



DBI - 1661029

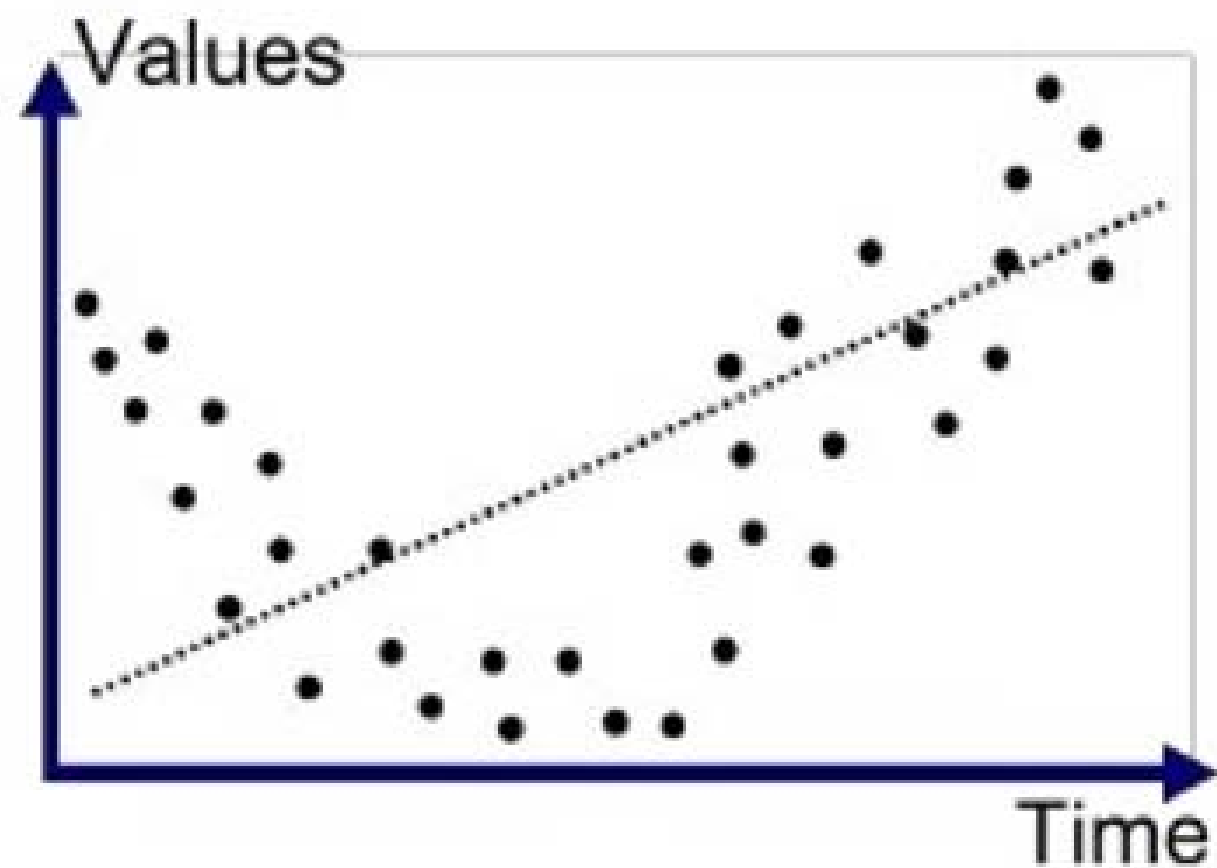
Posterior Predictive Checks in Bayesian Phylogenetics

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)

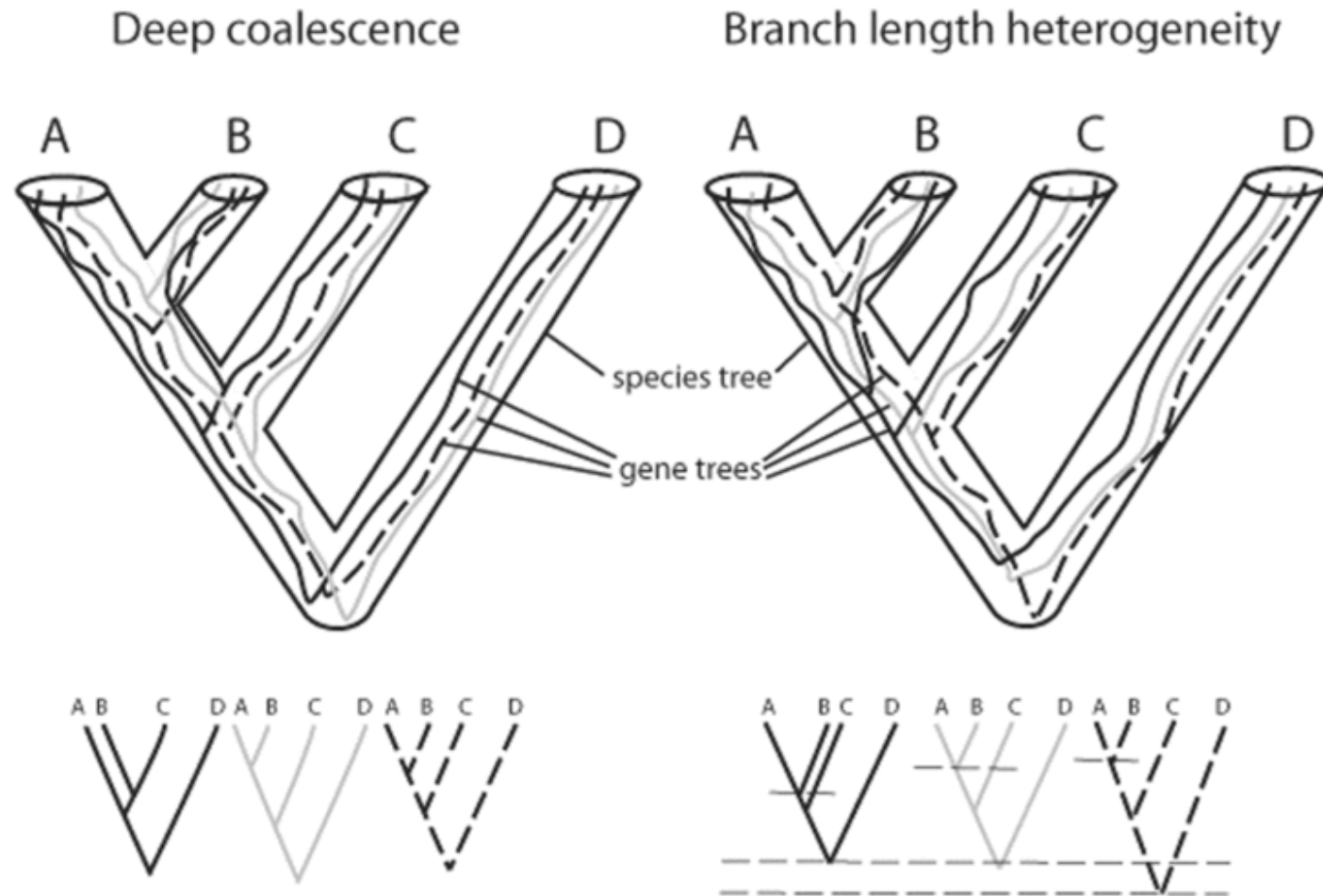
- developed to estimate species trees, while accounting for the coalescent process that can lead to incongruence among gene trees (also known as **incomplete lineage sorting**)

Multi-species coalescent model (MSCM)

Deep coalescence

Branch length heterogeneity

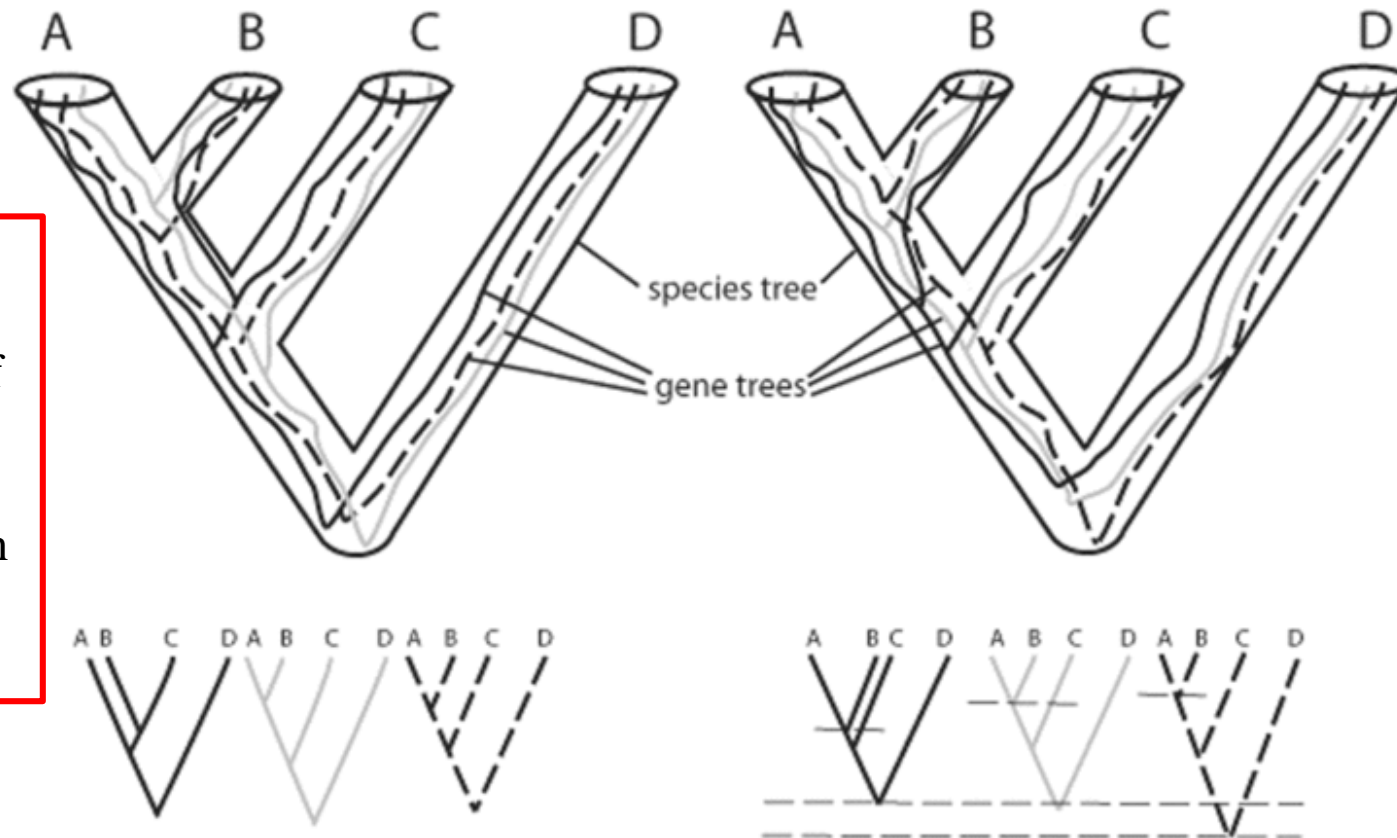
Multi-species coalescent model (MSCM)



Multi-species coalescent model (MSCM)

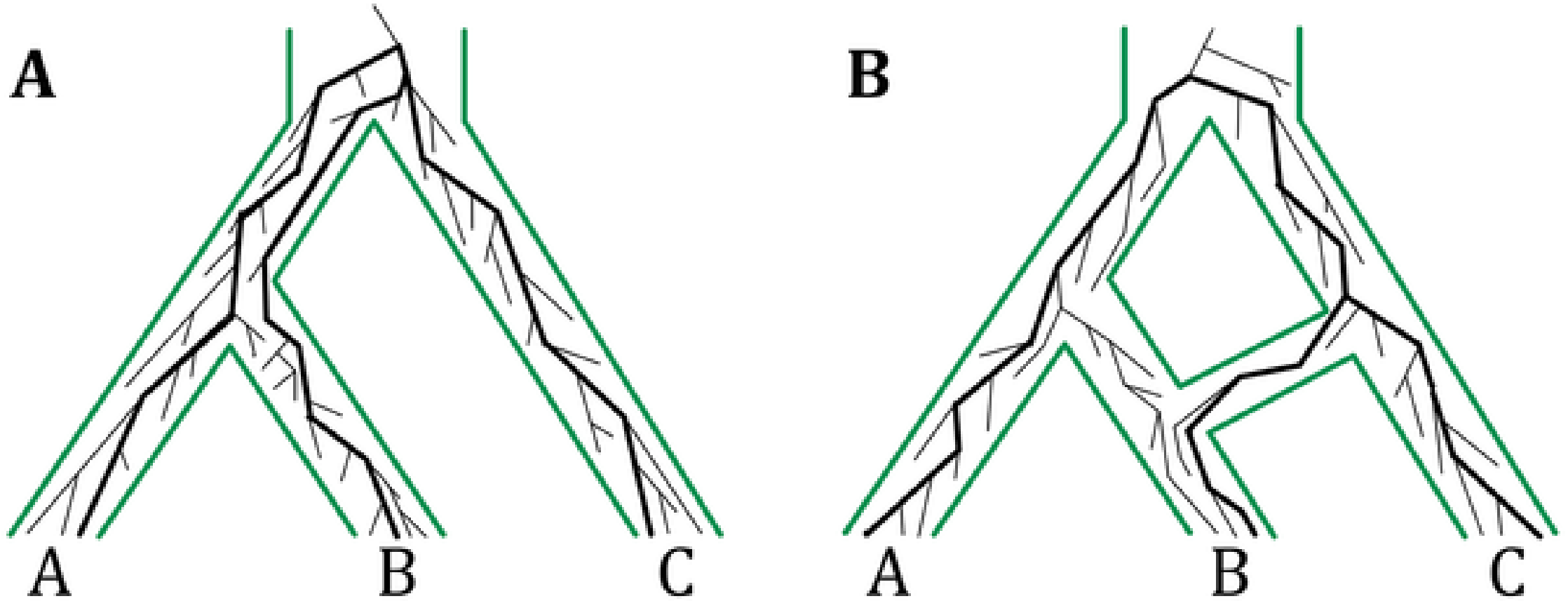
Deep coalescence

Branch length heterogeneity



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.

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

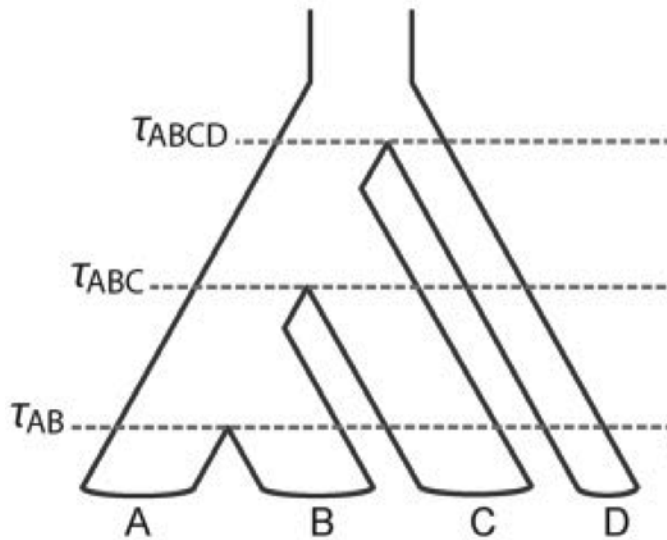
ADAM D. LEACHÉ^{1,*}, REBECCA B. HARRIS¹, BRUCE RANNALA^{2,3}, AND ZIHENG YANG^{3,4}

Starting species tree

Isolation-migration

Paraphyletic gene flow

Ancestral gene flow



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}

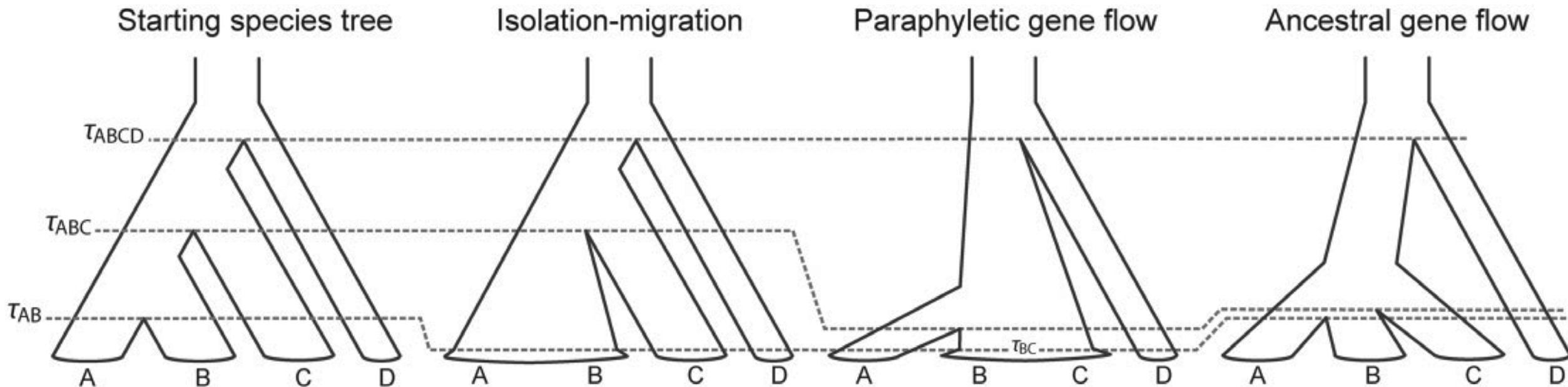


Figure 8 Species tree distortions caused by gene flow that can result from coalescent methods that only model ILS. Dashed lines illustrate species tree compression, and the widening of branches illustrates species tree dilation in relation to the starting species tree.

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 taxa;
4 Dimensions ntax=5;
5 TaxLabels
6 1
7 2
8 3
9 4
10 5
11 6
12 7
13 End;
14 Begin trees;
15 TransIs
16 1 1,
17 2 2,
18 3 3,
19 4 4,
20 5 5,
21 6 6
22 ;
23
24 tree STATE_0 = (((1[{theta=0.0600075182215585}]:1.0850363979945027E-4,2[{theta=0.14510478058433406}]:1.0850363979945027E-4){[theta=0.00971084629250842
25 tree STATE_1000 = (((1[{theta=0.0913459808322793}]:0.480964650605944E-5,2[{theta=0.08140563194055084}]:0.480964650605944E-5){[theta=0.00987174585862880
26 tree STATE_2000 = (((1[{theta=0.1409414473020646}]:0.84244333052133E-5,2[{theta=0.07236732547705981}]:0.84244333052133E-5){[theta=0.00247820488050532
27 tree STATE_3000 = (((1[{theta=0.05079016898997618}]:7.276296315931686E-5,2[{theta=0.09718331136592398}]:7.276296315931686E-5){[theta=0.01099191123343805
28 tree STATE_4000 = (((1[{theta=0.115252275259282}]:0.200870097025403E-5,2[{theta=0.0039302714605503}]:0.200870097025403E-5){[theta=0.0100057092249781
29 tree STATE_5000 = (((1[{theta=0.1303959558846023}]:0.51354577291440E-5,2[{theta=0.050178322747980}]:0.51354577291440E-5){[theta=0.0180113077616937
30 tree STATE_6000 = (((1[{theta=0.087780455263284}]:6.788189453075756E-5,2[{theta=0.1233907138208676}]:6.788189453075756E-5){[theta=0.00945155569785859
31 tree STATE_7000 = (((1[{theta=0.1120646577970807}]:7.02045457244023E-5,2[{theta=0.12267427830806153}]:7.02045457244023E-5){[theta=0.0102327210222556
32 tree STATE_8000 = (((1[{theta=0.0770821369519469}]:0.99068452603994E-5,2[{theta=0.07840896335929689}]:0.99068452603994E-5){[theta=0.01103086270808299
33 tree STATE_9000 = (((1[{theta=0.1147261991153843}]:7.7922942102634E-5,2[{theta=0.2000355302104985}]:7.7922942102634E-5){[theta=0.0110573224441848}]:0
34 tree STATE_10000 = (((1[{theta=0.1307453471202522}]:0.80773637406132E-5,2[{theta=0.0940702354614080}]:0.80773637406132E-5){[theta=0.0116404462107230
35 tree STATE_11000 = (((1[{theta=0.1119582886189199}]:7.51619657015127E-5,2[{theta=0.06183917835615091}]:7.51619657015127E-5){[theta=0.00900941870570813
36 tree STATE_12000 = (((1[{theta=0.12142023594786}]:2.10096181092249E-4,2[{theta=0.1003520184470121}]:2.10096181092249E-4){[theta=0.0090024797432184
37 tree STATE_13000 = (((1[{theta=0.0784044152609781}]:2.7403796701859087E-4,2[{theta=0.0351242874608867}]:2.7403796701859087E-4){[theta=0.0160800230997
38 tree STATE_14000 = (((1[{theta=0.14440974724999376}]:7.14721047951013E-5,2[{theta=0.1848758638554118}]:7.14721047951013E-5){[theta=0.009414442034260
39 tree STATE_15000 = (((1[{theta=0.0701822008700805}]:6.138012920084095E-5,2[{theta=0.1186537437147321}]:6.138012920084095E-5){[theta=0.007845204611908
40 tree STATE_16000 = (((1[{theta=0.06608013925252922}]:6.199578720496414E-5,2[{theta=0.17251335983201084}]:6.199578720496414E-5){[theta=0.0092053739354384
41 tree STATE_17000 = (((1[{theta=0.0633856569464707}]:2.08162120490405E-5,2[{theta=0.062609598070152}]:2.08162120490405E-5){[theta=0.0091293170970654
42 tree STATE_18000 = (((1[{theta=0.1195815878389110}]:7.183102130498403E-5,2[{theta=0.0916264891226695}]:7.183102130498403E-5){[theta=0.010793797232529
43 tree STATE_19000 = (((1[{theta=0.071361110438353}]:6.789401977422256E-5,2[{theta=0.0924871863712587}]:6.789401977422256E-5){[theta=0.010676746147468
44 tree STATE_20000 = (((1[{theta=0.068045627339383}]:2.44870007575707E-4,2[{theta=0.1345408065455418}]:2.44870007575707E-4){[theta=0.013353105390
45 tree STATE_21000 = (((1[{theta=0.12965993465275584}]:2.064900675253233E-4,2[{theta=0.04261831992186875}]:2.064900675253233E-4){[theta=0.01085320894425
46 tree STATE_22000 = (((1[{theta=0.16406179620054170}]:2.370391140676809E-4,2[{theta=0.0708973539172869}]:2.370391140676809E-4){[theta=0.009205112571519
47 tree STATE_23000 = (((1[{theta=0.0490607544609593}]:2.08737372203413E-4,2[{theta=0.14529183644701501}]:2.08737372203413E-4){[theta=0.007208724772
48 tree STATE_24000 = (((1[{theta=0.1698278385659487}]:1.6863163914577329E-4,2[{theta=0.10014592088693827}]:1.6863163914577329E-4){[theta=0.00784670606034
```

Posterior predictive simulation (PPS)

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

```
#R#JUS
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=0.14510478058433406]:1.08503639799450276-4)[theta=0.09971084629250842]
tree STATE_1000 = ((([theta=0.0934598083272793]:0.480964650605944-5,2[theta=0.08148563194855884]:0.480964650605944-5)[theta=0.08087174585362880]
tree STATE_2000 = ((([theta=0.1409414473026641]:0.482443330952113E-5,2[theta=0.07236732547290981]:0.482443330952113E-5)[theta=0.09247820488050332]
tree STATE_3000 = ((([theta=0.05079015898997618]:7.276296315931688E-5,2[theta=0.0971833136592398]:7.276296315931688E-5)[theta=0.0189191123343805]
tree STATE_4000 = ((([theta=0.11525227529250282]:0.208070097025403E-5,2[theta=0.0829382714605583]:0.208070097025403E-5)[theta=0.0186057892249781]
tree STATE_5000 = ((([theta=0.13539295988464923]:0.51354577291440E-5,2[theta=0.058178322747980]:0.51354577291440E-5)[theta=0.01881138072516937]
tree STATE_6000 = ((([theta=0.087780455263284]:0.78818945375756E-5,2[theta=0.1233987138280676]:0.78818945375756E-5)[theta=0.084515556978585]
tree STATE_7000 = ((([theta=0.112065673798073]:0.8284657246023E-5,2[theta=0.12027423830808153]:0.8284657246023E-5)[theta=0.018232213222556]
tree STATE_8000 = ((([theta=0.07708213696194869]:0.99808452603994E-5,2[theta=0.07840896355929689]:0.99808452603994E-5)[theta=0.0113080270808299]
tree STATE_9000 = ((([theta=0.114721911515843]:7.7922942162634E-5,2[theta=0.2040355302104985]:7.7922942162634E-5)[theta=0.01825732244418481]
tree STATE_10000 = ((([theta=0.1387454571202522]:0.407774627486132E-5,2[theta=0.094879334614088]:0.407774627486132E-5)[theta=0.016484462187230]
tree STATE_11000 = ((([theta=0.1119582886189198]:7.51619657015127E-5,2[theta=0.06183917835515891]:7.51619657015127E-5)[theta=0.08080941875978813]
tree STATE_12000 = ((([theta=0.1214428233927818]:2.10606181092240E-4,2[theta=0.108358284471212]:2.10606181092240E-4)[theta=0.08998224797432184]
tree STATE_13000 = ((([theta=0.0784481415208978]:2.748379670185987E-4,2[theta=0.03512482874688697]:2.748379670185987E-4)[theta=0.0186086230997]
tree STATE_14000 = ((([theta=0.1444974724999376]:7.14721454795101E-5,2[theta=0.1848758383554118]:7.14721454795101E-5)[theta=0.0894144426334269]
tree STATE_15000 = ((([theta=0.07812280870805]:6.13801292080489E-5,2[theta=0.118637043747322]:6.13801292080489E-5)[theta=0.087845204513908]
tree STATE_16000 = ((([theta=0.0668881392525222]:6.19957872049641E-5,2[theta=0.1725133983261884]:6.19957872049641E-5)[theta=0.089253739354384]
tree STATE_17000 = ((([theta=0.063385621664277]:7.18210213849485E-5,2[theta=0.062069598010152]:7.18210213849485E-5)[theta=0.0891329317079654]
tree STATE_18000 = ((([theta=0.1195815878338114]:7.18310213849485E-5,2[theta=0.091624891226695]:7.18310213849485E-5)[theta=0.01879372723529]
tree STATE_19000 = ((([theta=0.071356118438353]:6.78948197422256E-5,2[theta=0.092487186372587]:6.78948197422256E-5)[theta=0.018676746147448]
tree STATE_20000 = ((([theta=0.068465627333938]:2.46870800753707E-4,2[theta=0.134548840465518]:2.46870800753707E-4)[theta=0.0183351485380]
tree STATE_21000 = ((([theta=0.129659346575584]:2.6649806752533E-4,2[theta=0.0426183199218687]:2.6649806752533E-4)[theta=0.0185238994425]
tree STATE_22000 = ((([theta=0.1648517962864178]:2.37831148676808E-4,2[theta=0.0768973539172869]:2.37831148676808E-4)[theta=0.089285112717519]
tree STATE_23000 = ((([theta=0.049860784689593]:2.48737372283453E-4,2[theta=0.1452913646470158]:2.48737372283453E-4)[theta=0.0897268724772]
tree STATE_24000 = ((([theta=0.16893163914577329E-4,2[theta=0.18014592388693827]:1.6863163914577329E-4,2[theta=0.0878670660834
```

```
Simulation "State 0"
Begin
  N 1000
  Seed 123456789
  State 0
  Parameters
  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
  48
  49
  50
  51
  52
  53
  54
  55
  56
  57
  58
  59
  60
  61
  62
  63
  64
  65
  66
  67
  68
  69
  70
  71
  72
  73
  74
  75
  76
  77
  78
  79
  80
  81
  82
  83
  84
  85
  86
  87
  88
  89
  90
  91
  92
  93
  94
  95
  96
  97
  98
  99
  100
  101
  102
  103
  104
  105
  106
  107
  108
  109
  110
  111
  112
  113
  114
  115
  116
  117
  118
  119
  120
  121
  122
  123
  124
  125
  126
  127
  128
  129
  130
  131
  132
  133
  134
  135
  136
  137
  138
  139
  140
  141
  142
  143
  144
  145
  146
  147
  148
  149
  150
  151
  152
  153
  154
  155
  156
  157
  158
  159
  160
  161
  162
  163
  164
  165
  166
  167
  168
  169
  170
  171
  172
  173
  174
  175
  176
  177
  178
  179
  180
  181
  182
  183
  184
  185
  186
  187
  188
  189
  190
  191
  192
  193
  194
  195
  196
  197
  198
  199
  200
  201
  202
  203
  204
  205
  206
  207
  208
  209
  210
  211
  212
  213
  214
  215
  216
  217
  218
  219
  220
  221
  222
  223
  224
  225
  226
  227
  228
  229
  230
  231
  232
  233
  234
  235
  236
  237
  238
  239
  240
  241
  242
  243
  244
  245
  246
  247
  248
  249
  250
  251
  252
  253
  254
  255
  256
  257
  258
  259
  260
  261
  262
  263
  264
  265
  266
  267
  268
  269
  270
  271
  272
  273
  274
  275
  276
  277
  278
  279
  280
  281
  282
  283
  284
  285
  286
  287
  288
  289
  290
  291
  292
  293
  294
  295
  296
  297
  298
  299
  300
  301
  302
  303
  304
  305
  306
  307
  308
  309
  310
  311
  312
  313
  314
  315
  316
  317
  318
  319
  320
  321
  322
  323
  324
  325
  326
  327
  328
  329
  330
  331
  332
  333
  334
  335
  336
  337
  338
  339
  340
  341
  342
  343
  344
  345
  346
  347
  348
  349
  350
  351
  352
  353
  354
  355
  356
  357
  358
  359
  360
  361
  362
  363
  364
  365
  366
  367
  368
  369
  370
  371
  372
  373
  374
  375
  376
  377
  378
  379
  380
  381
  382
  383
  384
  385
  386
  387
  388
  389
  390
  391
  392
  393
  394
  395
  396
  397
  398
  399
  400
  401
  402
  403
  404
  405
  406
  407
  408
  409
  410
  411
  412
  413
  414
  415
  416
  417
  418
  419
  420
  421
  422
  423
  424
  425
  426
  427
  428
  429
  430
  431
  432
  433
  434
  435
  436
  437
  438
  439
  440
  441
  442
  443
  444
  445
  446
  447
  448
  449
  450
  451
  452
  453
  454
  455
  456
  457
  458
  459
  460
  461
  462
  463
  464
  465
  466
  467
  468
  469
  470
  471
  472
  473
  474
  475
  476
  477
  478
  479
  480
  481
  482
  483
  484
  485
  486
  487
  488
  489
  490
  491
  492
  493
  494
  495
  496
  497
  498
  499
  500
  501
  502
  503
  504
  505
  506
  507
  508
  509
  510
  511
  512
  513
  514
  515
  516
  517
  518
  519
  520
  521
  522
  523
  524
  525
  526
  527
  528
  529
  530
  531
  532
  533
  534
  535
  536
  537
  538
  539
  540
  541
  542
  543
  544
  545
  546
  547
  548
  549
  550
  551
  552
  553
  554
  555
  556
  557
  558
  559
  560
  561
  562
  563
  564
  565
  566
  567
  568
  569
  570
  571
  572
  573
  574
  575
  576
  577
  578
  579
  580
  581
  582
  583
  584
  585
  586
  587
  588
  589
  590
  591
  592
  593
  594
  595
  596
  597
  598
  599
  600
  601
  602
  603
  604
  605
  606
  607
  608
  609
  610
  611
  612
  613
  614
  615
  616
  617
  618
  619
  620
  621
  622
  623
  624
  625
  626
  627
  628
  629
  630
  631
  632
  633
  634
  635
  636
  637
  638
  639
  640
  641
  642
  643
  644
  645
  646
  647
  648
  649
  650
  651
  652
  653
  654
  655
  656
  657
  658
  659
  660
  661
  662
  663
  664
  665
  666
  667
  668
  669
  670
  671
  672
  673
  674
  675
  676
  677
  678
  679
  680
  681
  682
  683
  684
  685
  686
  687
  688
  689
  690
  691
  692
  693
  694
  695
  696
  697
  698
  699
  700
  701
  702
  703
  704
  705
  706
  707
  708
  709
  710
  711
  712
  713
  714
  715
  716
  717
  718
  719
  720
  721
  722
  723
  724
  725
  726
  727
  728
  729
  730
  731
  732
  733
  734
  735
  736
  737
  738
  739
  740
  741
  742
  743
  744
  745
  746
  747
  748
  749
  750
  751
  752
  753
  754
  755
  756
  757
  758
  759
  760
  761
  762
  763
  764
  765
  766
  767
  768
  769
  770
  771
  772
  773
  774
  775
  776
  777
  778
  779
  780
  781
  782
  783
  784
  785
  786
  787
  788
  789
  790
  791
  792
  793
  794
  795
  796
  797
  798
  799
  800
  801
  802
  803
  804
  805
  806
  807
  808
  809
  810
  811
  812
  813
  814
  815
  816
  817
  818
  819
  820
  821
  822
  823
  824
  825
  826
  827
  828
  829
  830
  831
  832
  833
  834
  835
  836
  837
  838
  839
  840
  841
  842
  843
  844
  845
  846
  847
  848
  849
  850
  851
  852
  853
  854
  855
  856
  857
  858
  859
  860
  861
  862
  863
  864
  865
  866
  867
  868
  869
  870
  871
  872
  873
  874
  875
  876
  877
  878
  879
  880
  881
  882
  883
  884
  885
  886
  887
  888
  889
  890
  891
  892
  893
  894
  895
  896
  897
  898
  899
  900
  901
  902
  903
  904
  905
  906
  907
  908
  909
  910
  911
  912
  913
  914
  915
  916
  917
  918
  919
  920
  921
  922
  923
  924
  925
  926
  927
  928
  929
  930
  931
  932
  933
  934
  935
  936
  937
  938
  939
  940
  941
  942
  943
  944
  945
  946
  947
  948
  949
  950
  951
  952
  953
  954
  955
  956
  957
  958
  959
  960
  961
  962
  963
  964
  965
  966
  967
  968
  969
  970
  971
  972
  973
  974
  975
  976
  977
  978
  979
  980
  981
  982
  983
  984
  985
  986
  987
  988
  989
  990
  991
  992
  993
  994
  995
  996
  997
  998
  999
  1000
End
Simulation "State 1000"
Begin
  N 1000
  Seed 123456789
  State 1000
  Parameters
  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
  48
  49
  50
  51
  52
  53
  54
  55
  56
  57
  58
  59
  60
  61
  62
  63
  64
  65
  66
  67
  68
  69
  70
  71
  72
  73
  74
  75
  76
  77
  78
  79
  80
  81
  82
  83
  84
  85
  86
  87
  88
  89
  90
  91
  92
  93
  94
  95
  96
  97
  98
  99
  100
  101
  102
  103
  104
  105
  106
  107
  108
  109
  110
  111
  112
  113
  114
  115
  116
  117
  118
  119
  120
  121
  122
  123
  124
  125
  126
  127
  128
  129
  130
  131
  132
  133
  134
  135
  136
  137
  138
  139
  140
  141
  142
  143
  144
  145
  146
  147
  148
  149
  150
  151
  152
  153
  154
  155
  156
  157
  158
  159
  160
  161
  162
  163
  164
  165
  166
  167
  168
  169
  170
  171
  172
  173
  174
  175
  176
  177
  178
  179
  180
  181
  182
  183
  184
  185
  186
  187
  188
  189
  190
  191
  192
  193
  194
  195
  196
  197
  198
  199
  200
  201
  202
  203
  204
  205
  206
  207
  208
  209
  210
  211
  212
  213
  214
  215
  216
  217
  218
  219
  220
  221
  222
  223
  224
  225
  226
  227
  228
  229
  230
  231
  232
  233
  234
  235
  236
  237
  238
  239
  240
  241
  242
  243
  244
  245
  246
  247
  248
  249
  250
  251
  252
  253
  254
  255
  256
  257
  258
  259
  260
  261
  262
  263
  264
  265
  266
  267
  268
  269
  270
  271
  272
  273
  274
  275
  276
  277
  278
  279
  280
  281
  282
  283
  284
  285
  286
  287
  288
  289
  290
  291
  292
  293
  294
  295
  296
  297
  298
  299
  300
  301
  302
  303
  304
  305
  306
  307
  308
  309
  310
  311
  312
  313
  314
  315
  316
  317
  318
  319
  320
  321
  322
  323
  324
  325
  326
  327
  328
  329
  330
  331
  332
  333
  334
  335
  336
  337
  338
  339
  340
  341
  342
  343
  344
  345
  346
  347
  348
  349
  350
  351
  352
  353
  354
  355
  356
  357
  358
  359
  360
  361
  362
  363
  364
  365
  366
  367
  368
  369
  370
  371
  372
  373
  374
  375
  376
  377
  378
  379
  380
  381
  382
  383
  384
  385
  386
  387
  388
  389
  390
  391
  392
  393
  394
  395
  396
  397
  398
  399
  400
  401
  402
  403
  404
  405
  406
  407
  408
  409
  410
  411
  412
  413
  414
  415
  416
  417
  418
  419
  420
  421
  422
  423
  424
  425
  426
  427
  428
  429
  430
  431
  432
  433
  434
  435
  436
  437
  438
  439
  440
  441
  442
  443
  444
  445
  446
  447
  448
  449
  450
  451
  452
  453
  454
  455
  456
  457
  458
  459
  460
  461
  462
  463
  464
  465
  466
  467
  468
  469
  470
  471
  472
  473
  474
  475
  476
  477
  478
  479
  480
  481
  482
  483
  484
  485
  486
  487
  488
  489
  490
  491
  492
  493
  494
  495
  496
  497
  498
  499
  500
  501
  502
  503
  504
  505
  506
  507
  508
  509
  510
  511
  512
  513
  514
  515
  516
  517
  518
  519
  520
  521
  522
  523
  524
  525
  526
  527
  528
  529
  530
  531
  532
  533
  534
  535
  536
  537
  538
  539
  540
  541
  542
  543
  544
  545
  546
  547
  548
  549
  550
  551
  5
```

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

```
#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=0.080871585362880]
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=0.0189919123343865]
tree STATE_4000 = ((([theta=0.115252725295282]:0.280870897825483-5,2[theta=0.083935271466558]:0.280870897825483-5)[theta=0.0186057892349781]
tree STATE_5000 = ((([theta=0.1353959588464923]:0.515154577291448E-5,2[theta=0.058178322747980]:0.515154577291448E-5)[theta=0.01881138072516937]
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=0.129274238388615]:7.0285457246236E-5)[theta=0.018232213222556]
tree STATE_8000 = ((([theta=0.0770821369619486]:0.99808452683994E-5,2[theta=0.0784898635592968]: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=0.011648448187230]
tree STATE_11000 = ((([theta=0.1159582886189198]:7.51619657015127E-5,2[theta=0.0618391783551589]:7.51619657015127E-5)[theta=0.0880894187578818]
tree STATE_12000 = ((([theta=0.121428283922498]:2.10686189392249E-4,2[theta=0.188358184471212]:2.10686189392249E-4)[theta=0.0899821479432184]
tree STATE_13000 = ((([theta=0.0784481415268978]:7.748379670185987E-4,2[theta=0.0351242874688667]:7.748379670185987E-4)[theta=0.0186888238997]
tree STATE_14000 = ((([theta=0.1444897474999378]:7.14721454795191E-5,2[theta=0.1848758383554118]:7.14721454795191E-5)[theta=0.089414442634260]
tree STATE_15000 = ((([theta=0.078122888788885]:6.138129288888E-5,2[theta=0.118613743174322]:6.138129288888E-5)[theta=0.08784520431988]
tree STATE_16000 = ((([theta=0.066888139252922]:6.19957872849641E-5,2[theta=0.1725135983261884]:6.19957872849641E-5)[theta=0.08925373935484]
tree STATE_17000 = ((([theta=0.083385526647077]:7.18310213849848E-5,2[theta=0.186206588888888]:7.18310213849848E-5)[theta=0.089129317897854]
tree STATE_18000 = ((([theta=0.1195815878388140]:7.18310213849848E-5,2[theta=0.091264891226695]:7.18310213849848E-5)[theta=0.018783797232529]
tree STATE_19000 = ((([theta=0.071356118438353]:6.78948197422256E-5,2[theta=0.092487186372587]:6.78948197422256E-5)[theta=0.018676746147488]
tree STATE_20000 = ((([theta=0.068845627339385]:2.46748888753978E-4,2[theta=0.134548848455813]:2.46748888753978E-4)[theta=0.013353185380]
tree STATE_21000 = ((([theta=0.1296593946575584]:2.66498867525333E-4,2[theta=0.0426183199218887]:2.66498867525333E-4)[theta=0.0185328894425]
tree STATE_22000 = ((([theta=0.1648519628645178]:2.37831148676888E-4,2[theta=0.078873591728691]:2.37831148676888E-4)[theta=0.08938511271519]
tree STATE_23000 = ((([theta=0.049860784689593]:2.4873737228343E-4,2[theta=0.145278364476155]:2.4873737228343E-4)[theta=0.08728724772]
tree STATE_24000 = ((([theta=0.169827835559487]:1.6861631914577739E-4,2[theta=0.18814592888693827]:1.6861631914577739E-4)[theta=0.087867686834
```

```
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
```

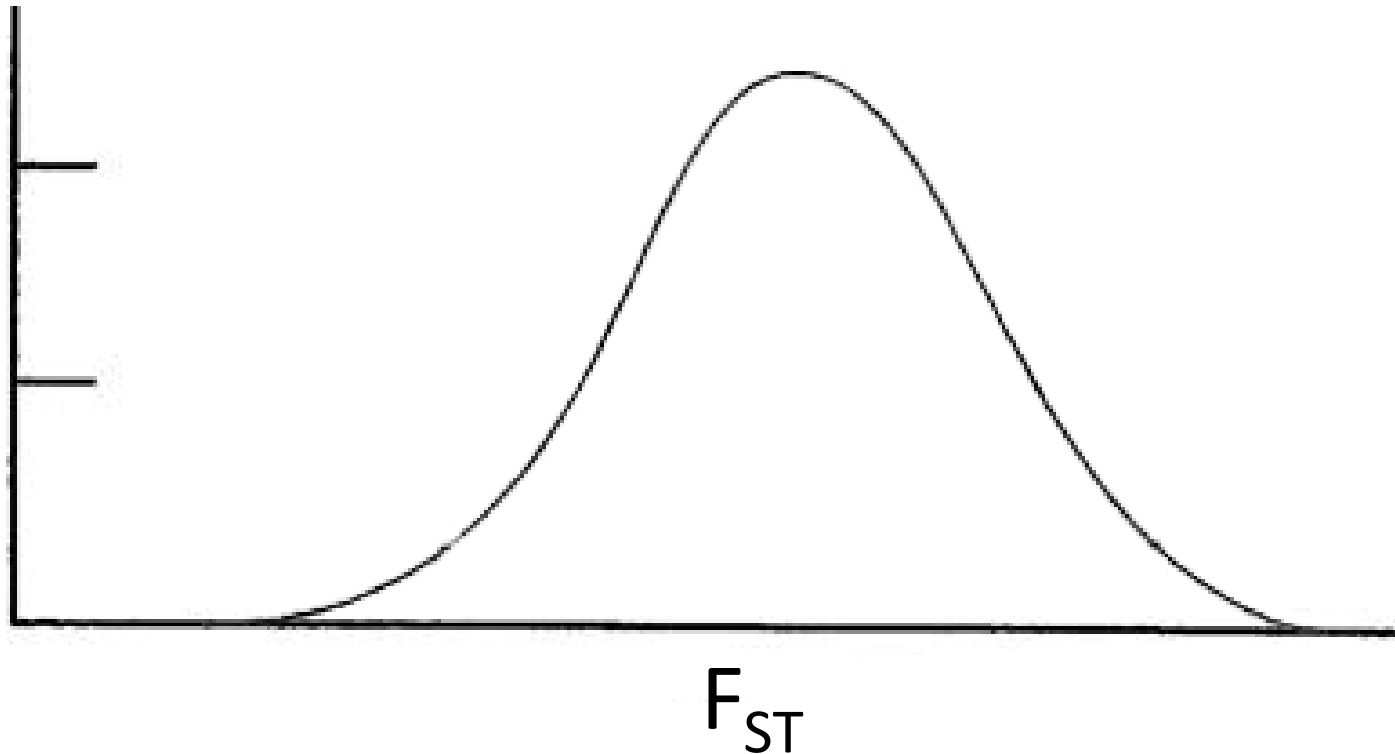
Posterior predictive distribution



Summary statistic

Posterior predictive distribution (PPD)

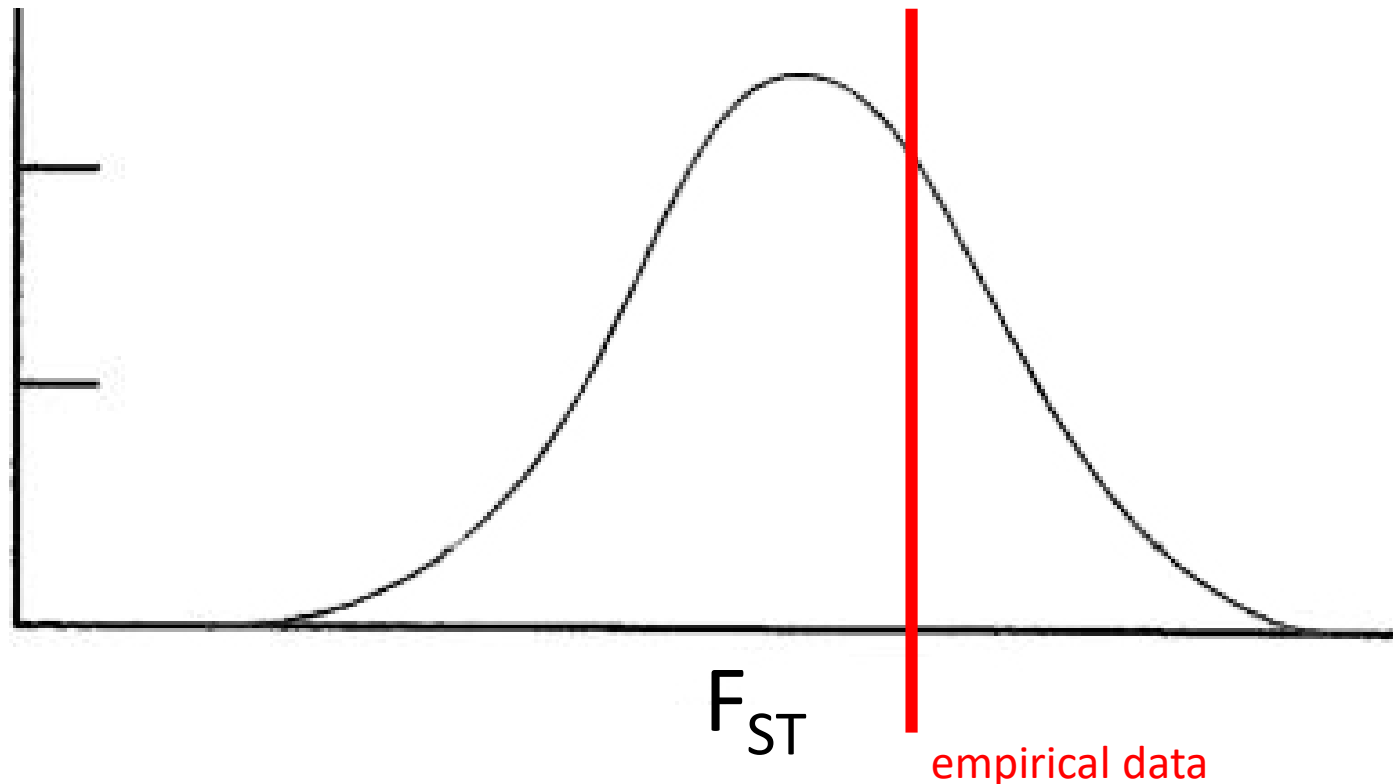
- How well does the empirical data fit this distribution?



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)

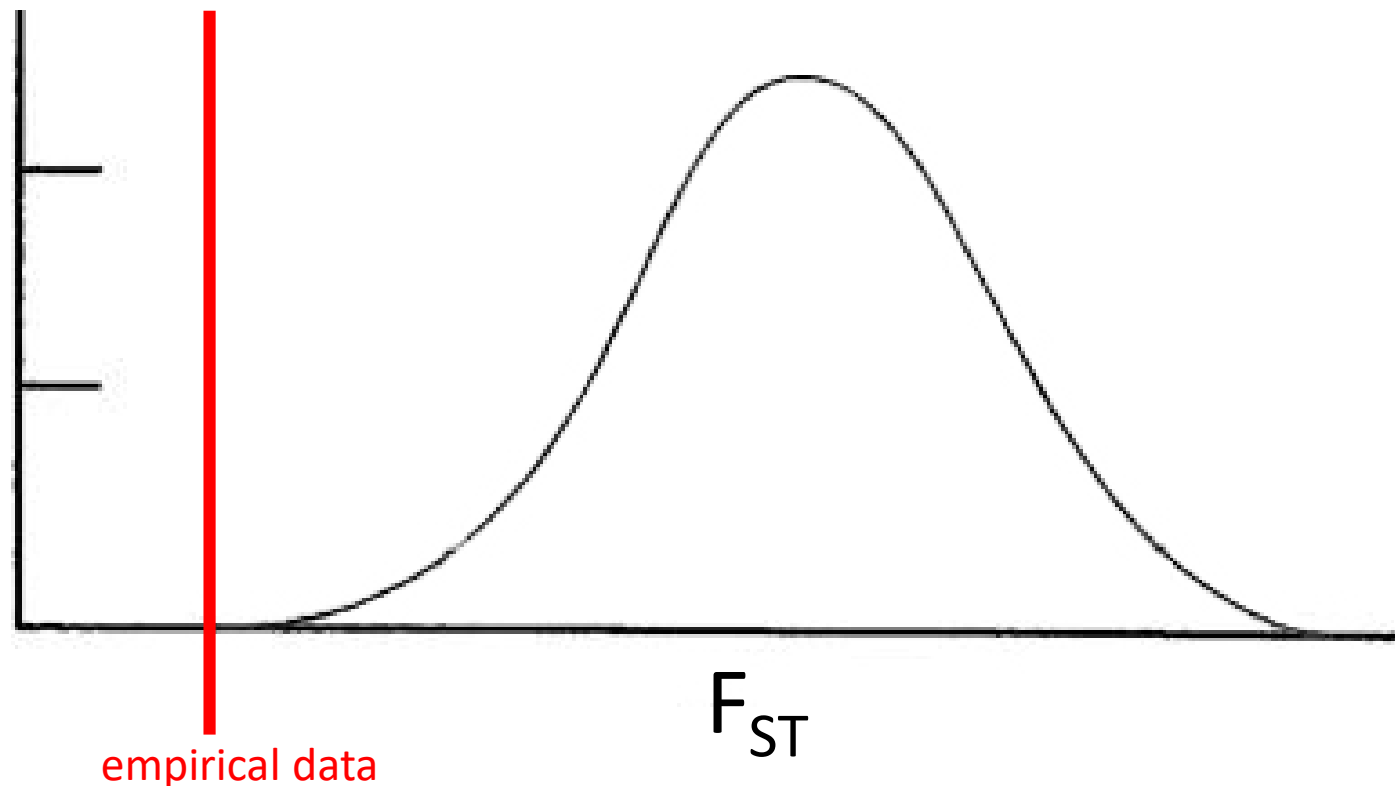
- How well does the empirical data fit this distribution?



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)

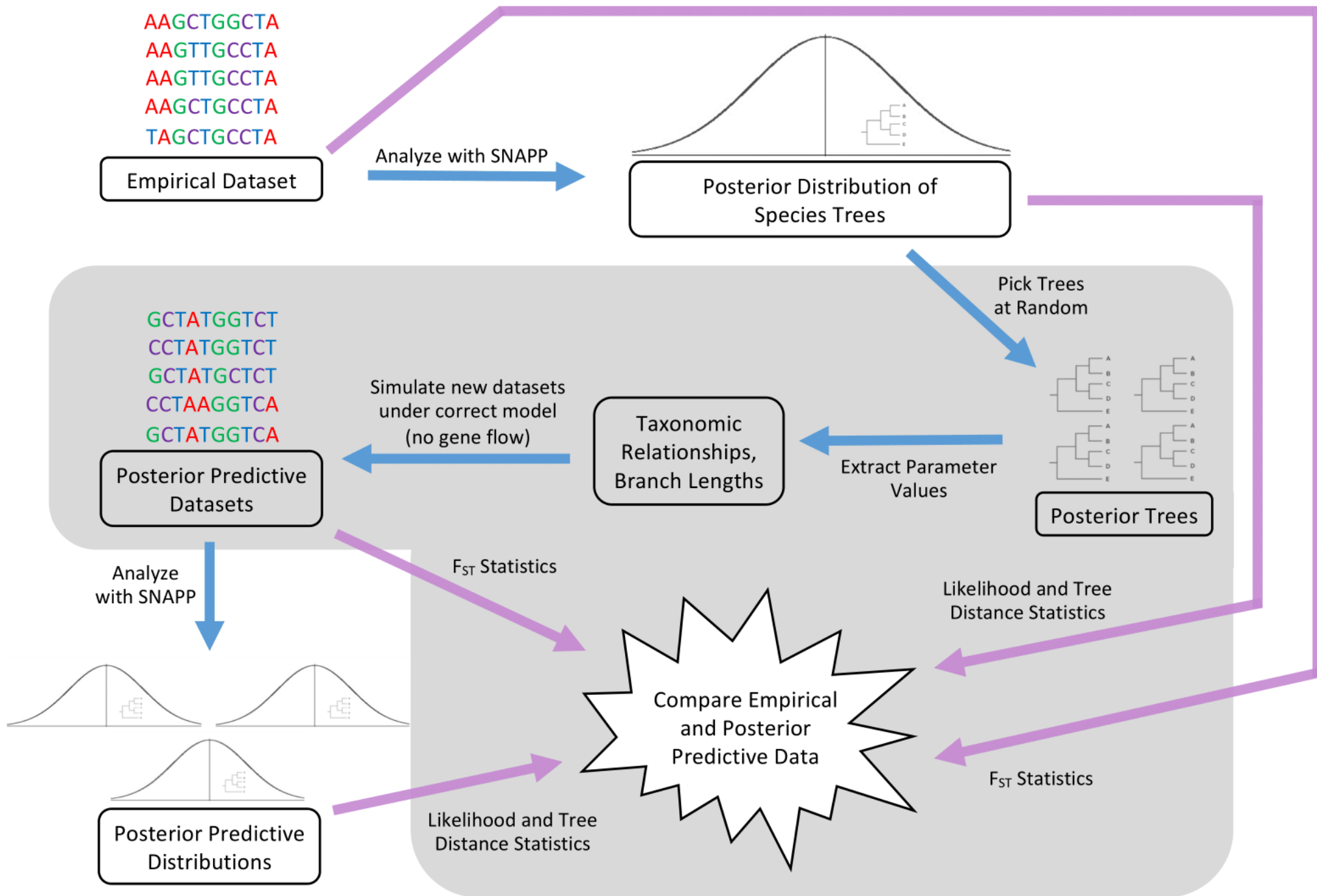
- Interfaces with the BEAST package
- Algorithm bypasses the gene trees and computes species tree likelihoods directly from the markers
- Returns a sample of species trees with (relative) divergence times and population sizes (posterior distribution)

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

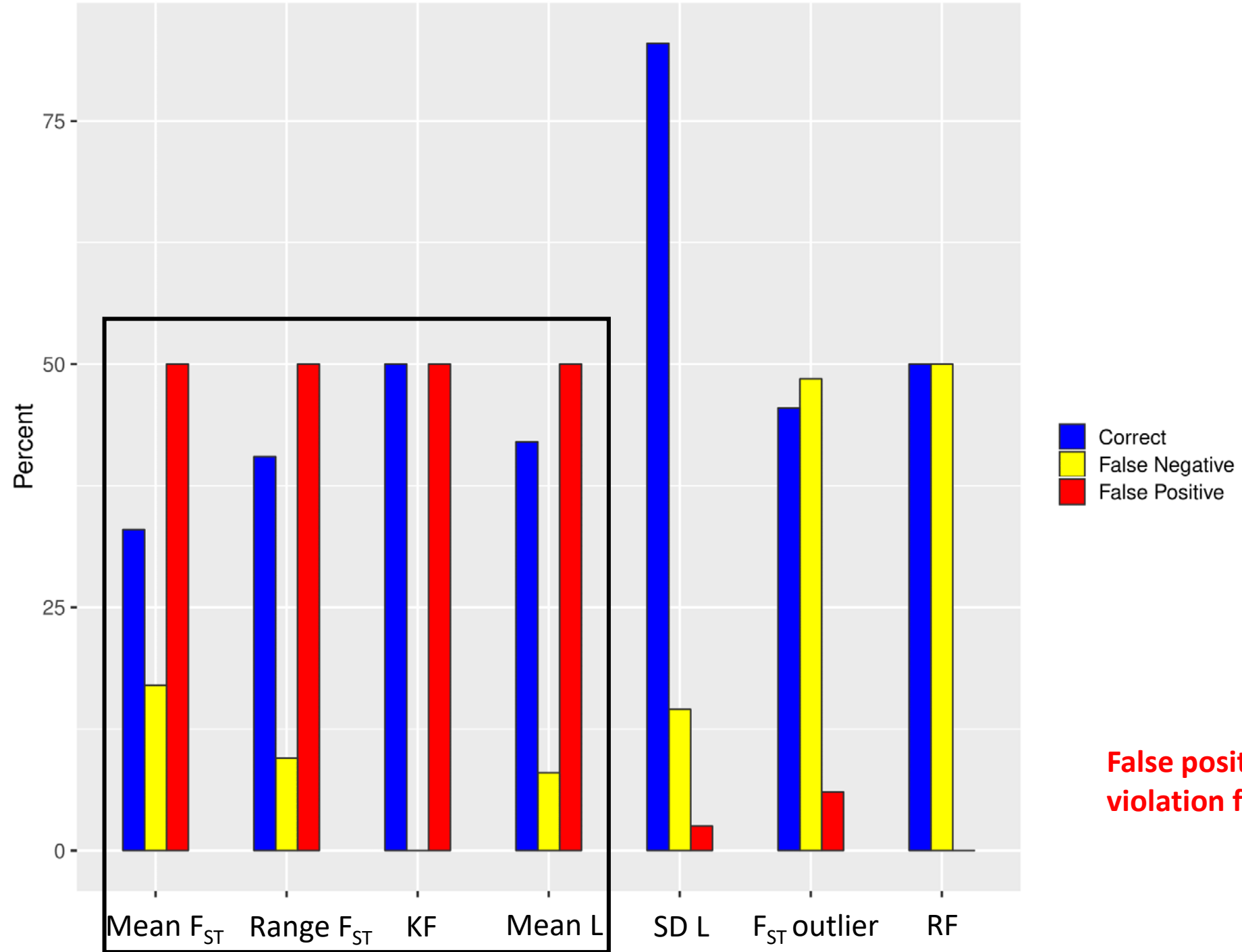
- **Summary statistics tested:**

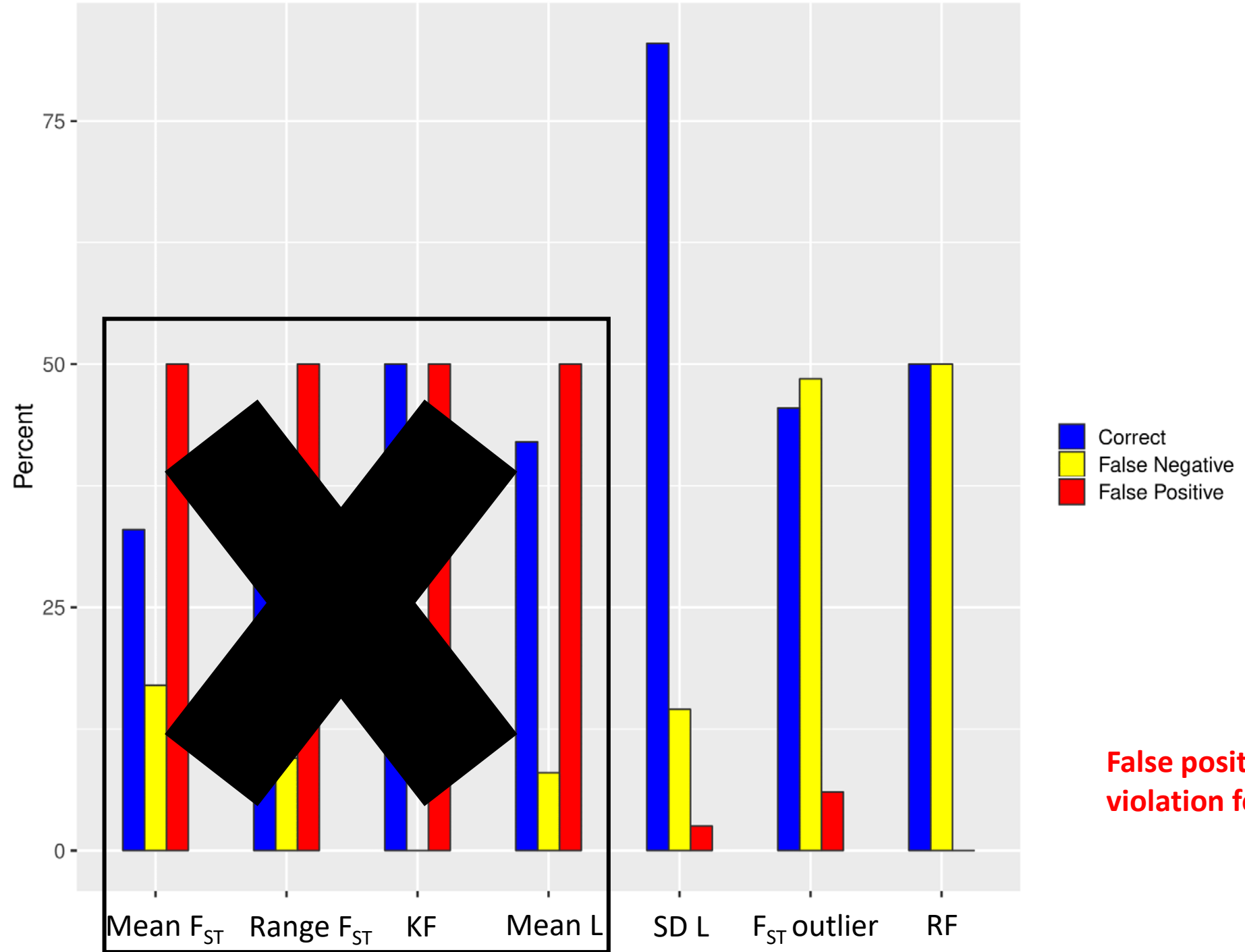
- Mean pairwise F_{ST}
- Range of pairwise F_{ST}
- F_{ST} outlier test
- Robinson-Foulds distance (topological distance only)
- Kuhner-Felsenstein distance (includes branch lengths)
- Mean of tree likelihood
- Standard deviation of tree likelihood

P2C2M.SNAPP simulation testing

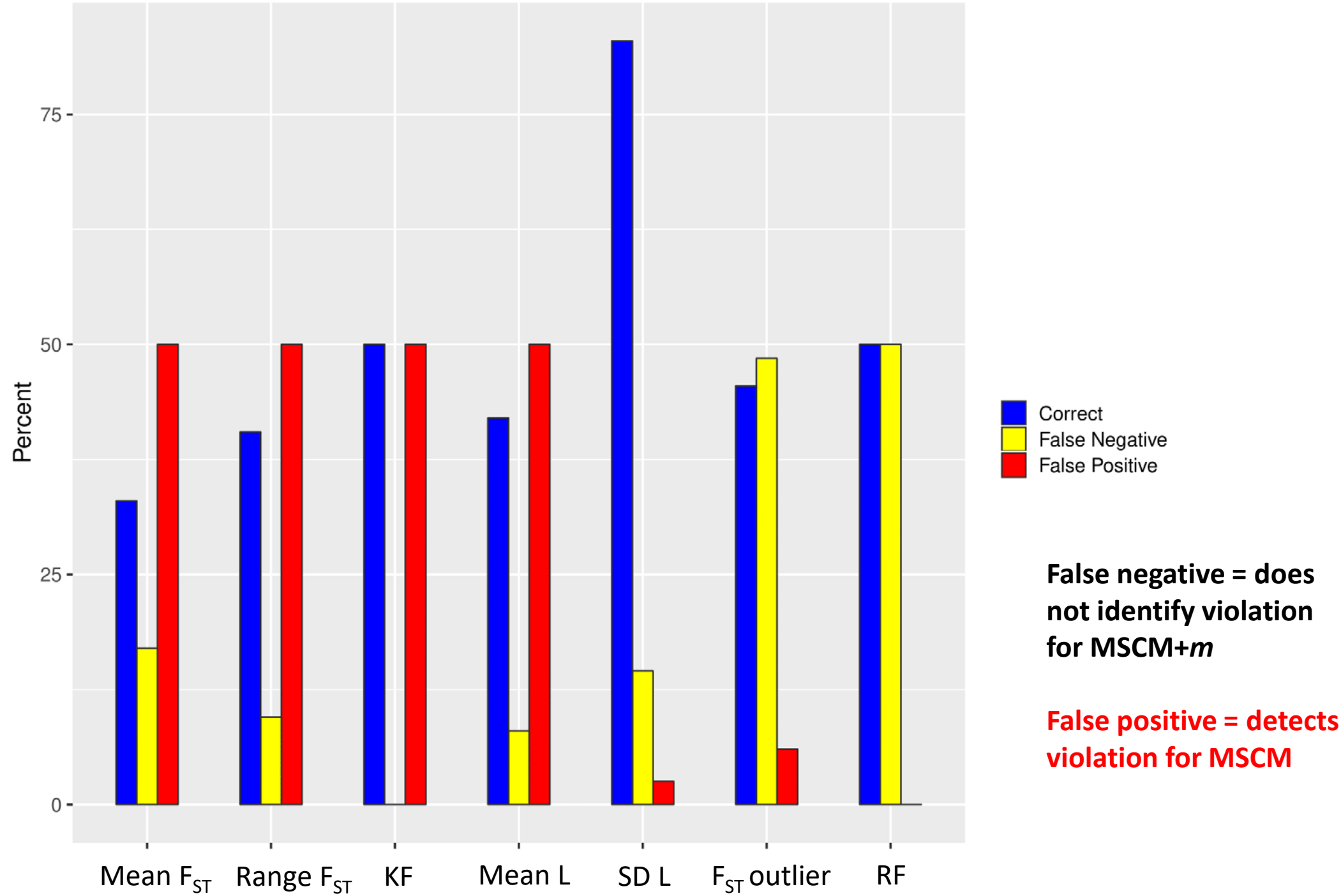
- **Summary statistics tested:**

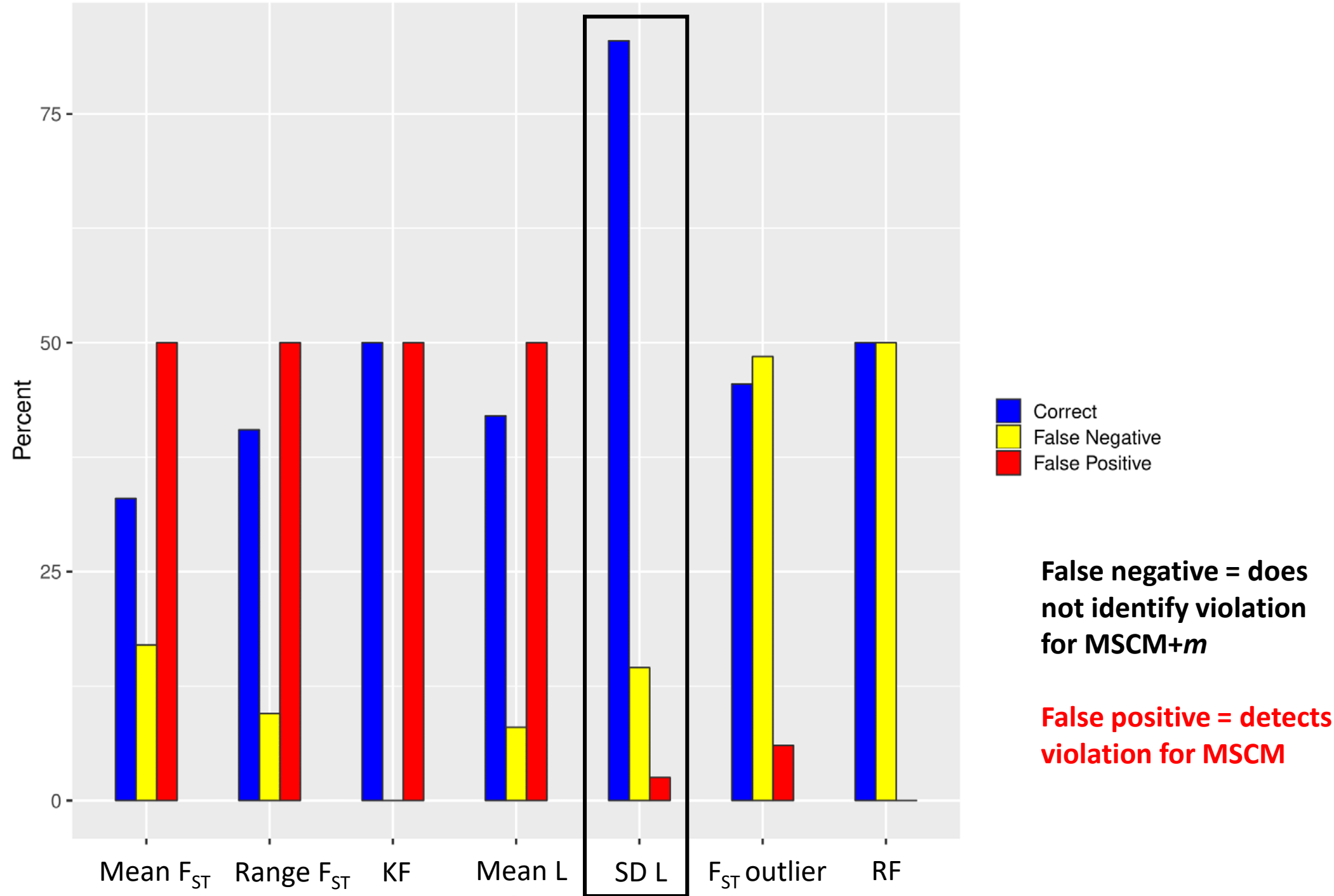
- ~~• Mean pairwise F_{ST}~~
- ~~• Range of pairwise F_{ST}~~
- **F_{ST} outlier test**
- **Robinson-Foulds distance (topological distance only)**
- ~~• Kuhner-Felsenstein distance (includes branch lengths)~~
- ~~• Mean of tree likelihood~~
- **Standard deviation of tree likelihood**

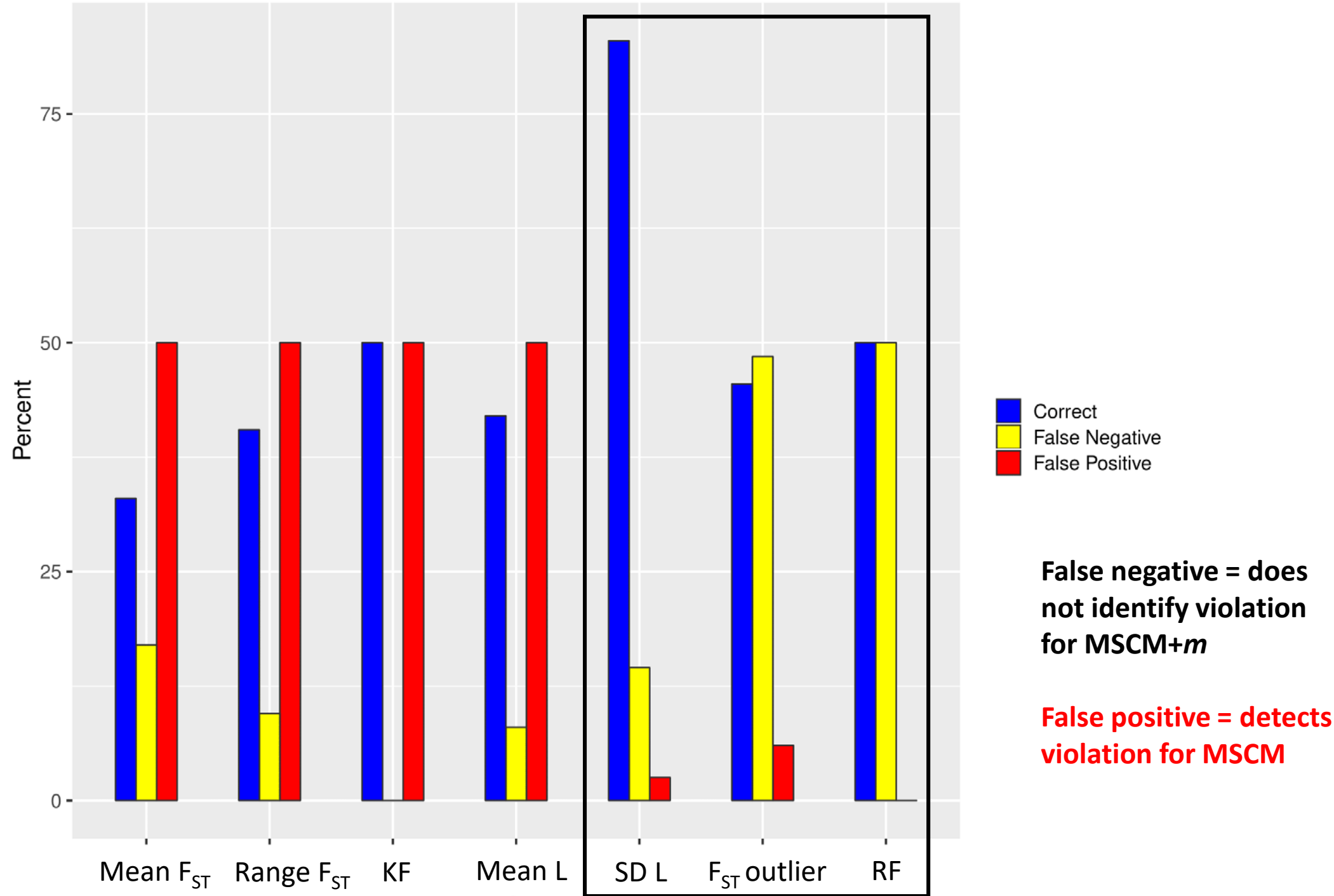




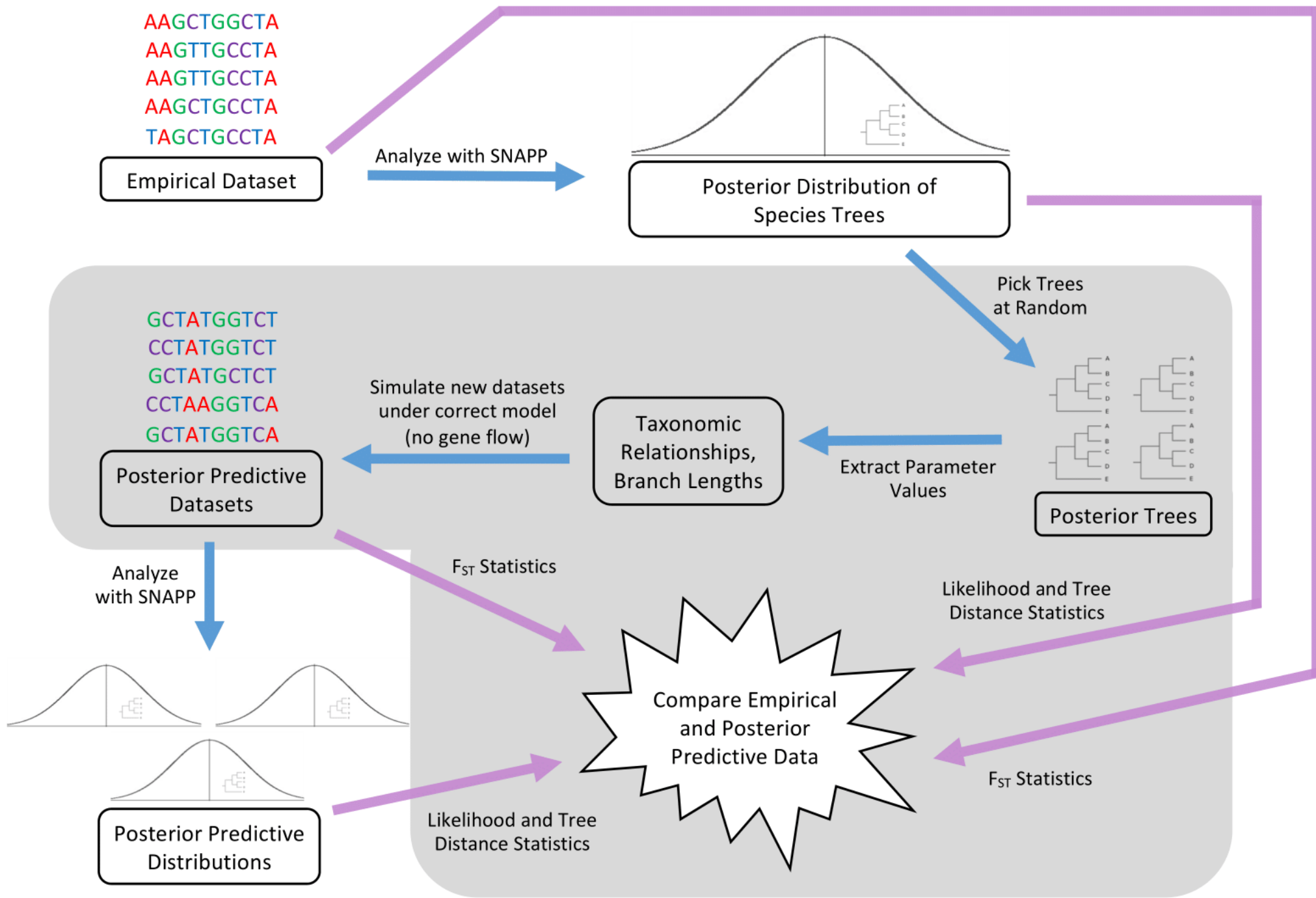
False positive = detects violation for MSCM







- **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