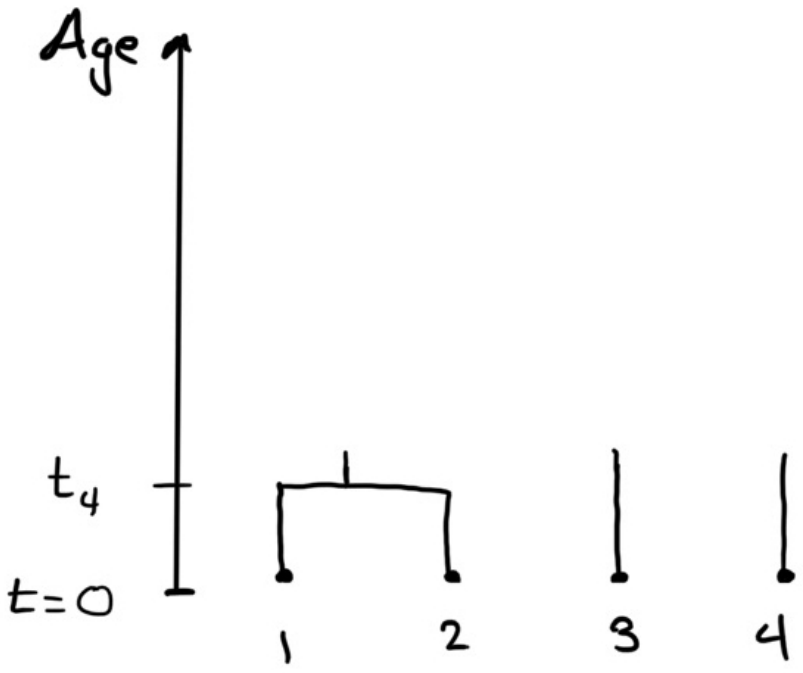
The Coalescent

- Lines = Ancestral lineages
- Each pair of lineages must have had a MRCA who lived at some time in the past.
- **Coalescence Event:** When two lineages reach the time of their MRCA, we *join* the two lineages at that time (next slide).



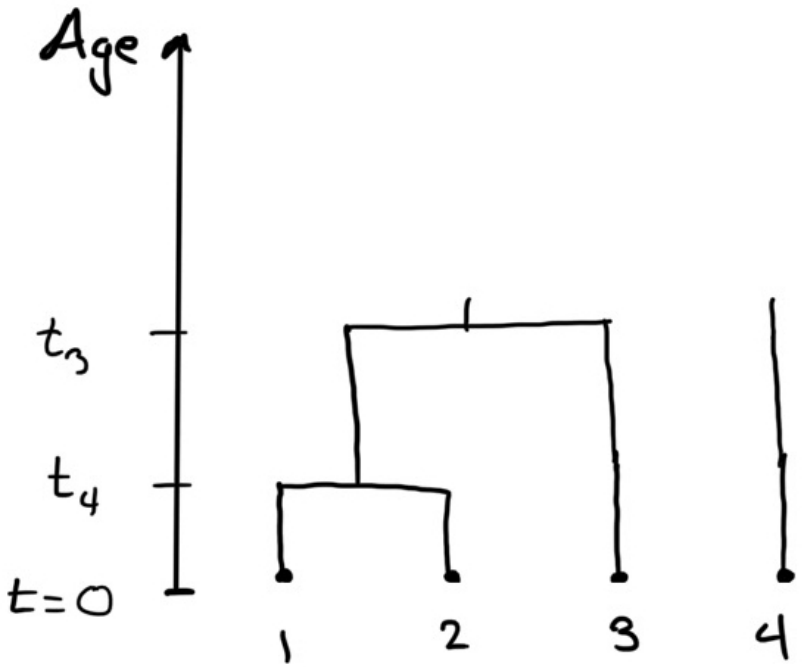
The Coalescent

- Time in "coalescent units" (1 c.u. $= 2N_e$ generations)
- Each pair of lineages coalesces independently at rate 1.
  - When N lineages, time until the next coalescence $\sim \exp\left(\binom{N}{2}\right)$.



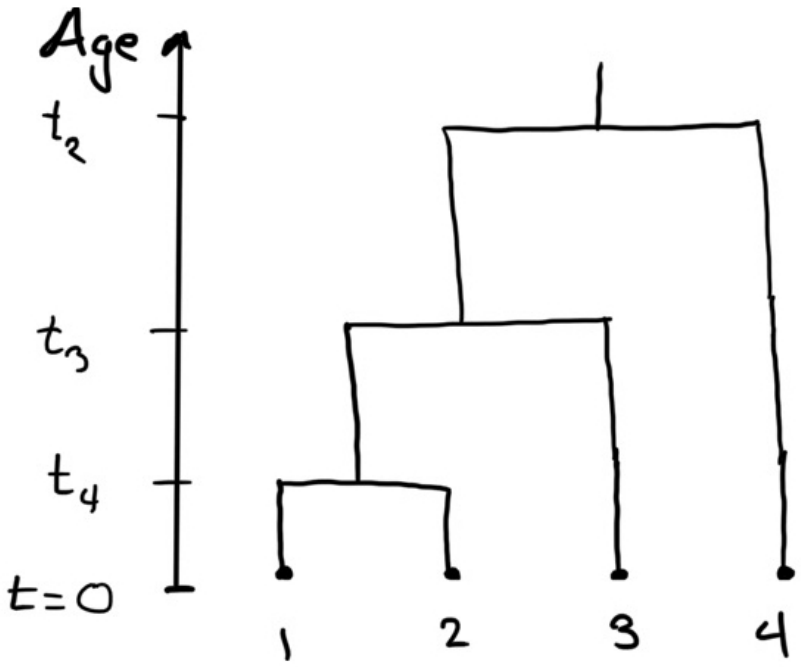
The Coalescent

- Going back further in time, another pair of lineages coalesce.
- Continue until there is only a single lineage remaining (next slide).



The Coalescent

- **Output:** a gene tree representing the genealogy of our four sampled individuals.
- Recombination **between** genes reduces linkage between them
- This justifies modeling genealogy of different genes using **independent** gene trees.



Intralocus Recombination: Definition

Natural question: What if recombination occurs *within* genes, not just *between* them?

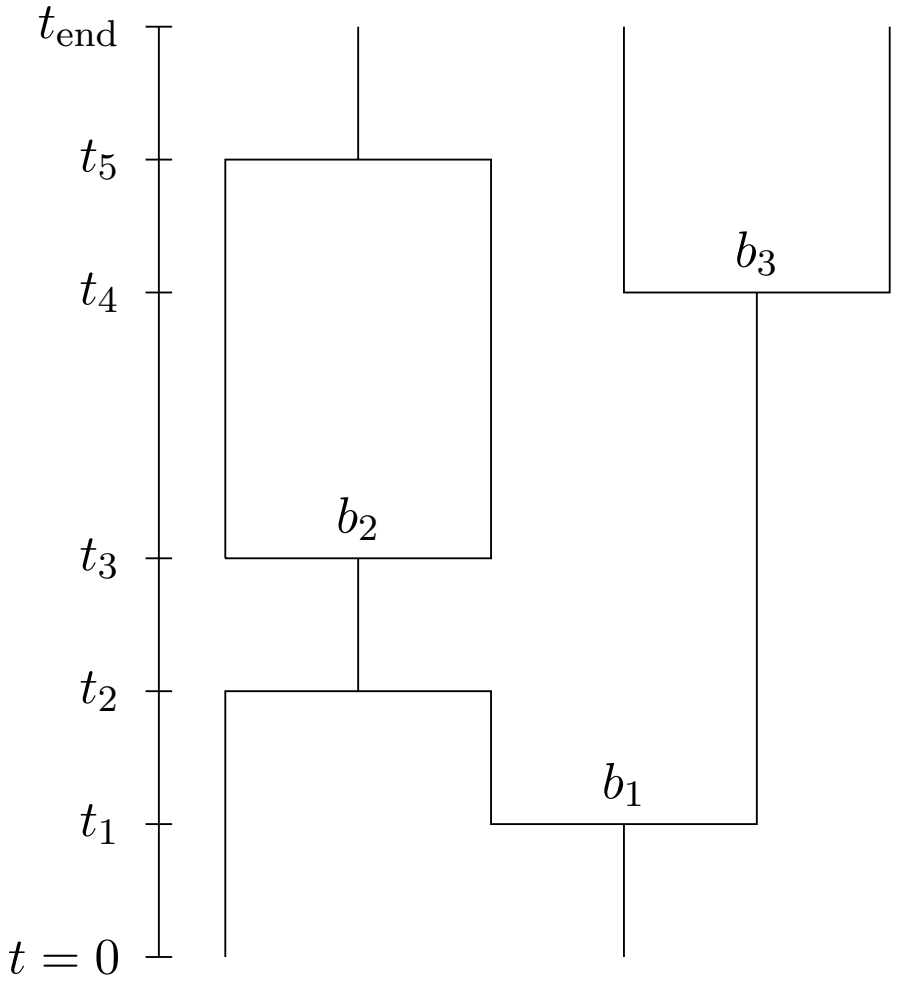
- For example, a *single* observed gene might consist of several components with distinct evolutionary histories, having been inherited from different ancestors. We call this phenomenon **intralocus recombination**.
- Common in real data (e.g. 80% of protein-coding genes in Eukaryotes) [1].

Impact of Intralocus Recombination

- The coalescent assumes *tree-like* gene ancestries. Consequently, inference methods which assume evolution is well-approximated by such a model might be misleading.
- Significant debate on this question [5, 8, 6].
- **Question:** Do inference procedures need to be designed to account for intralocus recombination?
- **Approach:** We consider a modification of the coalescent which allows for intralocus recombination, and consider in a simple setting how this modification impacts the ability to infer the species topology.

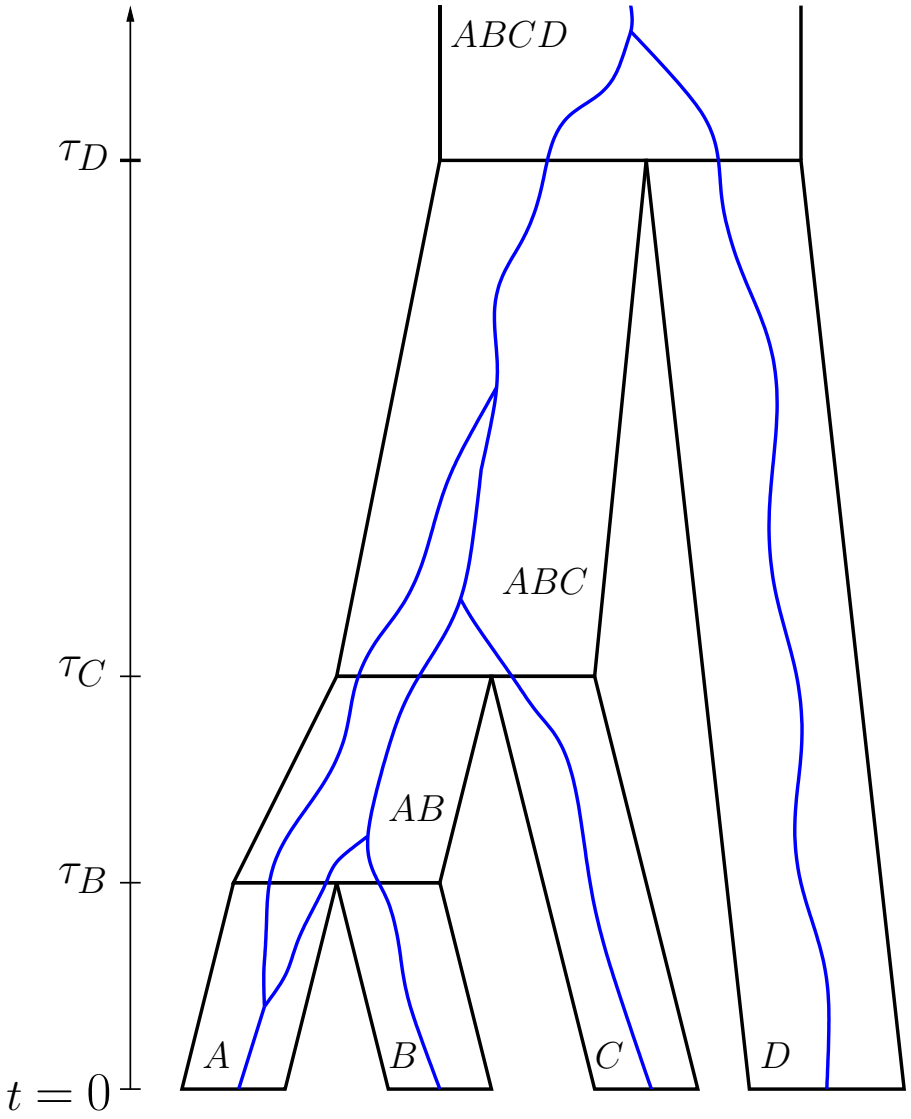
The Ancestral Recombination Graph Model (ARG)

- Similar to the coalescent, but includes **recombination events**, in which a lineage splits into two lineages.
- The number N of lineages is a birth-death process:
  - **Deaths** (coalescent events) occur at rate $\binom{N}{2}$.
  - **Births** (recombination events) occur at rate ρN.
- Recombinations are labeled by **breakpoints** $b \sim \text{unif}([0,1])$.

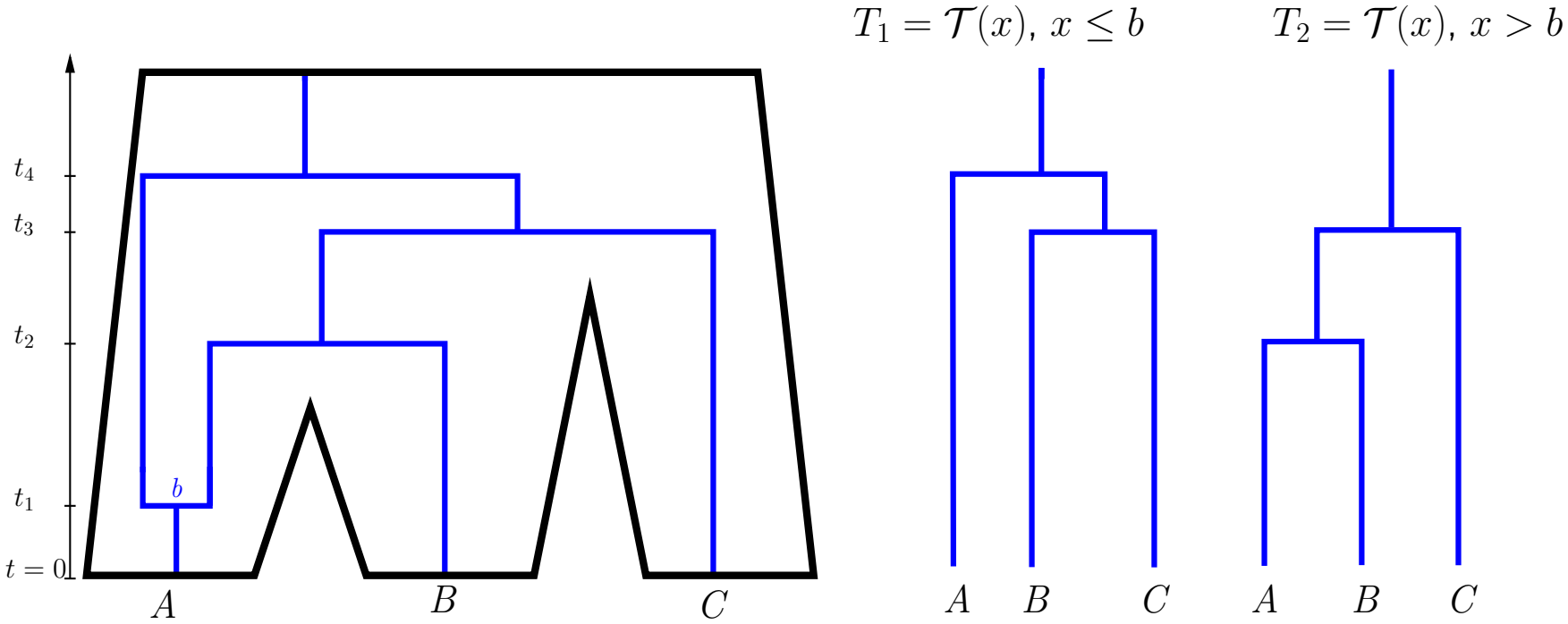


Multispecies Coalescent with Recombination (MSCR)

- Fixed species phylogeny S (here, $n = 4$ species)
- ARG is in blue
- Start with 1 lineage per leaf



Decomposing the ARG into Trees



For each site $x \in [0, 1]$, the **marginal gene tree** $\mathcal{T}(x)$ is defined by tracing up from the leaves; when a recombination breakpoint b is reached, continue tracing along left edge if $x \leq b$, or the right edge if $x > b$.

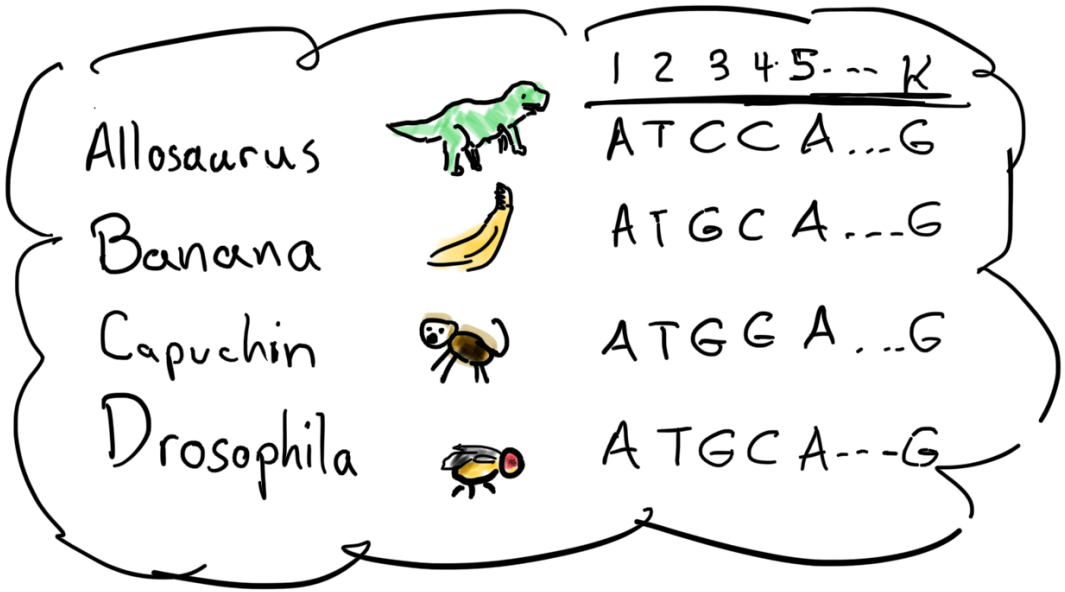
Model of Evolution: Two Parts

Two-part model:

1. **Gene Tree Process:** We use a generalization of the multi-species coalescent (MSC) to allow for intralocus recombination recombination.
2. **Sequence Evolution Process:** We run the Jukes-Cantor substitution process independently on each gene tree.

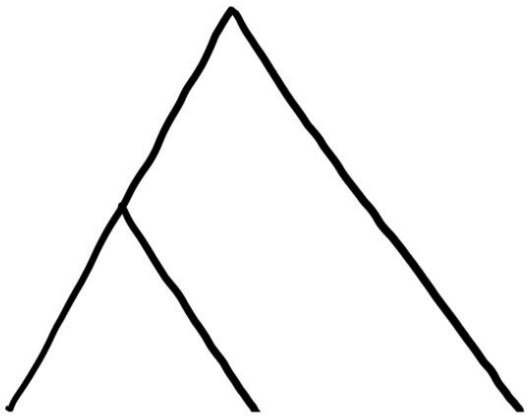
Sequence Evolution Process: Overview

- **Goal:** reconstruct a species phylogeny from *DNA sequence data* sampled from the leaves of S.
- **Approach:** run a mutation process independently for each site x on the marginal gene tree corresponding to that site.



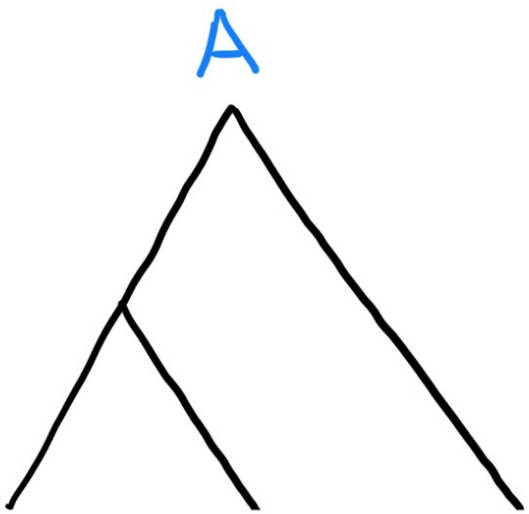
Sequence Evolution Process: JC Process

- **Input:** A marginal gene tree (representing the genealogy of a single DNA site).
- **Output:** A nucleotide letter at each tip of the tree (representing 1 column of the MSA).
- **Initialization:** Randomly assigning a nucleotide letter at the root. This is the "ancestral" state.



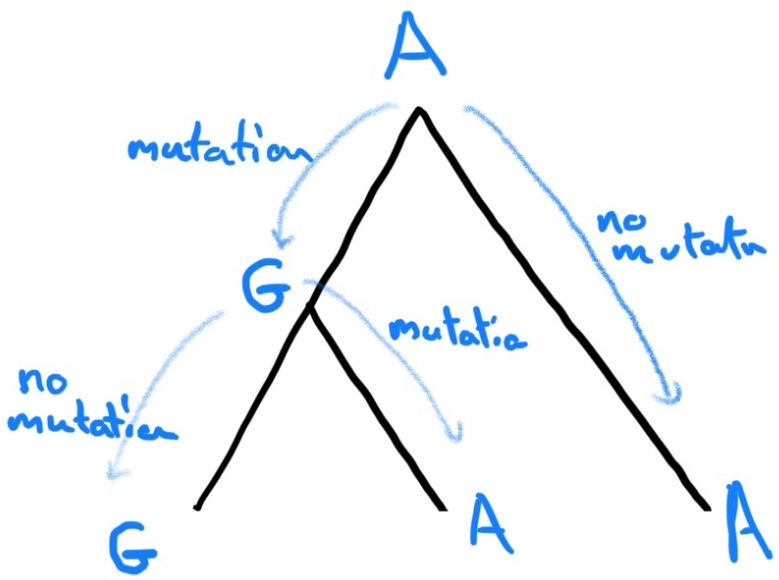
Sequence Evolution Process: JC Process

- **Input:** A marginal gene tree (representing the genealogy of a single DNA site).
- **Output:** A nucleotide letter at each tip of the tree (representing 1 column of the MSA).
- **Initialization:** Randomly assigning a nucleotide letter at the root. This is the "ancestral" state.



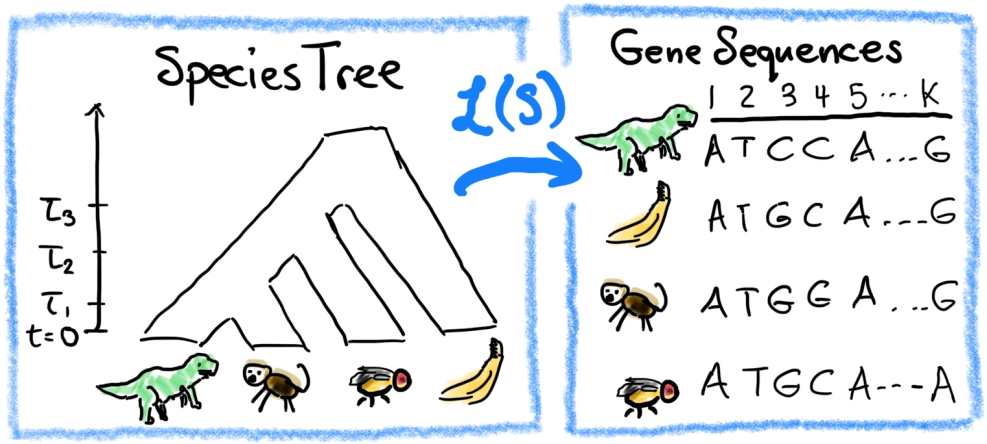
Sequence Evolution Process: JC Process

- Along each edge e, mutate to a different letter with probability p_e
  - p_e is a function of edge length and mutation rate
- We obtain a nucleotide at each tip
- Repeat this independently for all k sites to obtain an MSA



Model of Evolution: Summary

- We have sketched a two-part model: (1) gene tree process and (2) sequence evolution process
  - Input: a fixed species phylogeny S
  - Output: an MSA (aligned sequences at the tips of S).



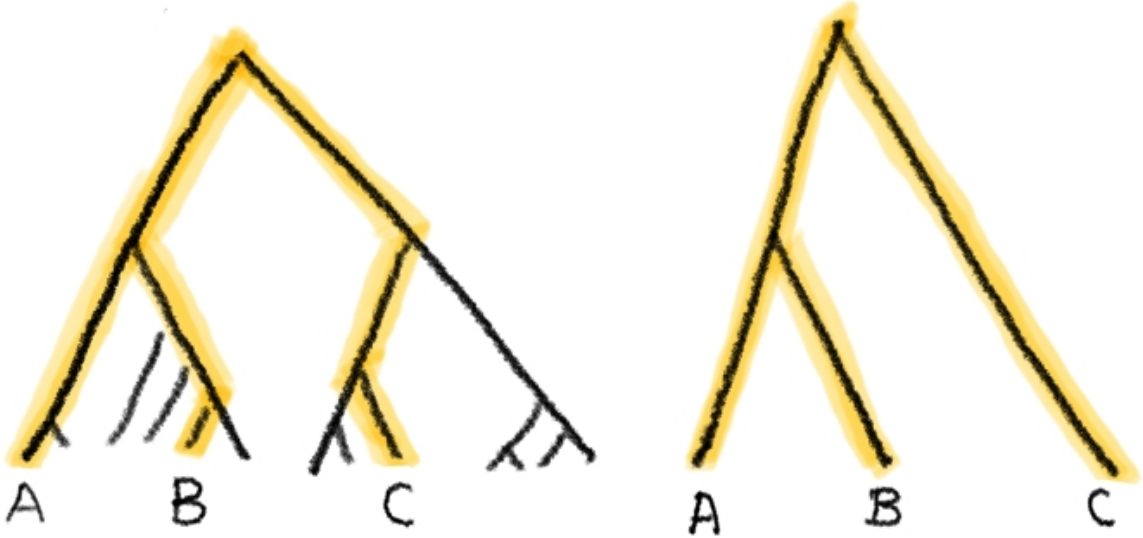
- The gene tree process explicitly accounts for intralocus recombination.
- "MSCR-JC(k) process on S"

Global Roadmap

1. Part 1: Key Definitions
2. Part 2: Model of Evolution
3. Part 3: Inference Problem & Analytic Results
4. Part 4: Simulation Study

Inference Procedure: Introduction

- **Phylogenetic reconstruction problem:** Recover the topology of S from m independent samples of M generated according to the MSCR-JC(k) process on S.
- We'll focus on a method based on reconstructing **species triplets** individually:



- Distance-based: based on the number of mismatching nucleotides δ_{XY} between the sequences from species X and Y.

Inference Procedure: R^* with Sequence Distances

Majority-rule Rooted Triple (R^*) inference pipeline:

1. For each sampled gene, infer a rooted triple for **every** species triplet in S.
   - In particular, the species triplet with leaves X, Y, Z is inferred to have rooted triple topology $XY|Z$ if $\delta_{XY} < \delta_{XZ} \wedge \delta_{YZ}$.
2. Make a list of those rooted triples which were uniquely favored (i.e., most-frequently inferred from the m sampled genes).
3. Construct the most-resolved topology containing only uniquely favored triples.

Inference Procedure: R^* with Sequence Distances

Majority-rule Rooted Triple (R^*) inference pipeline:

1. For each sampled gene, infer a rooted triple for **every** species triplet in S.
   - In particular, the species triplet with leaves X, Y, Z is inferred to have rooted triple topology $XY|Z$ if $\delta_{XY} < \delta_{XZ} \wedge \delta_{YZ}$.
2. Make a list of those rooted triples which were uniquely favored (i.e., most-frequently inferred from the m sampled genes).
3. Construct the most-resolved topology containing only uniquely favored triples.

Inference Procedure: R^* with Sequence Distances

Majority-rule Rooted Triple (R^*) inference pipeline:

1. For each sampled gene, infer a rooted triple for **every** species triplet in S.
   - In particular, the species triplet with leaves X, Y, Z is inferred to have rooted triple topology $XY|Z$ if $\delta_{XY} < \delta_{XZ} \wedge \delta_{YZ}$.
2. Make a list of those rooted triples which were uniquely favored (i.e., most-frequently inferred from the m sampled genes).
3. Construct the most-resolved topology containing only uniquely favored triples.

Inference Procedure: R^* Justification and Motivation

- R^* was designed to account for certain properties of the coalescent (the "anomaly zone") [3, 4, 2]
- R^* utilizes the fact that the full topology of S is uniquely determined by, and hence can be recovered from, its rooted triples [9].
- It is known that under the model with no intralocus recombination, R^* is statistically consistent estimator of the topology of S [4].

Question: *Does R^* still work if intralocus recombination is incorporated into the model?*

Theorem 1:

R^* is **not** statistically consistent when intralocus recombination is allowed.

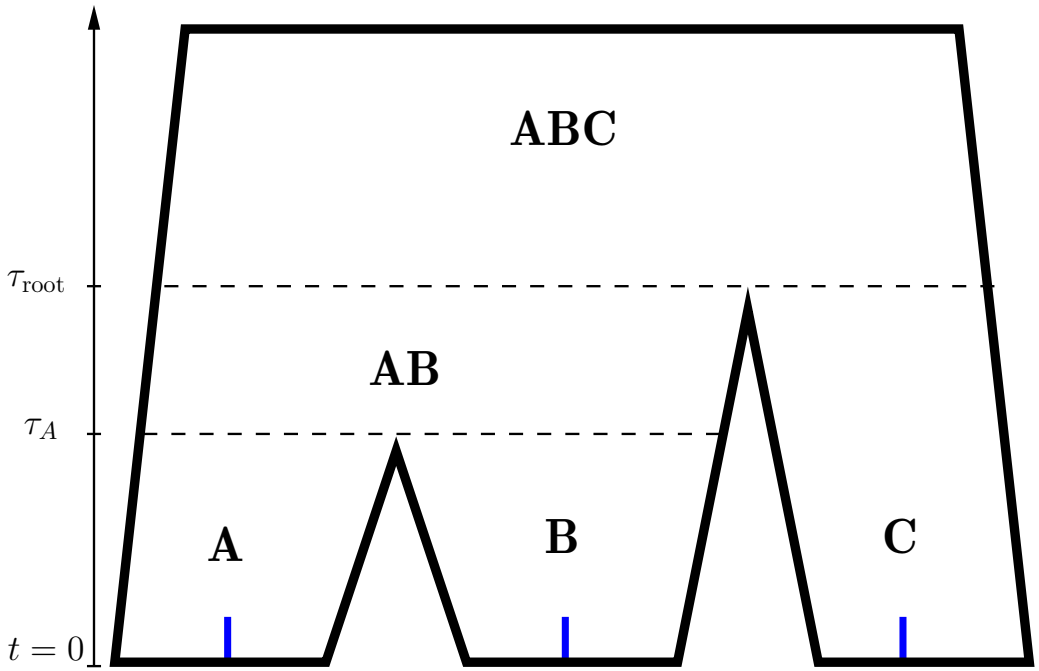
Theorem

For k sufficiently large, R^* using sequence distances is not statistically consistent under the MSCR-JC(k) model. That is, there exists a species phylogeny S such that the topology of the output of R^* using sequence distances does not converge in probability to the topology of the species tree.

(We also prove a related result for inference using unrooted quartets via the 4-point condition.)

Proof of Theorem 1: The Setting

- S has 3 leaves A, B, C, and topology $AB|C$
- Recombination **only** in population A
- Internal branch length $f := \tau_{\text{root}} - \tau_A$ is chosen to be very short.



We show that in this setting R^* converges to $BC|A$.

Proof of Theorem 1

Let $E_{XY|Z}$ be the event that the species triplet with leaves X, Y, Z is inferred to have the rooted triple $XY|Z$ from a **single** MSA.

- In symbols, $E_{XY|Z} = [\delta_{XY} < \delta_{XZ} \wedge \delta_{YZ}]$.
- Simplifying assumption: assume k is large (to minimize the role of randomness coming from the site substitution process.)

Lemma

A necessary and sufficient condition for R^* to converge to the (correct) topology of S as $m \to \infty$ is that

$$P\left[E_{AB|C}\right] > P\left[E_{BC|A}\right] \vee \mathbb{P}\left[E_{AC|B}\right] \tag{1}$$

By lemma, it suffices to show there exists some S such that

$$\mathbb{P}\left[E_{BC|A}\right] > \mathbb{P}\left[E_{AB|C}\right]. \tag{2}$$

Proof of Theorem 1

Let $E_{XY|Z}$ be the event that the species triplet with leaves X, Y, Z is inferred to have the rooted triple $XY|Z$ from a **single** MSA.

- In symbols, $E_{XY|Z} = [\delta_{XY} < \delta_{XZ} \wedge \delta_{YZ}]$.
- Simplifying assumption: assume k is large (to minimize the role of randomness coming from the site substitution process.)

Lemma

A necessary and sufficient condition for R^* to converge to the (correct) topology of S as $m \to \infty$ is that

$$P\left[E_{AB|C}\right] > P\left[E_{BC|A}\right] \vee \mathbb{P}\left[E_{AC|B}\right] \tag{1}$$

By lemma, it suffices to show there exists some S such that

$$\mathbb{P}\left[E_{BC|A}\right] > \mathbb{P}\left[E_{AB|C}\right]. \tag{2}$$

Proof of Theorem 1

Let $E = E_{AB|C}$ and $F = E_{BC|A}$. Need to show:

$$\mathbb{P}[F] - \mathbb{P}[E] > 0. \tag{3}$$

Let $R_k =$ exactly k recombinations occur. Then

$$\begin{aligned}\mathbb{P}[F] - \mathbb{P}[E] = &\left(\mathbb{P}[F|R_0] - \mathbb{P}[E|R_0]\right)\mathbb{P}[R_0] \\ &+ \left(\mathbb{P}[F|R_1] - \mathbb{P}[E|R_1]\right)\mathbb{P}[R_1] \\ &+ O(\rho_A^2)\end{aligned} \tag{4}$$

- First term: negative but $O(f)$ as $f \to 0$.
- Third term: negligible.
- Second term: $\mathbb{P}[R_1] = O(\rho_A)$.

Bottom line: Suffices to show that

$$\mathbb{P}[F|R_1] - \mathbb{P}[E|R_1] > C \tag{5}$$

where $C > 0$ does not depend on f, ρ_A.

Proof of Theorem 1

Let $E = E_{AB|C}$ and $F = E_{BC|A}$. Need to show:

$$\mathbb{P}[F] - \mathbb{P}[E] > 0. \tag{3}$$

Let $R_k =$ exactly k recombinations occur. Then

$$\begin{aligned}\mathbb{P}[F] - \mathbb{P}[E] = &\left(\mathbb{P}[F|R_0] - \mathbb{P}[E|R_0]\right)\mathbb{P}[R_0] \\ &+ \left(\mathbb{P}[F|R_1] - \mathbb{P}[E|R_1]\right)\mathbb{P}[R_1] \\ &+ O(\rho_A^2)\end{aligned} \tag{4}$$

- First term: negative but $O(f)$ as $f \to 0$.
- Third term: negligible.
- Second term: $\mathbb{P}[R_1] = O(\rho_A)$.

Bottom line: Suffices to show that

$$\mathbb{P}[F|R_1] - \mathbb{P}[E|R_1] > C \tag{5}$$

where $C > 0$ does not depend on f, ρ_A.

Proof of Theorem 1

Let $E = E_{AB|C}$ and $F = E_{BC|A}$. Need to show:

$$\mathbb{P}[F] - \mathbb{P}[E] > 0. \tag{3}$$

Let $R_k =$ exactly k recombinations occur. Then

$$\begin{aligned}\mathbb{P}[F] - \mathbb{P}[E] = &\left(\mathbb{P}[F|R_0] - \mathbb{P}[E|R_0]\right)\mathbb{P}[R_0] \\ &+ \left(\mathbb{P}[F|R_1] - \mathbb{P}[E|R_1]\right)\mathbb{P}[R_1] \\ &+ O(\rho_A^2)\end{aligned} \tag{4}$$

- First term: negative but $O(f)$ as $f \to 0$.
- Third term: negligible.
- Second term: $\mathbb{P}[R_1] = O(\rho_A)$.

Bottom line: Suffices to show that

$$\mathbb{P}[F|R_1] - \mathbb{P}[E|R_1] > C \tag{5}$$

where $C > 0$ does not depend on f, ρ_A.

Proof of Theorem 1

Let $E = E_{AB|C}$ and $F = E_{BC|A}$. Need to show:

$$\mathbb{P}[F] - \mathbb{P}[E] > 0. \tag{3}$$

Let $R_k =$ exactly k recombinations occur. Then

$$\begin{aligned}\mathbb{P}[F] - \mathbb{P}[E] = &\left(\mathbb{P}[F|R_0] - \mathbb{P}[E|R_0]\right)\mathbb{P}[R_0] \\ &+ \left(\mathbb{P}[F|R_1] - \mathbb{P}[E|R_1]\right)\mathbb{P}[R_1] \\ &+ O(\rho_A^2)\end{aligned} \tag{4}$$

- First term: negative but $O(f)$ as $f \to 0$.
- Third term: negligible.
- Second term: $\mathbb{P}[R_1] = O(\rho_A)$.

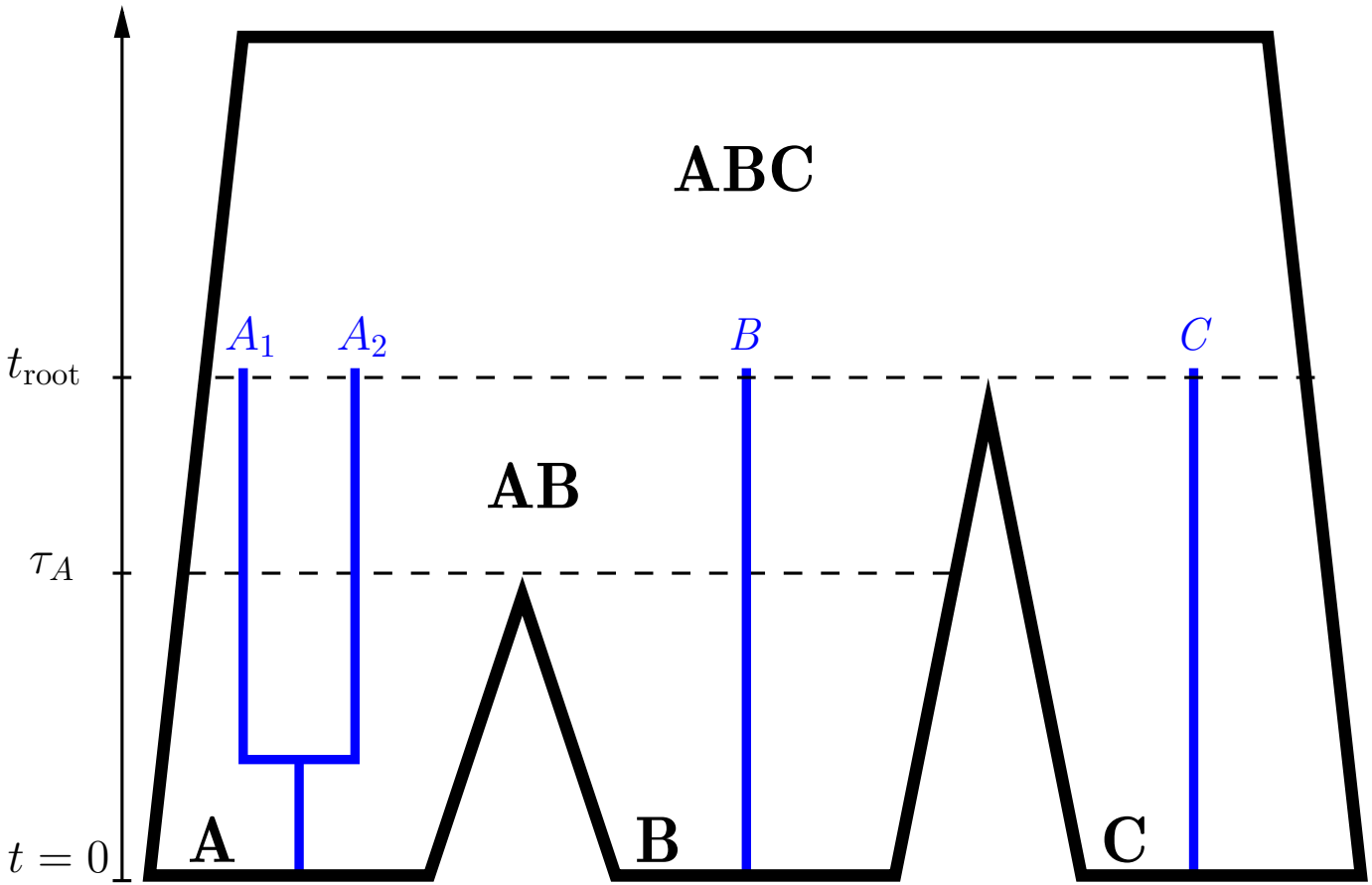
Bottom line: Suffices to show that

$$\mathbb{P}[F|R_1] - \mathbb{P}[E|R_1] > C \tag{5}$$

where $C > 0$ does not depend on f, ρ_A.

Proof of Theorem 1: Key Case

$R_1C_0 =$ exactly one recombination and no coalescence below the root



Global Roadmap

1. Part 1: Key Definitions
2. Part 2: Model of Evolution
3. Part 3: Inference Problem & Analytic Results
4. Part 4: Simulation Study

Simulation Study

Goal: Characterize the inconsistency results about R^*

1. Same setting as theorem:
   - S has three species: A, B, C
   - rooted topology $AB|C$
   - constant mutation rate θ
2. DNA sequences generated according to the MSCR-JC(500) process.
3. Biologically plausible rates of recombination $\rho \in (0, 20)$ and mutation $\theta \in \{0.01, 0.1\}$ were used.

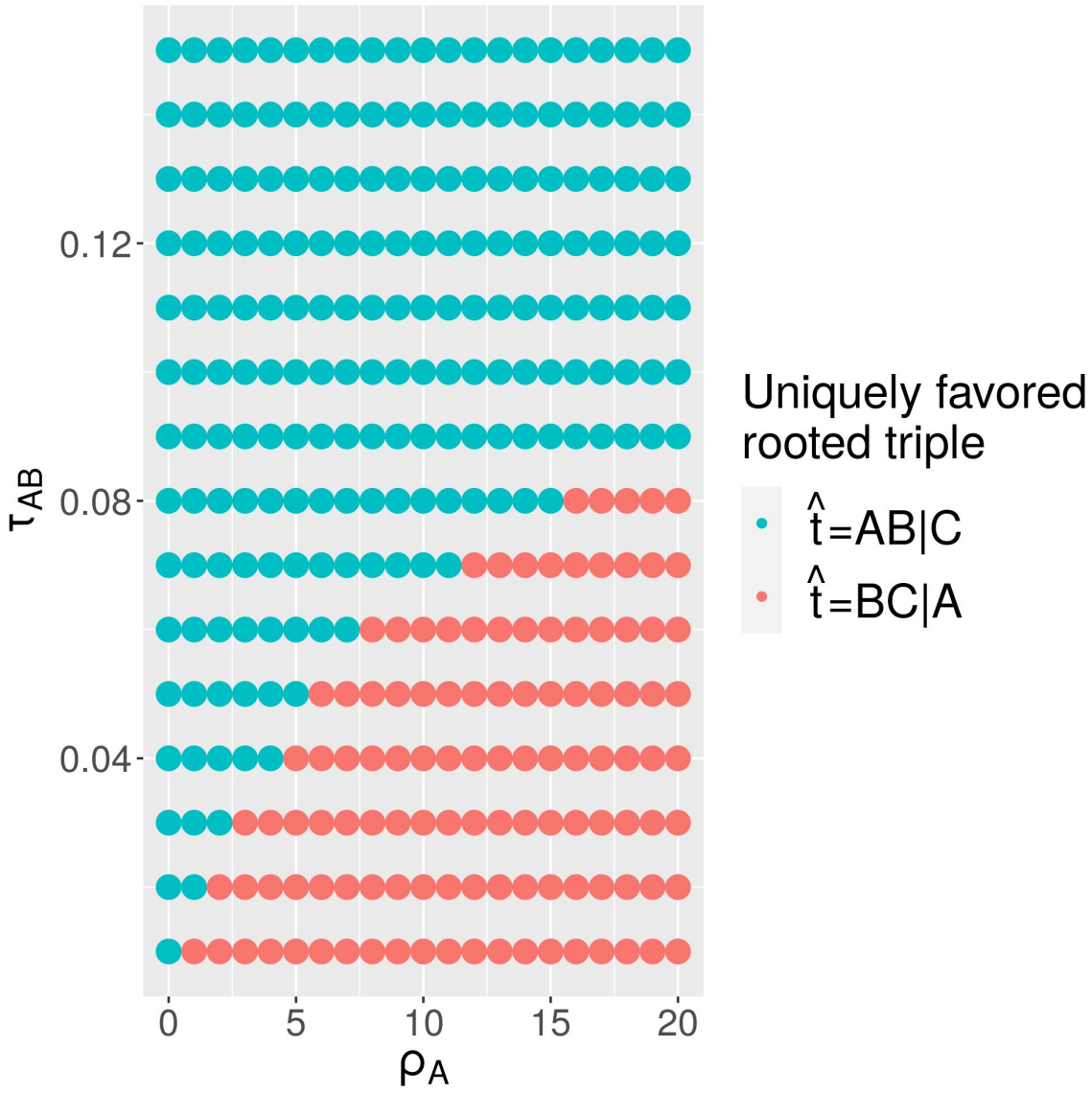
Simulation Study - Notation

- $\hat{p}_{XY|Z}$:= the proportion of samples from which the rooted triple $XY|Z$ was inferred
- $\hat{t}$:= the rooted triple most frequently inferred among the m samples
- Hence $\hat{t} = AB|C \iff \hat{p}_{AB|C} > \max\{\hat{p}_{AC|B}, \hat{p}_{BC|A}\}$.

Idea: By simulating a large number of samples, $\hat{t}$ estimates what topology we expect R^* inference to converge to.

Simulation Results 1: R^* inconsistency zone

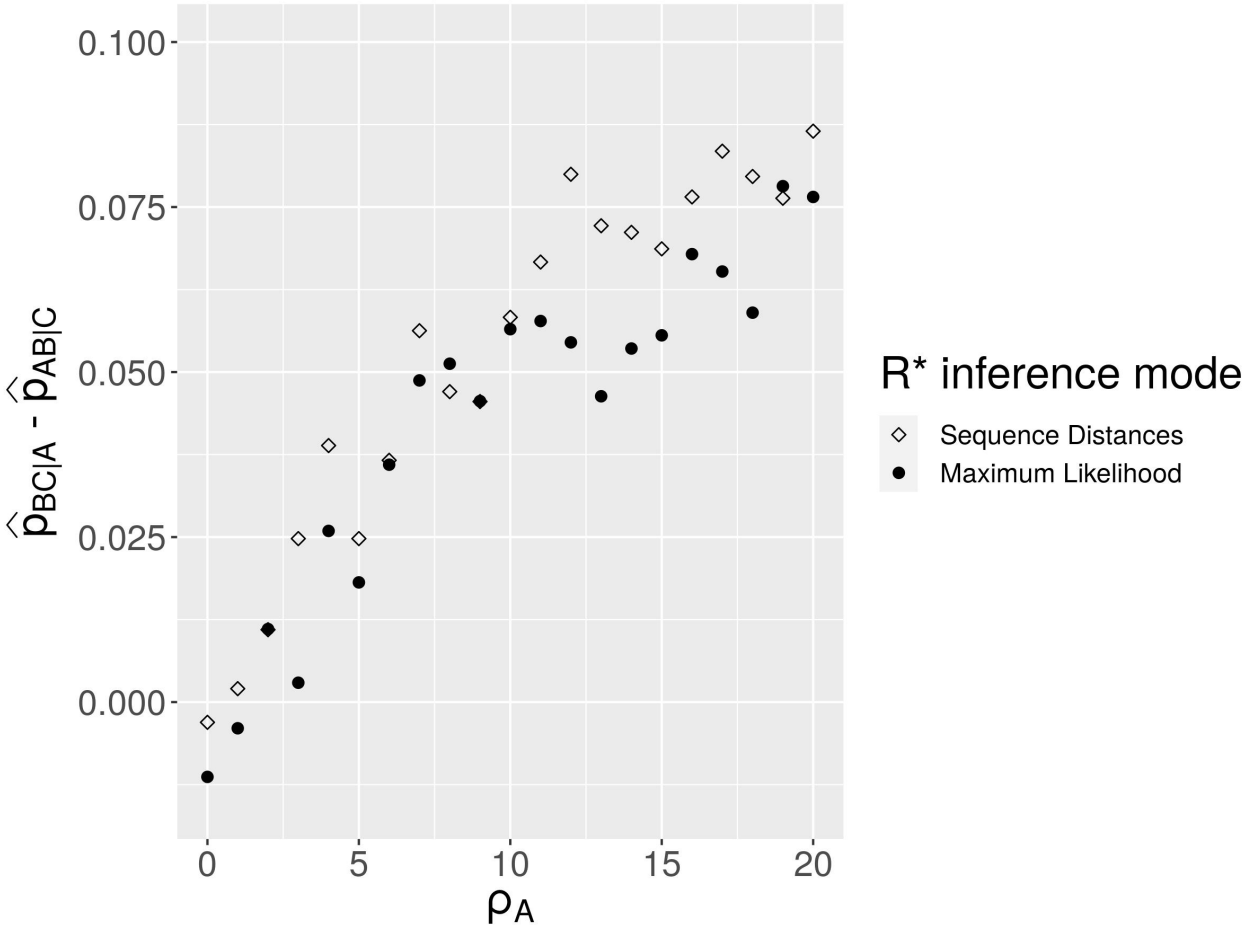
- Recombination **only** in population A
- Each dot $=$ 1 parameter regime
- Simulated $m = 10^6$ MSAs per parameter regime
- Dot color $=$ topology estimate
- Mutation rate $\theta = 0.1$



Simulation Results 2

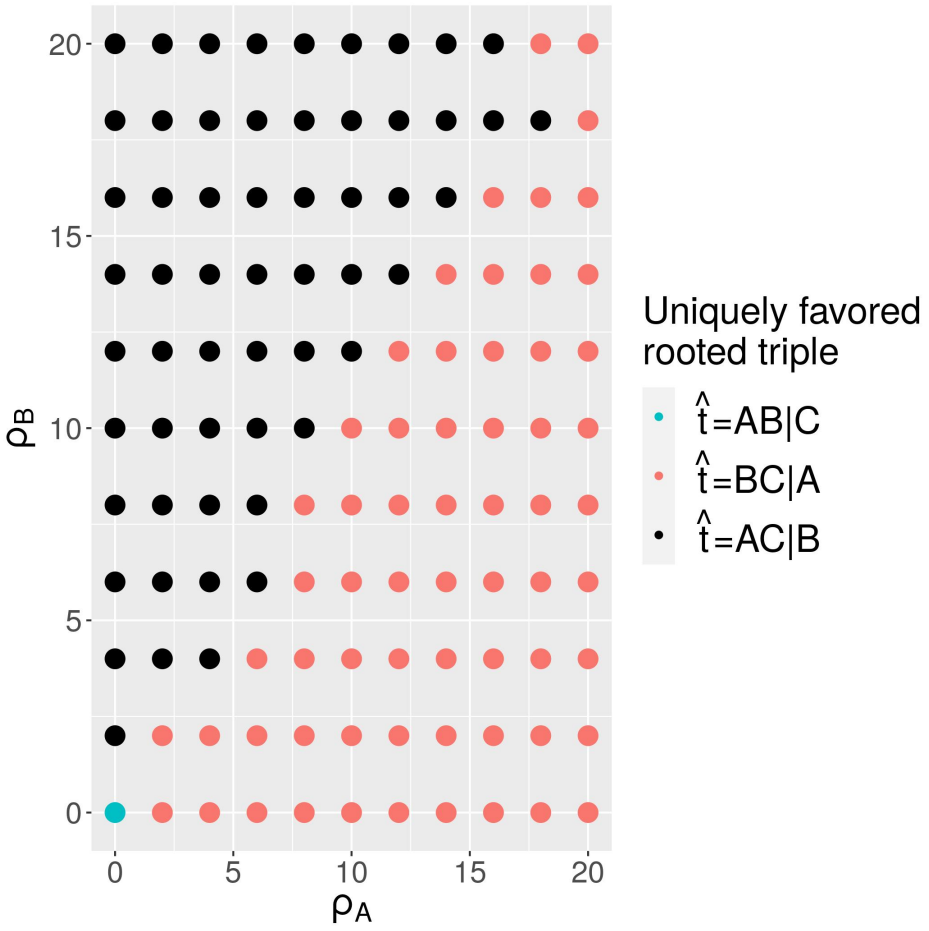
Effect of increasing ρ_A

- $m = 15,000$
- $\theta = 0.1$
- Short internal branch: $f = 0.01$
- Effect increases with recombination rate
- Small effect: ≤ 0.1



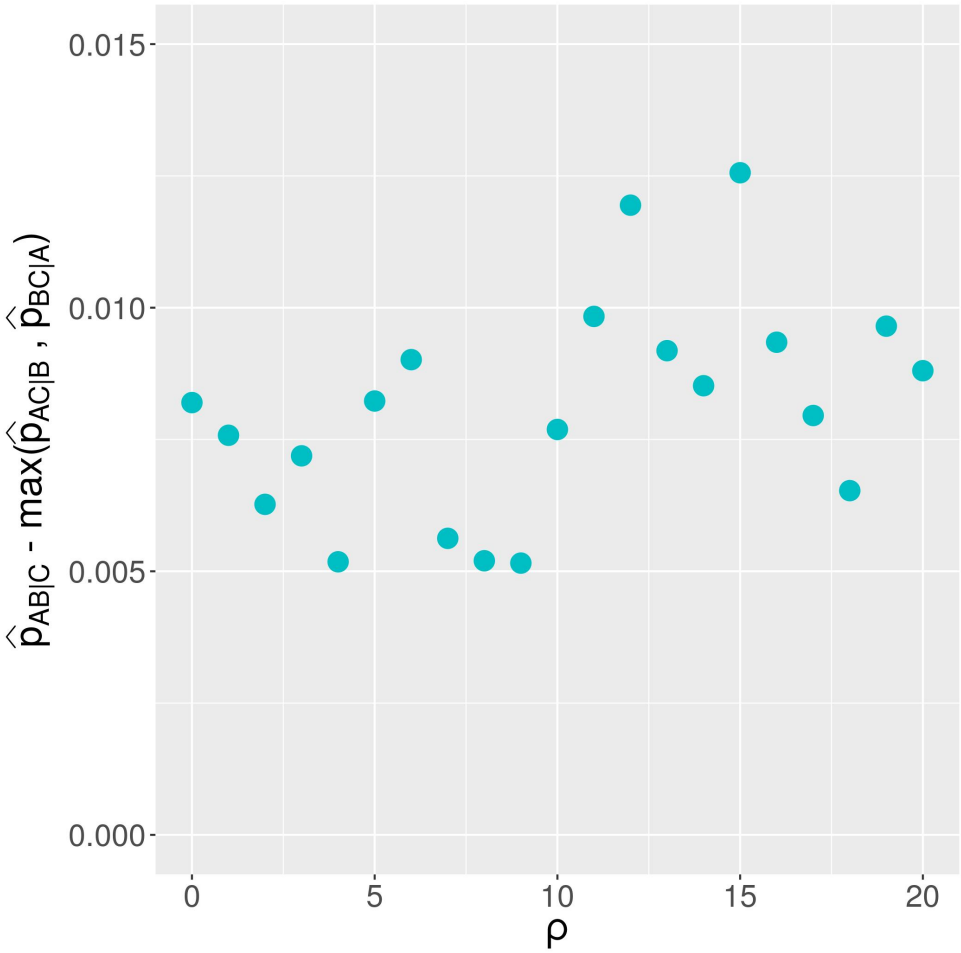
Simulation Results 3

- Recombination in **both** populations A and B.
- $m = 10^5$
- $f = 0.01$
- $\theta = 0.01$
- Higher relative rate of recombination in a population leads to overestimation of divergence time



Simulation Results 4

- Equal recombination rates in A, B, C and AB.
- $m = 10^6$
- $\theta = 0.1$
- $f = 0.01$
- When recombination rates are comparable across taxa, there is no impact on inference



Conclusions

- Key finding: overestimation of the divergence times of those taxa exhibiting disproportionately high intralocus recombination rates.
- Impact is minor?
  - relatively small effect
  - short internal branch lengths
  - effect only arises with *differential* rates of recombination between taxa

Other Projects and Future Work

1. Recombination project
   - Related to the recombination project: are other methods resilient to this effect? (STEAC appears so...)
   - Extension to unrooted quartets, other inference methods (e.g. maximum-likelihood)
   - F-statistics in admixture graphs (ongoing) – we expect this to have some connection to the recombination project.
2. GDL-ILS – impact of gene duplication and loss on quartet-based methods (joint with B. Legried and S. Roch)
3. Tree Depth – impact of tree depth on sample complexity (ongoing)

References I

[1] David Bryant and Matthew W Hahn. "The concatenation question". In: *Phylogenetics in the Genomic Era*. Ed. by C. Scornavacca, F. Delsuc, and N. Galtier. https://hal.inria.fr/PGE, 2020. Chap. 3.4, 3.4:1–3.4:23.

[2] James H Degnan. "Anomalous unrooted gene trees". In: *Systematic biology* 62.4 (2013), pp. 574–590.

[3] James H Degnan and Noah A Rosenberg. "Discordance of species trees with their most likely gene trees". In: *PLoS genet* 2.5 (2006), e68.

References II

[4] James H. Degnan and Noah A. Rosenberg. "Gene tree discordance, phylogenetic inference and the multispecies coalescent". In: *Trends in Ecology & Evolution* 24.6 (2009), pp. 332–340. ISSN: 0169-5347. DOI: https://doi.org/10.1016/j.tree.2009.01.009. URL: https://www.sciencedirect.com/science/article/pii/S0169534709000846.

[5] Hayley C Lanier and L Lacey Knowles. "Is recombination a problem for species-tree analyses?" In: *Systematic Biology* 61.4 (2012), pp. 691–701.

[6] Fábio K Mendes, Andrew P Livera, and Matthew W Hahn. "The perils of intralocus recombination for inferences of molecular convergence". In: *Philosophical Transactions of the Royal Society B* 374.1777 (2019), p. 20180244.

References III

[7] David E Sadava et al. *Principles of Life*. Sinauer Associates, 2014. URL: https://www.macmillanhighered.com/BrainHoney/Resource/6716/digital_first_content/trunk/test/hillis2e/hillis2e_ch16_2.html.

[8] Mark S Springer and John Gatesy. "Delimiting coalescence genes (c-genes) in phylogenomic data sets". In: *Genes* 9.3 (2018), p. 123.

[9] Mike Steel. *Phylogeny: discrete and random processes in evolution*. SIAM, 2016.

References IV

[10] Kazumasa Takemoto et al. "Genetic Exchange through Meiotic Recombination (IMAGE)". In: *Cell Reports* 31.8 (2020), p. 107686. ISSN: 2211-1247. DOI: https://doi.org/10.1016/j.celrep.2020.107686. URL: https://www.sciencedirect.com/science/article/pii/S2211124720306392;%20https://www.eurekalert.org/multimedia/549565.