SNAPP Analysis of Algonquian Language Data

Ben Teo

Cognate data

Cognates: words with common etymology

Not cognates: “emoticon” (English) & “絵文字” (Japanese) 😟
Cognates: “night” (English), “nacht” (Dutch, German)

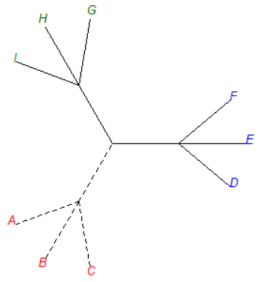
E.g. Suppose we have 9 languages (A, B, C, …, H, I), and 1 meaning, say “Fire”:

- Languages A-C use words for “Fire” that are cognates.
- Languages D-F …
- Languages G-I …

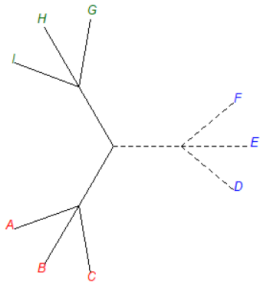
We say that there are 3 *cognate classes* for “Fire”.

This is coded as:

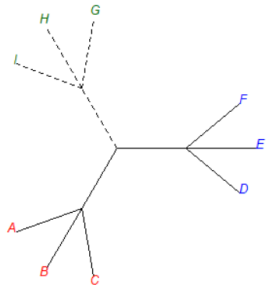
Language	Site1	Site2	Site3
A	1	0	0
B	1	0	0
C	1	0	0
D	0	1	0
E	0	1	0
F	0	1	0
G	0	0	1
H	0	0	1
I	0	0	1



Site 1



Site 2



Site 3

SNAPP (SNP and AFLP Package for Phylogenies)

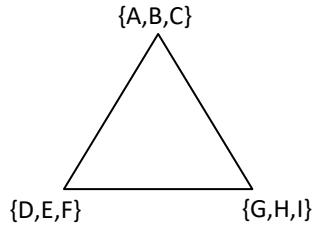
- Models the evolution of independent biallelic markers under the MSC process.
- Samples from the posterior species-tree distribution more efficiently!
- Implemented as a BEAST2 package!

MSC re-cap

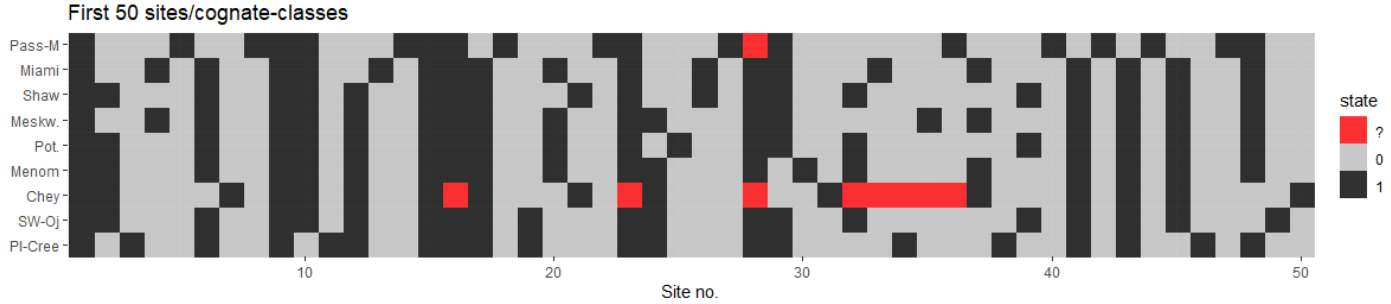
- Individual markers evolve along their respective gene-trees.
- Gene-trees are embedded within a common species-tree.
- Accounts for within-population variation due to ILS.

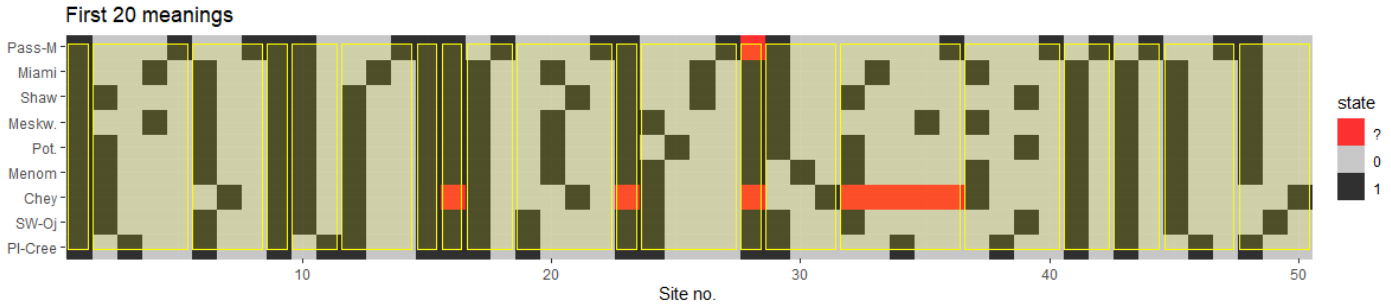
Applying SNAPP to Cognate data

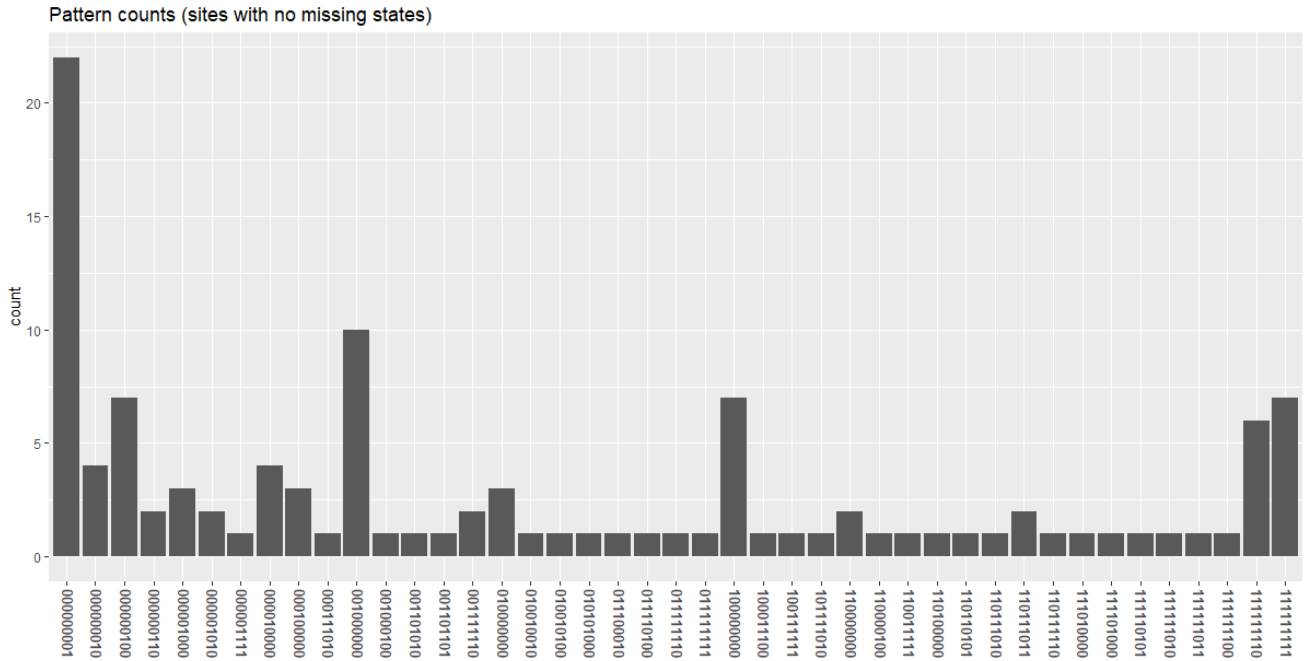
- Treat the cognate classes / sites as “markers”.
- At each site, a language is either at State-0 or State-1, so the “markers” are bi-allelic.



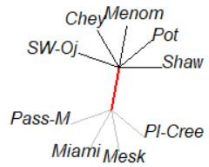
Visualizing the cognate data

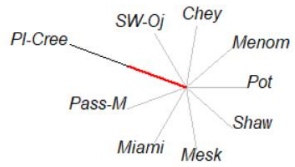


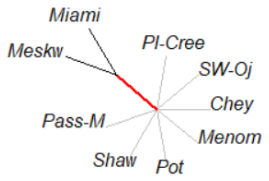


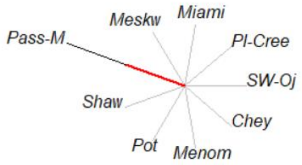


4 bipartitions corresponding to “meaning” 2 (See columns 2-5)









No. sites (w/o missing states): 113
No. of unique observed patterns: 43
No. of unique possible patterns: $2^9 = 512$

Evidence of signal?: $\frac{43}{512} \approx 0.08$
Insufficient data?: $113/43 \approx 2.6$ columns per unique pattern

Most frequent pattern: “000000001” => Pass-M is the outgroup
2nd most frequent pattern: “001000000” => Shaw is the outgroup

Complementary patterns: “000000001”, “111111110” => evidence for the same bipartition, but pushes rates for $0 \rightarrow 1,\ 1 \leftarrow 0$ to be more equal?

Uninformative (for phylogeny?): “111111111” => Not biallelic!

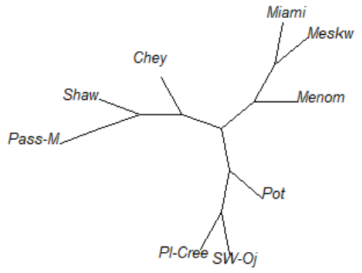
Setup

- Run SNAPP for 4 iterations.
- Each iteration produces an MCMC chain of 1-1.5 million states.
- Log the state of the MCMC chain every 1000 states.
- Each state consists of a value for:
  i. u, rate of $0 \rightarrow 1$
  ii. v, rate of $1 \leftarrow 0$
  iii. θ_i, $i \in \{1, 2, \ldots, 17\}$ (effective population sizes at the species-tree nodes)
  iv. λ, the birth-rate for the Yule-model prior on the species-tree
  v. The species-tree (topology + branch lengths)
- Use the default prior settings for parameters (i) – (iv).

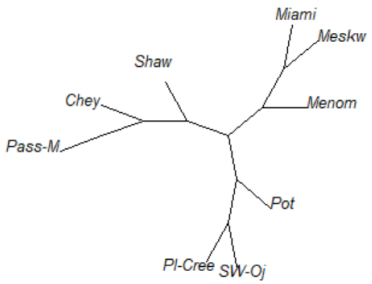
Species-trees ranked by posterior support

- Using TreeSetAnalyzer (packaged with SNAPP)
- Top 6% of topologies (Tree 1-20) capture about 38% of posterior support

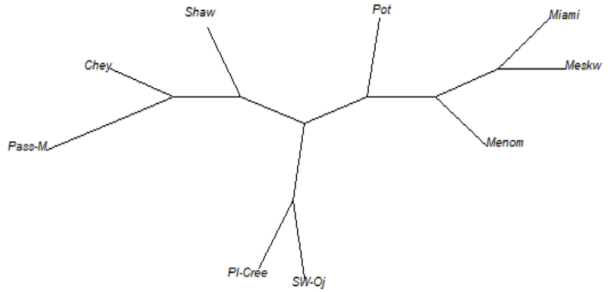
```
95% HPD contains 303 topologies,... out of a total of 348 topologies
C:\Users\totem\Documents\BOT 563\phylo-class-project\results\run_snapp_bg\run6_passed\snap.trees
48#nr coverage tree
Tree 1: 3.662597% (Pass-M_,(((((Pl-Cree_,SW-Oj_),Pot._),(Menom_,(Meskw._,Miami_))),Chey_),Shaw_))
Tree 2: 3.4406214% (Pass-M_,(((((Pl-Cree_,SW-Oj_),Pot._),(Menom_,(Meskw._,Miami_))),Shaw_),Chey_))
Tree 3: 3.329634% (Pass-M_,((((Pl-Cree_,SW-Oj_),((Menom_,(Meskw._,Miami_)),Pot._)),Shaw_),Chey_))
Tree 4: 3.329634% (Pass-M_,((((Pl-Cree_,SW-Oj_),((Menom_,(Meskw._,Miami_)),Pot._)),Chey_),Shaw_))
Tree 5: 2.8856826% (Pass-M_,(((Pl-Cree_,SW-Oj_),Pot._),(Menom_,(Meskw._,Miami_))),(Chey_,Shaw_)))
Tree 6: 2.3307438% (Pass-M_,(((Pl-Cree_,SW-Oj_),Pot._),((Menom_,(Meskw._,Miami_)),Chey_)),Shaw_))
Tree 7: 1.8867924% (Pass-M_,((Pl-Cree_,SW-Oj_),((Menom_,(Meskw._,Miami_)),Pot._)),(Chey_,Shaw_)))
Tree 8: 1.7758048% (Pass-M_,(((((Pl-Cree_,SW-Oj_),Pot._),Menom_),(Meskw._,Miami_)),Chey_),Shaw_))
Tree 9: 1.7758048% (Pass-M_,((Pl-Cree_,SW-Oj_),((Menom_,((Meskw._,Miami_),Pot._)),Chey_)),Shaw_))
Tree 10: 1.664817% (Pass-M_,(((Pl-Cree_,SW-Oj_),(Menom_,((Meskw._,Miami_),Pot._))),Chey_),Shaw_))
Tree 11: 1.4428413% (Pass-M_,((Pl-Cree_,SW-Oj_),(((Menom_,(Meskw._,Miami_)),Pot._),Chey_)),Shaw_))
Tree 12: 1.3318535% (Pass-M_,(((Pl-Cree_,SW-Oj_),((Menom_,Pot._),(Meskw._,Miami_))),Chey_),Shaw_))
Tree 13: 1.3318535% (Pass-M_,((Pl-Cree_,SW-Oj_),(((Menom_,(Meskw._,Miami_)),Pot._),Shaw_)),Chey_))
Tree 14: 1.2208657% (Pass-M_,(((((Pl-Cree_,SW-Oj_),Pot._),Chey_),(Menom_,(Meskw._,Miami_))),Shaw_))
Tree 15: 1.1098778% (Pass-M_,(((Pl-Cree_,SW-Oj_),((Menom_,Pot._),(Meskw._,Miami_))),Shaw_),Chey_))
Tree 16: 0.9988901% (Pass-M_,(((Pl-Cree_,SW-Oj_),Pot._),((Menom_,Chey_),(Meskw._,Miami_))),Shaw_))
Tree 17: 0.9988901% (Pass-M_,((Pl-Cree_,SW-Oj_),((Menom_,Pot._),(Meskw._,Miami_))),(Chey_,Shaw_)))
Tree 18: 0.9988901% (Pass-M_,(((((Pl-Cree_,SW-Oj_),Pot._),Shaw_),(Menom_,(Meskw._,Miami_))),Chey_))
Tree 19: 0.9988901% (Pass-M_,(((((Pl-Cree_,SW-Oj_),(Menom_,(Meskw._,Miami_))),Pot._),Shaw_),Chey_))
Tree 20: 0.9988901% (Pass-M_,(((Pl-Cree_,SW-Oj_),(Menom_,((Meskw._,Miami_),Pot._))),(Chey_,Shaw_)))
```



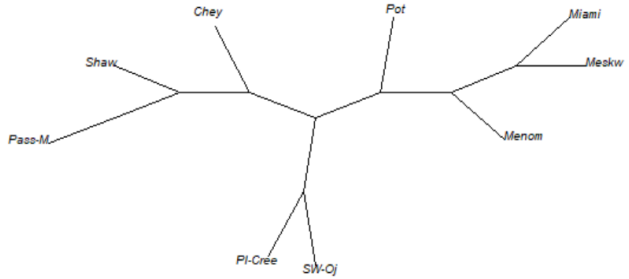
Tree 1



Tree 2



Tree 3

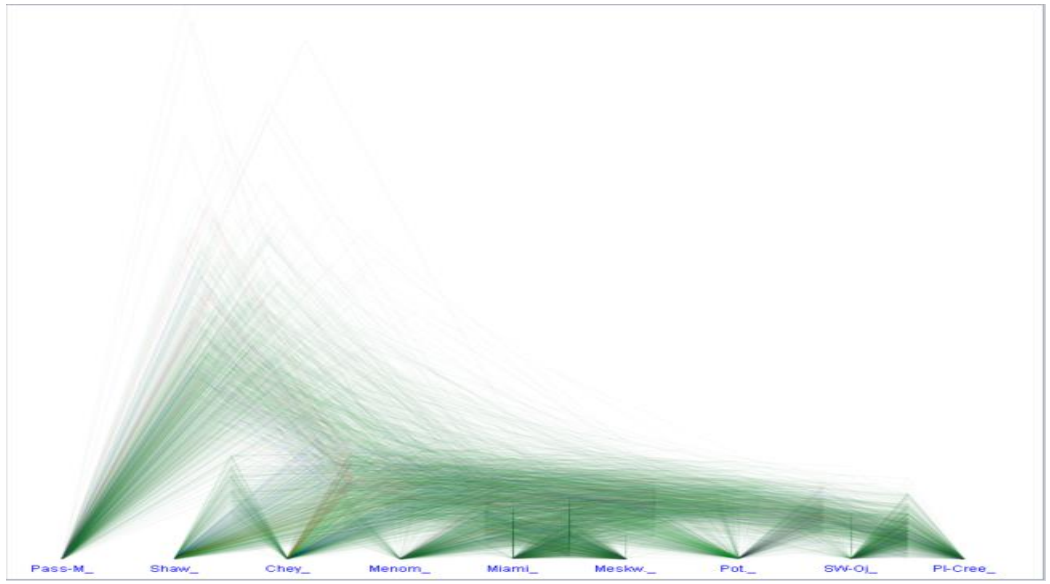


Tree 4

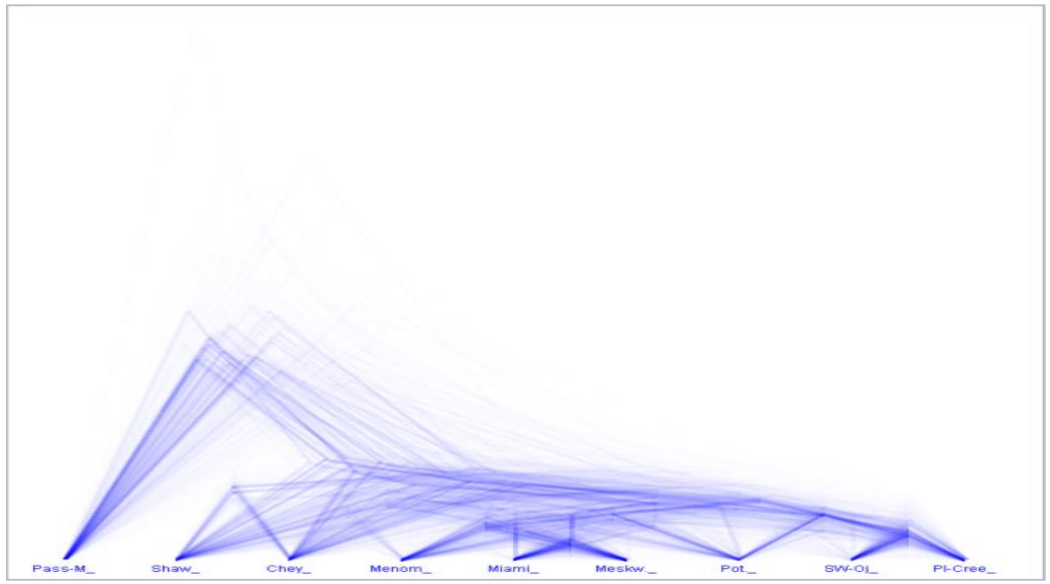
- Uncertainty in the positions of Shaw, Chey, Pot, Menom?
- Posterior distribution for species-tree topology is diffuse?
- *Or rather*, tree-space is large (# of unrooted 9-species trees = 135,135 and 348/135,135 ≈ 0.003)?

DensiTrees

All trees

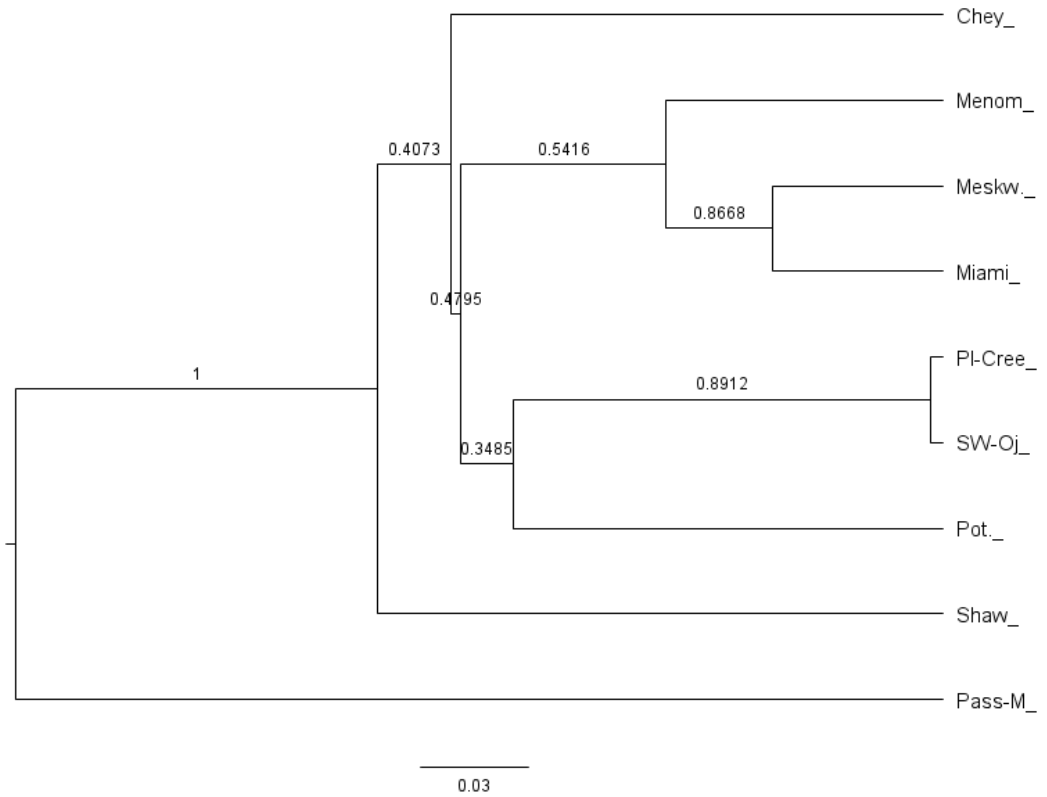


Consensus trees



Maximum clade credibility (MCC) tree

- Using TreeAnnotator (packaged with BEAST2) and FigTree.
- Branches labelled by posterior support.

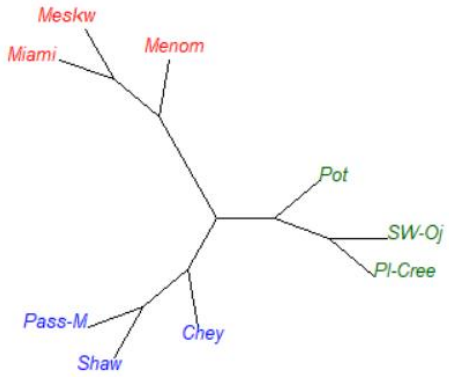


- This is the same as Tree 1 (modal topology sampled) in the previous slide!
- Uncertainty in the positions of Shaw, Chey, Pot, Menom …

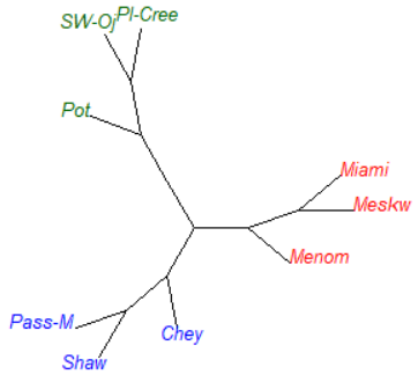
Bonus: Species-trees ranked by posterior support *for fake data*

- Can we cook up data that concentrates the posterior distribution for species-tree topology?
- Part of sanity-checks …

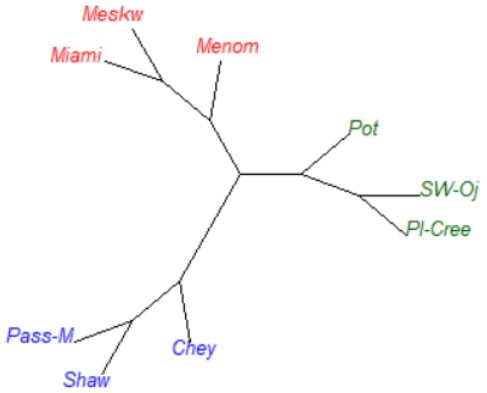
```
95% HPD contains 3 topologies,... out of a total of 3 topologies
C:\Users\totem\Documents\BOT 563\phylo-class-project\results\run_snapp_bg\run7_fake_passed\snap.trees
3#nr coverage tree
Tree 1: 34.5172% ((((Pass-M_,Shaw_),Chey_),((Pl-Cree_,SW-Oj_),Pot._)),(Menom_,(Meskw._,Miami_)))
Tree 2: 34.07325% ((((Pass-M_,Shaw_),Chey_),(Menom_,(Meskw._,Miami_))),((Pl-Cree_,SW-Oj_),Pot._))
Tree 3: 31.409544% (((Pass-M_,Shaw_),Chey_),(((Pl-Cree_,SW-Oj_),Pot._),(Menom_,(Meskw._,Miami_))))
```



Tree 1



Tree 2



Tree 3

- Yes, we can!