Review of "Comparing two Bayesian methods for gene tree/speies tree reconstruction: Simulations with incomplete lineage sorting and horizontal gene trasnfer" by Chung and Ané

Danielle Campbell

April 7, 2016

Background

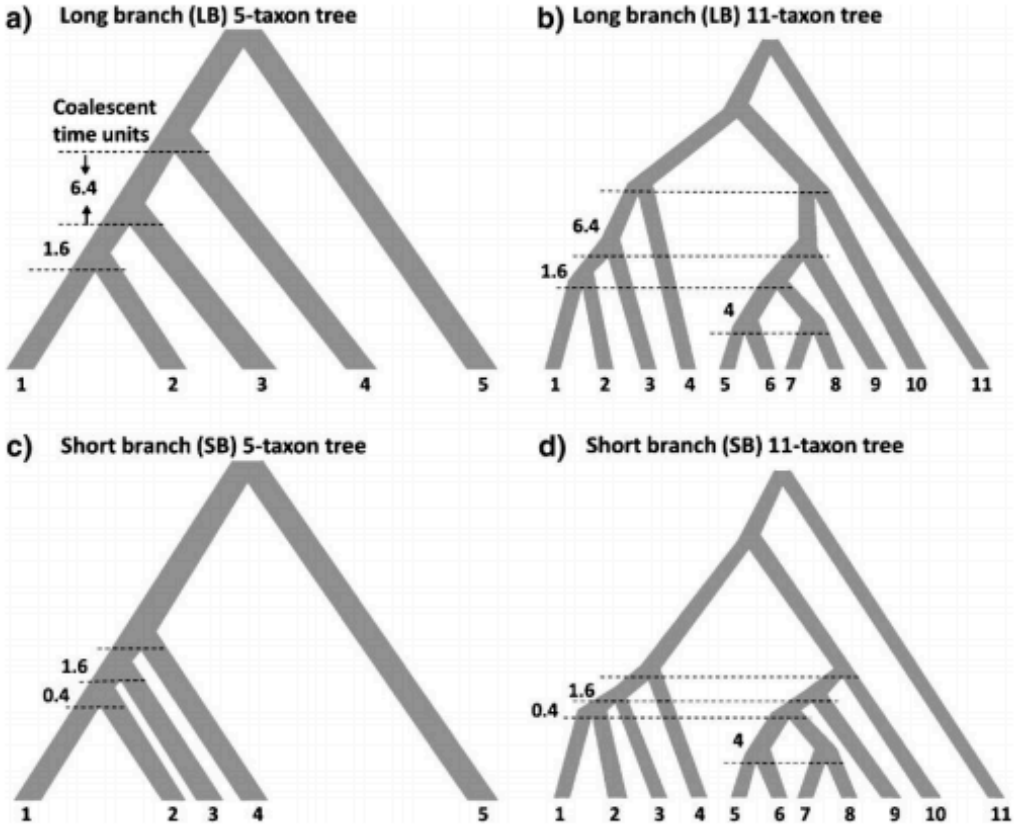
- Several biological processes can result in incongruent gene trees
- Some methods assume ILS is the only cause of gene tree discordance
- HGT plays an extensive role in bacterial evolution

Goals of this study

1. Thoroughly compare BEST and BUCKy on simulation data
2. Assess the accuracy of BUCKy at estimating concordance fatcors (CFs)
   - CF: the proportion of genes that truly have a given clade in their trees (distinct from statistical supports like bootstrap values)
3. Can BUCKy be used to test whether the coalescent model is an adequate explanation of gene tree discordance (even when HGT is the primary source of discordance)?

The Data

- 10, 50, or 100 unlinked gene trees were generated along the species trees



The Data

- Dataset 1: Only ILS
- Dataset 2: ILS + HGT
  - 0.5 HGT events per gene
  - Control for HGT by forcing exactly 70% of gene topologies to be unaffected by HGT events
  - Genomic rate change events simulated an average of three times per gene tree (modify branch lengths)
- Dataset 3: ILS + Uneven HGT

The Data

- ILS + HGT dataset has more discordant gene trees
- ILS + HGT gene trees are more different from the species tree

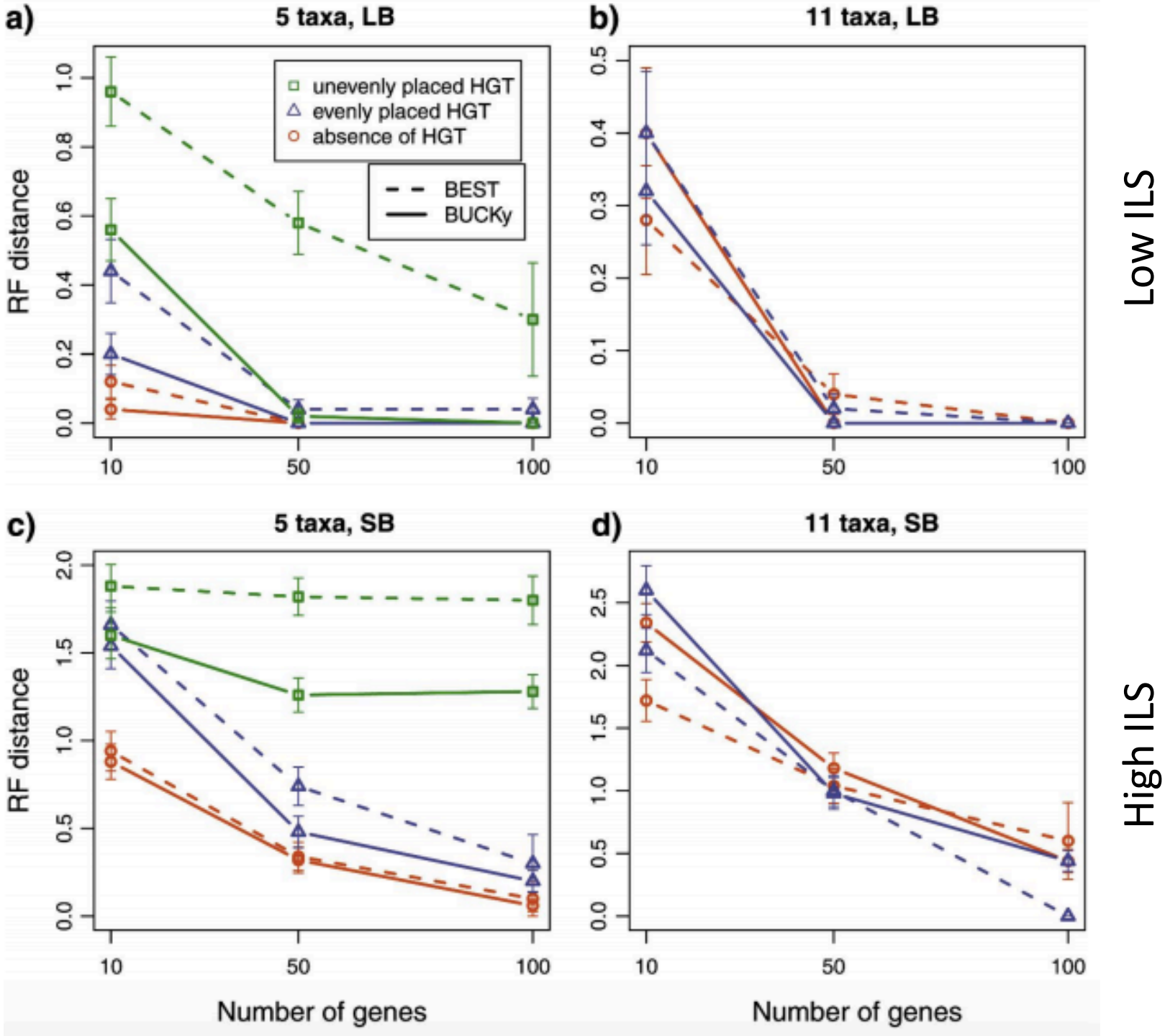
ILS	HGT	5-taxon case		11-taxon case	
		Proportion	RF distance	Proportion	RF distance
Weak (LB)	No	0.14	1.98	0.28	2.19
	Yes	0.32	2.67	0.49	4.32
Strong (SB)	No	0.54	2.34	0.79	3.26
	Yes	0.66	2.73	0.84	4.41

Methods

- BEST
  - Jukes-Cantor model of evolution
  - Accounts for gene tree discordance
  - Estimates species tree under coalescent-only model
- BUCKy
  - Jukes-Cantor model of evolution
  - Accounts for gene tree discordance
  - Estimates a concordance tree made of clades supported by the largest proportions of gene trees

Comparison of BEST and BUCKy

- *Hypothesis: BUCKy will perform better than BEST in the presence of HGT*
  - BEST assumes ILS only
  - BUCKy makes not assumptions


a)
5 taxa, LB
b)
11 taxa, LB
c)
5 taxa, SB
d)
11 taxa, SB
unevenly placed HGT
evenly placed HGT
absence of HGT
BEST
BUCKy
RF distance
Number of genes
Low ILS
High ILS

Comparison of BEST and BUCKy

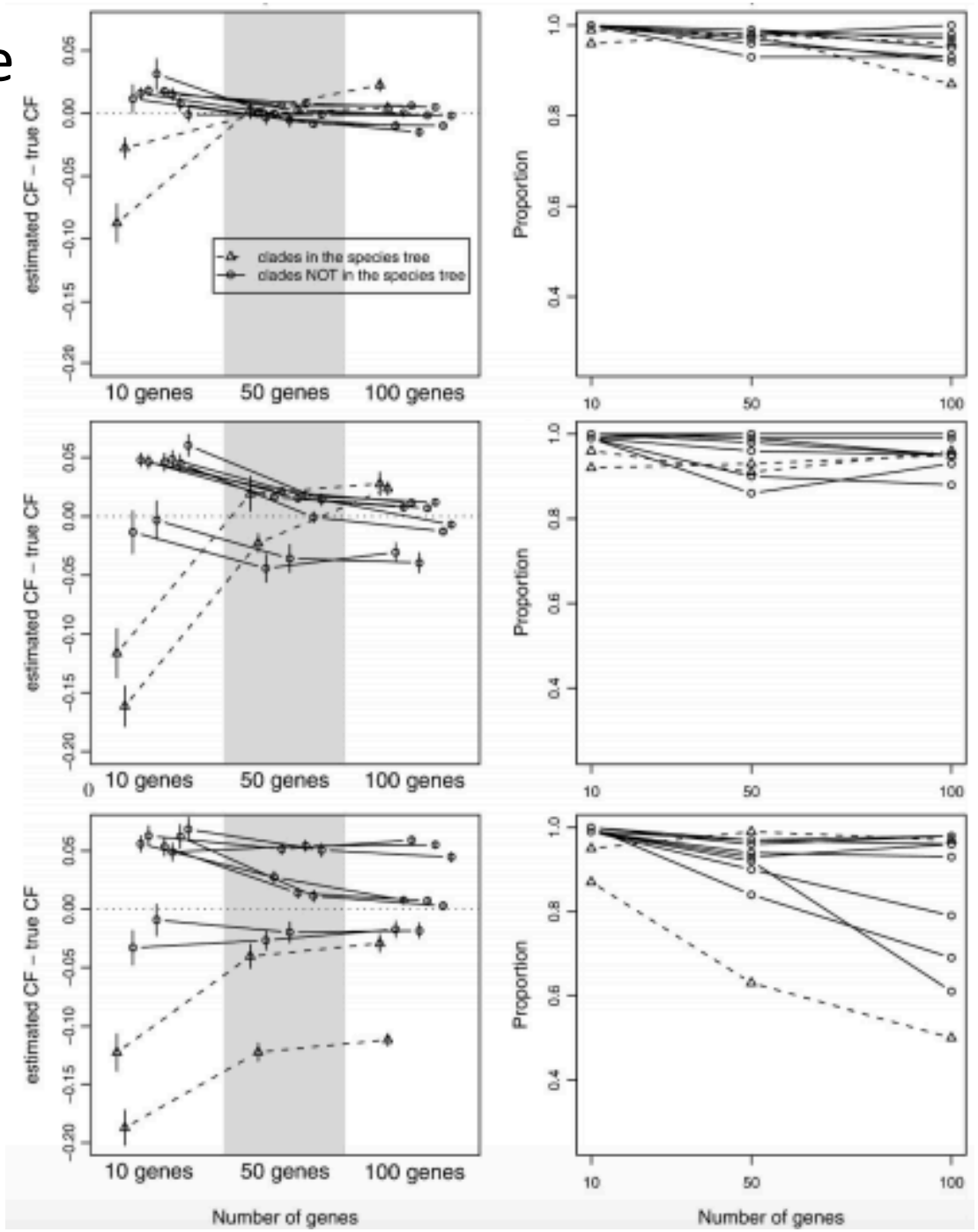
- *Hypothesis: BUCKy will perform better than BEST in the presence of HGT*
- BUCKy performed better than BEST under unevenly distributed HGT
- BEST performed better than BUCKy under high ILS for the 11-taxon tree (with and without HGT)
- Both methods are highly accurate with many genes in most cases

Accuracy of estimating CFs in BUCKy

- $CF_{estimated} - CF_{true}$
- Proportion of times that the 95% credibility intervals included the true CFs

Tree	Clades in the true species tree	Clades not in the true species tree
5-taxon tree	12; 1–3	13; 14; 15; 23; 24; 25; 34; 35
11-taxon tree	12; 1–3; 1–4; 56; 78; 5–8; 5–9	13; 23; 14; 24; 1–4,11; 1–4,9,11; 59; 69; 79; 89; 9,10; 569; 789; 56,10; 78,10

5-taxon tree


estimated CF – true CF
clades in the species tree
clades NOT in the species tree
10 genes
50 genes
100 genes
Proportion
Number of genes
LB, ILS + HGT
SB, ILS
SB, ILS + HGT

Accuracy of estimating CFs in BUCKy

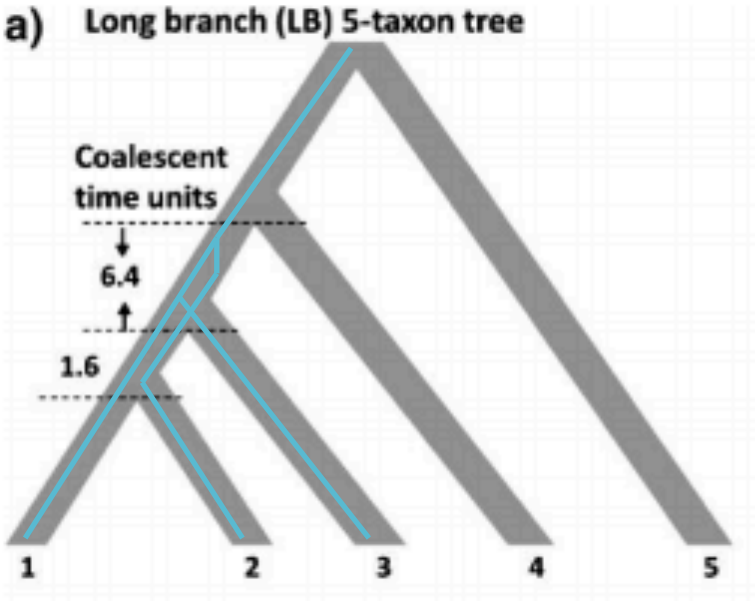
- CF accuracy decreases with increasing tree discordance
- Estimated CFs become more accurate as more genes are used
- CFs are more accurate for longer genes

Adequacy of the Coalescent Model

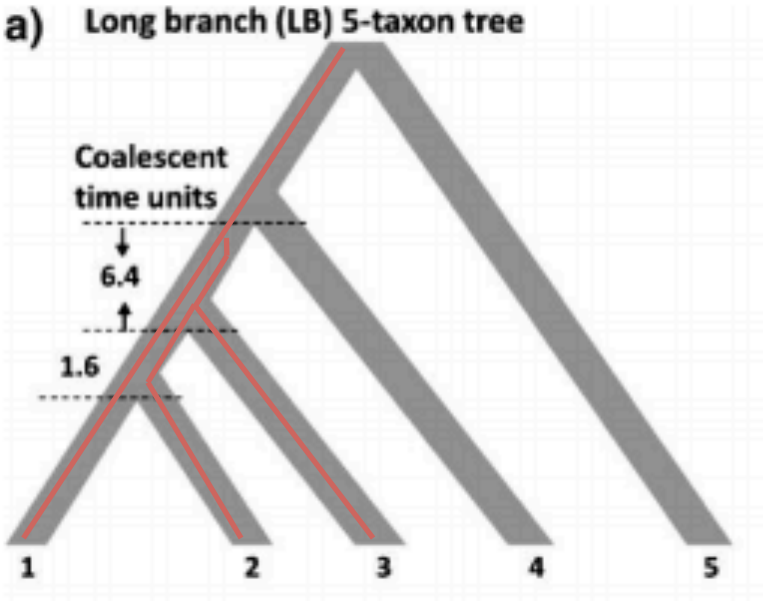
- Is the coalescent model alone able to explain gene tree discordance?
- Test the **symmetric signature of the coalescent**

Adequacy of the Coalescent Model

- CFs of **minor clades** from true sister taxa (*i.e.* taxa 1 and 2 below) should be equal when ILS is only factor causing tree discordance



Gene tree: 13|2



Gene tree: 23|1

$$CF_{13} = CF_{23}$$

Adequacy of the Coalescent Model

CFs used for testing	Number of genes	No HGT		Evenly distributed HGT		Unevenly distributed HGT	
		Weak ILS	Strong ILS	Weak ILS	Strong ILS	Weak ILS	Strong ILS
Genome wide	10	0	0	0	0	0	0
	50	0	0.01	0.02	0.02 (0.01)	0.04	0.05 (0.04)
	100	0	0.04	0	0.03 (0.02)	0.99	0.10 (0.09)
Sample wide	10	0	0.01	0.46 (0.45)	0.15 (0.14)	0.8	0.12
	50	0.02	0.04 (0.03)	0.52	0.18 (0.17)	1	0.54
	100	0 (0.03)	0.07	0.43	0.15 (0.13)	1	0.58 (0.52)

Notes: When the test based on the inferred species tree resulted in a different proportion of rejections, that proportion is indicated in parenthesis. Each value is based on 100 replicate data sets. In the absence of HGT, the values indicate the Type I error rate. In the presence of HGT, the values indicate the power of the test.

- Proportion of rejections of null hypothesis (symmetric signature of the coalescent not satisfied)

Adequacy of the Coalescent Model

- In absence of HGT, symmetric signature of the coalescent is present → accept null hypothesis
- In presence of symmetric HGT, symmetric signature is maintained → accept null hypothesis
- In present of uneven HGT, symmetric signature is lost → reject null hypothesis

Scalability of Bayesian Methods

- Running times should be considered
  - For largest dataset:
    - 30 days with BEST
    - 14.3 hours with BUCKy
  - Authors performed fewer replicates for largest datasets but with greater sampling

Summary

- BUCKy is most advantageous over BEST when HGT is unevenly distributed
  - Otherwise, equally powerful
  - Running time is greatly improved with BUCKy
- CF accuracy with BUCKy is dependent on gene tree discordance
- The coalescent model is not sufficient to explain gene tree discordance when uneven HGT is present