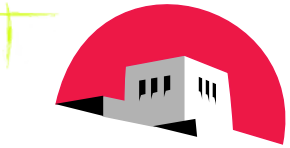


How Many Bootstrap Replicates are Necessary?

Nicholas D. Pattengale(*)[1], Masoud Alipour[2], Olaf R.P. Bininda-Emonds[3], Bernard M.E. Moret[2,4], Alexandros Stamatakis[5]

[1]Department of Computer Science, University of New Mexico, Albuquerque NM, USA

[2]Laboratory for Computational Biology and Bioinformatics, EPFL, Switzerland

[3] AG Systematik und Evolutionsbiologie, Institut für Biologie und Umweltwissenschaften, University of Oldenburg, Germany

[4] Swiss Institute of Bioinformatics, Lausanne, Switzerland

[5]The Exelixis Lab, Department of Computer Science, TU München, Germany

The University of New Mexico

Main Result/Contribution

Two criteria for *stopping numbers* in phylogenetic bootstrapping

- First empirical assessment of variability in support value, as a function of replicate count, in bootstrapping
- Validate our proposals for stopping criteria

The University of New Mexico

Table of Content

- Background
  - Phylogeny and Splits
  - The Phylogenetic Bootstrap
  - Stopping Numbers
- Our Technique
  - The Framework
    - Motivation – Permutation Test
    - Frequency Criterion (FC)
    - Weighted Criterion(WC)
  - The Experiment(s)
- Conclusion

The University of New Mexico

Phylogenetic Reconstruction

Human

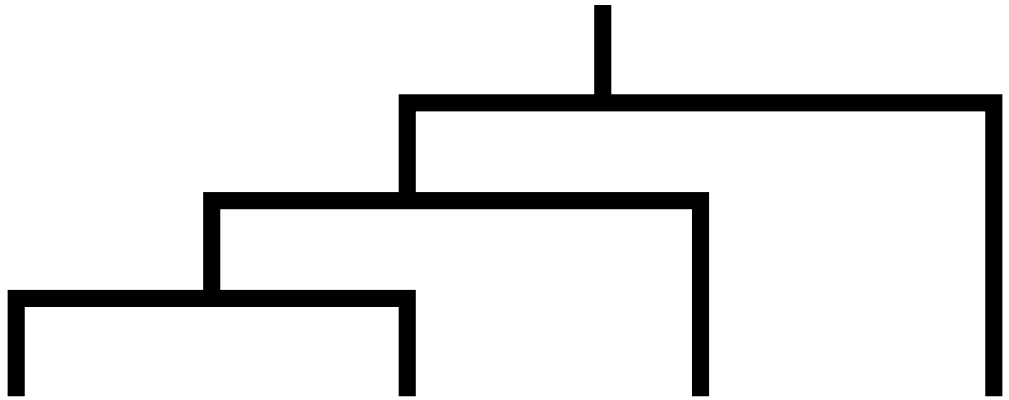
Chimpanzee

Gorilla

Orangutan

The University of New Mexico

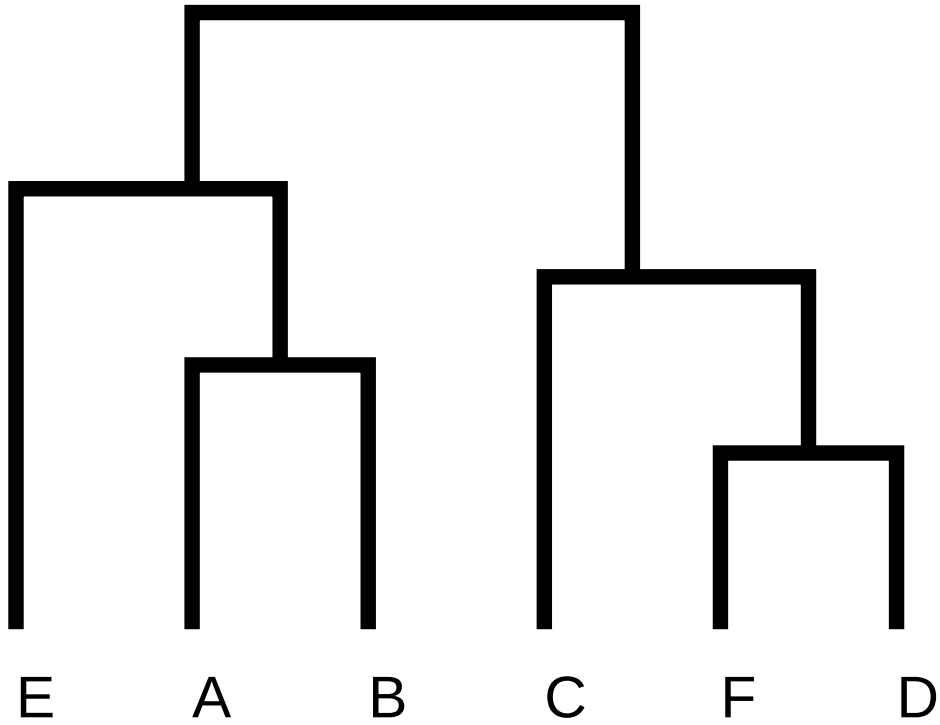
Phylogenetic Reconstruction



Human Chimpanzee Gorilla Orangutan

The University of New Mexico

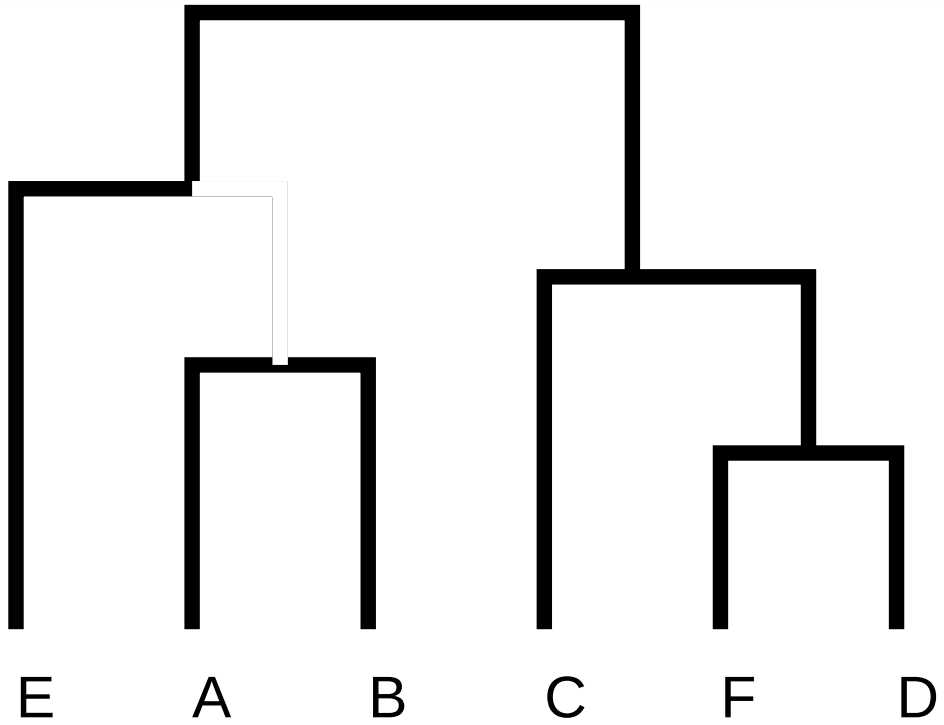
Canonical Representation - Splits



The University of New Mexico

Canonical Representation - Splits

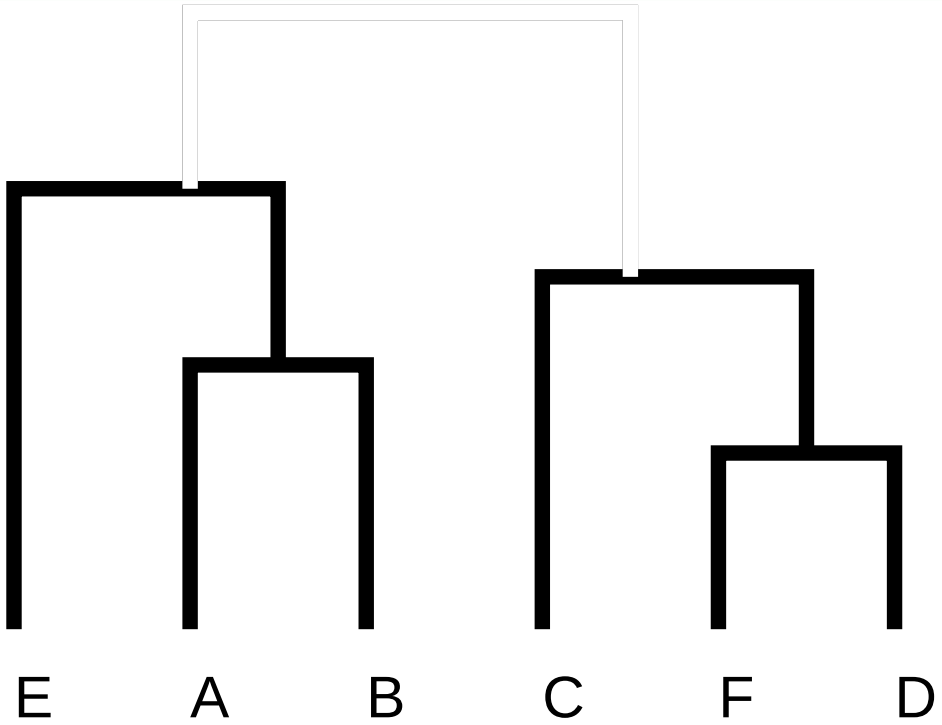
- AB|CDEF



The University of New Mexico

Canonical Representation - Splits

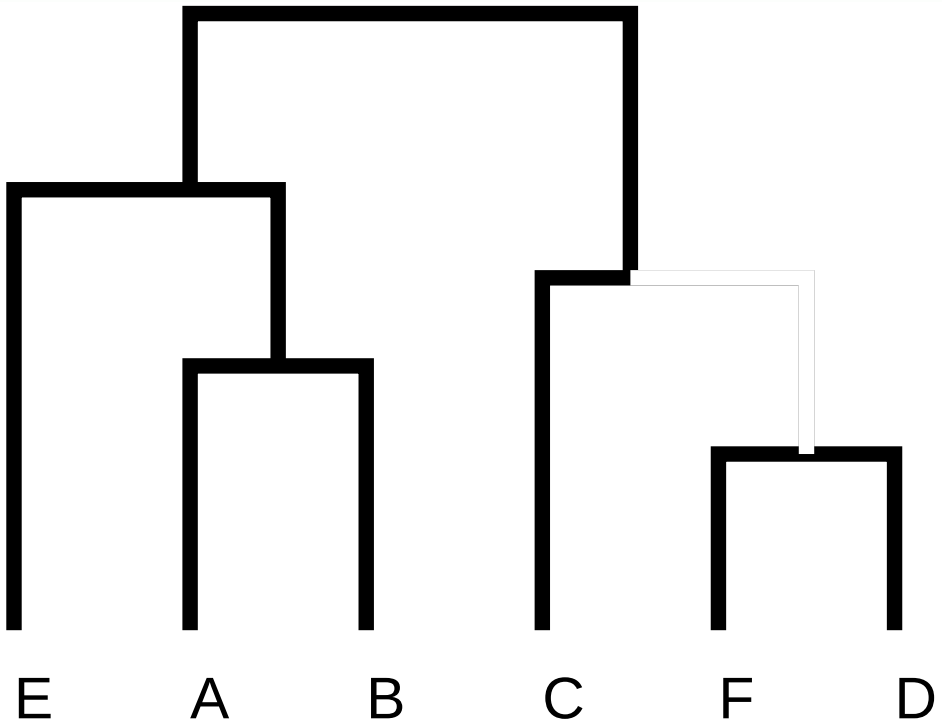
- AB|CDEF
- ABE|CDF



The University of New Mexico

Canonical Representation - Splits

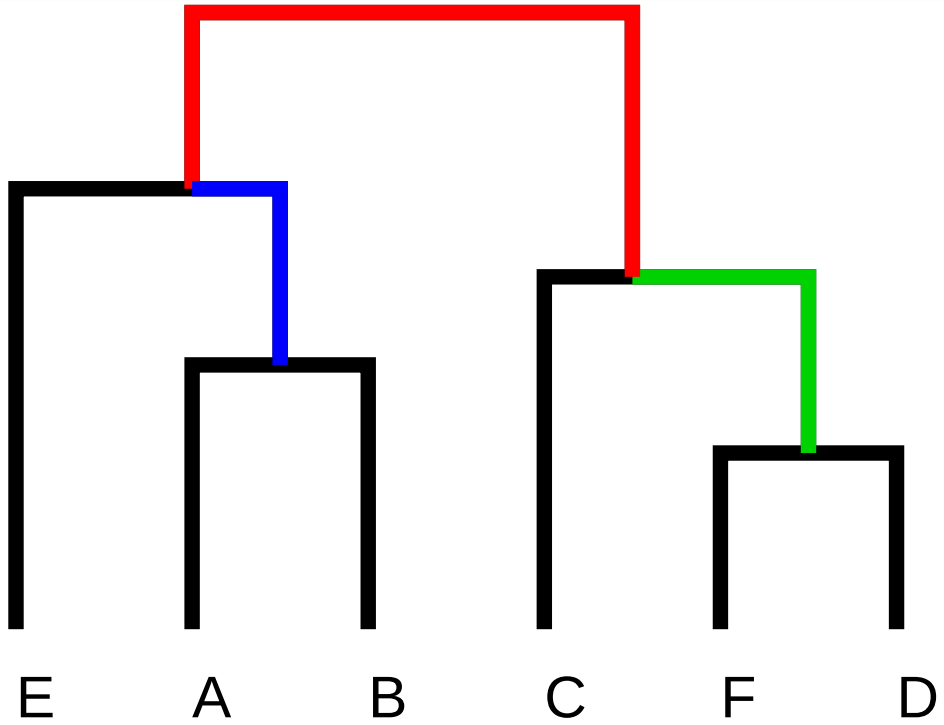
- AB|CDEF
- ABE|CDF
- DF|ABCE



The University of New Mexico

Canonical Representation - Splits

- AB|CDEF
- ABE|CDF
- DF|ABCE



The University of New Mexico

The Phylogenetic Bootstrap

- So you've reconstructed a tree via MP, or ML...

The University of New Mexico

The Phylogenetic Bootstrap

- So you've reconstructed a tree via MP, or ML...
- and you'd like to asses how well your data supports your tree

The University of New Mexico

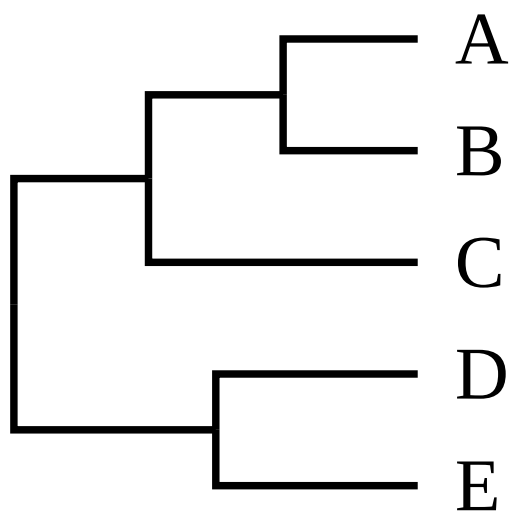
The Phylogenetic Bootstrap

- So you've reconstructed a tree via MP, or ML...
- and you'd like to asses how well your data supports your tree
- One answer: **the phylogenetic bootstrap**

The University of New Mexico

The Phylogenetic Bootstrap

Original Data	0	1	2	3
Species A	C	C	T	C
Species B	A	C	T	G
Species C	C	-	-	G
Species D	A	C	C	C
Species E	A	G	-	C

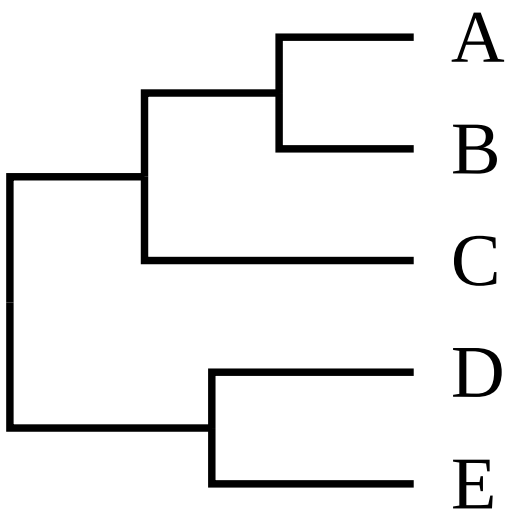


The University of New Mexico

The Phylogenetic Bootstrap

Original Data	0	1	2	3
Species A	C	C	T	C
Species B	A	C	T	G
Species C	C	-	-	G
Species D	A	C	C	C
Species E	A	G	-	C

Bootstrap 1	1	3	1	3
Species A	C	C	C	C
Species B	C	G	C	G
Species C	-	G	-	G
Species D	C	C	C	C
Species E	G	C	G	C

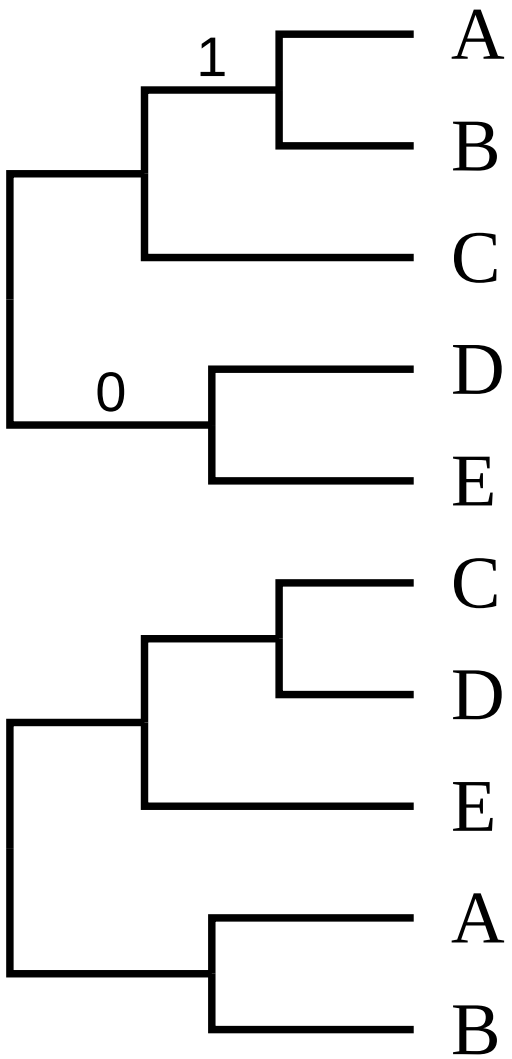


The University of New Mexico

The Phylogenetic Bootstrap

Original Data	0	1	2	3
Species A	C	C	T	C
Species B	A	C	T	G
Species C	C	-	-	G
Species D	A	C	C	C
Species E	A	G	-	C

Bootstrap 1	1	3	1	3
Species A	C	C	C	C
Species B	C	G	C	G
Species C	-	G	-	G
Species D	C	C	C	C
Species E	G	C	G	C

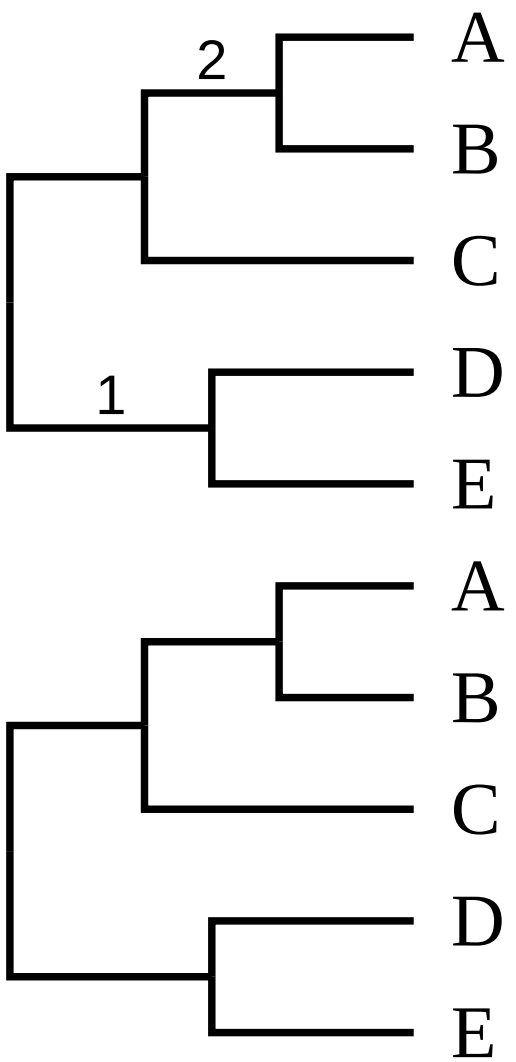


The University of New Mexico

The Phylogenetic Bootstrap

Original Data	0	1	2	3
Species A	C	C	T	C
Species B	A	C	T	G
Species C	C	-	-	G
Species D	A	C	C	C
Species E	A	G	-	C

Bootstrap 2	2	1	0	0
Species A	T	C	C	C
Species B	T	C	A	A
Species C	-	-	C	C
Species D	C	C	A	A
Species E	-	G	A	A

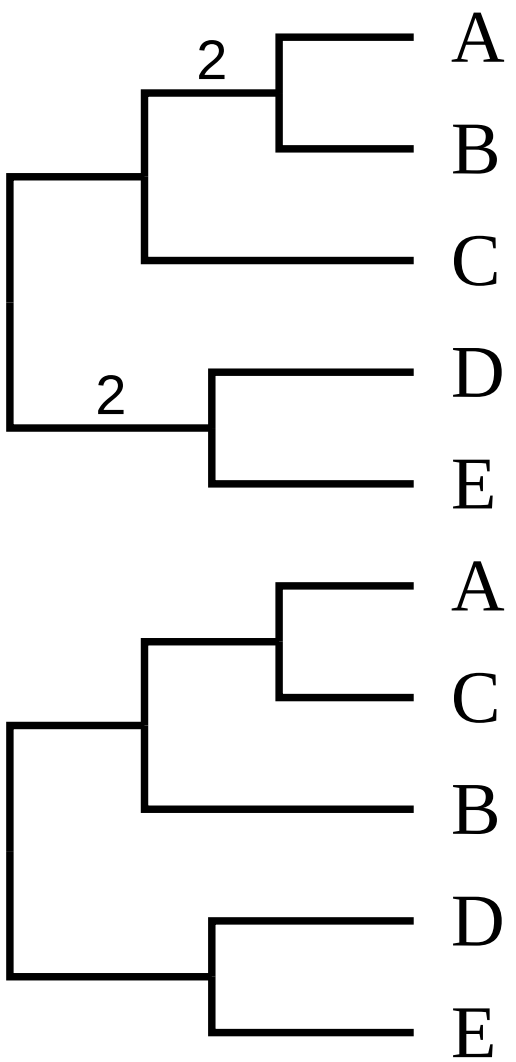


The University of New Mexico

The Phylogenetic Bootstrap

Original Data	0	1	2	3
Species A	C	C	T	C
Species B	A	C	T	G
Species C	C	-	-	G
Species D	A	C	C	C
Species E	A	G	-	C

Bootstrap 3	0	0	3	0
Species A	C	C	C	C
Species B	A	A	G	A
Species C	C	C	G	C
Species D	A	A	C	A
Species E	A	A	C	A

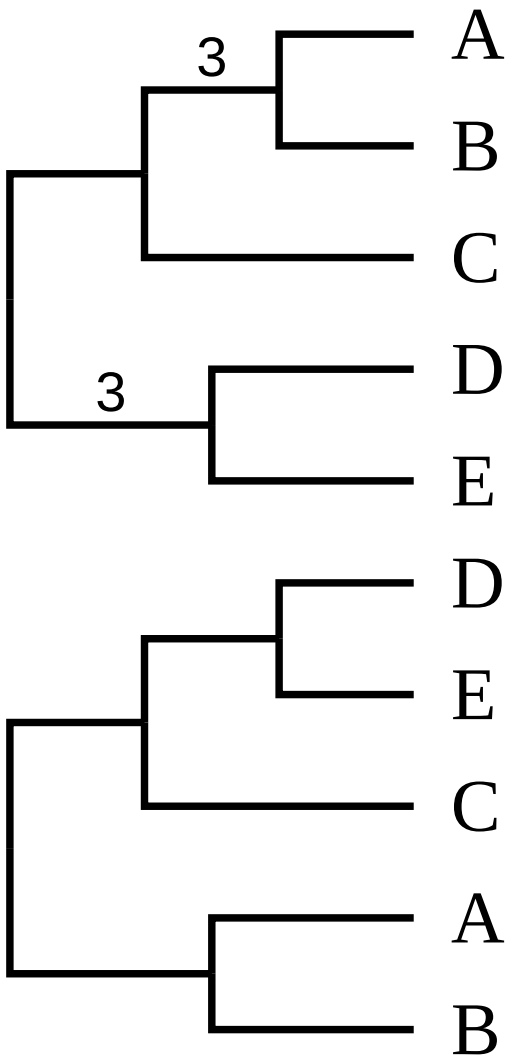


The University of New Mexico

The Phylogenetic Bootstrap

Original Data	0	1	2	3
Species A	C	C	T	C
Species B	A	C	T	G
Species C	C	-	-	G
Species D	A	C	C	C
Species E	A	G	-	C

Bootstrap 4	2	1	0	2
Species A	T	C	C	T
Species B	T	C	A	T
Species C	-	-	C	-
Species D	C	C	A	C
Species E	-	G	A	-

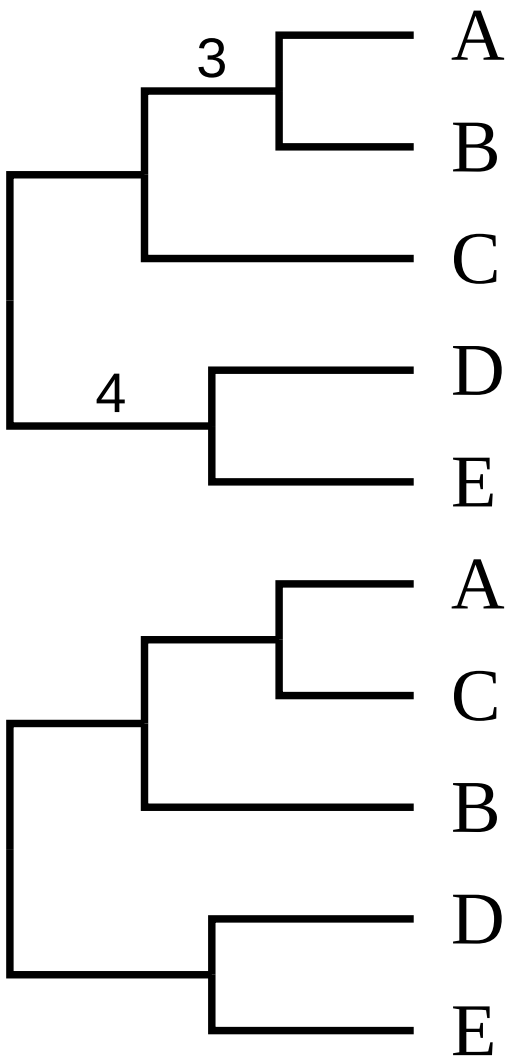


The University of New Mexico

The Phylogenetic Bootstrap

Original Data	0	1	2	3
Species A	C	C	T	C
Species B	A	C	T	G
Species C	C	-	-	G
Species D	A	C	C	C
Species E	A	G	-	C

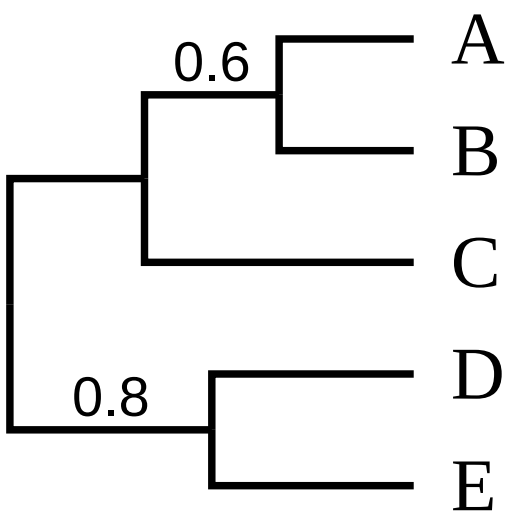
Bootstrap 5	2	2	1	3
Species A	T	T	C	C
Species B	T	T	C	G
Species C	-	-	-	G
Species D	C	C	C	C
Species E	-	-	G	C



The University of New Mexico

The Phylogenetic Bootstrap

Original Data	0	1	2	3
Species A	C	C	T	C
Species B	A	C	T	G
Species C	C	-	-	G
Species D	A	C	C	C
Species E	A	G	-	C



The University of New Mexico

The Phylogenetic Bootstrap

- Motivated by *resampling* technique from statistics
  - Used to assess the stability of simple summary statistics
- Computationally expensive – days to months

The University of New Mexico

Stopping Numbers

- Question we address: How many replicates?
- Theory exists for simpler estimators
  - In phylogeny, estimator is not only complex, but number of bipartitions grow
- State of the art in phylogeny: choose arbitrarily
- Hedges chooses *a priori* for a given level of significance
  - but ignores factors which greatly influence the estimator (the tree search algorithm) and hence the stability of BS replicates

The University of New Mexico

Our Framework

- Major goal: not be biased by current best tree
- Devise an *adaptive* criterion – to be used at run time
- Based on a *Permutation Test*
  - Typically used to reject that two samples arise from same distribution
  - We use to assess when a population subset sufficiently resembles full population

The University of New Mexico

Our Framework

- Bootstop()
  - With m replicates
    - Repeat $p = 100$ times
      - randomly split into two sets (of size $\frac{m}{2}$)
      - score similarity between two sets
    - Assess –
      - If $\frac{99}{100}$ scores beat threshold – DONE
      - Else – increment m (by, e.g. 50)

The University of New Mexico

Our Framework

- Bootstop()
  - With m replicates
    - Repeat $p = 100$ times
      - randomly split into two sets (of size $\frac{m}{2}$)
      - score similarity between two sets
    - Assess –
      - If $\frac{99}{100}$ scores beat threshold – DONE
      - Else – increment m (by, e.g. 50)
- Well less than 2^n possible

The University of New Mexico

Our Framework

- Bootstop()
  - With m replicates
    - Repeat $p = 100$ times
      - randomly split into two sets (of size $\frac{m}{2}$)
      - score similarity between two sets
    - Assess –
      - If $\frac{99}{100}$ scores beat threshold – DONE
      - Else – increment m (by, e.g. 50)
- Well less than 2^n possible
- Our two approaches differ in their defn. of similarity

The University of New Mexico

Scoring (Dis)similarity

- Frequency Criterion (FC)
  - Build vectors of edge support for the two subsets
  - Take Pearson's Correlation Coefficient between the two vectors

The University of New Mexico

Scoring (Dis)similarity

- Frequency Criterion (FC)
  - Build vectors of edge support for the two subsets
  - Take Pearson's Correlation Coefficient between the two vectors
- Weighted Criterion(WC)
  - Build (Majority Rules) Consensus trees for the two subsets
  - Take Weighted RF distance between the two trees

The University of New Mexico

Scoring (Dis)similarity

- Frequency Criterion (FC)
  - Build vectors of edge support for the two subsets
  - Take Pearson's Correlation Coefficient between the two vectors
- Weighted Criterion(WC)
  - Build (Majority Rules) Consensus trees for the two subsets
  - Take Weighted RF distance between the two trees
- What is the difference?
  - WC takes into account phylogenetically meaningful
  - WC is more conservative, but also sensitive

The University of New Mexico

Experimental Design

- For 17 diverse, real-world datasets with
  - 125 to 2,554 taxa
  - hudreds to tens of thousands of columns
- we did the following:

The University of New Mexico

Experimental Design

- For 17 diverse, real-world datasets with
  - 125 to 2,554 taxa
  - hudreds to tens of thousands of columns
- we did the following:
  - Generated $\geq 10,000$ BS replicates (serves $m \rightarrow \infty$)

The University of New Mexico

Experimental Design

- For 17 diverse, real-world datasets with
  - 125 to 2,554 taxa
  - hudreds to tens of thousands of columns
- we did the following:
  - Generated $\geq 10,000$ BS replicates (serves $m \rightarrow \infty$)
  - Applied our criteria to generate stopping numbers

The University of New Mexico

Experimental Design

- For 17 diverse, real-world datasets with
  - 125 to 2,554 taxa
  - hudreds to tens of thousands of columns
- we did the following:
  - Generated $\geq 10,000$ BS replicates (serves $m \rightarrow \infty$)
  - Applied our criteria to generate stopping numbers
  - Assessed quality of our stopping numbers w.r.t. $\geq 10,000$ tree set

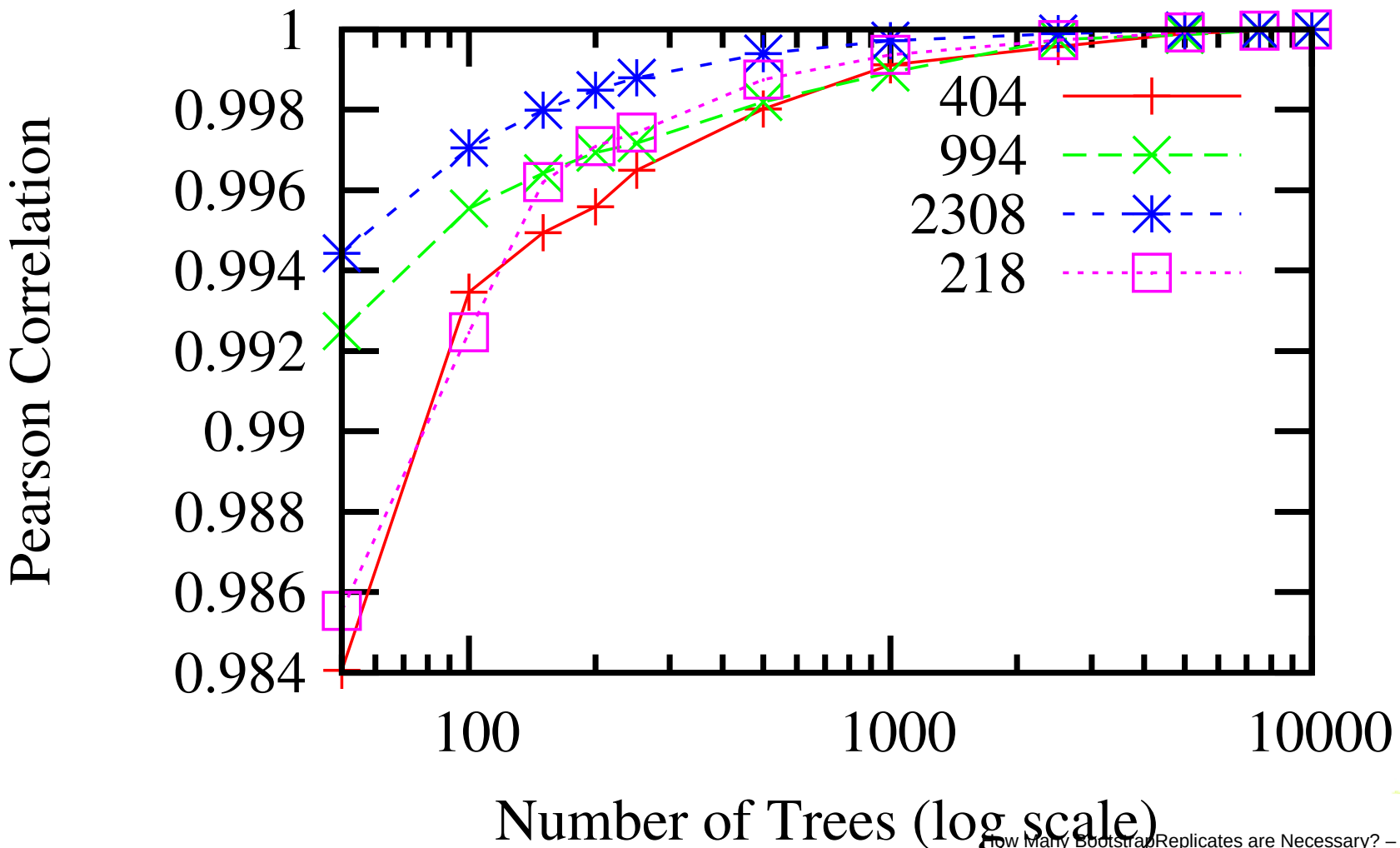
The University of New Mexico

Results

- Stopping numbers
  - FC: 150, 150, 150, 200, 200, 200, 200, 200, 250, 250, 250, 250, 250, 300, 300, 300, 450
  - WC: 50, 200, 300, 350, 400, 400, 400, 400, 450, 450, 500, 550, 600, 600, 650, 700, 1200
- Widely varying, dataset dependent (especially with WC)
- Correlation of support values always exceeds 99.5%
- WRF is smaller than the specified WC threshold value in all cases

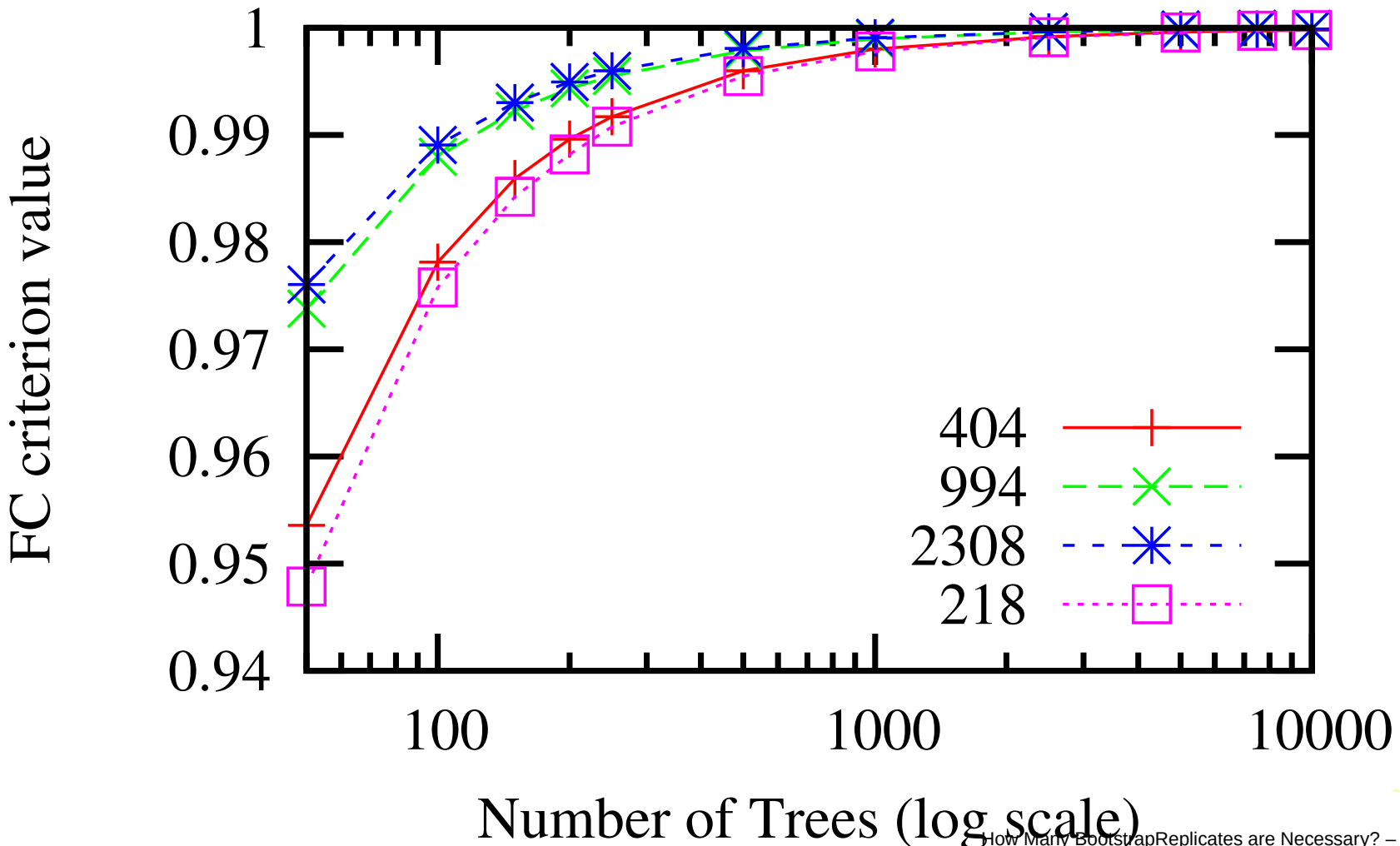
The University of New Mexico

Results



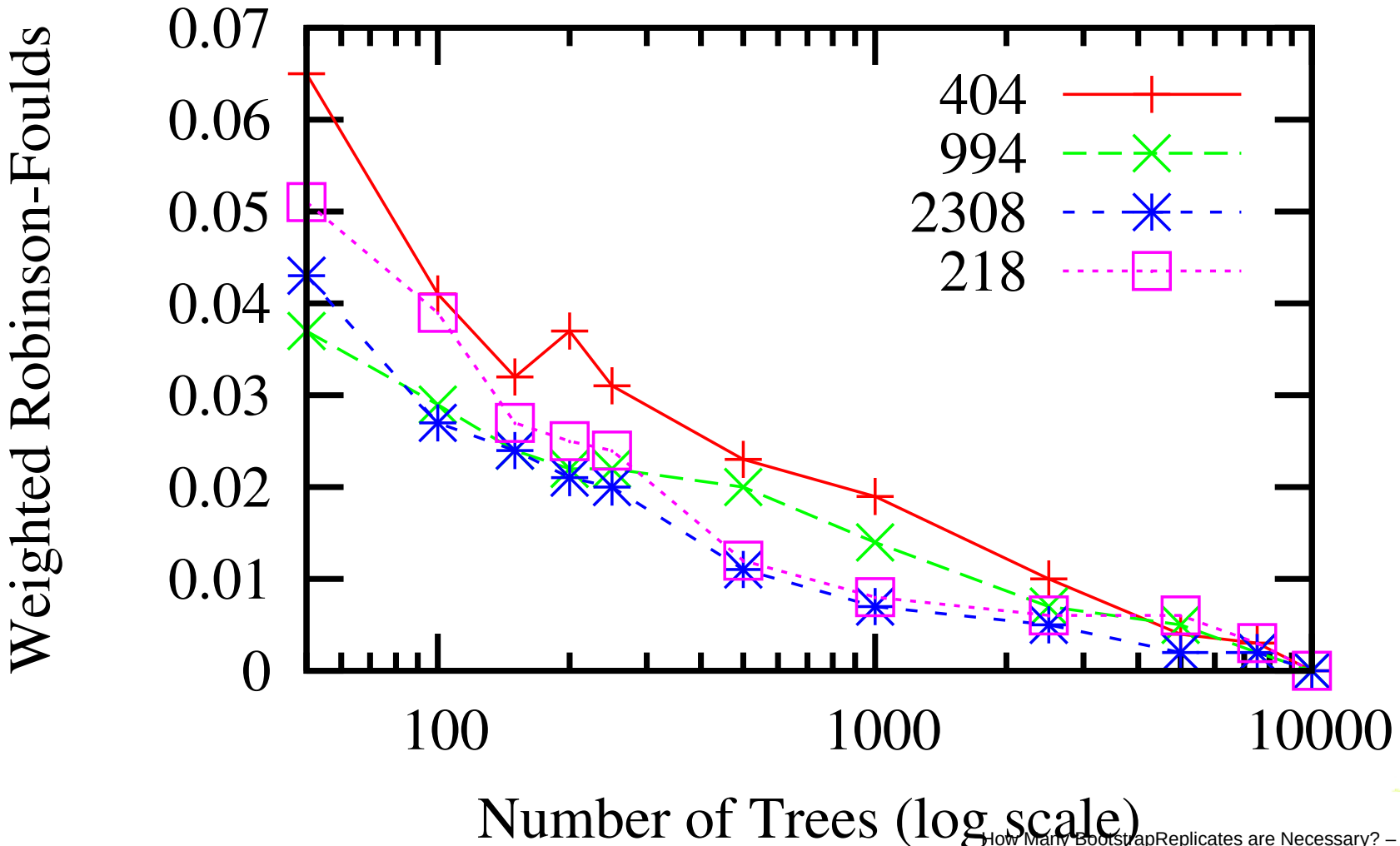
The University of New Mexico

Results



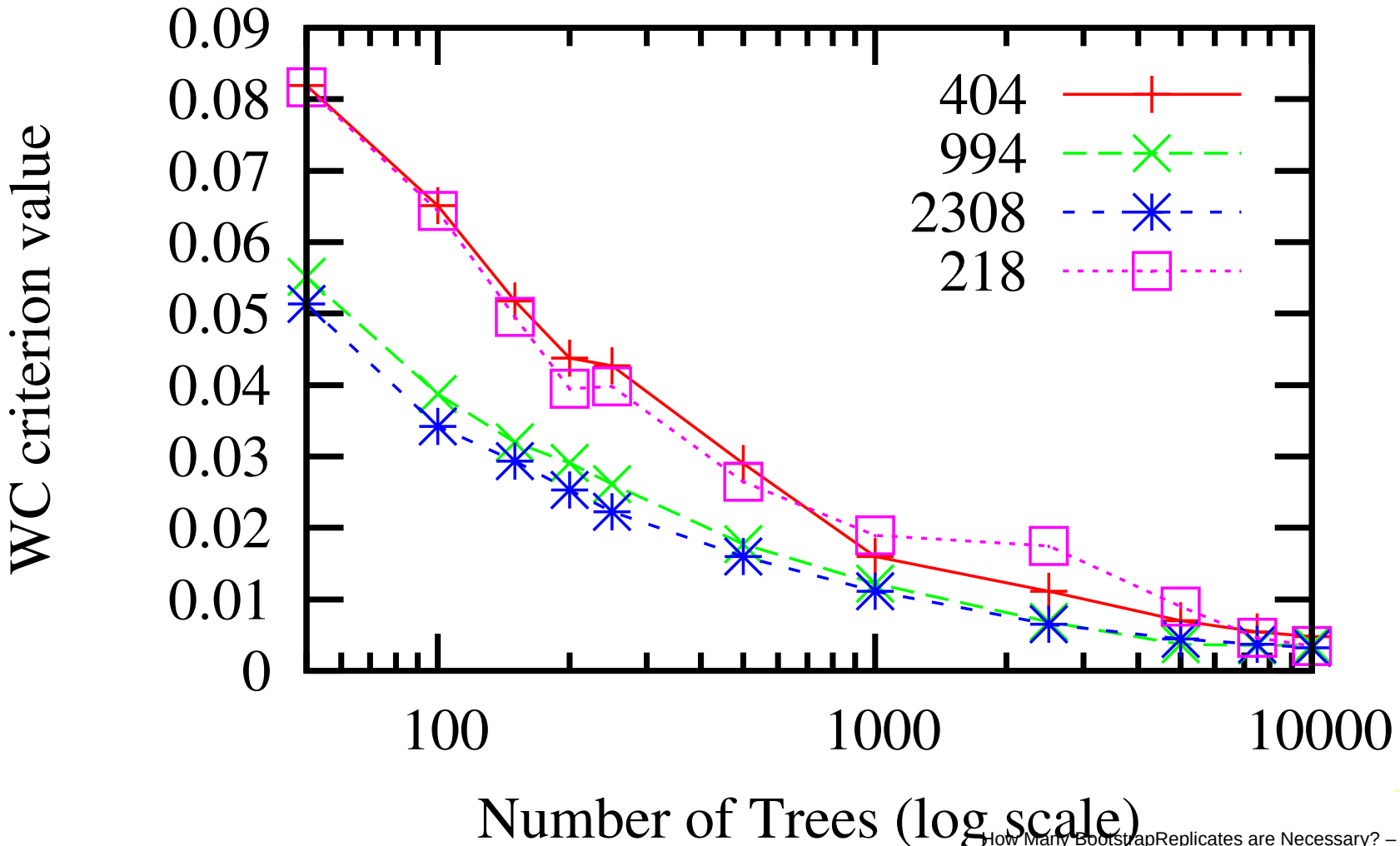
The University of New Mexico

Results



The University of New Mexico

Results



The University of New Mexico

Conclusion

- First large-scale empirical study of bootstrapping convergence
- Used biological datasets that cover a wide range of input alignment sizes and a broad variety of organisms and genes
- Developed and assessed two bootstopping criteria

The University of New Mexico

Conclusion

- Two criteria
  - Can be computed at run time
  - Do not rely on externally provided reference trees
  - Designed to capture stopping point providing sufficient accuracy for unambigoous biological interpretation of the resulting consensus trees or best-known ML trees with support values

The University of New Mexico

Conclusion

- WC criterion yields better performance and higher accuracy than FC
  - Correlates very well with the mean error of support values on the best-scoring tree.
- Advocate the use of WC over FC
  - Takes into account the BS support of “important” bipartitions which are subject to biological interpretation

The University of New Mexico

Conclusion

- Highly dataset dependent
- Only compute as many trees as needed
- Better methods (and ideally, some supporting theory) may exist

The University of New Mexico

That's All Folks

- Thanks
  - to my collaborators
  - to the organizers
  - for listening!
- Stopping Criteria are part of RAxML 7.1.0 alpha
  - http://wwwkramer.in.tum.de/exelixis/software.html
- Data for this study is also available
  - http://lcbb.epfl.ch/BS.tar.bz2