Coalescent Theory: An Introduction for Phylogenetics

Laura Salter Kubatko
Departments of Statistics and
Evolution, Ecology, and Organismal Biology
The Ohio State University
lkubatko@stat.ohio-state.edu

May 11, 2010

Relationship between population genetics and phylogenetics

- Population genetics: Study of genetic variation within a population

Relationship between population genetics and phylogenetics

- Population genetics: Study of genetic variation within a population
- Phylogenetics: Use genetic variation between taxa (species, populations) to infer evolutionary relationships

Relationship between population genetics and phylogenetics

- Population genetics: Study of genetic variation within a population
- Phylogenetics: Use genetic variation between taxa (species, populations) to infer evolutionary relationships
- So far, we've assumed:
  - Each taxon is represented by a single sequence – this is often called "exemplar sampling"
  - We have data for a single gene and wish to estimate the evolutionary history for that gene (the *gene tree* or *gene phylogeny*)

Relationship between population genetics and phylogenetics

- Given current technology, we could do much more:
  - Sample many individuals within each taxon (species, population, etc.)
  - Sequence many genes for all individuals

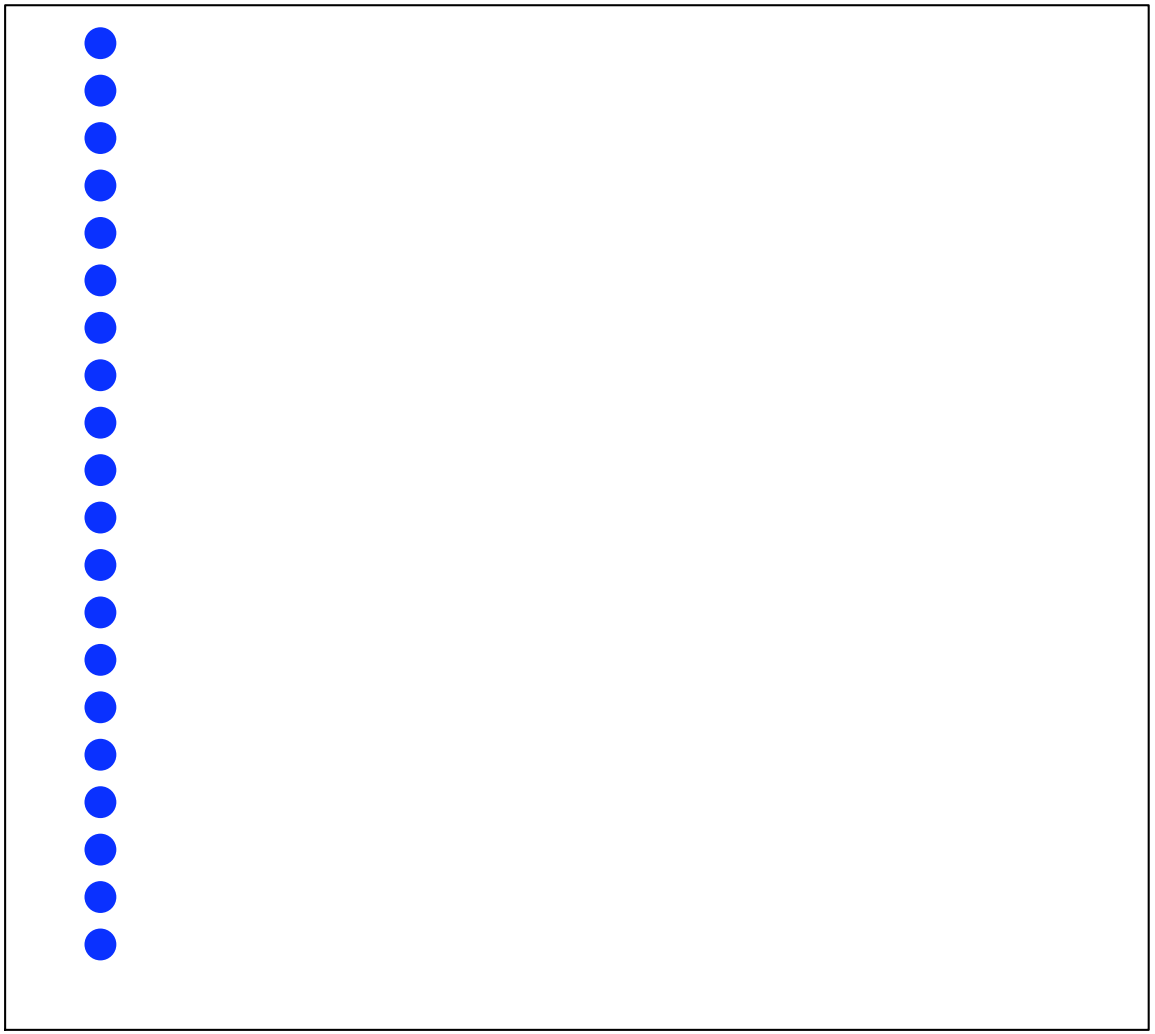
Relationship between population genetics and phylogenetics

- Given current technology, we could do much more:
  - Sample many individuals within each taxon (species, population, etc.)
  - Sequence many genes for all individuals
- Need models at two levels:
  - Model what happens within each population (standard population genetics)
  - Apply within-population models to each population represented on a phylogeny (more recent work)

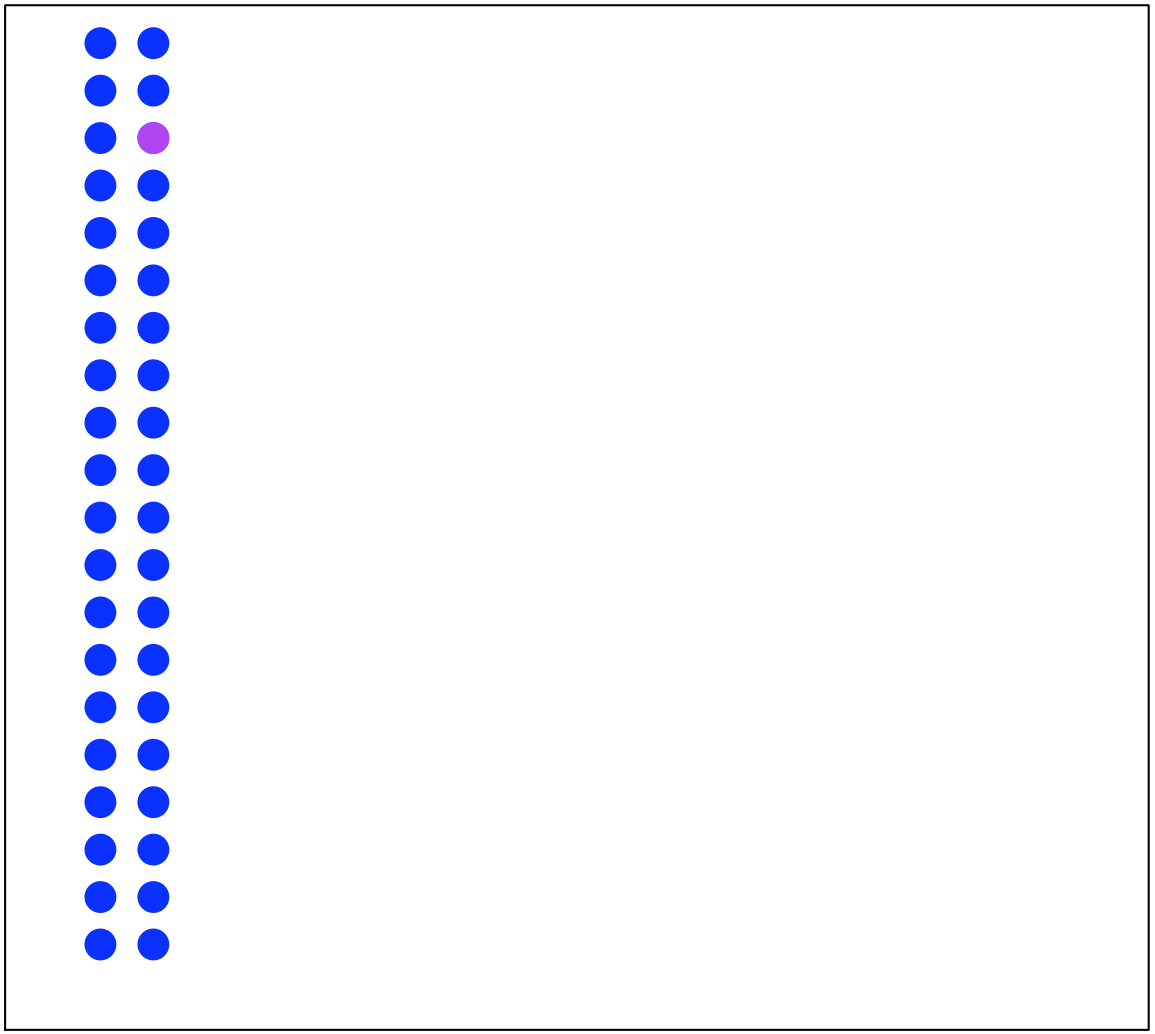
Wright-Fisher Model

- Assumptions:
  - Population of $2N$ gene copies
  - Discrete, non-overlapping generations of equal size
  - Parents of next generation of $2N$ genes are picked randomly with replacement from preceding generation (genetic differences have no fitness consequences)
  - Probability of a specific parent for a gene in the next generation is $\frac{1}{2N}$

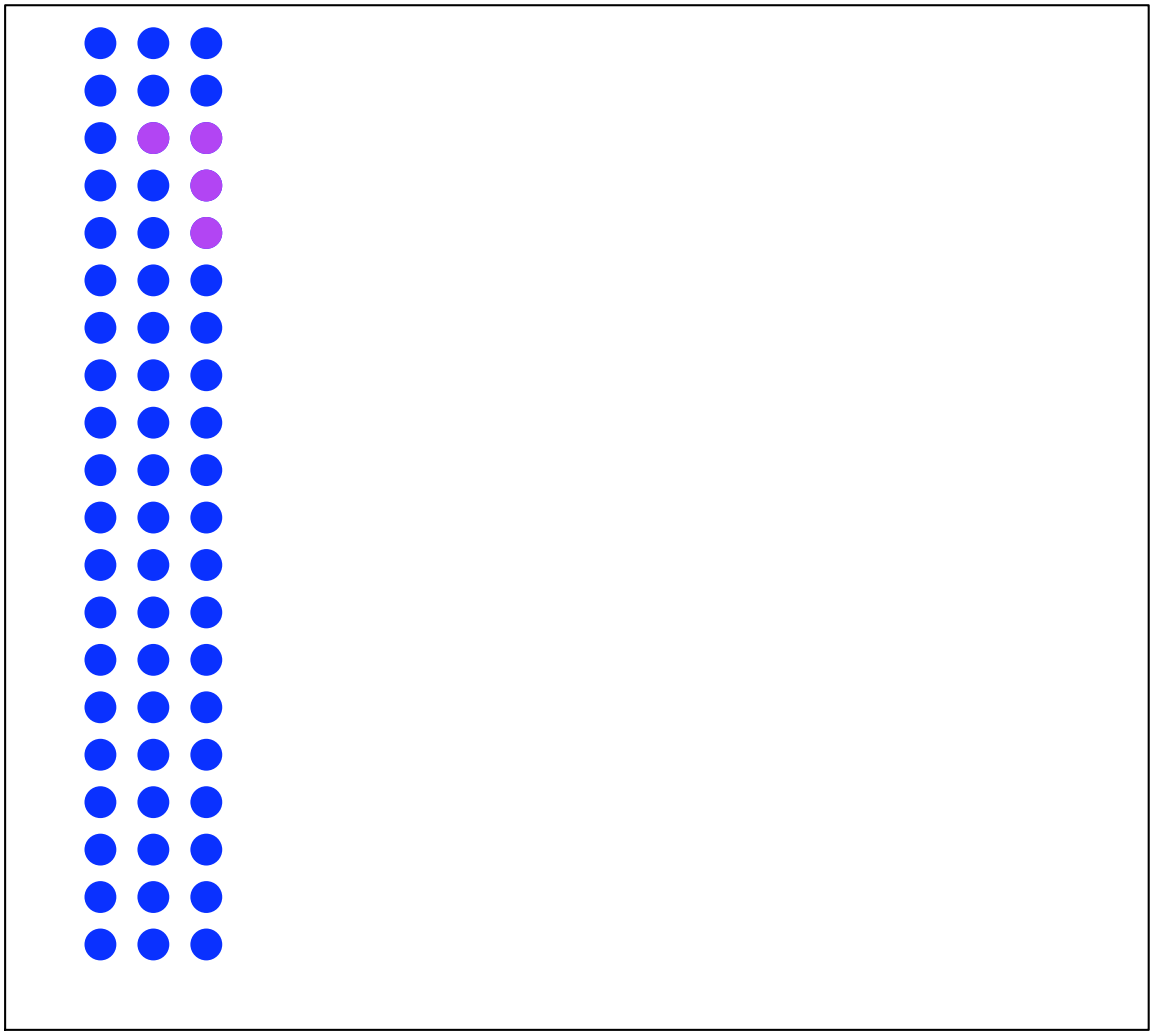
Wright-Fisher Model



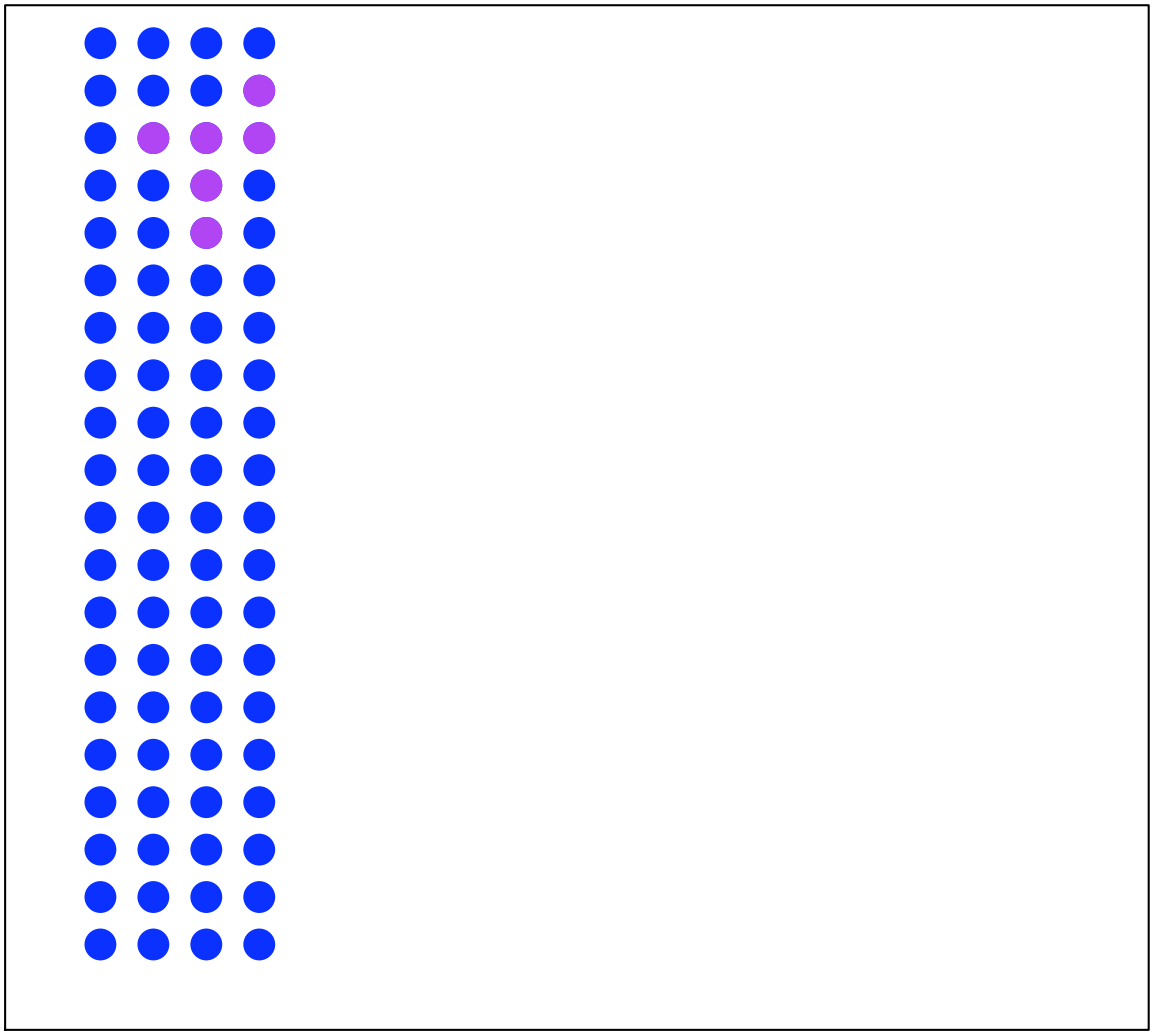
Wright-Fisher Model



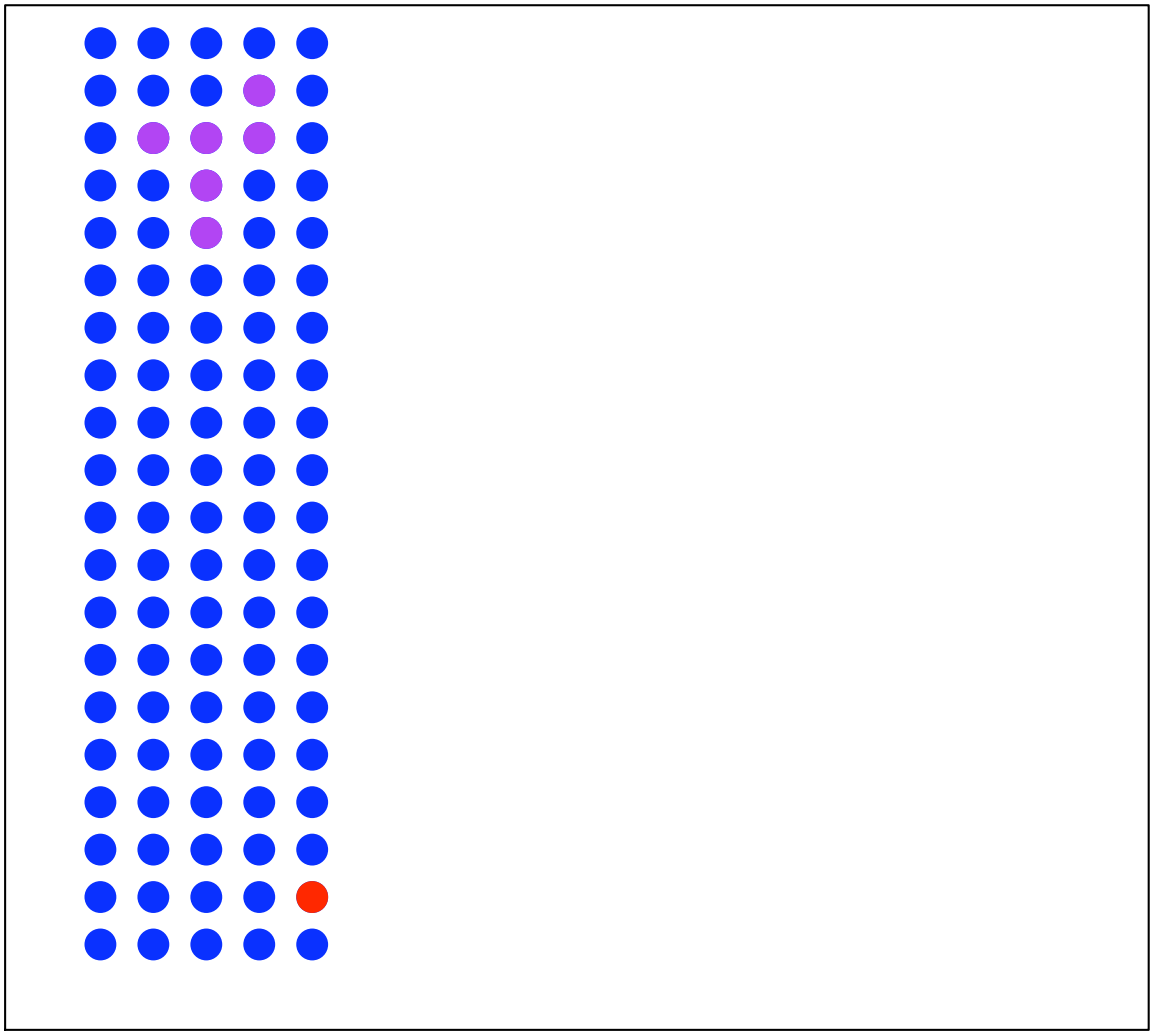
Wright-Fisher Model



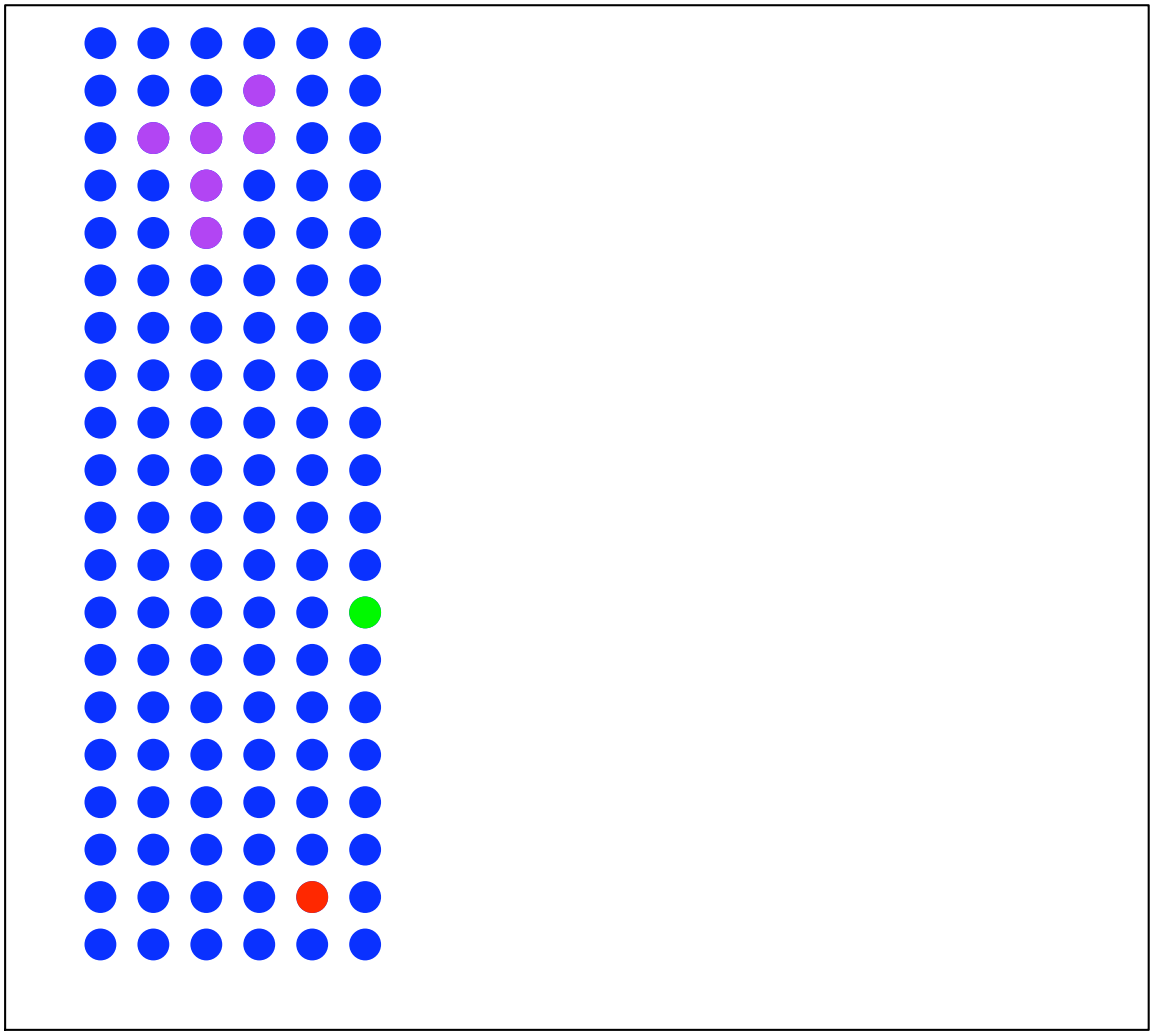
Wright-Fisher Model



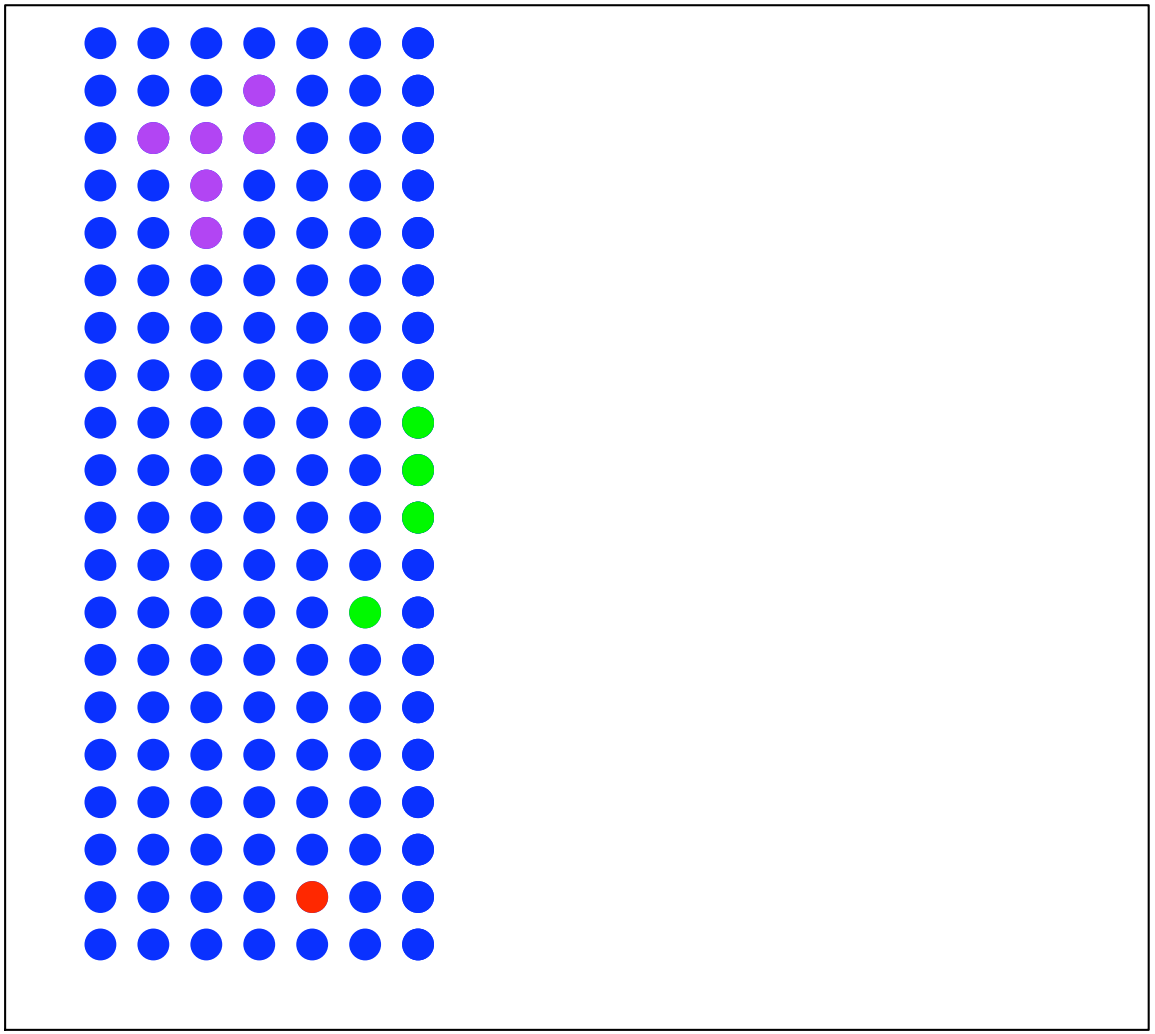
Wright-Fisher Model



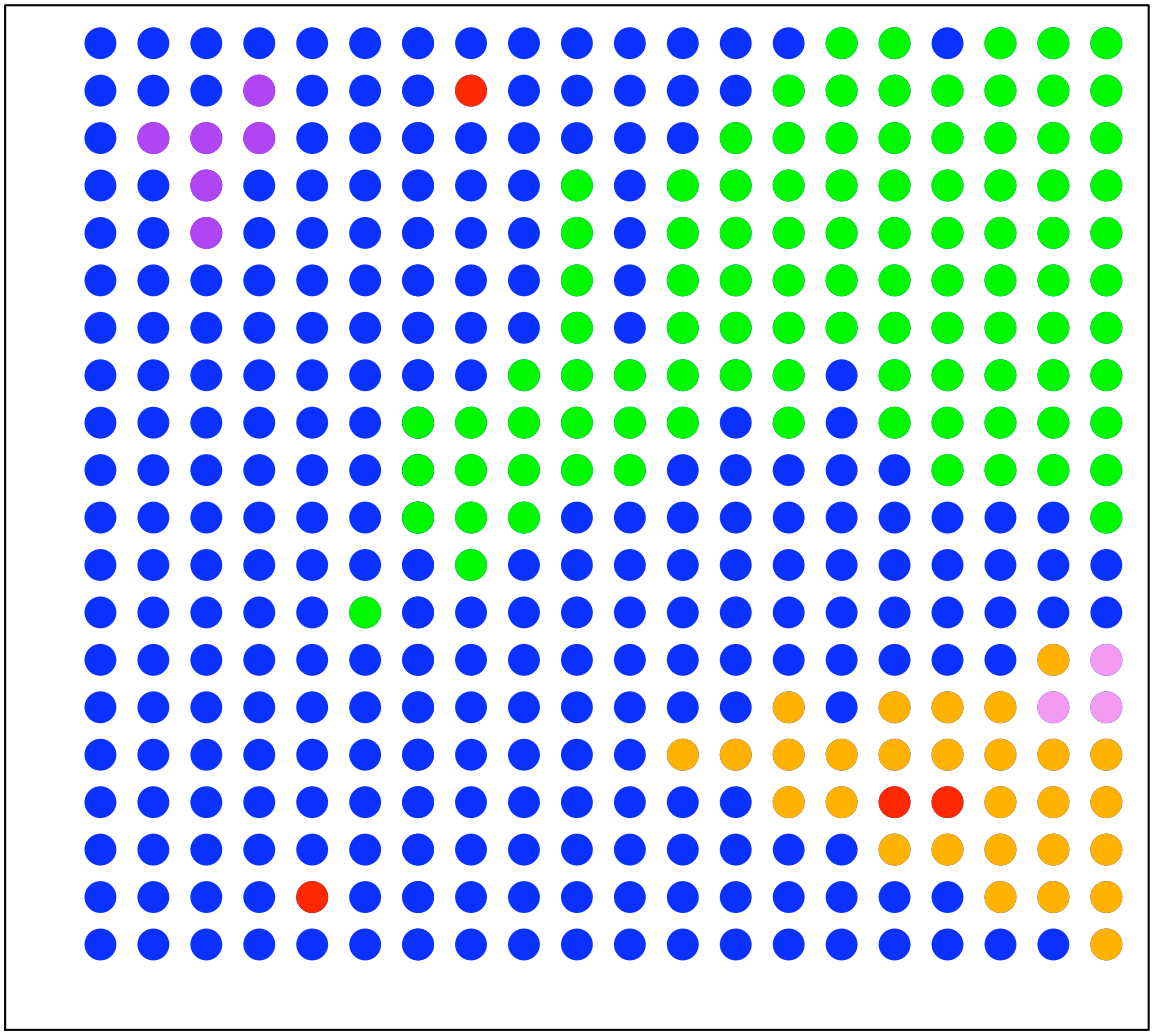
Wright-Fisher Model



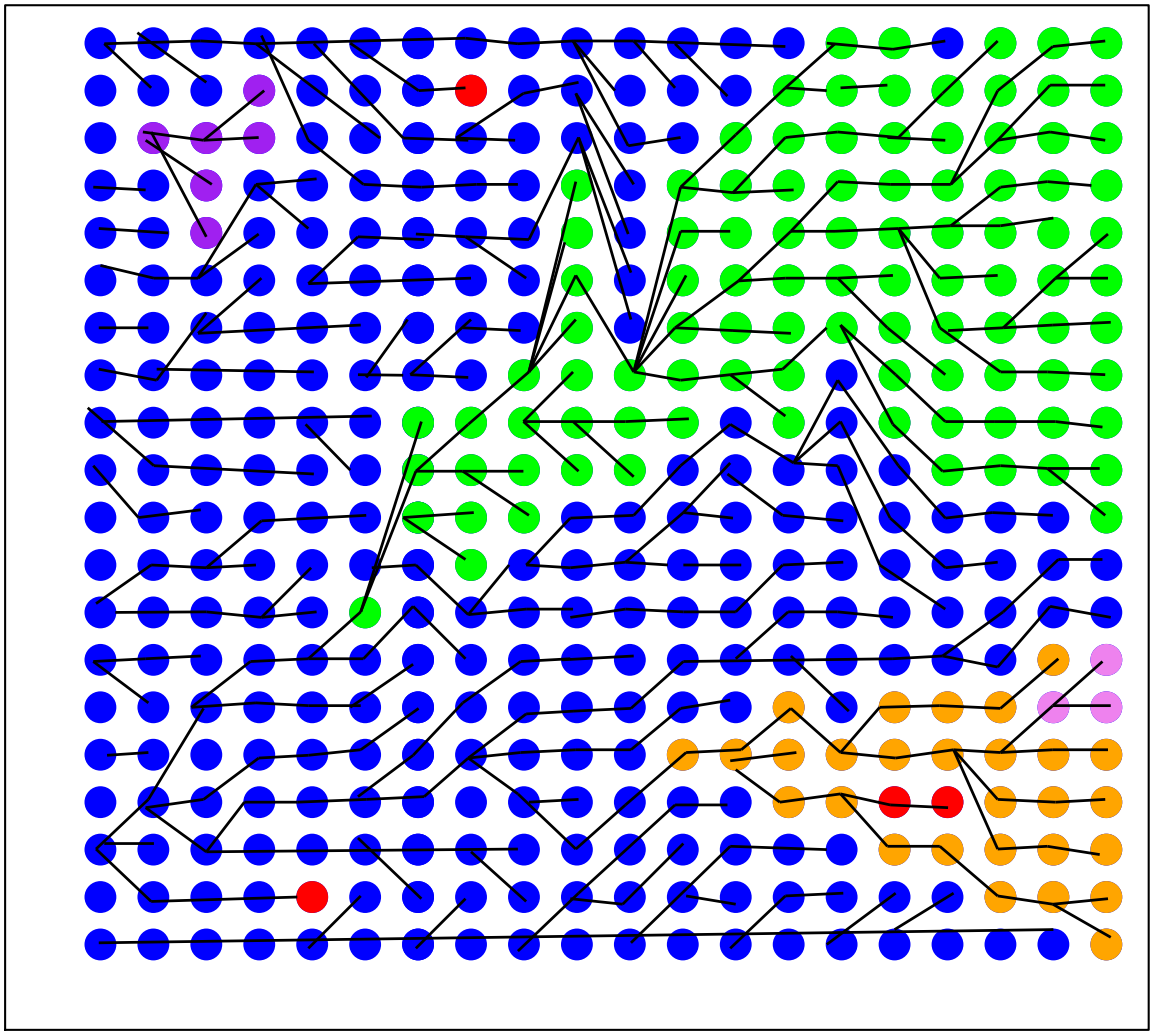
Wright-Fisher Model



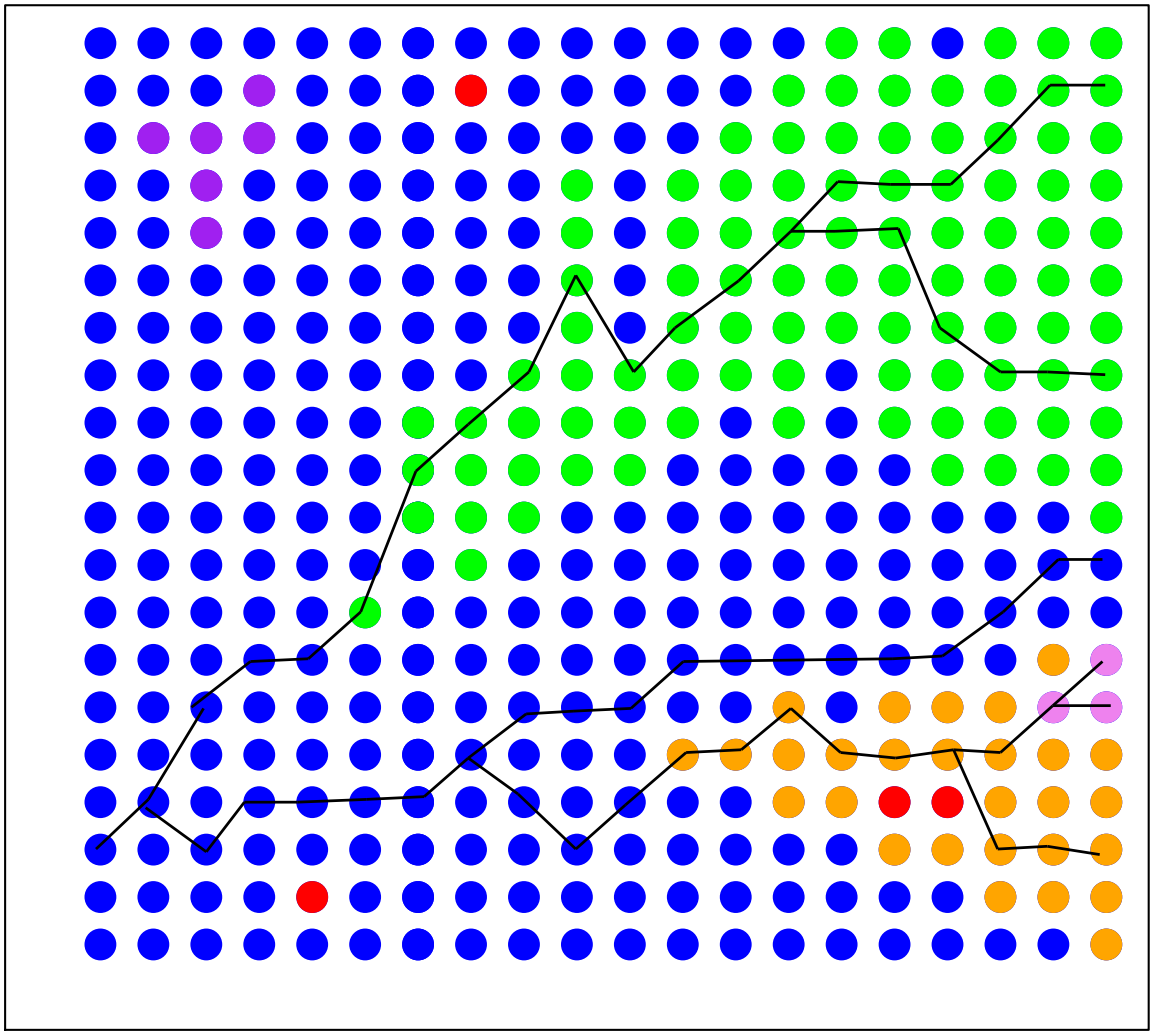
Wright-Fisher Model



Wright-Fisher Model



Wright-Fisher Model



The Coalescent Model

- Discrete Time Coalescent
  - P(two genes have same parent in the previous generation) is $\frac{1}{2N}$
  - Number of generations since two genes first shared a common ancestor $\sim$ Geometric($\frac{1}{2N}$)
  - Number of generations since at least two genes in a sample of k shared a common ancestor $\sim$ Geometric($\frac{k(k-1)}{4N}$)

Number of generations since at least two genes in a sample of k shared a common ancestor $\sim$ Geometric($\frac{k(k-1)}{4N}$)

- Define $G_{k,k}$ to be the probability that k genes have k distinct ancestors in the previous generation. Then

Number of generations since at least two genes in a sample of k shared a common ancestor $\sim$ Geometric($\frac{k(k-1)}{4N}$)

- Define $G_{k,k}$ to be the probability that k genes have k distinct ancestors in the previous generation. Then

$$G_{k,k} = \left(\frac{2N-1}{2N}\right)\left(\frac{2N-2}{2N}\right)\cdots\left(\frac{2N-(k-1)}{2N}\right)$$

Number of generations since at least two genes in a sample of k shared a common ancestor $\sim$ Geometric($\frac{k(k-1)}{4N}$)

- Define $G_{k,k}$ to be the probability that k genes have k distinct ancestors in the previous generation. Then

$$\begin{aligned} G_{k,k} &= \left(\frac{2N-1}{2N}\right)\left(\frac{2N-2}{2N}\right)\cdots\left(\frac{2N-(k-1)}{2N}\right) \\ &= \left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)\cdots\left(1-\frac{k-1}{2N}\right) \end{aligned}$$

Number of generations since at least two genes in a sample of k shared a common ancestor $\sim$ Geometric($\frac{k(k-1)}{4N}$)

- Define $G_{k,k}$ to be the probability that k genes have k distinct ancestors in the previous generation. Then

$$\begin{aligned} G_{k,k} &= \left(\frac{2N-1}{2N}\right)\left(\frac{2N-2}{2N}\right)\cdots\left(\frac{2N-(k-1)}{2N}\right) \\ &= \left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)\cdots\left(1-\frac{k-1}{2N}\right) \\ &= 1-\left(\frac{1+2+3+\cdots+(k-1)}{2N}\right)+\mathcal{O}\left(\frac{1}{N^2}\right) \end{aligned}$$

Number of generations since at least two genes in a sample of k shared a common ancestor $\sim$ Geometric($\frac{k(k-1)}{4N}$)

- Define $G_{k,k}$ to be the probability that k genes have k distinct ancestors in the previous generation. Then

$$\begin{aligned} G_{k,k} &= \left(\frac{2N-1}{2N}\right)\left(\frac{2N-2}{2N}\right)\cdots\left(\frac{2N-(k-1)}{2N}\right) \\ &= \left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)\cdots\left(1-\frac{k-1}{2N}\right) \\ &= 1-\left(\frac{1+2+3+\cdots+(k-1)}{2N}\right)+\mathcal{O}\left(\frac{1}{N^2}\right) \\ &= 1-\frac{k(k-1)}{4N}+\mathcal{O}\left(\frac{1}{N^2}\right) \end{aligned}$$

Number of generations since at least two genes in a sample of k shared a common ancestor $\sim$ Geometric($\frac{k(k-1)}{4N}$)

- Therefore, the probability that at least two genes share a common ancestor in the previous generation is

$$1 - G_{k,k} = \frac{k(k-1)}{4N} + \mathcal{O}\left(\frac{1}{N^2}\right)$$

Number of generations since at least two genes in a sample of k shared a common ancestor $\sim$ Geometric($\frac{k(k-1)}{4N}$)

- Therefore, the probability that at least two genes share a common ancestor in the previous generation is

$$1 - G_{k,k} = \frac{k(k-1)}{4N} + \mathcal{O}\left(\frac{1}{N^2}\right)$$

- Since this is the same in each generation, we have that the number of generations until at least two genes in a sample of k shared a common ancestor $\sim$ Geometric($\frac{k(k-1)}{4N}$)

Continuous-time Coalescent – Kingman's Approximation

- Kingman (1982a, b, c) considered the case where N (population size) is very large relative to k (sample size).
- Then, we can ignore the terms that are $\mathcal{O}(1/N^2)$ – this amounts to assuming that three or more genes coalescing in the same generation happens relatively rarely in comparison to two genes coalescing in one generation.

Continuous-time Coalescent – Kingman's Approximation

- Kingman (1982a, b, c) considered the case where N (population size) is very large relative to k (sample size).
- Then, we can ignore the terms that are $\mathcal{O}(1/N^2)$ – this amounts to assuming that three or more genes coalescing in the same generation happens relatively rarely in comparison to two genes coalescing in one generation.
- We have
  - Time since two gene copies had a common ancestor $\sim$ exponential ($\mu = 2N$)
  - Time to coalescence of k gene copies into $k-1$ $\sim$ exponential($\mu = 4N/(k(k-1))$)

  where time, T, is measured in number of generations.

Continuous-time Coalescent – Kingman's Approximation

- Kingman (1982a, b, c) considered the case where N (population size) is very large relative to k (sample size).
- Then, we can ignore the terms that are $\mathcal{O}(1/N^2)$ – this amounts to assuming that three or more genes coalescing in the same generation happens relatively rarely in comparison to two genes coalescing in one generation.
- We have
  - Time since two gene copies had a common ancestor $\sim$ exponential ($\mu = 2N$)
  - Time to coalescence of k gene copies into $k-1$ $\sim$ exponential($\mu = 4N/(k(k-1))$)

  where time, T, is measured in number of generations.
- This is generally a very good approximation, provided N is large enough.

Continuous-time Coalescent – Kingman's Approximation

- To generate a genealogy of k genes under Kingman's coalescent:
  - Draw an observation from an exponential distribution with mean $\mu = 4N/(k(k-1))$. This will be the time of the first coalescent event (looking from the present backwards in time).

Continuous-time Coalescent – Kingman's Approximation

- To generate a genealogy of k genes under Kingman's coalescent:
  - Draw an observation from an exponential distribution with mean $\mu = 4N/(k(k-1))$. This will be the time of the first coalescent event (looking from the present backwards in time).
  - Pick two lineages at random to coalescence.

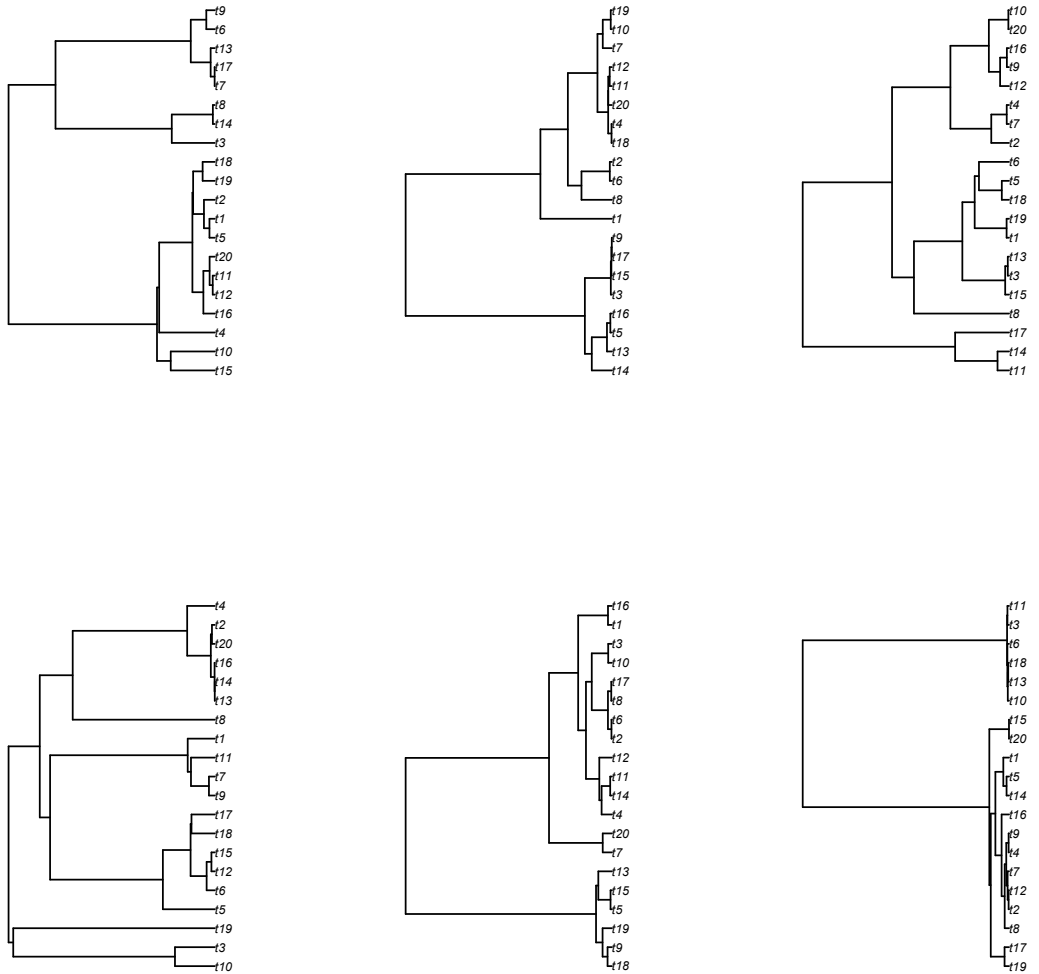
Continuous-time Coalescent – Kingman's Approximation

- To generate a genealogy of k genes under Kingman's coalescent:
  - Draw an observation from an exponential distribution with mean $\mu = 4N/(k(k-1))$. This will be the time of the first coalescent event (looking from the present backwards in time).
  - Pick two lineages at random to coalescence.
  - Decrease k by 1.

Continuous-time Coalescent – Kingman's Approximation

- To generate a genealogy of k genes under Kingman's coalescent:
  - Draw an observation from an exponential distribution with mean $\mu = 4N/(k(k-1))$. This will be the time of the first coalescent event (looking from the present backwards in time).
  - Pick two lineages at random to coalescence.
  - Decrease k by 1.
  - If $k = 1$, stop. Otherwise, repeat these steps.

Example Genealogies Under Kingman's Coalescent

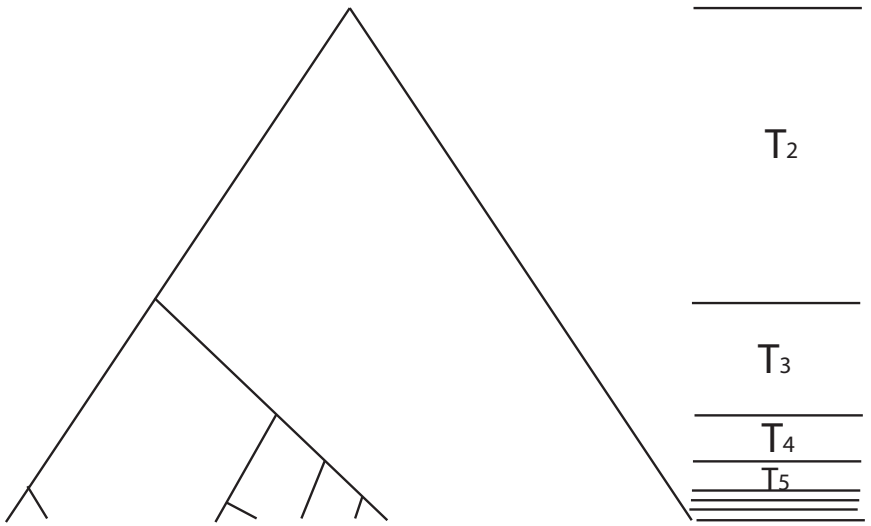


Properties of Genealogies

- Two measures of the size of a genealogy are commonly defined:
  - T_{MRCA} = the time of the most recent common ancestor of all lineages sampled
  - T_{total} = the total time represented by the geneaology
- Of interest are the mean, variance, and probability distribution of these.

Properties of Genealogies

- Define T_i to be the time in the history of the sample during which there were exactly i ancestral lineages.
- Note that $T_{MRCA} = \sum_{i=1}^{k} T_i$ and $T_{total} = \sum_{i=1}^{k} iT_i$



Properties of Genealogies - T_{MRCA}

- Note that $T_{MRCA} = \sum_{i=1}^{k} T_i$ and $T_i \sim Exp(\mu = \frac{4N}{i(i-1)})$
- Therefore, the mean is

$$\begin{aligned} E(T_{MRCA}) &= \sum_{i=2}^{k} E(T_i) = \sum_{i=2}^{k} \frac{4N}{i(i-1)} \\ &= 4N \sum_{i=2}^{k} \left(\frac{1}{i-1} - \frac{1}{i} \right) \\ &= 4N \left(1 - \frac{1}{k} \right) \end{aligned}$$

- If time is measured in units of $2N$ generations (*coalescent units*), then the mean is $2(1 - \frac{1}{k})$

Properties of Genealogies - T_{MRCA}

- Mean time to coalescence of all lineages is

$$4N\left(1-\frac{1}{k}\right)$$

- Notes:
  - When k is large, it takes $\approx 4N$ generations to reach the MRCA
  - When $k = 2$, it takes $\approx 2N$ generations to reach the MRCA
  - For a large sample, much of the total time represented in the genealogy will be spent waiting for the last coalescence to occur.

Properties of Genealogies - T_{MRCA}

- We can also show that $Var(T_{MRCA}) = (4N^2)\sum_{i=2}^{k} \frac{1}{i^2(i-1)^2}$
- We can show that as the sample size $k \to \infty$, $Var(T_{MRCA})$ converges to $4\pi^2/3 - 12 \approx 1.16$ (in coalescent units).
- Since T_{MRCA} is the sum of $k-1$ independent exponential random variables T_i, we have the following distribution for T_{MRCA}:

$$f_{T_{MRCA}}(t) = \sum_{i=2}^{k} \binom{i}{2} e^{-\binom{i}{2}t} \prod_{j=2, j\neq i}^{k} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}}$$

Properties of Genealogies - T_{total}

- Note that $T_{total} = \sum_{i=1}^{k} iT_i$ and $T_i \sim Exp(\mu = \frac{4N}{i(i-1)})$
- Therefore, the mean is

$$\begin{aligned} E(T_{total}) &= \sum_{i=2}^{k} iE(T_i) = \sum_{i=2}^{k} i\frac{4N}{i(i-1)} \\ &= 4N\sum_{i=1}^{k-1}\frac{1}{i} \end{aligned}$$

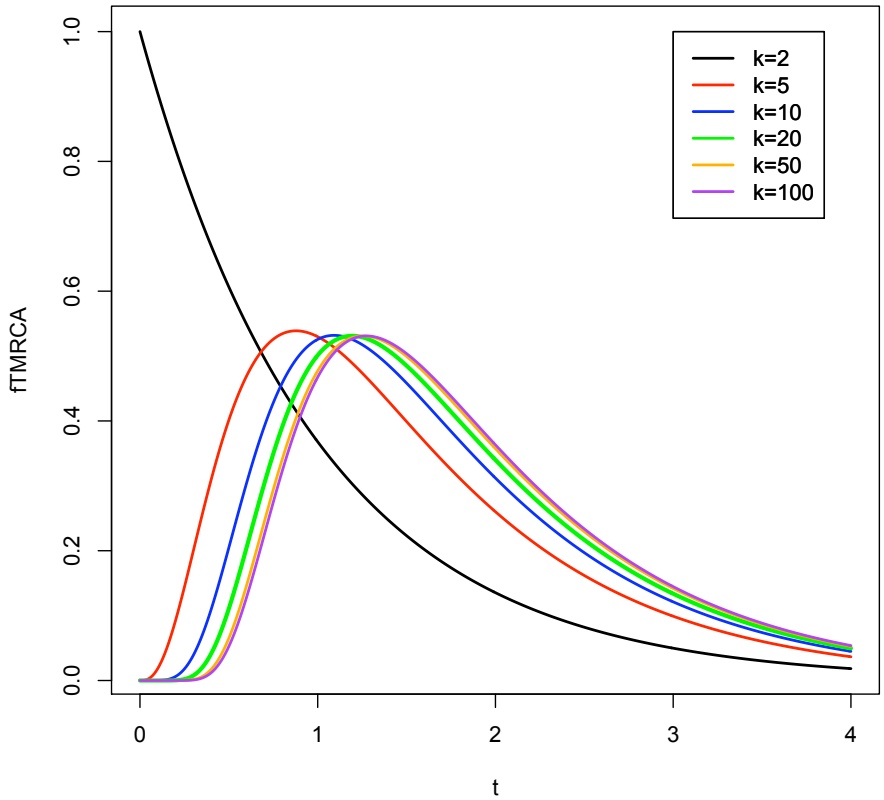
- If time is measured in units of $2N$ generations (*coalescent units*), then the mean is $2\sum_{i=1}^{k-1}\frac{1}{i}$

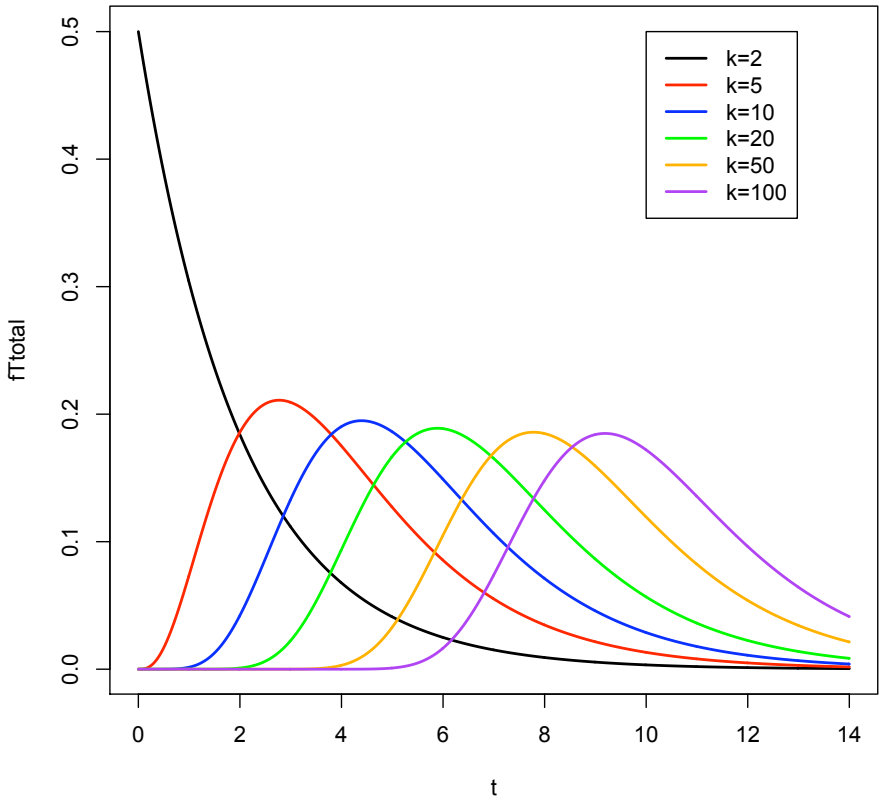
Properties of Genealogies - T_{total}

- We can also show that $Var(T_{total}) = (2N^2)\left[4\sum_{i=1}^{k-1}\frac{1}{i^2}\right]$
- Note that as the sample size $k \to \infty$, $Var(T_{total})$ converges to $2\pi^2/3 \approx 6.58$ (in coalescent units).
- Since T_{total} is the sum of $k-1$ independent exponential random variables iT_i, we have the following distribution for T_{total}:

$$f_{T_{total}}(t) = \sum_{i=2}^{k} \frac{i-1}{2} e^{-\frac{i-1}{2}t} \prod_{j=2, j\neq i}^{k} \frac{j-1}{j-i}$$

Properties of Genealogies - T_{MRCA} and T_{total}





Properties of Genealogies

- We need one more quantity to be able to link our population genetics model to our phylogenetic model – the probability that a specified number of coalescent events have occurred in a fixed amount of time, t.

Properties of Genealogies

- We need one more quantity to be able to link our population genetics model to our phylogenetic model – the probability that a specified number of coalescent events have occurred in a fixed amount of time, t.

- The probability that u lineages coalesce into v lineages in time t is given by (Tavare, 1984; Watterson, 1984; Takahata and Nei, 1985; Rosenberg, 2002)

$$P_{uv}(t) = \sum_{j=v}^{u} e^{-j(j-1)t/2} \frac{(2j-1)(-1)^{j-v}}{v!(j-v)!(v+j-1)} \prod_{y=0}^{j-1} \frac{(v+y)(u-y)}{u+y}$$

Properties of Genealogies

- When u and v are small, these are easy to compute. For example,

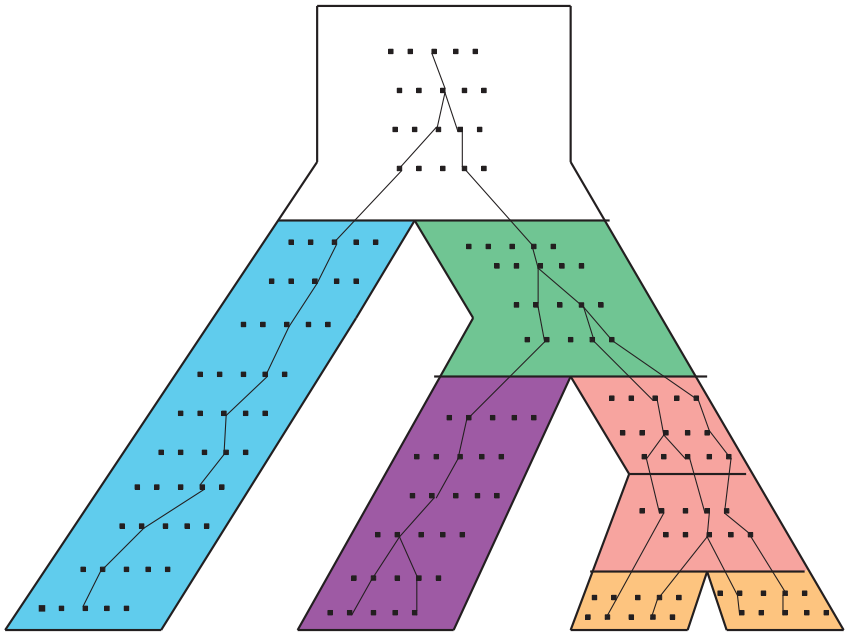
$$\begin{aligned} P_{21}(t) &= \text{probability that 2 lineages coalescence to 1 lineage in time t} \\ &= \text{probability of 1 coalescent event in time t when k=2} \\ &= P(T \leq t), \text{where } T \sim Exp(\mu = \frac{4N}{2(2-1)}) \\ &= \int_0^t \frac{1}{2N} e^{-\frac{x}{2N}} dx = 1 - e^{-\frac{t}{2N}} \end{aligned}$$

- Similarly,

$$\begin{aligned} P_{22}(t) &= \text{prob. of no coalescence in time t when k=2} \\ &= P(T > t) \\ &= \int_t^\infty \frac{1}{2N} e^{-\frac{x}{2N}} dx = e^{-\frac{t}{2N}} \end{aligned}$$

The Coalescent Model Along a Species Tree

- So far, we've considered the coalescent process within a single population.
- A phylogenetic tree consists of many populations followed throughout evolutionary time:



The Coalescent Model Along a Species Tree

- Goal is to apply coalescent model across the phylogeny. The basic assumption is that events that occur in one population are independent of what happens in other populations within the phylogeny.
- More specifically, given the number of lineages entering and leaving a population, coalescent events within populations are independent of one another.
- It is also important to recall an assumption we "inherit" from our population genetics model: all pairs of lineages are equally likely to coalesce within a population.

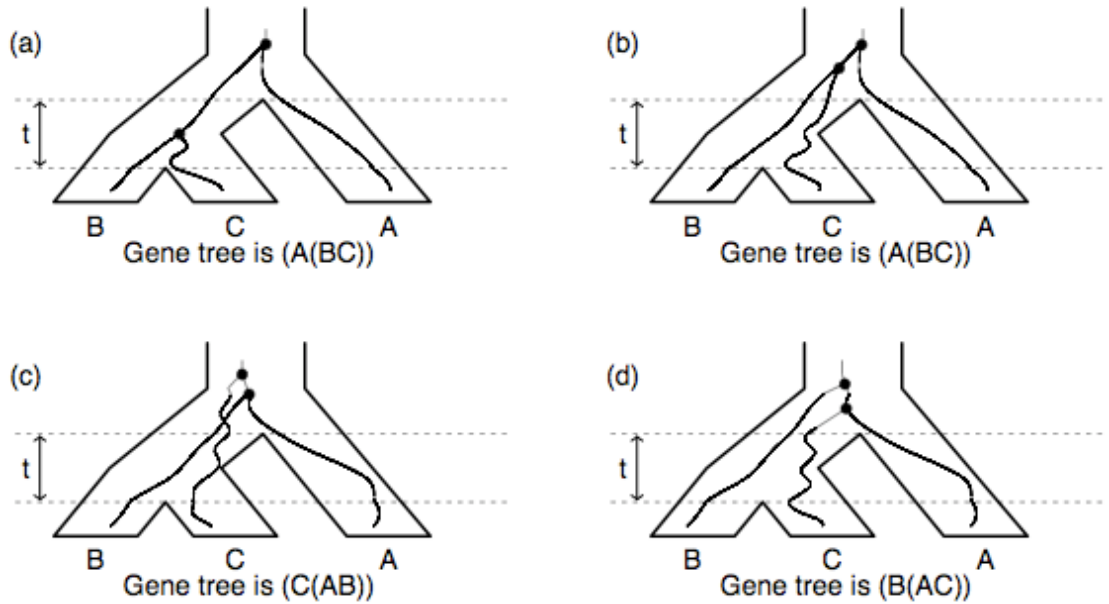
The Coalescent Model Along a Species Tree

- When talking about gene tree distributions, there are two cases of interest:
  - The gene tree topology distribution
  - The joint distribution of topologies and branch lengths
- Start with the simple case of 3 species with 1 lineage sampled in each and look at the gene tree topology distribution

Example: Computation of Gene Tree Topology Probabilities for the 3-taxon Case

Example of gene tree probability computation (for simplicity, let's use coalescent units for our time scale):

(a) Prob $= 1 - e^{-t}$; (b), (c), (d) Prob $= \frac{1}{3}e^{-t}$



Example: Computation of Gene Tree Topology Probabilities for the 3-taxon Case

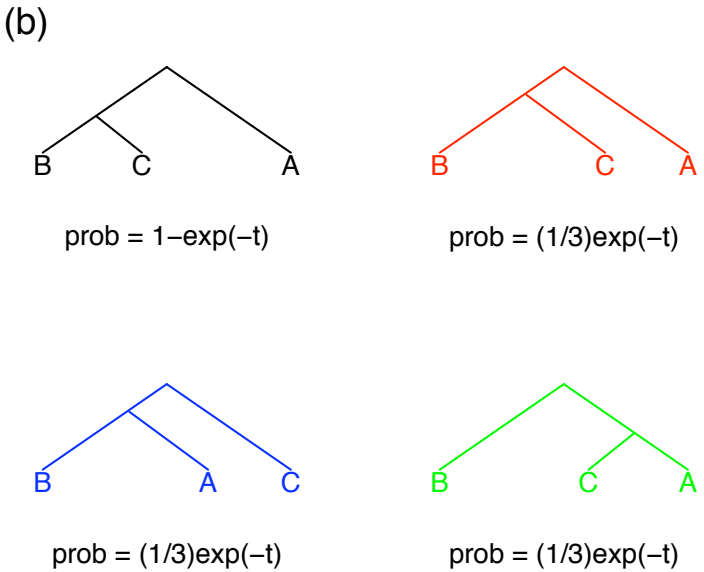
- Thus, we have the following probabilities:
  - Gene tree (A,(B,C)): prob $= 1 - e^{-t} + \frac{1}{3}e^{-t} = 1 - \frac{2}{3}e^{-t}$
  - Gene tree (B,(A,C)): prob $= \frac{1}{3}e^{-t}$
  - Gene tree (C,(A,B)): prob $= \frac{1}{3}e^{-t}$

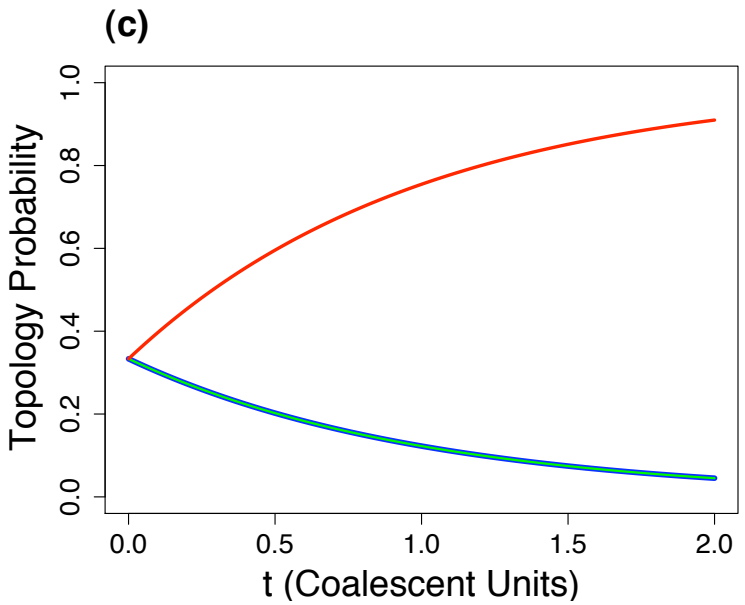
- Note: There are two ways to get the first gene tree. We call these histories.

- The probability associated with a gene tree topology will be the sum over all histories that have that topology.

Example: Computation of Gene Tree Topology Probabilities for the 3-taxon Case

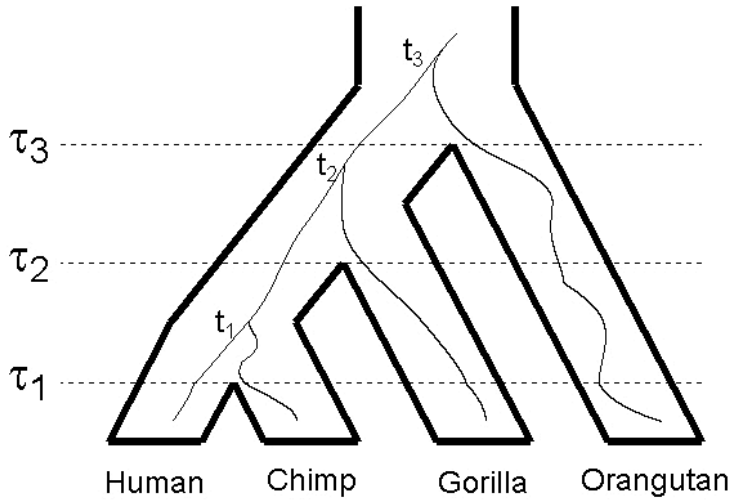
- What are these probabilities like as a function of t, the length of time between speciation events?





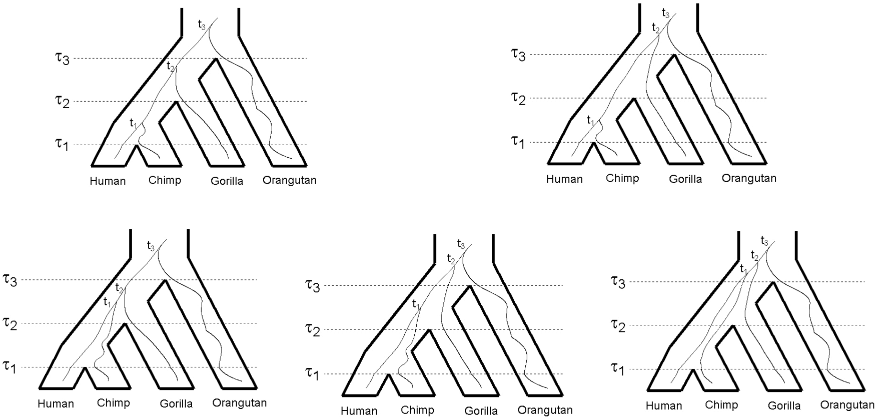
Example: A Slightly Larger Case

- Consider 4 taxa – the human-chimp-gorilla problem



Coalescent Histories for the 4-taxon Example

- There are 5 possibilities for this example:



Computing the Topology Distribution by Enumerating Histories

- In the general case, we have the following:
  The probability of gene tree g given species tree $\mathcal{S}$ is given by

$$P\{G = g|\mathcal{S}\} = \sum_{histories} P\{G = g, history|\mathcal{S}\}$$

Computing the Topology Distribution by Enumerating Histories

- In the general case, we have the following:
  The probability of gene tree g given species tree $\mathcal{S}$ is given by

$$\begin{aligned} P\{G = g|\mathcal{S}\} &= \sum_{histories} P\{G = g, history|\mathcal{S}\} \\ &= \sum_{histories} \prod_{b} w_b P_{u(b),v(b)}(t_b) \end{aligned}$$

Degnan and Salter, *Evolution*, 2005

Computing the Topology Distribution by Enumerating Histories

- The probability of gene tree g given species tree $\mathcal{S}$ is given by

$$\begin{aligned} P\{G = g|\mathcal{S}\} &= \sum_{histories} P\{G = g, history|\mathcal{S}\} \\ &= \sum_{histories} \prod_{b} w_b P_{u(b),v(b)}(t_b) \end{aligned}$$

Number of terms only known in special cases (Rosenberg, 2007)

Computing the Topology Distribution by Enumerating Histories

- The probability of gene tree g given species tree $\mathcal{S}$ is given by

$$\begin{aligned} P\{G = g|\mathcal{S}\} &= \sum_{histories} P\{G = g, history|\mathcal{S}\} \\ &= \sum_{histories} \prod_{b} w_b P_{u(b),v(b)}(t_b) \end{aligned}$$

Multiply probabilities associated with history over internal branches (once the number of lineages entering and leaving a branch is known – which is what is given by the histories – coalescence happens independently along branches)

Computing the Topology Distribution by Enumerating Histories

- The probability of gene tree g given species tree $\mathcal{S}$ is given by

$$\begin{aligned} P\{G = g|\mathcal{S}\} &= \sum_{histories} P\{G = g, history|\mathcal{S}\} \\ &= \sum_{histories} \prod_{b} w_b P_{u(b),v(b)}(t_b) \end{aligned}$$

Probability of getting sequence of coalescent events that is consistent with g

Computing the Topology Distribution by Enumerating Histories

- The probability of gene tree g given species tree $\mathcal{S}$ is given by

$$\begin{aligned} P\{G = g|\mathcal{S}\} &= \sum_{histories} P\{G = g, history|\mathcal{S}\} \\ &= \sum_{histories} \prod_{b} w_b P_{u(b),v(b)}(t_b) \end{aligned}$$

Probability that u lineages coalescent into v in time t_b

Computing the Topology Distribution by Enumerating Histories

- The probability of gene tree g given species tree $\mathcal{S}$ is given by

$$\begin{aligned} P\{G = g|\mathcal{S}\} &= \sum_{histories} P\{G = g, history|\mathcal{S}\} \\ &= \sum_{histories} \prod_{b} w_b P_{u(b),v(b)}(t_b) \end{aligned}$$

Length of branch b

Computing the Topology Distribution by Enumerating Histories

TABLE 3. The number of valid coalescent histories when the gene tree and species tree have the same topology. The number of histories is also the number of terms in the outer sum in equation (12).

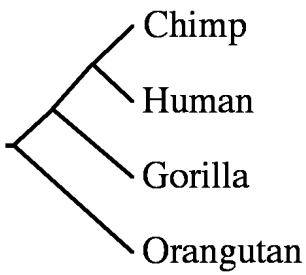
	Number of histories		
Taxa	Asymmetric trees	Symmetric trees	Number of topologies
4	5	4	15
5	14	10	105
6	42	25	945
7	132	65	10,395
8	429	169	135,135
9	1430	481	2,027,025
10	4862	1369	34,459,425
12	58,786	11,236	13,749,310,575
16	9,694,845	1,020,100	6.190×10^{15}
20	1,767,263,190	100,360,324	8.201×10^{21}

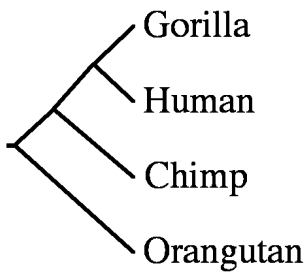
Degnan and Salter, *Evolution*, 2005

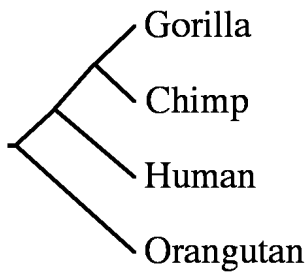
Applications of the Topology Distribution - Example 1

- Motivation: Paper by Ebersberger et al. 2007. *Mol. Biol. Evol.* 24:2266-2276
- Examined 23,210 distinct alignments for 5 primate taxa: Human, Chimp, Gorilla, Orangutan, Rhesus
- Looked at distribution of gene trees among these taxa - observed strongly supported incongruence only among the Human-Chimp-Gorilla clade.

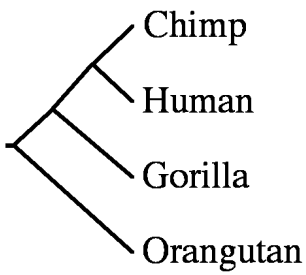
Applications of the Topology Distribution - Example 1



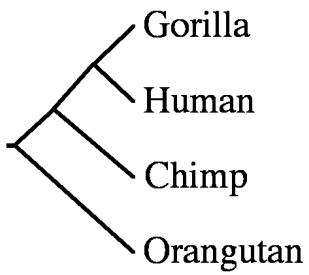




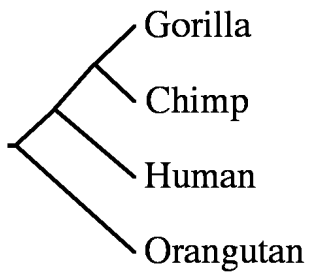
Applications of the Topology Distribution - Example 1



76.6%



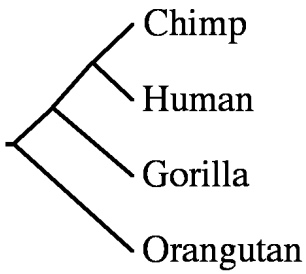
11.4%

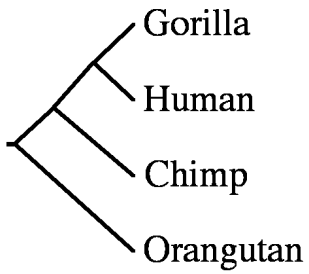


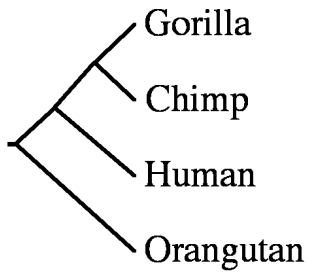
11.5%

Observed proportions of each gene tree among ML phylogenies

Applications of the Topology Distribution - Example 1







76.6%	11.4%	11.5%
79.1%	9.9%	9.9%

Observed proportions of each gene tree among ML phylogenies

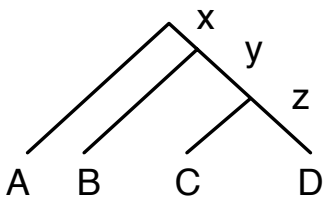
Predicted proportions using parameters from Rannala & Yang, 2003.

Applications of the Topology Distribution - Example 2

- In the previous example, one topology is clearly preferred
- Must the distribution always look this way?
- Examine entire distribution when the number of taxa is small

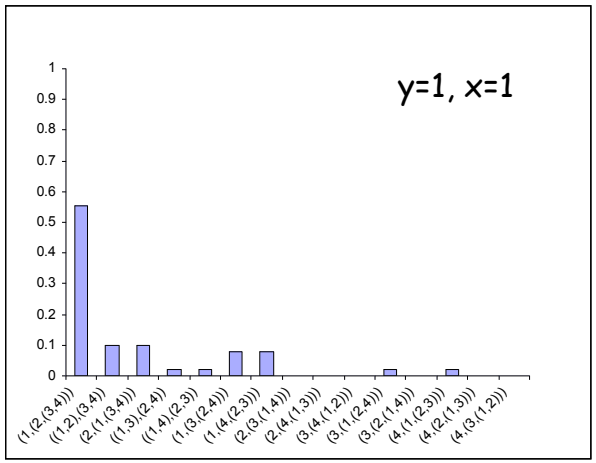
Applications of the Topology Distribution - Example 2

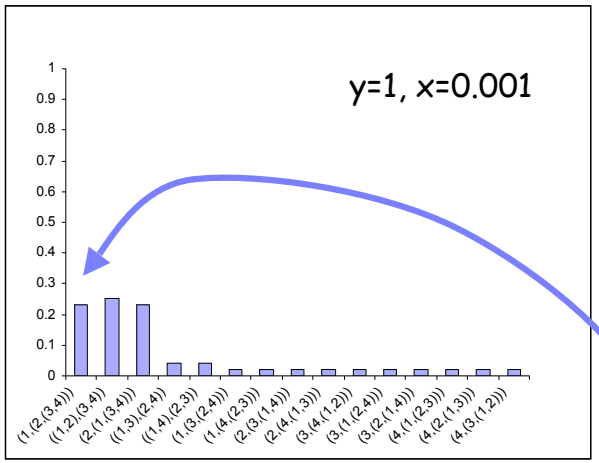
- Consider 4 taxa: A, B, C, and D
- Species tree:

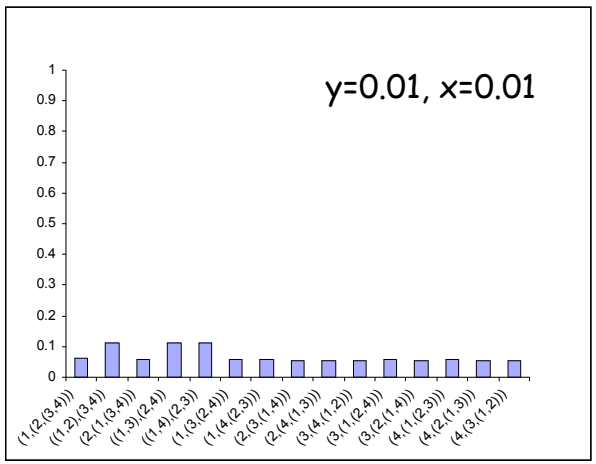


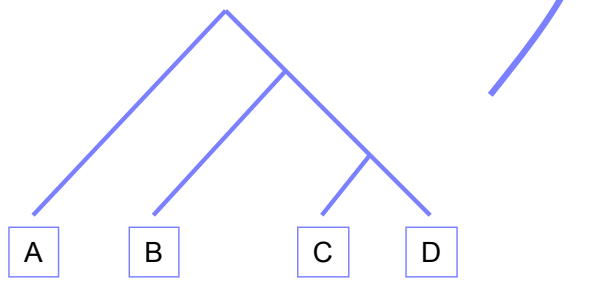
- Look at probabilities of all 15 tree topologies for values of x, y, and z

Applications of the Topology Distribution - Example 2

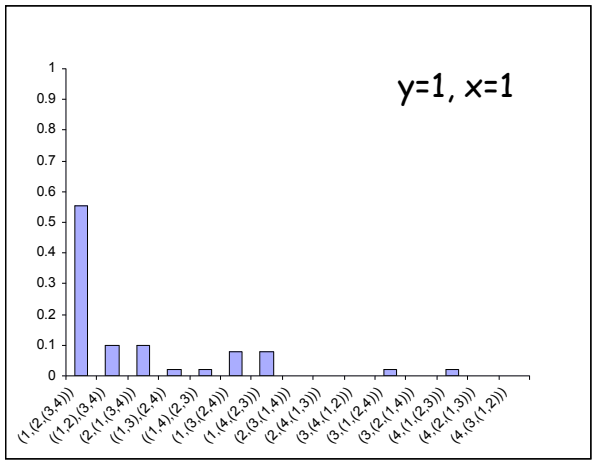


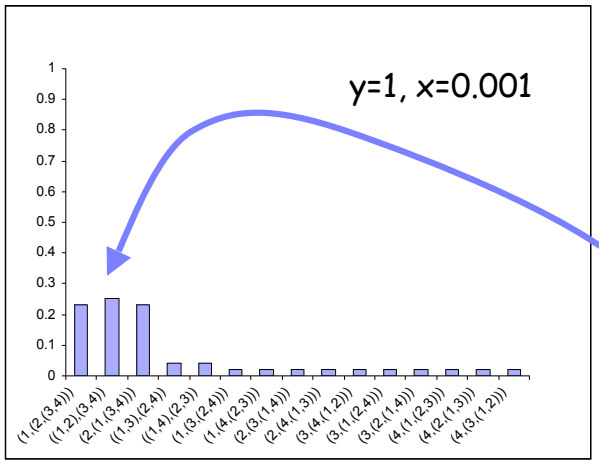


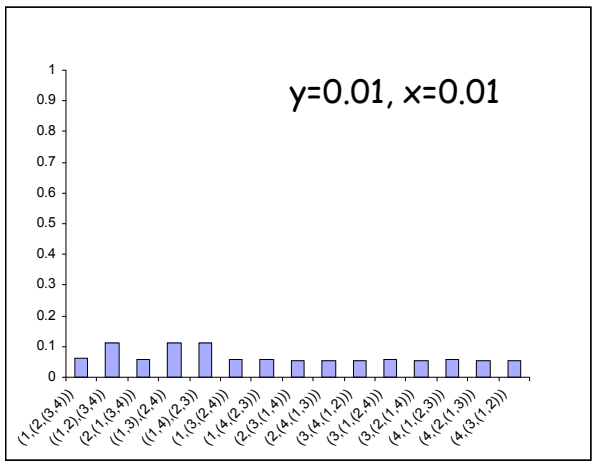


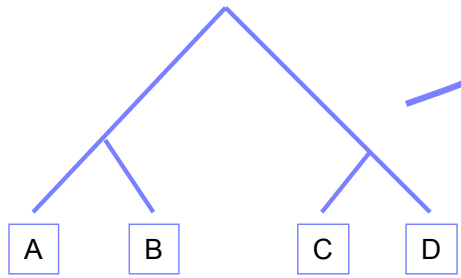


Applications of the Topology Distribution - Example 2

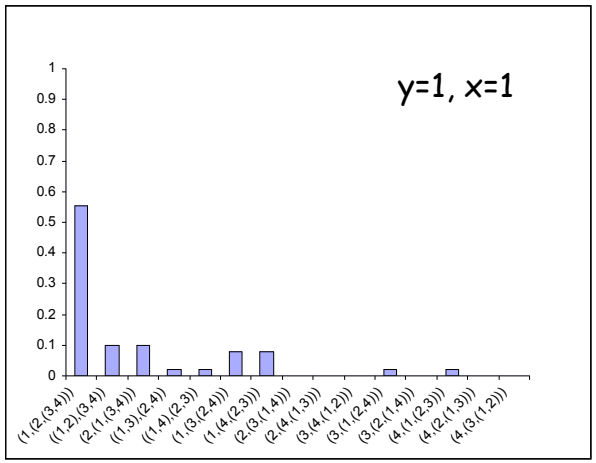


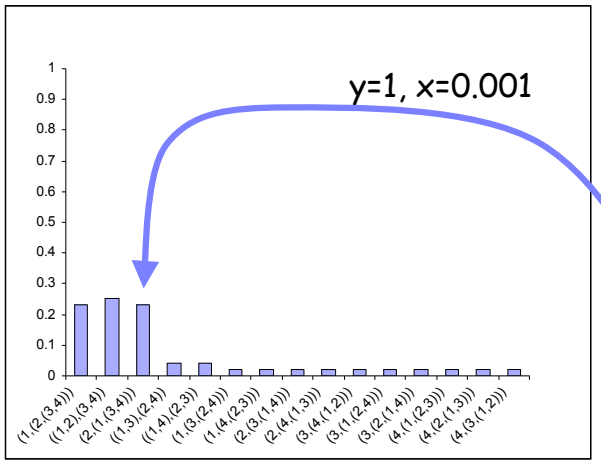


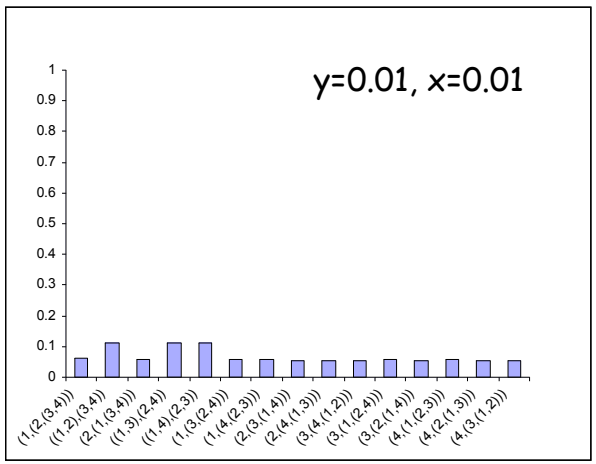


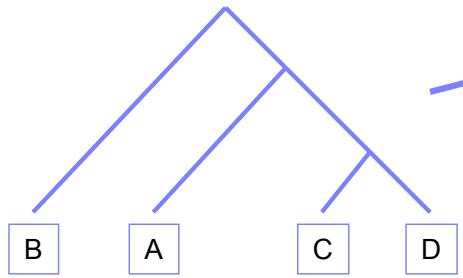


Applications of the Topology Distribution - Example 2

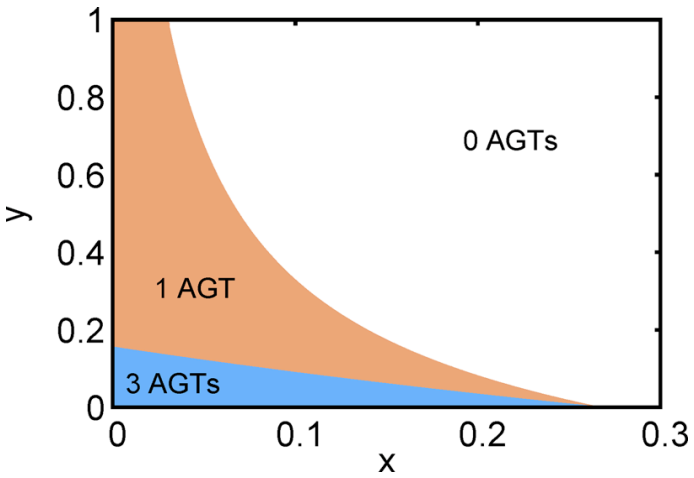








Applications of the Topology Distribution - Example 2



- The existence of anomalous gene trees has implications for the inference of species trees

Degnan and Rosenberg, *PLoS Genetics*, 2006

Rosenberg and Tao, *Systematic Biology*, 2008

Applications of the Topology Distribution - Example 3

- What about mutation? How does this affect data analysis?
- The coalescent gives a model for determining gene tree probabilities for each gene.
- View DNA sequence data as the result of a two-stage process:
  - Coalescent process generates a gene tree topology.
  - Given this gene tree topology, DNA sequences evolve along the tree.

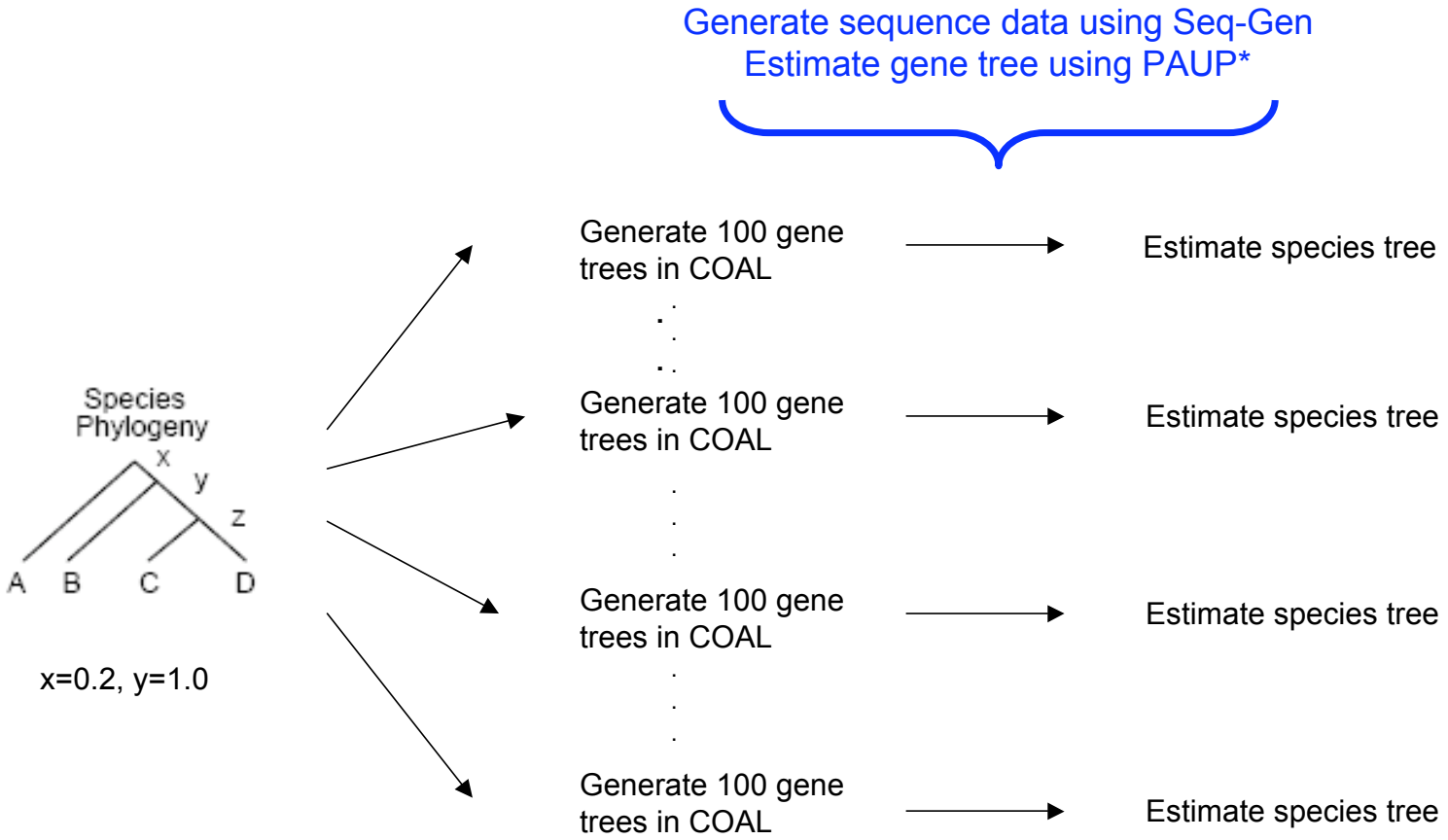
Applications of the Topology Distribution - Example 3

- Given this model, how should inference be carried out?

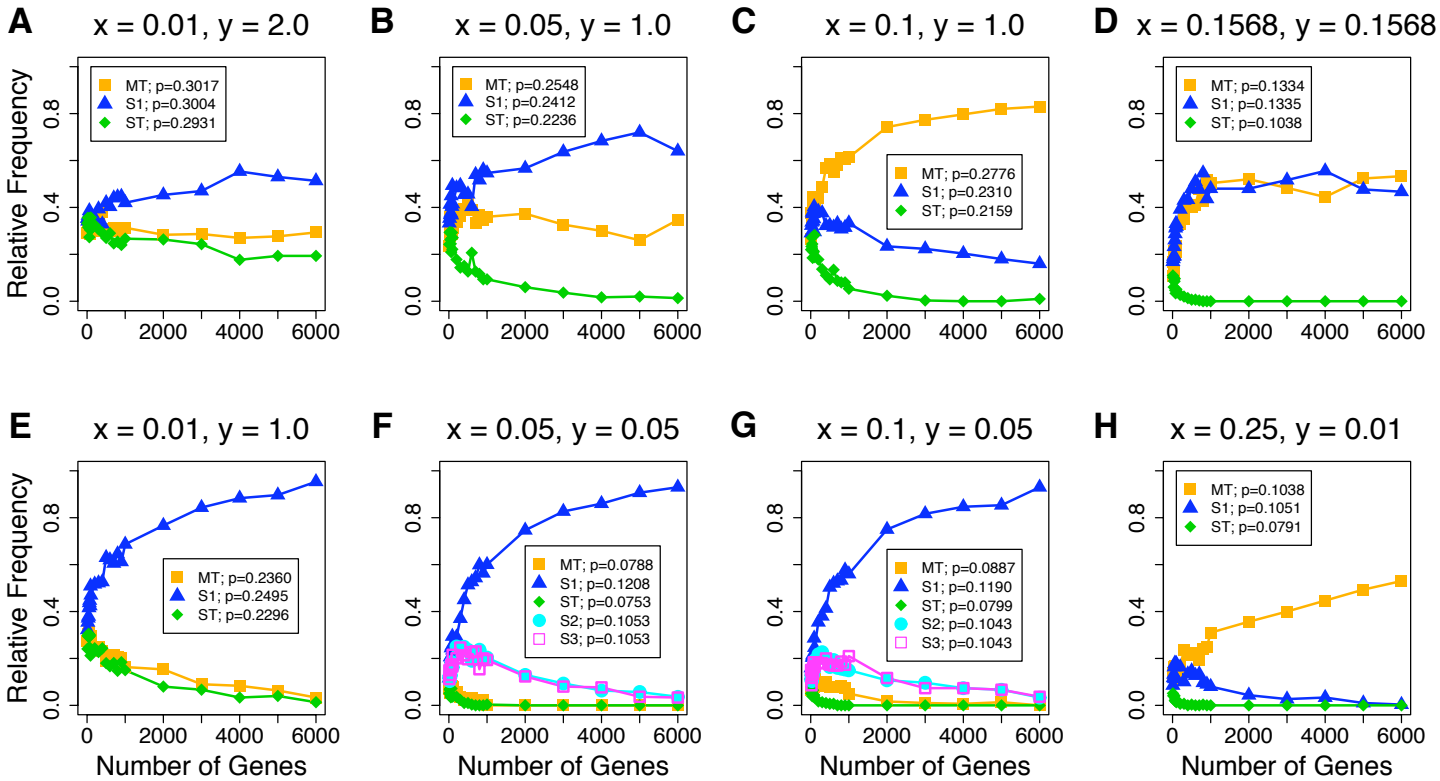
Applications of the Topology Distribution - Example 3

- Given this model, how should inference be carried out?
- Hypothesis: As more data (genes) are added, the process of estimating species trees from concatenated data can be statistically inconsistent
- May fail to converge to any single tree topology if there are many equally likely trees.
- May converge to the wrong tree when a gene tree that is topologically incongruent with the species tree has the highest probability.

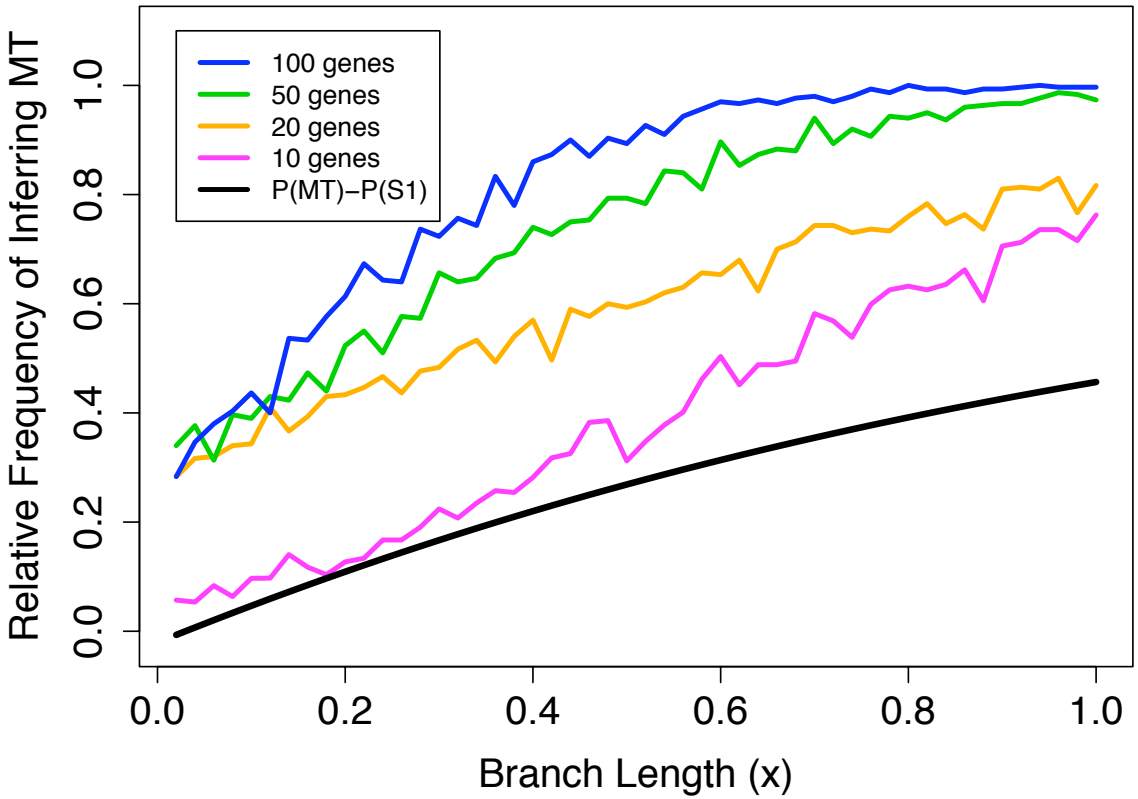
Applications of the Topology Distribution - Example 3



Applications of the Topology Distribution - Example 3
Simulation Study 1



Applications of the Topology Distribution - Example 3
Simulation Study 2



Applications of the Topology Distribution - Example 3

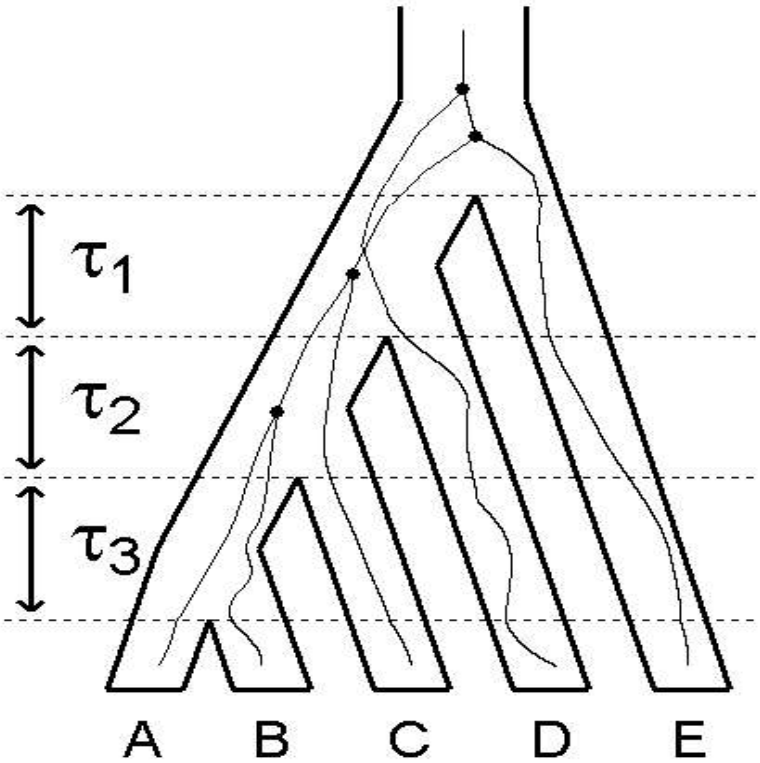
- Performance of the Concatenation Approach:
  - Can be statistically inconsistent when branch lengths in the species phylogeny are sufficiently small
  - May perform poorly even when branch lengths are only moderately short

Applications of the Topology Distribution - Example 3

- Performance of the Concatenation Approach:
  - Can be statistically inconsistent when branch lengths in the species phylogeny are sufficiently small
  - May perform poorly even when branch lengths are only moderately short
- What should we do? Need to design inference methods that incorporate the coalescent process.
  Dennis's lecture next week

Joint Density of Gene Tree Topology and Branch Lengths – An Example

Rannala and Yang 2003



$$2exp\{-2\tfrac{3(2)}{2}t_{ABCE}\}exp\{-2(t_{ABCE}-t_{ABCDE})\}$$

$$2exp\{-2\tfrac{3(2)}{2}t_{ABC}\}exp\{-2(\tau_1-t_{ABC})\}$$

$$2exp\{-2\tfrac{3(2)}{2}t_{AB}\}exp\{-2(\tau_2-t_{AB})\}$$

$$exp\{-2\tau_3\}$$

Gene Tree Branch Length Distributions

- We now have the following distributions
  - $p(g|\mathcal{S})$
  - $f(g, \mathbf{t}|\mathcal{S})$
- We can thus, in theory, get the distributions of gene tree branches by simply manipulating these quantities:

$$f(\mathbf{t}|G, \mathcal{S}) = \frac{f(g, \mathbf{t}|\mathcal{S})}{p(g|\mathcal{S})}$$

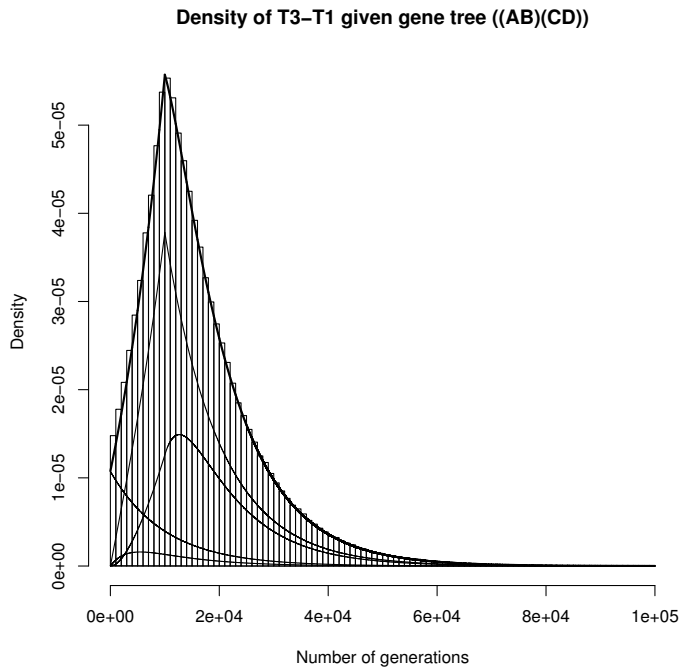
- Integrating out branches which aren't of interest gives joint or marginal distributions
- Can even examine correlations between branch lengths

Gene Tree Branch Length Distributions

- Complication: Region of integration will change for each history within a given gene tree
- Branch length densities are then a mixture over histories
- For the case of four taxa, James Degnan and I have worked out all joint and marginal distributions
- Simulate data and compare theoretical distribution to observed distribution
- Correlations are also well-approximated by simulation

Applications of Branch Length Distributions - Example 1

- Simulate 1,000,000 gene trees from species tree ((A:1.0,B:1.0):1.0,(C:1.5,D:1.5):0.5);
- Of these, 449,599 had the same topology as the species tree
- Compare observed distribution of branch length connecting (A,B) to root node to true distribution



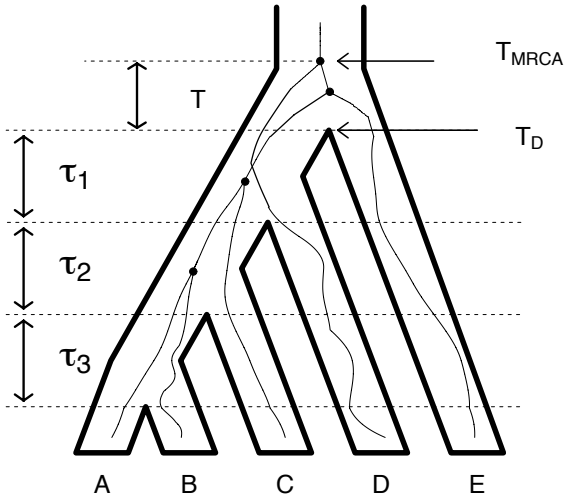
- Good fit between observed and true distributions

Applications of Branch Length Distributions - Example 2

- Estimation of speciation times using information in gene trees is often desirable
- Under the coalescent model (with no gene flow following speciation), it must the case that gene divergence times pre-date speciation times

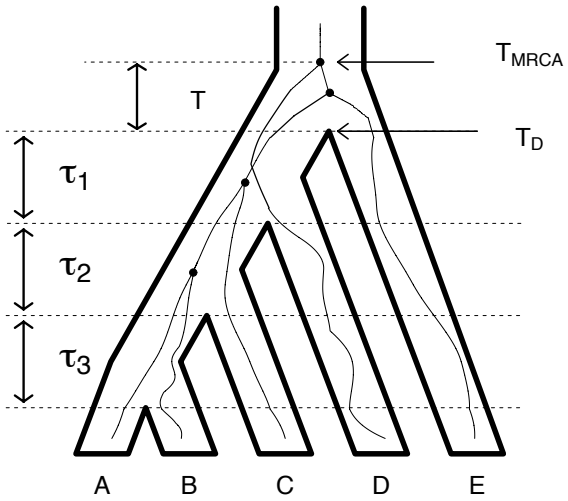
Applications of Branch Length Distributions - Example 2

- Estimation of speciation times using information in gene trees is often desirable
- Under the coalescent model (with no gene flow following speciation), it must the case that gene divergence times pre-date speciation times



Applications of Branch Length Distributions - Example 2

- Estimation of speciation times using information in gene trees is often desirable
- Under the coalescent model (with no gene flow following speciation), it must the case that gene divergence times pre-date speciation times



- What is the distribution of this difference?
- How does it depend on species tree shape (e.g., symmetry) and species tree branch lengths?

Applications of Branch Length Distributions - Example 2

- When considering the distribution of the MRCA of a sample of lineages is of interest, we can simplify computation of the distribution

Applications of Branch Length Distributions - Example 2

- When considering the distribution of the MRCA of a sample of lineages is of interest, we can simplify computation of the distribution
- Let T be the distribution of the difference between the speciation time and the time of the MRCA of all lineages. Note that

$$f_{T|\mathcal{S}}(t) = \sum_{n=2}^{k} Pr(L = n|\mathcal{S}) P_{n1}(t)$$

where L is the random number of lineages available to coalesce above the root of the species tree

Applications of Branch Length Distributions - Example 2

- When considering the distribution of the MRCA of a sample of lineages is of interest, we can simplify computation of the distribution
- Let T be the distribution of the difference between the speciation time and the time of the MRCA of all lineages. Note that

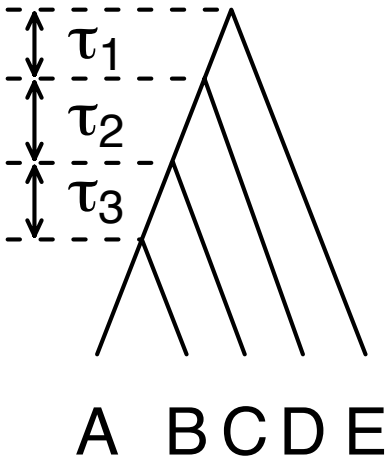
$$f_{T|\mathcal{S}}(t) = \sum_{n=2}^{k} Pr(L = n|\mathcal{S}) P_{n1}(t)$$

  where L is the random number of lineages available to coalesce above the root of the species tree
- $Pr(L = n|\mathcal{S})$ can be computed recursively in a peeling-type algorithm

Efromovich and Kubatko, SAGMB, 2008

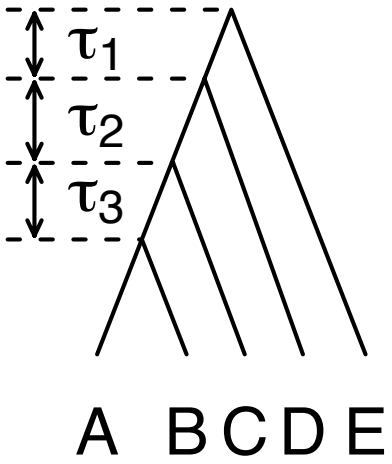
Applications of Branch Length Distributions - Example 2

- A particular example

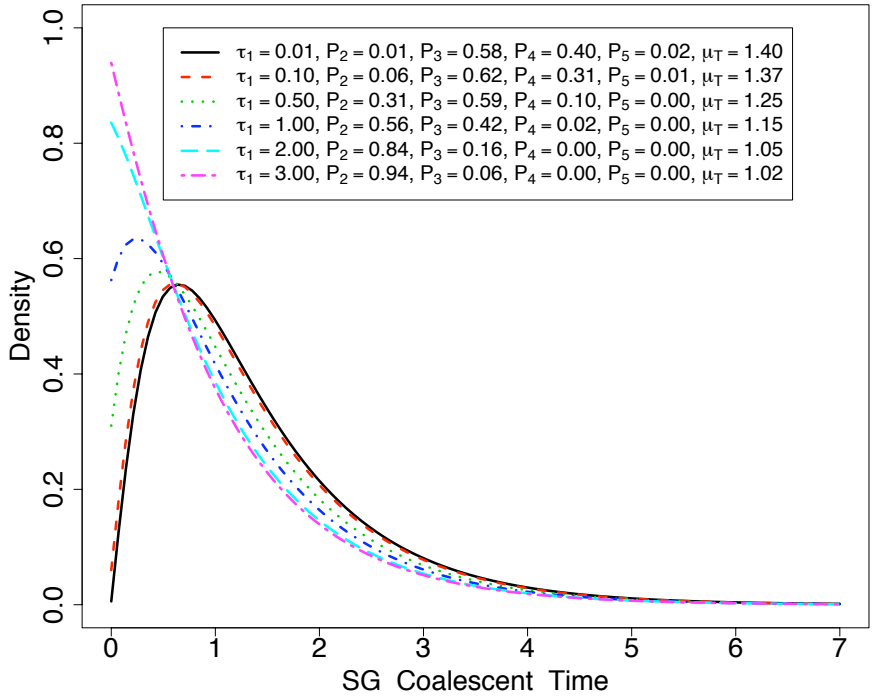


Applications of Branch Length Distributions - Example 2

- A particular example



- When $\tau_2 = \tau_3 = 1.0$, we have the following:



Applications of Branch Length Distributions - Example 2

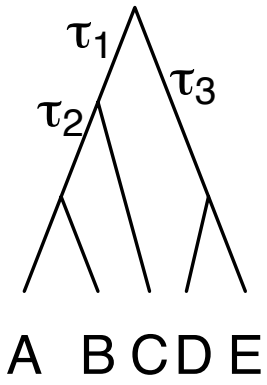
- What can we conclude from this?
  - Shorter branches lead to more potential for incomplete lineage sorting, which results in longer times to the MRCA
  - This effect will be most pronounced for branches that are close to the root of the tree

Applications of Branch Length Distributions - Example 2

- What can we conclude from this?
  - Shorter branches lead to more potential for incomplete lineage sorting, which results in longer times to the MRCA
  - This effect will be most pronounced for branches that are close to the root of the tree
- What about symmetry of the tree?

Applications of Branch Length Distributions - Example 2

- Consider a more symmetric tree:



- Two internal branches adjacent to the root node
- Suggests more possibility of incomplete lineage sorting $\Longrightarrow$ longer time to MRCA

Coalescent Theory

- We now have the main ideas of the coalescent model and how to apply it to a phylogeny.
  - But there are many things we haven't discussed: migration, recombination, etc.
- Next week: How can we take this model and use it to infer a species-level phylogeny?
- Thursday's lab: Using the program COAL to compute gene tree topology probabilities.