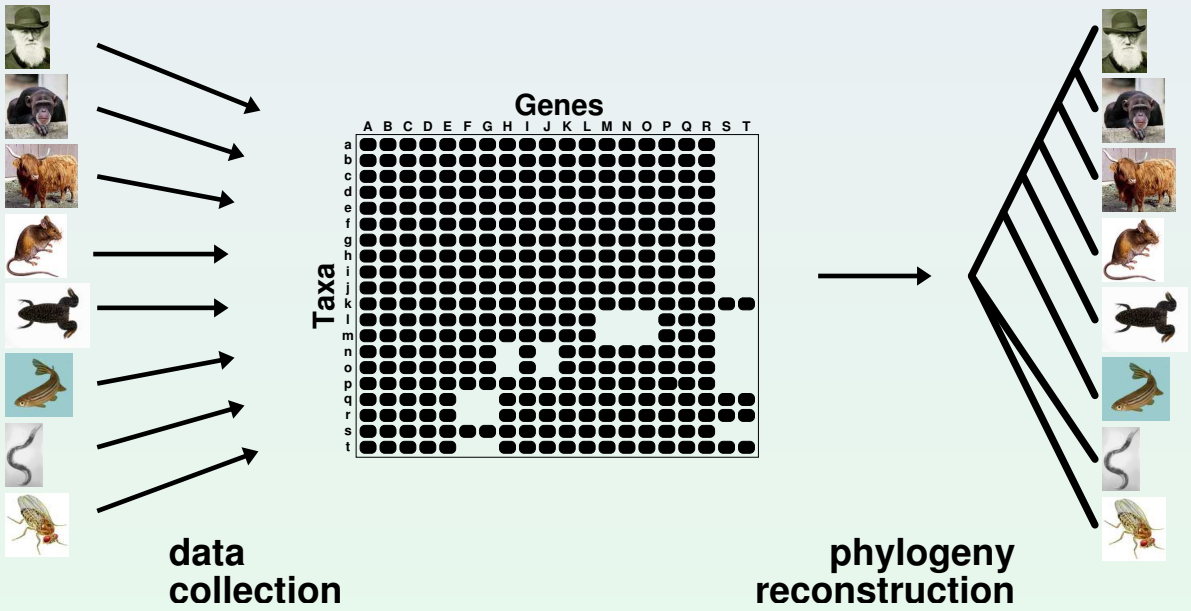
Comparison of commonly used methods for combining multiple phylogenetic data sets

Anne Kupczok, Heiko A. Schmidt and Arndt von Haeseler

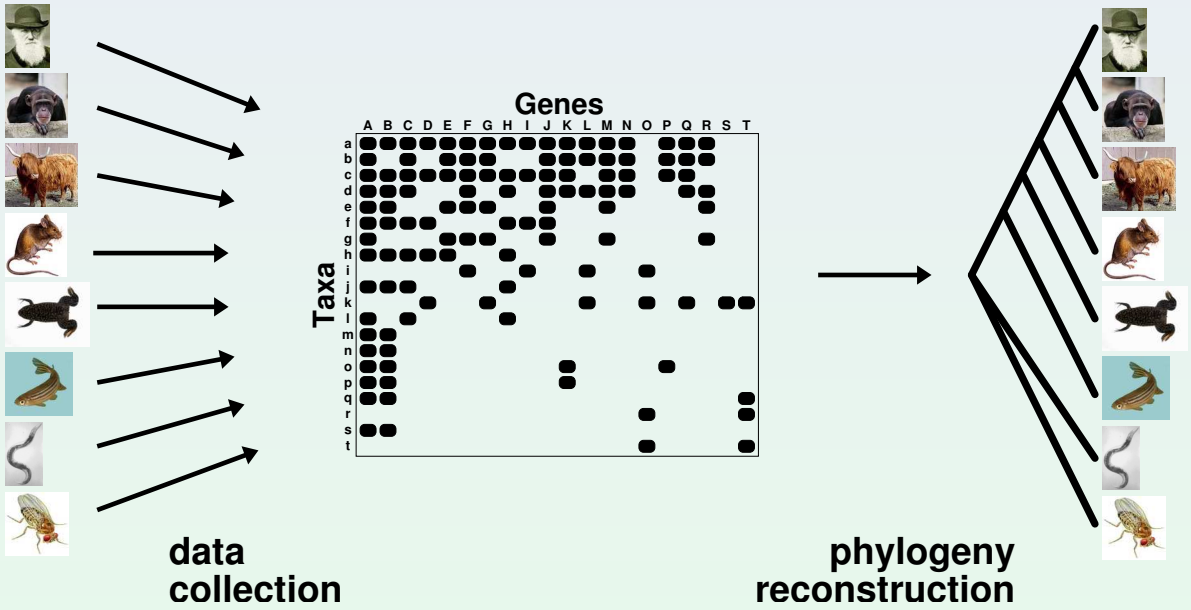
Center for Integrative Bioinformatics Vienna
Max F. Perutz Laboratories

June 12th, 2008

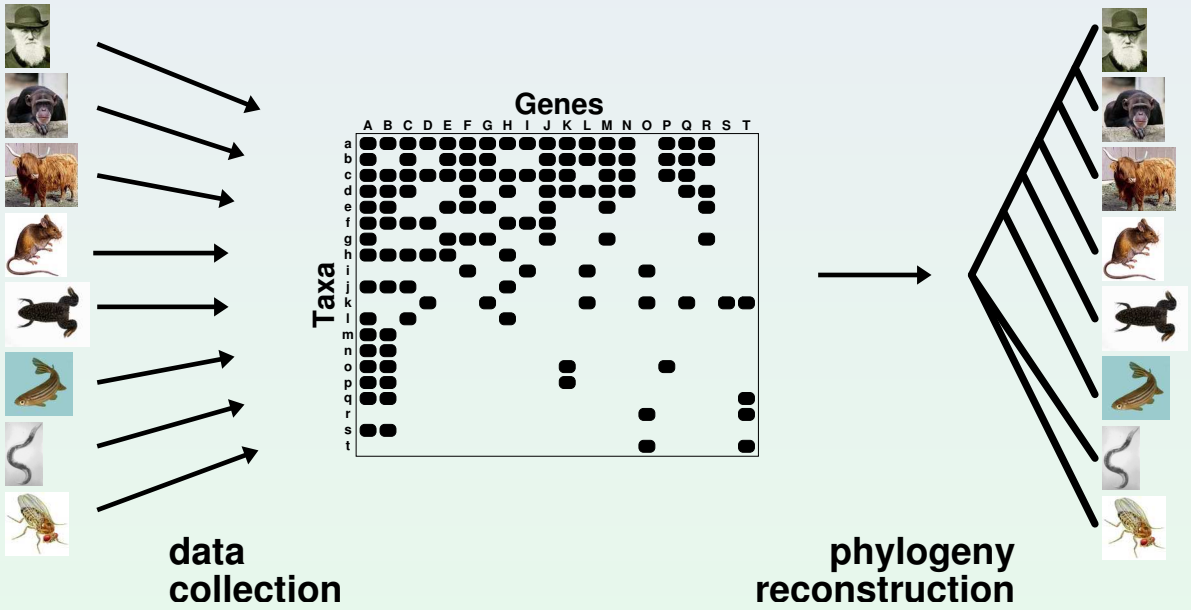
Multi-Locus Datasets



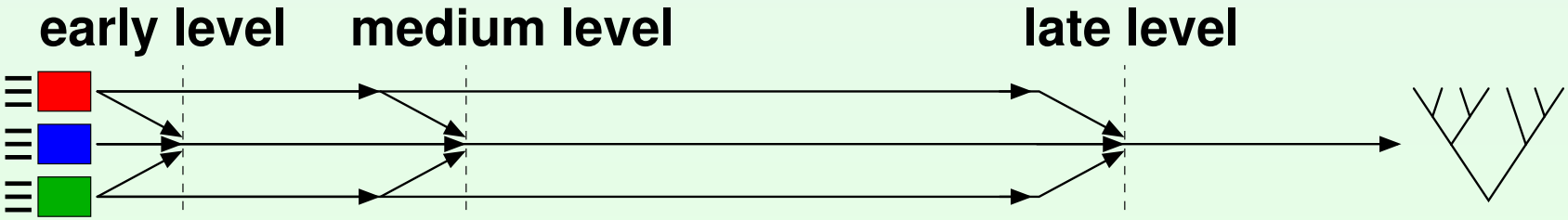
Multi-Locus Datasets



Multi-Locus Datasets



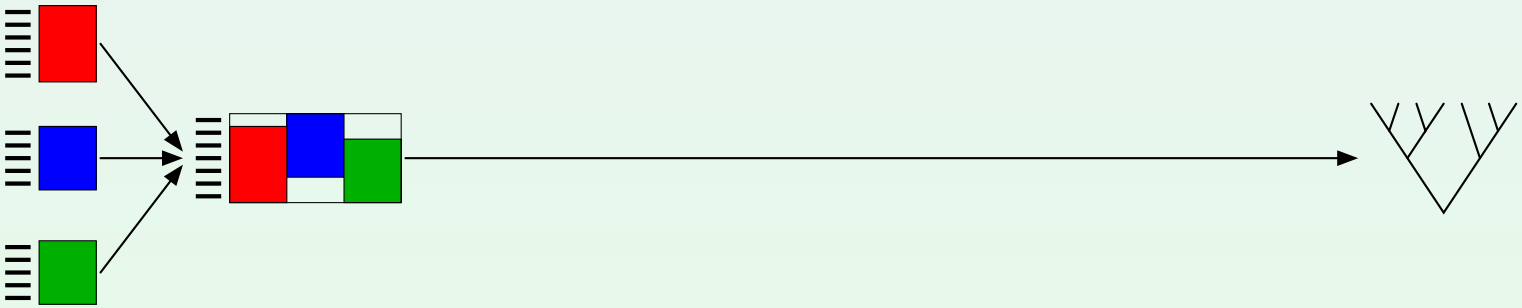
Approaches:



Early-level combination: Superalignment

= Supermatrix or 'Total Evidence'

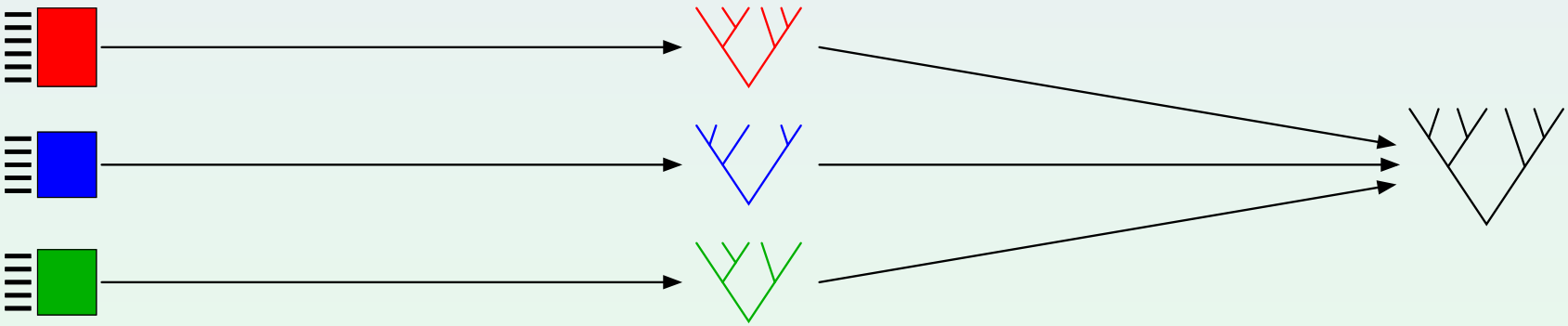
- Combination by concatenating data sets:



- Any tree reconstruction method can be applied to the data matrix

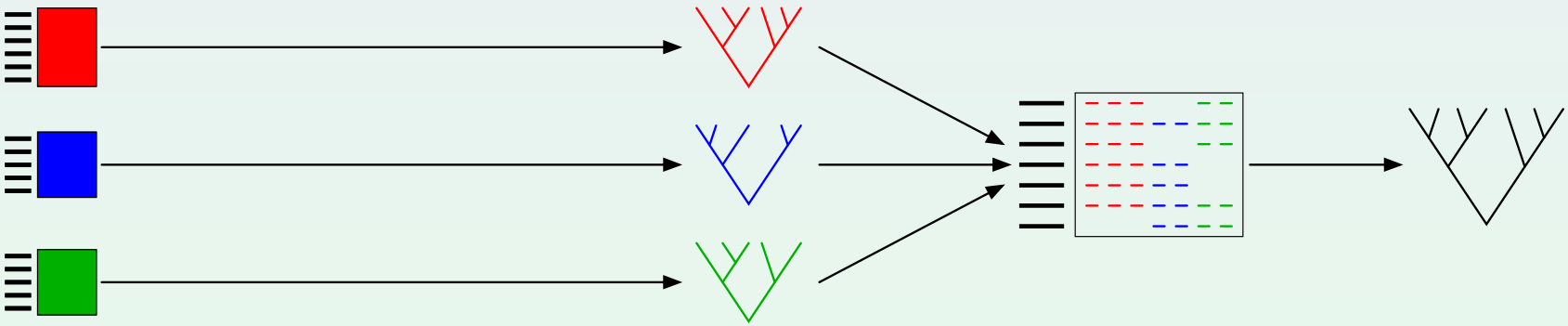
Late-level combination: Supertree

Construct separate trees for each gene and combine them to a supertree:



Late-level combination: Supertree

Construct separate trees for each gene and combine them to a supertree:



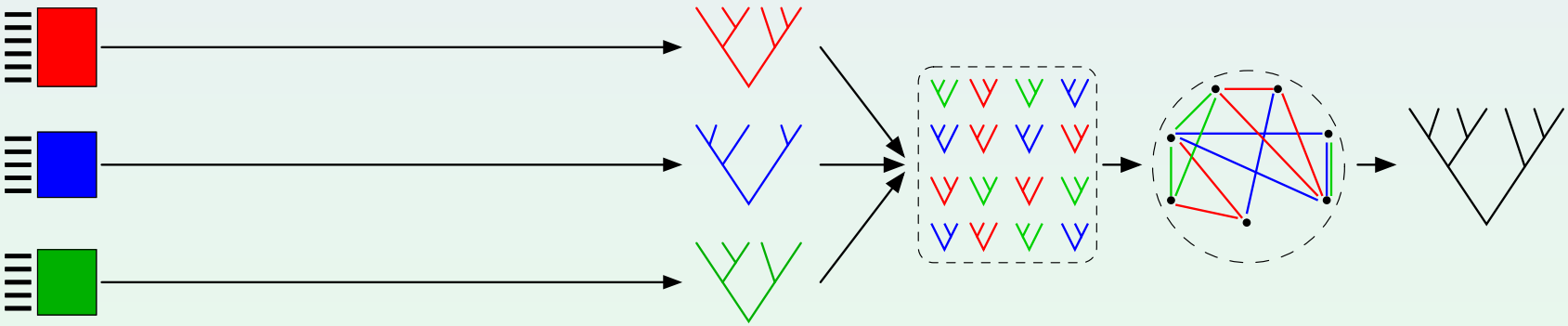
Supertree methods combine special kinds of information:

Split information → Matrix Representation:

- MR with Parsimony (MRP, Baum, 1992; Ragan, 1992)
- MR with Flipping (MRF, e.g. Chen et al., 2003)

Late-level combination: Supertree

Construct separate trees for each gene and combine them to a supertree:



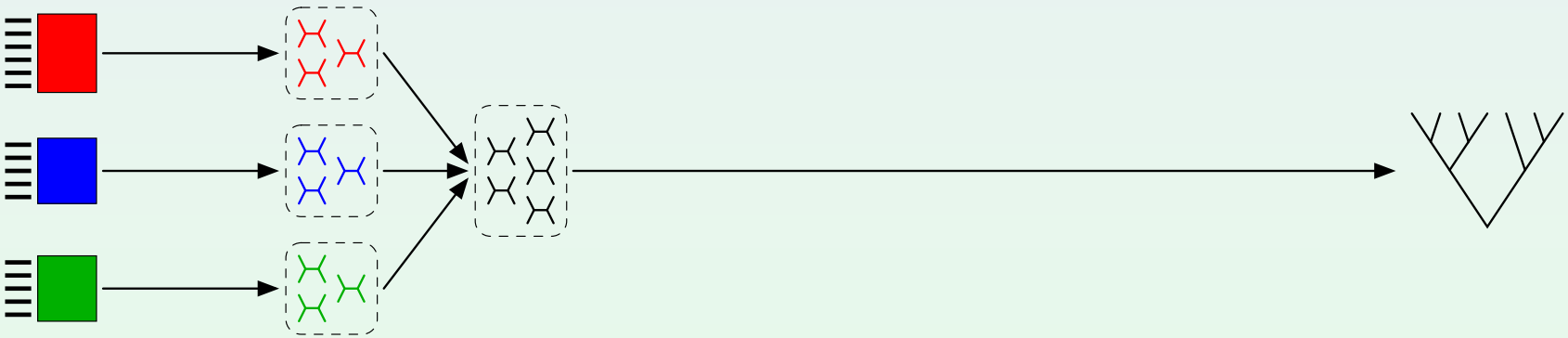
Supertree methods combine special kinds of information:

Triplet information → Rooted triplets:

- MinCut (Semple and Steel, 2000)
- Modified MinCut (Page, 2002)
- MaxCut (Snir and Rao, 2006)

Medium-level combination

Intermediate data (not final trees) is computed from every source alignment and subsequently combined to a tree.



SuperQP: Combination of quartet likelihoods (Schmidt, 2003)

Medium-level combination

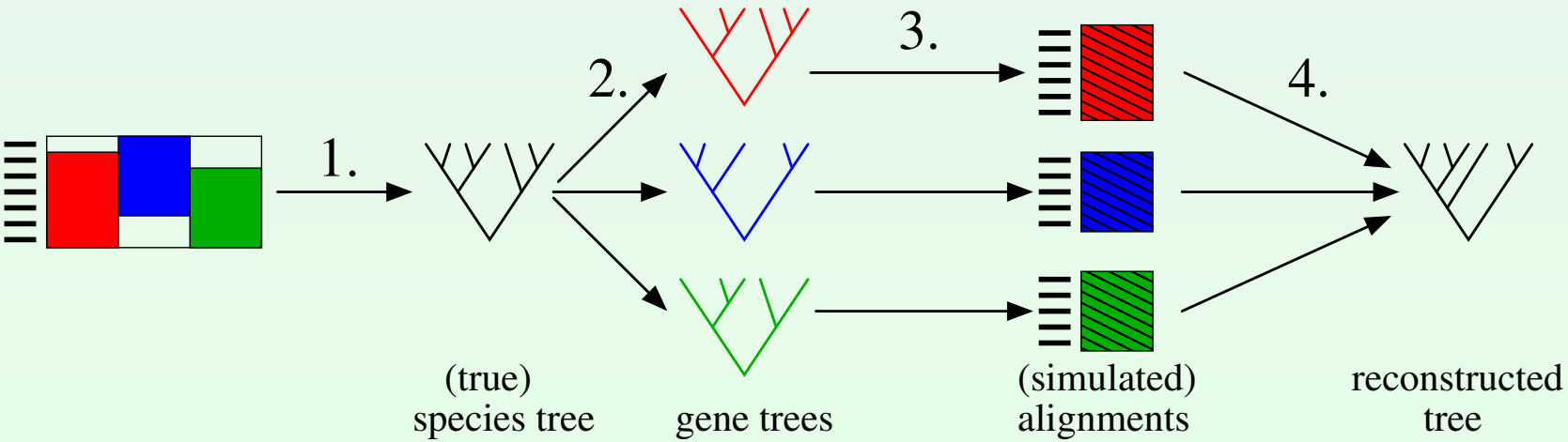
Intermediate data (not final trees) is computed from every source alignment and subsequently combined to a tree.

Average Consensus: Average over distance matrix for each gene (Lapointe and Cucumel, 1997)

SDM: Additional weights estimated (Criscuolo et al., 2006)

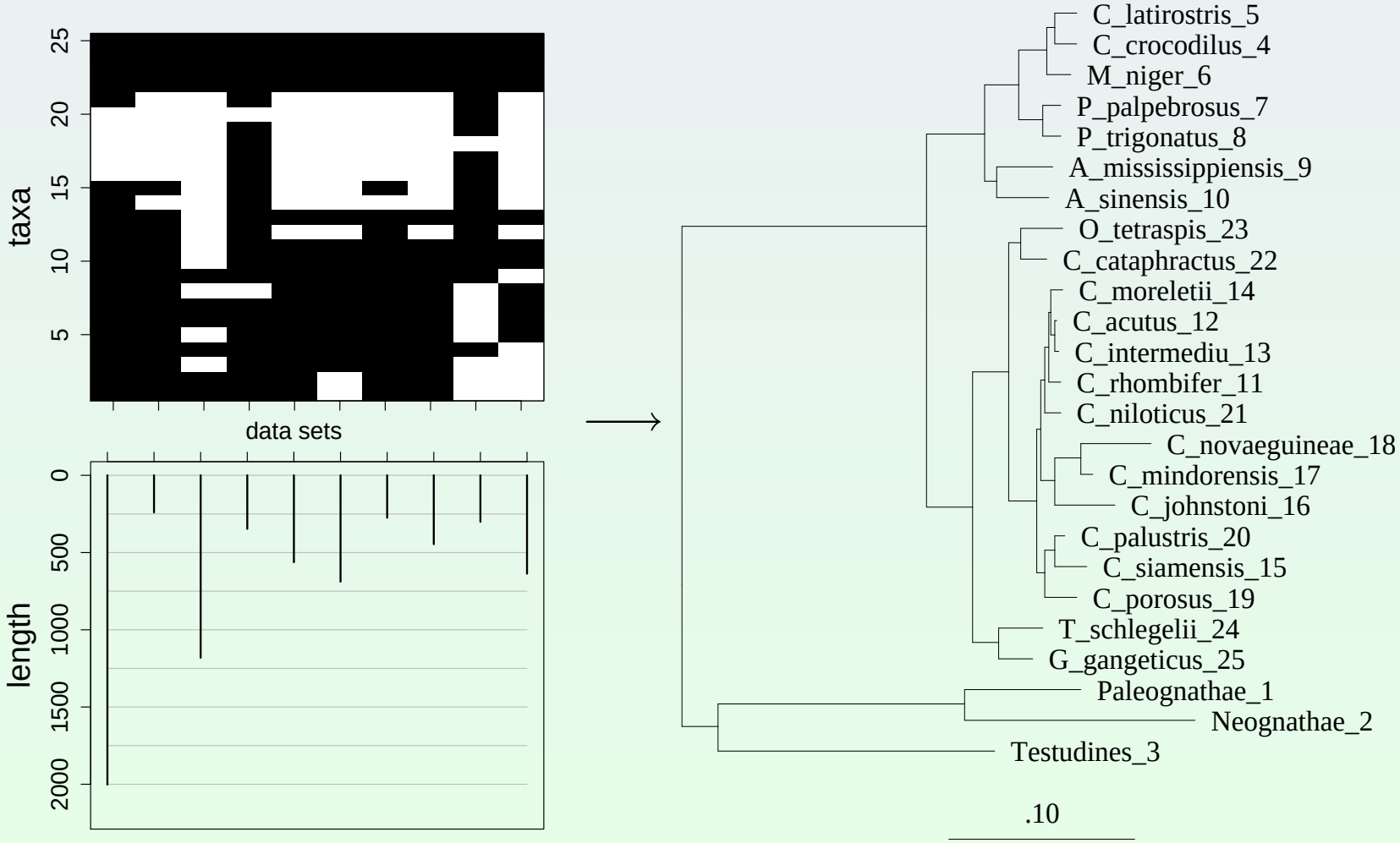
Simulation setting

1. Estimate an ML tree with branch lengths and model parameters from a data superalignment $\longrightarrow$ species tree
2. Generate gene trees
3. Simulate alignments along the gene trees
4. Apply the reconstruction methods to each data set and compare the result with the model tree



Species tree

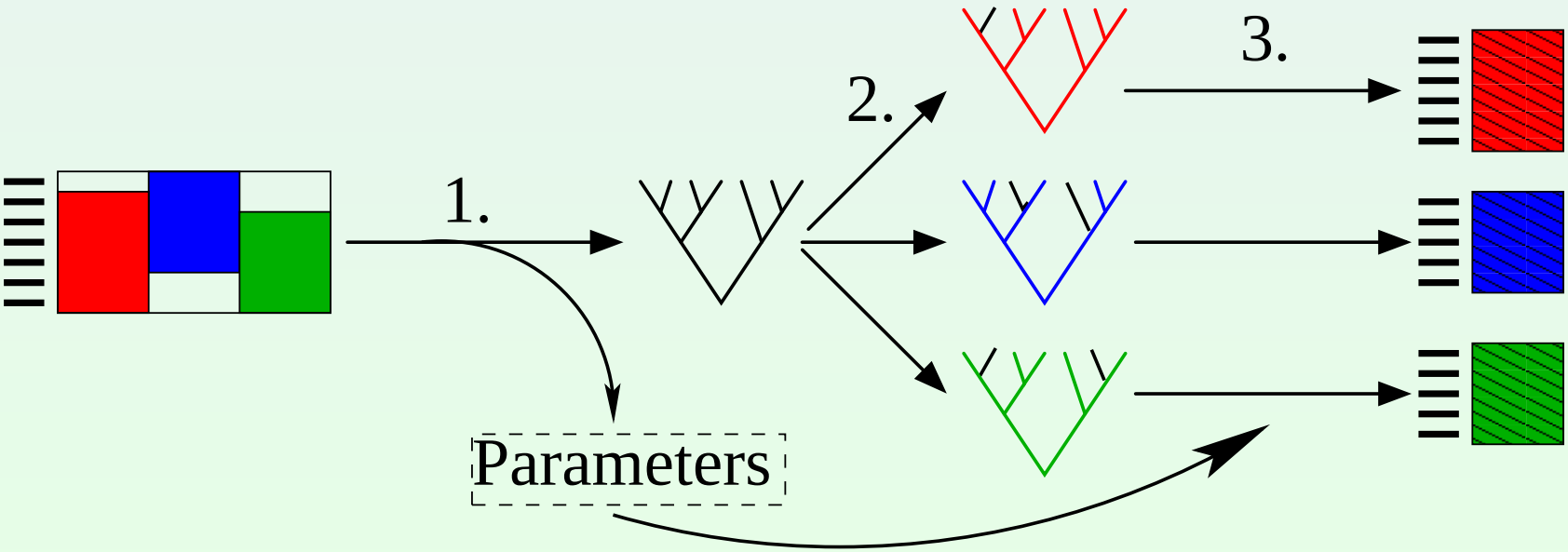
10 genes of 25 Crocodylia species (Gatesy et al., 2004)

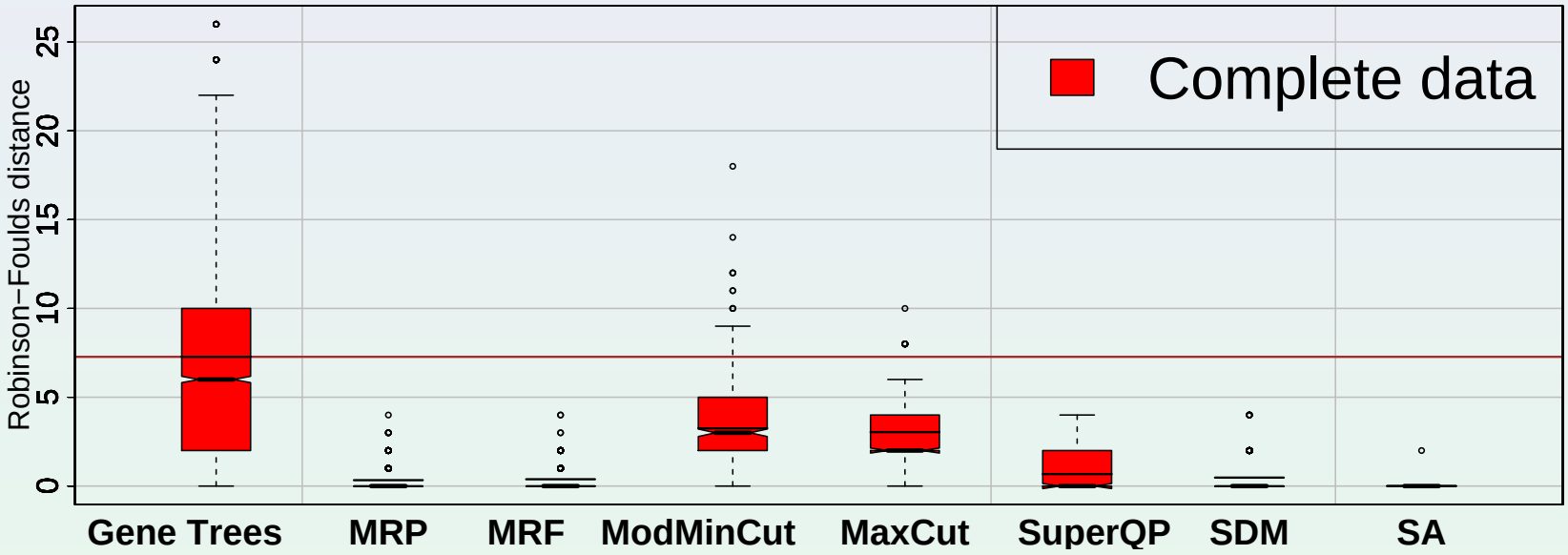


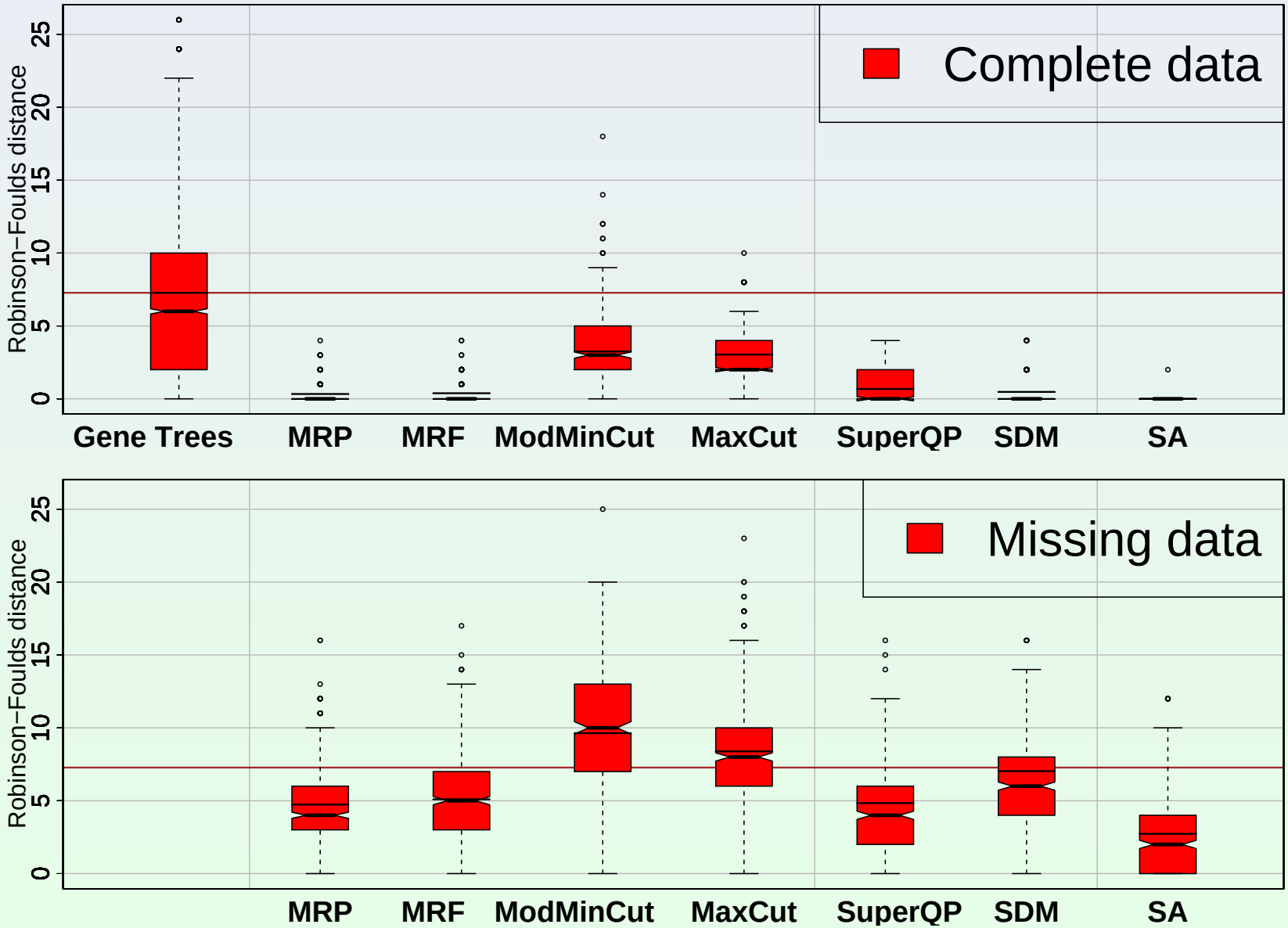
Complete and missing data

Step 2: Gene trees are the complete model tree (complete data) or the pruned model tree (missing data)

Step 3: Simulation with the parameters estimated with the superalignment



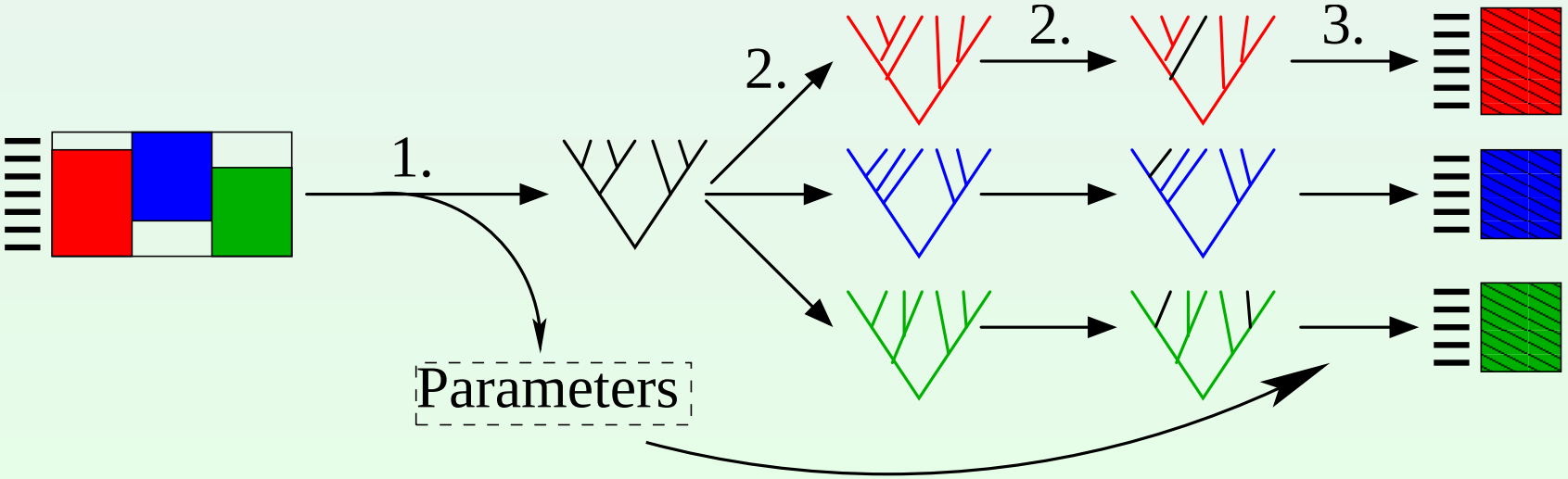

Robinson−Foulds distance
0
5
10
15
20
25
Complete data
Gene Trees
MRP
MRF
ModMinCut
MaxCut
SuperQP
SDM
SA

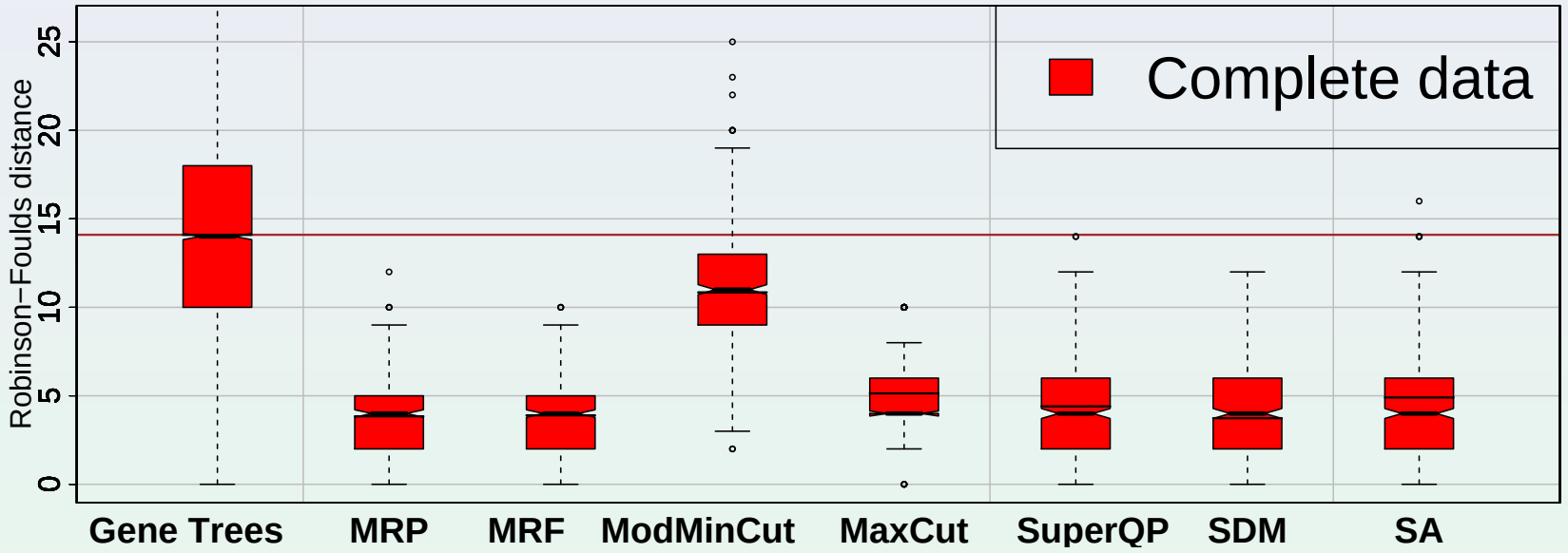

Robinson−Foulds distance
0
5
10
15
20
25
Complete data
Gene Trees
MRP
MRF
ModMinCut
MaxCut
SuperQP
SDM
SA
Robinson−Foulds distance
0
5
10
15
20
25
Missing data
MRP
MRF
ModMinCut
MaxCut
SuperQP
SDM
SA

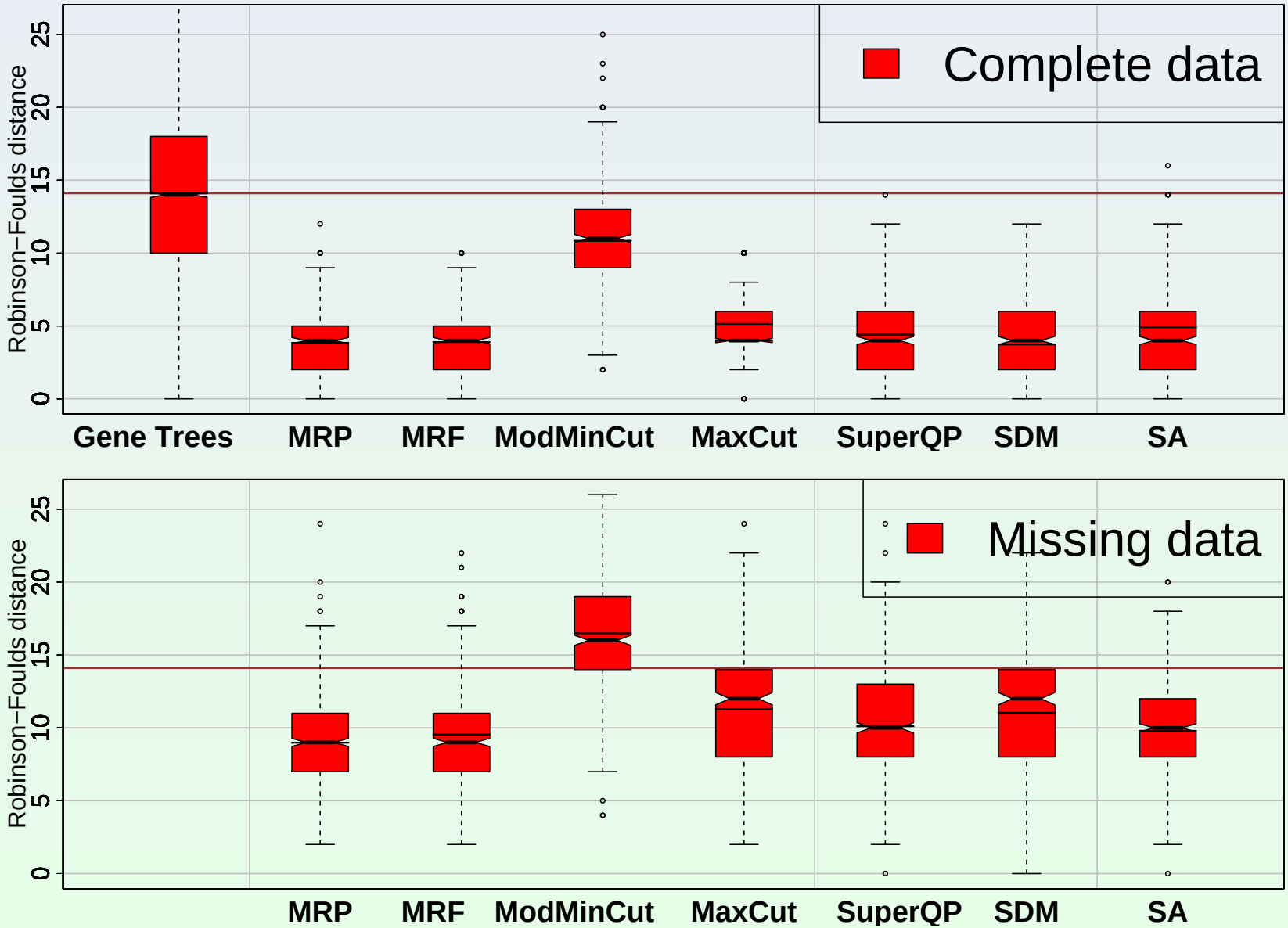
Incomplete lineage sorting

Step 2: For every simulation, a gene tree is generated from the species tree with a coalescent process ($\theta = 0.005$)

Step 3: Simulation with the parameters estimated with the superalignment




Robinson−Foulds distance
0
5
10
15
20
25
Complete data
Gene Trees
MRP
MRF
ModMinCut
MaxCut
SuperQP
SDM
SA


Robinson−Foulds distance
0
5
10
15
20
25
Complete data
Gene Trees
MRP
MRF
ModMinCut
MaxCut
SuperQP
SDM
SA
Robinson−Foulds distance
0
5
10
15
20
25
Missing data
MRP
MRF
ModMinCut
MaxCut
SuperQP
SDM
SA

Summary

- Simulation of sequence-based phylogenetic analysis for multiple data sets
- With the assumption of tree-like evolution for most genes, superalignment yields the highest accuracy
- In case of high incongruency among gene trees other methods may outperform superalignment
- Matrix Representation methods are the best choice for supertree reconstruction

Summary

- Simulation of sequence-based phylogenetic analysis for multiple data sets
- With the assumption of tree-like evolution for most genes, superalignment yields the highest accuracy
- In case of high incongruency among gene trees other methods may outperform superalignment
- Matrix Representation methods are the best choice for supertree reconstruction

Acknowledgements:

- Gregory Ewing (CIBIV)
- WWTF for funding