



DBI - 1661029

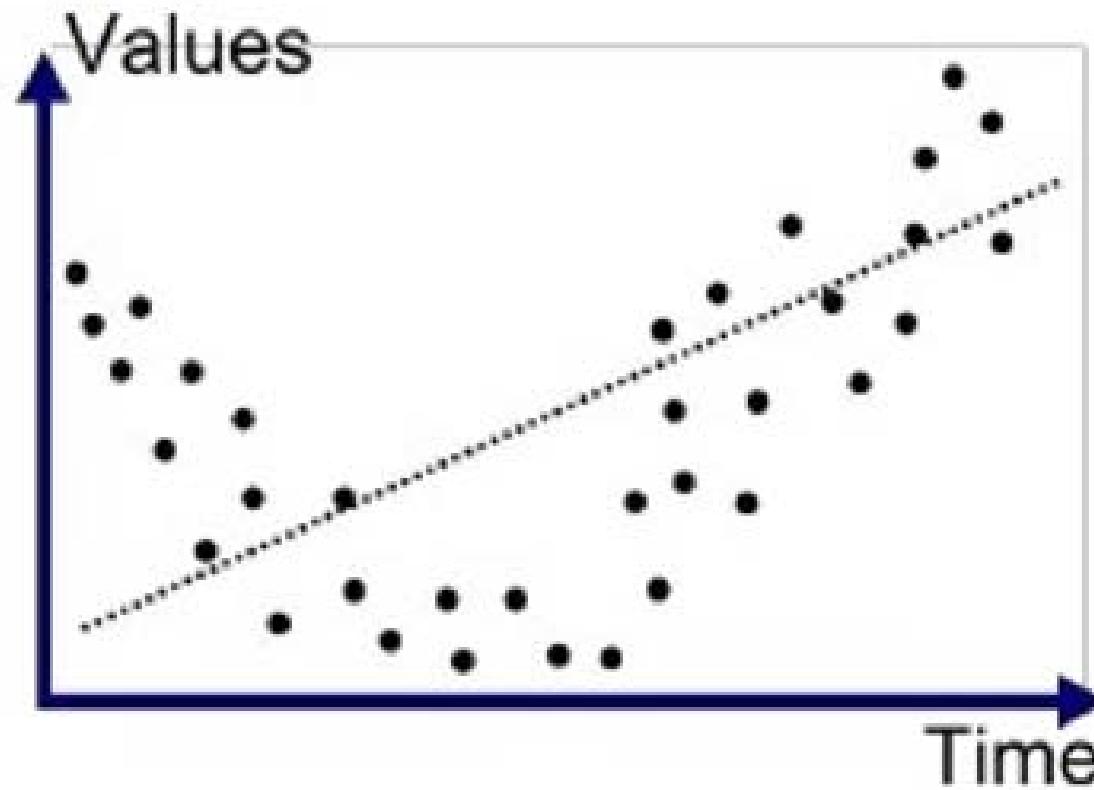
Posterior Predictive Checks in Bayesian Phylogenetics

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. Bollback¹

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 ($\Pr[\text{Tree}]$) with the likelihood ($\Pr[\text{Data} \mid \text{Tree}]$) to produce a posterior probability distribution on trees ($\Pr[\text{Tree} \mid \text{Data}]$). The posterior probability of a tree



Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology

John P. Huelsenbeck,^{1*} Fredrik Ronquist,² Rasmus Nielsen,³ Jonathan P. Bollback¹

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.

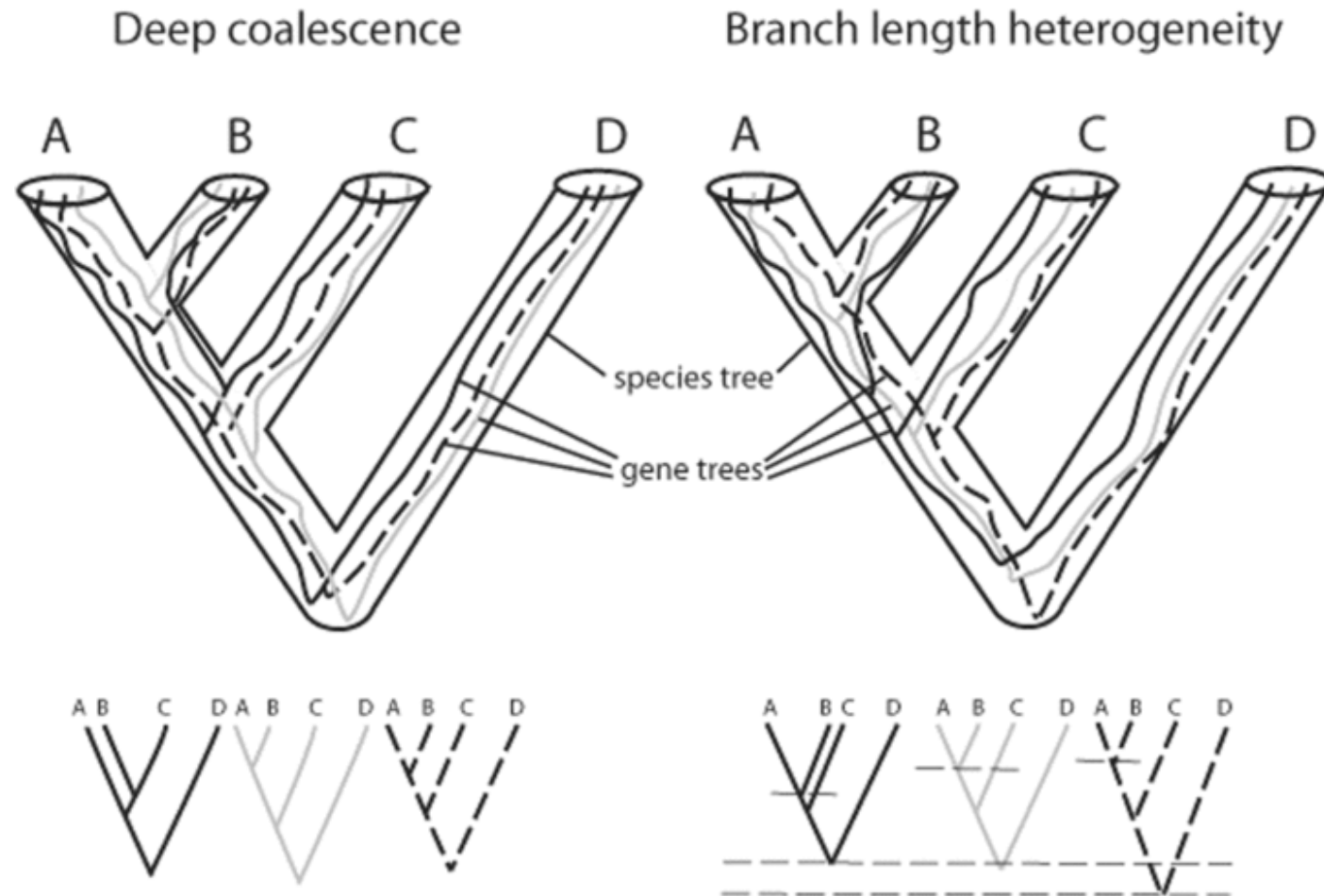
ity of a tree (Fig. 1). Bayes's theorem

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

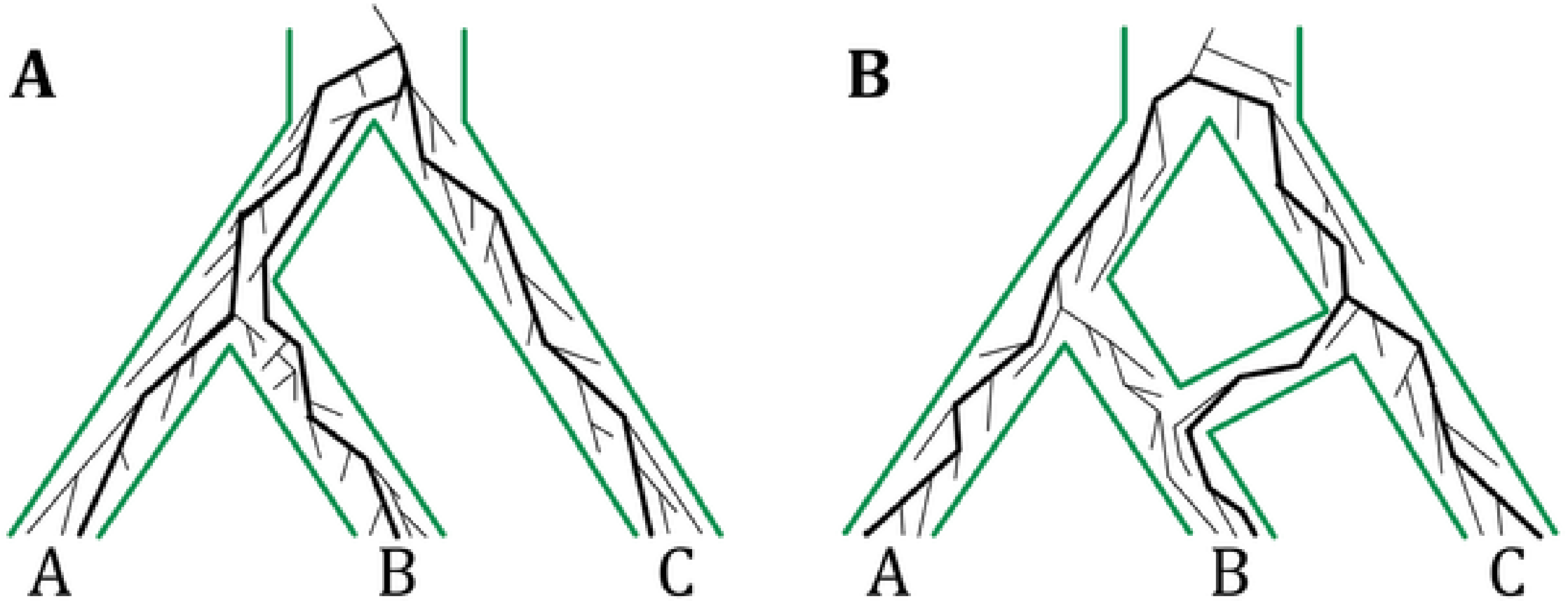
(where the vertical bar should be read as “given”) is used to combine the prior probability of a phylogeny ($\Pr[\text{Tree}]$) with the likelihood ($\Pr[\text{Data} \mid \text{Tree}]$) to produce a posterior probability distribution on trees ($\Pr[\text{Tree} \mid \text{Data}]$). The posterior probability 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)



Multi-species coalescent model (MSCM)



Syst. Biol. 63(1):17–30, 2014

© The Author(s) 2013. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.

For Permissions, please email: journals.permissions@oup.com

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. HARRIS¹, BRUCE RANNALA^{2,3}, AND ZIHENG YANG^{3,4}

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

ADAM D. LEACHÉ^{1,*}, REBECCA B. HARRIS¹, BRUCE RANNALA^{2,3}, AND ZIHENG YANG^{3,4}

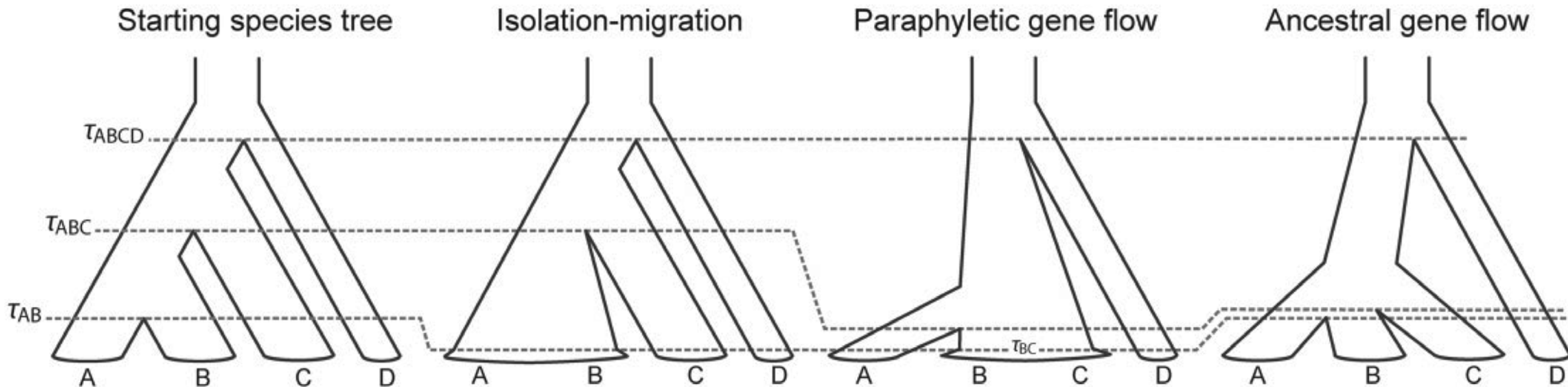


Figure 8 Species tree distortions caused by gene flow that can result from coalescent methods that only model ILS. Dashed lines illustrate species tree compression, and the widening of branches illustrates species tree dilation in relation to the starting species tree.

Syst. Biol. 63(3):322–333, 2014

© The Author(s) 2013. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.

For Permissions, please email: journals.permissions@oup.com

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. HIRD¹, JEREMY M. BROWN¹, TARA A. PELLETIER², JOHN D. McVAY¹, JORDAN D. SATLER²,
AND BRYAN C. CARSTENS²

Syst. Biol. 67(2):269–284, 2018

© The Author(s) 2017. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.

For Permissions, please email: journals.permissions@oup.com

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. BROWN², AND ROBERT C. THOMSON¹

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

```
1 #RDIUS
2
3 Begin taxaj;
4 Dimensions ntaxaj;
5
6 1
7 2
8 4
9 4
10 5
11 6
12 5
13
14 End;
15 Begin trees;
16 Translats
17 1 1,
18 2 2,
19 3 3,
20 4 4,
21 5 5,
22 6 6
23
24 tree STATE_0 = (((1[<theta=0.0600075182215585]1.08503639795450276-4,2[<theta=0.14518478858433406]1.08583630795450276-4)[<theta=0.00971084629250842
25 tree STATE_1000 = (((1[<theta=0.0913459888322793]0.488964658685944E-5,2[<theta=0.08148563194855884]0.488964658685944E-5)[<theta=0.00987174585862880
26 tree STATE_2000 = (((1[<theta=0.1409414473520646]0.482443338952133E-5,2[<theta=0.0723673254775981]0.482443338952133E-5)[<theta=0.00247824888585332
27 tree STATE_3000 = (((1[<theta=0.05679516889897618]7.276296315931688E-5,2[<theta=0.09718331136592398]7.276296315931688E-5)[<theta=0.01889191123343865
28 tree STATE_4000 = (((1[<theta=0.1152522752592852]0.208078097825483E-5,2[<theta=0.0839382714665583]0.208078097825483E-5)[<theta=0.0188057892249781
29 tree STATE_5000 = (((1[<theta=0.13539595588484923]0.513545772914485E-5,2[<theta=0.058178322747980]0.513545772914485E-5)[<theta=0.0188113877516837
30 tree STATE_6000 = (((1[<theta=0.087788455263284]0.78818945375756E-5,2[<theta=0.1233687188288676]0.78818945375756E-5)[<theta=0.08945155569785859
31 tree STATE_7000 = (((1[<theta=0.1120646779708873]7.02845457244025E-5,2[<theta=0.12267427838886153]7.02845457244025E-5)[<theta=0.018232721822556
32 tree STATE_8000 = (((1[<theta=0.07768213695194863]0.99868452683994E-5,2[<theta=0.0784898635529588]0.99868452683994E-5)[<theta=0.018388627888299
33 tree STATE_9000 = (((1[<theta=0.114721891153843]7.7922942162634E-5,2[<theta=0.208355302184985]7.7922942162634E-5)[<theta=0.0182573224441848]0
34 tree STATE_10000 = (((1[<theta=0.1387453471202522]0.8077363748832E-5,2[<theta=0.0948793234614888]0.8077363748832E-5)[<theta=0.0164844482187235
35 tree STATE_11000 = (((1[<theta=0.1139582886189198]7.51619657015127E-5,2[<theta=0.06183917835515891]7.51619657015127E-5)[<theta=0.08880841878578813
36 tree STATE_12000 = (((1[<theta=0.12142823594788]7.21068618392249E-4,2[<theta=0.1883582818447121]7.21068618392249E-4)[<theta=0.0898921479432184
37 tree STATE_13000 = (((1[<theta=0.0784844152689781]7.248379670185987E-4,2[<theta=0.03512482874688667]7.248379670185987E-4)[<theta=0.01868886238997
38 tree STATE_14000 = (((1[<theta=0.1444974724999378]7.147214547951013E-5,2[<theta=0.1848758838554118]7.147214547951013E-5)[<theta=0.089414442834269
39 tree STATE_15000 = (((1[<theta=0.078182288788883]0.13812928848895E-5,2[<theta=0.118613743714732]0.13812928848895E-5)[<theta=0.087845258613898
40 tree STATE_16000 = (((1[<theta=0.06688813925252922]0.199578728496414E-5,2[<theta=0.17251335983261884]0.199578728496414E-5)[<theta=0.089253739354384
41 tree STATE_17000 = (((1[<theta=0.0833856561464787]7.483162128498485E-5,2[<theta=0.0626885988780152]7.483162128498485E-5)[<theta=0.0891393138789654
42 tree STATE_18000 = (((1[<theta=0.1195815878388114]7.183162128498485E-5,2[<theta=0.0916264891226695]7.183162128498485E-5)[<theta=0.018793797232529
43 tree STATE_19000 = (((1[<theta=0.071356118438353]0.789481977422256E-5,2[<theta=0.0924871863712587]0.789481977422256E-5)[<theta=0.0188676746147488
44 tree STATE_20000 = (((1[<theta=0.068845627339383]2.448788867557978E-4,2[<theta=0.1345488865855418]2.448788867557978E-4)[<theta=0.013353185386
45 tree STATE_21000 = (((1[<theta=0.12965993465275584]2.064988675253233E-4,2[<theta=0.04281831992188875]2.064988675253233E-4)[<theta=0.01885328894425
46 tree STATE_22000 = (((1[<theta=0.16486179628854178]2.2378391148676889E-4,2[<theta=0.07888735391728589]2.2378391148676889E-4)[<theta=0.089285112571519
47 tree STATE_23000 = (((1[<theta=0.0498869754869593]2.48373731283413E-4,2[<theta=0.14529136464761581]2.48373731283413E-4)[<theta=0.08728724772
48 tree STATE_24000 = (((1[<theta=0.1698278385559487]1.6863163914577329E-4,2[<theta=0.18814592888693827]1.6863163914577329E-4)[<theta=0.0878467686884
```

Posterior predictive simulation (PPS)

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

```
#RStudio
Begin taxa;
Dimensions ntax=5;
Taxa labels;
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
tree STATE_0 = ((([theta=0.0600075182215585]:1.08503639799450276-4,2[theta=0.14510478058433406]:1.08503639799450276-4)[theta=0.099771084629250842]
tree STATE_1000 = ((([theta=0.0935498803227793]:0.4809646506059446-5,2[theta=0.08148563194855884]:0.4809646506059446-5)[theta=0.0808771585362808]
tree STATE_2000 = ((([theta=0.1409414473026641]:0.482443330521133-5,2[theta=0.0723673254770981]:0.482443330521133-5)[theta=0.0924782048805032]
tree STATE_3000 = ((([theta=0.05079516898997618]:7.2762963159316886-5,2[theta=0.0971831136592398]:7.2762963159316886-5)[theta=0.0189191123343865]
tree STATE_4000 = ((([theta=0.1152522752952825]:0.2080700970254031-5,2[theta=0.0829382774865583]:0.2080700970254031-5)[theta=0.0186057892249781]
tree STATE_5000 = ((([theta=0.1353959588464923]:0.513545772914405-5,2[theta=0.058178322747980]:0.513545772914405-5)[theta=0.0188113077516937]
tree STATE_6000 = ((([theta=0.087780455263284]:0.788189453757566-5,2[theta=0.1233987138288676]:0.788189453757566-5)[theta=0.0894551556978589]
tree STATE_7000 = ((([theta=0.112065673798073]:7.02854572240235-5,2[theta=0.1202742383080615]:7.02854572240235-5)[theta=0.018232713222556]
tree STATE_8000 = ((([theta=0.07708213696194869]:0.998084526039948-5,2[theta=0.07840896355929689]:0.998084526039948-5)[theta=0.0113080270808299]
tree STATE_9000 = ((([theta=0.114721991153843]:7.79229421626346-5,2[theta=0.2040355302104985]:7.79229421626346-5)[theta=0.0182573224441848]
tree STATE_10000 = ((([theta=0.1387454571202522]:0.407736274861325-5,2[theta=0.0948789334614088]:0.407736274861325-5)[theta=0.0116494462187230]
tree STATE_11000 = ((([theta=0.1135928886189198]:7.516196570151276-5,2[theta=0.06183917835615891]:7.516196570151276-5)[theta=0.08080941870578813]
tree STATE_12000 = ((([theta=0.1214428233927881]:2.100681810922408-4,2[theta=0.1083528184471212]:2.100681810922408-4)[theta=0.08998224797432184]
tree STATE_13000 = ((([theta=0.0784481415208978]:2.7483796701859876-4,2[theta=0.03512482874688697]:2.7483796701859876-4)[theta=0.0186806230997]
tree STATE_14000 = ((([theta=0.14448974724999376]:7.1472145479519136-5,2[theta=0.1848758838554118]:7.1472145479519136-5)[theta=0.0894144426334269]
tree STATE_15000 = ((([theta=0.078152280870805]:6.138012920804895-5,2[theta=0.118653743747321]:6.138012920804895-5)[theta=0.087845204513908]
tree STATE_16000 = ((([theta=0.06688813925252922]:6.199578720496414E-5,2[theta=0.1725133983261084]:6.199578720496414E-5)[theta=0.089253739354384]
tree STATE_17000 = ((([theta=0.0633856526464277]:1.481162136494858-5,2[theta=0.062069580870152]:1.481162136494858-5)[theta=0.0891329317079654]
tree STATE_18000 = ((([theta=0.11958158783389140]:7.183102138498482E-5,2[theta=0.091624891226695]:7.183102138498482E-5)[theta=0.018793797232529]
tree STATE_19000 = ((([theta=0.0713561198438353]:6.78948197422256E-5,2[theta=0.092487186372587]:6.78948197422256E-5)[theta=0.0186767461474488]
tree STATE_20000 = ((([theta=0.0684656273339385]:2.468780807537978E-4,2[theta=0.134548846455818]:2.468780807537978E-4)[theta=0.0183351485380]
tree STATE_21000 = ((([theta=0.1296599346575584]:2.06498067525333E-4,2[theta=0.0426183199218687]:2.06498067525333E-4)[theta=0.0185328894425]
tree STATE_22000 = ((([theta=0.16486179628645178]:2.37831148676889E-4,2[theta=0.0788973539172869]:2.37831148676889E-4)[theta=0.0892851127157519]
tree STATE_23000 = ((([theta=0.049860784689593]:2.48737372283453E-4,2[theta=0.145278364470150]:2.48737372283453E-4)[theta=0.089758724772]
tree STATE_24000 = ((([theta=0.1689278365559487]:1.6863163914577329E-4,2[theta=0.1081459288693827]:1.6863163914577329E-4)[theta=0.0878670680834
```

```
Simulation: "State 0"
Begin taxa;
Dimensions ntax=5;
Taxa labels;
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
tree STATE_0 = ((([theta=0.0600075182215585]:1.08503639799450276-4,2[theta=0.14510478058433406]:1.08503639799450276-4)[theta=0.099771084629250842]
tree STATE_1000 = ((([theta=0.0935498803227793]:0.4809646506059446-5,2[theta=0.08148563194855884]:0.4809646506059446-5)[theta=0.0808771585362808]
tree STATE_2000 = ((([theta=0.1409414473026641]:0.482443330521133-5,2[theta=0.0723673254770981]:0.482443330521133-5)[theta=0.0924782048805032]
tree STATE_3000 = ((([theta=0.05079516898997618]:7.2762963159316886-5,2[theta=0.0971831136592398]:7.2762963159316886-5)[theta=0.0189191123343865]
tree STATE_4000 = ((([theta=0.1152522752952825]:0.2080700970254031-5,2[theta=0.0829382774865583]:0.2080700970254031-5)[theta=0.0186057892249781]
tree STATE_5000 = ((([theta=0.1353959588464923]:0.513545772914405-5,2[theta=0.058178322747980]:0.513545772914405-5)[theta=0.0188113077516937]
tree STATE_6000 = ((([theta=0.087780455263284]:0.788189453757566-5,2[theta=0.1233987138288676]:0.788189453757566-5)[theta=0.0894551556978589]
tree STATE_7000 = ((([theta=0.112065673798073]:7.02854572240235-5,2[theta=0.1202742383080615]:7.02854572240235-5)[theta=0.018232713222556]
tree STATE_8000 = ((([theta=0.07708213696194869]:0.998084526039948-5,2[theta=0.07840896355929689]:0.998084526039948-5)[theta=0.0113080270808299]
tree STATE_9000 = ((([theta=0.114721991153843]:7.79229421626346-5,2[theta=0.2040355302104985]:7.79229421626346-5)[theta=0.0182573224441848]
tree STATE_10000 = ((([theta=0.1387454571202522]:0.407736274861325-5,2[theta=0.0948789334614088]:0.407736274861325-5)[theta=0.0116494462187230]
tree STATE_11000 = ((([theta=0.1135928886189198]:7.516196570151276-5,2[theta=0.06183917835615891]:7.516196570151276-5)[theta=0.08080941870578813]
tree STATE_12000 = ((([theta=0.1214428233927881]:2.100681810922408-4,2[theta=0.1083528184471212]:2.100681810922408-4)[theta=0.08998224797432184]
tree STATE_13000 = ((([theta=0.0784481415208978]:2.7483796701859876-4,2[theta=0.03512482874688697]:2.7483796701859876-4)[theta=0.0186806230997]
tree STATE_14000 = ((([theta=0.14448974724999376]:7.1472145479519136-5,2[theta=0.1848758838554118]:7.1472145479519136-5)[theta=0.0894144426334269]
tree STATE_15000 = ((([theta=0.078152280870805]:6.138012920804895-5,2[theta=0.118653743747321]:6.138012920804895-5)[theta=0.087845204513908]
tree STATE_16000 = ((([theta=0.06688813925252922]:6.199578720496414E-5,2[theta=0.1725133983261084]:6.199578720496414E-5)[theta=0.089253739354384]
tree STATE_17000 = ((([theta=0.0633856526464277]:1.481162136494858-5,2[theta=0.062069580870152]:1.481162136494858-5)[theta=0.0891329317079654]
tree STATE_18000 = ((([theta=0.11958158783389140]:7.183102138498482E-5,2[theta=0.091624891226695]:7.183102138498482E-5)[theta=0.018793797232529]
tree STATE_19000 = ((([theta=0.0713561198438353]:6.78948197422256E-5,2[theta=0.092487186372587]:6.78948197422256E-5)[theta=0.0186767461474488]
tree STATE_20000 = ((([theta=0.0684656273339385]:2.468780807537978E-4,2[theta=0.134548846455818]:2.468780807537978E-4)[theta=0.0183351485380]
tree STATE_21000 = ((([theta=0.1296599346575584]:2.06498067525333E-4,2[theta=0.0426183199218687]:2.06498067525333E-4)[theta=0.0185328894425]
tree STATE_22000 = ((([theta=0.16486179628645178]:2.37831148676889E-4,2[theta=0.0788973539172869]:2.37831148676889E-4)[theta=0.0892851127157519]
tree STATE_23000 = ((([theta=0.049860784689593]:2.48737372283453E-4,2[theta=0.145278364470150]:2.48737372283453E-4)[theta=0.089758724772]
tree STATE_24000 = ((([theta=0.1689278365559487]:1.6863163914577329E-4,2[theta=0.1081459288693827]:1.6863163914577329E-4)[theta=0.0878670680834
```

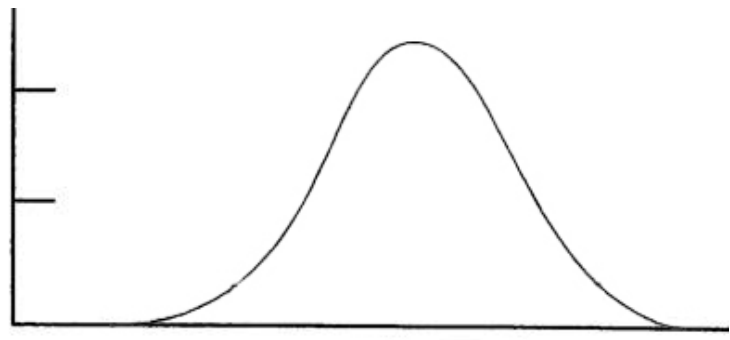
Posterior predictive simulation (PPS)

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

```
#RStudio
Begin taxa;
Dimensions ntax=5;
TaxLabels;
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
tree STATE_0 = ((([theta=0.0600075182215585]:1.08503639799450276-4,2[theta=0.14518478858433406]:1.08503639799450276-4)[theta=0.09971084629250842]
tree STATE_1000 = ((([theta=0.0913459883272793]:0.488964658685944-5,2[theta=0.0814853194855884]:0.488964658685944-5)[theta=0.080871588362880]
tree STATE_2000 = ((([theta=0.140941447324664]:0.40244338952113E-5,2[theta=0.0723672954720981]:0.40244338952113E-5)[theta=0.082478248885032]
tree STATE_3000 = ((([theta=0.05679516898997618]:7.276296315931686E-5,2[theta=0.0971833136592398]:7.276296315931686E-5)[theta=0.018991912334385]
tree STATE_4000 = ((([theta=0.115252275259282]:0.280878097825483-5,2[theta=0.083935271466558]:0.280878097825483-5)[theta=0.018605789224978]
tree STATE_5000 = ((([theta=0.1353959588464923]:0.515154577291448E-5,2[theta=0.058178322747980]:0.515154577291448E-5)[theta=0.01881138077516937]
tree STATE_6000 = ((([theta=0.087786455263284]:0.78818945375756E-5,2[theta=0.1233687188288676]:0.78818945375756E-5)[theta=0.084515556978585]
tree STATE_7000 = ((([theta=0.112065673798873]:7.0285457246236E-5,2[theta=0.1292742383886153]:7.0285457246236E-5)[theta=0.018232213222556]
tree STATE_8000 = ((([theta=0.07708213696194863]:0.99808452683994E-5,2[theta=0.07848986355929689]:0.99808452683994E-5)[theta=0.0113088270888299]
tree STATE_9000 = ((([theta=0.114721991153843]:7.7922942162634E-5,2[theta=0.208355302184985]:7.7922942162634E-5)[theta=0.0182573224441848]
tree STATE_10000 = ((([theta=0.138745471202522]:0.40777687486132E-5,2[theta=0.094879334614088]:0.40777687486132E-5)[theta=0.016484462187235]
tree STATE_11000 = ((([theta=0.1159582886189198]:7.51619657015127E-5,2[theta=0.06183917835615891]:7.51619657015127E-5)[theta=0.0880894187578813]
tree STATE_12000 = ((([theta=0.121428235927881]:2.10686181892249E-4,2[theta=0.188358184471212]:2.10686181892249E-4)[theta=0.0899821479432184]
tree STATE_13000 = ((([theta=0.0784481415268978]:7.748379670185987E-4,2[theta=0.03512482874688697]:7.748379670185987E-4)[theta=0.0186880238997]
tree STATE_14000 = ((([theta=0.1444974724999378]:7.14721454795191E-5,2[theta=0.1848758383554118]:7.14721454795191E-5)[theta=0.089414442634269]
tree STATE_15000 = ((([theta=0.07815228878085]:6.13801292880485E-5,2[theta=0.118613743174732]:6.13801292880485E-5)[theta=0.087845204513908]
tree STATE_16000 = ((([theta=0.066888139252922]:6.19957872849641E-5,2[theta=0.1725135983261884]:6.19957872849641E-5)[theta=0.089253739354384]
tree STATE_17000 = ((([theta=0.063385621664277]:7.18310213849485E-5,2[theta=0.2065898010152]:7.18310213849485E-5)[theta=0.089129317879654]
tree STATE_18000 = ((([theta=0.1195815878338114]:7.18310213849485E-5,2[theta=0.091264891226695]:7.18310213849485E-5)[theta=0.019783797232529]
tree STATE_19000 = ((([theta=0.071356118438353]:6.78948197427256E-5,2[theta=0.092487186372587]:6.78948197427256E-5)[theta=0.018676746147468]
tree STATE_20000 = ((([theta=0.068465627339385]:2.46878808753797E-4,2[theta=0.134548846455418]:2.46878808753797E-4)[theta=0.0183351485380]
tree STATE_21000 = ((([theta=0.129659346575584]:2.66498067525323E-4,2[theta=0.0426183199218687]:2.66498067525323E-4)[theta=0.0185328994425]
tree STATE_22000 = ((([theta=0.1648617962864178]:2.37831148676888E-4,2[theta=0.078873591723691]:2.37831148676888E-4)[theta=0.089385112571519]
tree STATE_23000 = ((([theta=0.049860784689593]:2.48737372283433E-4,2[theta=0.1452913646476158]:2.48737372283433E-4)[theta=0.087528724772]
tree STATE_24000 = ((([theta=0.169827835559487]:1.686163191457732E-4,2[theta=0.1881459288693827]:1.686163191457732E-4)[theta=0.0878675686834
```

```
Simulation - State 0
No parameters
Simulation - State 1
No parameters
Simulation - State 2
No parameters
Simulation - State 3
No parameters
Simulation - State 4
No parameters
Simulation - State 5
No parameters
Simulation - State 6
No parameters
Simulation - State 7
No parameters
Simulation - State 8
No parameters
Simulation - State 9
No parameters
Simulation - State 10
No parameters
Simulation - State 11
No parameters
Simulation - State 12
No parameters
Simulation - State 13
No parameters
Simulation - State 14
No parameters
Simulation - State 15
No parameters
Simulation - State 16
No parameters
Simulation - State 17
No parameters
Simulation - State 18
No parameters
Simulation - State 19
No parameters
Simulation - State 20
No parameters
Simulation - State 21
No parameters
Simulation - State 22
No parameters
Simulation - State 23
No parameters
Simulation - State 24
No parameters
```

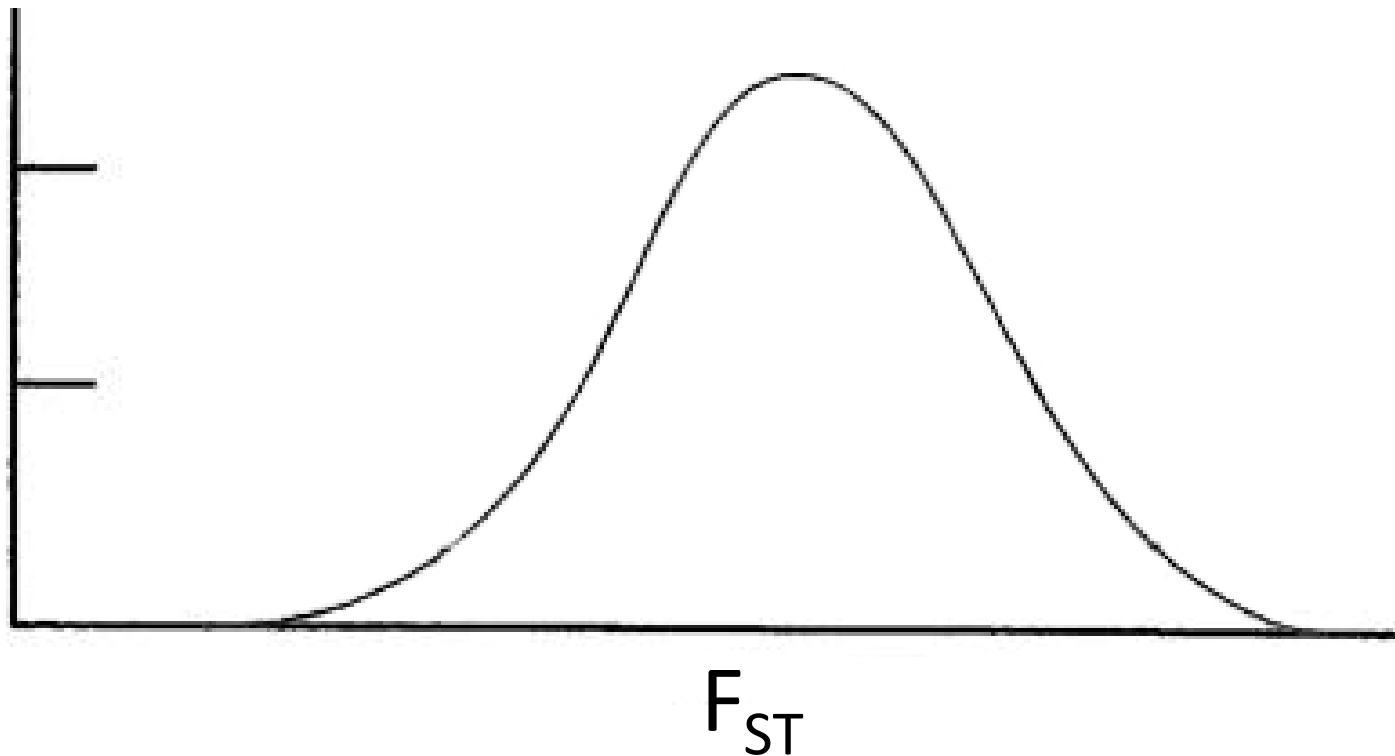
Posterior predictive distribution



Summary statistic

Posterior predictive distribution (PPD)

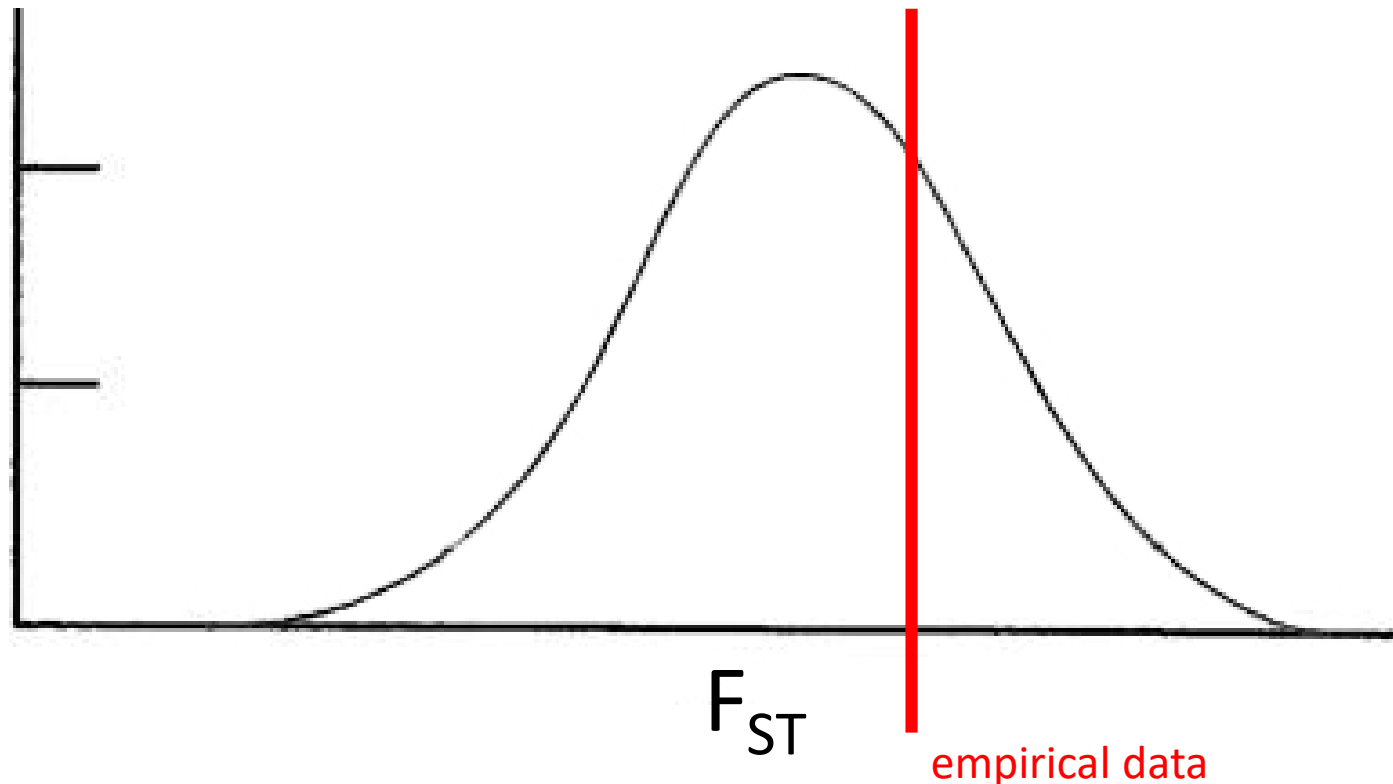
- How well does the empirical data fit this distribution?



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)

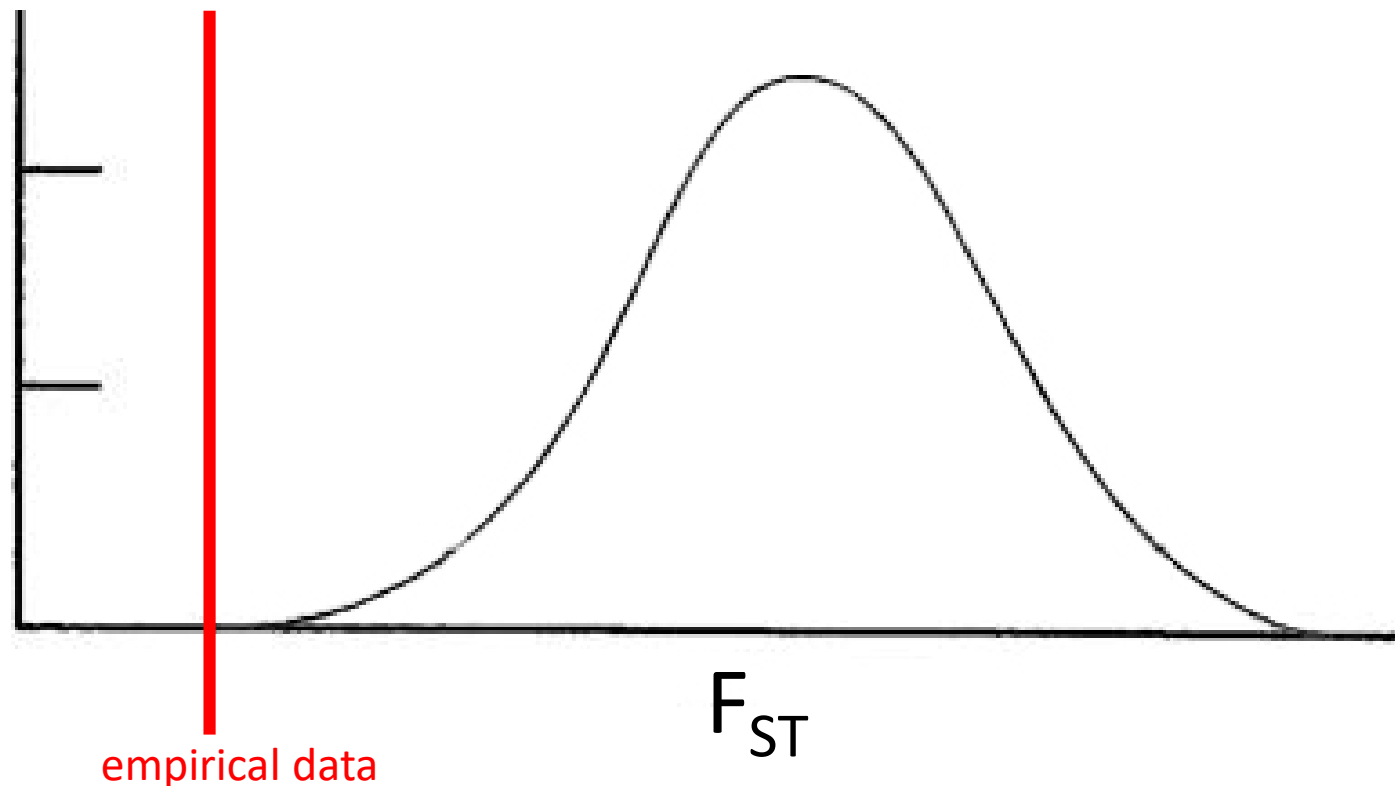
- How well does the empirical data fit this distribution?



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

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)

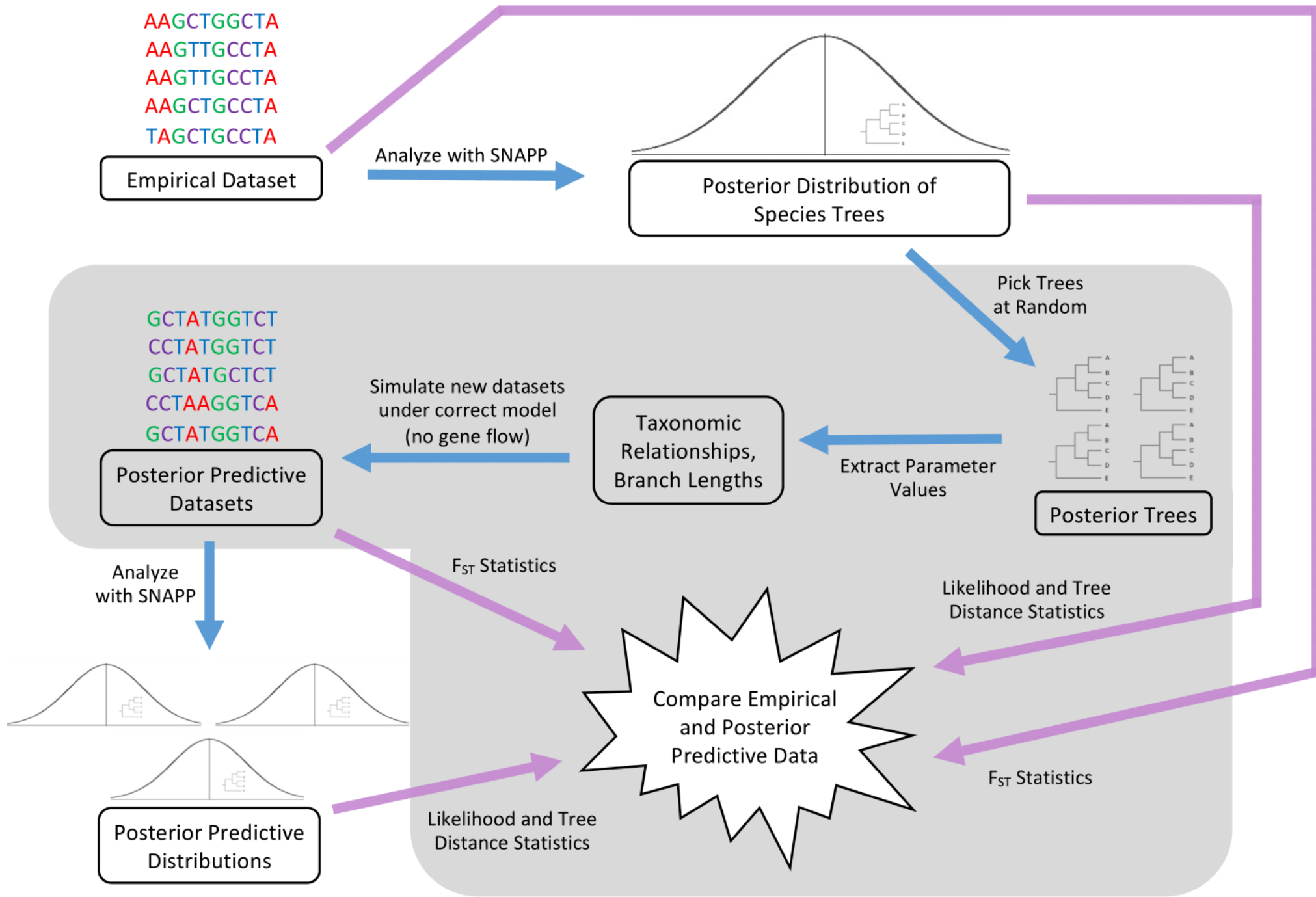
- Interfaces with the BEAST package
- Algorithm bypasses the gene trees and computes species tree likelihoods directly from the markers
- Returns a sample of species trees with (relative) divergence times and population sizes (posterior distribution)

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



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

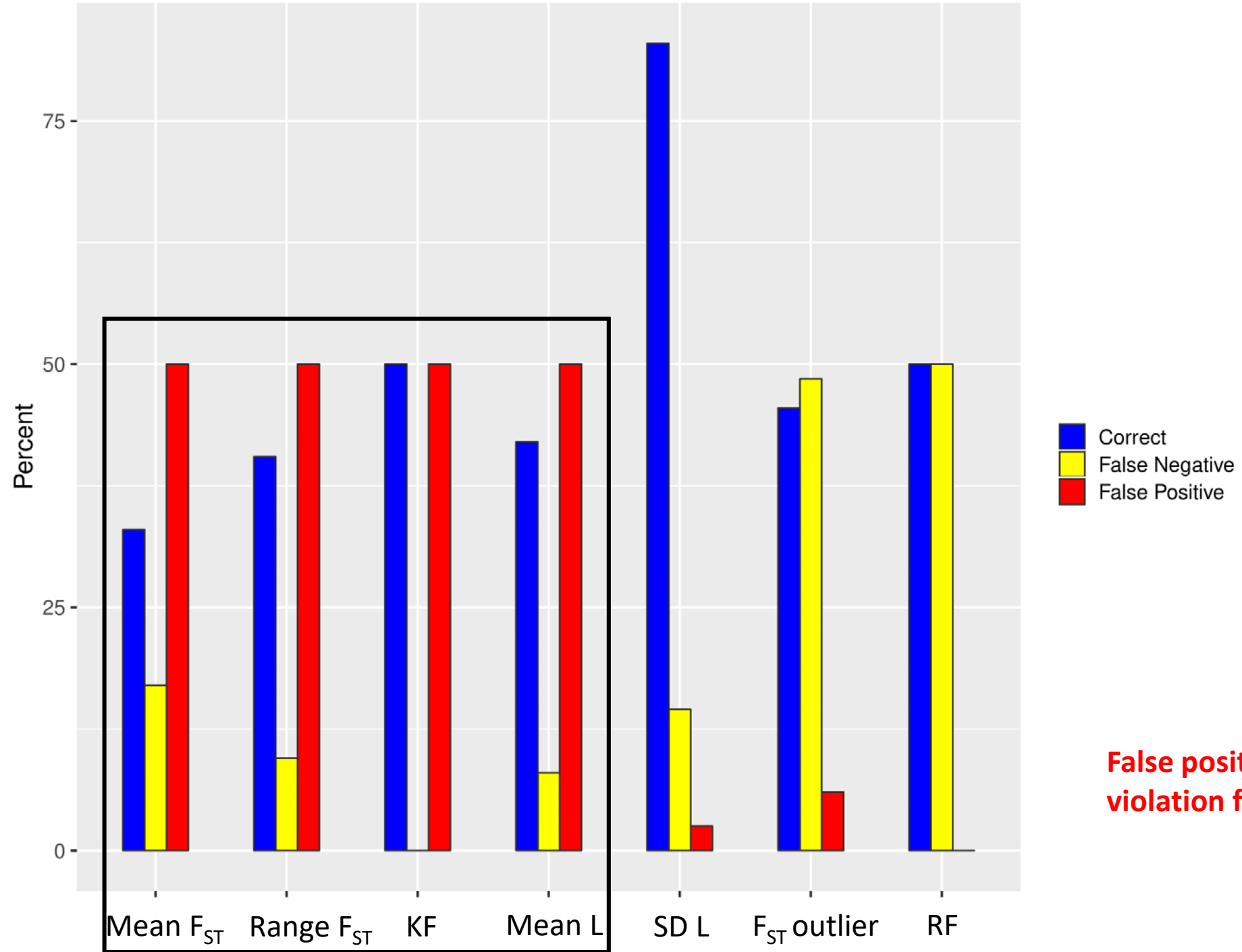
- **Summary statistics tested:**

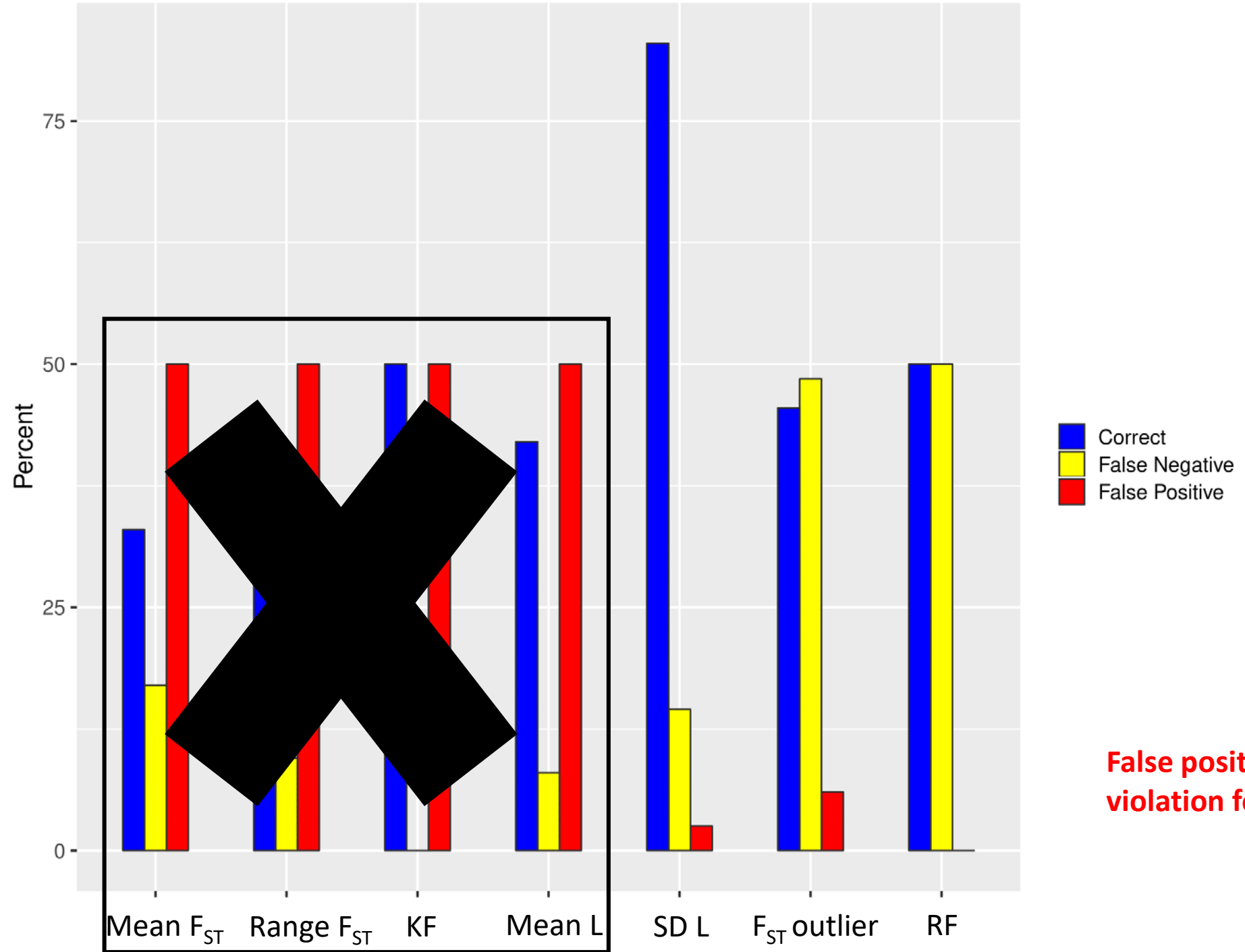
- Mean pairwise F_{ST}
- Range of pairwise F_{ST}
- F_{ST} outlier test
- Robinson-Foulds distance (topological distance only)
- Kuhner-Felsenstein distance (includes branch lengths)
- Mean of tree likelihood
- Standard deviation of tree likelihood

P2C2M.SNAPP simulation testing

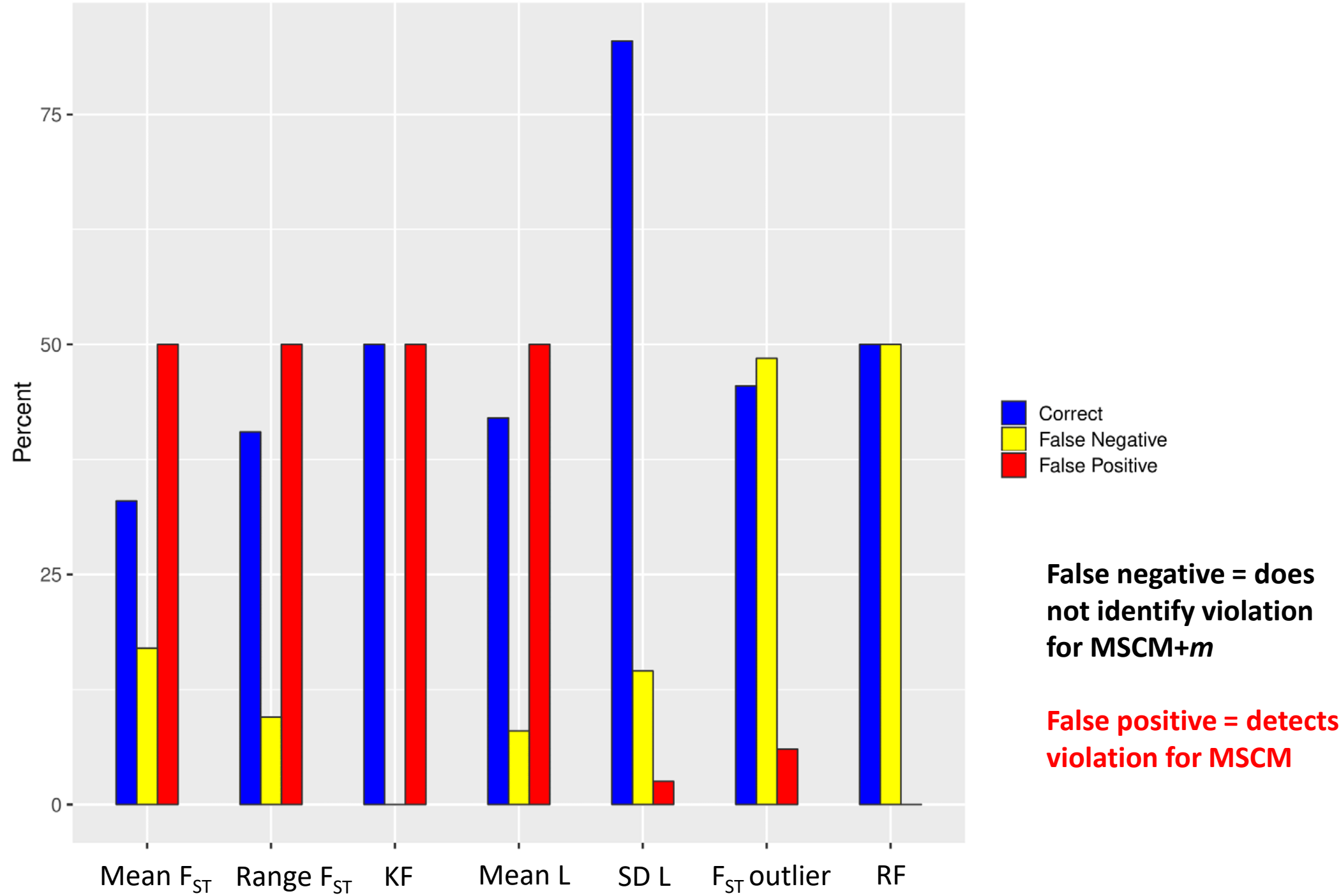
- **Summary statistics tested:**

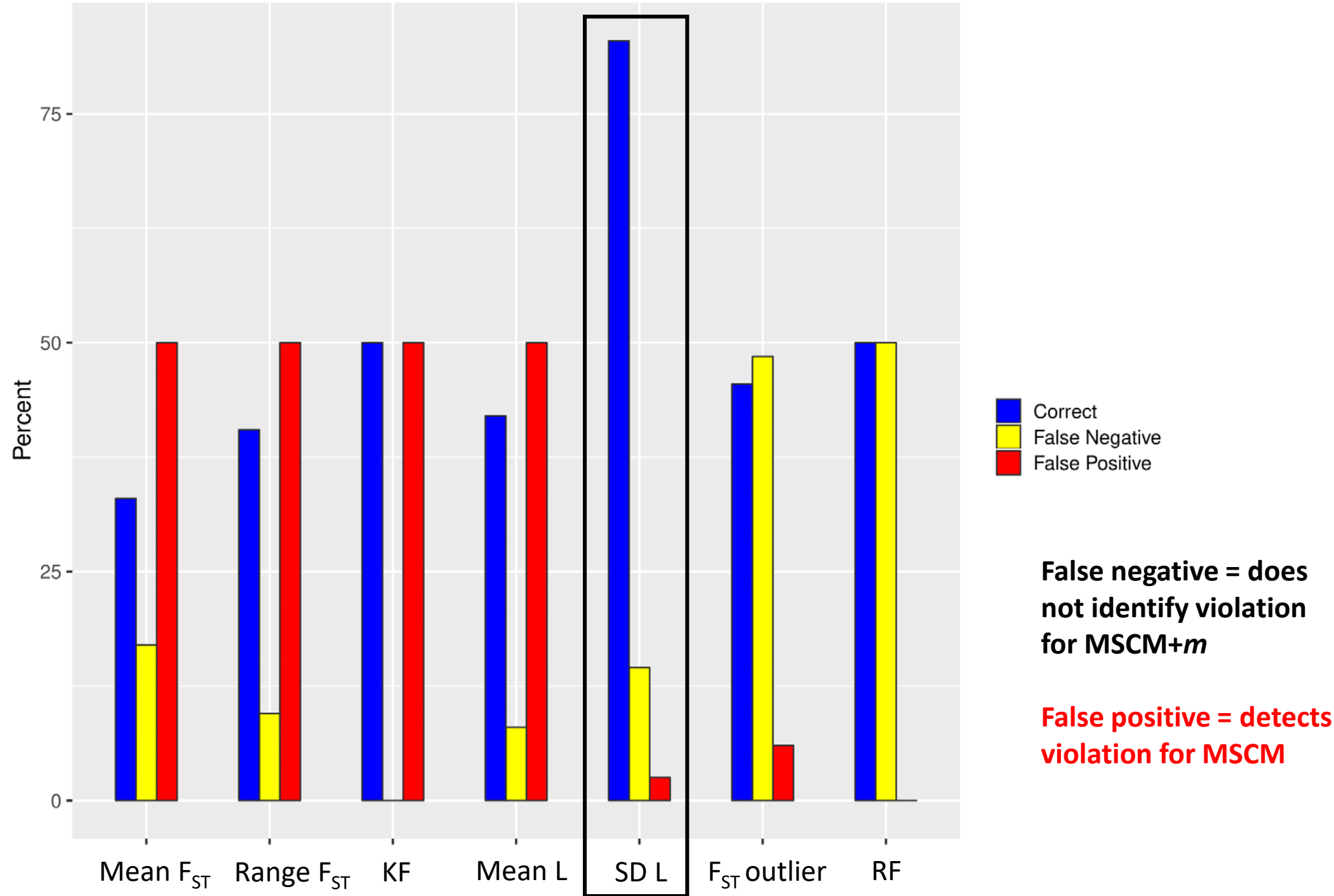
- ~~Mean pairwise F_{ST}~~
- ~~Range of pairwise F_{ST}~~
- **F_{ST} outlier test**
- **Robinson-Foulds distance (topological distance only)**
- ~~Kuhner-Felsenstein distance (includes branch lengths)~~
- ~~Mean of tree likelihood~~
- **Standard deviation of tree likelihood**

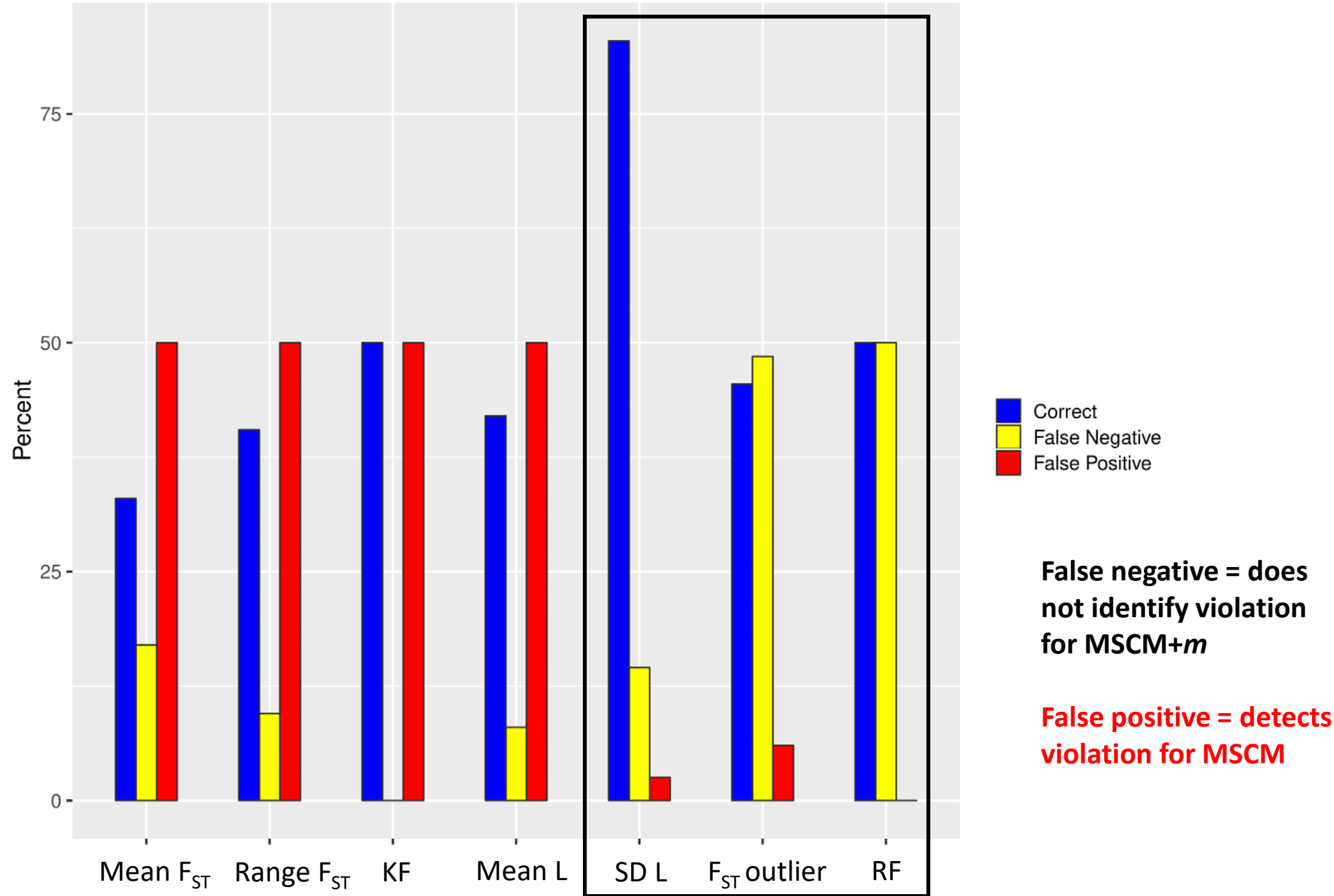




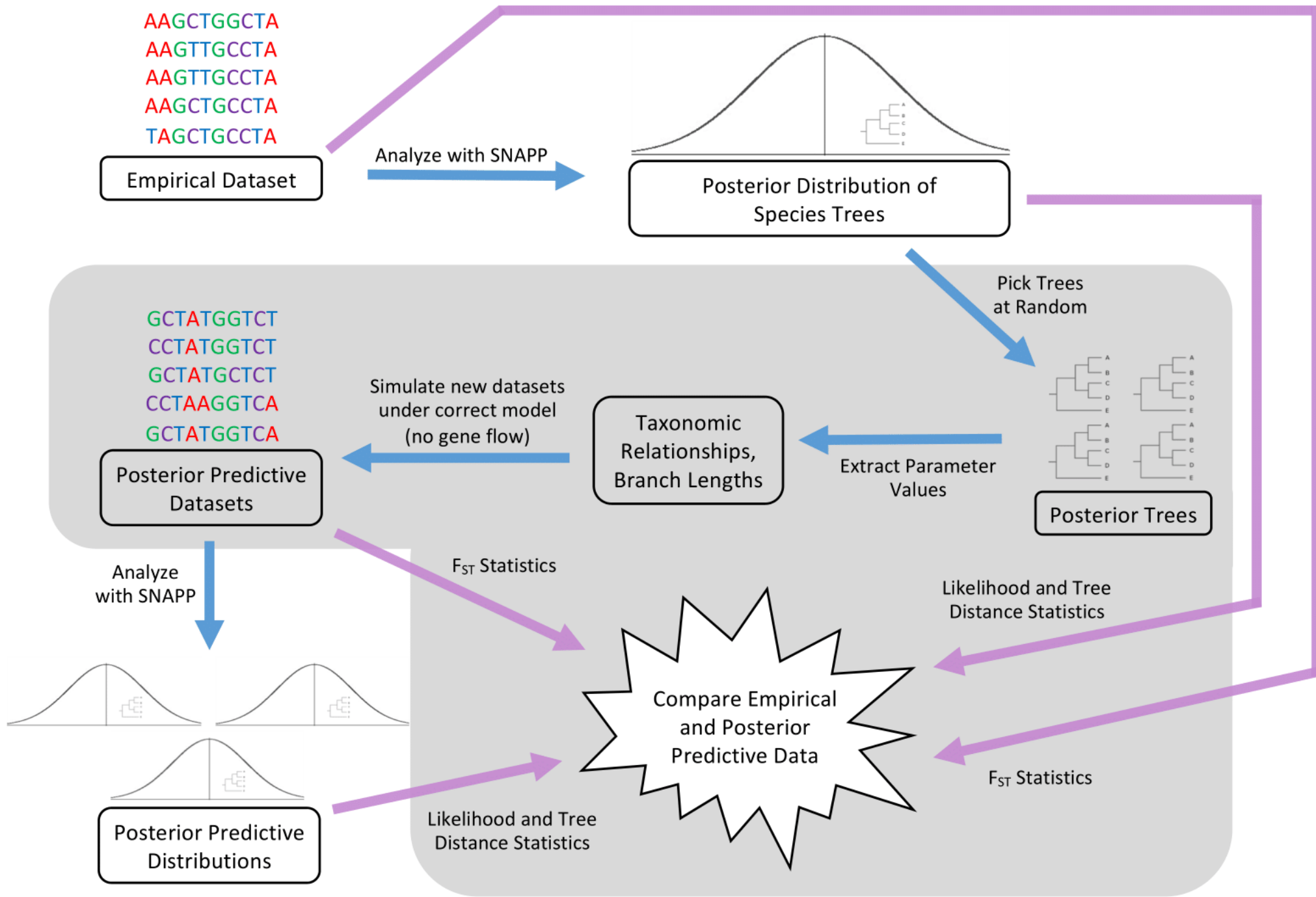
False positive = detects violation for MSCM







- **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



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