Phylogenetics I – BUCKy

NESCent Academy summer course

July 2014

Cécile Ané

Departments of Statistics and of Botany
University of Wisconsin - Madison

Plan for today

- gene tree incongruence: **Why bother**?
- **Gene tree models**

  STEM, *BEAST, BEST: coalescent process

  BUCKy: clustering prior on gene trees
- **BUCKy**, model assumptions and goals:

  concordance factors, concordance tree, population tree
- **Comparisons** between methods

  from simulations
- **Tutorial**

Why bother?

- Why not just **concatenate** all loci?

  When we do so, we usually get strong support.

- Some discordance due to estimation errors

  Sampling error (wrong tree but poorly supported)

  model mis-specification (wrong tree, high support), e.g: LBA

- Concatenation: support for wrong tree can be high

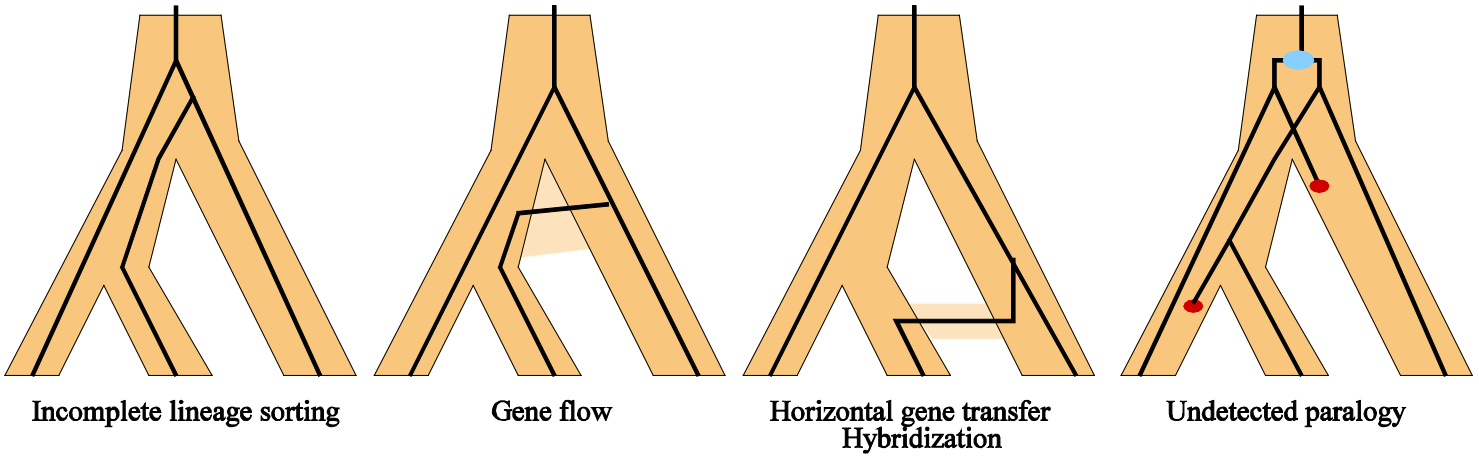
  Amplification of ‘systematic biases’ e.g. LBA

  Even without any sampling error or systematic bias, the presence of ILS can cause damage (“Anomaly zone”).

Why bother?

Biological processes cause real discordance

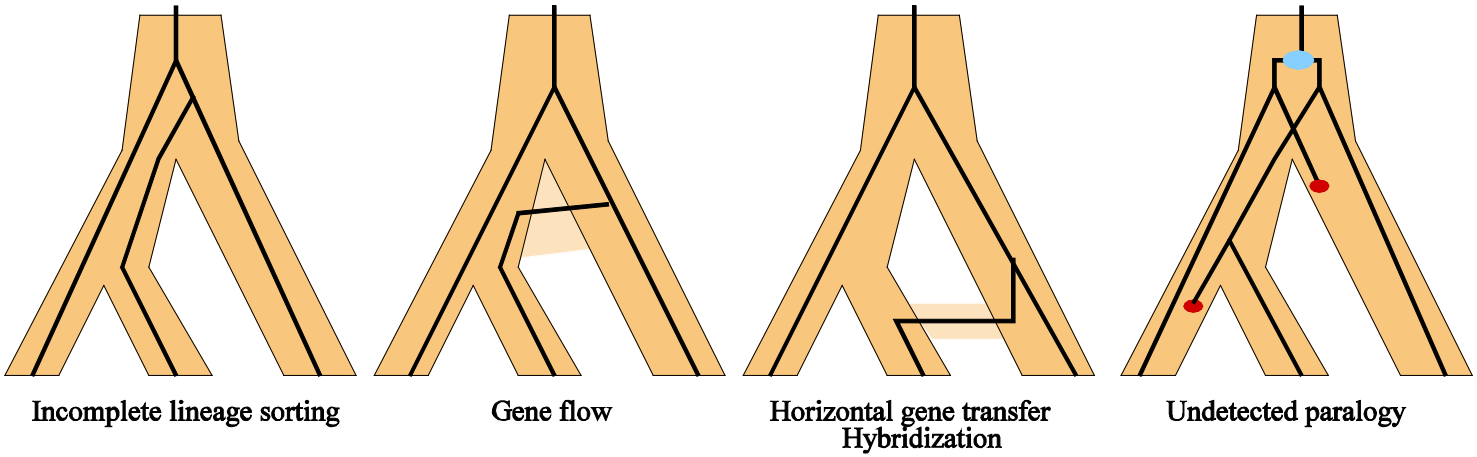
- Incomplete lineage sorting
- Gene flow, Hybridization, Horizontal gene transfers
- Unrecognized paralogy: duplications and losses combined may go unseen.



Why bother?

Two reasons to estimate species trees:

- Avoid highly supported wrong tree from concatenation
- Study what caused discordance



Plan for today

- gene tree incongruence: **Why bother**?
- **Gene tree models**

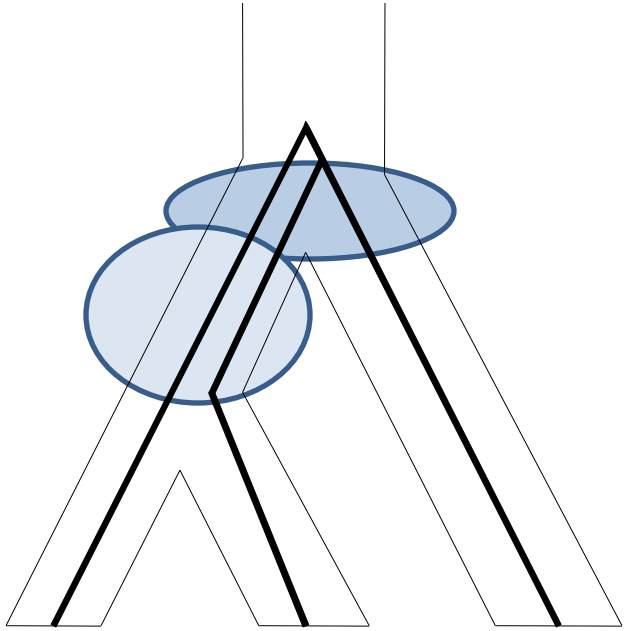
  STEM, *BEAST, BEST: coalescent process

  BUCKy: clustering prior on gene trees
- **BUCKy**, model assumptions and goals:

  concordance factors, concordance tree, population tree
- **Comparisons** between methods

  from simulations
- **Tutorial**

The coalescent process



- Wright-Fisher model: forward in time.

  N_e diploid individuals.
- Description backwards in time: time in **coalescent units**:

  $$u = \#\text{generations}/ (2N_e)$$

Time to next coalescent event: Exponential distribution. P(T>u) = exp(-rate*u)

All **pairs** have **equal probabilities** and equal rates to coalesce.

STEM, BEST, *BEAST, STAR, MP-EST: coalescent model

Discordance assumed from the **coalescent** model: each species is panmictic, no gene flow, no population structure.

Can include several individual per species. **Species assignment** needs to be known without error.

Estimates **divergence times** and N_e along each branch, through

$$\theta = 4N_e\mu.$$

Genes: clock-like trees, possibly different **relative mutation rates** (r_i)

Branch length in:	tree for gene *i*		Species tree
	$r_i t$ (subst/site)	$\longleftrightarrow$	t/θ (coalescent units)
	$r_i u\theta$ (subst/site)		u (coalescent units)

BUCKy: Bayesian concordance analysis

Prior distribution on gene trees: **not** from coalescent model, but by a **Dirichlet** process.

No assumption regarding the **source** of discordance.

Could be: horizontal transfers, hybridization, incomplete lineage sorting, unrecognized paralogy, systematic bias.

Based on *clustering* of genes with the same *topology*

no branch length assumption: genes trees do not need to be clock-like, different genes can have different rates.

No prior assignment of individuals to species.

Dirichlet process prior on gene trees

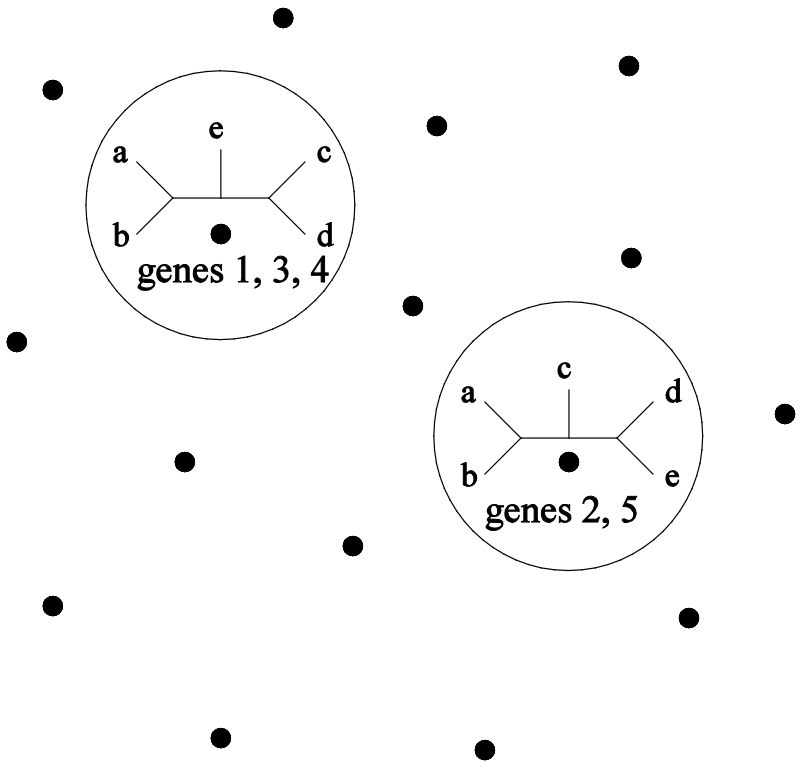
Prior probability of a set of gene trees: depends on how many genes share the same topology, and on a parameter α.

Ex: genes 1,3,4 have T_1=((a,b),e,(c,d))

genes 2,5 have T_2= ((a,b),c,(d,e))

Prior prob. = f(# and size of clusters, α)

2 "clusters" of genes, of sizes 3 and 2.



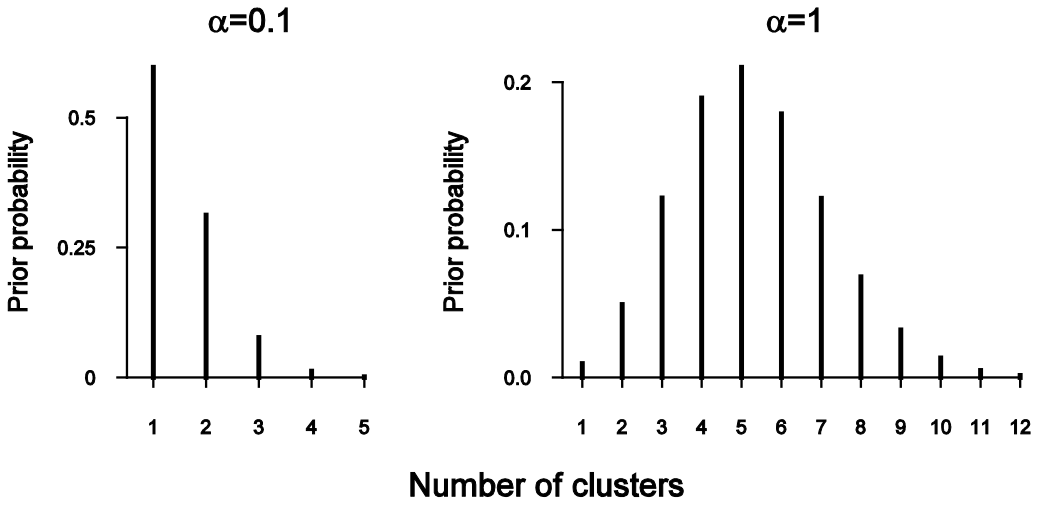
Probability that 2 randomly sampled genes share the same topology $=(1+\alpha/T)/(\alpha+1) \approx 1/(\alpha+1)$ with moderately many taxa

Dirichlet process prior on gene trees

α: **a-priori** level of **discordance**

α = 0: forces a single cluster. All genes assumed to share the same topology. Like concatenated approach in MrBayes, but all parameters 'unlinked'.

α = infinite: # clusters doesn't matter. Independent gene trees. Like consensus approach with concordance factors estimation.



α = 0.1 and α = 1: a priori # of clusters from 106 genes on 8 taxa.

Plan for today

- gene tree incongruence: **Why bother**?
- **Gene tree models**

  STEM, *BEAST, BEST: coalescent process

  BUCKy: clustering prior on gene trees
- **BUCKy**, model assumptions and goals:

  concordance factors, concordance tree, population tree
- **Comparisons** between methods

  from simulations
- **Tutorial**

BUCKy: Bayesian concordance analysis

Goal: infer the **primary concordance tree**, along with

Concordance factors: measures of **genomic support**,
% of the genome having a clade.
Concordance tree built from clades with largest CFs (greedy)

Credibility intervals: measures of **statistical support.**

Then: use CFs to estimate a **population tree**, assuming that the coalescent is the source of all discordance.

BUCKy output

Main result: sample of "gene-to-tree maps". At each MCMC generation: each gene is 'mapped' to a certain tree.

Posterior distribution on Gene-to-Tree maps is summarized:

- **Concordance factor (CF)** of each clade and its credibility interval, Concordance factor of each quartet.

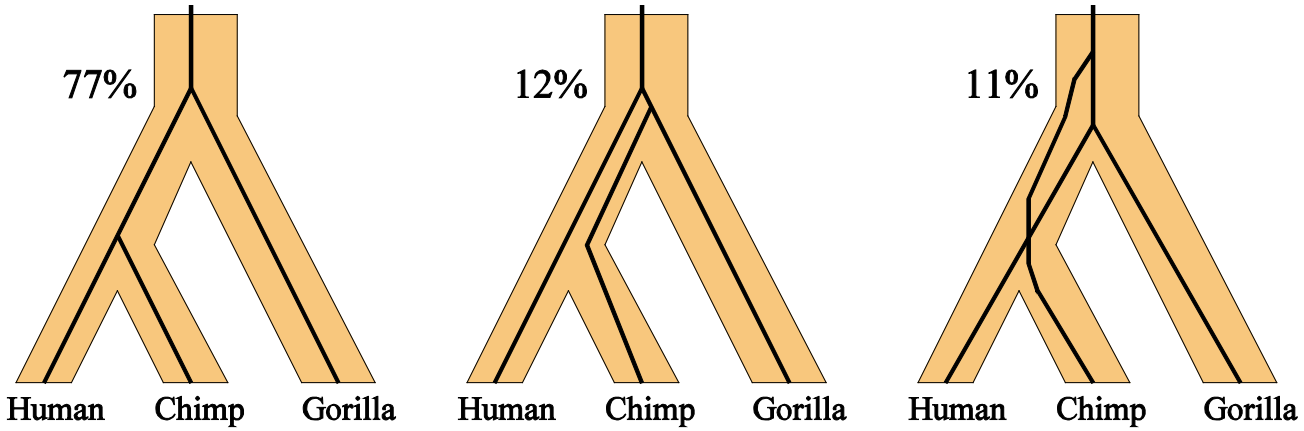
  Sample-wide CF: % genes in the sample

  Genome-wide CF: % genes in the genome

- **Primary concordance tree**: made of clades with highest CFs
- **Estimated population tree**: made of quartets with highest CFs, branch lengths estimated in coalescent units.
  Assumes the coalescent to estimate the tree from CFs.

Example: Great apes

30,040 alignments on 5 taxa, from Ebersberger et al. (2007) analysis.
1/3 alignments were clock-like and informative:

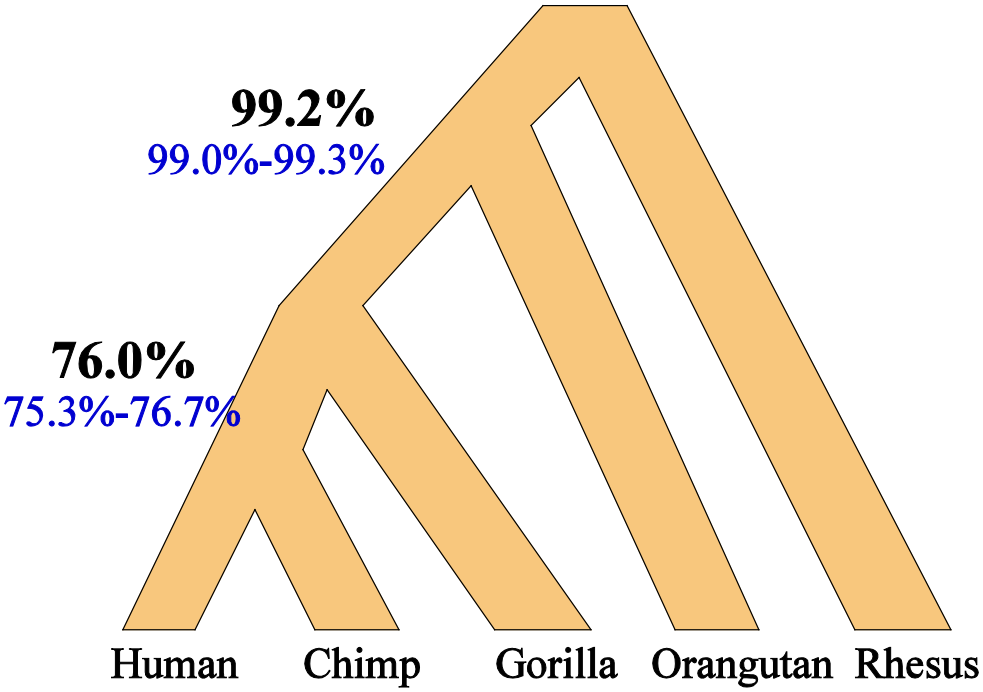


Pattern compatible with incomplete lineage sorting only?
Population tree with branch lengths in coalescent units?

Great apes: concordance tree

Re-analysis with BUCKy, using all 30,040 alignments (including those with phylogenetic uncertainty, and those with non-clock trees)

Primary concordance tree:

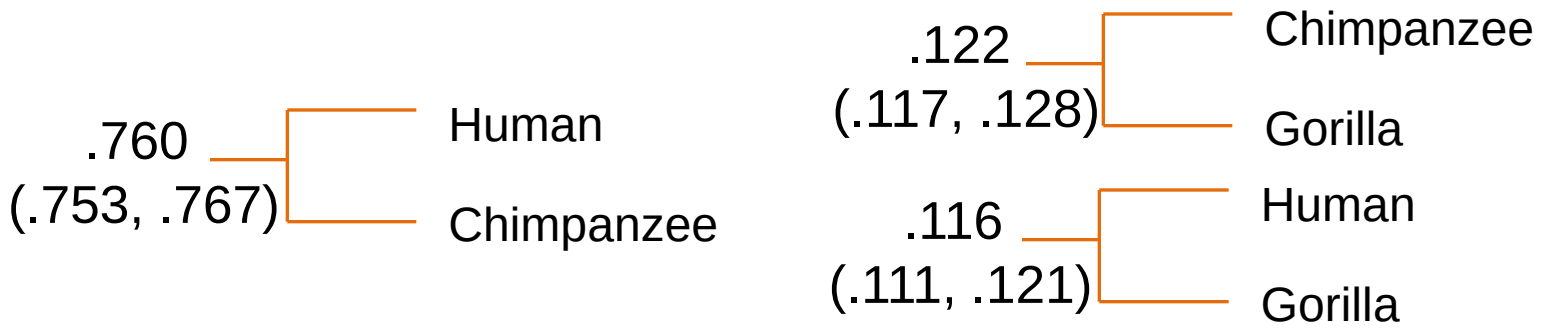


Genomic support: concordance factors (values on edges)

Statistical support:
95% credibility interval for CFs. This is the concordance tree with 1.0 posterior probability.

These clades' CF are > any conflicting clade's CF with 1.0 posterior probability (compare credibility intervals)

Great apes: concordance factors to test ILS



Concordance factors are **compatible with ILS-only** model:

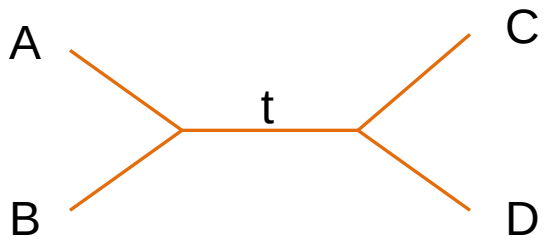
the 2 minor clades' CFs have overlapping credibility intervals
$\rightarrow$ do not differ significantly.

Equal concordance factors for (CG|…) and (HG|…) expected under the coalescent model.

Great apes: population tree

Population tree: branch lengths in coalescent units

Branch lengths from concordance factors on quartets, assuming coalescent model:

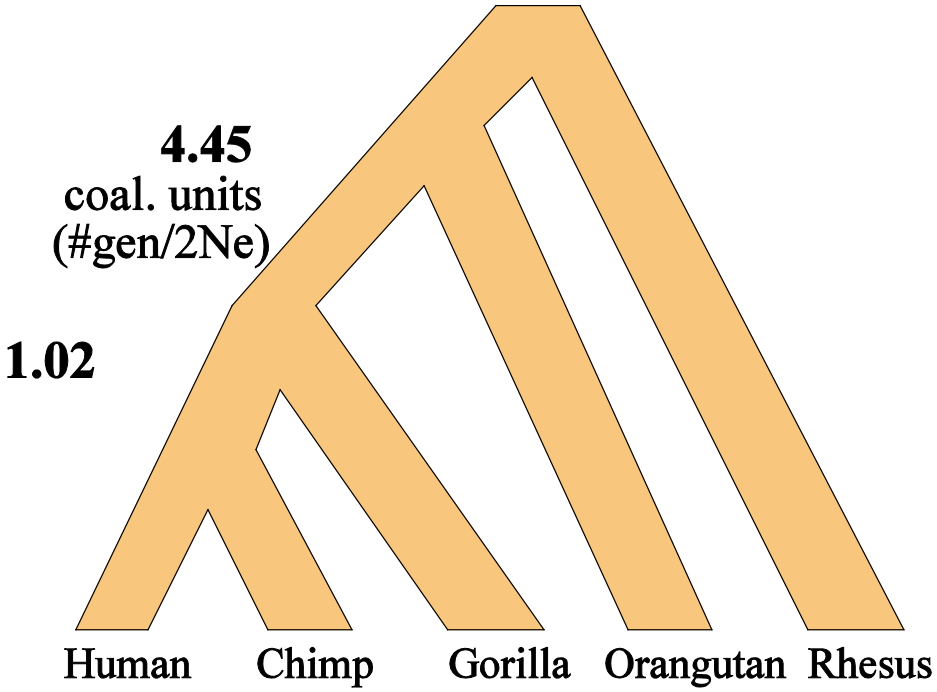


under coalescent model:

$CF_{AB|CD} = 1-(2/3)\ e^{-t}$

$CF_{AC|BD} = \quad (1/3)\ e^{-t} = CF_{AD|BC}$

So $t = -\log (3/2)*(1- CF_{AB|CD})$



BUCKy: model assumptions

- Gene trees: no assumption of molecular clock.

  Branch lengths completely unlinked across genes.
- Each locus has its own:

  substitution model (e.g. JC, HKY+Γ, GTR+I, WAG)

  substitution rates (e.g. ti/tv)

  base frequencies

  rate heterogeneity (α)

  branch lengths
- Dirichlet prior on the number of clusters and cluster sizes, controlled by α. Uniform prior for the tree of each cluster.

BUCKy: model assumptions

Bayesian analysis to **jointly estimate** gene trees.

$$P(\text{gene trees } \tau_1,\ldots,\tau_k \mid \text{data } D_1,\ldots,D_k \text{, prior, model})$$

Posterior prob. of the *k* gene trees

$$= \quad f(\tau_1,\ldots,\tau_k) \quad \prod P(\tau_i \mid D_i) \quad / \quad P(\text{ data } D_1,\ldots,D_k)$$

Dirichlet prior on gene trees

Single gene posterior

Two-step algorithm:

1. get $P(\tau_i \mid D_i)$ values: **MrBayes** on each gene separately
2. combine all tree files (.t) using **bucky**

Caveats

$$P(\text{gene trees } \tau_1,\ldots,\tau_k \mid \text{data } D_1,\ldots,D_k \text{, prior, model})$$

Posterior prob. of the *k* gene trees

$$= \quad f(\tau_1,\ldots,\tau_k) \quad \prod P(\tau_i \mid D_i) \quad / \quad P(\text{ data } D_1,\ldots,D_k)$$

Dirichlet prior on gene trees

Single gene posterior

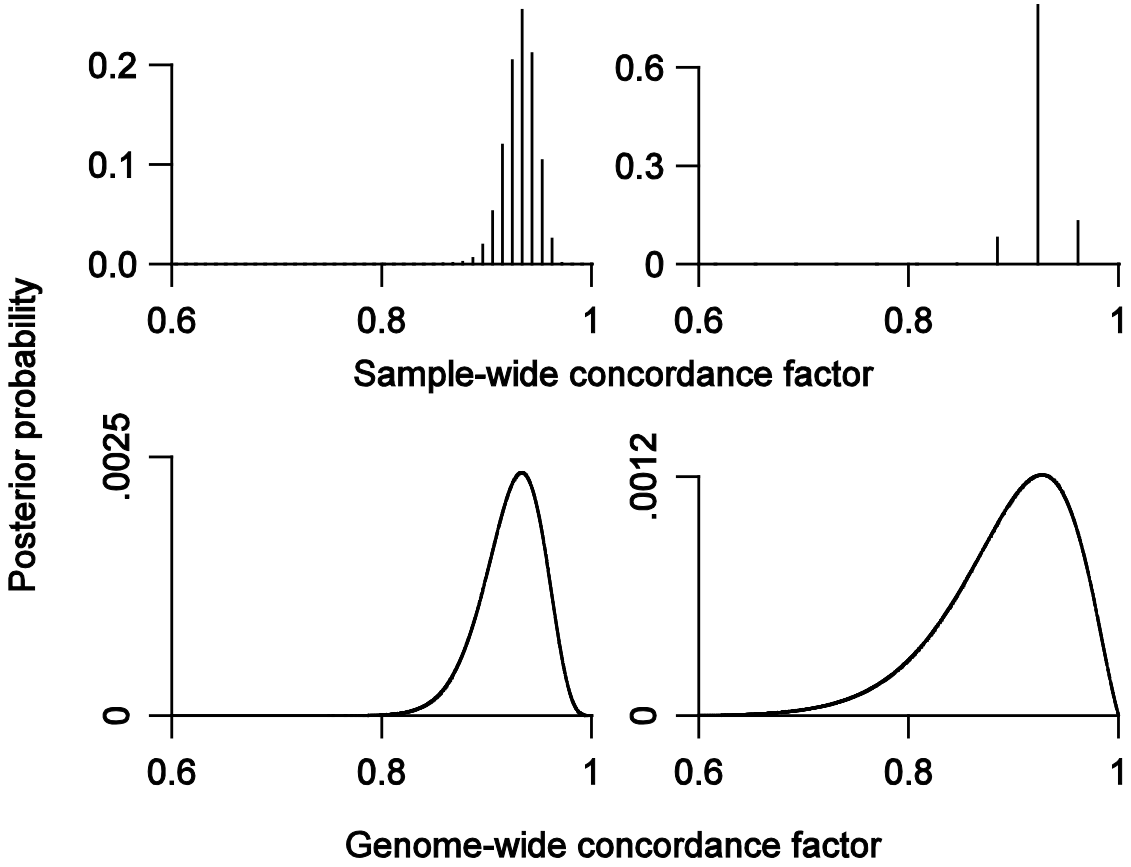
- $P(\tau_i \mid D_i)$ may not be well approximated *for all likely* topologies τ_i with short alignments on many taxa.

  Consequence: underestimated concordance factors on many taxa
- Dirichlet prior: discordant trees are drawn at random from all possible trees. Unrealistic under HGT, or hybridization, etc.
- New version to fix both issues: new prior, and estimation of $P(\tau_i \mid D_i)$ using conditional clade probabilities (Larget 2013 Syst. Biol.).

Genome-wide concordance factors

- Proportion of genes **in the genome** that have a clade.
- Even if PP(6 of my 10 genes have (abe|cd)) = 1, there is still uncertainty about the genome-wide CF.
- Analytical formula.

Sampling 106 genes versus 26 genes



Plan for today

- gene tree incongruence: **Why bother**?
- **Gene tree models**

  STEM, *BEAST, BEST: coalescent process

  BUCKy: clustering prior on gene trees
- **BUCKy**, model assumptions and goals:

  concordance factors, concordance tree, population tree
- **Comparisons** between methods

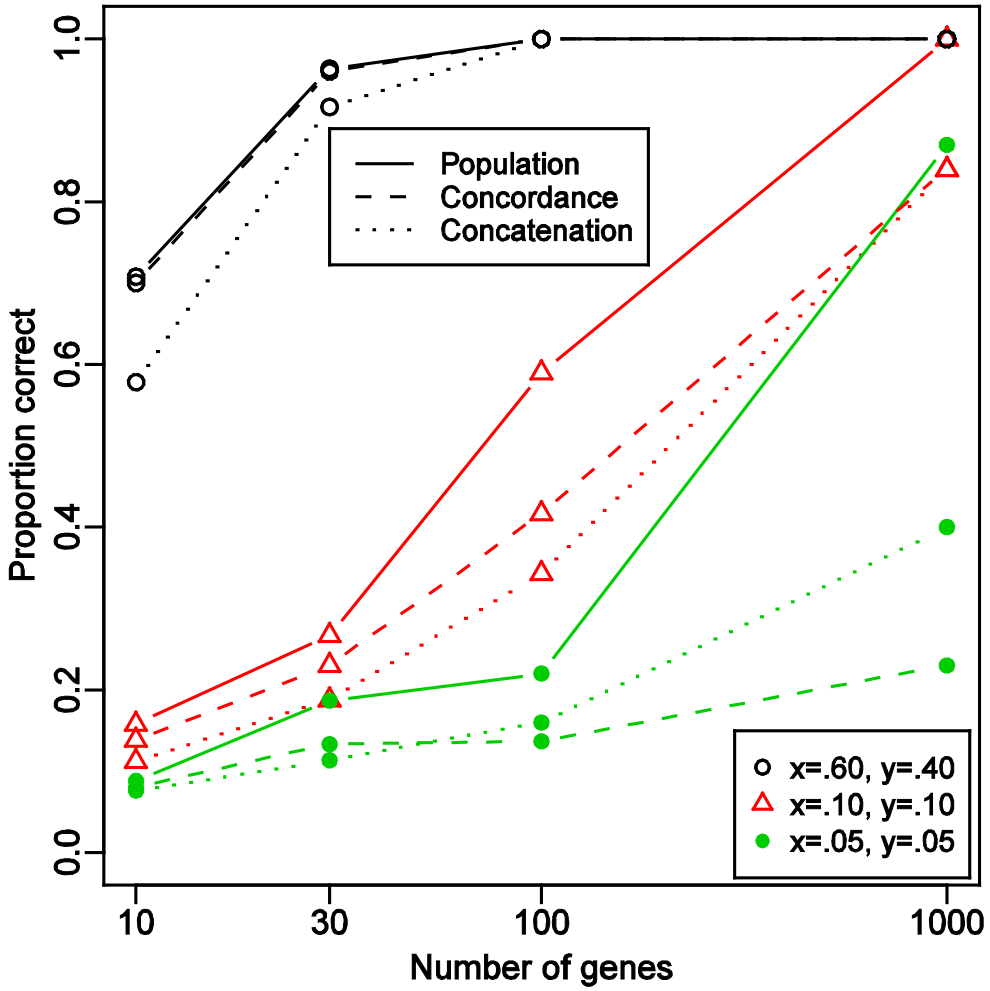
  from simulations
- **Tutorial**

Method comparisons from simulations

- Coalescent model, 5-taxon asymmetric population tree

- Comparing BUCKy's **population** tree, **concordance** tree, and tree from MrBayes on **concatenated** genes

- If the coalescent is true, consistent population tree

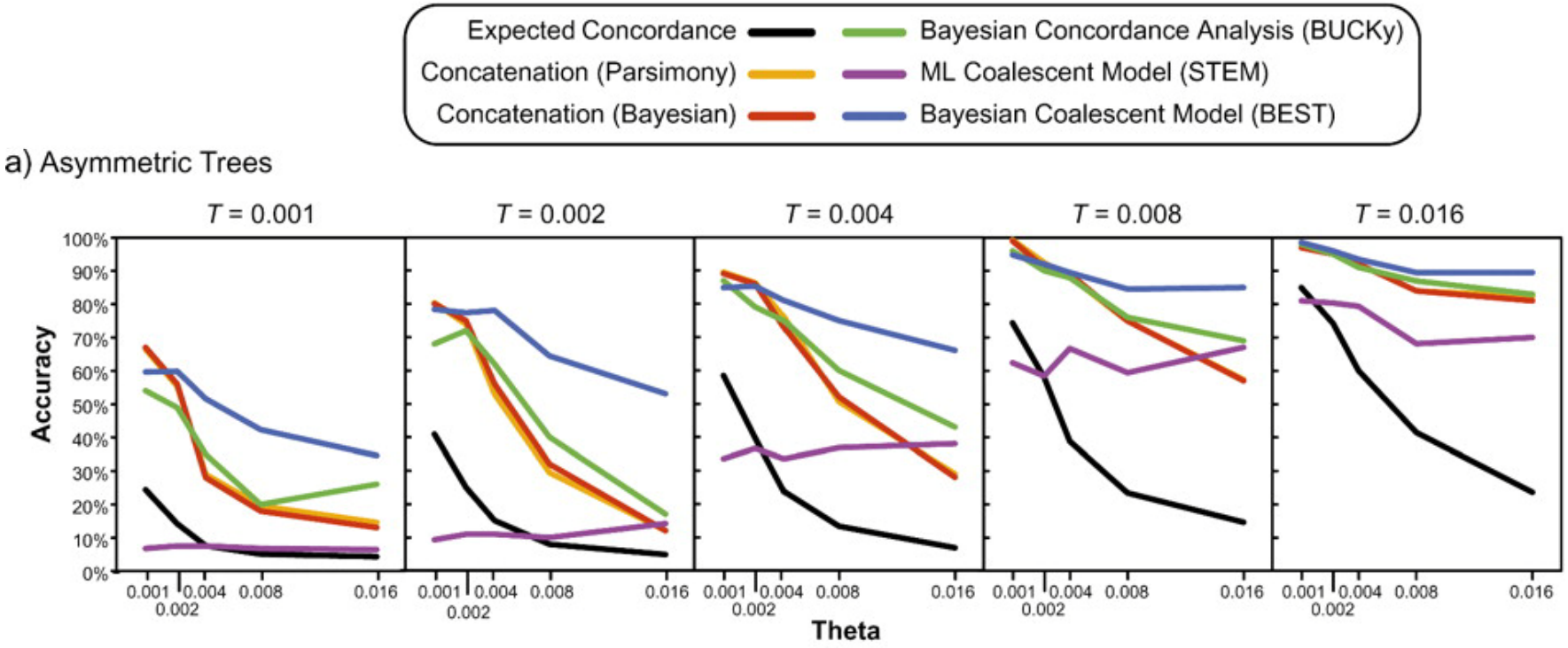


B. Larget, S.K. Kotha, C.N. Dewey, C. Ané (2010). *Bioinformatics* 26: 2910-2911

Method comparisons from simulations

- Coalescent model, 5-taxa, (a) asymmetric species tree, then (b) symmetric species tree.
- Comparing trees from BEST, STEM, BUCKy's concordance tree and tree from concatenation.

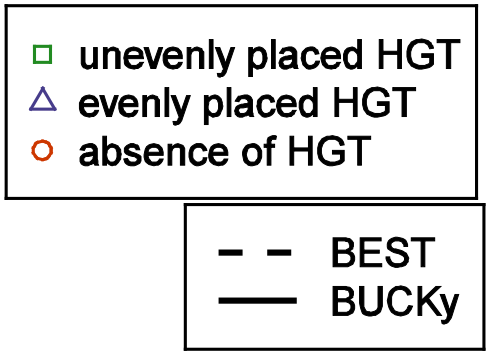
Leaché A D , Rannala B (2011) *Systematic Biology* 60:126-137

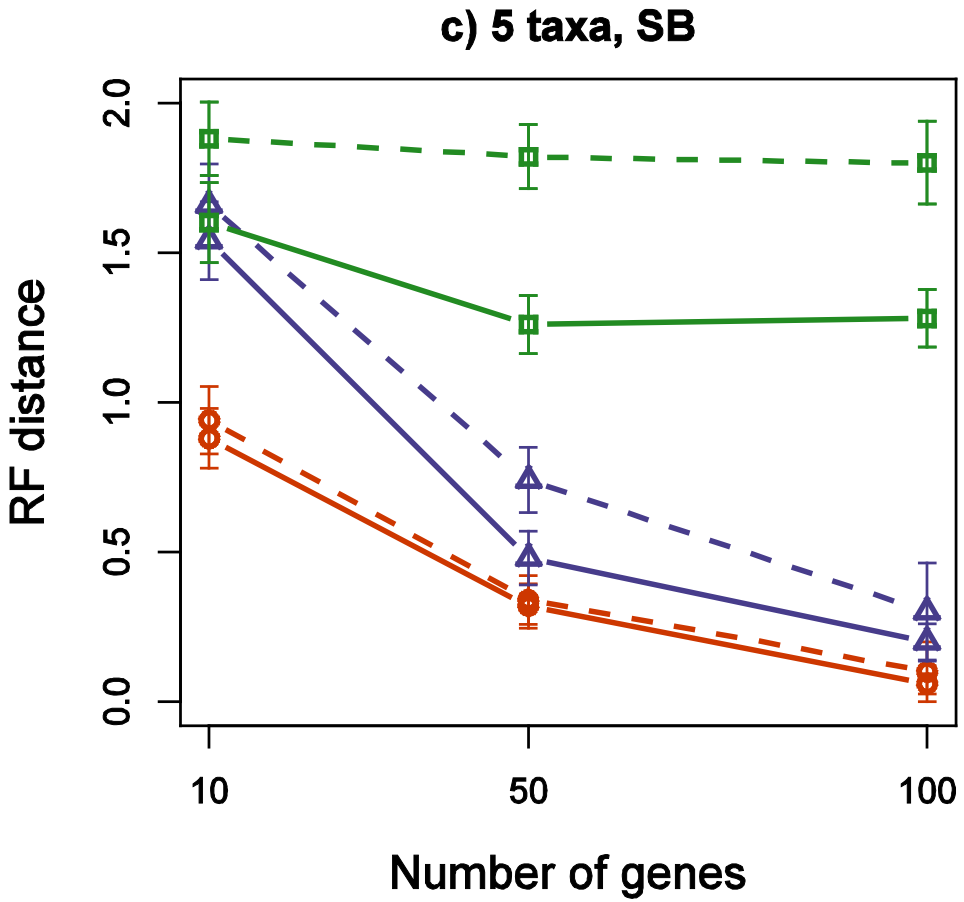


Method comparisons from simulations

- Coalescent model **with HGT**, 5-taxon asymmetric population tree

- Comparing BEST and BUCKy's concordance tree





Y. Chung, C. Ané (2011). *Systematic Biology*. 60: 261-275

Plan for today

- gene tree incongruence: **Why bother**?
- **Gene tree models**

  STEM, *BEAST, BEST: coalescent process

  BUCKy: clustering prior on gene trees
- **BUCKy**, model assumptions and goals:

  concordance factors, concordance tree, population tree
- **Comparisons** between methods

  from simulations
- **Tutorial**: your turn!