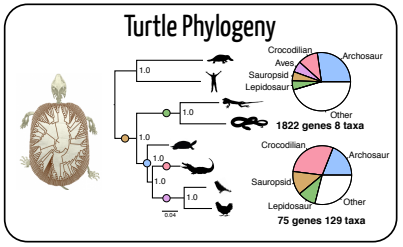
Species Tree Inference

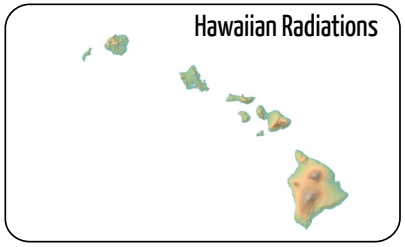
2014 Bodega Bay Applied Phylogenetics Workshop

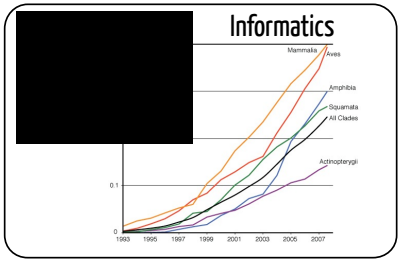
Bob Thomson
thomsonr@hawaii.edu
thomsonlab.org

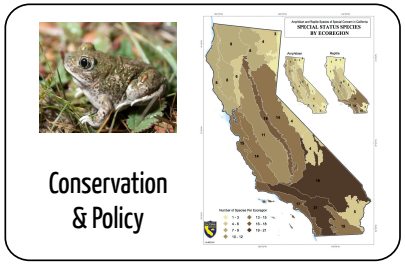
UNIVERSITY of HAWAI'I MĀNOA

Research interests







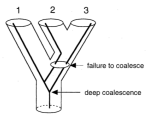


Overview

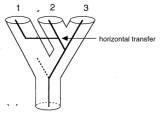
- Think a bit about phylogenetic reconstruction
- Do our simplifications cause problems?
- A few cases where they might, and how we might deal with those issues when they arise.

Sources of gene tree variation

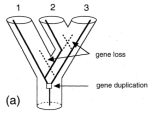
- Incomplete coalescence



- Horizontal transfer

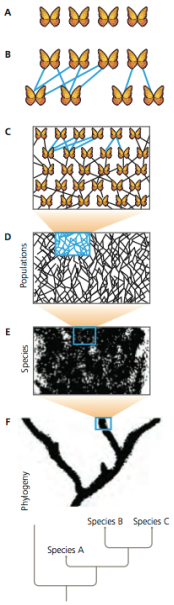


- Gene duplication



Anatomy of a tree

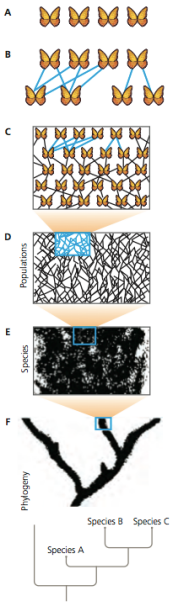
- What are phylogenetic lineages?
  - Each species lineage implicitly contains populations of reproducing populations
  - Phylogenies among species are simplifying this process



5

Anatomy of a tree

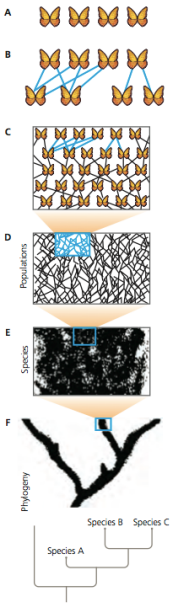
- Let's say we want to infer a phylogeny of these 3 butterfly species
  - We collect data for an individual from each species and infer a phylogeny



6

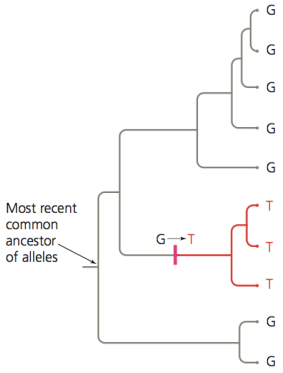
Anatomy of a tree

- Let's say we want to infer a phylogeny of these 3 butterfly species
  - We collect data for an individual from each species and infer a phylogeny
  - Implicitly, we're saying that the evolutionary relationships among those three individuals match the evolutionary relationships among the three species
  - Can this cause problems?



7

Molecular Phylogenetics



- This is a simplification
  - The G->T substitution is a population genetic process
  - i.e., a single mutation occurred in one individual in an ancestral population. It then increased in frequency until it became fixed in the whole species.

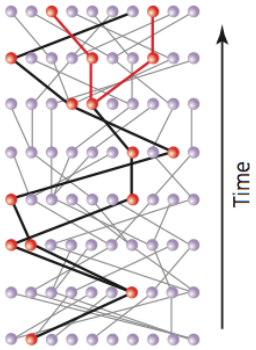
8

Molecular Phylogenetics

- We need to connect our simplified phylogenies of species to trees of individuals
  - We can build a model for this
  - Will start with a case involving only a single species
    - The coalescent
  - Then extend to multiple species
    - The multispecies coalescent

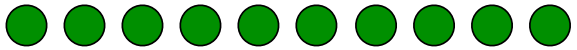
The coalescent model

- Imagine a single species made up of N diploid individuals (2n total alleles)
  - Let's think about the relationships between all of those alleles
    - Here alleles simply refer to physical copies of a particular locus, not distinct forms of that locus



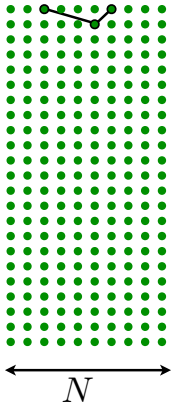
The coalescent model

- How many generations ago did these alleles last share a common ancestor?



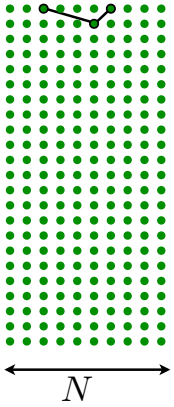
- We can model this in a very simple way...

The coalescent model



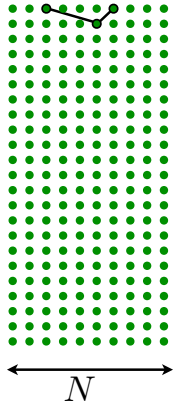
- The probability of 2 lineages coalescing is

The coalescent model



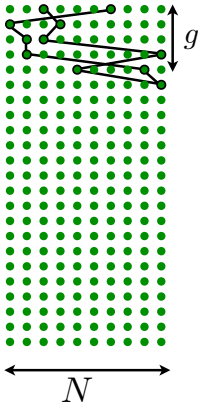
- The probability of 2 lineages coalescing is the probability of them choosing the same ancestor: $\frac{1}{N}$

The coalescent model



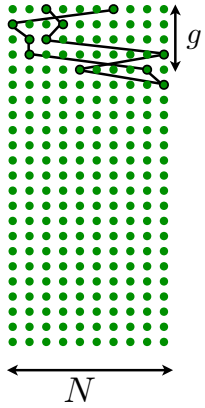
- The probability of 2 lineages coalescing is the probability of them choosing the same ancestor: $\frac{1}{N}$
- Because of this, the expected time until coalescence is simply N

The coalescent model



- Probability that coalescence occurs g+1 generations back:

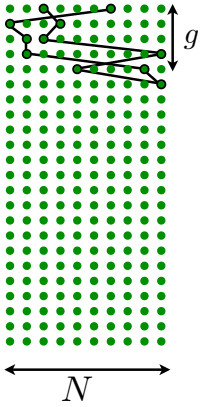
The coalescent model



- Probability that coalescence occurs g+1 generations back:
  - Probability of no coalescence for g generations

$$(1-\frac{1}{N})\times(1-\frac{1}{N})...=(1-\frac{1}{N})^g$$

The coalescent model

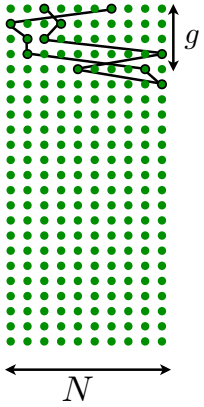


- Probability that coalescence occurs g+1 generations back:
  - Probability of no coalescence for g generations

$$(1-\frac{1}{N})\times(1-\frac{1}{N})...=(1-\frac{1}{N})^g$$

  - followed by coalescence $\frac{1}{N}$

The coalescent model



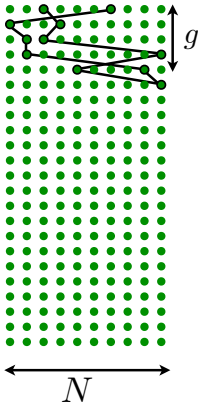
- Probability that coalescence occurs g+1 generations back:
  - Probability of no coalescence for g generations

$$(1-\frac{1}{N})\times(1-\frac{1}{N})...=(1-\frac{1}{N})^g$$

  - followed by coalescence $\frac{1}{N}$

$$=\frac{1}{N}(1-\frac{1}{N})^g$$

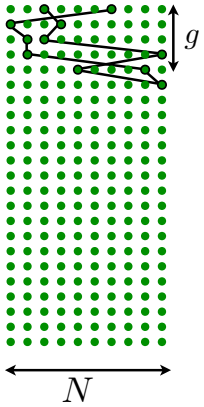
The coalescent model



$$=\frac{1}{N}(1-\frac{1}{N})^g$$

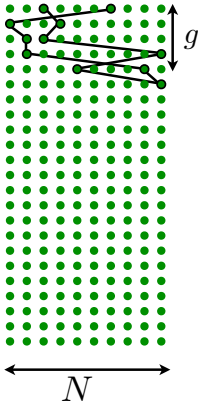
- This is the geometric distribution
  - Describes the time of the first success for independent trials with probability of success p and probability of failure (1-p)
  - Rate = p or 1/N
  - Mean = 1/p or N

The coalescent model



- Probability of coalescence event (or success rate) among n sampled lineages is

The coalescent model



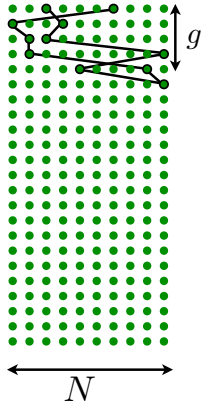
- Probability of coalescence event (or success rate) among n sampled lineages is

$$\frac{\binom{n}{2}}{N}$$

- n choose 2 accounts for the variety of ways that coalescence can occur

$$\frac{n!}{2!(n-2)!}$$

The coalescent model



- Probability of coalescence event (or success rate) among n sampled lineages is

$$\frac{\binom{n}{2}}{N}$$

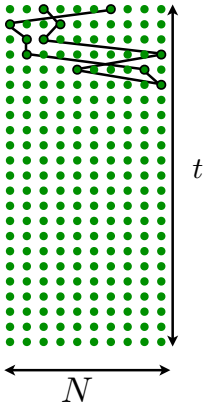
- n choose 2 accounts for the variety of ways that coalescence can occur

$$\frac{n!}{2!(n-2)!}$$

- Probability of event g+1 generations back:

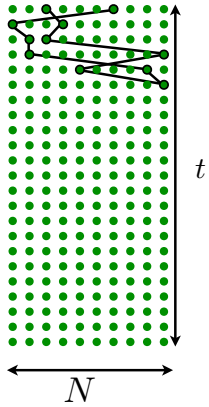
$$\frac{\binom{n}{2}}{N}(1-\frac{\binom{n}{2}}{N})^g$$

The coalescent model



- Geometric distribution is a discrete time distribution

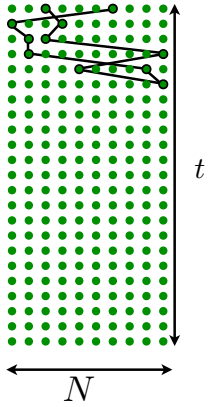
The coalescent model



- Geometric distribution is a discrete time distribution
- Continuous time version is the exponential distribution

$$\lambda e^{-\lambda t}$$

The coalescent model



- Geometric distribution is a discrete time distribution
- Continuous time version is the exponential distribution

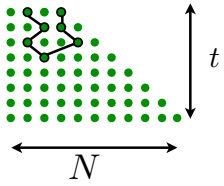
$$\lambda e^{-\lambda t}$$

- As N goes to infinity, the coalescent process converges to a continuous time markov process with instantaneous rate of coalescence:

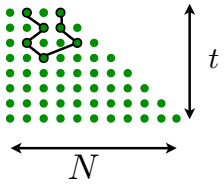
$$\lambda = \frac{\binom{n}{2}}{N}$$

The coalescent model

- We've been assuming constant population size



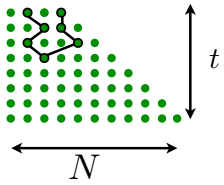
The coalescent model



- We've been assuming constant population size
- Instead of N, we can specify a function that describes a changing population size through time

$$N \rightarrow N(t)$$

The coalescent model



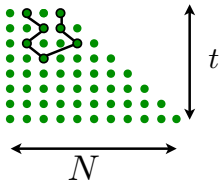
- We've been assuming constant population size
- Instead of N, we can specify a function that describes a changing population size through time

$$N \rightarrow N(t)$$

- Our instantaneous rate of coalescence is a function of N, so we need to integrate the rate of coalescence across the function for N

$$\frac{\binom{n}{2}}{N} e^{-\frac{\binom{n}{2}}{N} t} \rightarrow \frac{\binom{n}{2}}{N(t)} \exp\left(- \int_0^t \frac{\binom{n}{2}}{N(t)} dt \right)$$

The coalescent model

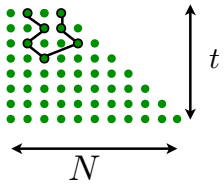


- We have a nice function to calculate the probability of one coalescent event occurring at time t, given a demographic function of t:

$$\mathrm{P}(t) = \frac{\binom{n}{2}}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(t)} dt\right)$$

29

The coalescent model



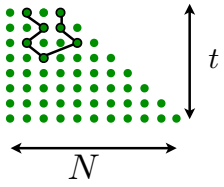
- We have a nice function to calculate the probability of one coalescent event occurring at time t, given a demographic function of t:

$$\mathrm{P}(t) = \frac{\binom{n}{2}}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(t)} dt\right)$$

- What is the probability of <u>all</u> coalescent events observed in a sample?

30

The coalescent model



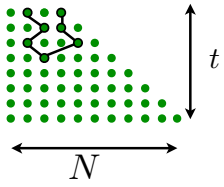
- We have a nice function to calculate the probability of one coalescent event occurring at time t, given a demographic function of t:

$$\mathrm{P}(t) = \frac{\binom{n}{2}}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(t)} dt\right)$$

- What is the probability of <u>all</u> coalescent events observed in a sample?
  - Given a demographic function and a list of coalescence times $\mathrm{L} = (0, t_n, t_{n-1}, \ldots)$

31

The coalescent model



- We have a nice function to calculate the probability of one coalescent event occurring at time t, given a demographic function of t:

$$\mathrm{P}(t) = \frac{\binom{n}{2}}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(t)} dt\right)$$

- What is the probability of <u>all</u> coalescent events observed in a sample?
  - Given a demographic function and a list of coalescence times $\mathrm{L} = (0, t_n, t_{n-1}, \ldots)$
  - Each event is independent, so take the product

$$\mathrm{P}(L|N(t)) = \prod_{i=2}^{n} \frac{\binom{n}{2}}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(t)} dt\right)$$

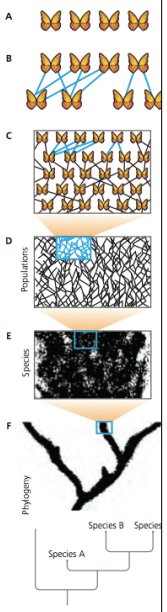
32

The coalescent model

- Starting with first principles, we can derive a model that describes the probability of coalescence histories within a lineage

The coalescent model

- Starting with first principles, we can derive a model that describes the probability of coalescence histories within a lineage
- Connects our simplified idea of a phylogenetic lineage back to the underlying individual sampling



The coalescent model

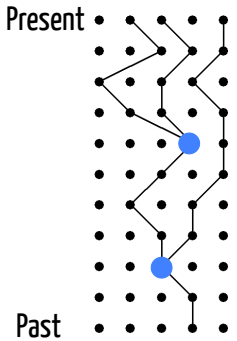
- Starting with first principles, we can derive a model that describes the probability of coalescence histories within a lineage
- Connects our simplified idea of a phylogenetic lineage back to the underlying individual sampling
- We end up with an equation that allows us to calculate the likelihood of an observed set of coalescence times within a lineage

$$\mathrm{P}(L|N(t)) = \prod_{i=2}^{n} \frac{\binom{n}{2}}{N(t)} \exp\left(- \int_0^t \frac{\binom{n}{2}}{N(t)} dt \right)$$

The coalescent model

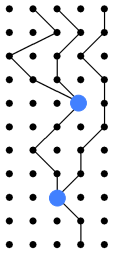
THE IMPORTANT THING TO REMEMBER:

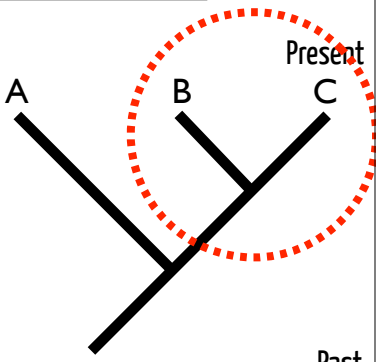
- Probability of coalescence within a population depends on:
  - Population size
  - Number of generations



The coalescent model

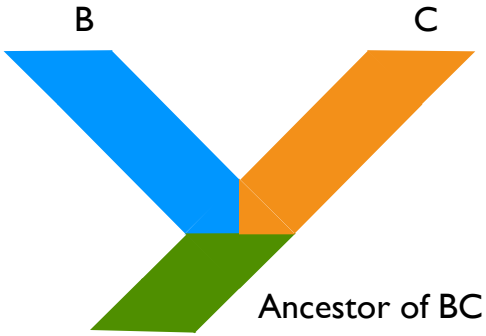
- Connecting this to multiple species
  - A phylogenetic tree of species is simply a collection of these population lineages





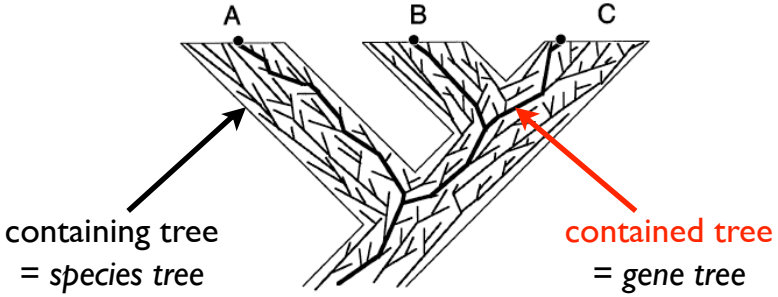
37

The multispecies coalescent



38

The multispecies coalescent

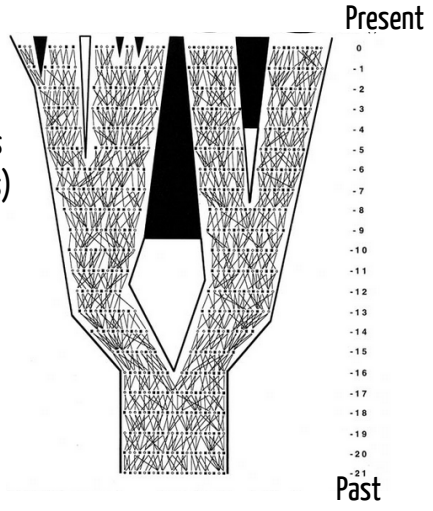


Maddison 1997

39

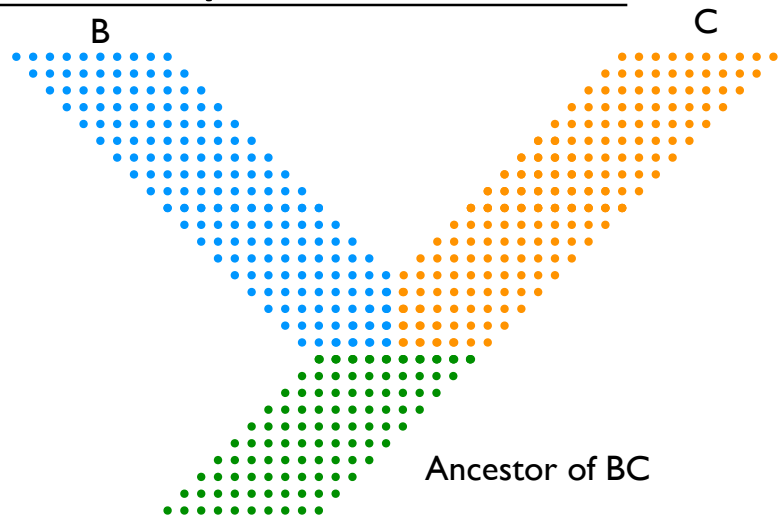
The multispecies coalescent

- Each branch in the species tree has a duration (Number of generations) and a population size
- The multispecies coalescent joins each of these together



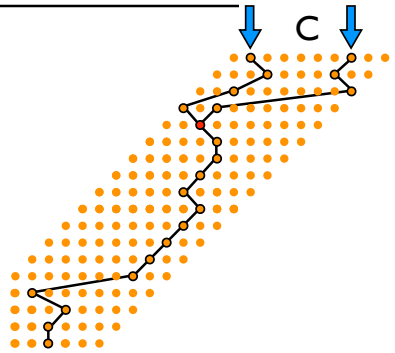
40

The multispecies coalescent



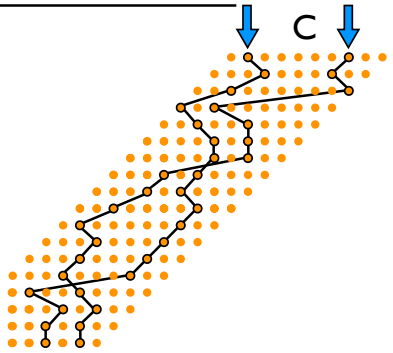
The multispecies coalescent

- If we sample 2 alleles, they have some probability of coalescing before the population 'ends' at the ancestor



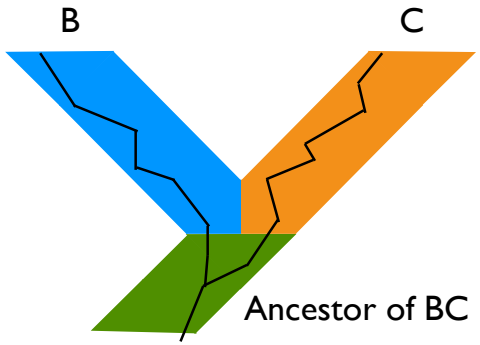
The multispecies coalescent

- If we sample 2 alleles, they have some probability of coalescing before the population 'ends' at the ancestor
- Which means they also have some probability of not coalescing
  - Depends on the population size and the number of generations



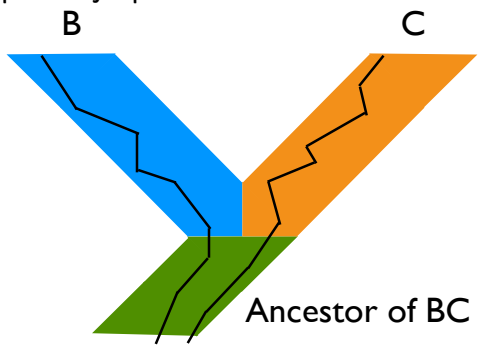
The multispecies coalescent

- Likewise, if we sample one allele from each of two different species, there is some probability that the two alleles will coalesce in the ancestor



The multispecies coalescent

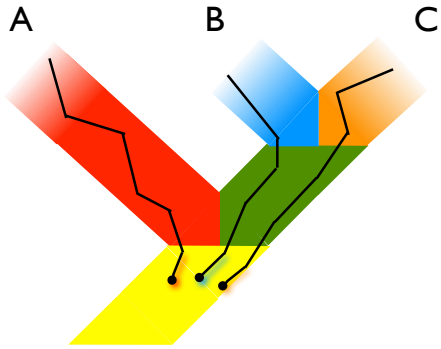
- But there is also some probability that they will not.
  - This is called incomplete coalescence
- What does the probability depend on?



45

The multispecies coalescent

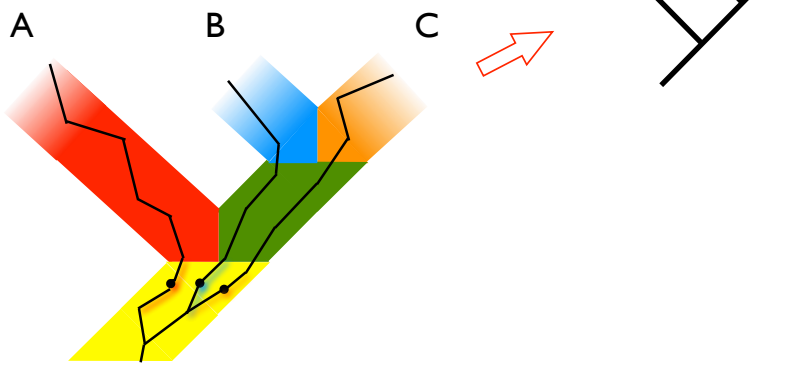
- If they don't coalesce within the ancestor, they move down into the next ancestral population



46

The multispecies coalescent

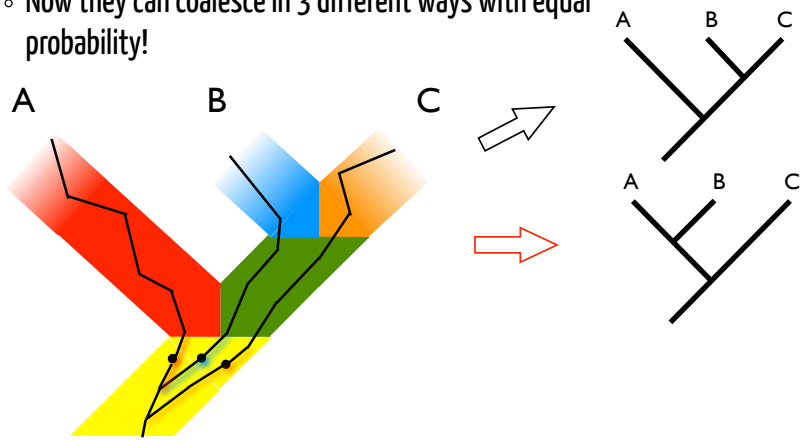
- Now they can coalesce in 3 different ways with equal probability!



47

The multispecies coalescent

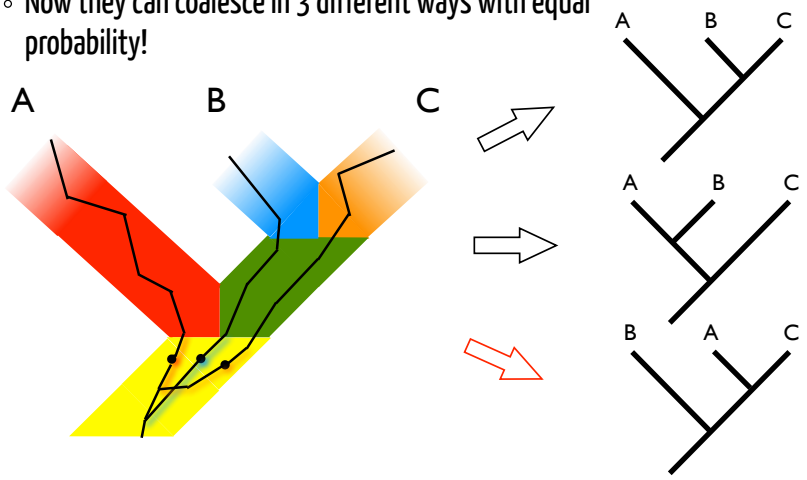
- Now they can coalesce in 3 different ways with equal probability!



48

The multispecies coalescent

- Now they can coalesce in 3 different ways with equal probability!



49

The multispecies coalescent

- Only 1 of the 3 matches the actual species phylogeny
  - So if there is an incomplete coalescence event in the alleles that we sampled, we have a 2/3rds chance of getting the wrong tree
- How do we determine the probability of incomplete coalescence?

50

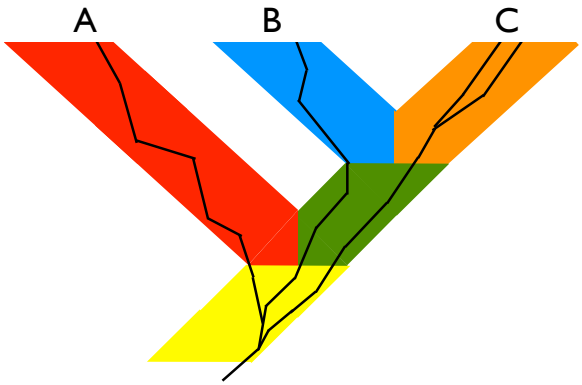
The multispecies coalescent

- The probability depends on the coalescent process that occurs within each lineage
  - We can break up the tree into its component parts
  - Each part has an 'input' and an 'output' number of lineages
    - Inherits the input from what happens above it
    - Output depends on the population size and the duration of the branch

51

The multispecies coalescent

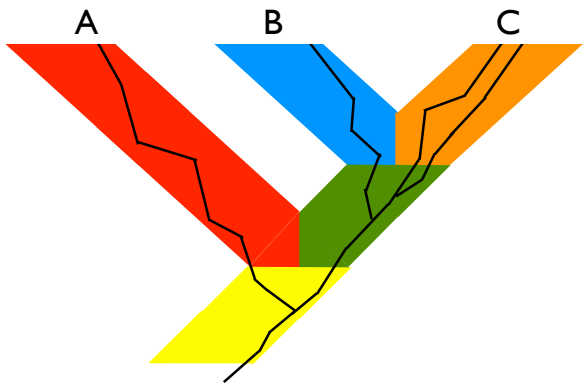
- Gene 1 might look like this



52

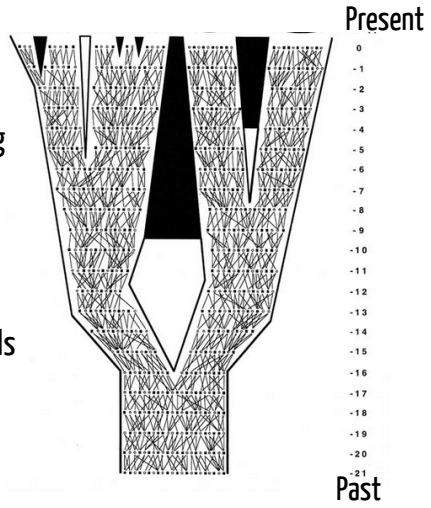
The multispecies coalescent

- While another gene looks like this. Each unlinked gene tree is an independent sample.



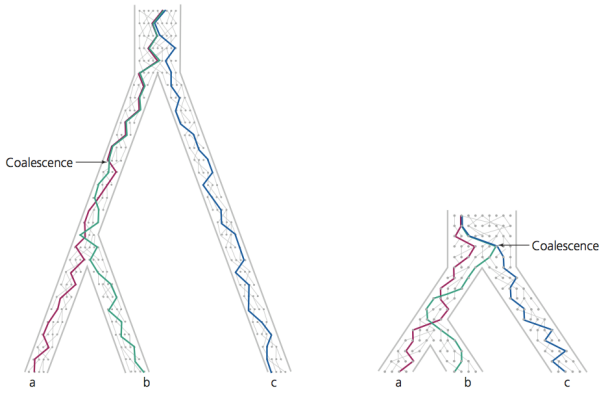
The multispecies coalescent

- So does our phylogenetic sampling cause problems?
  - I.e., does our tree of individuals match our tree of species?
- Answer: Not necessarily, it depends on the population sizes and durations



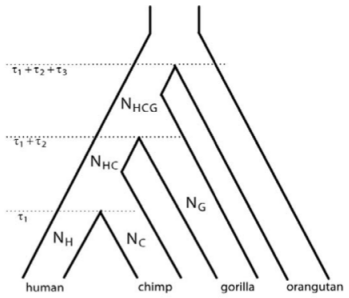
The multispecies coalescent

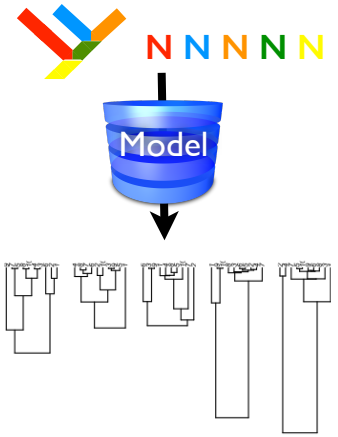
- For some species, gene trees are fantastic estimates of the species tree
- But in other cases they aren't



The multispecies coalescent

- We can simulate this





Empirical Example

OPEN ACCESS Freely available online PLoS GENETICS

Widespread Discordance of Gene Trees with Species Tree in *Drosophila:* Evidence for Incomplete Lineage Sorting

Daniel A. Pollard[1], Venky N. Iyer[2☯], Alan M. Moses[1☯], Michael B. Eisen[1,2,3,4*]

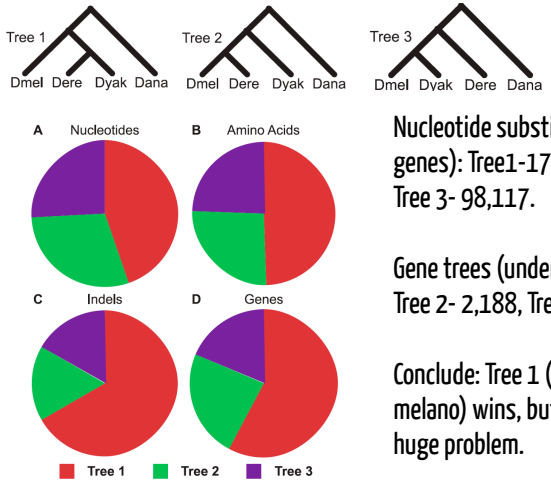
- Genomic data for each of:
  - D. ananassae - outgroup
  - D. melanogaster
  - D. erecta
  - D. yakuba

Pollard et al. 2006

Empirical Example

Pollard et al. 2006

Empirical Example



Nucleotide substitutions (in 9405 genes): Tree1-170,002, Tree 2-112,278, Tree 3- 98,117.

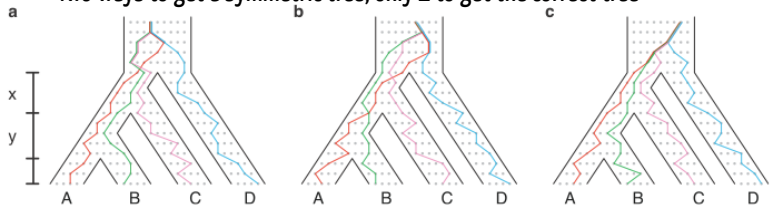
Gene trees (under ML): Tree 1- 5,381, Tree 2- 2,188, Tree 3-1,746

Conclude: Tree 1 ((erecta,yakuba), melano) wins, but lineage sorting is a huge problem.

Pollard et al. 2006

Anomolous gene trees

- There are cases where the wrong tree is more likely to be inferred than the correct tree
- 4 taxon asymmetric tree
  - If branches x and y are short -> lineages join randomly
  - Two ways to get a symmetric tree, only 1 to get the correct tree



Degnan and Rosenberg 2006

Empirical Example

- Human, Chimp, Gorilla
- Look at distribution of genome trees across the entire genome

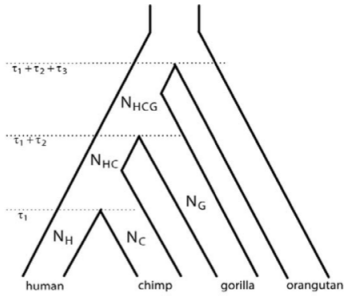
OPEN ACCESS Freely available online PLoS GENETICS

Genomic Relationships and Speciation Times of Human, Chimpanzee, and Gorilla Inferred from a Coalescent Hidden Markov Model

Asger Hobolth[1*], Ole F. Christensen[2], Thomas Mailund[2,3], Mikkel H. Schierup[2]

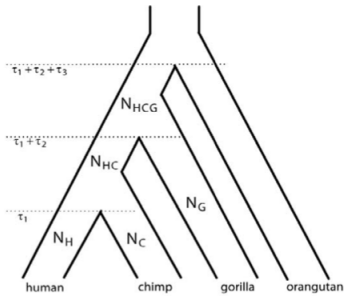
1 Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, United States of America, 2 Bioinformatics Research Center, University of Aarhus, Aarhus, Denmark, 3 Department of Statistics, University of Oxford, Oxford, United Kingdom

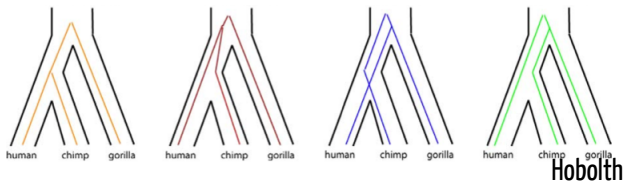
Empirical Example



Hobolth et al. 2007

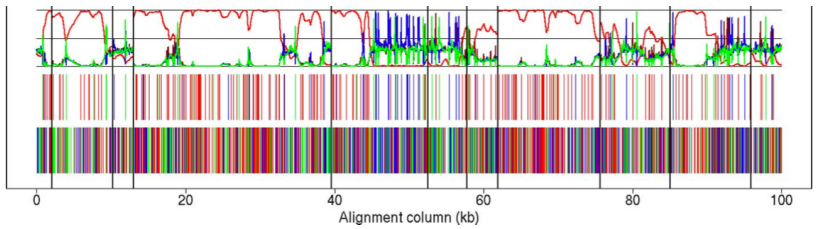
Empirical Example

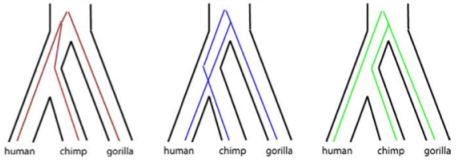




Hobolth et al. 2007

Empirical Example

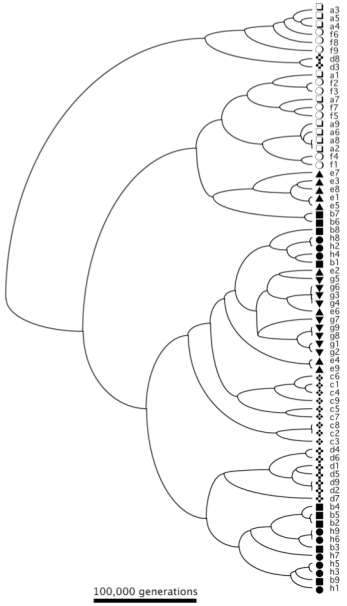




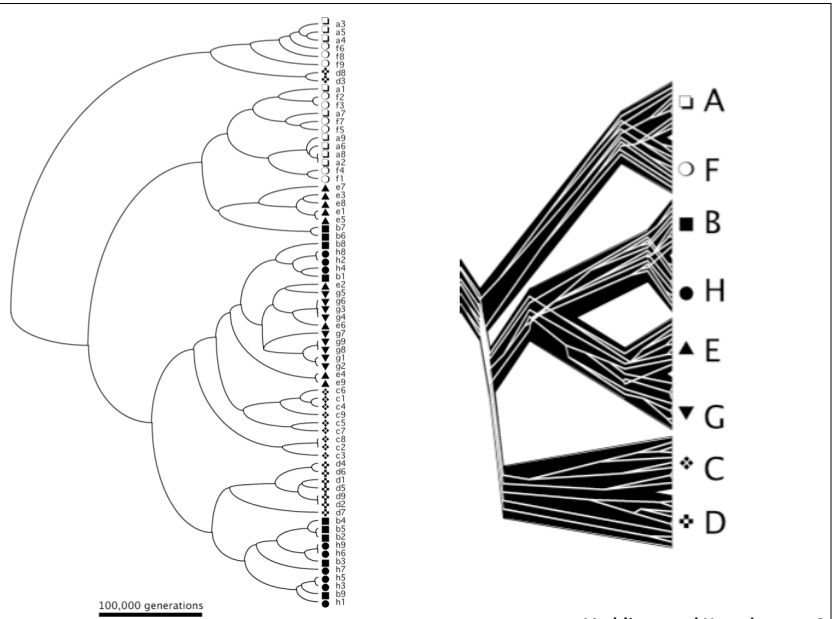
Hobolth et al. 2007

Minimize Deep Coalescence

- given a set of gene trees, find the species tree that minimizes the implied number of deep coalescences (Maddison 1997, Maddison and Knowles 2006)



Maddison and Knowles 2006



Maddison and Knowles 2006

Minimize Deep Coalescence

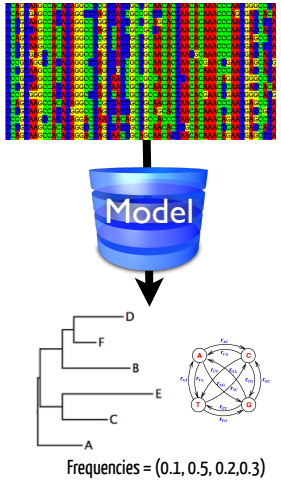
- simple and intuitive
- but ignores important information (branch lengths), no measure of support
- software packages
  - mesquite
  - deep
  - Phylonet
- Doesn't explicitly model the coalescent process, places all probably on single histories

Multispecies coalescent inference

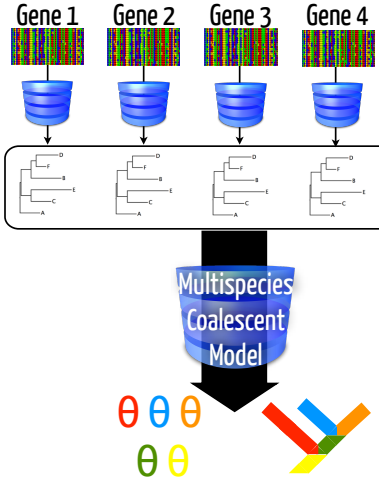
- Perhaps a better solution:
  - We have this nice model, we can use statistical inference to infer species trees from gene trees and/or alignments

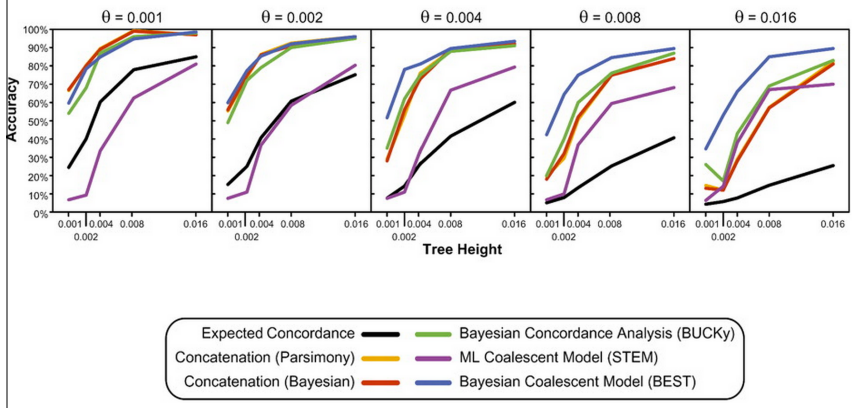
Statistical inference

- What we've been doing:
- Inference under the MC, in the most general case, involves adding another level to this model.



The multispecies coalescent





Leache and Rannala 2011

*BEAST Model

$$P(S|D) = \frac{\prod_{i=1}^{n} P(d_i|g_i)P(g_i|S)P(S)}{P(D)}$$

- 4 components:

Heled and Drummond 2010

*BEAST Model

$$P(S|D) = \frac{\prod_{i=1}^{n} P(d_i|g_i)P(g_i|S)P(S)}{P(D)}$$

- 4 components:
  - $P(d_i|g_i)$- standard likelihood for alignment and gene tree i

Heled and Drummond 2010

*BEAST Model

$$P(S|D) = \frac{\prod_{i=1}^{n} P(d_i|g_i)P(g_i|S)P(S)}{P(D)}$$

- 4 components:
  - $P(d_i|g_i)$- standard likelihood for alignment and gene tree i
  - $P(g_i|S)$- coalescent likelihood of gene trees

Heled and Drummond 2010

*BEAST Model

$$P(S|D) = \frac{\prod_{i=1}^{n} P(d_i|g_i)P(g_i|S)P(S)}{P(D)}$$

- 4 components:
  - $P(d_i|g_i)$- standard likelihood for alignment and gene tree i
  - $P(g_i|S)$- coalescent likelihood of gene trees
  - $P(S)$ - uniform topology
    - birth-death or Yule branching
    - gamma pop sizes with hyperprior

$P(D)$ - normalizing constant

Heled and Drummond 2010

*BEAST

$P(g_i|S)$

- Likelihood of gene trees given the species tree
- We have an equation to calculate the likelihood of coalescent histories within a lineage

$$\mathrm{P}(L|N(t)) = \prod_{i=2}^{n} \frac{\binom{n}{2}}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(t)} dt\right)$$

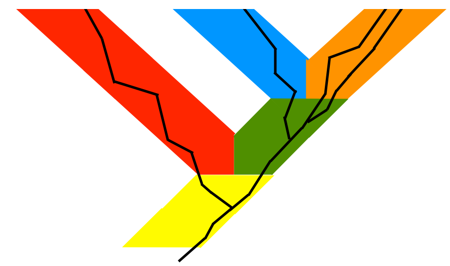
- How might we extend this to a whole tree?

Heled and Drummond 2010

*BEAST

- Answer: take the product of the coalescent likelihood along each branch

$$= \prod_{i=2}^{n} \frac{\binom{n}{2}}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(t)} dt\right) \times \prod_{i=2}^{n} \frac{\binom{n}{2}}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(t)} dt\right) \times \prod_{i=2}^{n} \frac{\binom{n}{2}}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(t)} dt\right) \times \prod_{i=2}^{n} \frac{\binom{n}{2}}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(t)} dt\right) \times \prod_{i=2}^{n} \frac{\binom{n}{2}}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(t)} dt\right)$$

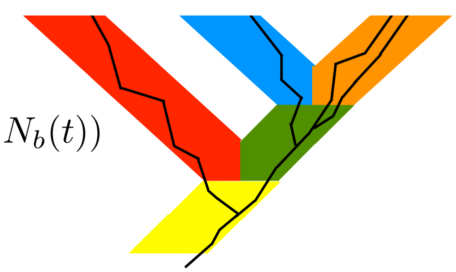


Heled and Drummond 2010

*BEAST

- Answer: take the product of the coalescent likelihood along each branch

$$\mathrm{P}(g|S) = \prod_{b \in S} \mathrm{P}(L_b(g)|N_b(t))$$



Heled and Drummond 2010

*BEAST

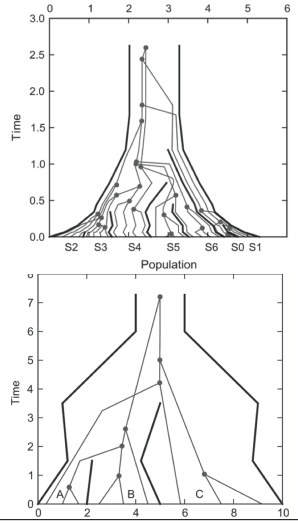
$$P(S|D) = \frac{\prod_{i=1}^{n} P(d_i|g_i)P(g_i|S)P(S)}{P(D)}$$

- Assumptions and limitations:
  - lineage sorting only source of incongruence
  - no gene flow following speciation
  - Implements a couple of demographic functions

Heled and Drummond 2010

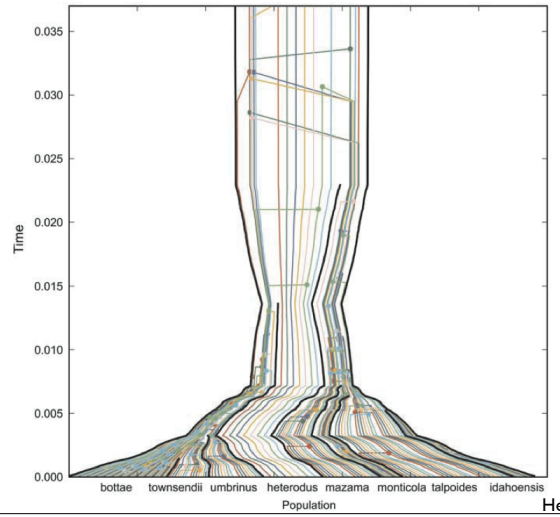
*BEAST demographic functions

- Constant size
- Linear change



Heled and Drummond 2010

*BEAST - Pocket Gophers



Heled and Drummond 2010

*BEAST

```
<!-- The list of taxa to be analysed (can also include dates/ages).         -->
<!-- ntax=26                                                                 -->
<taxa id="taxa">
    <taxon id="Orthogeomys_heterodus">
        <attr name="species">
            heterodus
        </attr>
    </taxon>
    <taxon id="Thomomys_bottae_awahnee_a">
        <attr name="species">
            bottae
        </attr>
    </taxon>
```

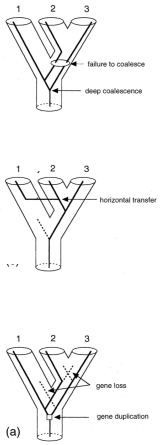
```
<alignment id="alignment1" dataType="nucleotide">
    <sequence>
        <taxon idref="Orthogeomys_heterodus"/>
        ATTCTAGGCAAAAAG-AGCAATGCTGGAGGTATTACAATACCAGACTTCAAACTTTACTATAGAGCCATAATAACAAA
    </sequence>
    <sequence>
        <taxon idref="Thomomys_bottae_awahnee_a"/>
        ??????????????????????ATGCTGGTGGTATTACAATACCAGACTTCAAACTCTACTATAGAGCCATAATAACAAA
    </sequence>
    <sequence>
        <taxon idref="Thomomys_bottae_awahnee_b"/>
        ??????????????????????ATGCTGGTGGTATTACAATACCAGACTTCAAACTCTACTATAGAGCCATAATAACAAA
    </sequence>
```


*BEAST

```
<!-- Species definition: binds taxa, species and gene trees                 -->
<species id="species">
    <sp id="bottae">
        <taxon idref="Thomomys_bottae_awahnee_a"/>
        <taxon idref="Thomomys_bottae_awahnee_b"/>
        <taxon idref="Thomomys_bottae_xerophilus"/>
        <taxon idref="Thomomys_bottae_cactophilus"/>
        <taxon idref="Thomomys_bottae_albatus"/>
        <taxon idref="Thomomys_bottae_ruidosae"/>
        <taxon idref="Thomomys_bottae_bottae"/>
        <taxon idref="Thomomys_bottae_alpinus"/>
        <taxon idref="Thomomys_bottae_riparius"/>
        <taxon idref="Thomomys_bottae_mewa"/>
        <taxon idref="Thomomys_bottae_saxatilis"/>
        <taxon idref="Thomomys_bottae_laticeps"/>
    </sp>
    <sp id="heterodus">
        <taxon idref="Orthogeomys_heterodus"/>
    </sp>
    <sp id="idahoensis">
        <taxon idref="Thomomys_idahoensis_pygmaeus_a"/>
        <taxon idref="Thomomys_idahoensis_pygmaeus_b"/>
    </sp>
```

Sources of gene tree variation

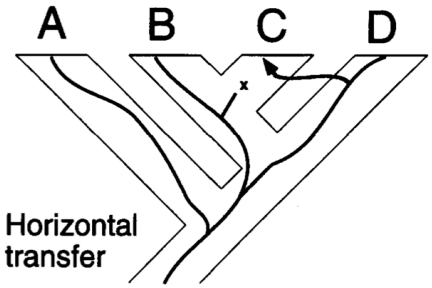
- Incomplete coalescence
- Horizontal transfer
- Gene duplication



85

Horizontal gene transfer

- Caused by hybridization or transfer via vectors
- Leads to a network like species history
- Can occur in conjunction with incomplete coalescence

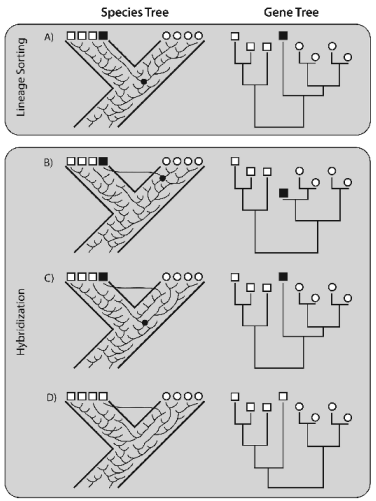


Maddison 1997

86

Horizontal gene transfer

- Work on this is also emerging
- One basic idea is to use the distribution of branching times to detect shallow branching events that are unlikely under the coalescent

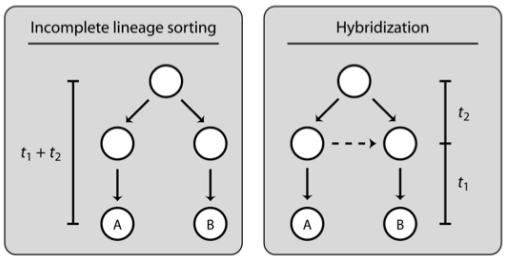


Joly et al. 2012

87

Horizontal gene transfer

- One approach: uses a technique called posterior predictive simulation to assess the probability of observing "young" nodes under the multispecies coalescent by itself

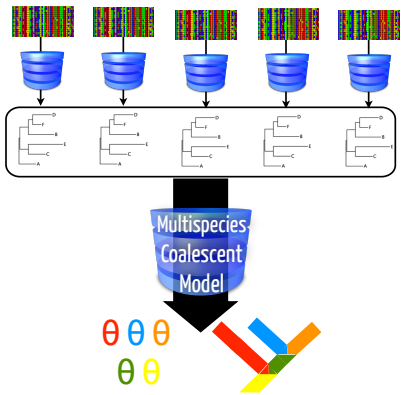


Joly et al. 2012

88

Horizontal gene transfer

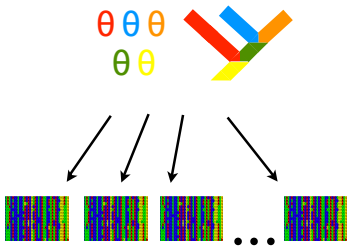
- Steps for posterior predictive simulation:
  - perform species tree analyses (*BEAST)



Joly et al. 2012

Horizontal gene transfer

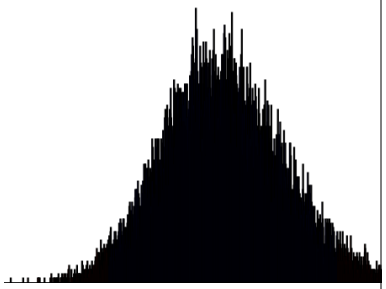
- Steps for posterior predictive simulation:
  - perform species tree analyses (*BEAST)
  - Sample species trees, branch lengths, and population sizes from the posterior distribution
  - Use these samples to simulate sequences



Joly et al. 2012

Horizontal gene transfer

- Steps for posterior predictive simulation:
  - perform species tree analyses (*BEAST)
  - Sample species trees, branch lengths, and population sizes from the posterior distribution
  - Use these samples to simulate sequences
  - Find the minimum pairwise distance between simulated sequences for your species of interest

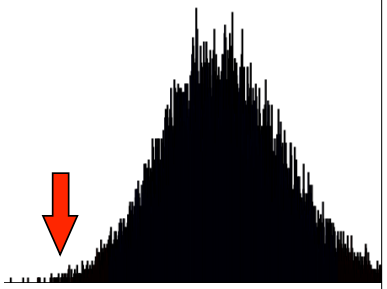


Posterior predictive distribution of minimum pairwise divergences

Joly et al. 2012

Horizontal gene transfer

- Steps for posterior predictive simulation:
  - perform species tree analyses (*BEAST)
  - Sample species trees, branch lengths, and population sizes from the posterior distribution
  - Use these samples to simulate sequences
  - Find the minimum pairwise distance between simulated sequences for your species of interest
  - Compare the minimum observed pairwise difference to construct p-value

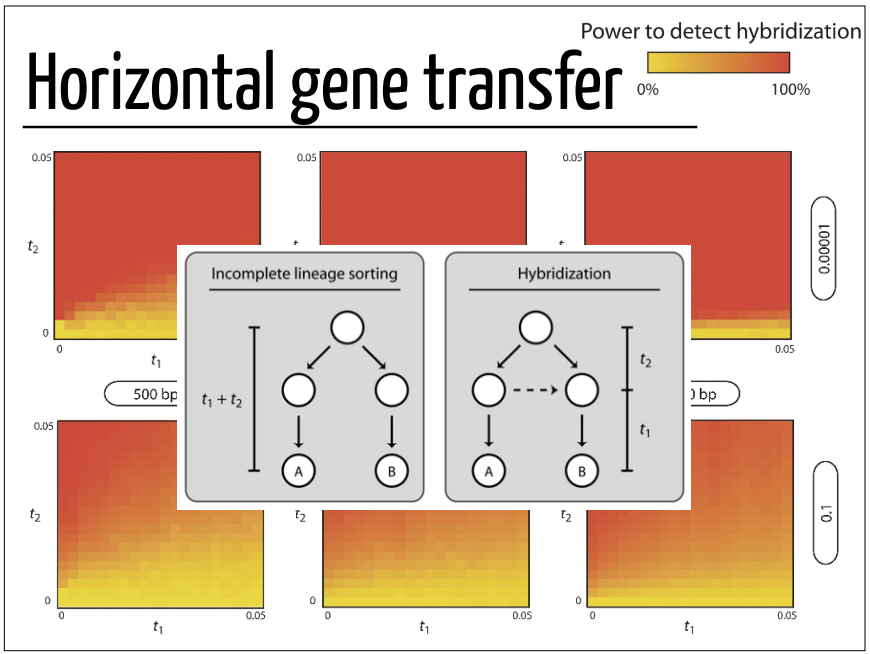


Posterior predictive distribution of minimum pairwise divergences

$$p = \mathrm{P}(minDist(AB) < mindDist(AB)^{sim})$$

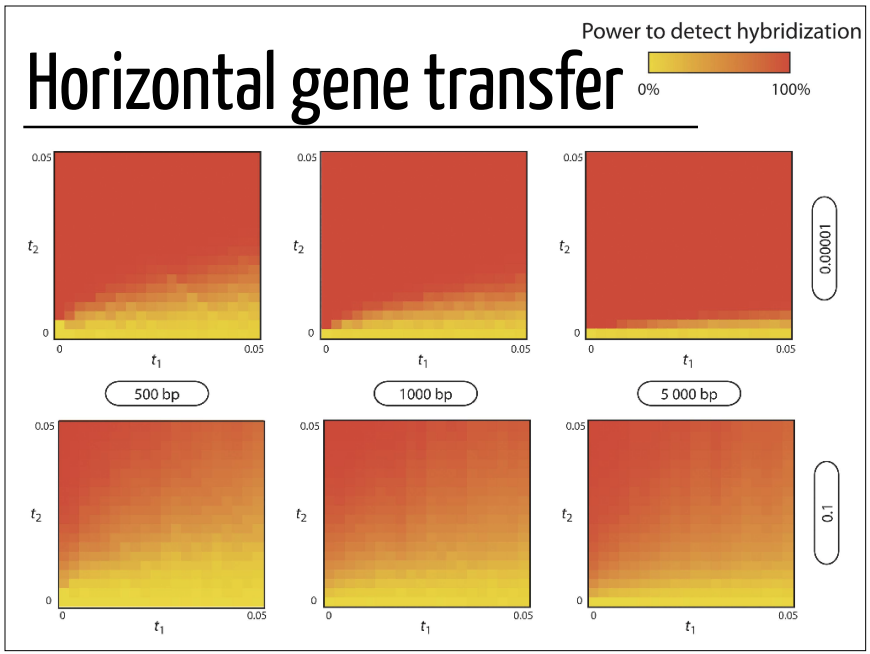
Joly et al. 2012

Horizontal gene transfer



93

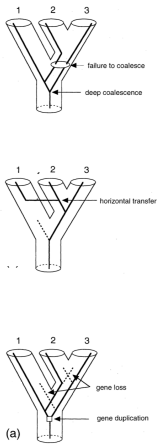
Horizontal gene transfer



94

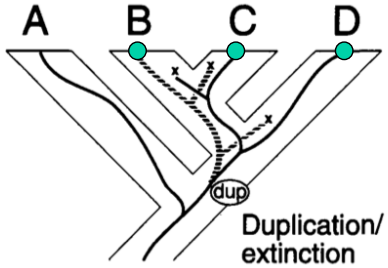
Sources of gene tree variation

- Incomplete coalescence
- Horizontal transfer
- Gene duplication



95

Gene Duplication



Surviving sequence

- gene duplications and extinctions can yield misleading gene trees.
- parsimony and likelihood approaches for addressing this

Maddison 1997

96

Gene Duplication

- One solution: Just avoid the problem altogether
  - This may often be the best option

Gene Duplication

- One solution: Just avoid the problem altogether
  - This may often be the best option
- For well characterized genomes, focus on known single copy genes

Gene Duplication

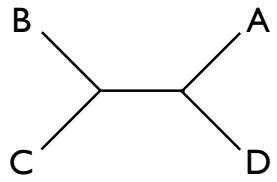
- One solution: Just avoid the problem altogether
  - This may often be the best option
- For well characterized genomes, focus on known single copy genes
- More problematic with large genome scale datasets
  - Need to be careful about automated homology assignment

Gene Tree Parsimony

- input a collection of rooted gene trees, find the species tree that minimizes the reconciliation cost
  - reconciliation cost is number of duplications, or duplications and losses, summed across gene trees

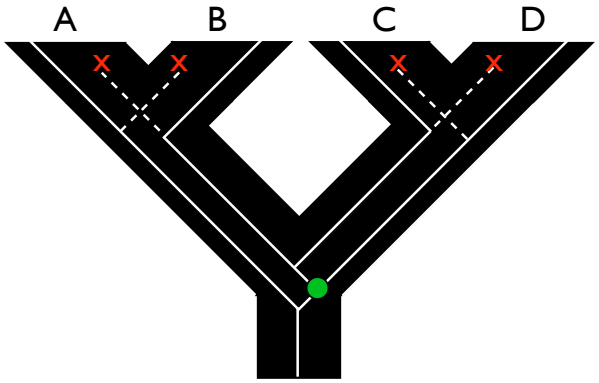
Gene Tree Parsimony

- example gene tree:



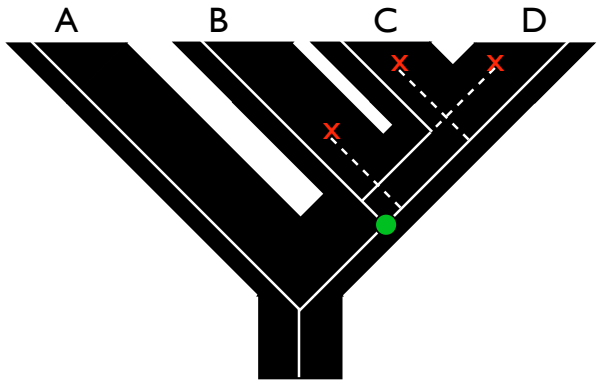
- calculate reconciliation costs for species trees

Gene Tree Parsimony



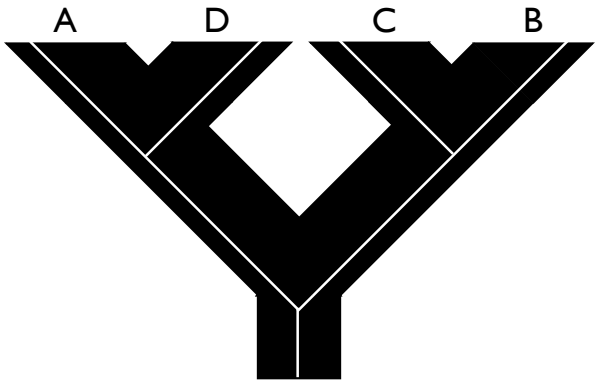
Reconciliation score = 5

Gene Tree Parsimony



Reconciliation score = 4

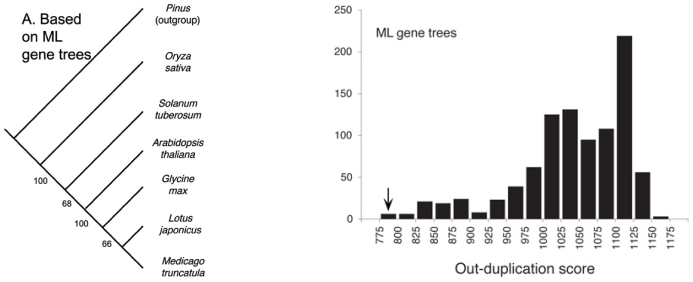
Gene Tree Parsimony



Reconciliation score = 0

Empirical Example

- Sanderson and McMahon 2007
  - GTP analysis of 576 gene trees for 6 angiosperm species (plus outgroup)
  - known species tree recovered successfully despite massive gene duplication



105

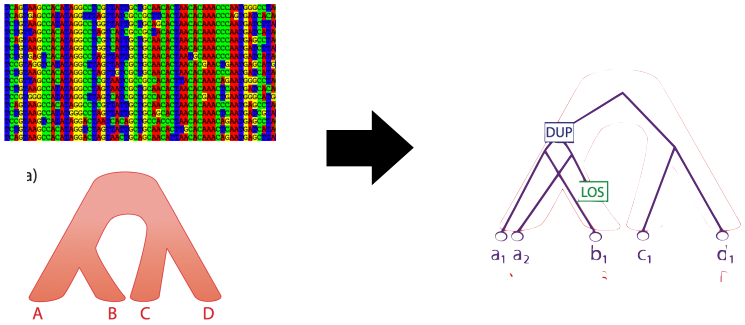
Gene Tree Parsimony

- small trees - Gtp (Sanderson and McMahon 2007)
- large trees - DupTree (Wehe et al 2008)

106

Statistical Approaches

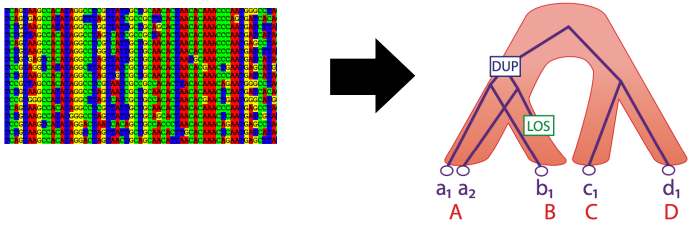
- Likelihood methods for inferring gene trees and duplication and loss history given a species tree have existed for some time



107

Statistical Approaches

- Likelihood methods for inferring gene trees and duplication and loss history given a species tree have existed for some time
- Until recently, no methods available to do the joint inference



Boussau et al. 2013

108

Statistical Approaches

- Likelihood methods for inferring gene trees and duplication and loss history given a species tree have existed for some time
- Until recently, no methods available to do the joint inference

$$L(T, S, N|A) = \prod_{i \in \mathcal{G}} L(S, N|T_i) L(T_i|A_i)$$

Boussau et al. 2013

Statistical Approaches

- Likelihood methods for inferring gene trees and duplication and loss history given a species tree have existed for some time
- Until recently, no methods available to do the joint inference

$$L(T, S, N|A) = \prod_{i \in \mathcal{G}} L(S, N|T_i) L(T_i|A_i)$$

Boussau et al. 2013

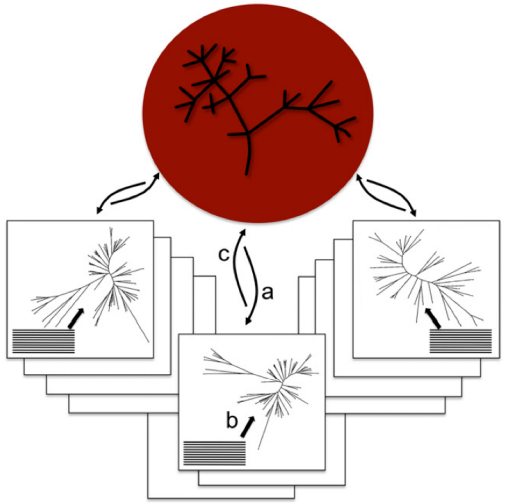
Statistical Approaches

- Likelihood methods for inferring gene trees and duplication and loss history given a species tree have existed for some time
- Until recently, no methods available to do the joint inference

$$L(T, S, N|A) = \prod_{i \in \mathcal{G}} L(S, N|T_i) L(T_i|A_i)$$

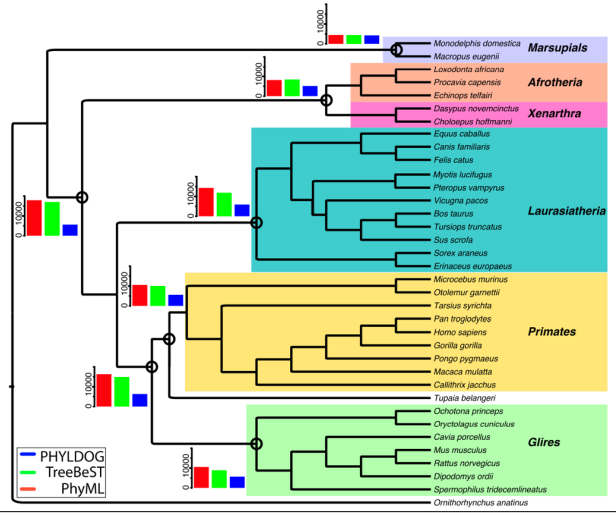
Boussau et al. 2013

Statistical Approaches



Boussau et al. 2013

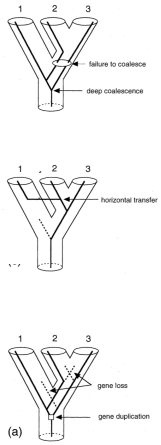
Statistical Approaches



Boussau et al. 2013

Sources of gene tree variation

- Incomplete coalescence
- Horizontal transfer
- Gene duplication



Sources of gene tree variation

The inference of gene trees with species trees

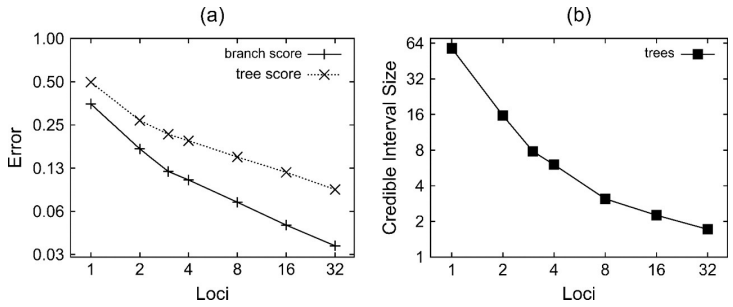
Gergely J. Szöllősi[1], Eric Tannier[2,3,4], Vincent Daubin[2,3], Bastien Boussau[2,3]

arXiv: 1311.0651v1

Wrapping up

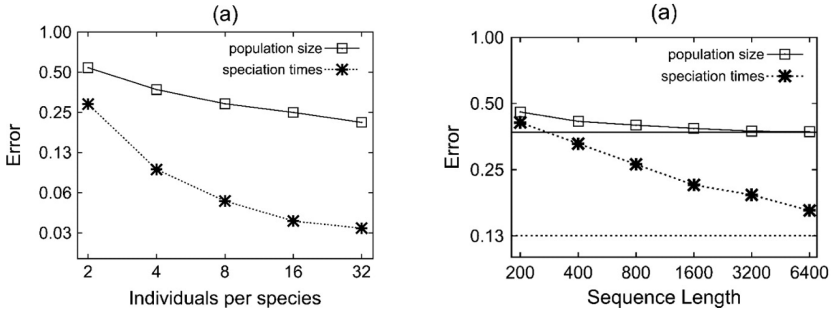
- Some thoughts:
  - There are several options here, you should carefully choose a model based on biological knowledge
  - Need for more simulation studies
    - Sensitivities to priors and demographic functions
  - Data needs are substantial

*BEAST - dataset design



Heled and Drummond 2010

*BEAST - dataset design

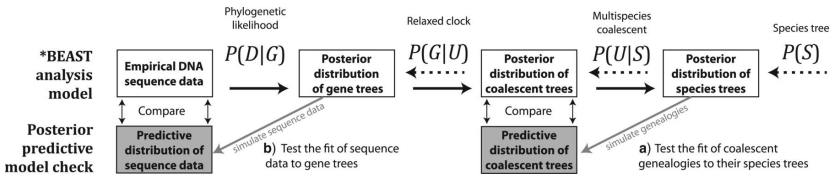


Heled and Drummond 2010

Difficulties

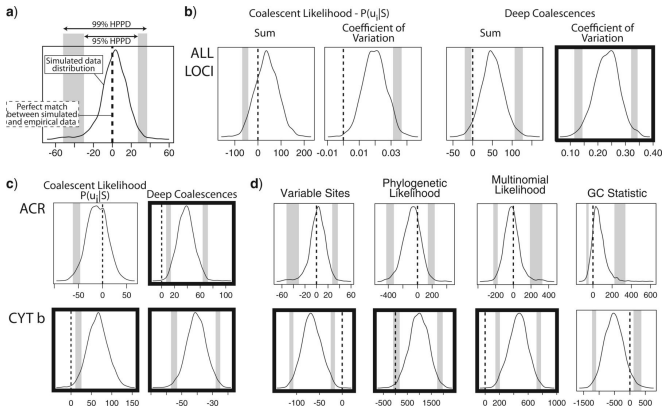
- Often making some strong assumptions:
  - changes or constancy of population size
  - species membership and assignments
  - sources of gene tree variation
  - Not always well known how robust it is to deviations from the correct model

Coalescent model plausibility



Reid et al. 2013

Coalescent model plausibility



Reid et al. 2013

the end