- More formally,

$$\mathbb{E}\left(\begin{array}{c}n \text{ lineages coalesce} \\ \text{to } n-1\end{array}\right) = \int_0^\infty u \binom{n}{2} e^{-\binom{n}{2}u} \, du \quad \nearrow \text{ exponential} \quad = \frac{1}{\binom{n}{2}} = \frac{2}{n(n-1)}$$

Theorem: Under the Kingman coalescent model, given n lineages in a population, the expected time to the next coalescent event is
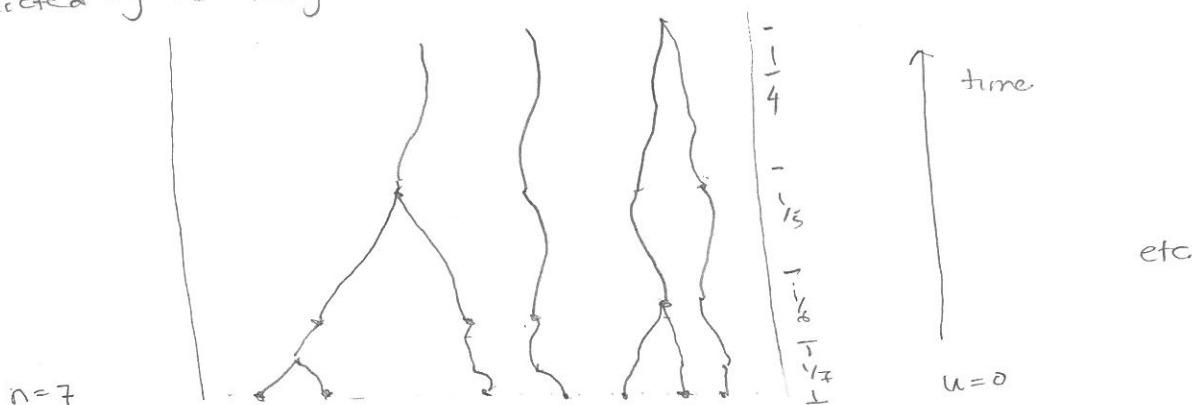
$$\mathbb{E}(n \text{ lineages coalesce to } n-1 \text{ lineages}) = \frac{2}{n(n-1)}$$

$$\frac{1}{\binom{n}{2}} = \frac{2}{n(n-1)}$$

n	2	3	4	5	6	..
	1	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{10}$	$\frac{1}{15}$	$\frac{1}{21}$ etc.

```
   1
   1 1
  1 2 1
 1 3 3 1
1 4 6 4 1
```

This has a profound effect on the types of gene trees predicted by the Kingman coalescent model.

$n = 7$

$\frac{1}{4}$, $\frac{1}{5}$, $\frac{1}{6}$, $\frac{1}{7}$

↑ time

$u = 0$

etc

i.e. Coalescence events are more numerous near the present.
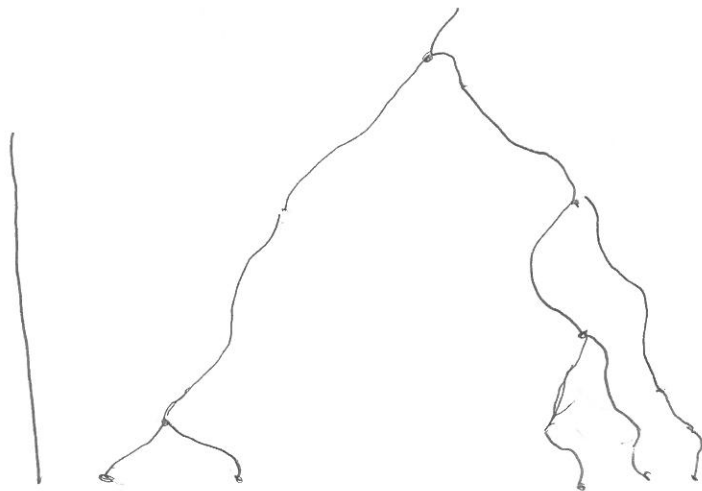
R example.

Finally, given n lineages then the expected time for them to coalesce down to one lineage is

$$\mathbb{E}(\text{gene tree w/ } n \text{ lineage formed}) = \sum_{i=2}^{n} \frac{1}{\binom{i}{2}} = \sum_{i=2}^{n} \frac{2}{i(i-1)} = 2\sum_{i=2}^{n}\left[\frac{1}{i-1} - \frac{1}{i}\right] = 2\left(1 - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \cdots \frac{1}{n-1} - \frac{1}{n}\right)$$

partial fractions

$$= 2\left(1 - \frac{1}{n}\right) \quad \text{and} \quad \lim_{n\to\infty} 2\left(1 - \frac{1}{n}\right) = 2 \quad \text{coalescent units}$$

It follows that modeling the process of n gene lineages coalescing to form a gene tree (down to 1 lineage) is a piecewise process

Demo:

$\bar{T}$ rate $= \binom{2}{2} = 1$ $\mathbb{E}\left(\begin{smallmatrix} \text{time to} \\ c. \end{smallmatrix}\right) = 1$

rate $\binom{3}{2} = 3$ $\mathbb{E}\left(\begin{smallmatrix} \text{time to} \\ c \end{smallmatrix}\right) = 1/3$

rate $\binom{4}{2}$ $\mathbb{E}\left(\begin{smallmatrix} \text{time} \\ \text{to c.} \end{smallmatrix}\right) = \frac{1}{6}$

rate $\binom{5}{2}$ $\mathbb{E}\left(\begin{smallmatrix} \text{time} \\ \text{to c.} \end{smallmatrix}\right) = 1/10$
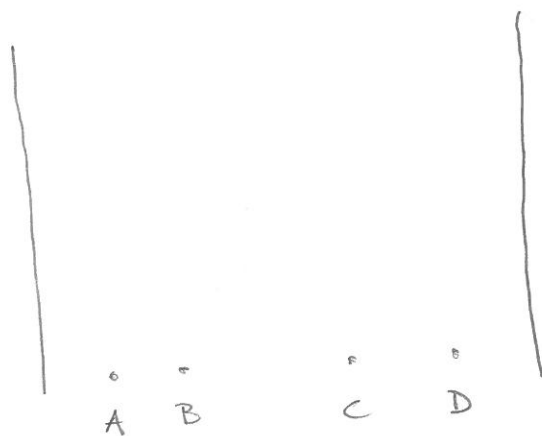
Computing gene tree probabilities under the Coalescent Model

Example: 4-taxon TOPOLOGICAL gene trees in a single population

General case: Degnan and Kubatko 2005 [Pamilo + Nei, Rosenberg Tavare]

Label the gene lineages A, B, C, D

All rooted **RANKED** gene trees are equally probable under the coalescent model

Shapes: balanced

Caterpillar or pectinate single ranking

12 total

6 total

2 rankings

Thus $\mathbb{P}\left(\underset{A\ B\ C\ D}{\wedge}\right) = \frac{1}{18}$ $\mathbb{P}(\text{any caterpillar}) = 1/18$

$\mathbb{P}\left(\underset{A\ B\ C\ D}{\wedge}\right) = 2/18$ 2 since two ranked versions of shape
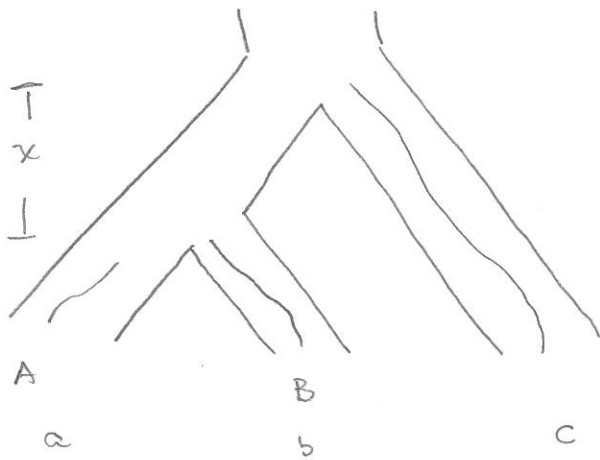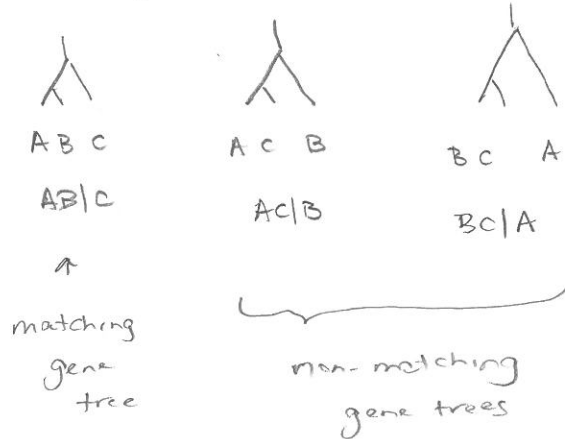
Example 2: Probability of gene tree topologies on a 3-taxon

Species tree:

Find the probabilities of the topological gene trees

A B C   A C B   B C   A

AB|C   AC|B   BC|A

↑   ⎵_____⎵

matching   non-matching
gene   gene trees
tree

Easiest to compute $\mathbb{P}(AC|B)$ (non-matching gene tree first)
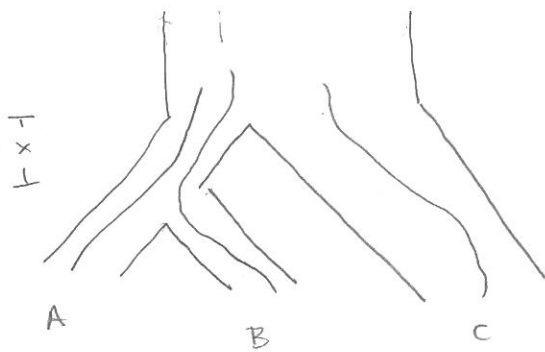
The only way AC|B can form on σ
is if A, B do <u>not</u> coalesce in the
immediately ancestral population. I.e
all three lineages enter the "above the
root" population

$$\mathbb{P}(AB|C) = \frac{1}{3} \underbrace{e^{-x}}$$

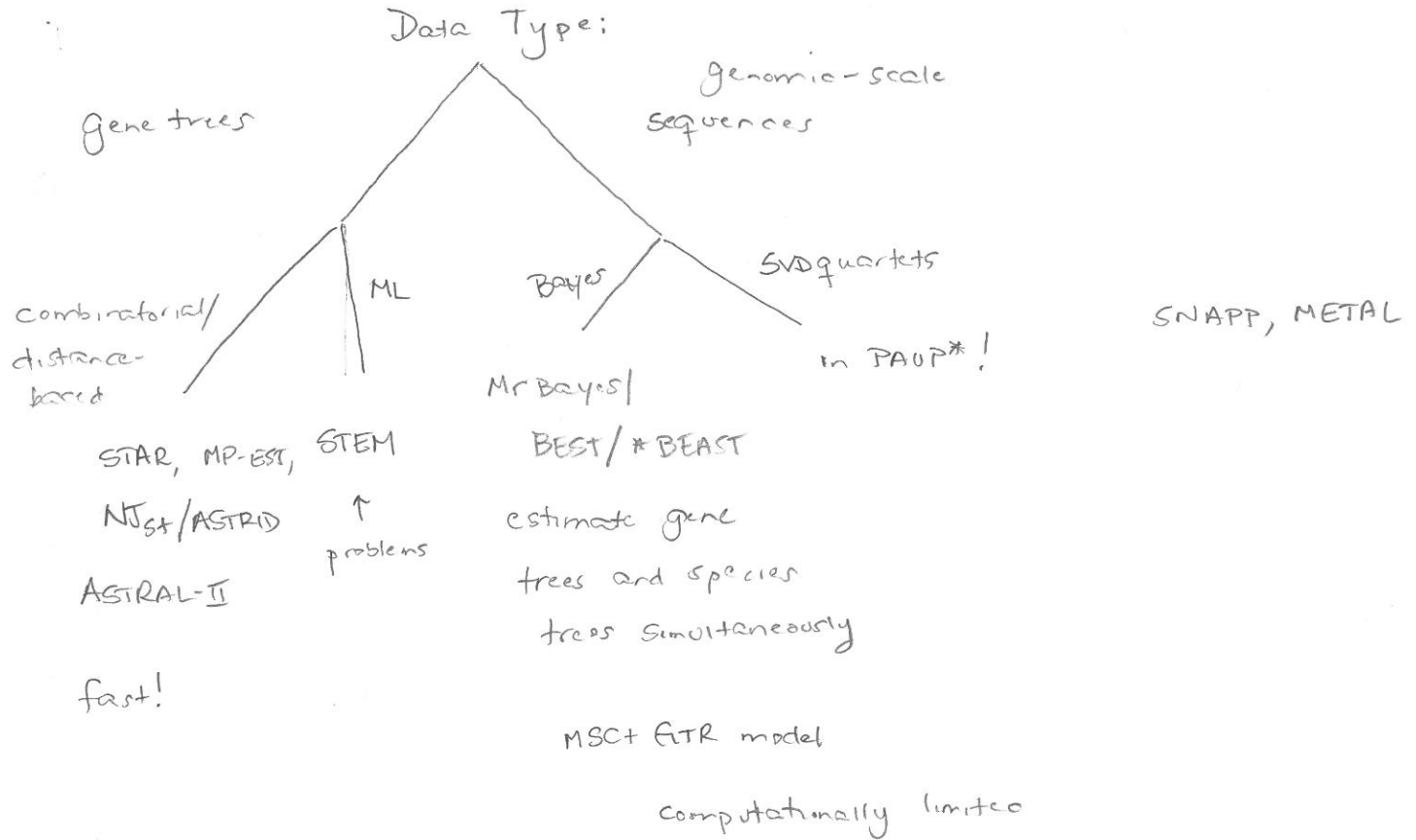⤷   A,B did <u>not</u> coalesce before time x

⤷ Each of the 3 topologies are equally likely by
exchangeability once they enter the population
above the root

$\sigma: ((a,b):x, c)$

$\mathbb{P}(AB|C) = 1 - 2e^{-x}$ $P(AC|B) = P(BC|A) = e^{-x}$!

This has been used to estimate divergence times   (Australian finches,
                                                  HC, G)

Inference: of Species Trees:

Data Type:

Gene trees

genomic-scale
sequences

combinatorial/
distance-
based
     ML

Bayes

SVDquartets

in PAUP* !

SNAPP, METAL

STAR, MP-EST, STEM

NJst/ASTRID

ASTRAL-II

↑
problems

fast!

MrBayes

BEST / *BEAST

estimate gene
trees and species
trees simultaneously

MSC+ GTR model

computationally limited

? ? ?