BEAST2 tree (space) tools
Convergence + Means

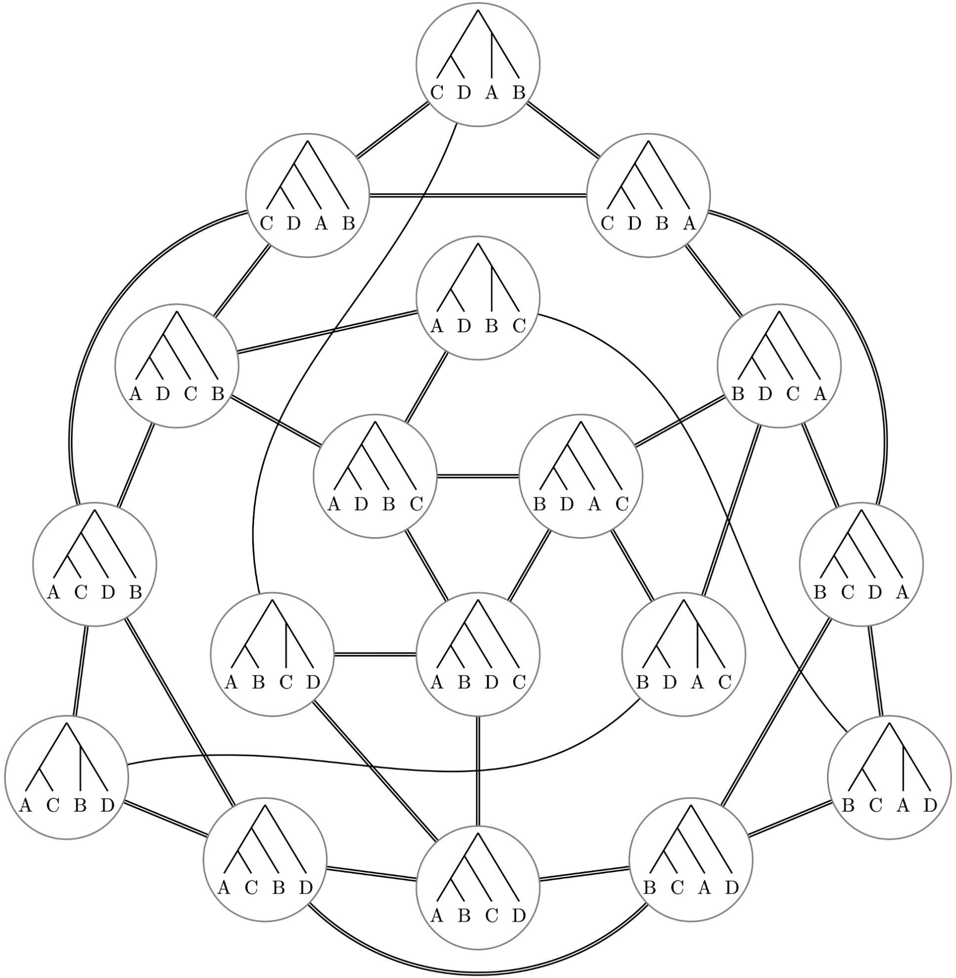
ASM and CCD

Lars Berling

Joint work with: Remco Bouckaert, Lena Collienne, Alexei Drummond, Alex Gavryushkin, Jonathan Klawitter, Walter Xie

RNNI space with 4 taxa

- Rank moves – single edges
- NNI moves – double edges



Convergence of Markov chain Monte Carlo

- In theory MCMC converges to the target distribution asymptotically
- Consider: convergence = indistinguishable samples form independent runs
- Phylogenetics rule of thumb for effective sample size (ESS):
  - ESS < 100 is bad
  - ESS > 100 is okay
  - ESS > 200 is converged
  - !NOTE: ESS is only for continuous parameter traces

→ No one is checking the trees!

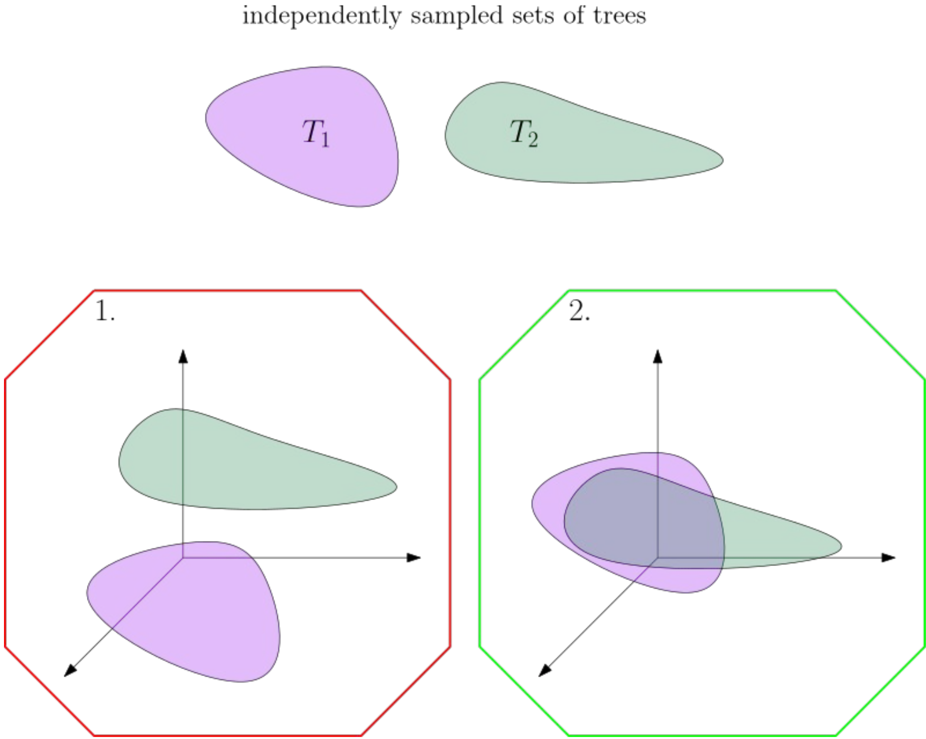
The ASM package automates best practises and adds tree space based diagnosis

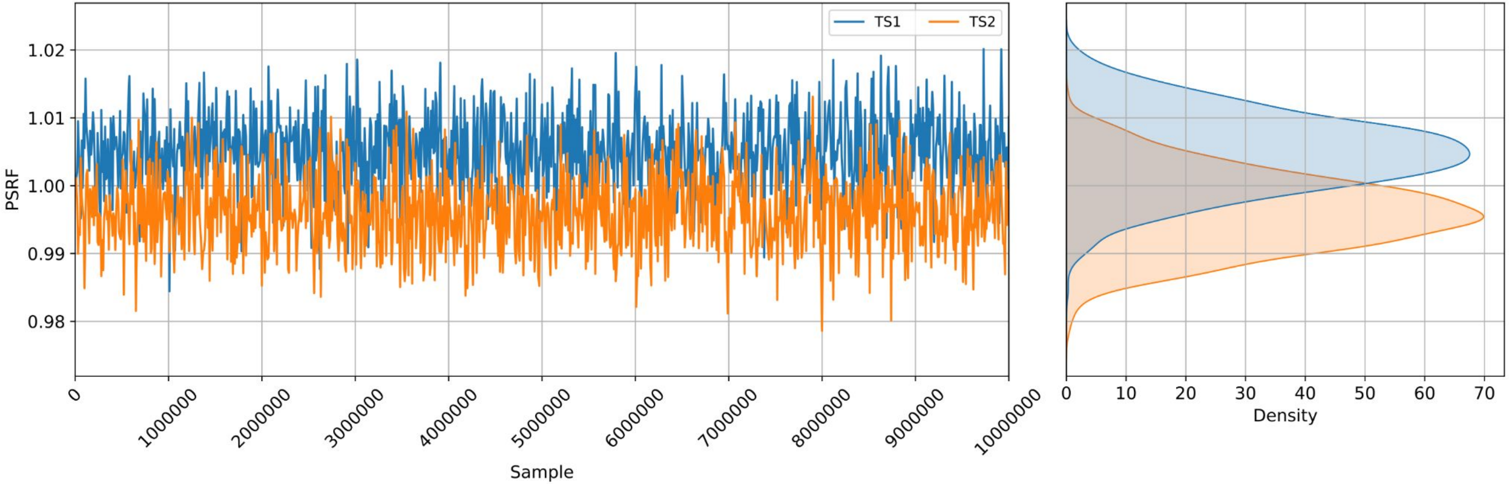
Use the RNNI space to define a convergence diagnostic:

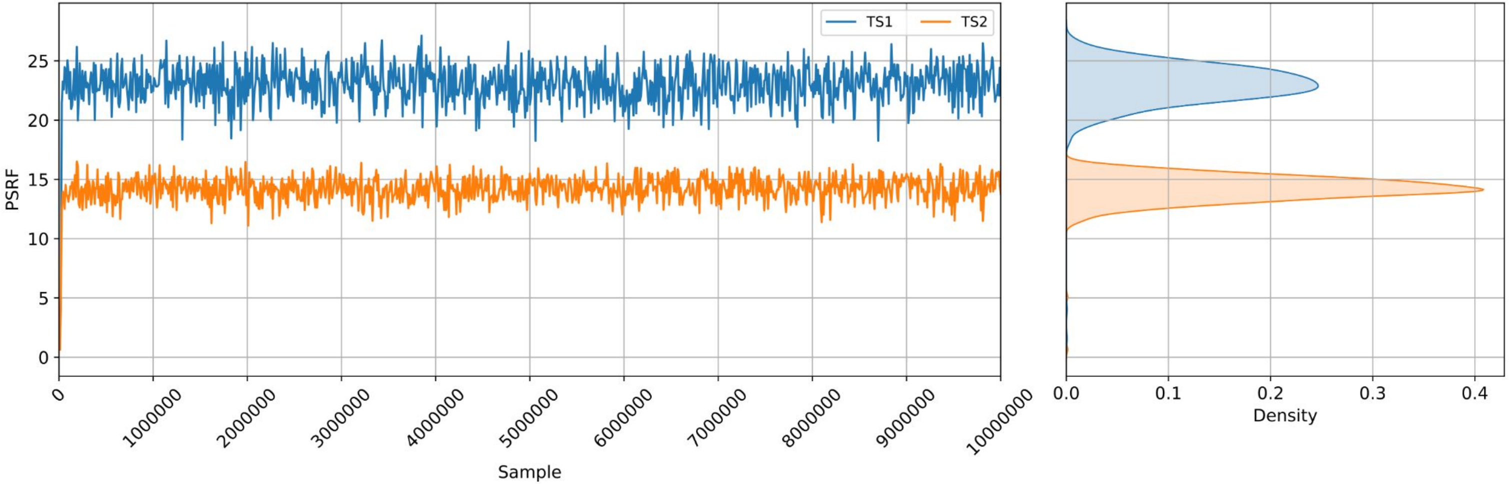
$$\mathbf{Var}(t)_{\mathcal{T}} = \frac{\Sigma_{t_i \in \mathcal{T}} d(t_i, t)^2}{|\mathcal{T}|}.$$

$$\mathbf{PSRF}(t; \mathcal{T}_1, \mathcal{T}_2) = \sqrt{\frac{\mathbf{Var}(t)_{\mathcal{T}_2}}{\mathbf{Var}(t)_{\mathcal{T}_1}}}, t \in \mathcal{T}_1$$

$$\mathbf{GR}(t_i^1; \mathcal{T}_1, \mathcal{T}_2) = \frac{1}{i} \sum_{s=1}^{i} \mathbf{PSRF}(t_s; \mathcal{T}_1, \mathcal{T}_2)$$




TS1
TS2
PSRF
1.02
1.01
1.00
0.99
0.98
Sample
0
1000000
2000000
3000000
4000000
5000000
6000000
7000000
8000000
9000000
10000000
Density
0
10
20
30
40
50
60
70


TS1
TS2
PSRF
25
20
15
10
5
0
Sample
0
1000000
2000000
3000000
4000000
5000000
6000000
7000000
8000000
9000000
10000000
Density
0.0
0.1
0.2
0.3
0.4

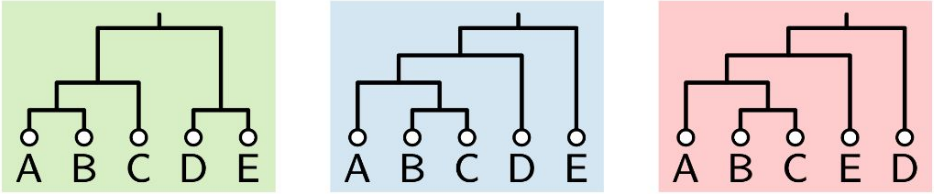
Summary/Consensus/Mean tree

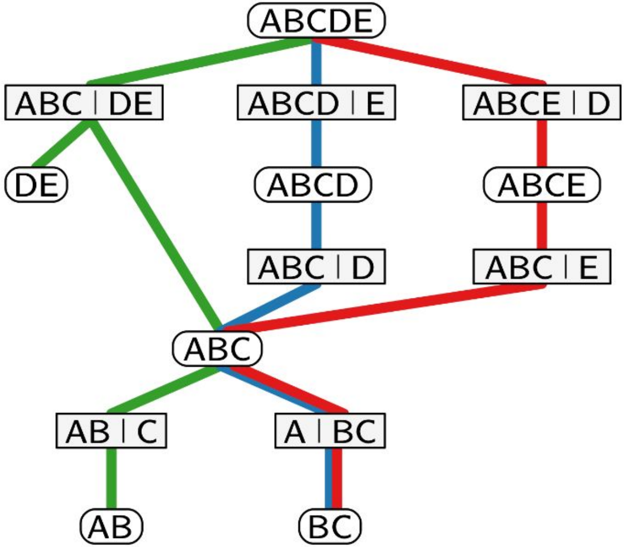
- Trees are now only topologies
- Tractable tree distribution:
  - A simple example would be a MCMC sample
  - Here we use: Conditional clade distributions (CCDs)
- We want an accurate representation of the posterior
- Use a MCMC sample to generate a CCD

CCD graph & CCD1

- Based on observed clade splits

sampled trees

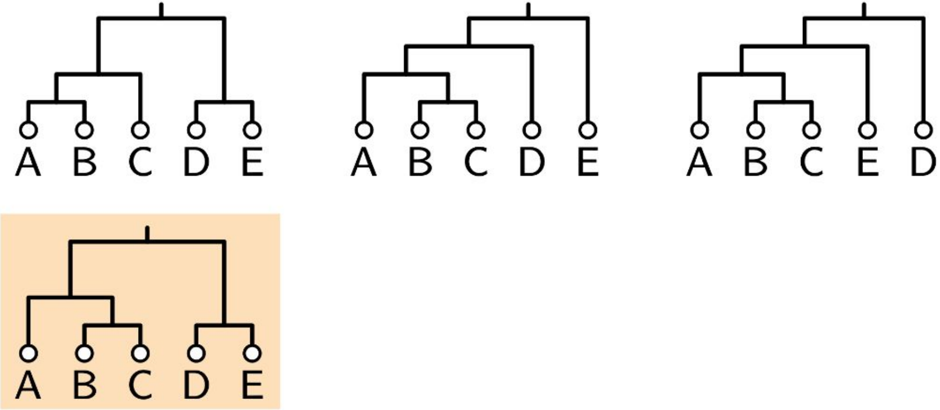


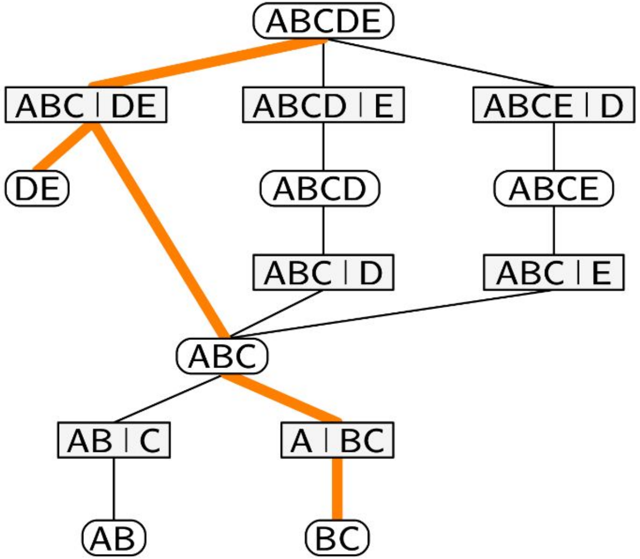


CCD graph & CCD1

- Based on observed clade splits

sampled trees




ABCDE

ABC | DE
ABCD | E
ABCE | D

DE
ABCD
ABCE

ABC | D
ABC | E

ABC

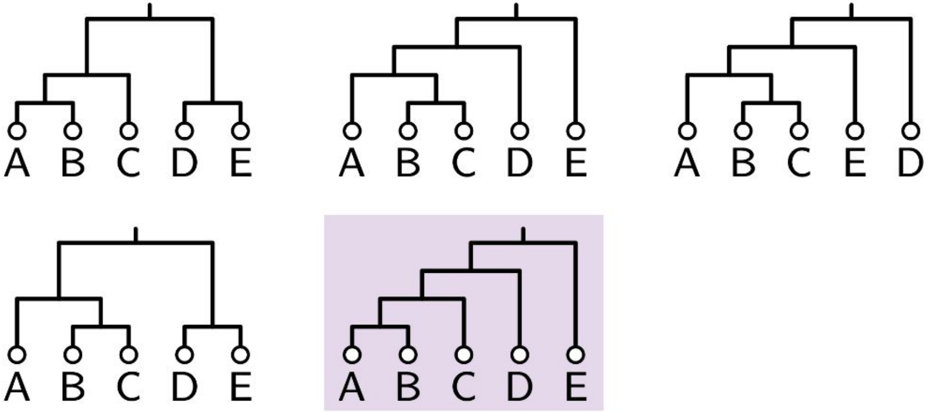
AB | C
A | BC

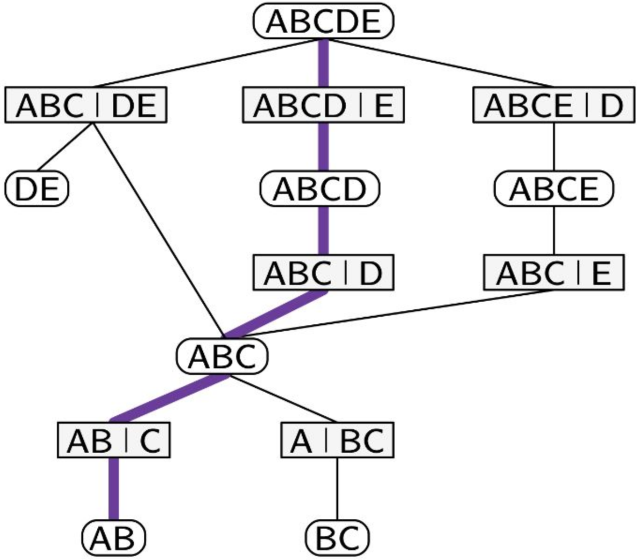
AB
BC

CCD graph & CCD1

- Based on observed clade splits

sampled trees

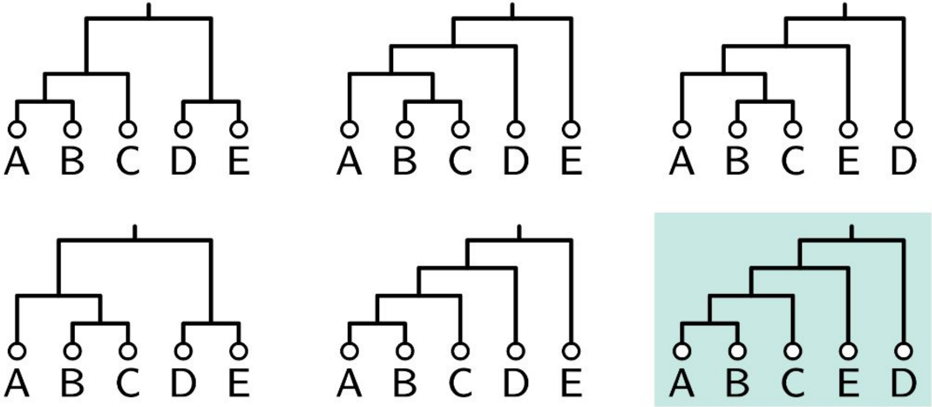


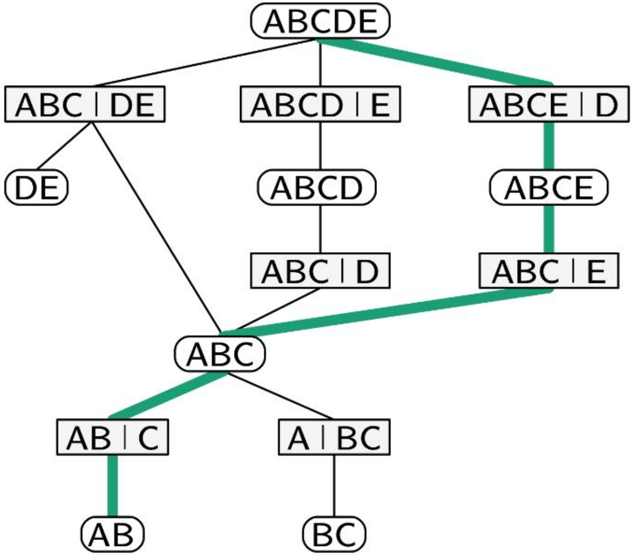


CCD graph & CCD1

- Based on observed clade splits

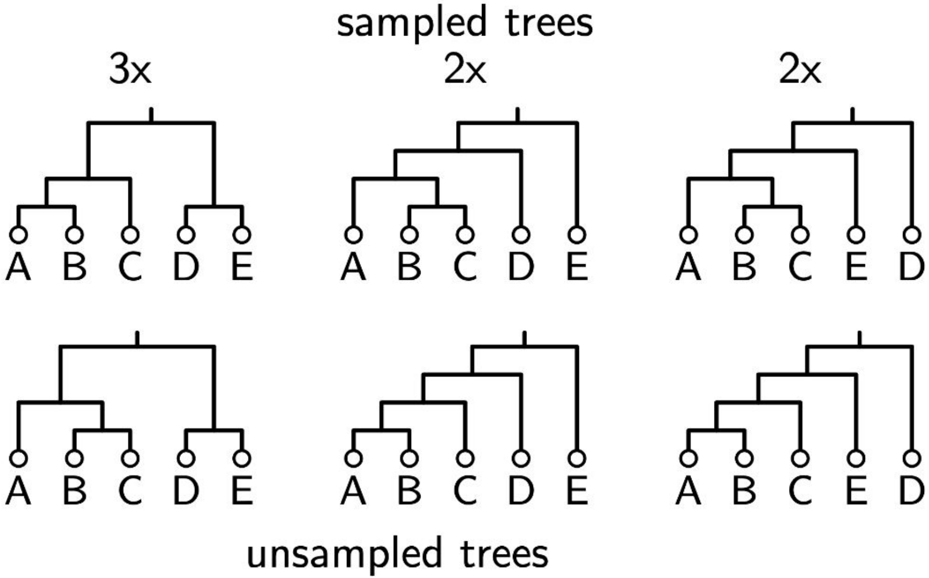
sampled trees

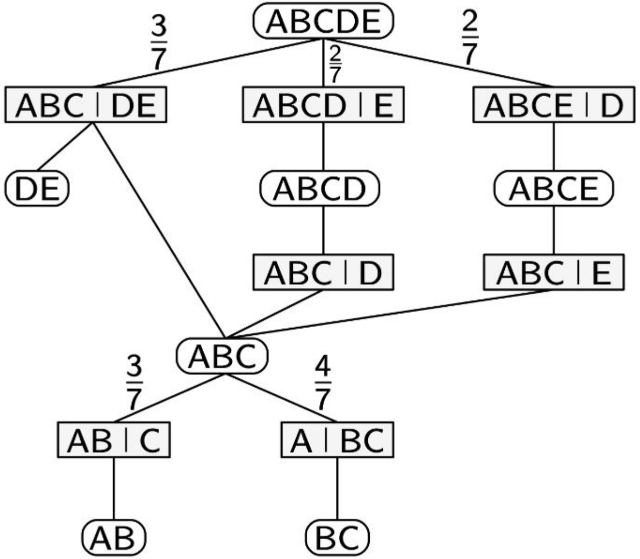




CCD graph & CCD1

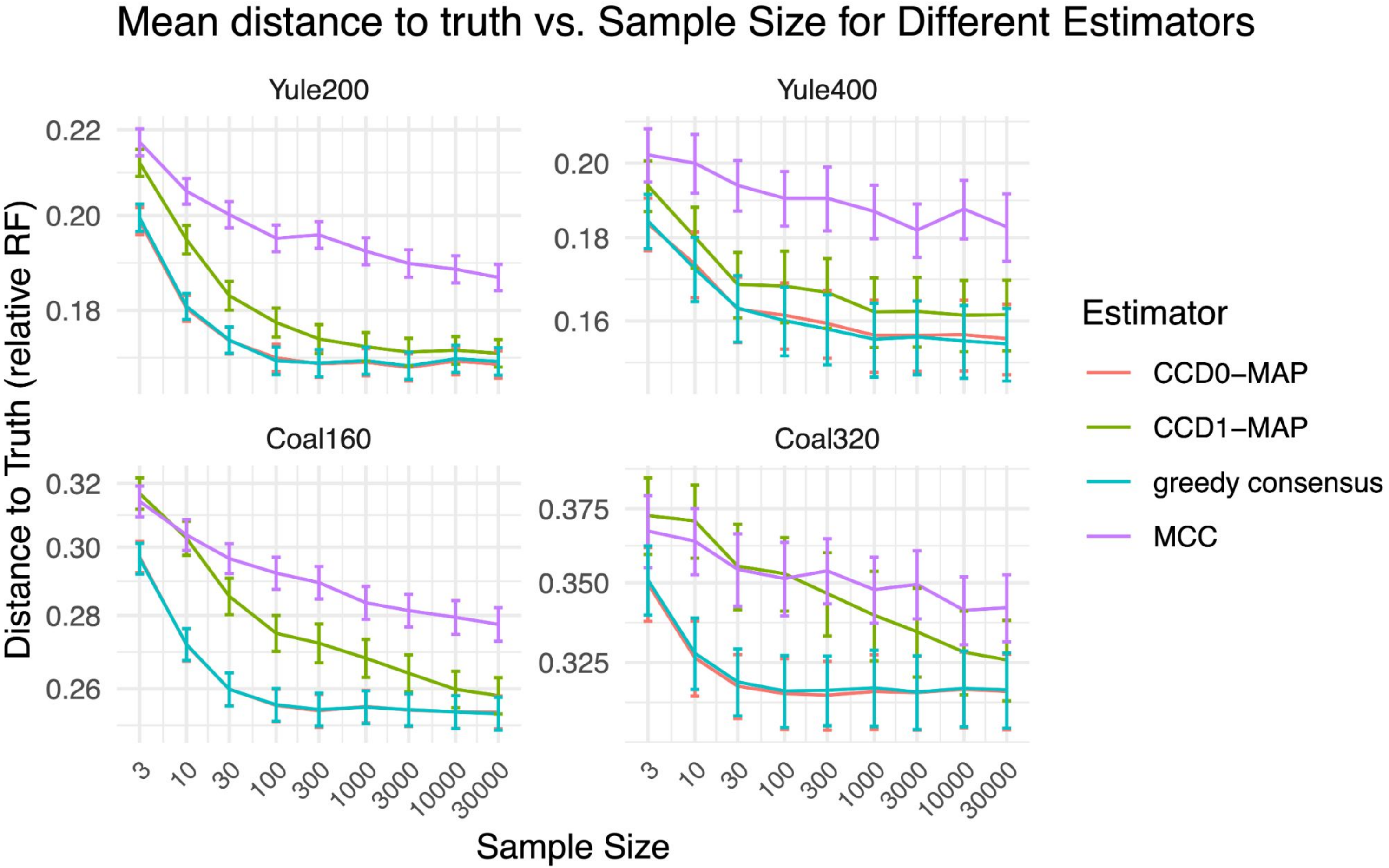
- Based on observed clade splits



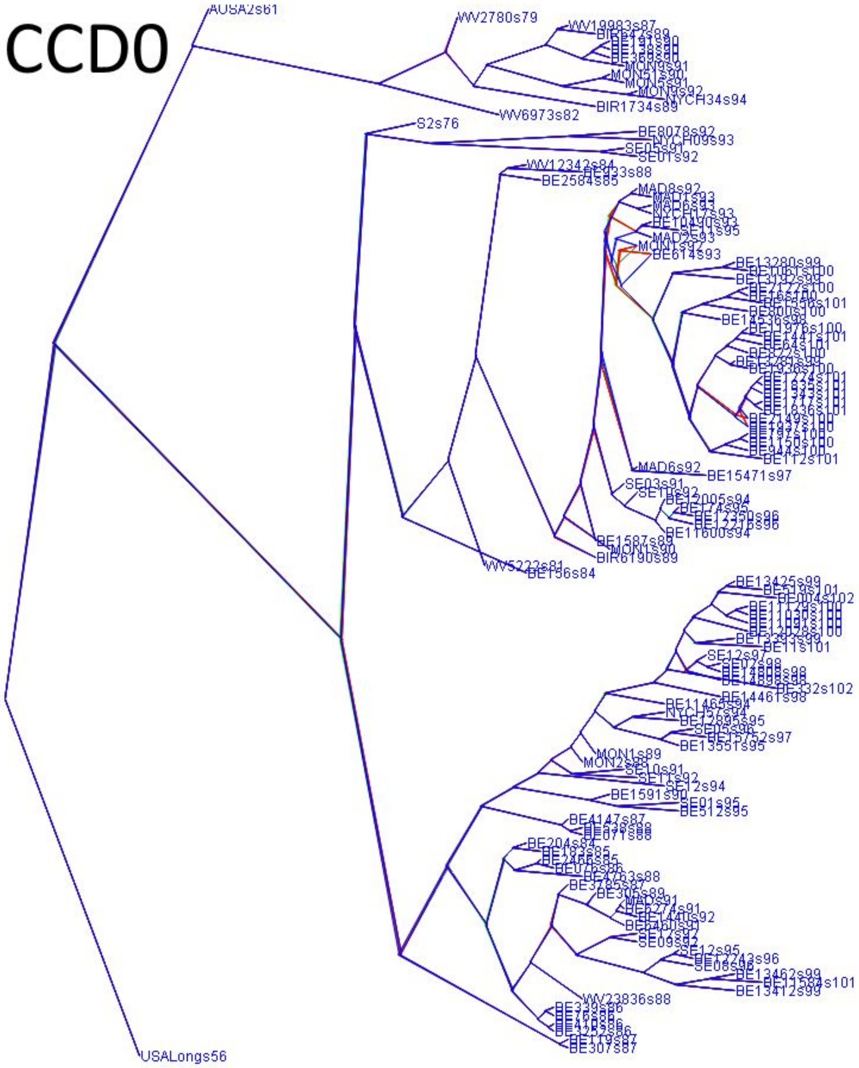


- Tree probabilities . . .

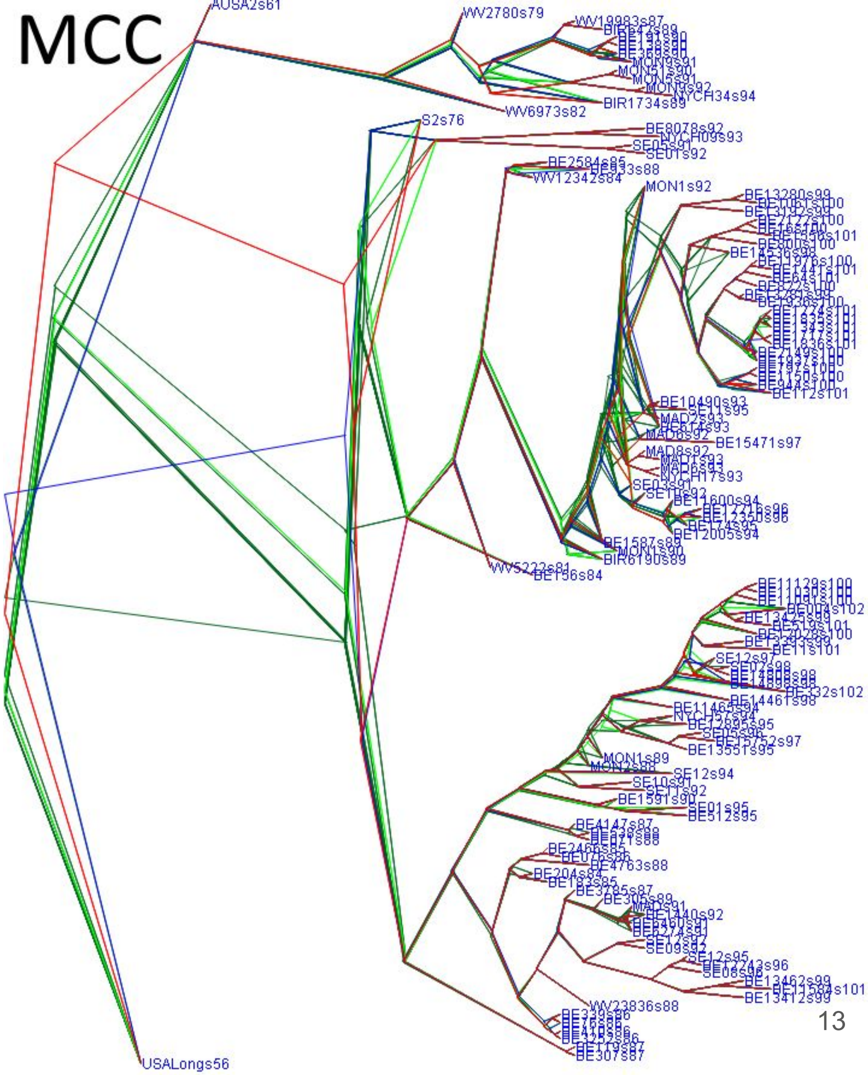
Accuracy of CCD MAP trees



CCD0


AUSA2s61
WV2780s79
WV19983s87
MON51s90
MON9s92
NYCH34s94
BIR1734s89
WV6973s82
S2s76
BE8078s92
NYCH09s93
SE05s91
SE01s92
WV12342s84
BE933s88
BE2584s85
MAD8s92
MAD1s93
MAD2s93
MON1s92
BE614s93
BE13280s99
BE11061s100
BE13192s99
MAD6s92
BE15471s97
SE03s91
SE11s92
BE12005s94
BE174s95
BE12350s96
BE11600s94
BE1587s89
MON1s90
BIR6190s89
WV5222s81
BE156s84
BE13425s99
BE004s102
BE11s101
SE12s97
BE14461s98
BE232s102
BE11465s94
NYCH17s94
BE12695s95
SE05s96
BE15752s97
BE13551s95
MON1s89
MON2s88
SE10s91
SE11s92
SE12s94
BE1591s90
SE01s95
BE512s95
BE4147s87
BE071s88
BE204s84
BE4763s88
BE3785s87
BE305s89
BE6274s91
BE1440s92
SE09s92
SE12s95
BE12743s96
SE08s96
BE13462s99
BE11584s101
BE13412s99
WV23836s88
BE339s86
BE307s87
USALongs56

MCC


AUSA2s61
WV2780s79
WV19983s87
MON9s91
MON51s90
MON9s92
NYCH34s94
BIR1734s89
WV6973s82
S2s76
BE8078s92
NYCH09s93
SE05s91
SE01s92
BE2584s85
BE933s88
WV12342s84
MON1s92
BE13280s99
BE11061s100
BE13192s99
BE10490s93
SE01s95
MAD2s93
BE614s93
MAD6s92
BE15471s97
MAD8s92
MAD1s93
NYCH17s93
SE03s91
SE11s92
BE11600s94
BE12716s96
BE12350s96
BE174s95
BE12005s94
BE1587s89
MON1s90
BIR6190s89
WV5222s81
BE156s84
BE11129s100
BE004s102
BE11s101
SE12s97
BE14461s98
BE232s102
BE11465s94
NYCH17s94
BE12695s95
BE15752s97
BE13551s95
MON1s89
MON2s88
SE12s94
SE10s91
SE11s92
BE1591s90
SE01s95
BE512s95
BE4147s87
BE071s88
BE2466s85
BE4763s88
BE204s84
BE183s85
BE3785s87
BE305s89
BE1440s92
BE6274s91
SE09s92
SE12s95
BE12743s96
SE08s96
BE13462s99
BE11584s101
BE13412s99
WV23836s88
BE339s86
BE307s87
USALongs56

- Check trees for convergence!
- Don’t use the MCC tree!

Thank you!

ASM package

"An automated convergence diagnostic for phylogenetic MCMC analyses." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2024)

CCD package

"Accurate Bayesian phylogenetic point estimation using a tree distribution parameterized by clade probabilities." *PLOS Computational Biology (2025)*