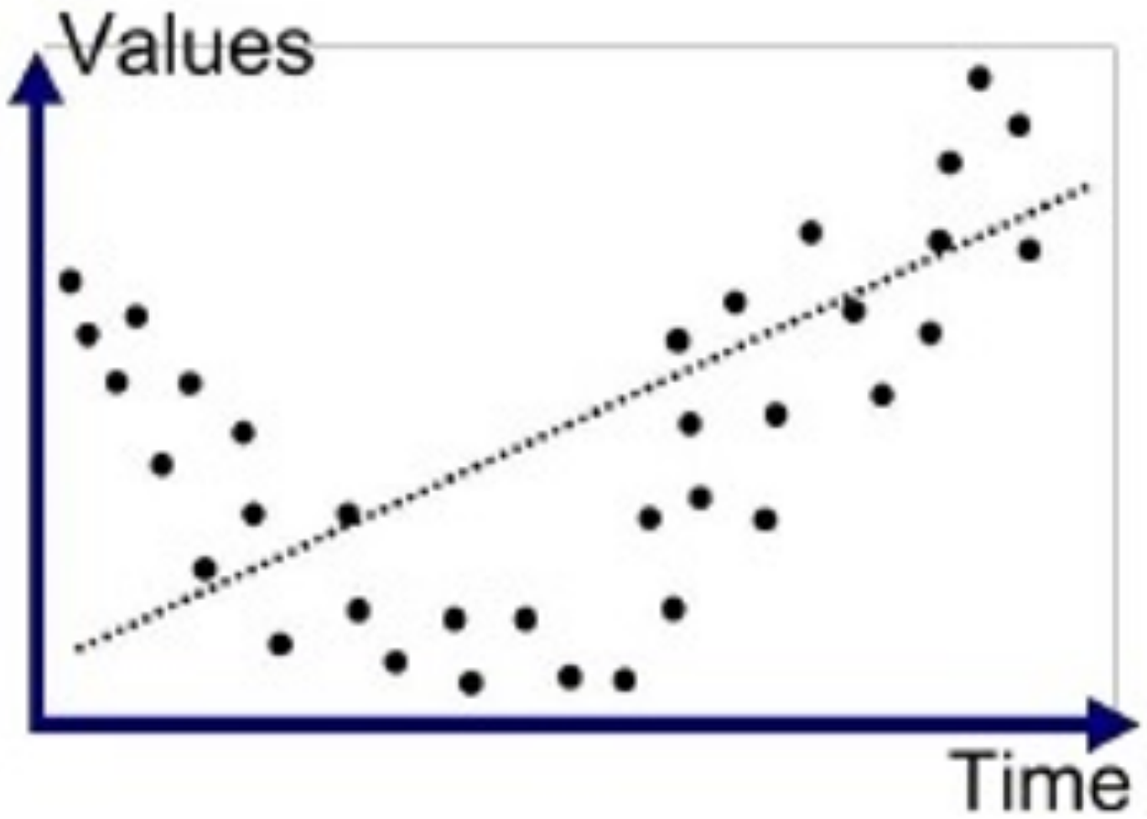
DBI - 1661029

Posterior Predictive Checks in Bayesian Phylogenetics

R package: P2C2M.SNAPP

Drew Duckett, Tara A Pelletier, Bryan C Carstens

Models matter!



Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology

John P. Huelsenbeck,[1]* Fredrik Ronquist,[2] Rasmus Nielsen,[3] Jonathan P. Bollback[1]

As a discipline, phylogenetics is becoming transformed by a flood of molecular data. These data allow broad questions to be asked about the history of life, but also present difficult statistical and computational problems. Bayesian inference of phylogeny brings a new perspective to a number of outstanding issues in evolutionary biology, including the analysis of large phylogenetic trees and complex evolutionary models and the detection of the footprint of natural selection in DNA sequences.

ity of a tree (Fig. 1). Bayes's theorem

$$\Pr[\text{Tree} \mid \text{Data}] = \frac{\Pr[\text{Data} \mid \text{Tree}] \times \Pr[\text{Tree}]}{\Pr[\text{Data}]}$$

(where the vertical bar should be read as "given") is used to combine the prior probability of a phylogeny (Pr[Tree]) with the likelihood (Pr[Data | Tree]) to produce a posterior probability distribution on trees (Pr[Tree | Data]). The posterior probability of a tree

SCIENCE'S COMPASS • REVIEW

REVIEW: EVOLUTION

Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology

John P. Huelsenbeck,[1]* Fredrik Ronquist,[2] Rasmus Nielsen,[3] Jonathan P. Bollback[1]

As a discipline, phylogenetics is becoming transformed by a flood of molecular data. These data allow broad questions to be asked about the history of life, but also present difficult statistical and computational problems. Bayesian inference of phylogeny brings a new perspective to a number of outstanding issues in evolutionary biology, including the analysis of large phylogenetic trees and complex evolutionary models and the detection of the footprint of natural selection in DNA sequences.

ity of a tree (Fig. 1). Bayes's theorem

$$\Pr[\text{Tree} \mid \text{Data}] = \frac{\Pr[\text{Data} \mid \text{Tree}] \times \Pr[\text{Tree}]}{\Pr[\text{Data}]}$$

(where the vertical bar should be read as "given") is used to combine the prior probability of a phylogeny (Pr[Tree]) with the likelihood (Pr[Data | Tree]) to produce a posterior probability distribution on trees (Pr[Tree | Data]). The posterior probability of a tree

"The posterior probability of a tree can be interpreted as the probability that the tree is correct", **given the model**.

Multi-species coalescent model (MSCM)

- developed to estimate species trees, while accounting for the coalescent process that can lead to incongruence among gene trees (due to **incomplete lineage sorting**)

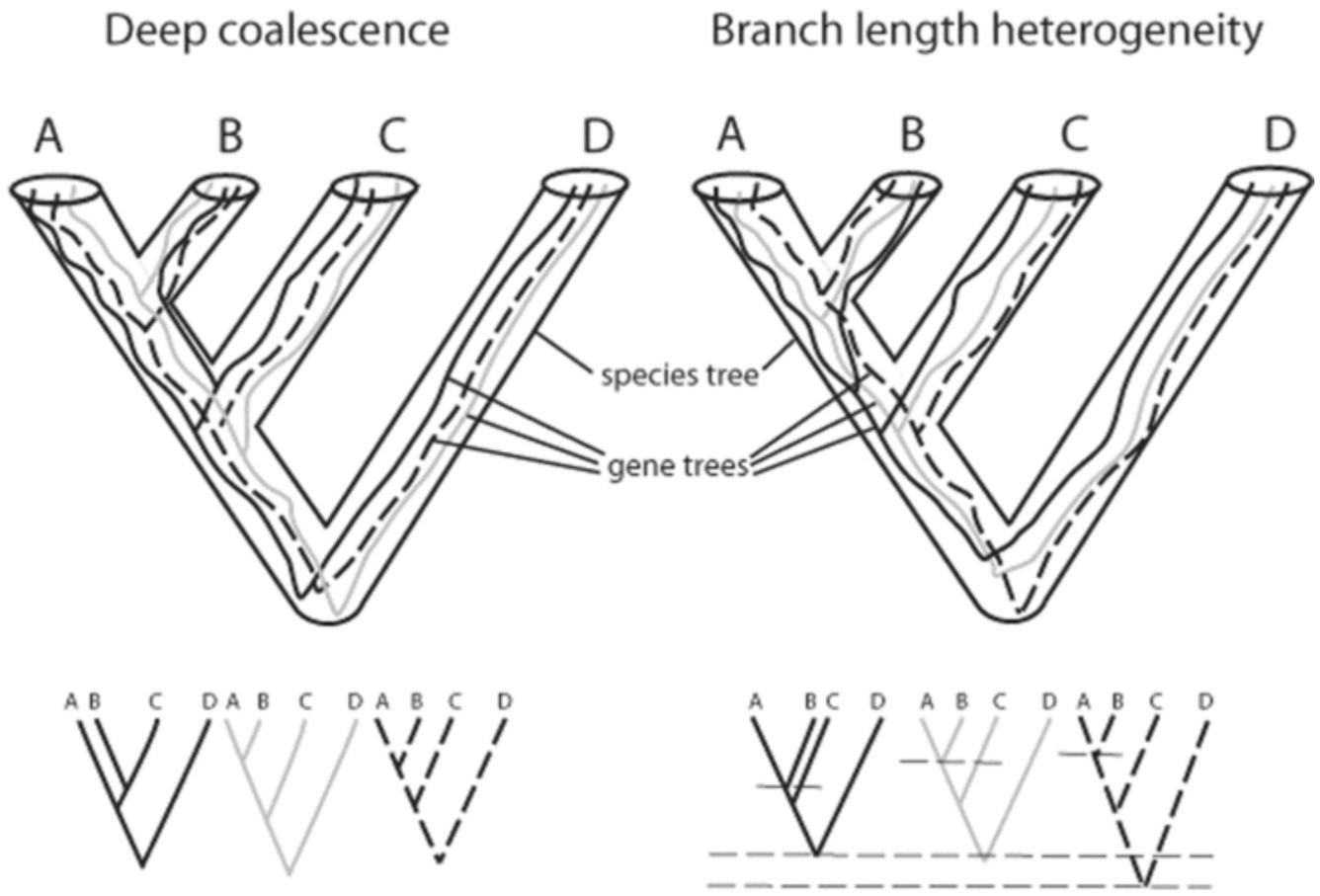
Multi-species coalescent model (MSCM)

Deep coalescence

Branch length heterogeneity

Evolution, Volume: 63, Issue: 1, Pages: 1-19, First published: 14 January 2009, DOI: (10.1111/j.1558-5646.2008.00549.x)

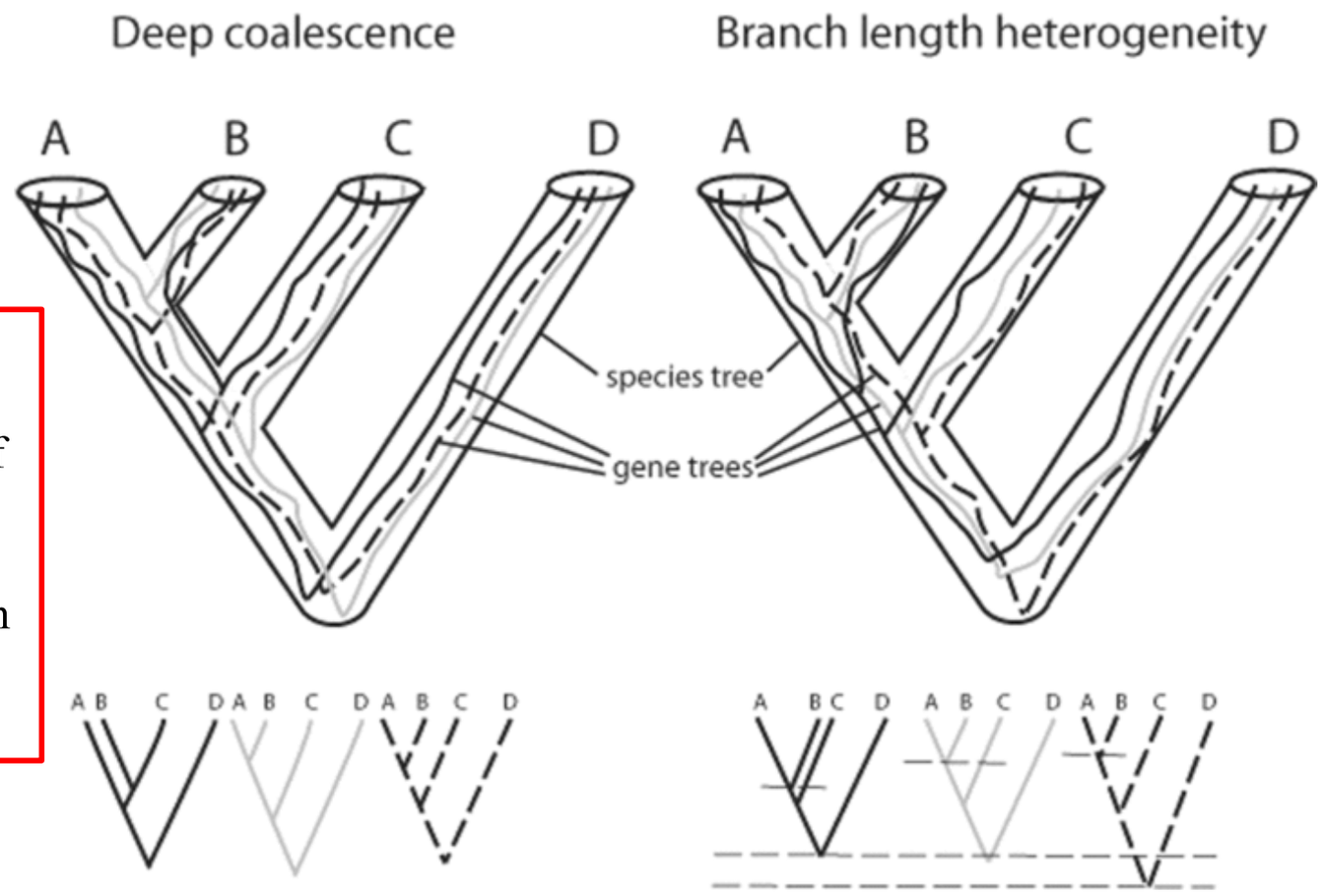
Multi-species coalescent model (MSCM)



Evolution, Volume: 63, Issue: 1, Pages: 1-19, First published: 14 January 2009, DOI: (10.1111/j.1558-5646.2008.00549.x)

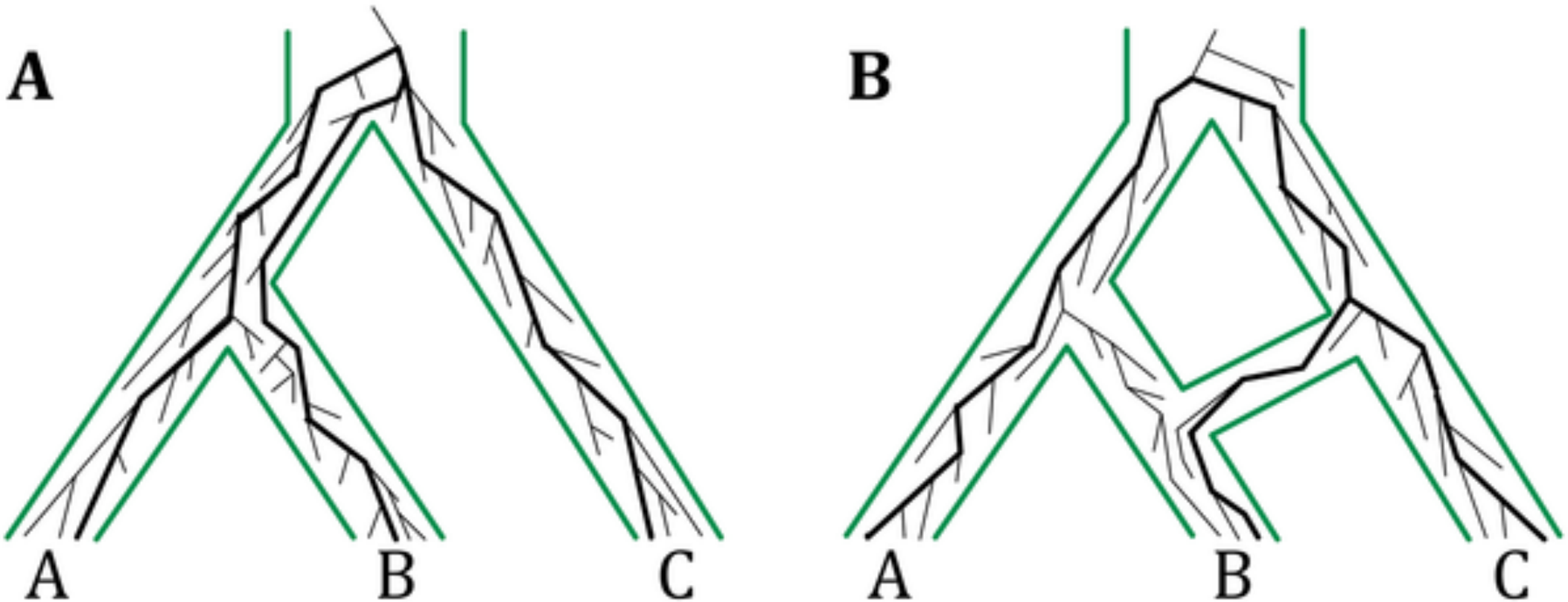
Multi-species coalescent model (MSCM)

Given a species tree, we can calculate the probability distribution of gene trees, then use this information to estimate the best species tree given our data under the coalescent model.



Evolution, Volume: 63, Issue: 1, Pages: 1-19, First published: 14 January 2009, DOI: (10.1111/j.1558-5646.2008.00549.x)

Multi-species coalescent model (MSCM)



Wen D, Yu Y, Nakhleh L (2016) Bayesian Inference of Reticulate Phylogenies under the Multispecies Network Coalescent. PLOS Genetics 12(5): e1006006. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006006

Syst. Biol. 63(1):17–30, 2014

DOI:10.1093/sysbio/syt049
Advance Access publication August 13, 2013

The Influence of Gene Flow on Species Tree Estimation: A Simulation Study

ADAM D. LEACHÉ[1,*], REBECCA B. HARRIS[1], BRUCE RANNALA[2,3], AND ZIHENG YANG[3,4]

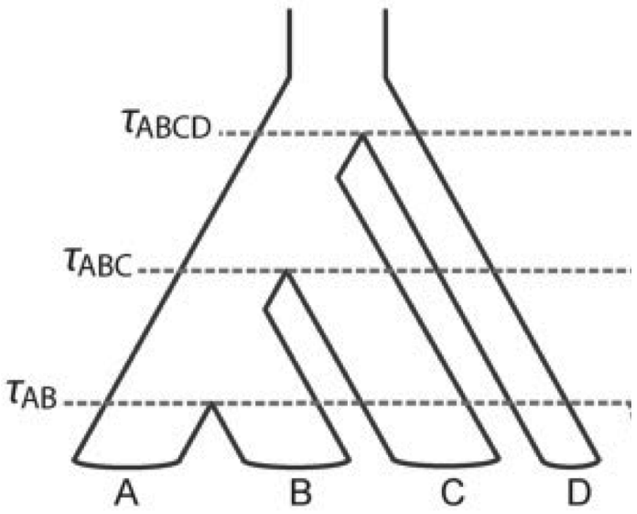
Syst. Biol. 63(1):17–30, 2014

DOI:10.1093/sysbio/syt049
Advance Access publication August 13, 2013

The Influence of Gene Flow on Species Tree Estimation: A Simulation Study

ADAM D. LEACHÉ[1,*], REBECCA B. HARRIS[1], BRUCE RANNALA[2,3], AND ZIHENG YANG[3,4]

Starting species tree



Isolation-migration

Paraphyletic gene flow

Ancestral gene flow

Syst. Biol. 63(1):17–30, 2014

DOI:10.1093/sysbio/syt049
Advance Access publication August 13, 2013

The Influence of Gene Flow on Species Tree Estimation: A Simulation Study

ADAM D. LEACHÉ[1,*], REBECCA B. HARRIS[1], BRUCE RANNALA[2,3], AND ZIHENG YANG[3,4]

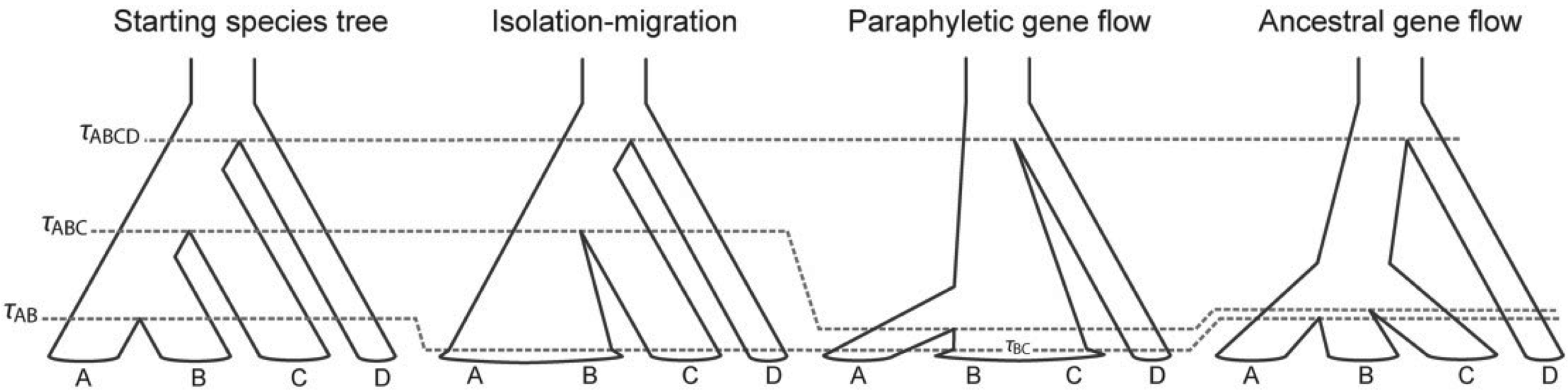


Figure 8 Species tree distortions caused by gene flow that can result from coalescent methods that only model ILS. Dashed lines illustrate species tree compression, and the widening of branches illustrates species tree dilation in relation to the starting species tree.

Syst. Biol. 63(3):322–333, 2014

DOI:10.1093/sysbio/syt057
Advance Access publication August 28, 2013

Poor Fit to the Multispecies Coalescent is Widely Detectable in Empirical Data

NOAH M. REID[1,*], SARAH M. HIRD[1], JEREMY M. BROWN[1], TARA A. PELLETIER[2], JOHN D. MCVAY[1], JORDAN D. SATLER[2], AND BRYAN C. CARSTENS[2]

Syst. Biol. 67(2):269–284, 2018

DOI:10.1093/sysbio/syx073
Advance Access publication September 4, 2017

Impact of Model Violations on the Inference of Species Boundaries Under the Multispecies Coalescent

ANTHONY J. BARLEY[1,*], JEREMY M. BROWN[2], AND ROBERT C. THOMSON[1]

Posterior predictive simulation (PPS)

- We can approximate the **posterior predictive distribution** of a model by **simulating new observations** from parameter values sampled from the posterior distribution of your Bayesian phylogenetic analysis.

Posterior predictive simulation (PPS)

- We can approximate the **posterior predictive distribution** of a model by **simulating new observations** from parameter values sampled from the posterior distribution of your Bayesian phylogenetic analysis.

- If an evolutionary model is a good fit to your data (i.e., it does a good job of explaining patterns in the DNA), then data simulated under that model (**PPD**) should be similar to the **empirical data**.

Posterior predictive simulation (PPS)

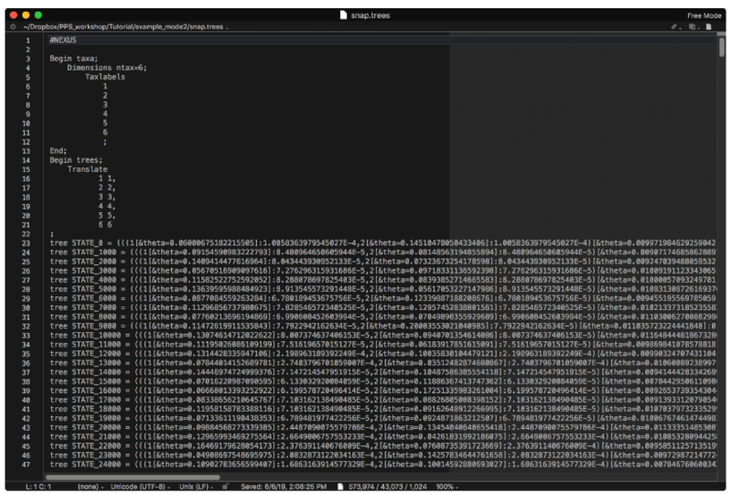
- We can approximate the **posterior predictive distribution** of a model by **simulating new observations** from parameter values sampled from the posterior distribution of your Bayesian phylogenetic analysis.

- If an evolutionary model is a good fit to your data (i.e., it does a good job of explaining patterns in the DNA), then data simulated under that model (**PPD**) should be similar to the **empirical data**.

- We can ask: does a particular model adequately describe an individual empirical data set?

Posterior predictive simulation (PPS)

1) Sample tree from the posterior distribution

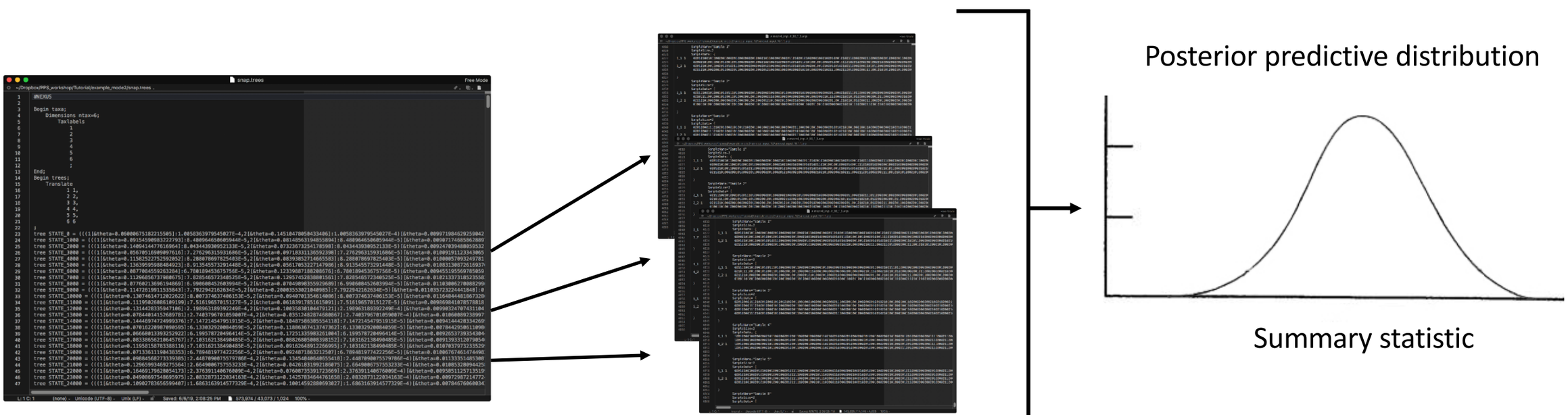


Posterior predictive simulation (PPS)

1) Sample tree from the posterior distribution
2) Simulate data using this tree/parameters under the MSCM

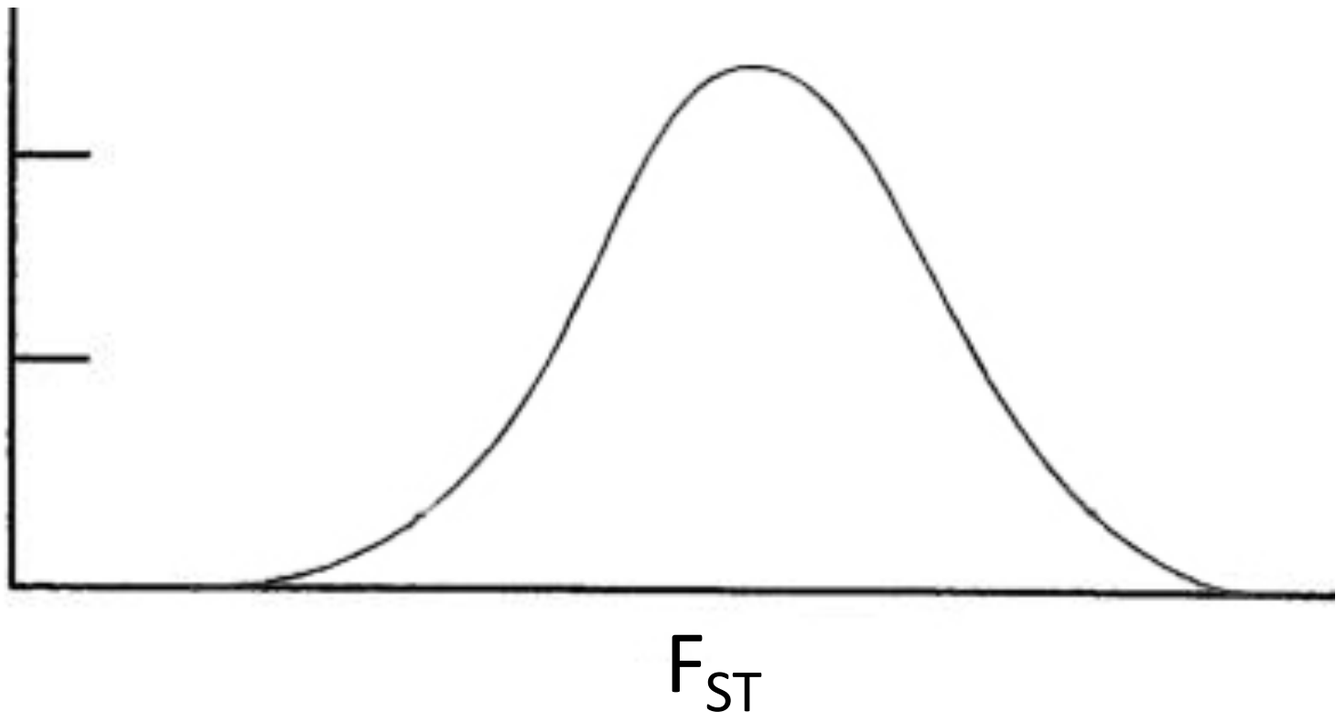
Posterior predictive simulation (PPS)

1) Sample tree from the posterior distribution
2) Simulate data using this tree/parameters under the MSCM
3) Summarize simulated datasets with test statistic



Posterior predictive distribution (PPD)

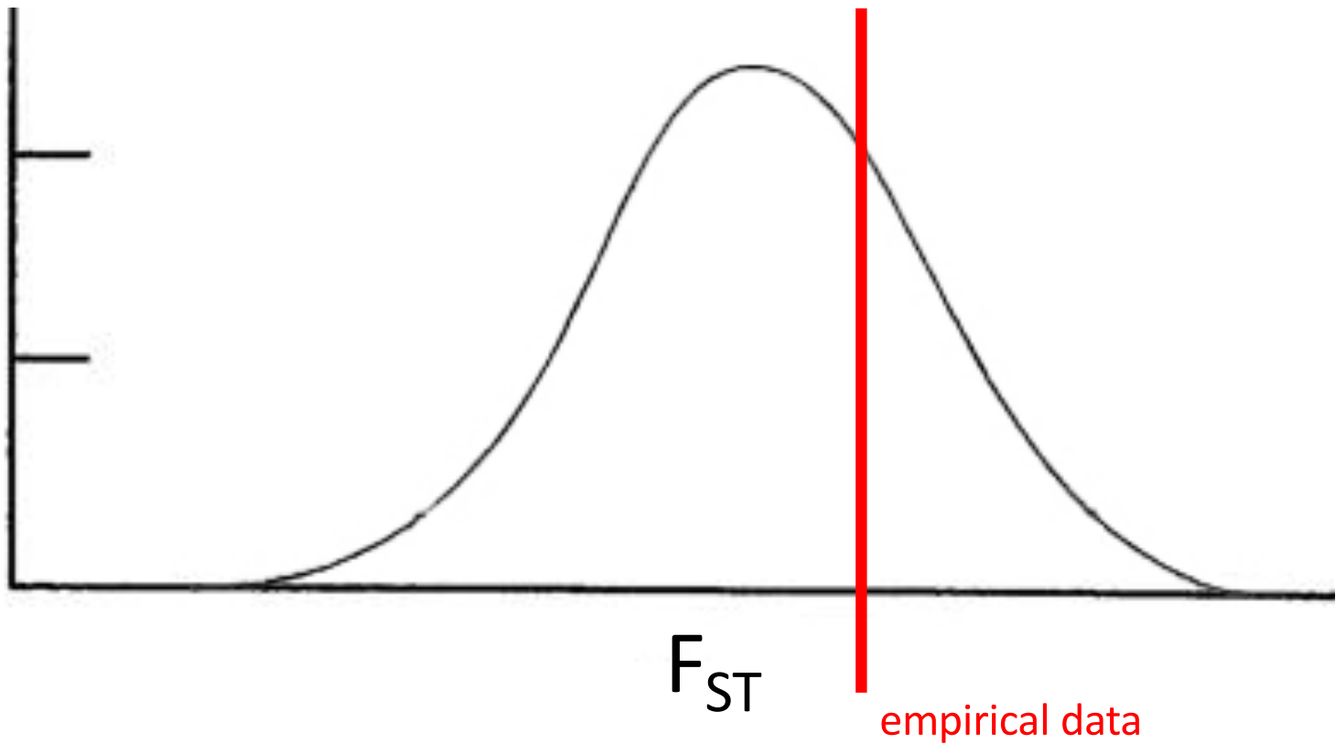
- How well does the empirical data fit this distribution?



This distribution is a representation of your data if the model were true – if the MSC were in fact generating your empirical data

Posterior predictive distribution (PPD)

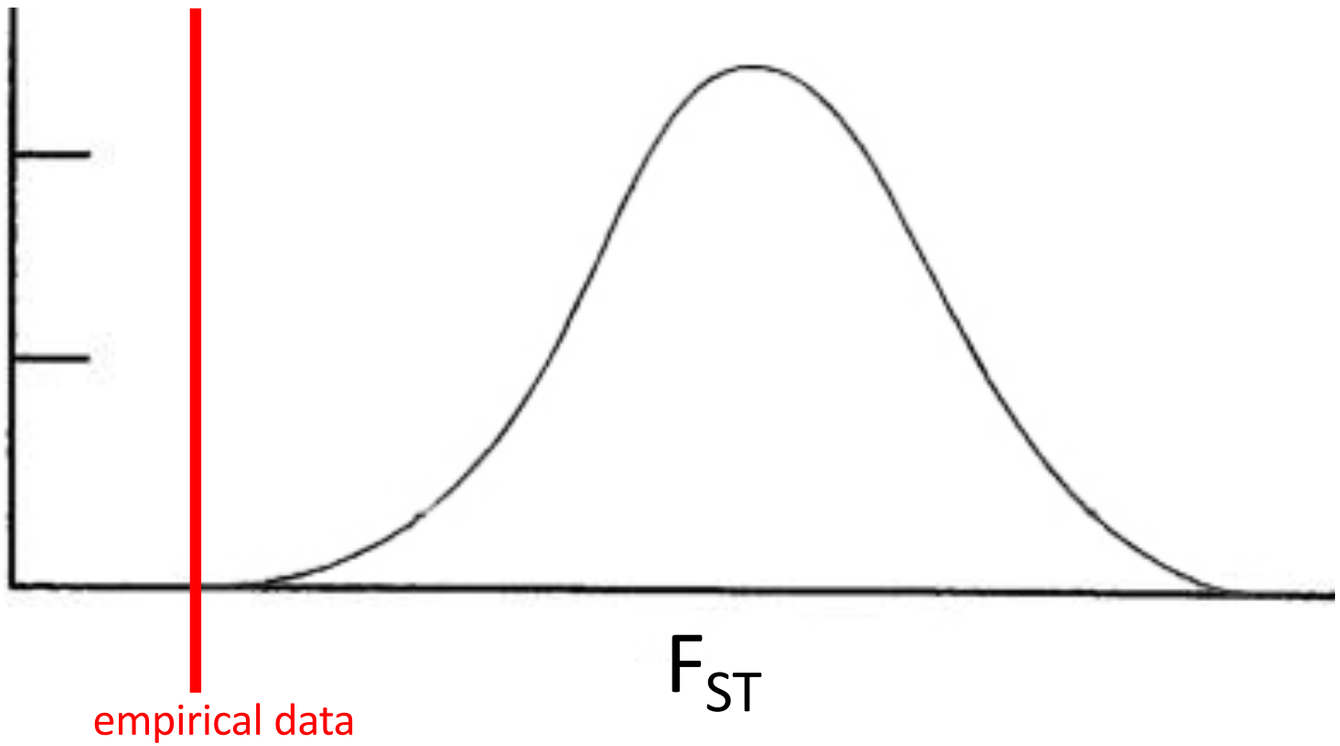
- How well does the empirical data fit this distribution?



This distribution is a representation of your data if the model were true – if the MSC were in fact generating your empirical data

Posterior predictive distribution (PPD)

- How well does the empirical data fit this distribution?



This distribution is a representation of your data if the model were true – if the MSC were in fact generating your empirical data

MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2016) **16**, 193–205 doi: 10.1111/1755-0998.12435

Posterior predictive checks of coalescent models: P2C2M, an R package

MICHAEL GRUENSTAEUDL,*‡ NOAH M. REID,† GREGORY L. WHEELER* and BRYAN C. CARSTENS*

**Department of Evolution, Ecology & Organismal Biology, Ohio State University, Columbus, OH 43210, USA†Department of Environmental Toxicology, University of California, Davis, CA 95616, USA*

*BEAST – multi-locus sequence data

SNAPP (SNP and AFLP Phylogenies)

- Interfaces with the BEAST package
- Algorithm bypasses the gene trees and computes species tree likelihoods directly from the markers
- Returns a sample of species trees with (relative) divergence times and population sizes (posterior distribution)

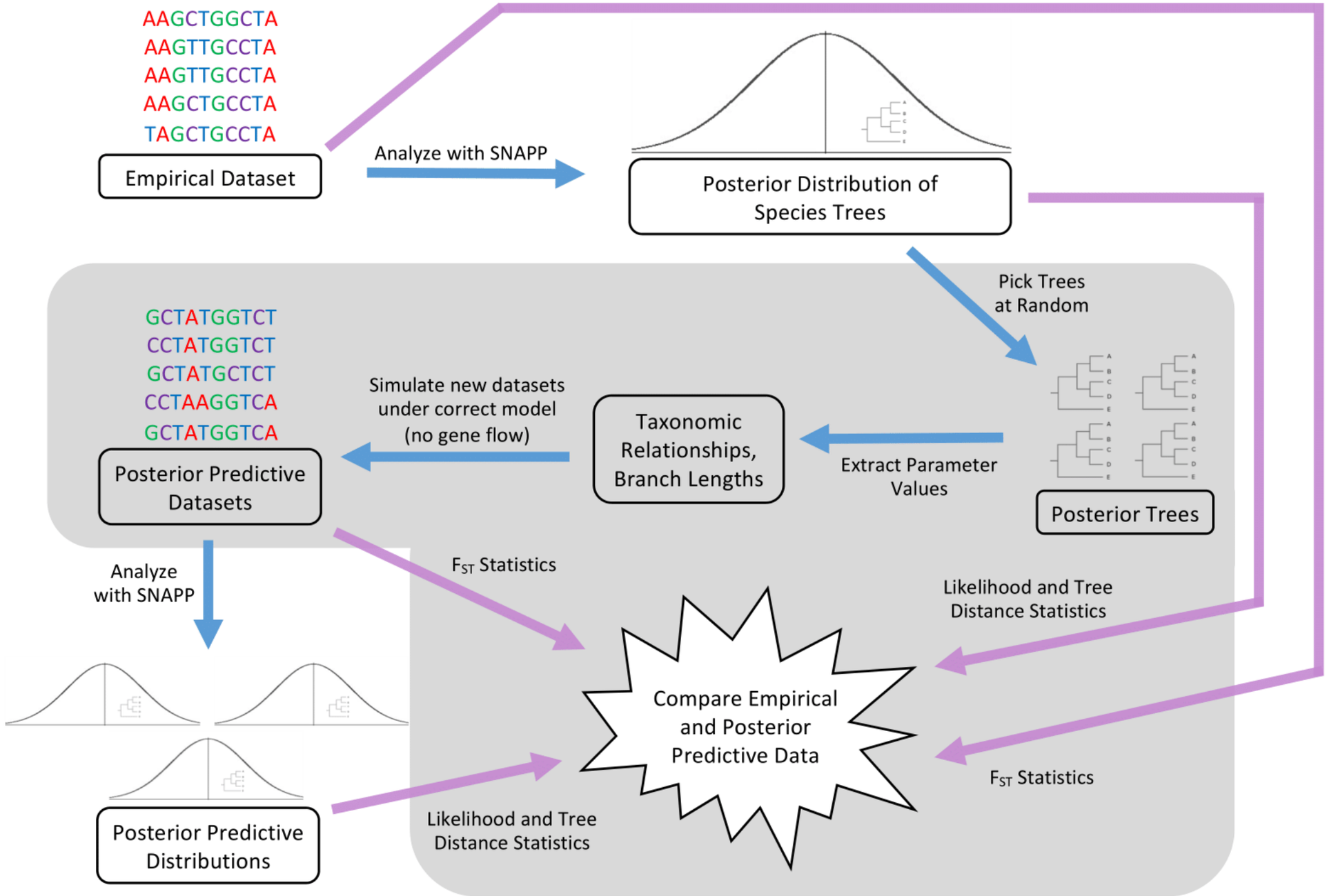
SNAPP assumptions

- Those of the coalescent process (shared polymorphism is due to ILS)
- Each marker is a single biallelic character
- The genealogies for separate markers are conditionally independent (satisfied for SNPs that are well spaced along the genome)

R package: P2C2M.SNAPP

- Conducts posterior predictive checks on your analysis from the program SNAPP

- We are about to submit both this paper and the package to CRAN


AAGCTGGCTA
AAGTTGCCTA
AAGTTGCCTA
AAGCTGCCTA
TAGCTGCCTA
Empirical Dataset
Analyze with SNAPP
Posterior Distribution of Species Trees
Pick Trees at Random
Posterior Trees
Extract Parameter Values
Taxonomic Relationships, Branch Lengths
Simulate new datasets under correct model (no gene flow)
GCTATGGTCT
CCTATGGTCT
GCTATGCTCT
CCTAAGGTCA
GCTATGGTCA
Posterior Predictive Datasets
Analyze with SNAPP
Posterior Predictive Distributions
FST Statistics
Likelihood and Tree Distance Statistics
Compare Empirical and Posterior Predictive Data
Likelihood and Tree Distance Statistics
FST Statistics

P2C2M.SNAPP simulation testing

- **Which summary statistics work best?**

- We simulated two trees 100X: **MSCM** and **MSCM+*m***
  - 6 species (symmetrical tree)
  - 2 individuals per species
  - 2000 SNPs
  - N_e = 100,000
  - Speciation times at 5N, 10N, and 20N
  - When gene flow was incorporated into the model it happened at 2.5N generations in the past and *m* was drawn from a uniform prior between 0.5 and 5 migrants per generation

P2C2M.SNAPP simulation testing

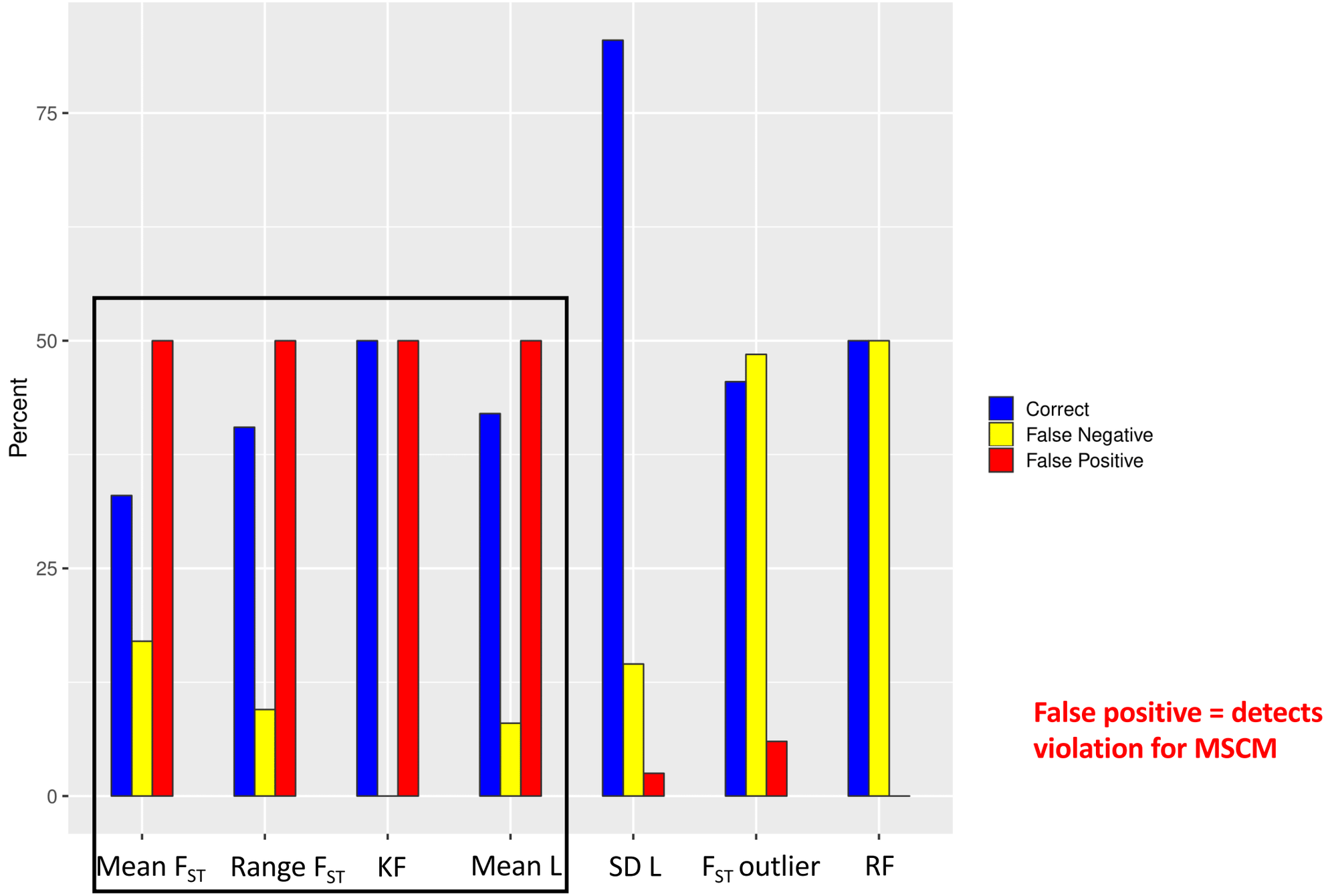
- **Summary statistics tested:**

  - Mean pairwise F_{ST}
  - Range of pairwise F_{ST}
  - F_{ST} outlier test
  - Robinson-Foulds distance (topological distance only)
  - Kuhner-Felsenstein distance (includes branch lengths)
  - Mean of tree likelihood
  - Standard deviation of tree likelihood

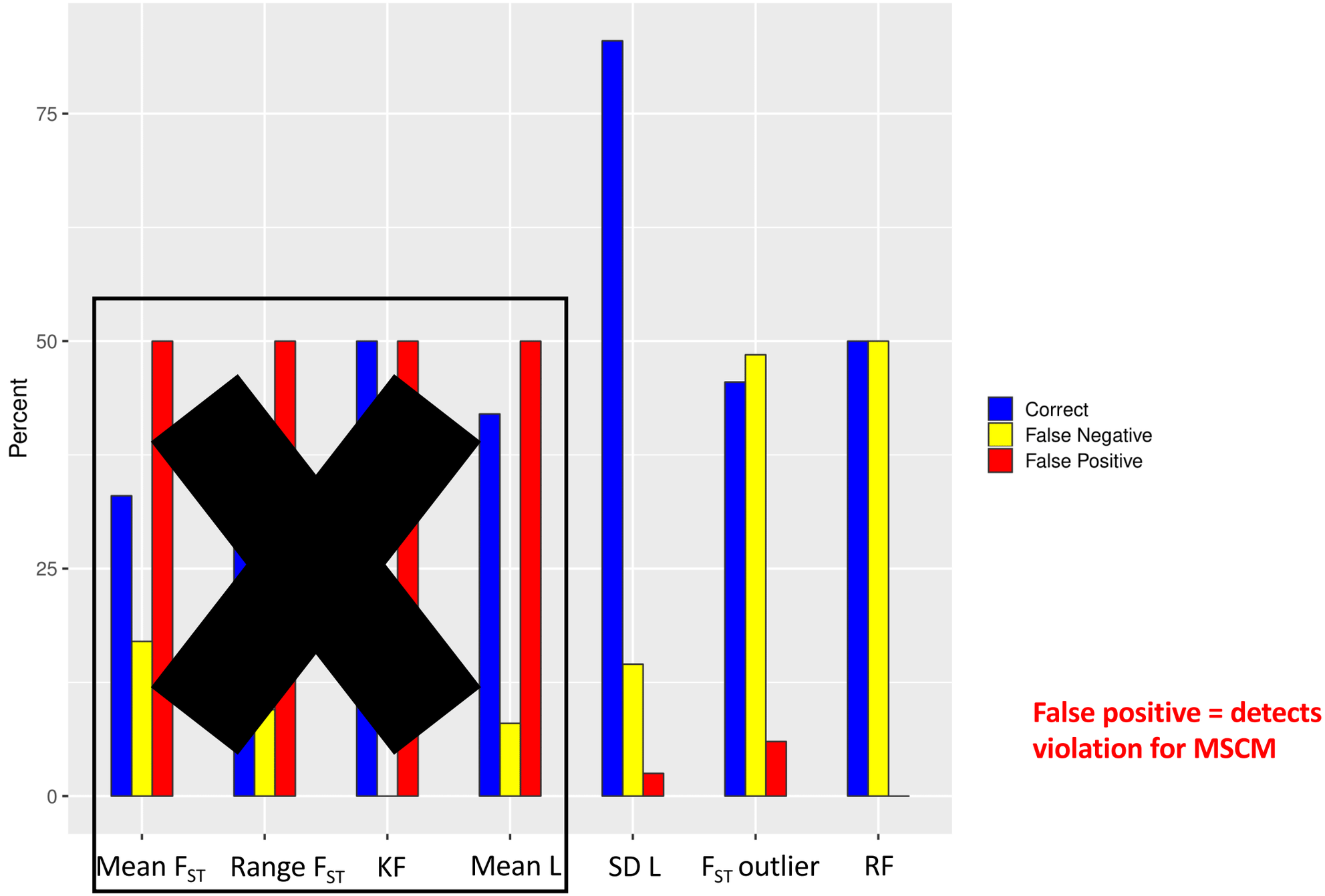
P2C2M.SNAPP simulation testing

- **Summary statistics tested:**

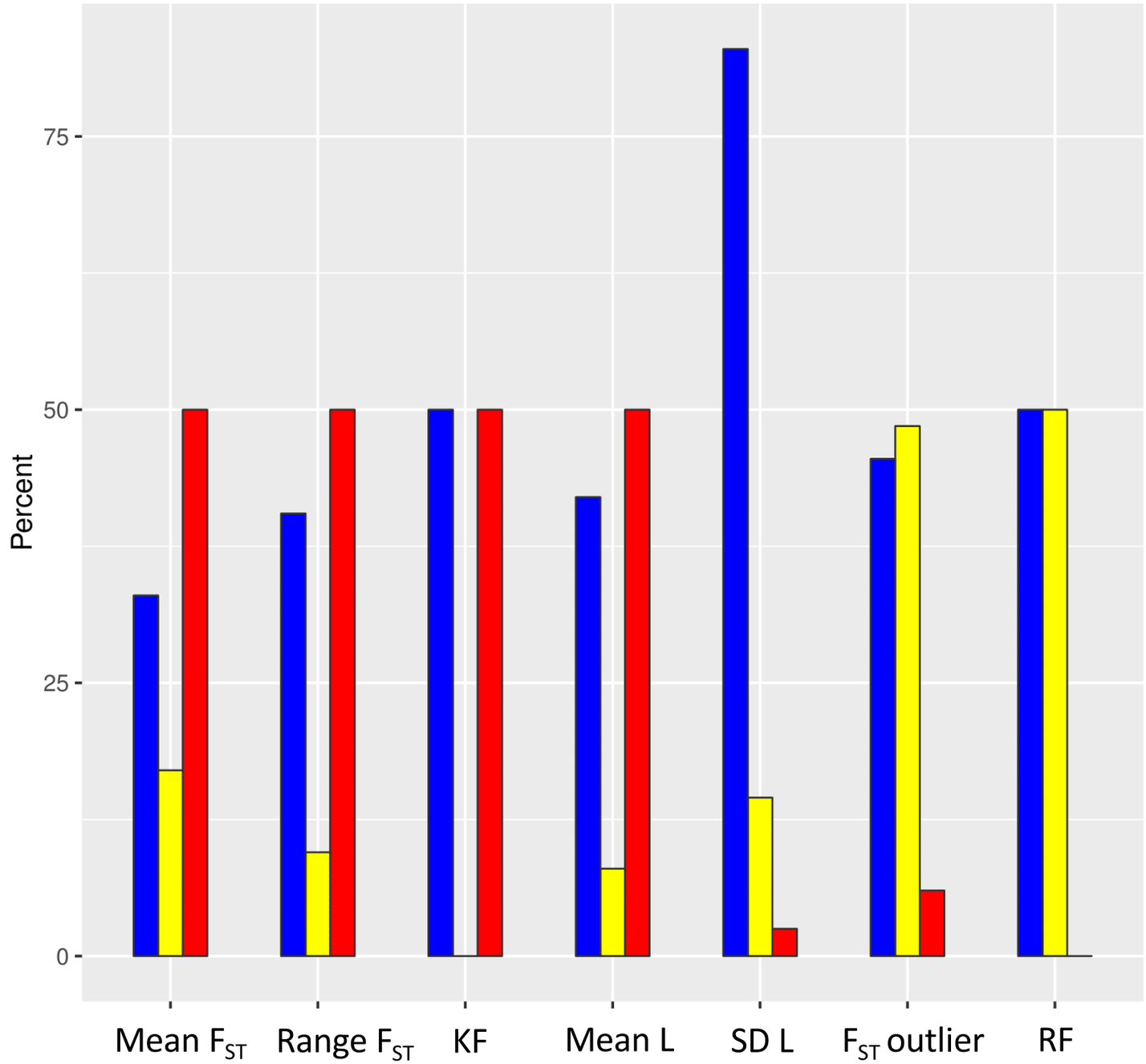
  - ~~Mean pairwise F_{ST}~~
  - ~~Range of pairwise F_{ST}~~
  - **F_{ST} outlier test**
  - **Robinson-Foulds distance (topological distance only)**
  - ~~Kuhner-Felsenstein distance (includes branch lengths)~~
  - ~~Mean of tree likelihood~~
  - **Standard deviation of tree likelihood**



False positive = detects violation for MSCM

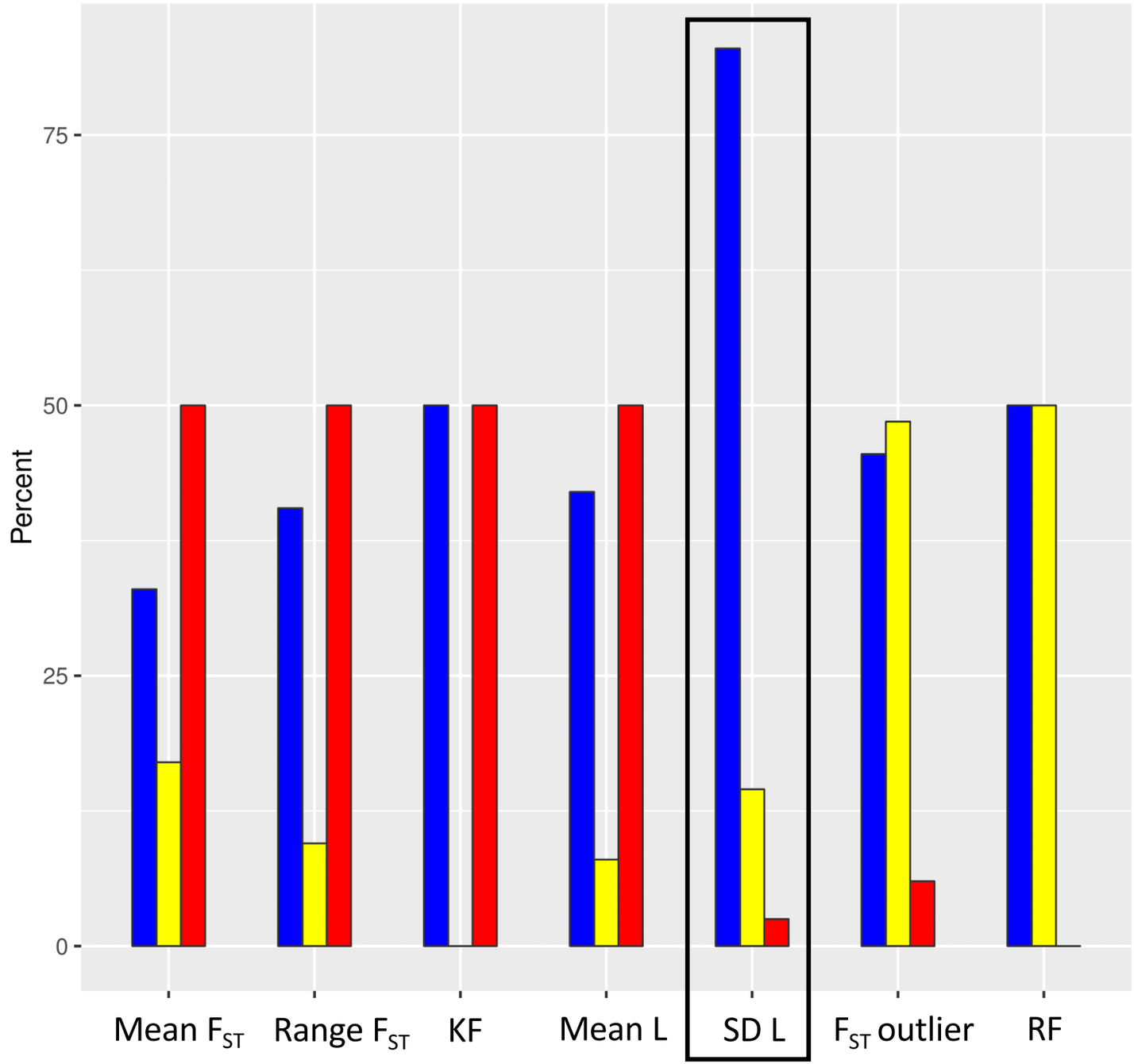


False positive = detects violation for MSCM



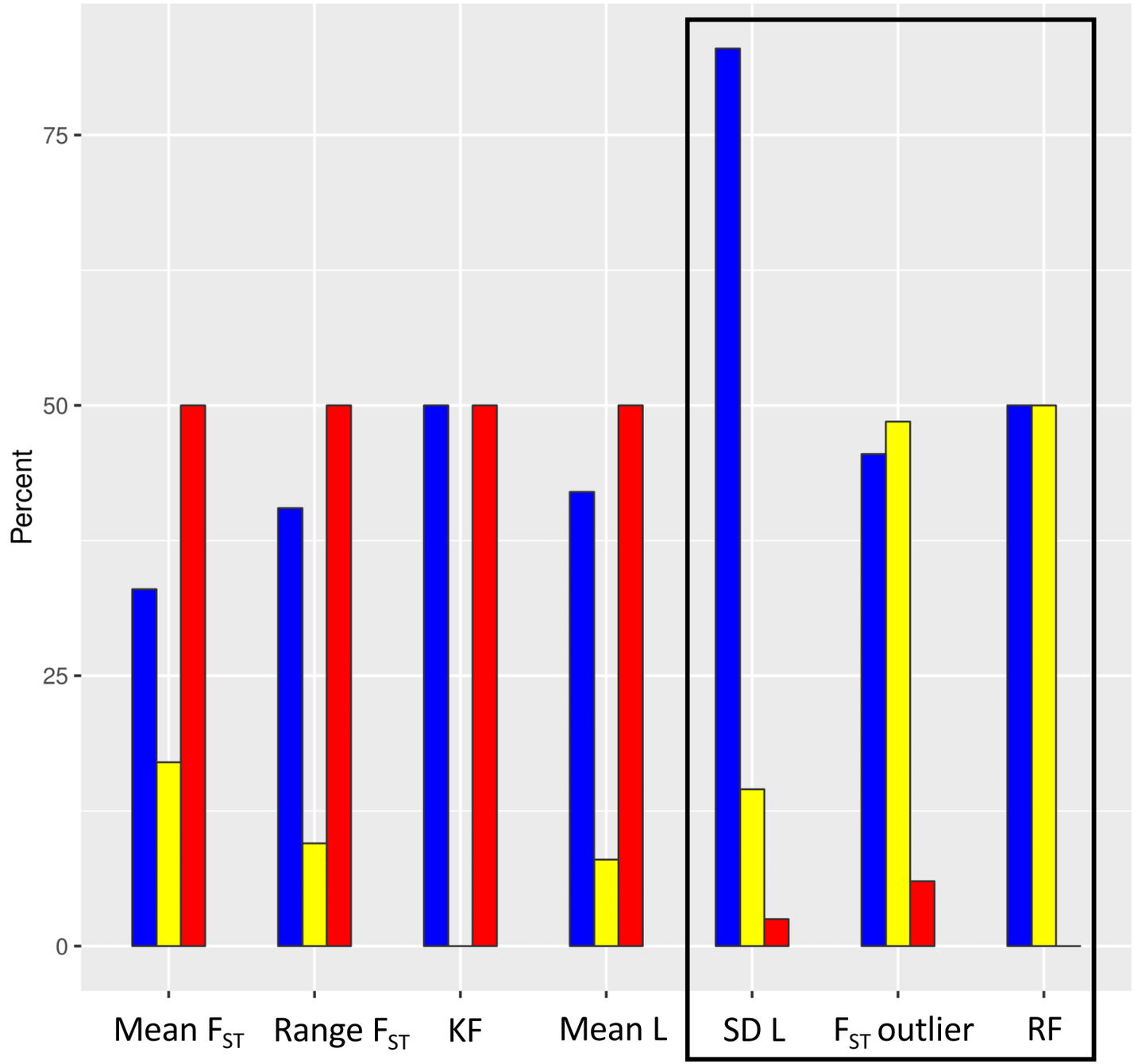
False negative = does not identify violation for MSCM+*m*

False positive = detects violation for MSCM



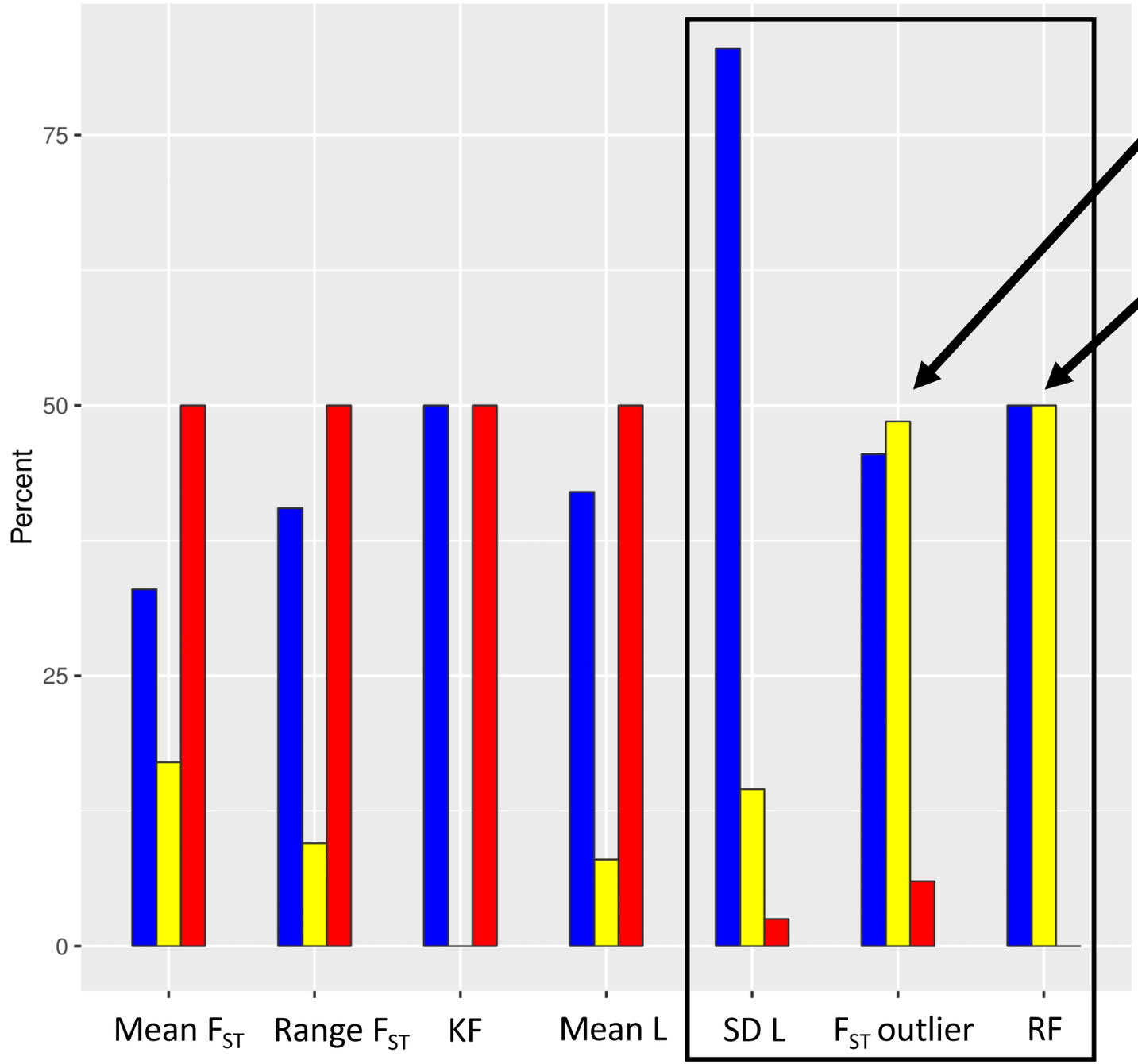
False negative = does not identify violation for MSCM+*m*

False positive = detects violation for MSCM



False negative = does not identify violation for MSCM+*m*

False positive = detects violation for MSCM



Might have more power if more OTUs are included

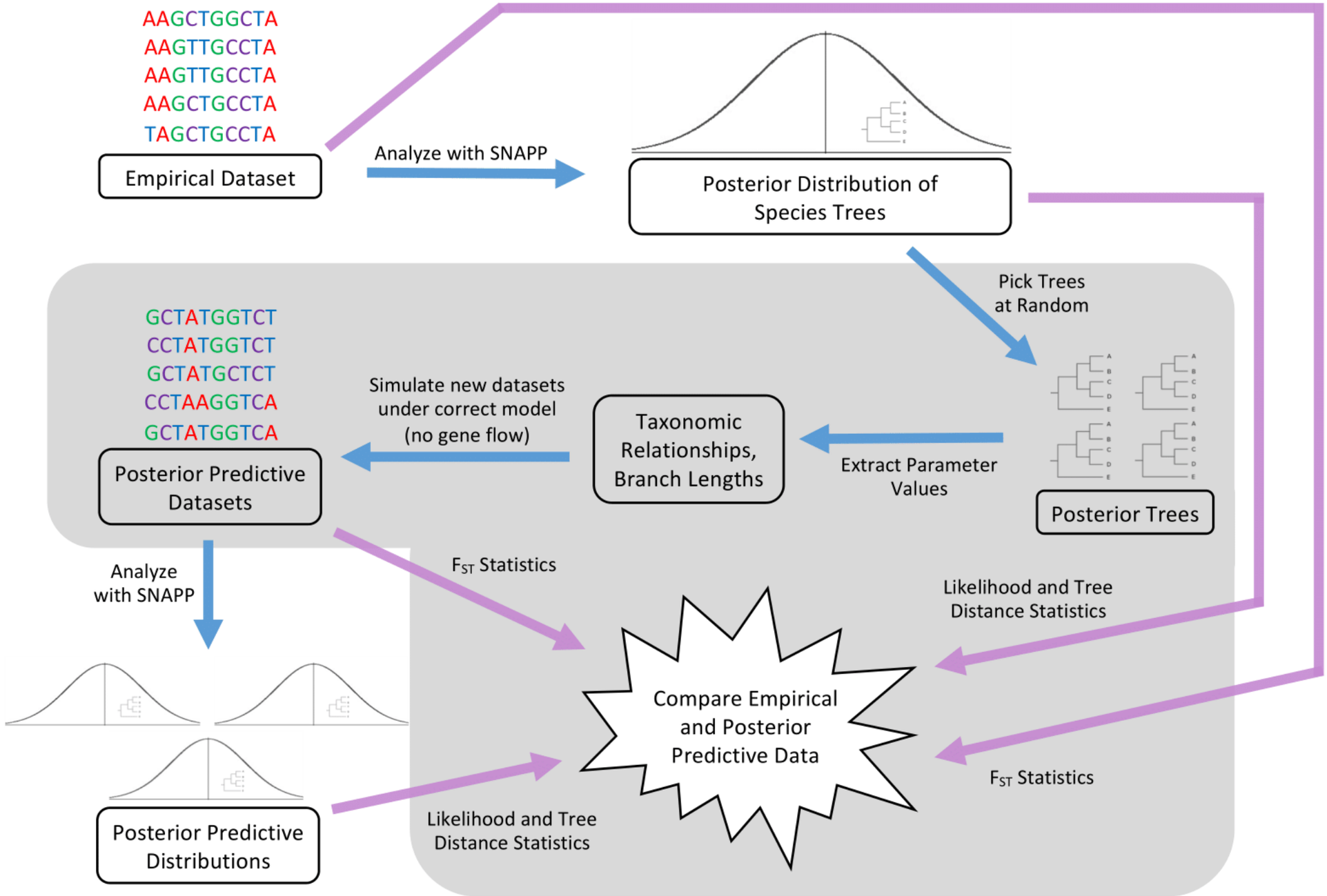
Although not statistically significant at 0.05, distances greater than 0 were only detected when there was a model violation

False negative = does not identify violation for MSCM+*m*

False positive = detects violation for MSCM

- **Data-based test statistics**: assess whether the observed and posterior predictive data sets exhibit similar characteristics (estimated directly from the data)
  - F_{ST} outliers

- **Inference-based test statistics**: where inferences drawn from the observed posterior distribution and posterior predictive distributions are compared
  - SD likelihood, RF


AAGCTGGCTA
AAGTTGCCTA
AAGTTGCCTA
AAGCTGCCTA
TAGCTGCCTA
Empirical Dataset
Analyze with SNAPP
Posterior Distribution of Species Trees
Pick Trees at Random
Posterior Trees
Extract Parameter Values
Taxonomic Relationships, Branch Lengths
Simulate new datasets under correct model (no gene flow)
GCTATGGTCT
CCTATGGTCT
GCTATGCTCT
CCTAAGGTCA
GCTATGGTCA
Posterior Predictive Datasets
Analyze with SNAPP
Posterior Predictive Distributions
FST Statistics
Likelihood and Tree Distance Statistics
Compare Empirical and Posterior Predictive Data
Likelihood and Tree Distance Statistics
FST Statistics

What if you detect violations?

- acknowledge the model violation and the effects it could have on your phylogeny estimate (we used a cut-off of 0.05 for simulation testing but you should consider the p-value)

- conduct additional analyses to examine the cause of the model violation, as such violations indicate interesting evolutionary processes not accounted for by the MSCM model
  - PhyloNet (Wen et al. 2018): MSNC

Running P2C2M.SNAPP

- First make sure your SNAPP analysis has converged (next presentation)! Otherwise you get false positives.

- Getting started (see tutorial)
  - Install P2C2M.SNAPP
  - Install Fastsimcoal
  - Put SNAPP output and fastsimcoal executable in working directory