Influence Function for Robust Phylogenetic Reconstruction

Mahendra Mariadassou

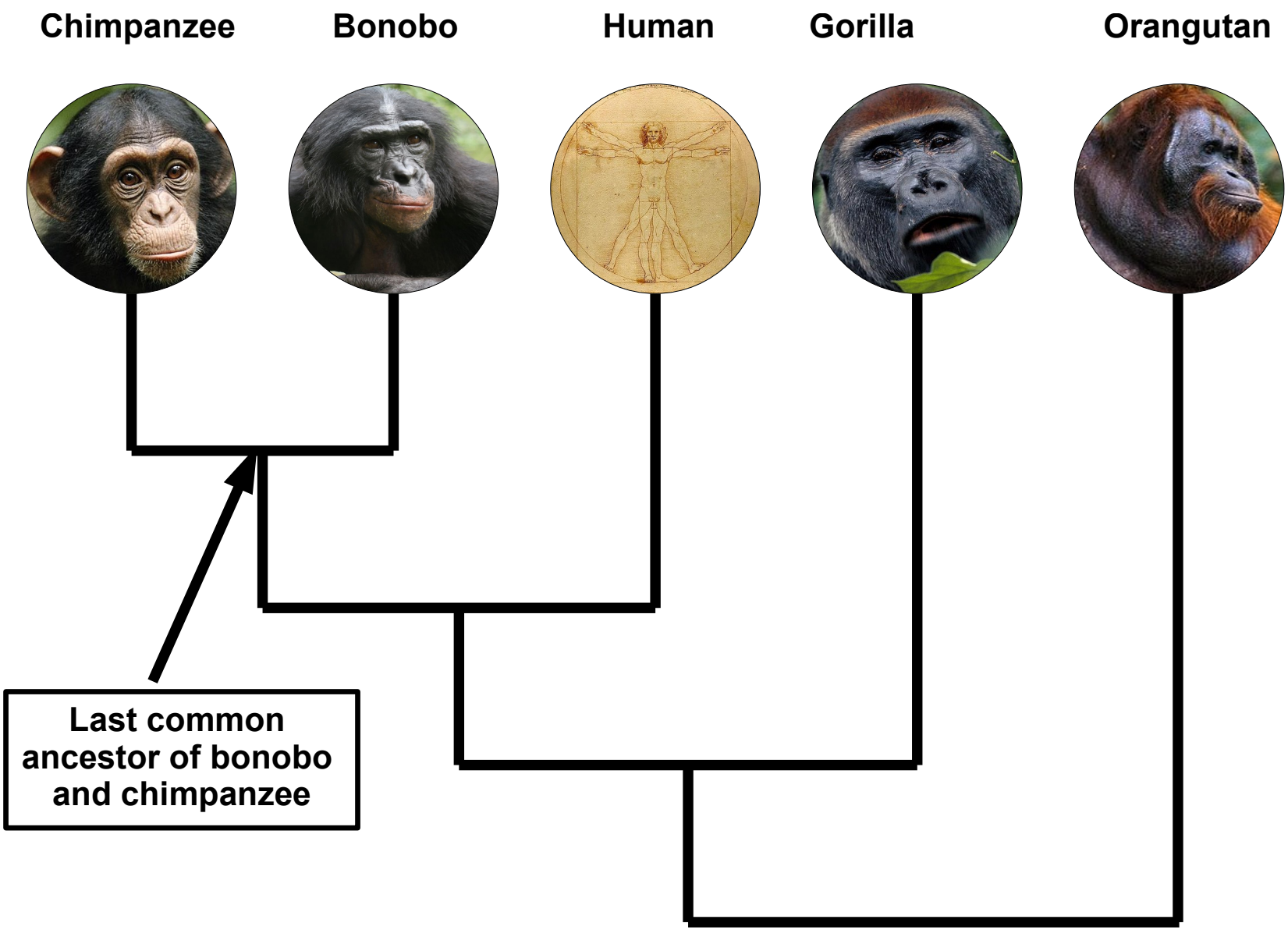
Unité MIG
INRA Jouy-en-Josas

JOBIM 2011
Institut Pasteur

Outline

1. Motivation
2. Outlier Sites
3. Taxon Influence Index

Goal of Phylogenetic Reconstruction



Applications in Many Domains

Including but not limited to:

Studies of an Evolutionary Group:

- Estimate Time to Most Recent Common Ancestor (TMRCA);
- Find genes under positive/purifying selection;
- Identifying Horizontal Gene Transfer;
- Testing evolutionary hypothesis.

Systematics:

- Reconstruct Tree of X, phylogeny of all living X;
- DNA barcoding: easily identify the species of a new organism;
- Natural way to measure biodiversity.

Most of these applications require “good” trees.

Applications in Many Domains

Including but not limited to:

Studies of an Evolutionary Group:

- Estimate Time to Most Recent Common Ancestor (TMRCA);
- Find genes under positive/purifying selection;
- Identifying Horizontal Gene Transfer;
- Testing evolutionary hypothesis.

Systematics:

- Reconstruct Tree of X, phylogeny of all living X;
- DNA barcoding: easily identify the species of a new organism;
- Natural way to measure biodiversity.

Most of these applications require “good” trees.

Applications in Many Domains

Including but not limited to:

Studies of an Evolutionary Group:

- Estimate Time to Most Recent Common Ancestor (TMRCA);
- Find genes under positive/purifying selection;
- Identifying Horizontal Gene Transfer;
- Testing evolutionary hypothesis.

Systematics:

- Reconstruct Tree of X, phylogeny of all living X;
- DNA barcoding: easily identify the species of a new organism;
- Natural way to measure biodiversity.

Most of these applications require “good” trees.

Reconstruction Methods

Three main families:

Distance based: Neighbor-Joining (NJ),

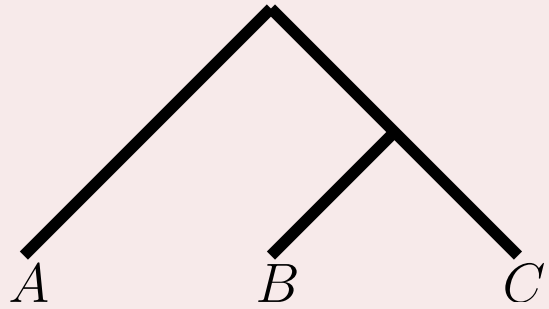
Parsimony based: Maximum Parcimony(MP),

Likelihood based: Maximum Likelihood(ML), Bayesian Inference (BI).

Alignment $\longrightarrow$ **Phylogenetic Tree**

A	C	C	T	T
B	G	G	A	A
C	G	G	A	C

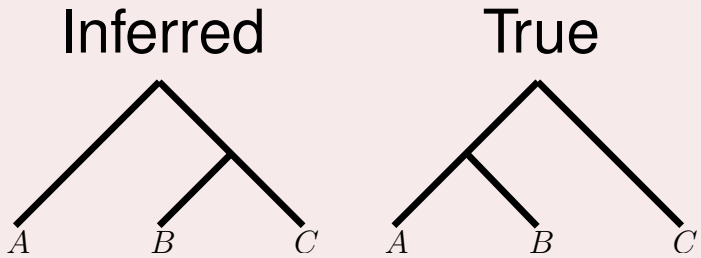
NJ, MP,
ML, BI,
...



Validating the Tree

Inference Problems:

- Compare inferred tree to true tree to assess how good it is,



- But the true tree is not available!

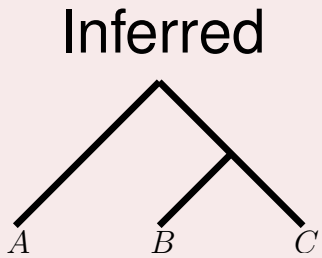
Confidence Issue:

- How confident are we on the inferred tree ?
- Which parts of the tree are reliable/not reliable ?
- How robust is the tree to small changes in the data and outliers ?

Validating the Tree

Inference Problems:

- Compare inferred tree to true tree to assess how good it is,



- But the true tree is not available!

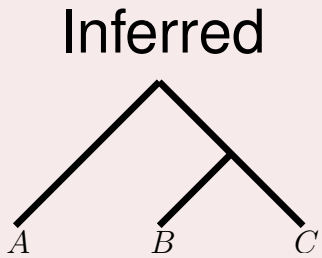
Confidence Issue:

- How confident are we on the inferred tree ?
- Which parts of the tree are reliable/not reliable ?
- How robust is the tree to small changes in the data and outliers ?

Validating the Tree

Inference Problems:

- Compare inferred tree to true tree to assess how good it is,



- But the true tree is not available!

Confidence Issue:

- How confident are we on the inferred tree ?
- Which parts of the tree are reliable/not reliable ?
- How robust is the tree to small changes in the data and outliers ?

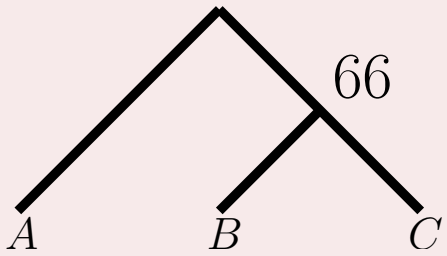
Bootstrap Values: the Theory

Original Dataset:

Alignment

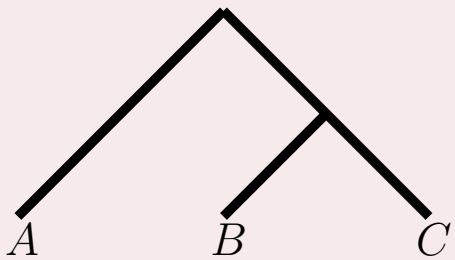
A	A	C	T	T
B	G	G	A	T
C	G	G	C	C

Phylogenetic tree

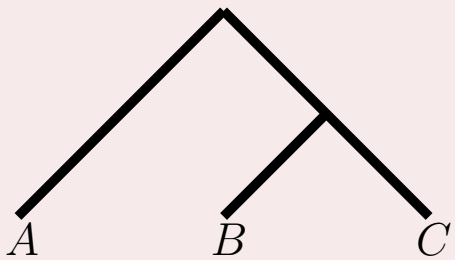


Bootstrap Datasets:

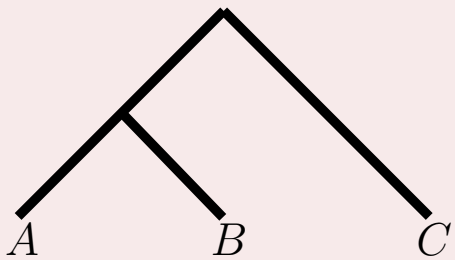
A	A	C	T	C
B	G	G	A	G
C	G	G	C	G



A	C	A	T	A
B	G	G	A	G
C	G	G	C	G



A	T	T	T	T
B	A	T	A	T
C	C	C	C	C



Bootstrap Values: A Robustness Index ?

Potential Causes for Uncertainty:

- Finite sequence lengths (sampling errors);
- Poor alignment quality (influent sites);
- Poor taxon sampling (rogue taxa);
- Model misspecification,
- . . .

Different Tools to Assess Them:

- Bootstrap: deals with global sources of uncertainty;
- Unable to pinpoint local sources of uncertainty;
- Need for other indexes to detect outliers

Bootstrap Values: A Robustness Index ?

Potential Causes for Uncertainty:

- Finite sequence lengths (sampling errors);
- Poor alignment quality (influent sites);
- Poor taxon sampling (rogue taxa);
- Model misspecification,
- . . .

Different Tools to Assess Them:

- Bootstrap: deals with global sources of uncertainty;
- Unable to pinpoint local sources of uncertainty;
- Need for other indexes to detect outliers

Outlier Sites: Motivation and Goal

Motivation: Filter Data

Identifying/filter out outliers corresponding to:

- Sequencing errors;
- Alignment errors;
- Presence of an atypical DNA segment;
- ...

Procedure

- Quantify influence of site i by
  - Removing site i from alignment;
  - Computing new tree T^{-i} from smaller (jackknife) alignment;
  - Comparing new tree to original tree;
- Compute phylogeny from "not too influent sites".

Outlier Sites: Motivation and Goal

Motivation: Filter Data

Identifying/filter out outliers corresponding to:

- Sequencing errors;
- Alignment errors;
- Presence of an atypical DNA segment;
- ...

Procedure

- Quantify influence of site i by
  - Removing site i from alignment;
  - Computing new tree T^{-i} from smaller (jackknife) alignment;
  - Comparing new tree to original tree;
- Compute phylogeny from "not too influent sites".

For Phylogenies...

Definition

- For alignment of size n, the influence value of site i is:

$$IF(\text{site } i) = (n-1)\left(\frac{LogLik(T^{-i})}{n-1} - \frac{LogLik(T)}{n}\right)$$

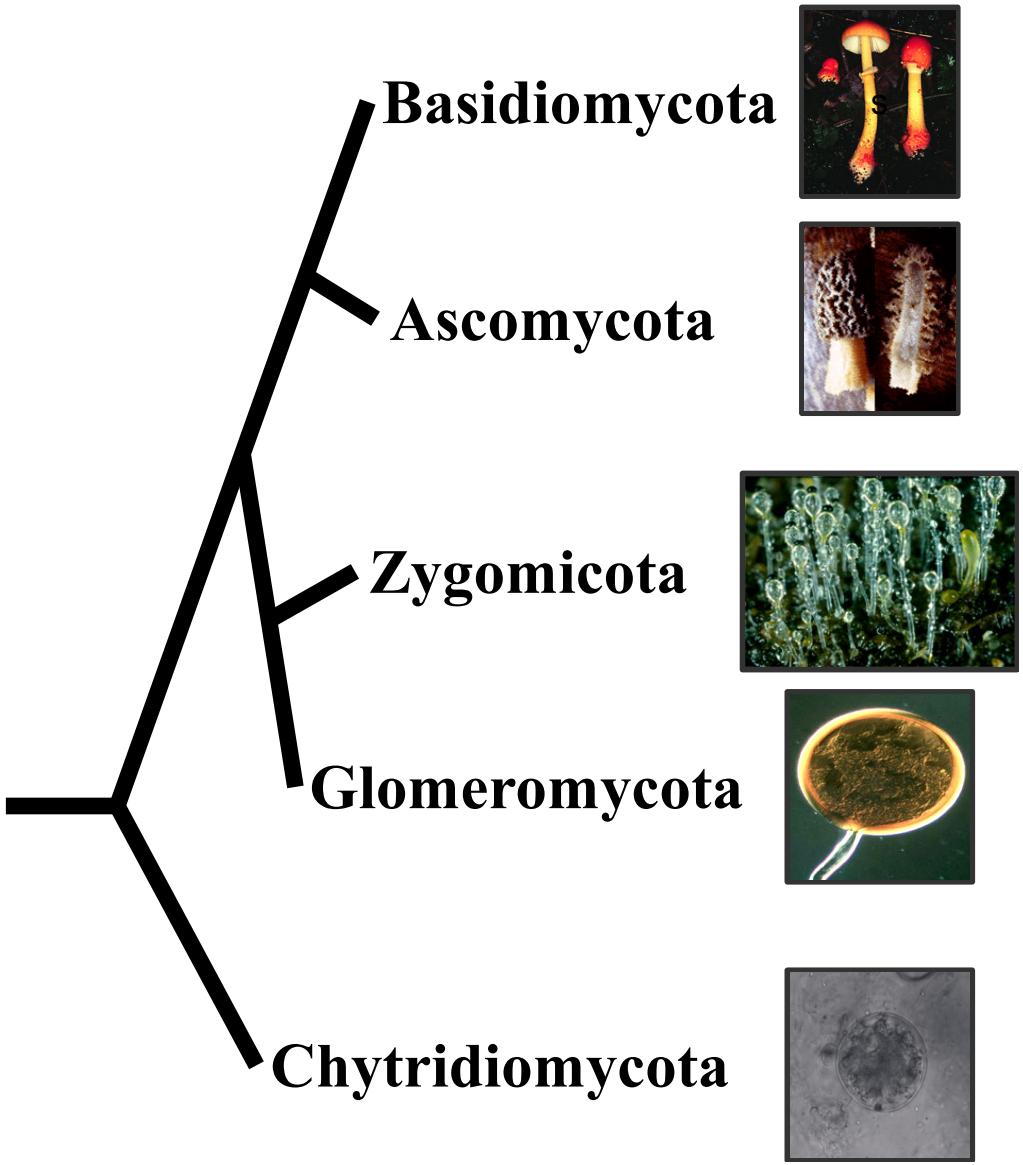
- Difference between support of alignments for their ML trees;
- Negative / Positive value: enhanced / weakened support (when adding the site);
- Expect most sites with small negative values and a few with large positive values.

Strategy towards greater stability

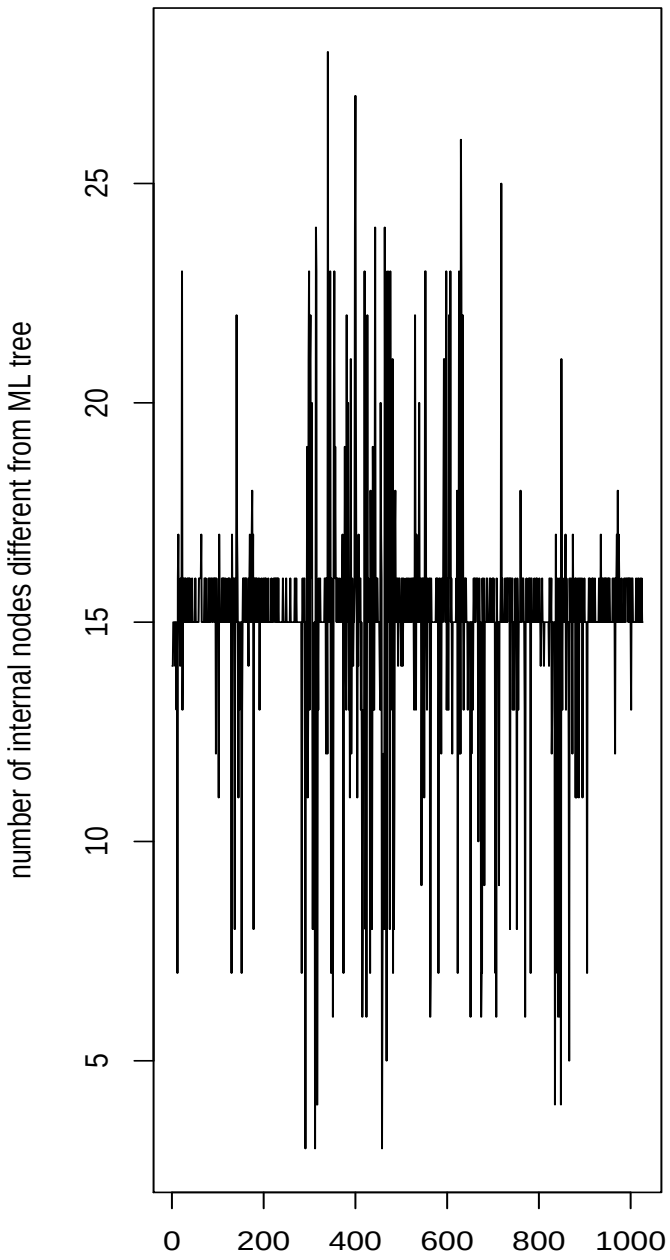
- Focus on outliers: sites with $IF(\text{site } i) > 0$;
- Rank them in decreasing $IF(\text{site } i)$;
- Remove them one at the time until a stable tree is found.

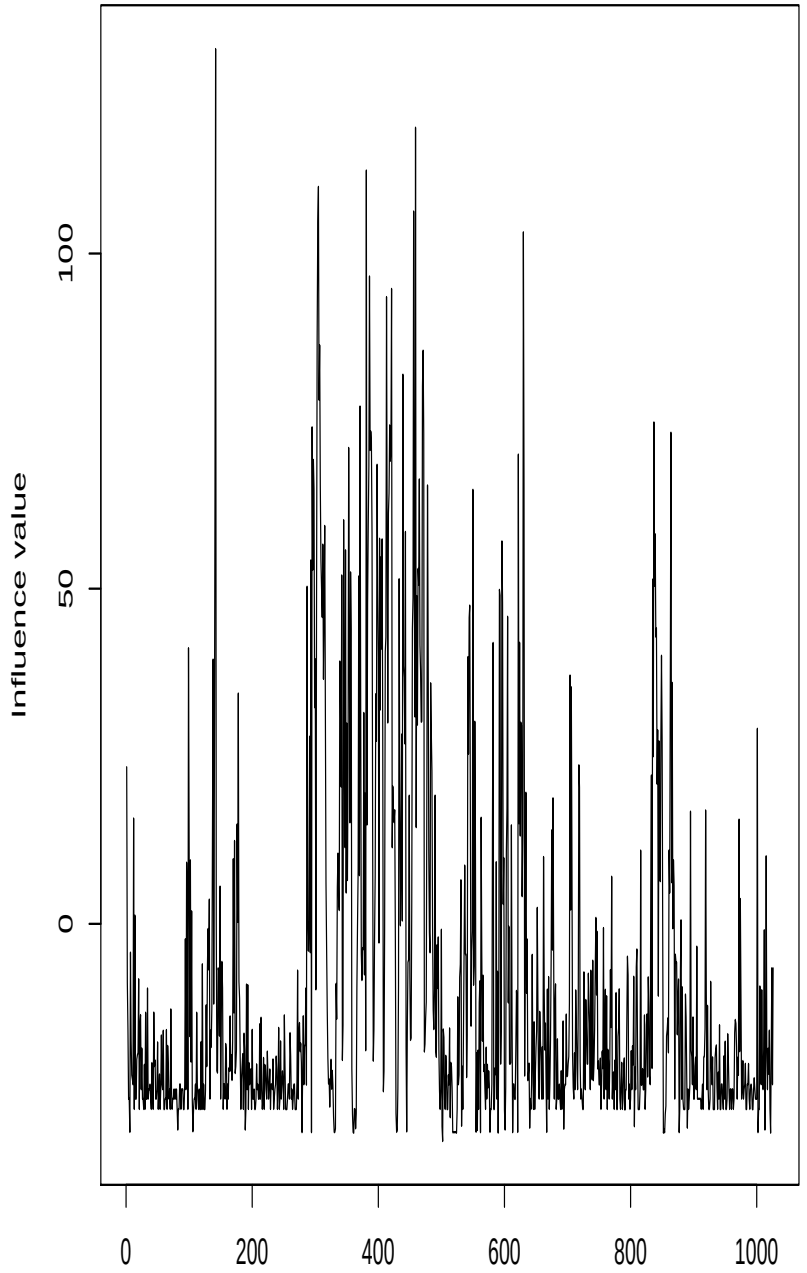
Data: Zygomycetes & Chytridiomycetes

- Lower mushrooms;
- Biology: unknown!
- Enough signal to resolve the topology;
- 1026 sites, 158 taxa, GTR model.

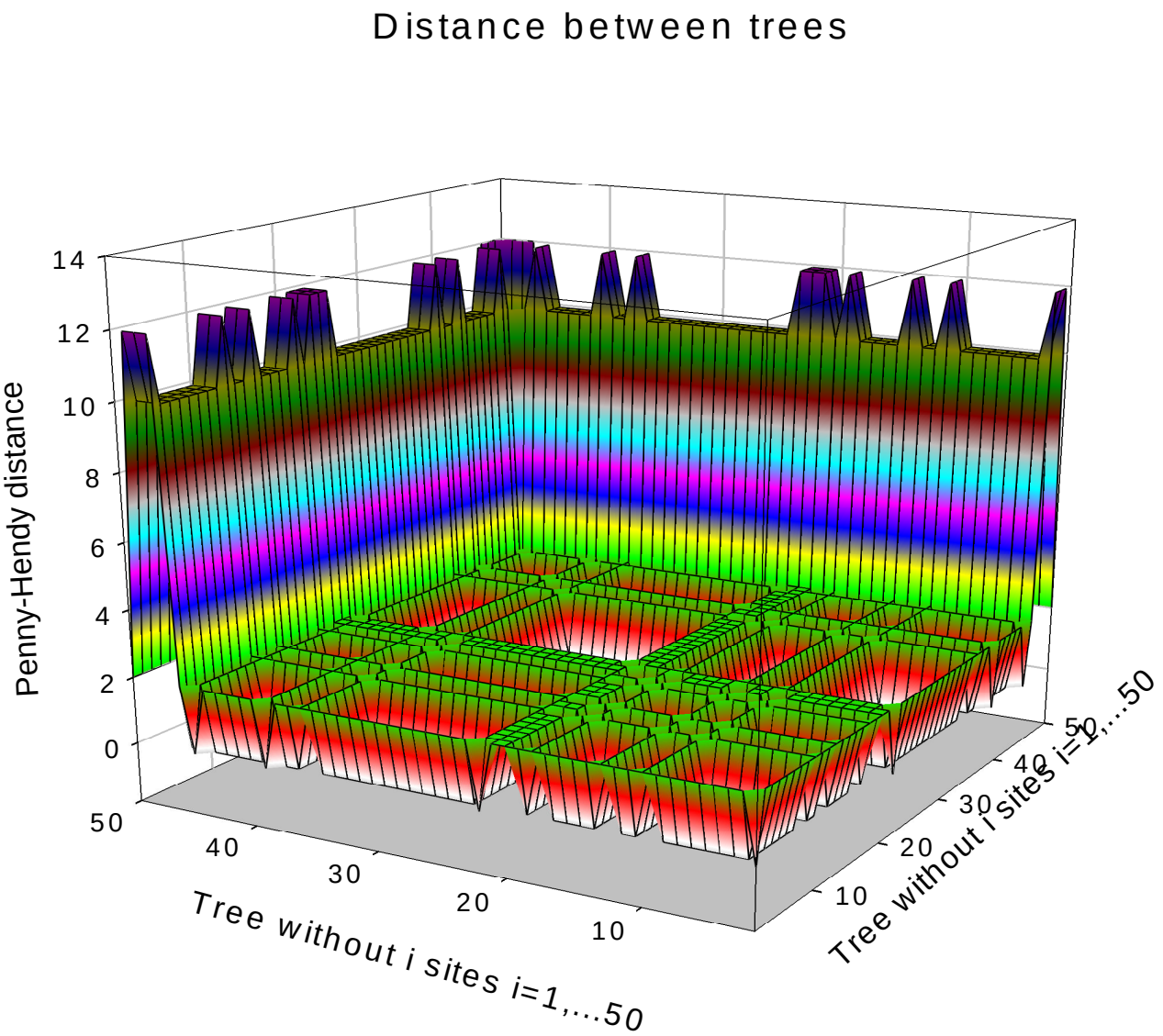


Information About Sites





Distance Between Trees



$d(T_0, T_i) \approx 18$ and $d(T_i, T_j) \leq 2$ for $i, j = 1..45$

Observations

- Strongest Outlier (position 142): Highly variable site located on a med loop (5nt) located on a conserved hairpin;
- Removing most influent sites leads to Increased bootstrap values and loss of 20% of inner nodes;
- Confirms monophyly of phyla Glomeromycota
- Reinforces polyphyletic status of phyla Chytridiomycota and Zygomycota

Taxon Influence Index: Motivation and Goal

Motivation: Filter Data

- Study the robustness of the tree with respect to the species
- Identify rogue taxa.

Procedure

- Quantify influence of taxon i by:
  - Removing site i from alignment;
  - Computing new tree T^{-i} from smaller (jackknife) alignment;
  - Comparing new tree to original tree;
- Compute phylogeny from "not too influent sites".

Taxon Influence Index: Motivation and Goal

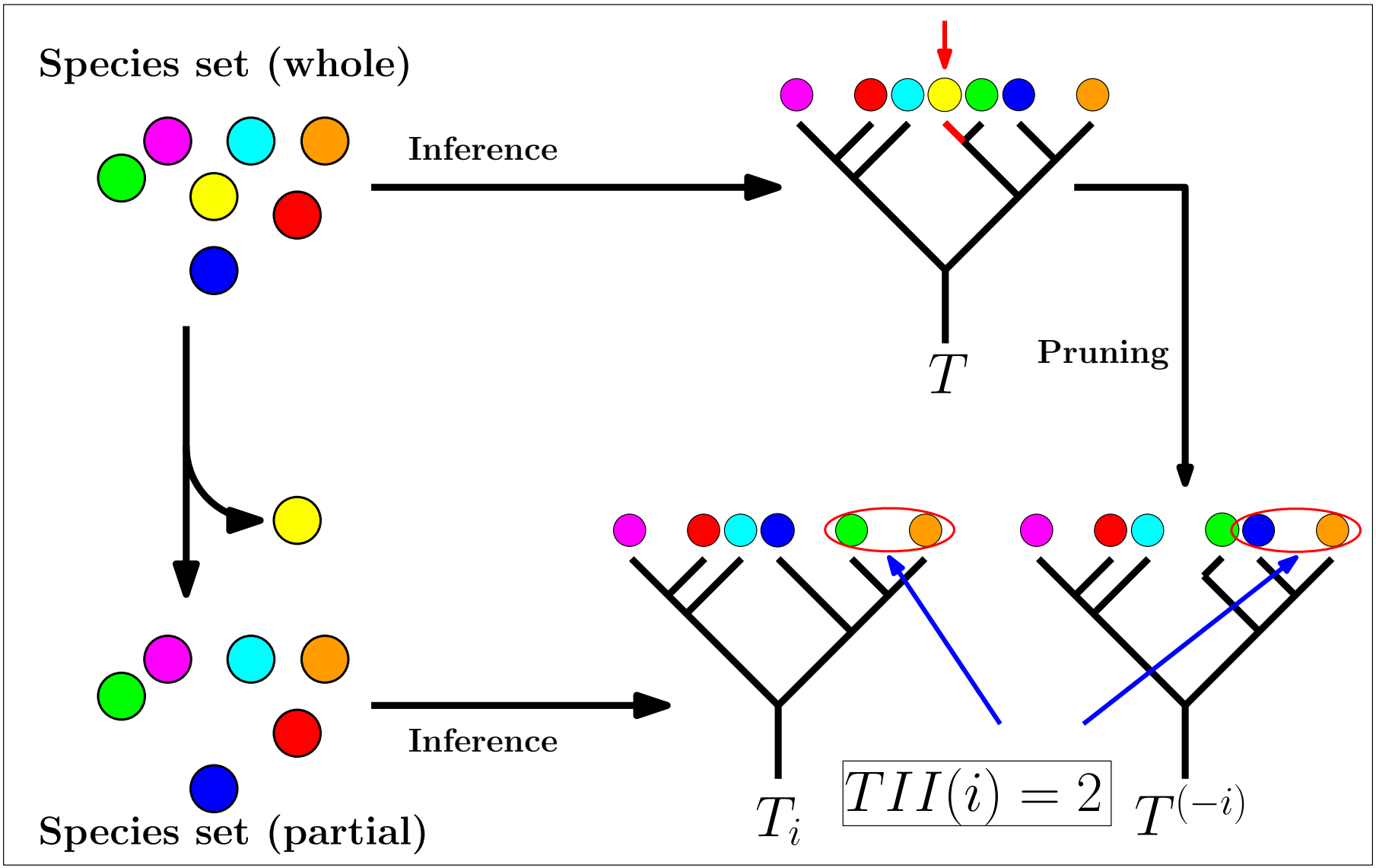
Motivation: Filter Data

- Study the robustness of the tree with respect to the species
- Identify rogue taxa.

Procedure

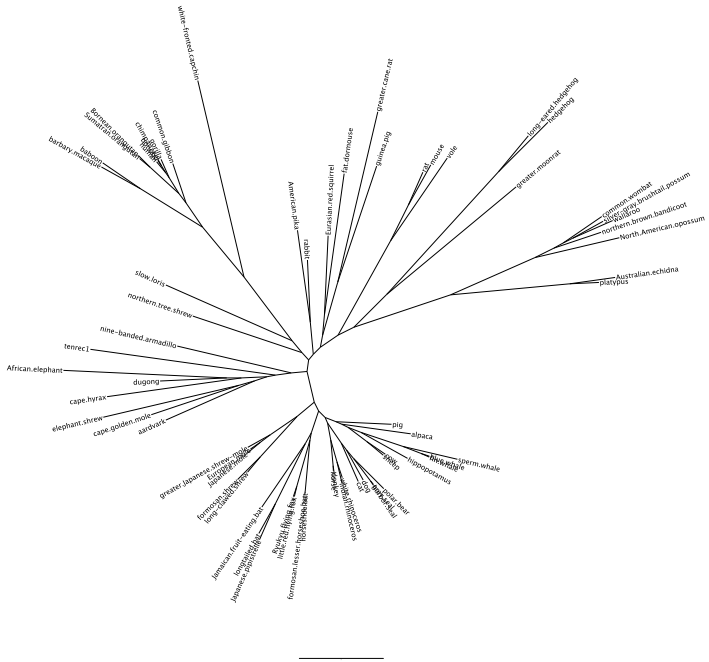
- Quantify influence of taxon i by:
  - Removing site i from alignment;
  - Computing new tree T^{-i} from smaller (jackknife) alignment;
  - Comparing new tree to original tree;
- Compute phylogeny from "not too influent sites".

Method

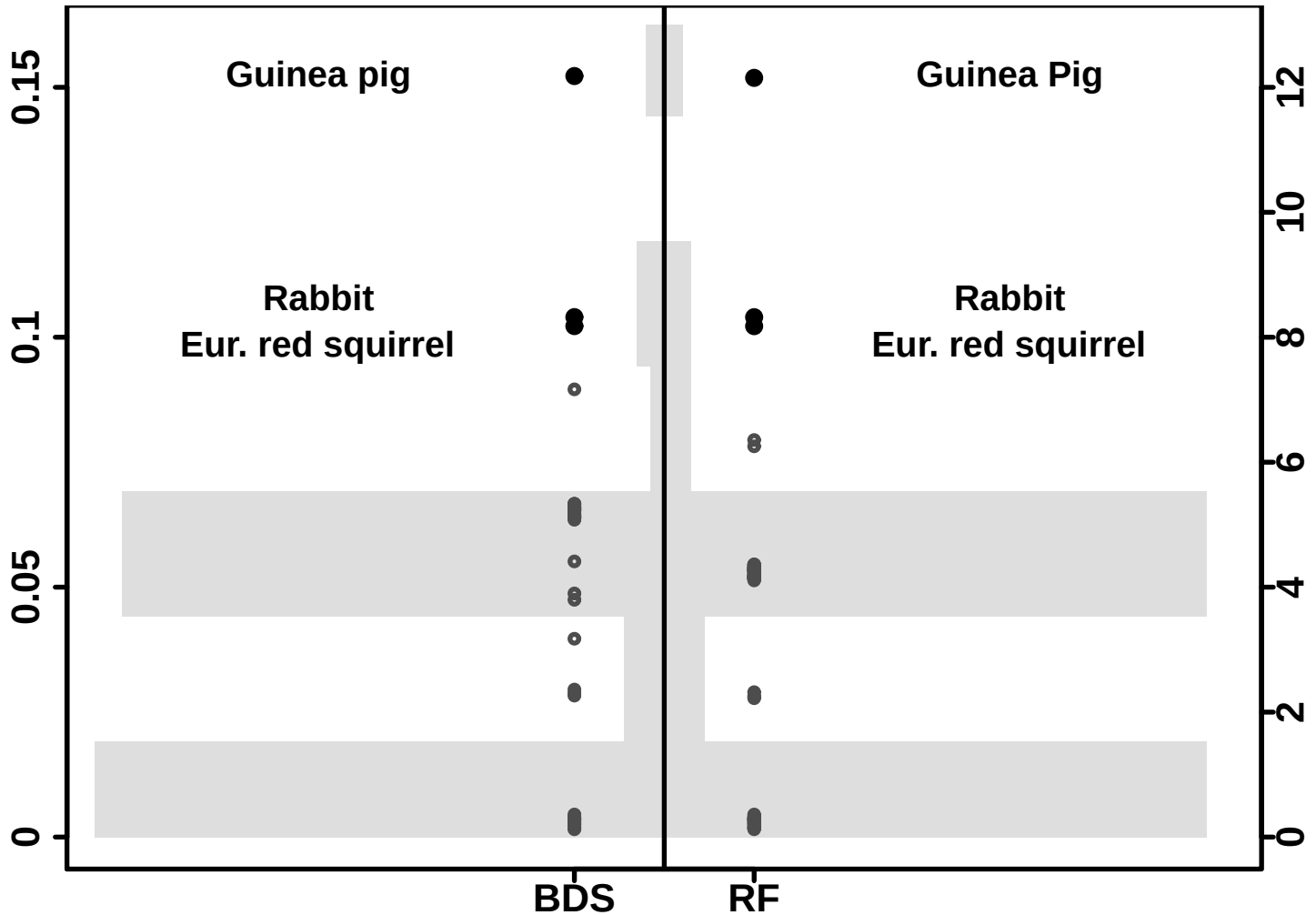

Species set (whole)
Inference
T
Pruning
Species set (partial)
Inference
T_i
$TII(i) = 2$
$T^{(-i)}$

Data: Placental Mammal Phylogeny

- Mitochondrial genome of 68 mammals;
- Amino Acids sequences;
- Sequences are 3658 AA long;
- MtMam + I + $\Gamma 4$ model;
- Phylogeny published in Nikaido *et al.* in 2003.

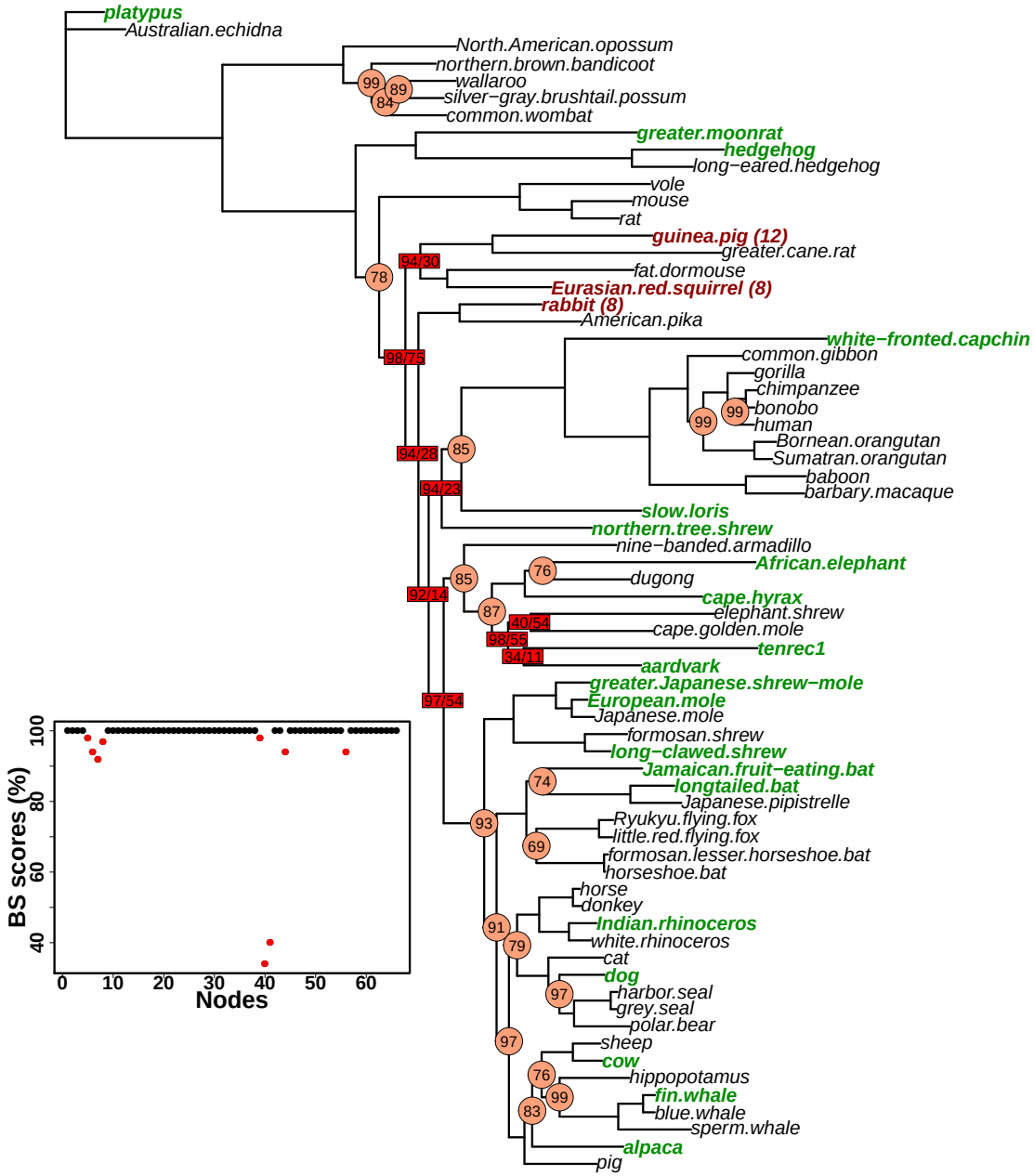


Taxon Influence Index

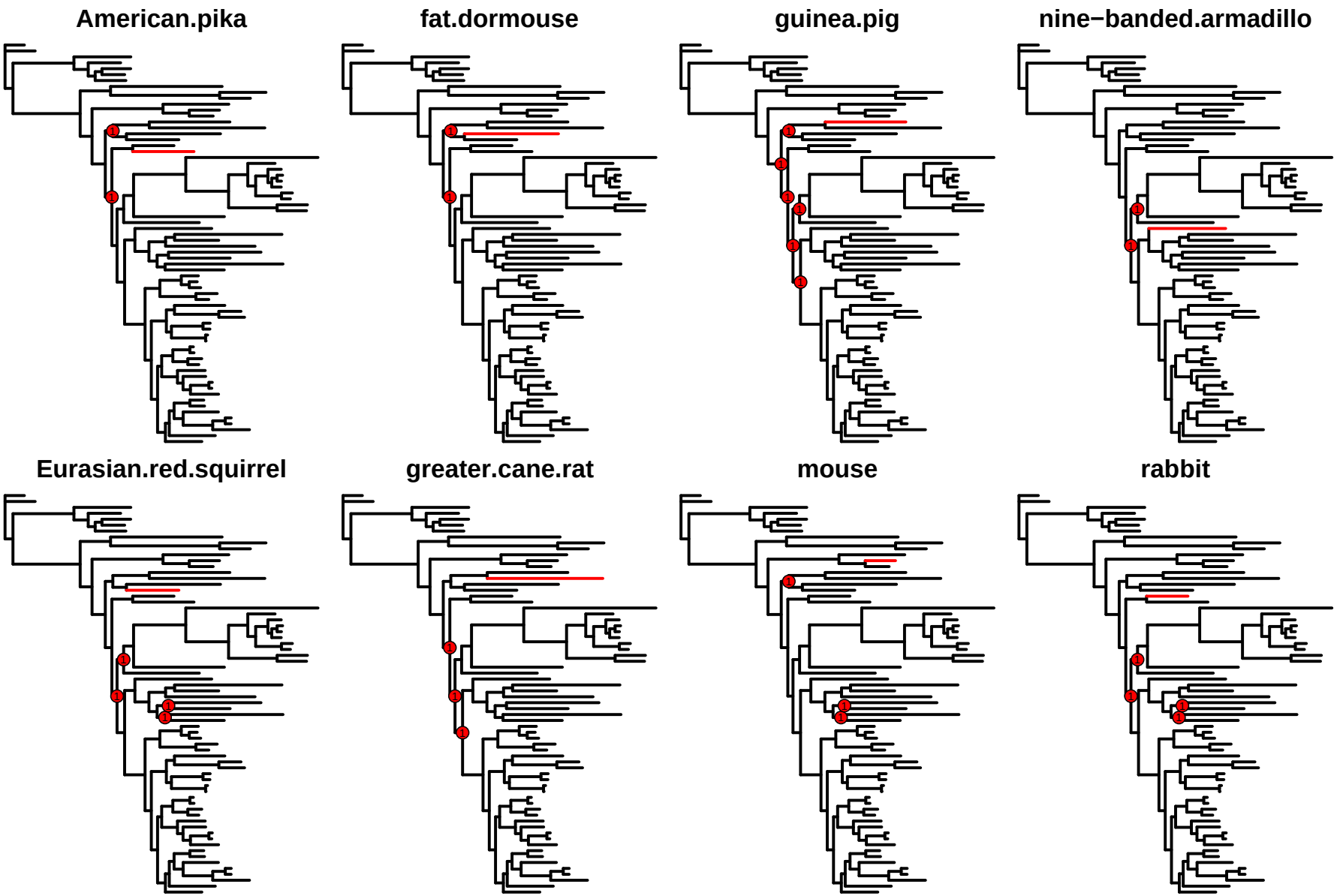


BSD: Branch-score distance / RF: Robinson-Foulds distance
mean edge length: 0.05

Complete Phylogeny

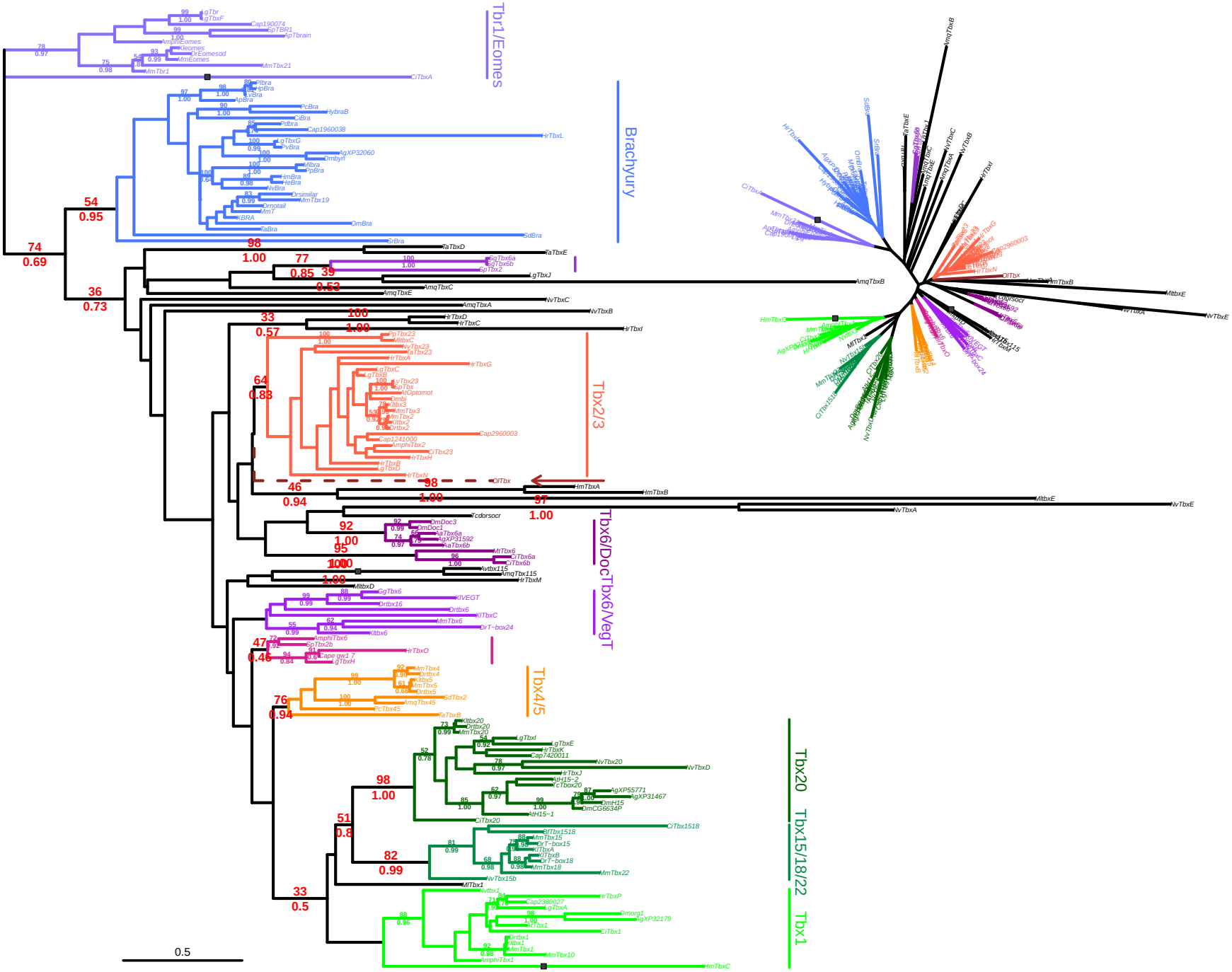


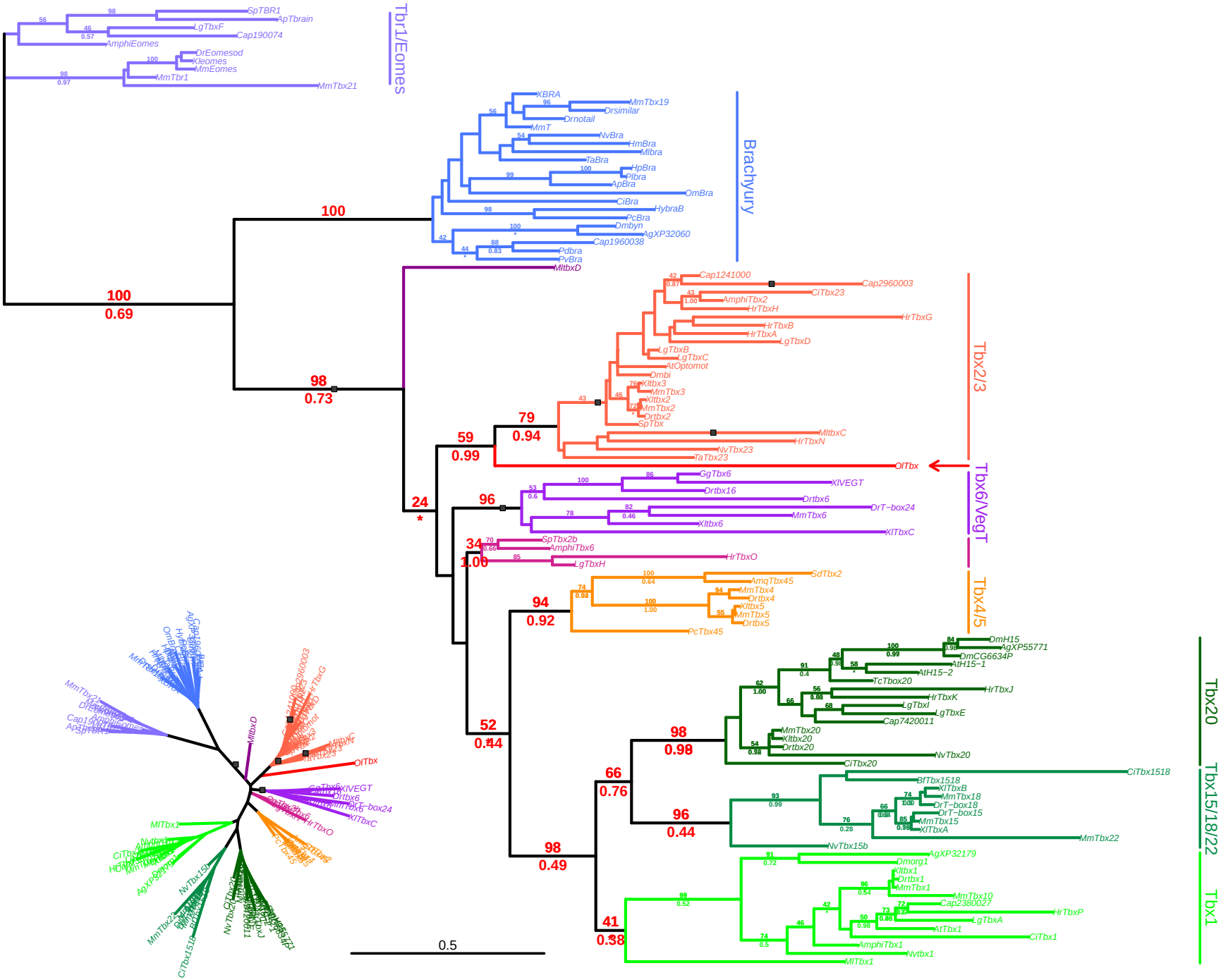
Influential Species (mostly in Afrotheria)



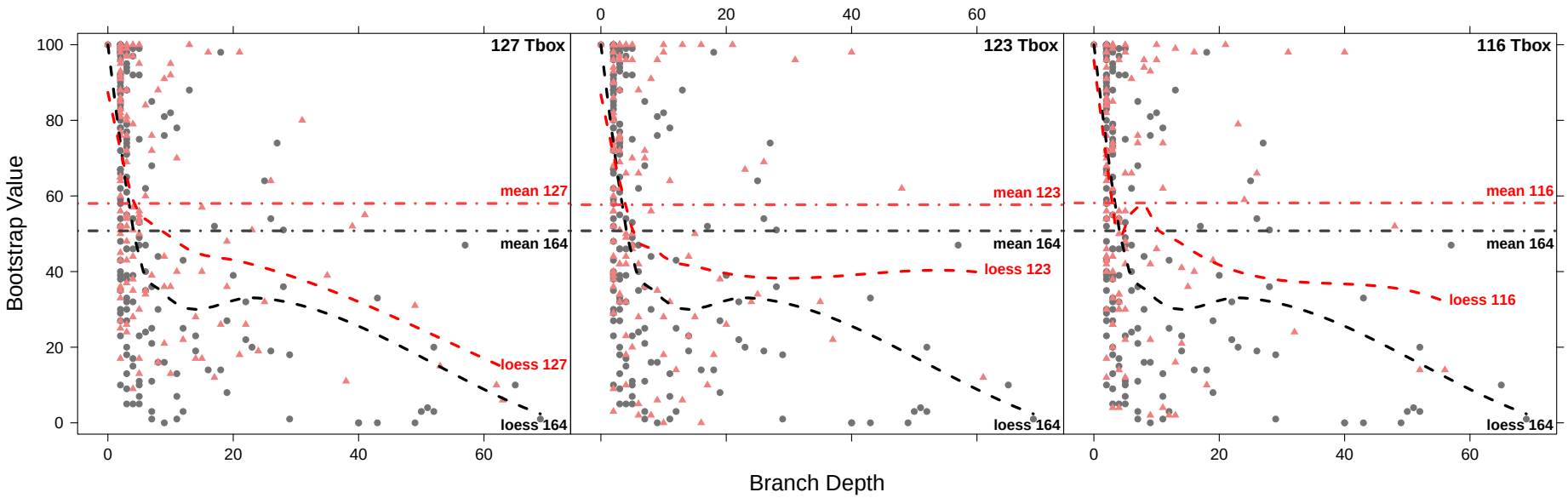
Data: Bilaterian Transcription Factor T-Box

- T-box TF of 164 metazoans, involved in gastrulation;
- Amino Acids sequences;
- Sequences are 296 AA long;
- LG + I + $\Gamma 4$ model;
- Ancient family with 8 subfamilies
  - *Brachyury*
  - *Tbx1/10*
  - *Tbx15/18/22*
  - *Tbx20*
  - *Tbx2/3*
  - *Tbx4/5*
  - *Tbx6/VegT*
  - *Eomes/Tbr1/Tbx21*
- Interest lies in the position of TF OlTbx present in a *Oscarella lobularis*.


Tbr1/Eomes
Brachyury
Tbx2/3
Tbx6/Doc
Tbx6/VegT
Tbx4/5
Tbx20
Tbx15/18/22
Tbx1
0.5


Tbr1/Eomes
Brachyury
Tbx2/3
Tbx6/VegT
Tbx4/5
Tbx20
Tbx15/18/22
Tbx1
SpTBR1
ApTbrain
LgTbxF
Cap190074
AmphiEomes
DrEomesod
Xleomes
MmEomes
MmTbr1
MmTbx21
XBRA
MmTbx19
Drsimilar
Drnotail
MmT
NvBra
HmBra
MlBra
TaBra
HpBra
PlBra
ApBra
OmBra
CiBra
HybraB
PcBra
Dmbyn
AgXP32060
Cap1960038
Pdbra
PvBra
MltbxD
Cap1241000
Cap2960003
CiTbx23
AmphiTbx2
HrTbxH
HrTbxG
HrTbxB
HrTbxA
LgTbxD
LgTbxB
LgTbxC
AtOptomot
Dmbi
Xltbx3
MmTbx3
Xltbx2
MmTbx2
Drtbx2
SpTbx
MltbxC
HrTbxN
NvTbx23
TaTbx23
OlTbx
GgTbx6
Drtbx16
XlVEGT
Drtbx6
DrT−box24
MmTbx6
Xltbx6
XlTbxC
SpTbx2b
AmphiTbx6
LgTbxH
HrTbxO
SdTbx2
AmqTbx45
MmTbx4
Drtbx4
Xltbx5
MmTbx5
Drtbx5
PcTbx45
DmH15
AgXP55771
DmCG6634P
AtH15−1
AtH15−2
TcTbox20
HrTbxJ
HrTbxK
LgTbxI
LgTbxE
Cap7420011
MmTbx20
Xltbx20
Drtbx20
NvTbx20
CiTbx20
CiTbx1518
BfTbx1518
XlTbxB
MmTbx18
DrT−box18
DrT−box15
MmTbx15
XlTbxA
MmTbx22
NvTbx15b
AgXP32179
Dmorg1
Xltbx1
Drtbx1
MmTbx1
MmTbx10
Cap2380027
HrTbxP
LgTbxA
AtTbx1
CiTbx1
AmphiTbx1
Nvtbx1
MlTbx1
0.5

T-Box TF Phylogeny



Number of Tbx in tree	164	127	123	116
BP of OlTbx in clade TBX2/3	19	32	34	59

No sponge (yet) identified as a member of the Tbx2/3 subfamily: new evolutionary hypothesis.

Summary

Potential Sources of Instability and Indexes to Detect Them

1. Data sampling: Bootstrap;
2. Outliers: Influence function for sites;
3. Rogue species: Taxon Influence Index.

Pros and Cons

1. Assess uncertainty coming from different sources;
2. Highlight potential outliers;
3. No rigorous statistical threshold for inclusion (p-values,...)