IQ-TREE
Methodology and Practice

Minh Bui
Australian National University

Workshop on Molecular Evolution
Woods Hole, August 2019

BIG THANKS TO THE MECHANICS!

IQ-TREE DEVELOPMENT TEAM

Lam Tung Nguyen

Google Scholar

Contribution: Tree search algorithm and parallelization.

Olga Chernomor

Google Scholar

Contribution: Partition models and phylogenomic search.

Heiko A. Schmidt

Google Scholar

Contribution: Integration of TREE-PUZZLE features.

Jana Trifinopoulos

Contribution: W-IQ-TREE web service.

Bui Quang Minh

Google Scholar

Contribution: Team leader, software core, ultrafast bootstrap, model selection.

Dominik Schrempf

Google Scholar

Contribution: Polymorphism-aware models (PoMo).

Michael Woodhams

Google Scholar

Contribution: Lie Markov models.

Arndt von Haeseler

Google Scholar

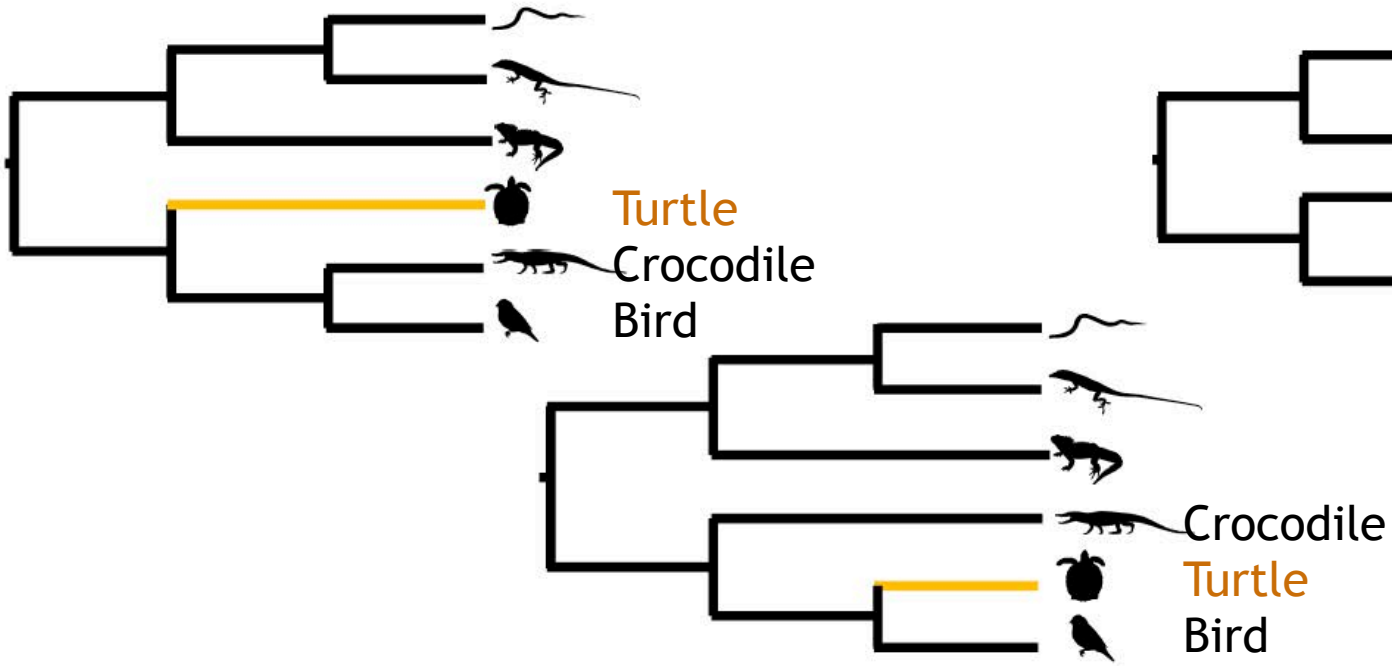
Contribution: Advice, ideas and financial support.

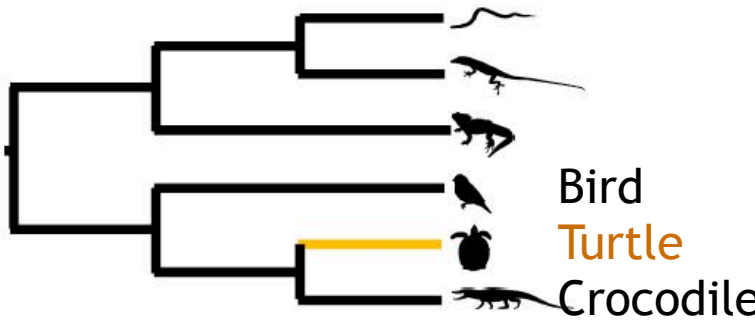
Diep Thi Hoang

Contribution: Improving ultrafast bootstrap.

Thanks to plenty of users for feedback and bug reports!

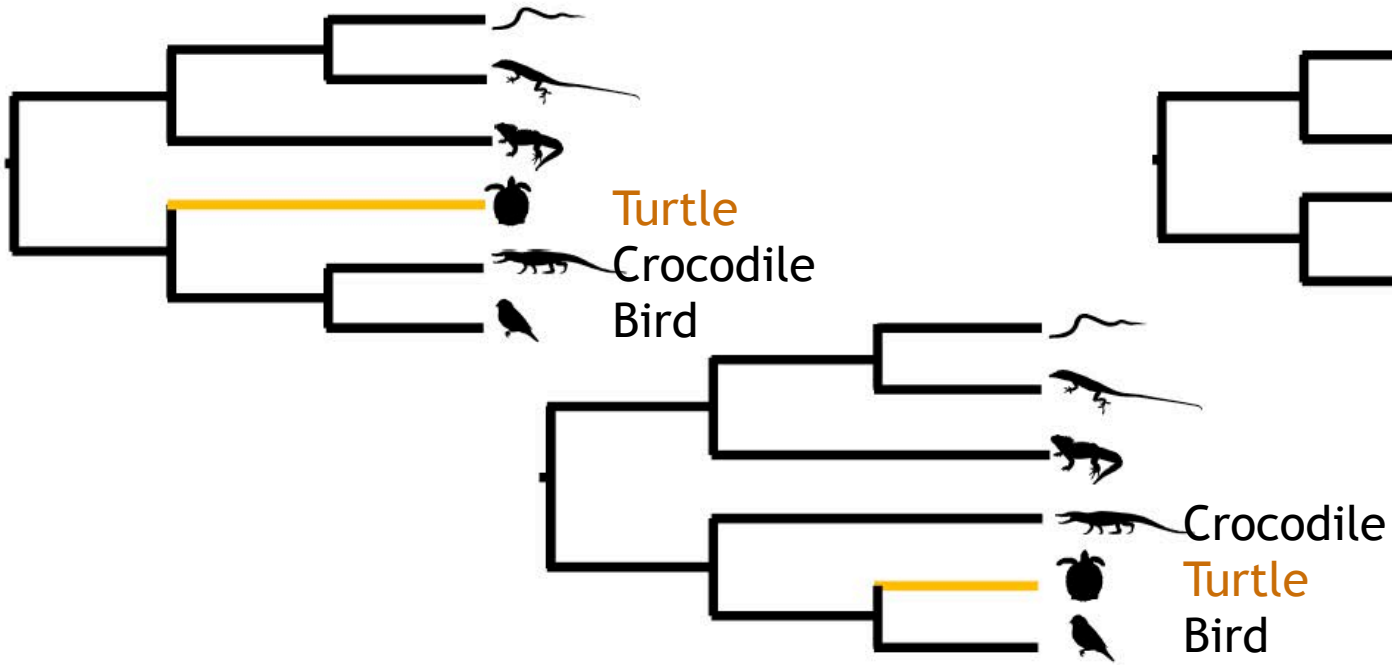
Practical data set: Where is Turtle in the tree?

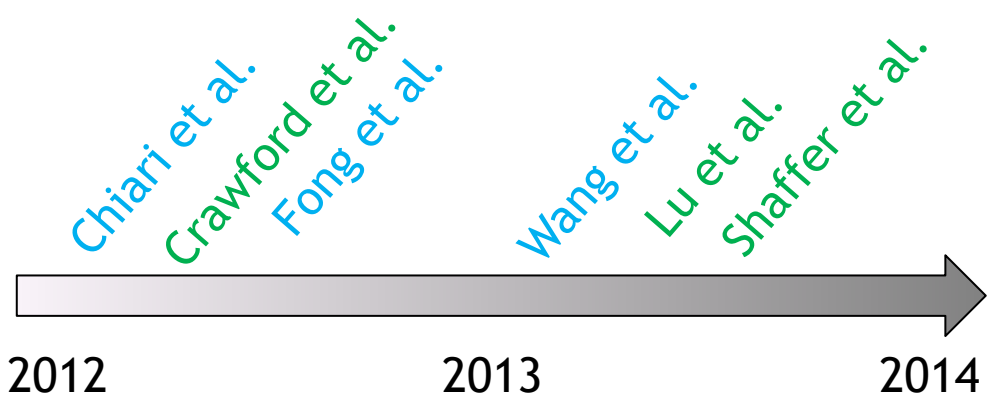




Thanks Jeremy Brown

Practical data set: Where is Turtle in the tree?

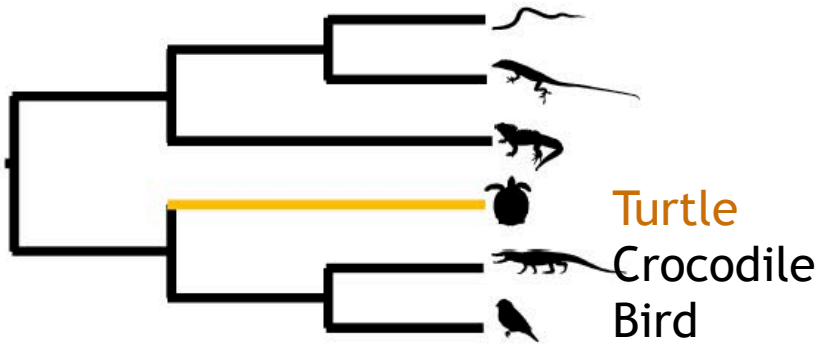


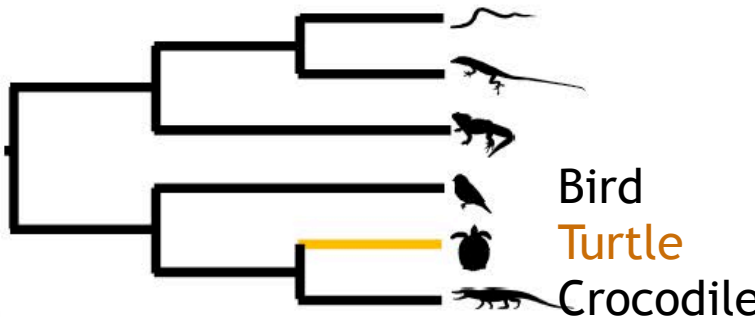


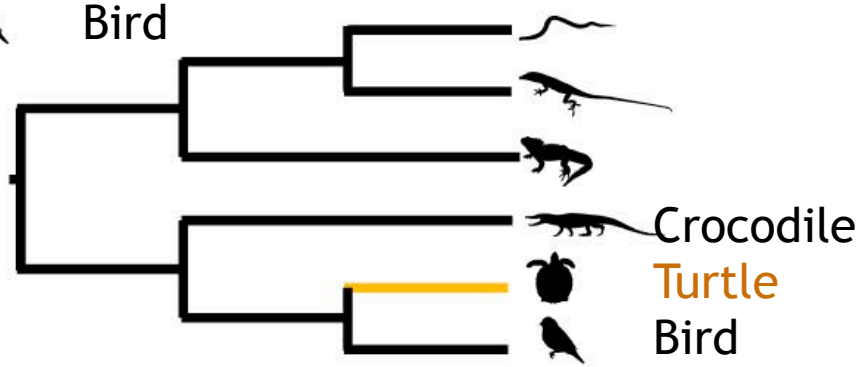
Different studies led to different trees!

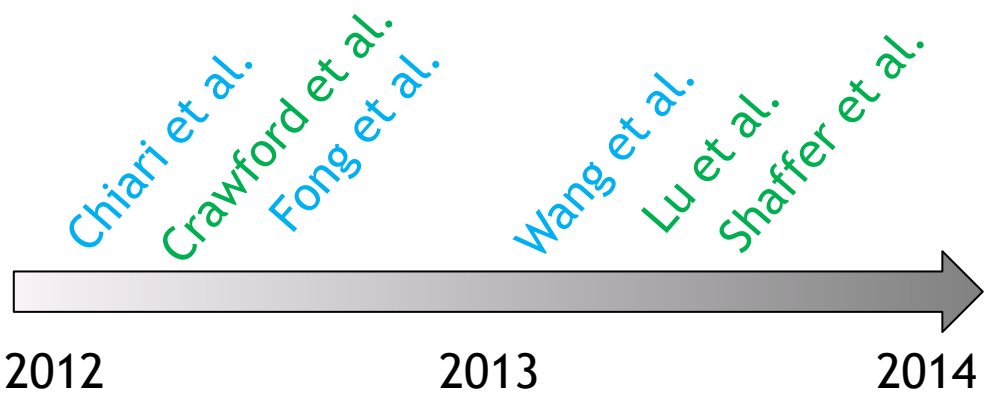
Thanks Jeremy Brown

Practical data set: Where is Turtle in the tree?









Different studies led to different trees!

Dataset: 16 species, 29 genes, 20,820 bp (a subset of Chiari et al. 2012)

Thanks Jeremy Brown

Why IQ-TREE?

Why IQ-TREE?

Next generation sequencing data represent both a blessing and a curse:

Why IQ-TREE?

Next generation sequencing data represent both a blessing and a curse:

- Blessing: (Phylo)genomic data help to elucidate many phylogenetic questions.

Why IQ-TREE?

Next generation sequencing data represent both a blessing and a curse:

- Blessing: (Phylo)genomic data help to elucidate many phylogenetic questions.
- Curse: Many model assumptions become increasingly distant from the truth due to growing data complexity.

"*All models are wrong, but some are useful*" (Box, 1976)

Why IQ-TREE?

Next generation sequencing data represent both a blessing and a curse:

- Blessing: (Phylo)genomic data help to elucidate many phylogenetic questions.
- Curse: Many model assumptions become increasingly distant from the truth due to growing data complexity.

"*All models are wrong, but some are useful*" (Box, 1976)

With IQ-TREE we aim to:

Why IQ-TREE?

Next generation sequencing data represent both a blessing and a curse:

- Blessing: (Phylo)genomic data help to elucidate many phylogenetic questions.
- Curse: Many model assumptions become increasingly distant from the truth due to growing data complexity.

"*All models are wrong, but some are useful*" (Box, 1976)

With IQ-TREE we aim to:

- Analyze ultra-large data sets.

Why IQ-TREE?

Next generation sequencing data represent both a blessing and a curse:

- Blessing: (Phylo)genomic data help to elucidate many phylogenetic questions.
- Curse: Many model assumptions become increasingly distant from the truth due to growing data complexity.

"*All models are wrong, but some are useful*" (Box, 1976)

With IQ-TREE we aim to:

- Analyze ultra-large data sets.
- Provide many (if not most) "useful" models of sequence evolution.

Why IQ-TREE?

Next generation sequencing data represent both a blessing and a curse:

- Blessing: (Phylo)genomic data help to elucidate many phylogenetic questions.
- Curse: Many model assumptions become increasingly distant from the truth due to growing data complexity.

"*All models are wrong, but some are useful*" (Box, 1976)

With IQ-TREE we aim to:

- Analyze ultra-large data sets.
- Provide many (if not most) "useful" models of sequence evolution.

But still, there are RAxML, PhyML out there, why do I need IQ-TREE?

Why IQ-TREE?

Next generation sequencing data represent both a blessing and a curse:

- Blessing: (Phylo)genomic data help to elucidate many phylogenetic questions.
- Curse: Many model assumptions become increasingly distant from the truth due to growing data complexity.

"*All models are wrong, but some are useful*" (Box, 1976)

With IQ-TREE we aim to:

- Analyze ultra-large data sets.
- Provide many (if not most) "useful" models of sequence evolution.

But still, there are RAxML, PhyML out there, why do I need IQ-TREE?

- We better have at least 2 software independently developed for similar purpose. Only then, the pros and cons (sometimes **bugs**) can be identified. This creates a *friendly* competition, which helps to advance the field!

Why IQ-TREE?

Next generation sequencing data represent both a blessing and a curse:

- Blessing: (Phylo)genomic data help to elucidate many phylogenetic questions.
- Curse: Many model assumptions become increasingly distant from the truth due to growing data complexity.

"*All models are wrong, but some are useful*" (Box, 1976)

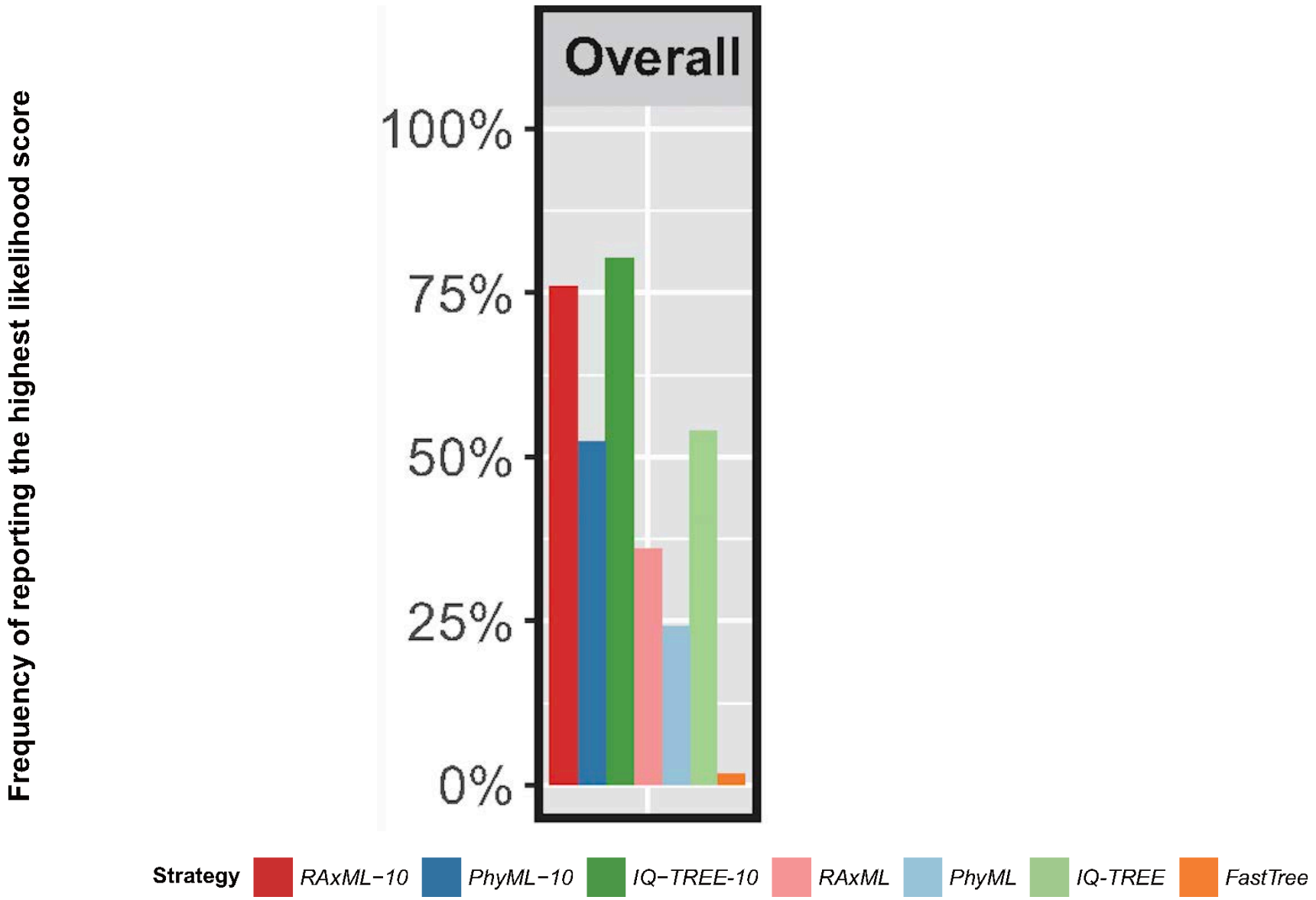
With IQ-TREE we aim to:

- Analyze ultra-large data sets.
- Provide many (if not most) "useful" models of sequence evolution.

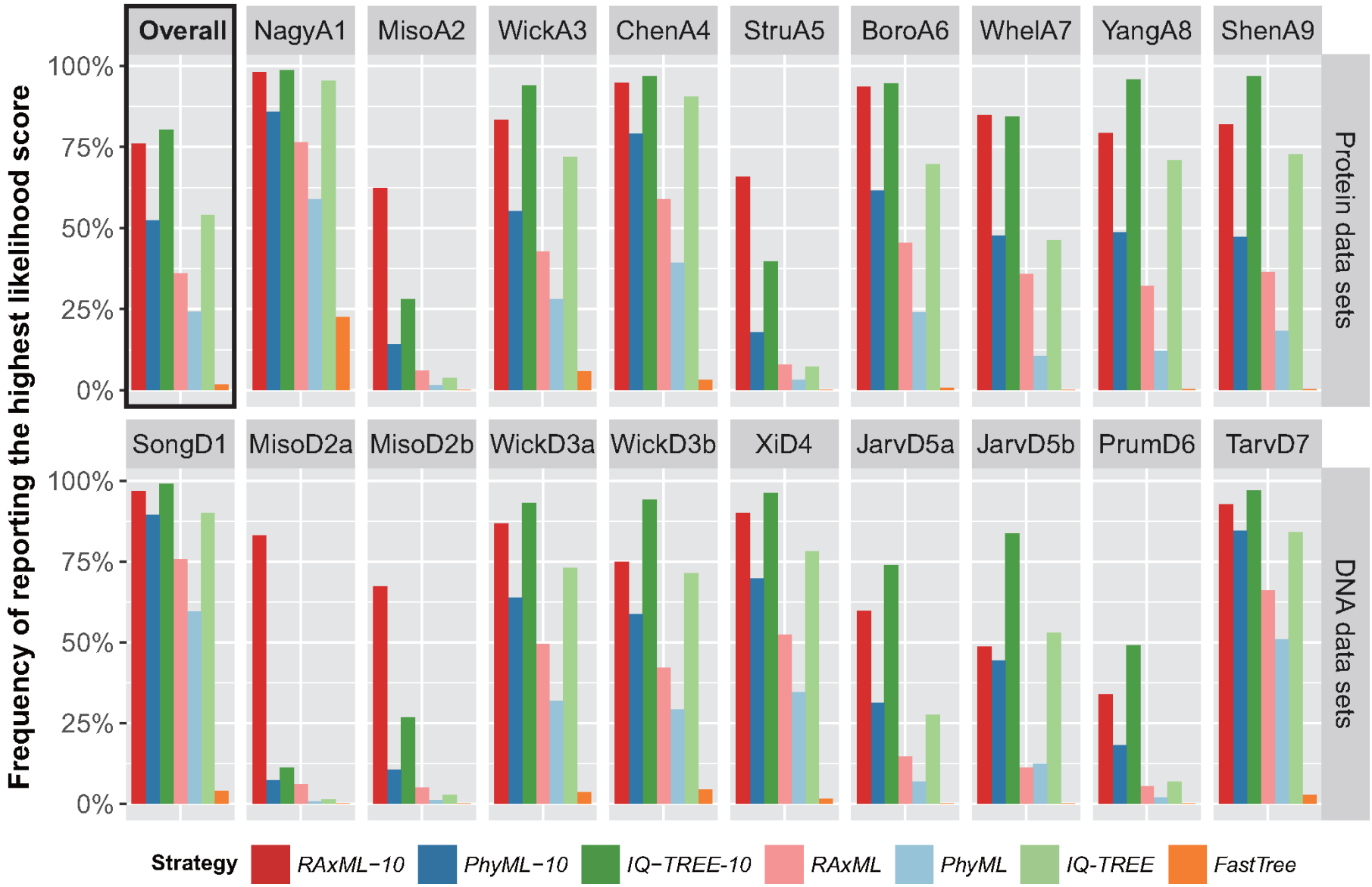
But still, there are RAxML, PhyML out there, why do I need IQ-TREE?

- We better have at least 2 software independently developed for similar purpose. Only then, the pros and cons (sometimes **bugs**) can be identified. This creates a *friendly* competition, which helps to advance the field!
- Same as having MrBayes, RevBayes, BEAST for Bayesian inference.

An independent benchmark by Zhou et al. (2018)



An independent benchmark by Zhou et al. (2018)



Tree search algorithms in RAxML and IQ-TREE

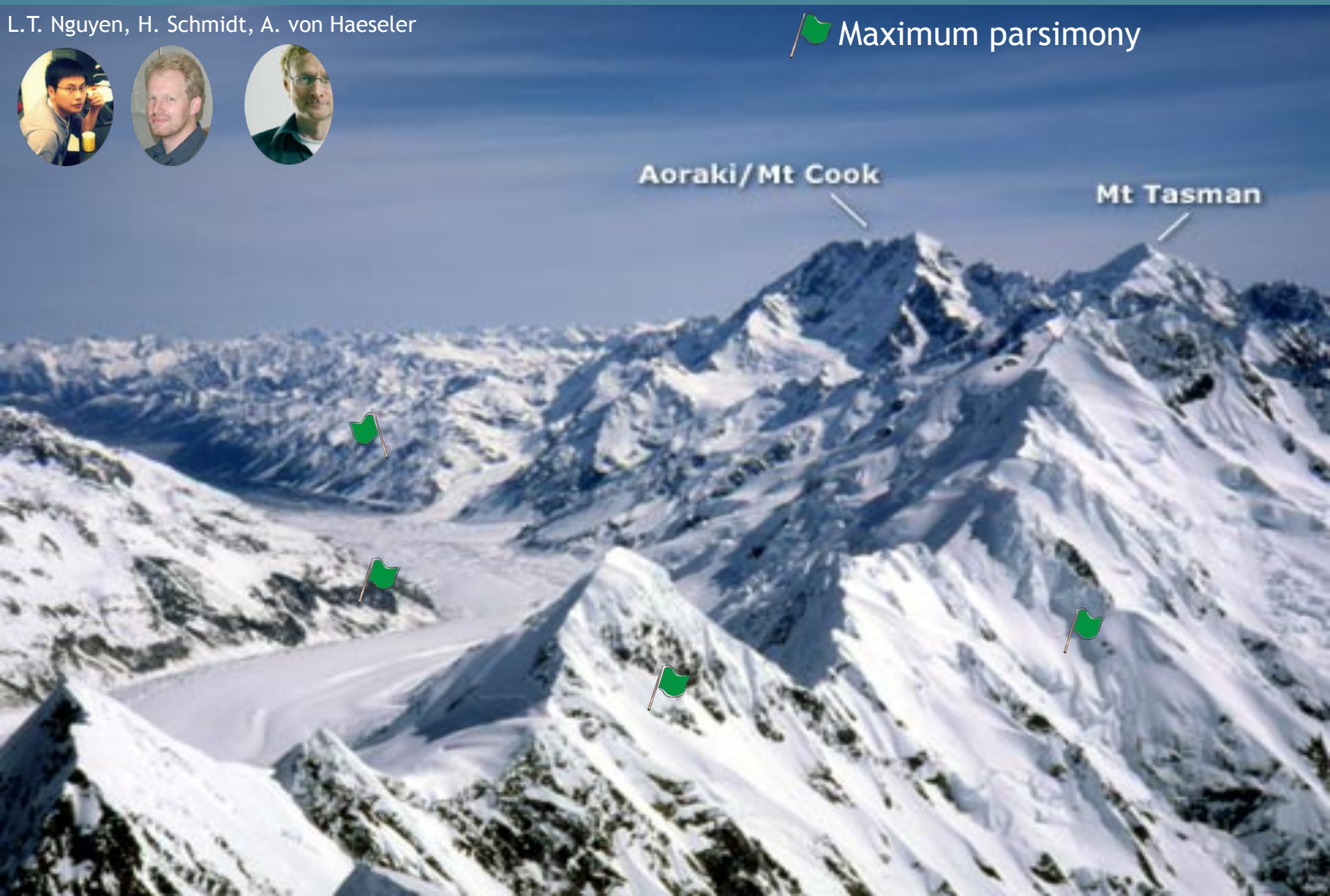
Feature	RAxML	IQ-TREE
Starting tree	Parsimony: Stepwise addition + subtree pruning and regrafting (SPR)	99 parsimony trees (like RAxML) and 1 Neighbor-joining tree
Tree search heuristics	Hill-climbing SPR	Stochastic: Hill-climbing Nearest Neighbor Interchange (NNI) and downhill NNI

IQ-TREE: A stochastic tree search algorithm

L.T. Nguyen, H. Schmidt, A. von Haeseler

IQ-TREE: A stochastic tree search algorithm

L.T. Nguyen, H. Schmidt, A. von Haeseler

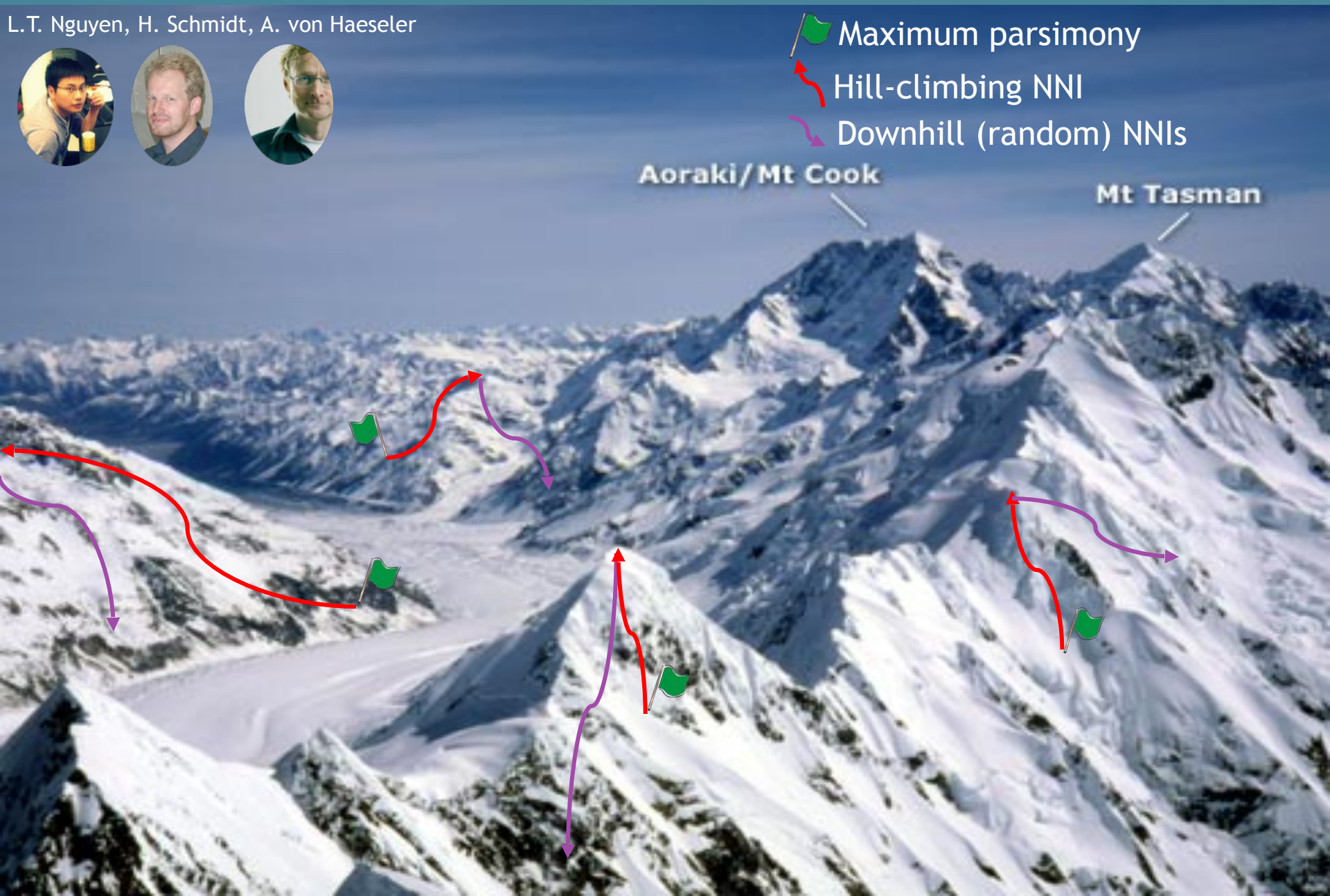


IQ-TREE: A stochastic tree search algorithm

L.T. Nguyen, H. Schmidt, A. von Haeseler

IQ-TREE: A stochastic tree search algorithm

L.T. Nguyen, H. Schmidt, A. von Haeseler



IQ-TREE: A stochastic tree search algorithm

L.T. Nguyen, H. Schmidt, A. von Haeseler

IQ-TREE: A stochastic tree search algorithm

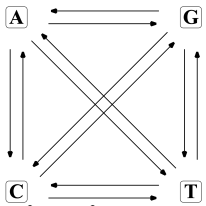
L.T. Nguyen, H. Schmidt, A. von Haeseler

Typical phylogenetic analysis

Sequence alignment

```
CACCTGTCGT  ----------  ----------  TCTGGTGCAG
CAGCTGTCGT  GCTCTTTCTG  TTGAGCCTGG  TCTGGTGCAG
CAGCTGCCGT  GTTTTCTCTG  TTGAGCCTGG  TCTGGTACAG
CAGCTGCCGC  GTTCTCTCCG  ----------  TCTGGTGCAA
CTCCTGCCGG  GTGCTCTCAG  ----------  ----------
CTCCTGCCGG  ----------  CTGAGCCGGG  TCTGGTGCAG
CTCTTGCCGG  ----------  CTGAGCCTTG  ----------
```

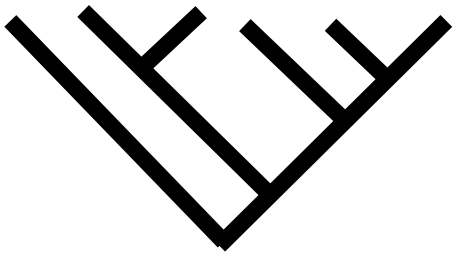
Model selection



Substitution models

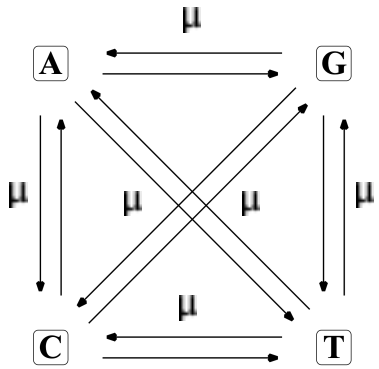
Tree search

Bootstrap



Phylogeny

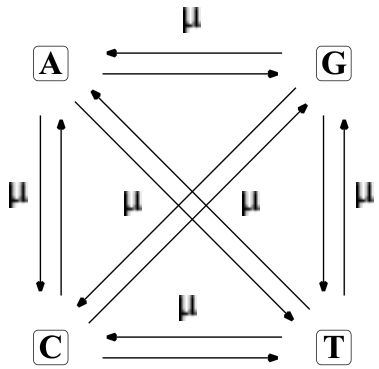
Models of sequence evolution



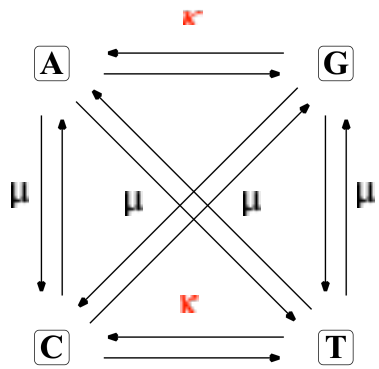
JC
(Jukes & Cantor 1969)

Rate model	Explanation
+I	Some sites are *invariable* (zero rate), e.g. due to selective force.
+G	Site rates follow a *Gamma* distribution.
+I+G	Some sites are invariable, the rest follow a Gamma distribution.
+R	Sites fall into several categories from slow to fast rates. No assumption of rate distribution (free-rate model).

Models of sequence evolution



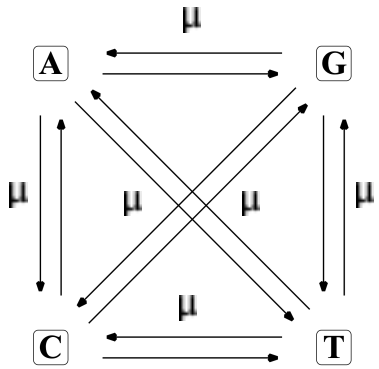
JC
(Jukes & Cantor 1969)



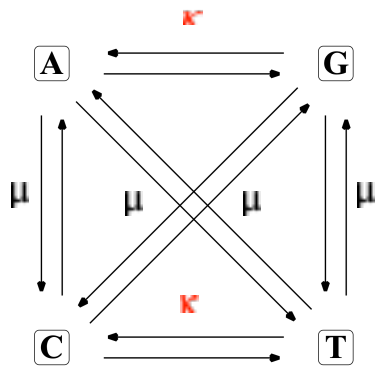
HKY
(Hasegawa, Kishino, Yano 1985)

Rate model	Explanation
+I	Some sites are *invariable* (zero rate), e.g. due to selective force.
+G	Site rates follow a *Gamma* distribution.
+I+G	Some sites are invariable, the rest follow a Gamma distribution.
+R	Sites fall into several categories from slow to fast rates. No assumption of rate distribution (free-rate model).

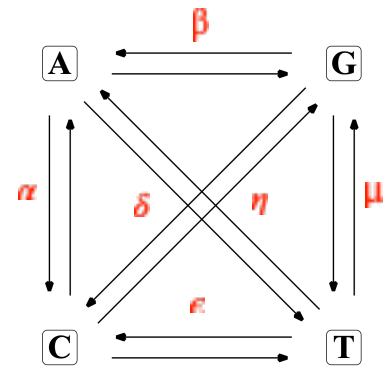
Models of sequence evolution



JC
(Jukes & Cantor 1969)



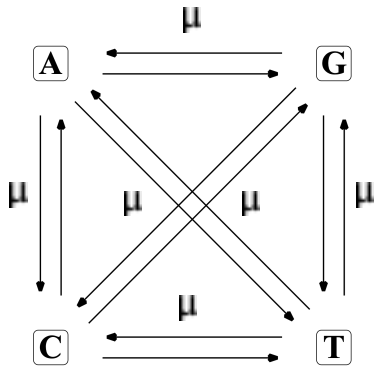
HKY
(Hasegawa, Kishino, Yano 1985)



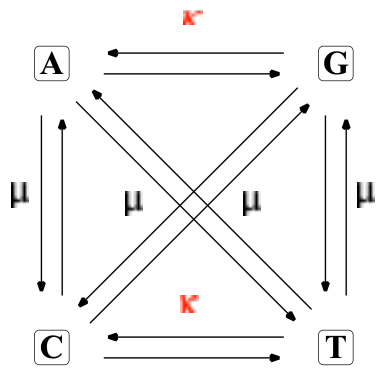
GTR
(General Time Reversible, 1986)

Rate model	Explanation
+I	Some sites are *invariable* (zero rate), e.g. due to selective force.
+G	Site rates follow a *Gamma* distribution.
+I+G	Some sites are invariable, the rest follow a Gamma distribution.
+R	Sites fall into several categories from slow to fast rates. No assumption of rate distribution (free-rate model).

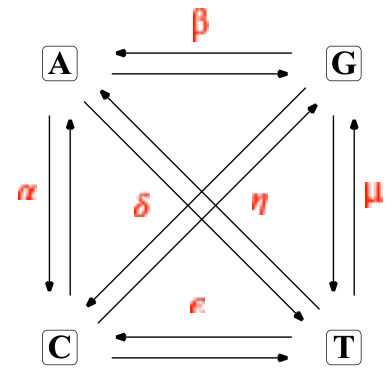
Models of sequence evolution



JC
(Jukes & Cantor 1969)



HKY
(Hasegawa, Kishino, Yano 1985)

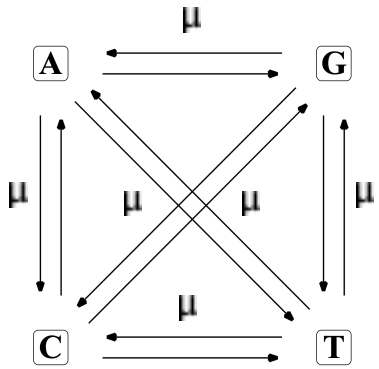


GTR
(General Time Reversible, 1986)

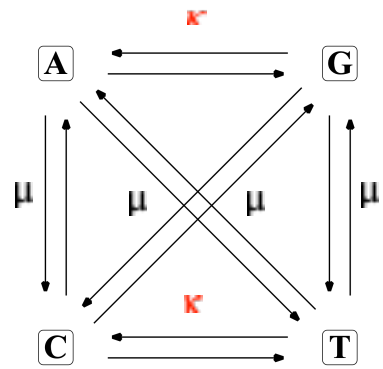
Rate heterogeneity: alignment sites evolved at different rates. Some slow, some fast.

Rate model	Explanation
+I	Some sites are *invariable* (zero rate), e.g. due to selective force.
+G	Site rates follow a *Gamma* distribution.
+I+G	Some sites are invariable, the rest follow a Gamma distribution.
+R	Sites fall into several categories from slow to fast rates. No assumption of rate distribution (free-rate model).

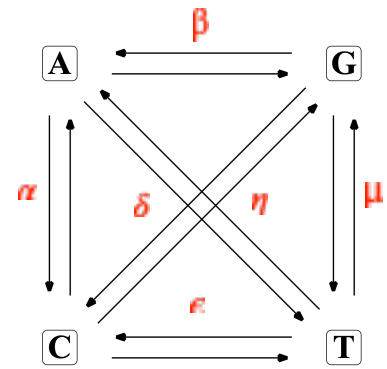
Models of sequence evolution



JC
(Jukes & Cantor 1969)



HKY
(Hasegawa, Kishino, Yano 1985)



GTR
(General Time Reversible, 1986)

Rate heterogeneity: alignment sites evolved at different rates. Some slow, some fast.

Rate model	Explanation
+I	Some sites are *invariable* (zero rate), e.g. due to selective force.
+G	Site rates follow a *Gamma* distribution.
+I+G	Some sites are invariable, the rest follow a Gamma distribution.
+R	Sites fall into several categories from slow to fast rates. No assumption of rate distribution (free-rate model).

A model = substitution model + rate heterogeneity, e.g. “GTR+G”

Model selection

JC		GTR
JC+G		GTR+G
JC+I		GTR+I
JC+I+G	.....	GTR+I+G
JC+R2		GTR+R2
…		…
JC+R10		GTR+R10

Which model is best?

Model selection

JC		GTR
JC+G		GTR+G
JC+I		GTR+I
JC+I+G	.....	GTR+I+G
JC+R2		GTR+R2
…		…
JC+R10		GTR+R10

Which model is best?

Problem:
More complex models always have higher *likelihood* than simpler models!

Model selection

JC		GTR
JC+G		GTR+G
JC+I		GTR+I
JC+I+G	.....	GTR+I+G
JC+R2		GTR+R2
…		…
JC+R10		GTR+R10

Which model is best?

Problem:
More complex models always have higher *likelihood* than simpler models!

Solution: Penalize a model *M* by the number of its parameters (*k*)

1. Akaike information criterion (AIC):

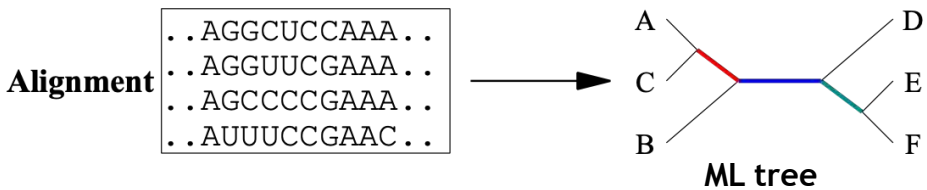
2. Bayesian information criterion (BIC):

where *n* is the number of alignment sites.

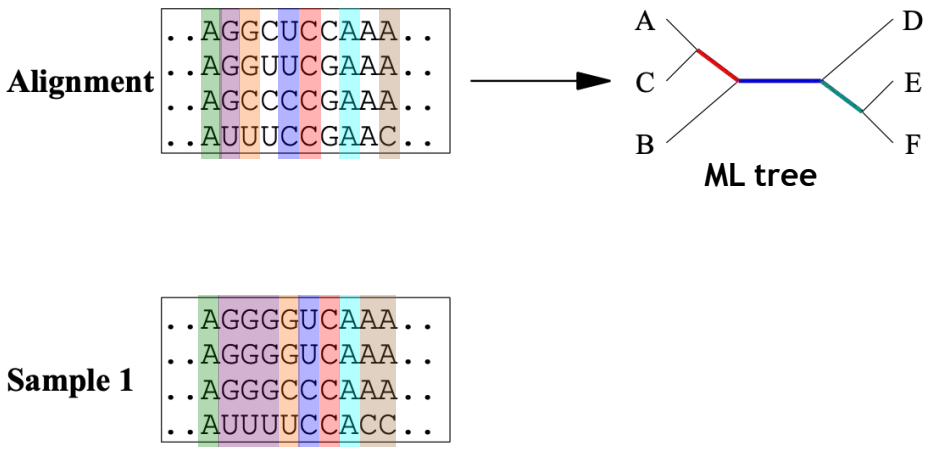
Select the model with **smallest AIC or BIC score**.

The default in IQ-TREE is BIC, but you should state that in the publication!

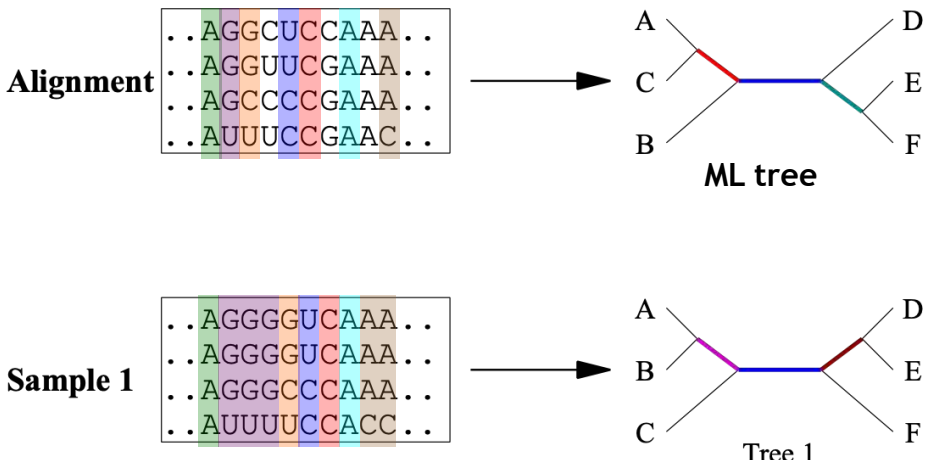
Bootstrap: How reliable are branches of the tree?



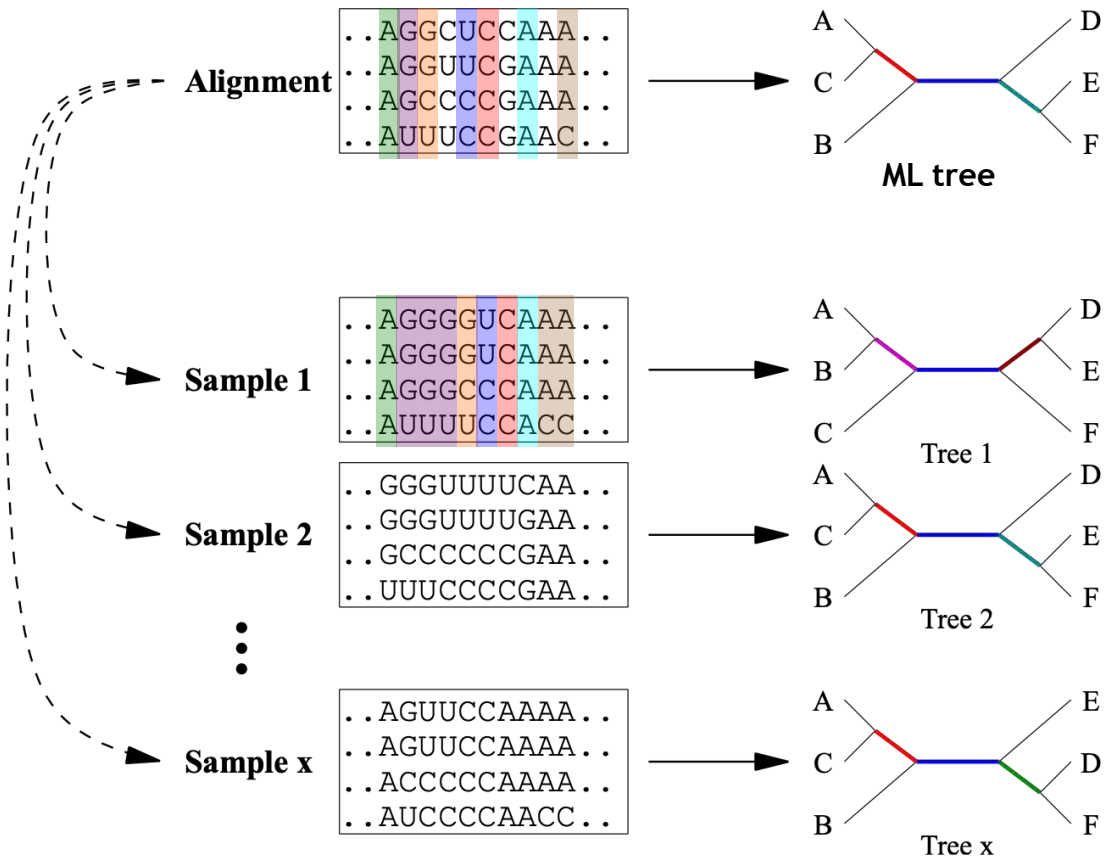
Bootstrap: How reliable are branches of the tree?



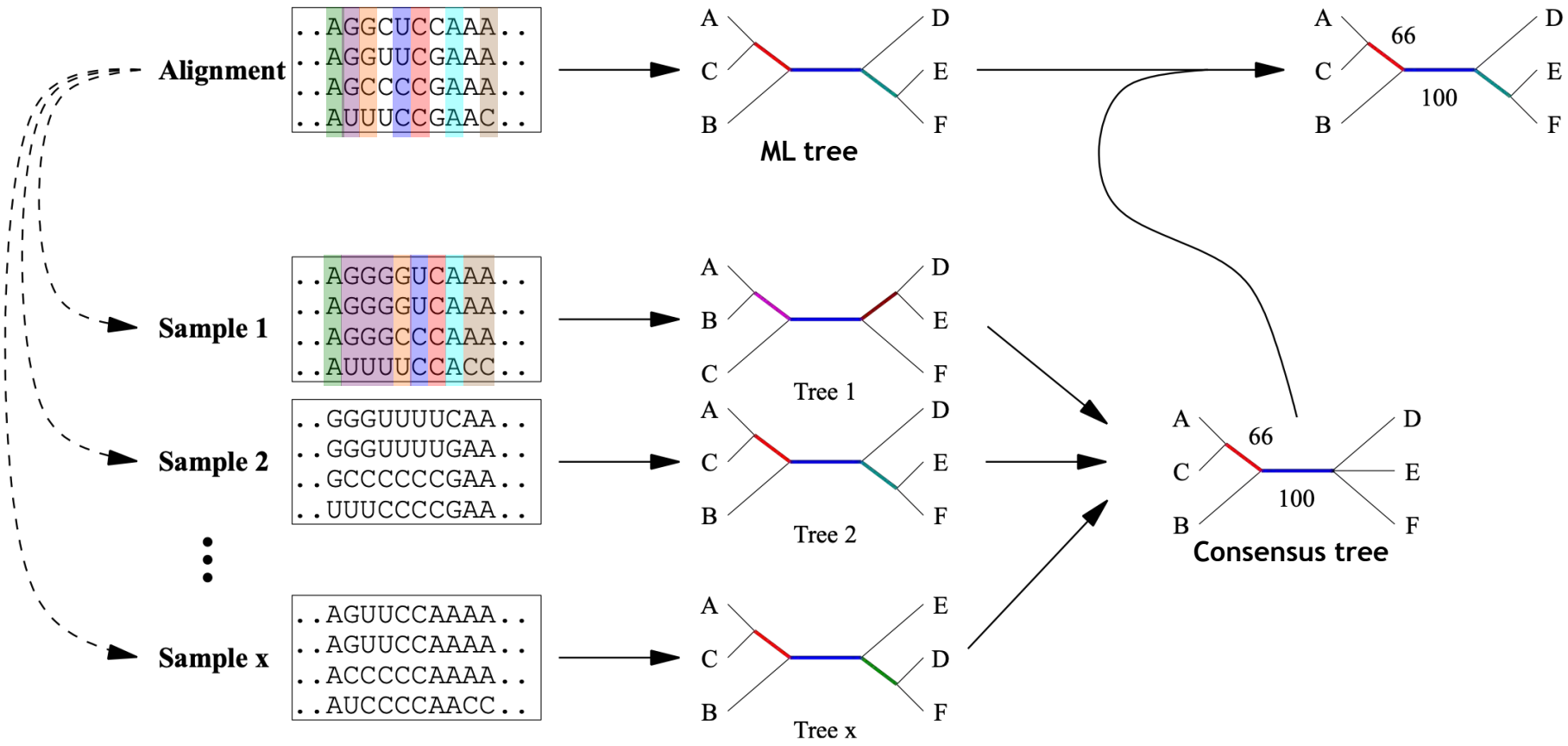
Bootstrap: How reliable are branches of the tree?



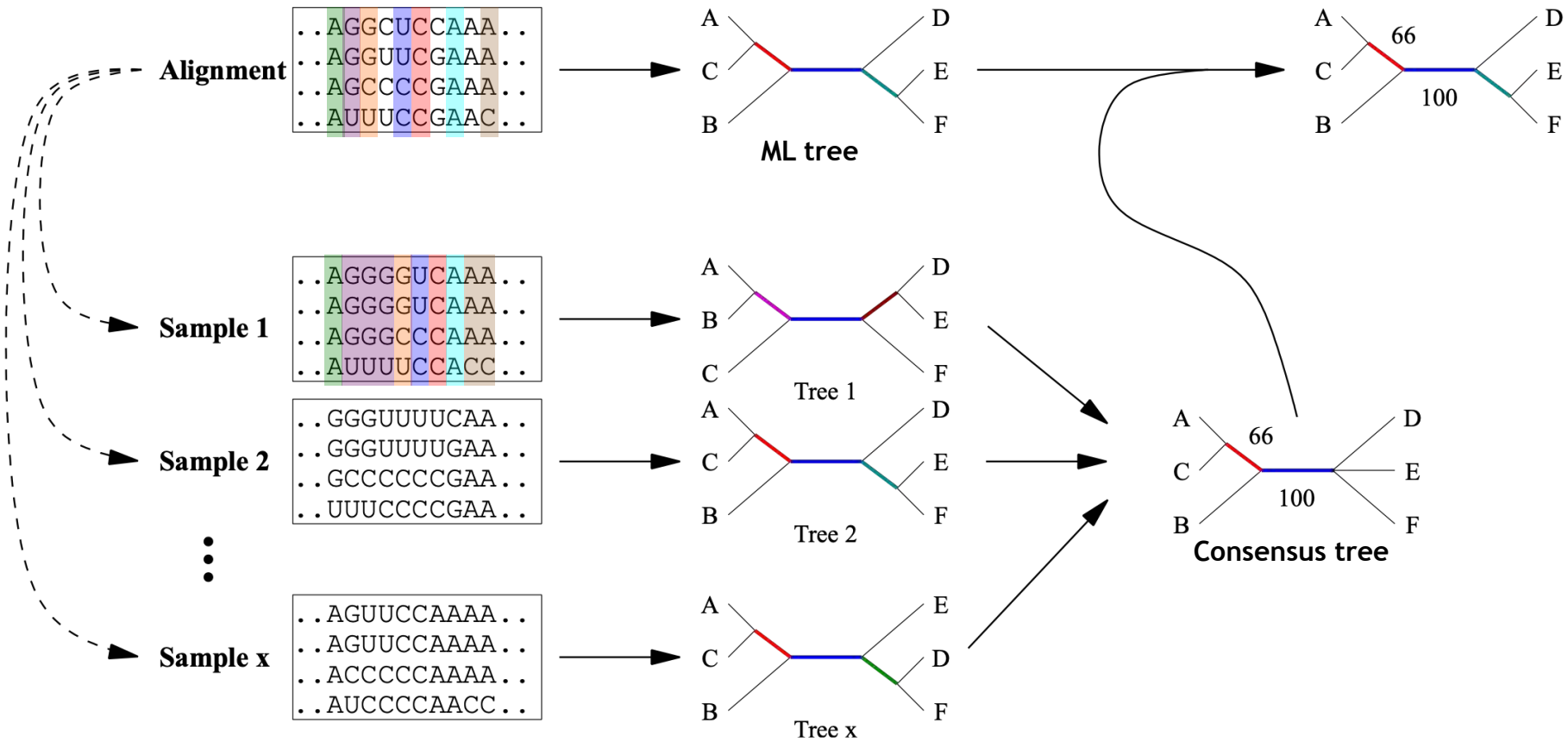
Bootstrap: How reliable are branches of the tree?



Bootstrap: How reliable are branches of the tree?

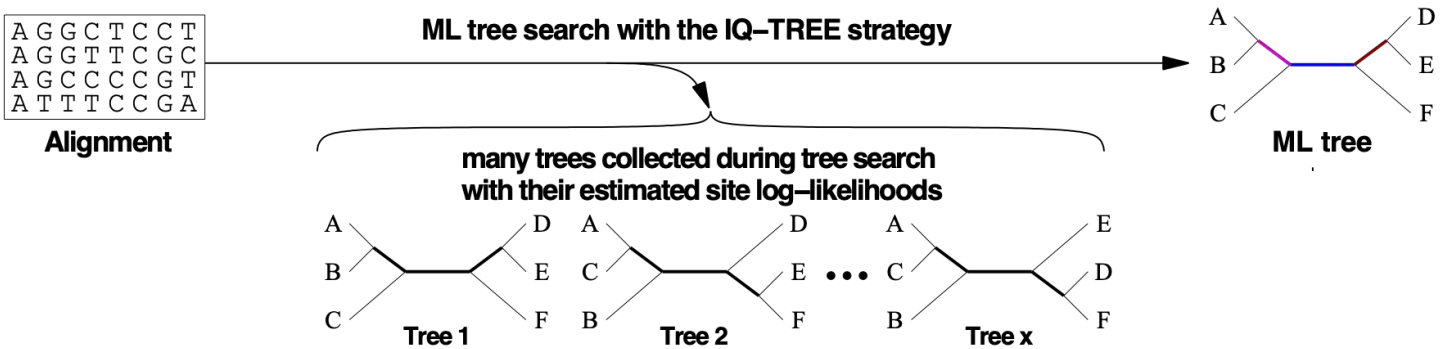


Bootstrap: How reliable are branches of the tree?



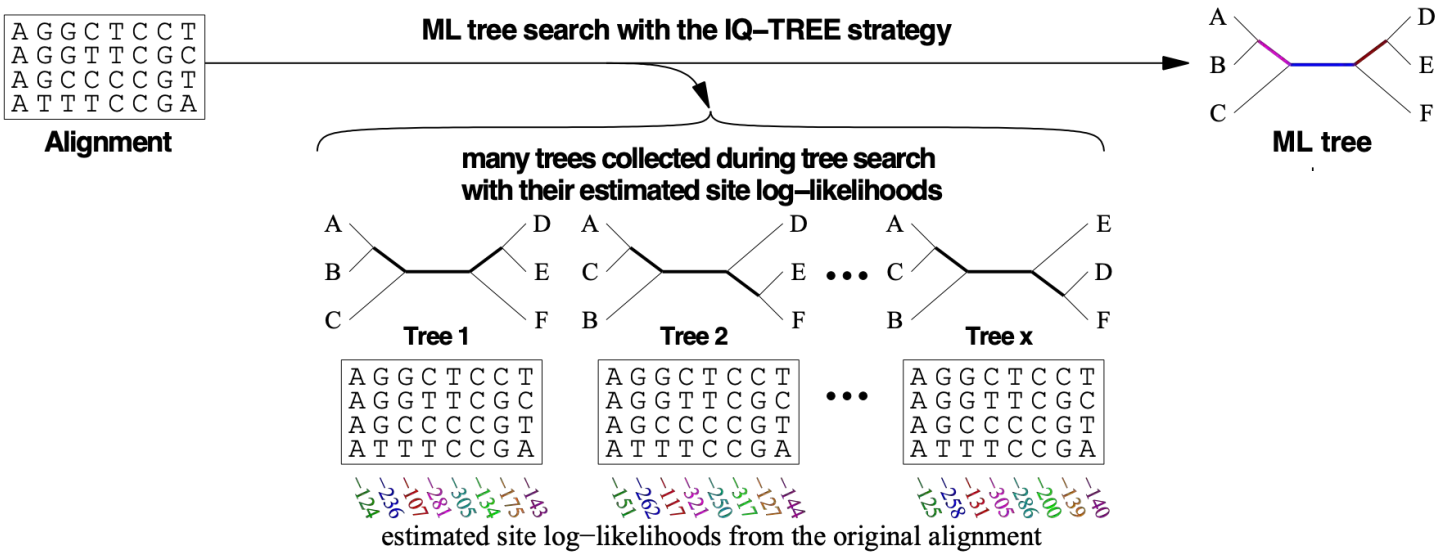
Bootstrap analysis is extremely time-consuming!

UFBoot: Ultrafast bootstrap approximation




M.A.T. Nguyen, A. von Haesel

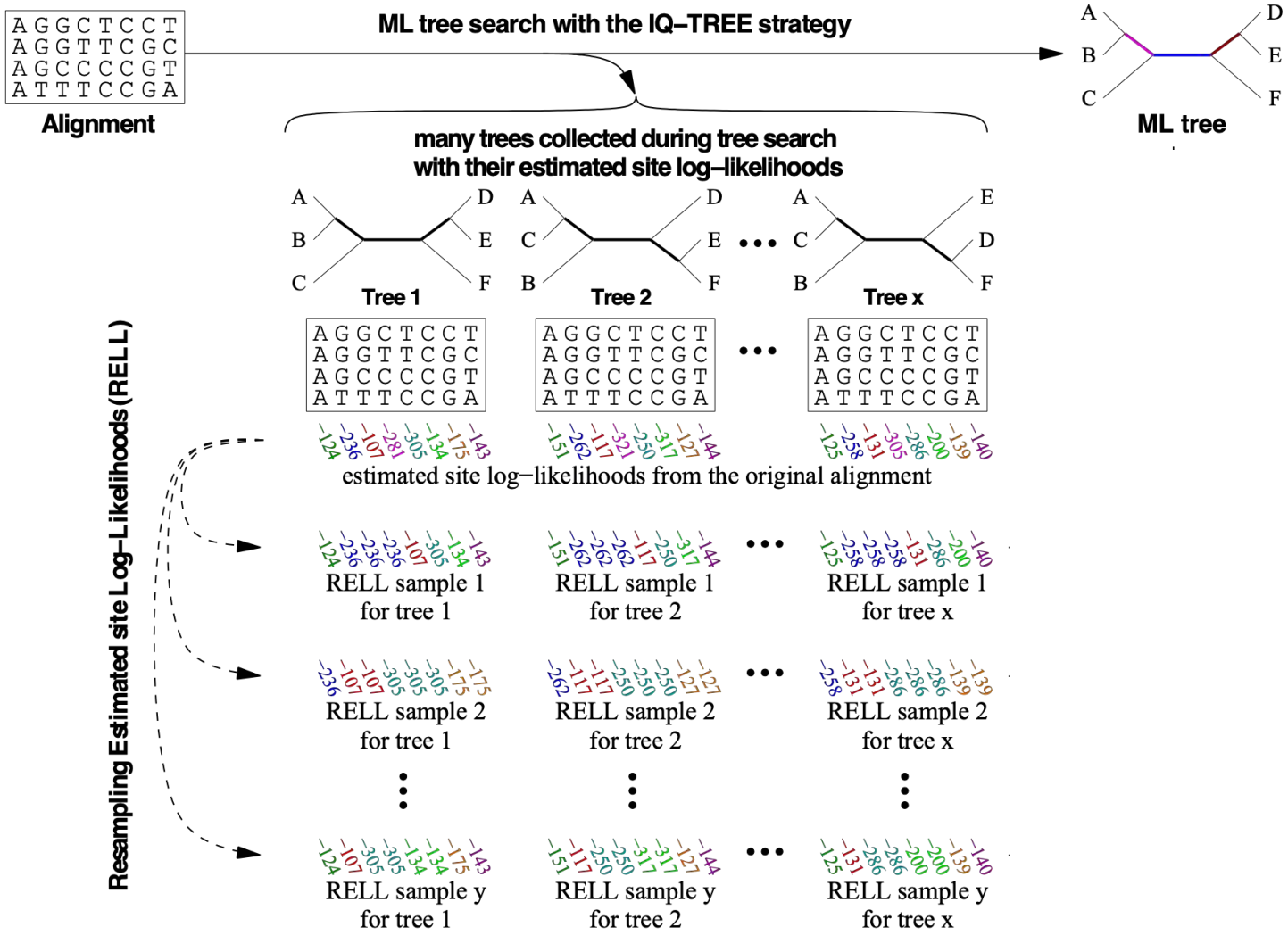
UFBoot: Ultrafast bootstrap approximation



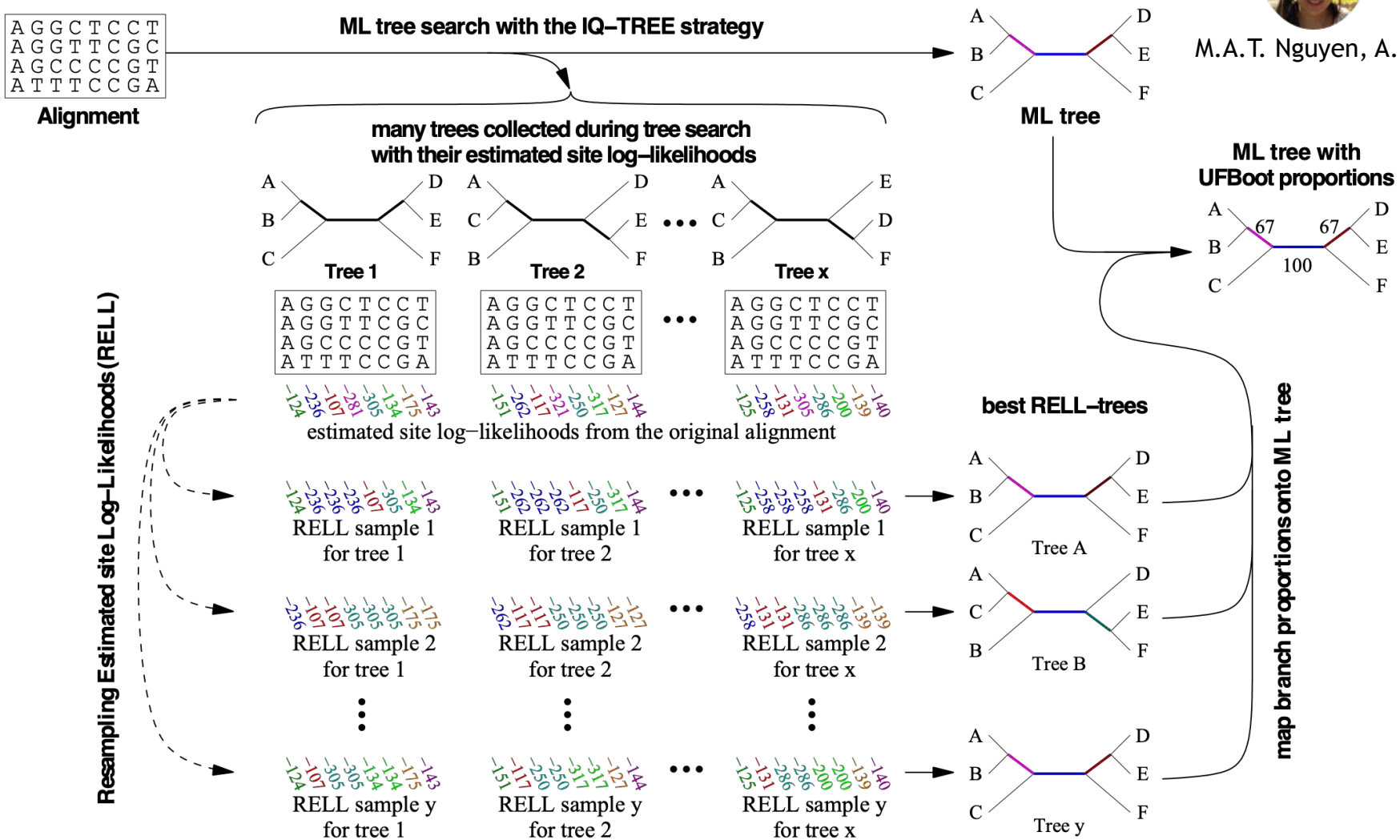

M.A.T. Nguyen, A. von Haesel

UFBoot: Ultrafast bootstrap approximation


M.A.T. Nguyen, A. von Haesel



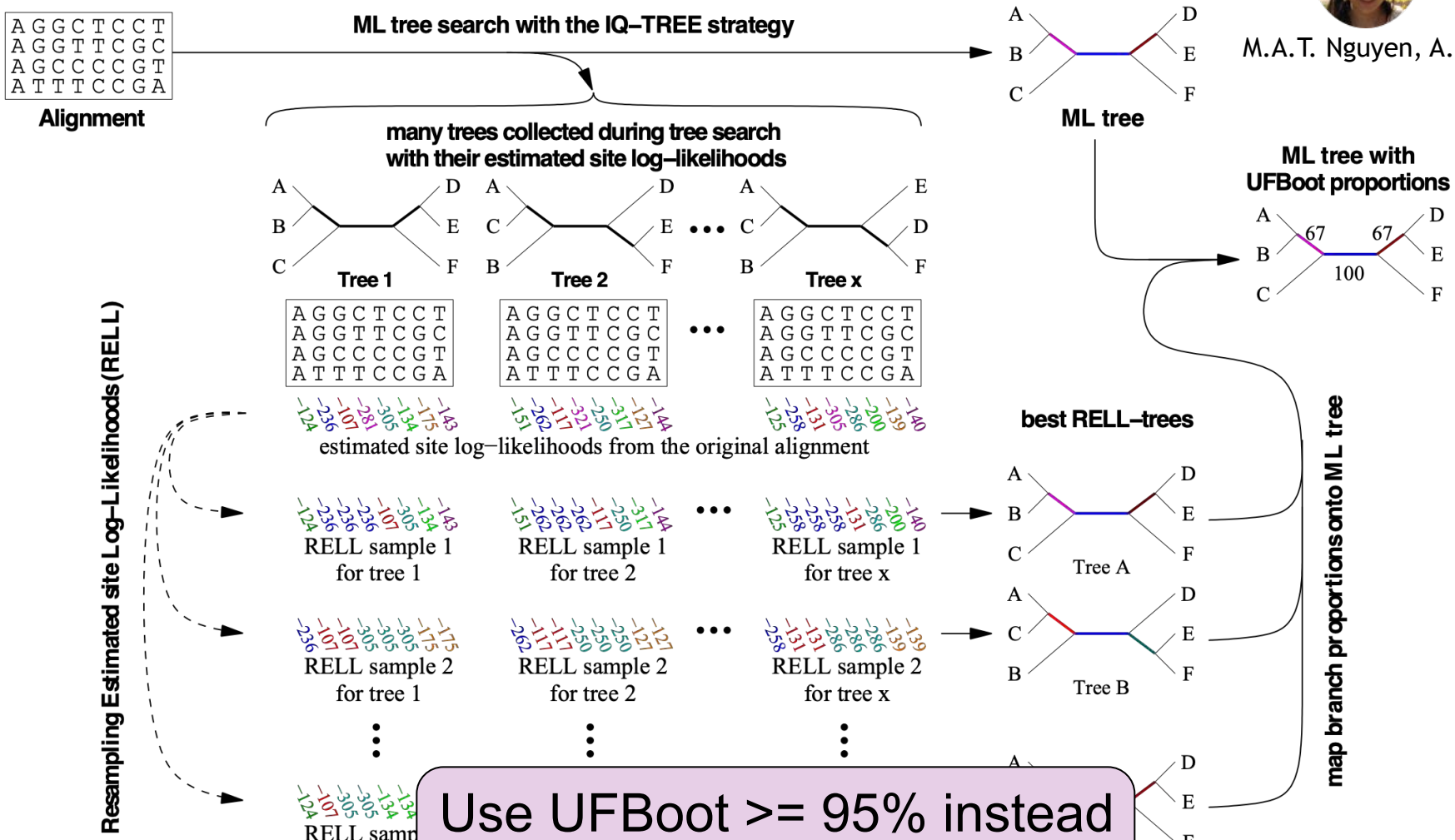
UFBoot: Ultrafast bootstrap approximation




M.A.T. Nguyen, A. von Haesel

UFBoot: Ultrafast bootstrap approximation


M.A.T. Nguyen, A. von Haesel



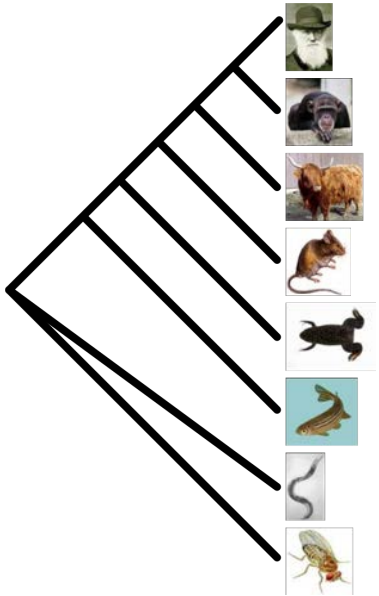
Use UFBoot >= 95% instead of 70% !

Genome-scale data: Concatenation methods

Supermatrix

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Phylogenomic Inference



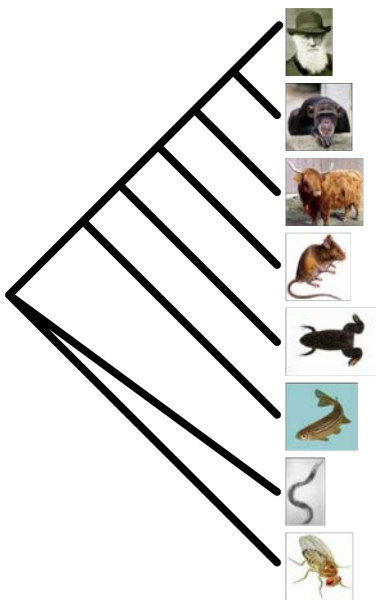
Species tree of life

Genome-scale data: Concatenation methods

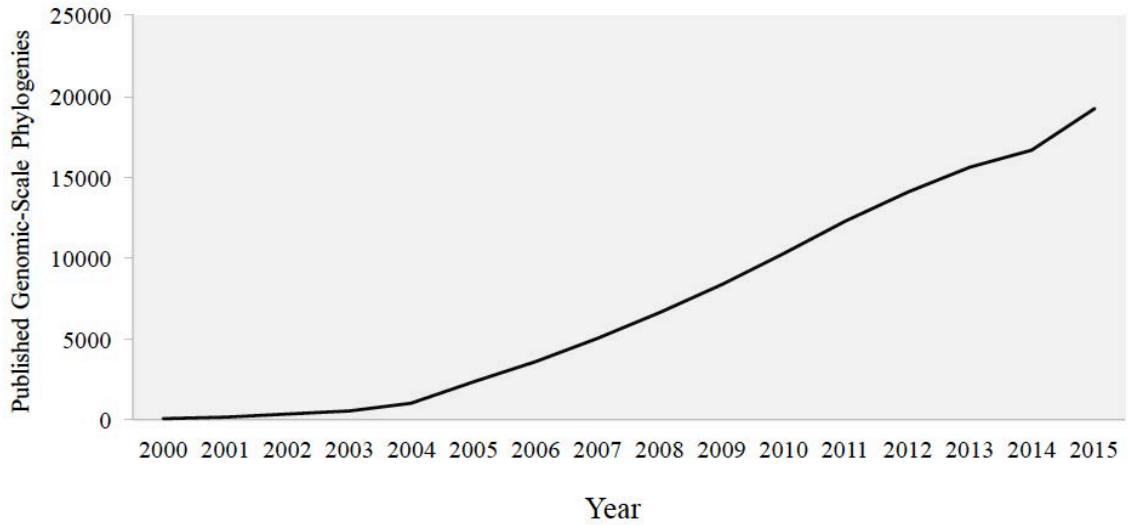
Supermatrix

Gene 1	Gene 2	……	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Phylogenomic Inference



Species tree of life

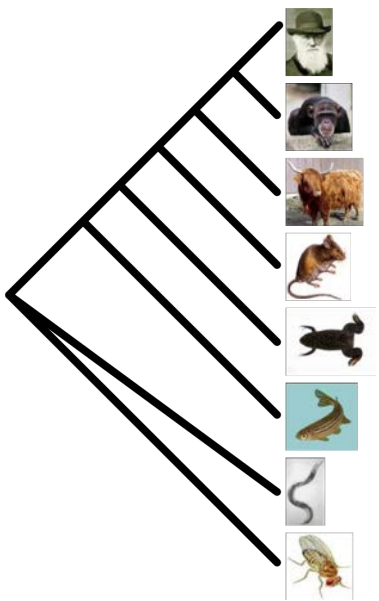


Genome-scale data: Concatenation methods

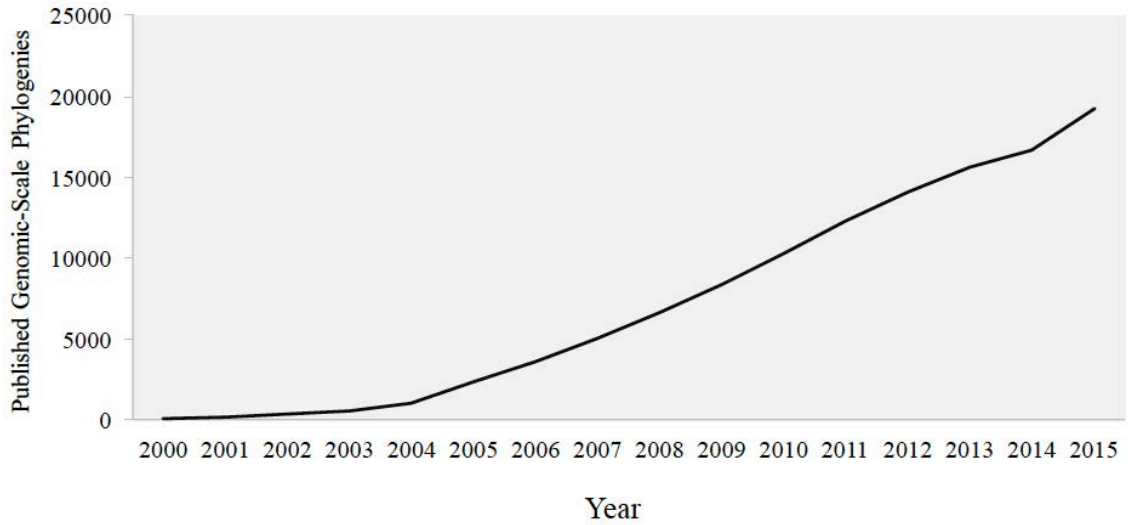
Supermatrix

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Phylogenomic Inference

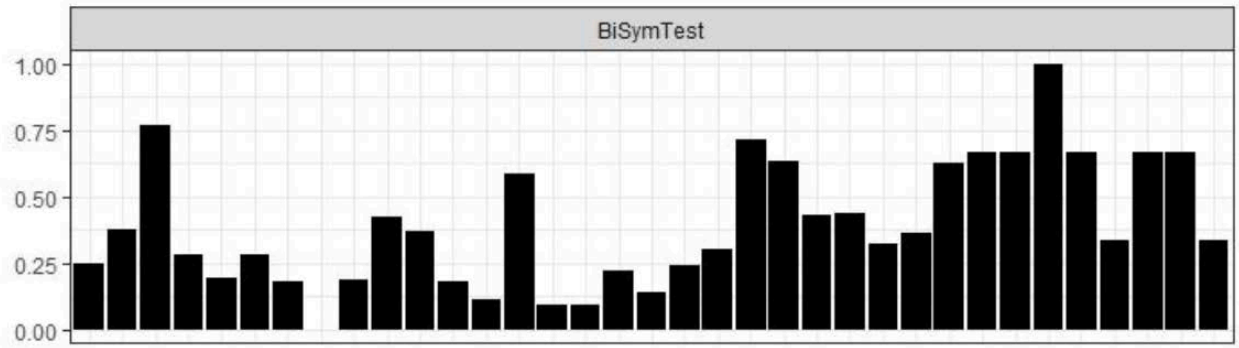


Species tree of life



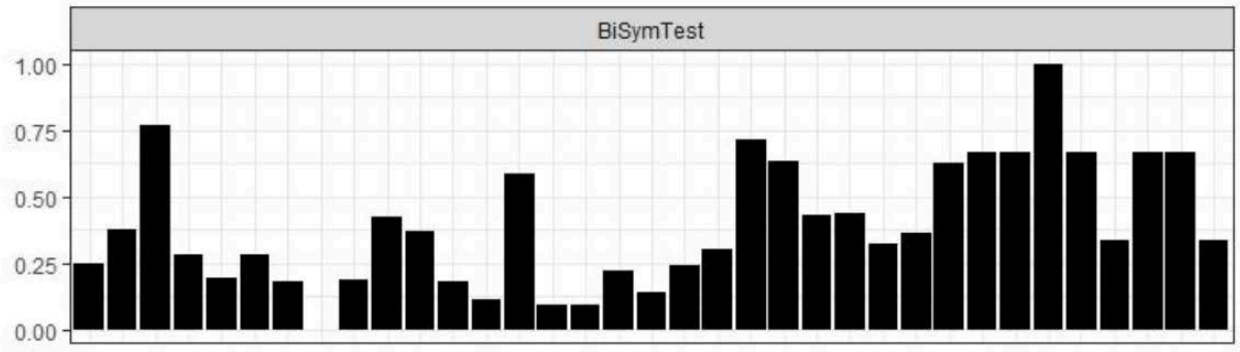
30 days of computation and 280 GB RAM for an insect data set!

“Data-model gap” is increasing!



Level of model violations in 35 phylogenomic datasets (https://doi.org/10.1101/460121)

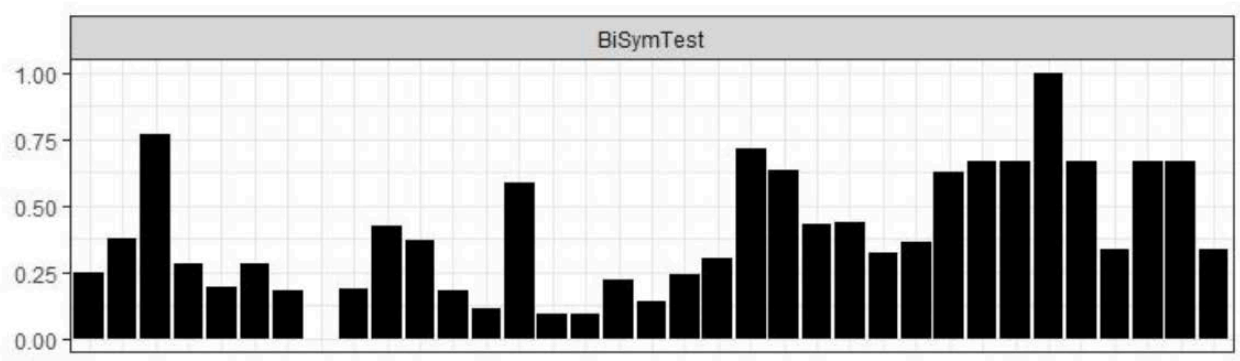
“Data-model gap” is increasing!



Level of model violations in 35 phylogenomic datasets (https://doi.org/10.1101/460121)

1. Resulting trees tend to be biased towards the genes that violated model assumptions.
2. Bootstrap supports tend to 100% as #genes increases.

“Data-model gap” is increasing!

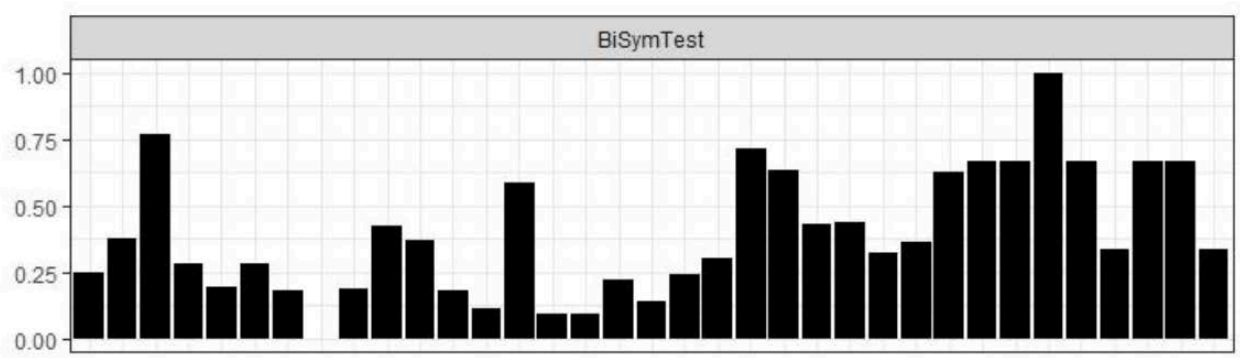


Level of model violations in 35 phylogenomic datasets (https://doi.org/10.1101/460121)

1. Resulting trees tend to be biased towards the genes that violated model assumptions.
2. Bootstrap supports tend to 100% as #genes increases.

Model violation → Systematic bias

“Data-model gap” is increasing!



Level of model violations in 35 phylogenomic datasets (https://doi.org/10.1101/460121)

1. Resulting trees tend to be biased towards the genes that violated model assumptions.
2. Bootstrap supports tend to 100% as #genes increases.

Model violation → Systematic bias

1. Remove “bad” loci
2. Use more realistic models

Partition model

Supermatrix

Gene 1	Gene 2	……	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

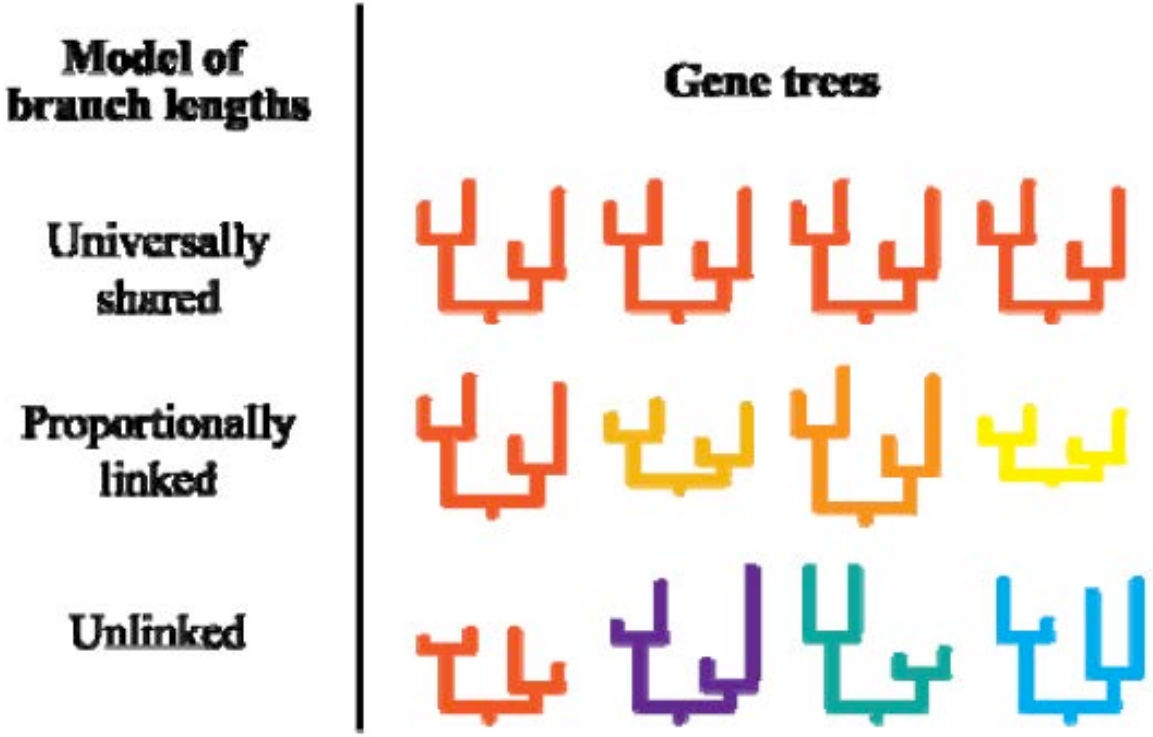
Substitution models: JC HKY+G …… GTR+G

Partition model

Supermatrix

Gene 1	Gene 2	……	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Substitution models: JC HKY+G …… GTR+G

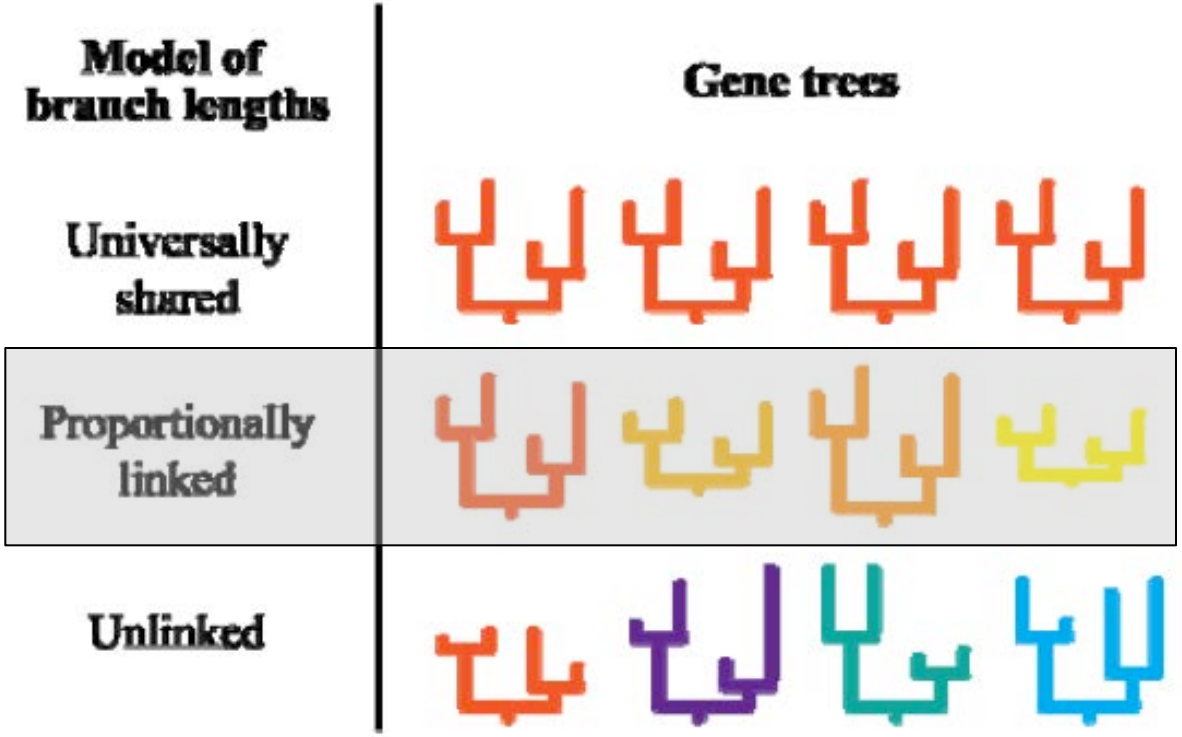


Partition model

Supermatrix

Gene 1	Gene 2	……	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Substitution models: JC HKY+G …… GTR+G



Recommended for typical analysis, confirmed by Dunchene et al. (2018) https://doi.org/10.1101/467449

How to reduce potential model overfitting?

Supermatrix

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

How to reduce potential model overfitting?

Supermatrix

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

PartitionFinder

Gene 1+2	.....	Gene 200+1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

How to reduce potential model overfitting?

Supermatrix

Gene 1	Gene 2	……	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

PartitionFinder

Gene 1+2	……	Gene 200+1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Substitution models: HKY …… GTR+G

How to reduce potential model overfitting?

Supermatrix

Gene 1	Gene 2	……	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

PartitionFinder

Gene 1+2	……	Gene 200+1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

PartitionFinder algorithm
(Lanfear et al. 2012):

1. Evaluate to merge all pairs of genes.
2. Choose the pair with the best score.
3. If score improves, merge two genes and repeat steps 1-3.
4. Otherwise, stop.

Substitution models: HKY …… GTR+G

How to reduce potential model overfitting?

Supermatrix

Gene 1	Gene 2	……	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

PartitionFinder

Gene 1+2	……	Gene 200+1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Substitution models: HKY …… GTR+G

PartitionFinder algorithm (Lanfear et al. 2012):

1. Evaluate to merge all pairs of genes.
2. Choose the pair with the best score.
3. If score improves, merge two genes and repeat steps 1-3.
4. Otherwise, stop.

Relaxed clustering algorithm (Lanfear et al. 2014):

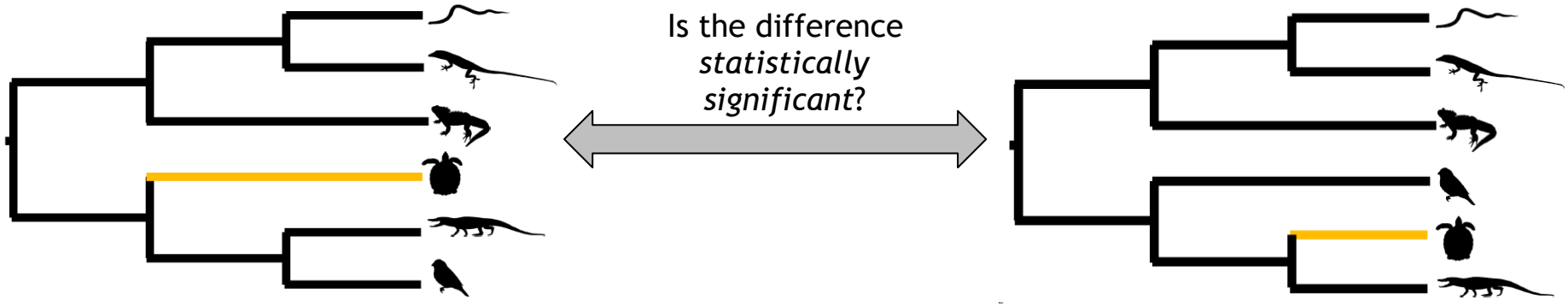
In step 1: only examine the top k% of most "promising" pairs.

How to bootstrap phylogenomic alignments?

Supermatrix

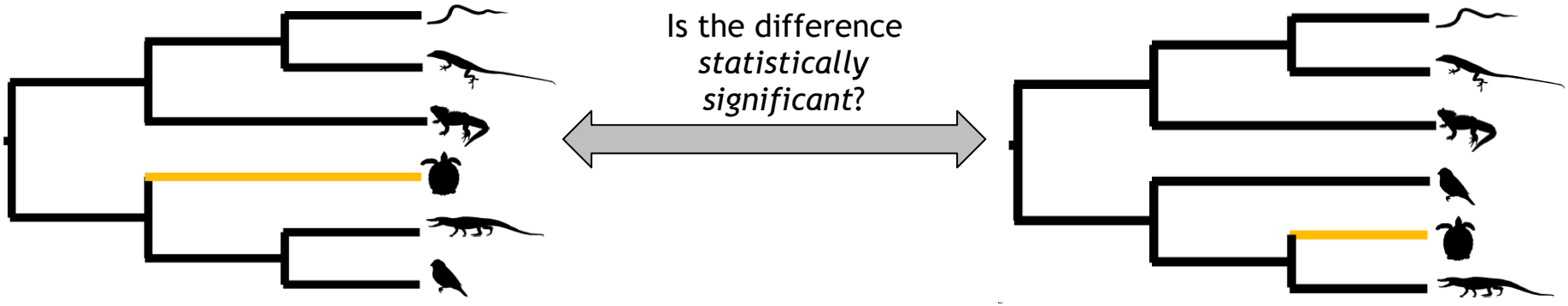
Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Tree topology tests



$$\delta = \log\Big(likelihood(T_1)\Big) - \log\Big(likelihood(T_0)\Big)$$

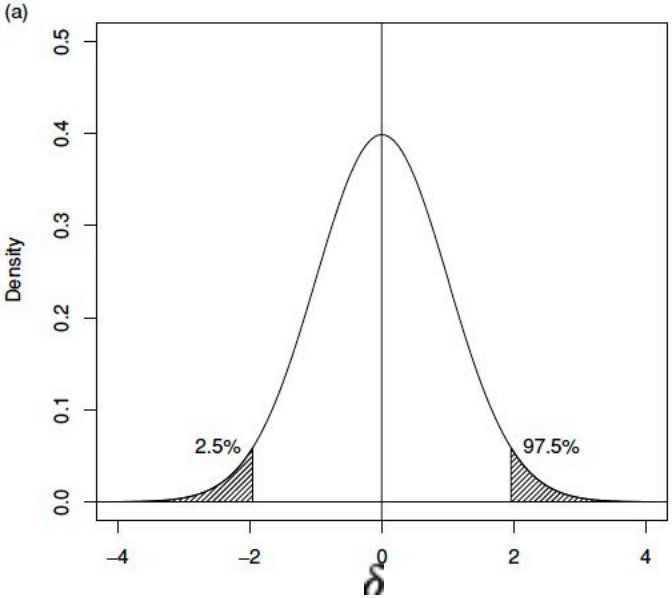
Tree topology tests



δ**Testing two trees** (Kishino & Hasegawa, 1989):

1. Statistic: .
2. Generate distribution of from many “random” data (e.g. by 1000 bootstrap resampling).
3. Compare the statistic between original and random data to obtain *p-value*.
4. If p-value < 0.05: YES! two trees are significantly different.

- If p-value >= 0.05: NO! they are not.

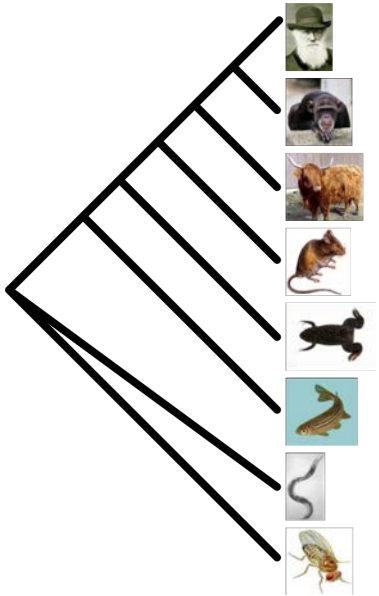


Concatenation methods: Limitation

Supermatrix

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Phylogenomic Inference



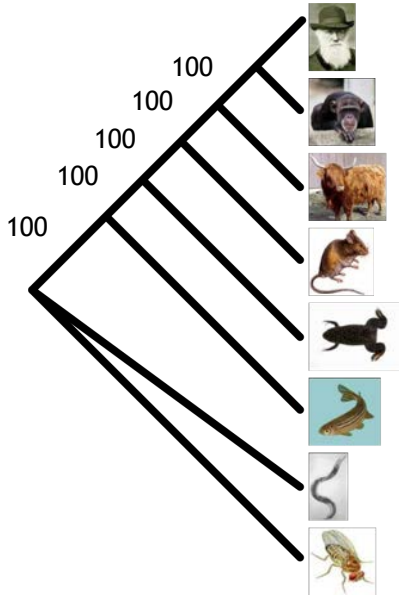
Species tree of life

Concatenation methods: Limitation

Supermatrix

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Phylogenomic Inference



Species tree of life

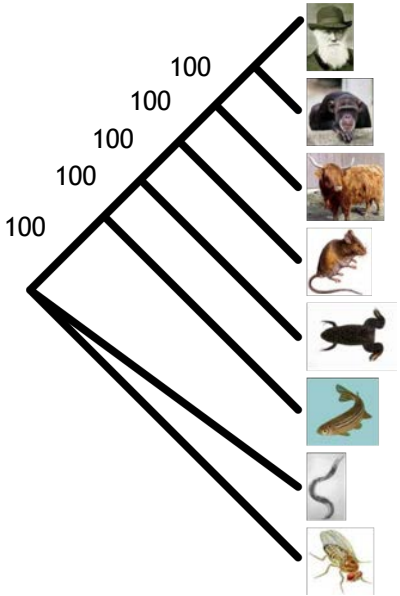
Bootstrap supports and Bayesian posteriors tend to 100% as #genes increases!

Concatenation methods: Limitation

Supermatrix

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Phylogenomic Inference



Species tree of life

Bootstrap supports and Bayesian posteriors tend to 100% as #genes increases!

Concatenation assumes a single tree across all loci

Potential *systematic bias*

Felsenstein (1985):

which not. Where the method of inferring phylogenies is one with undesirable statistical properties such as inconsistency, the bootstrap does not correct for these.

Coalescent/reconciliation methods

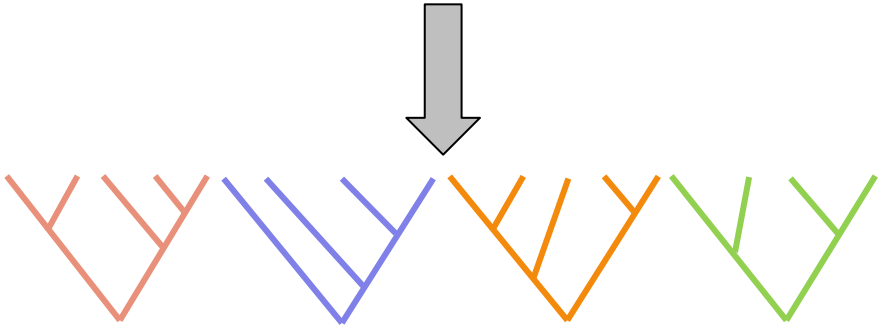
Supermatrix

Gene 1	Gene 2	……	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Coalescent/reconciliation methods

Supermatrix

Gene 1	Gene 2	……	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

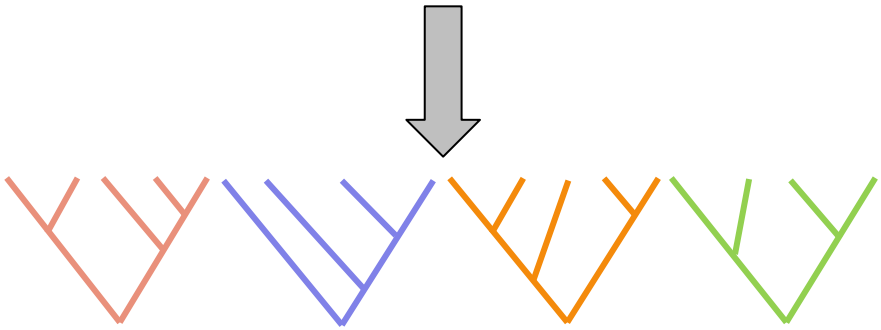


Gene tree 1 Gene tree 2 …… Gene tree 1,000

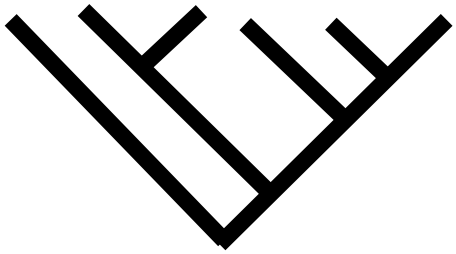
Coalescent/reconciliation methods

Supermatrix

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------



Gene tree 1 Gene tree 2 ..... Gene tree 1,000

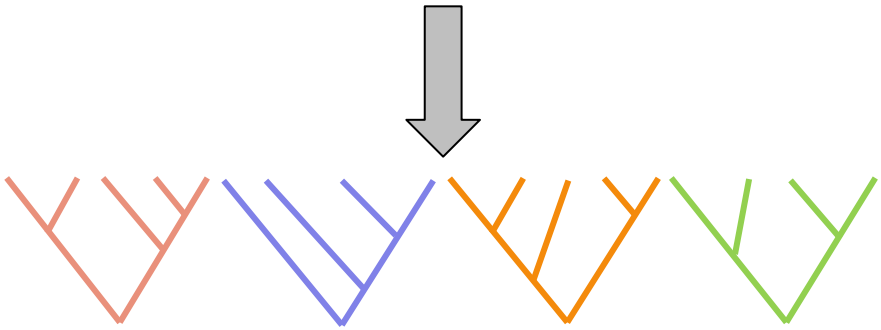


Species tree

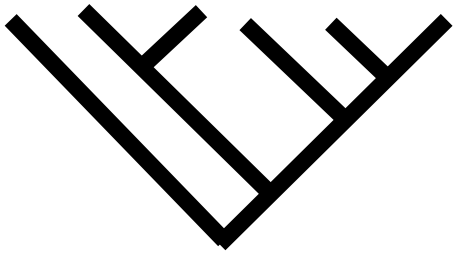
Coalescent/reconciliation methods

Supermatrix

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------



Gene tree 1 Gene tree 2 ..... Gene tree 1,000



Species tree

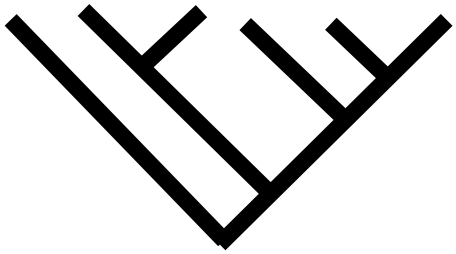
Gene Concordance Factor (gCF): How often a branch in species tree is found among gene trees?
0% ≤ gCF ≤ 100%

Coalescent/reconciliation methods

Supermatrix

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Gene tree 1 Gene tree 2 ..... Gene tree 1,000



Species tree

Gene Concordance Factor (gCF): How often a branch in species tree is found among gene trees?
0% ≤ gCF ≤ 100%

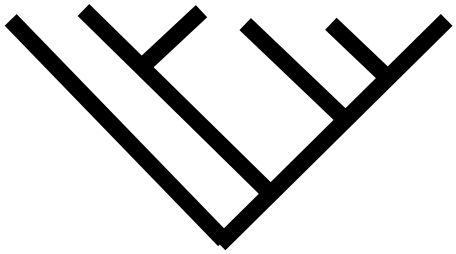
Implementation in IQ-TREE fully accounts for missing data

Coalescent/reconciliation methods

Supermatrix

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Gene tree 1 Gene tree 2 ..... Gene tree 1,000



Species tree

Gene Concordance Factor (gCF): How often a branch in species tree is found among gene trees?
0% ≤ gCF ≤ 100%

Implementation in IQ-TREE fully accounts for missing data

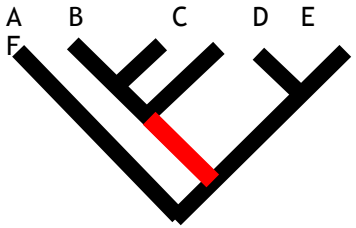
Problem: Uncertainties in gene trees!

Site Concordance Factor (sCF)

Supermatrix

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Site Concordance Factor (sCF):
How often a branch is "supported" by alignment sites?
33.3% ≅ sCF ≤ 100%

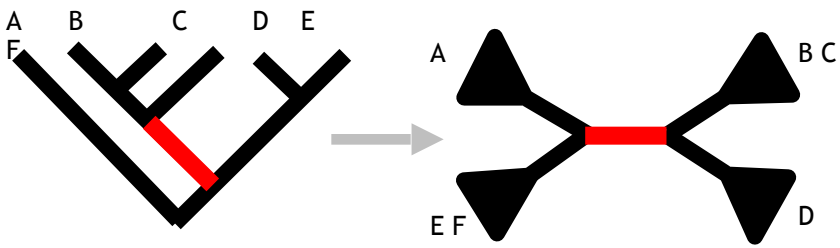


Site Concordance Factor (sCF)

Supermatrix

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Site Concordance Factor (sCF):
How often a branch is "supported" by alignment sites?
33.3% ≦ sCF ≤ 100%

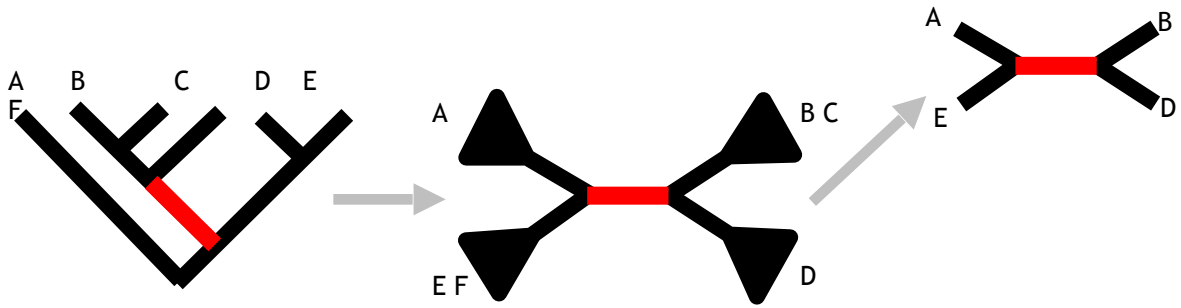


Site Concordance Factor (sCF)

Supermatrix

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Site Concordance Factor (sCF):
How often a branch is "supported" by alignment sites?
33.3% ≦ sCF ≤ 100%

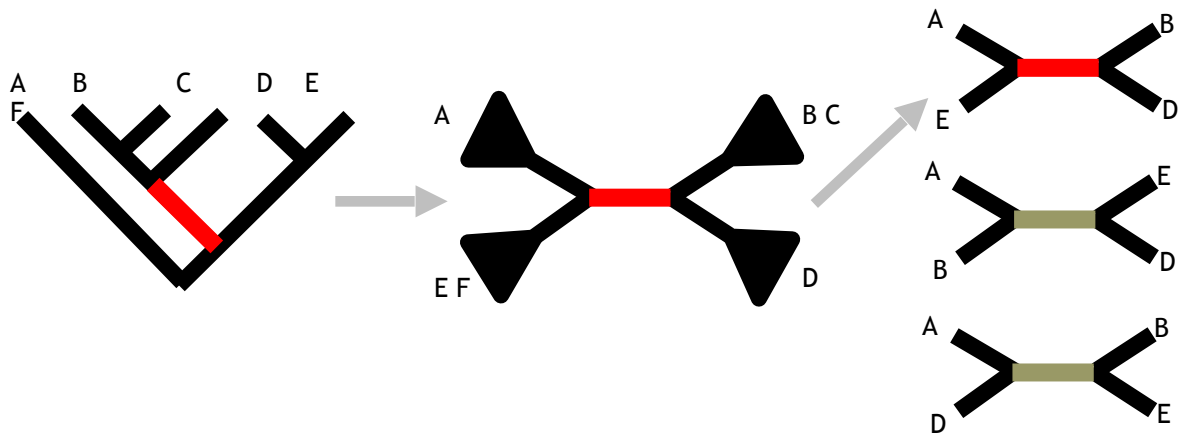


Site Concordance Factor (sCF)

Supermatrix

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Site Concordance Factor (sCF):
How often a branch is "supported" by alignment sites?
33.3% ≦ sCF ≤ 100%

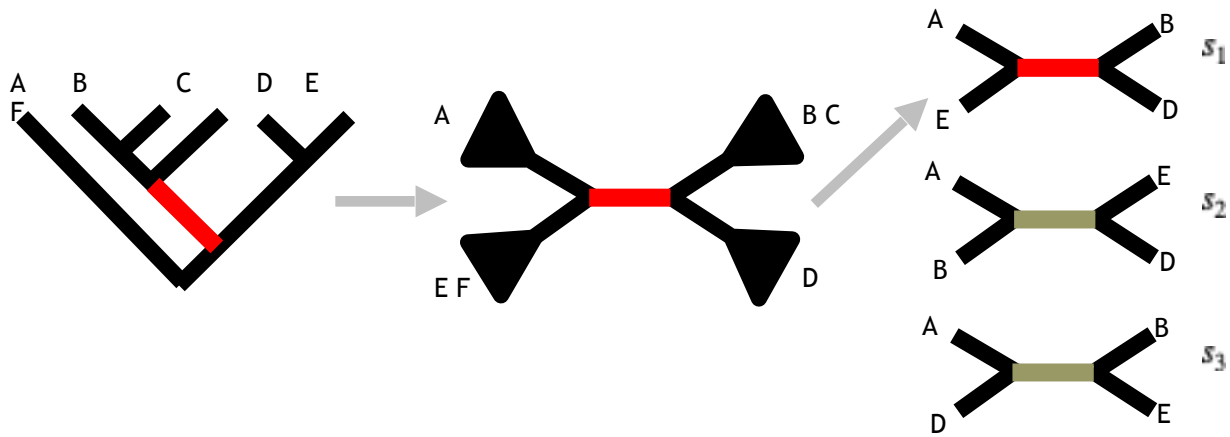


Site Concordance Factor (sCF)

Supermatrix

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Site Concordance Factor (sCF):
How often a branch is "supported" by alignment sites?
33.3% ≦ sCF ≤ 100%



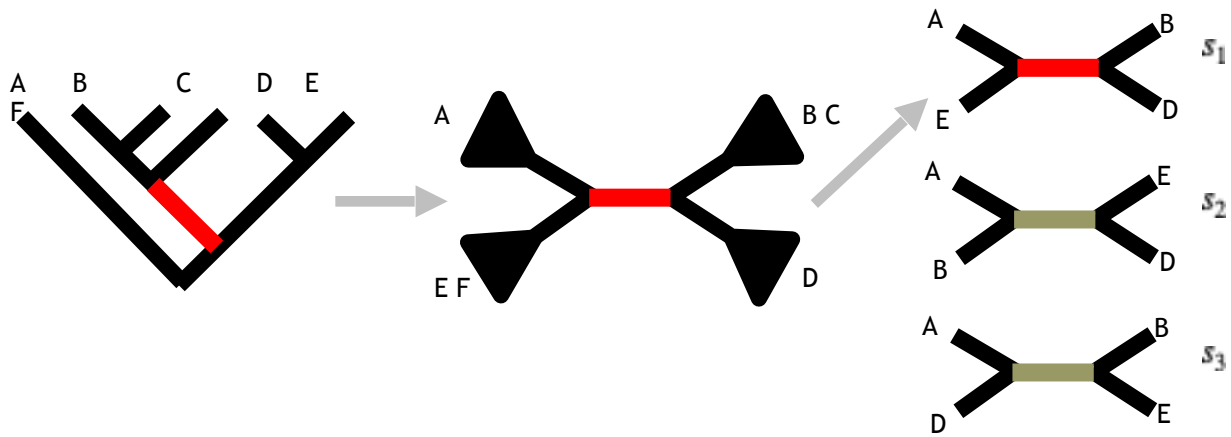
$$qCF(quartet) = \frac{s_1}{s_1 + s_2 + s_3}$$

Site Concordance Factor (sCF)

Supermatrix

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

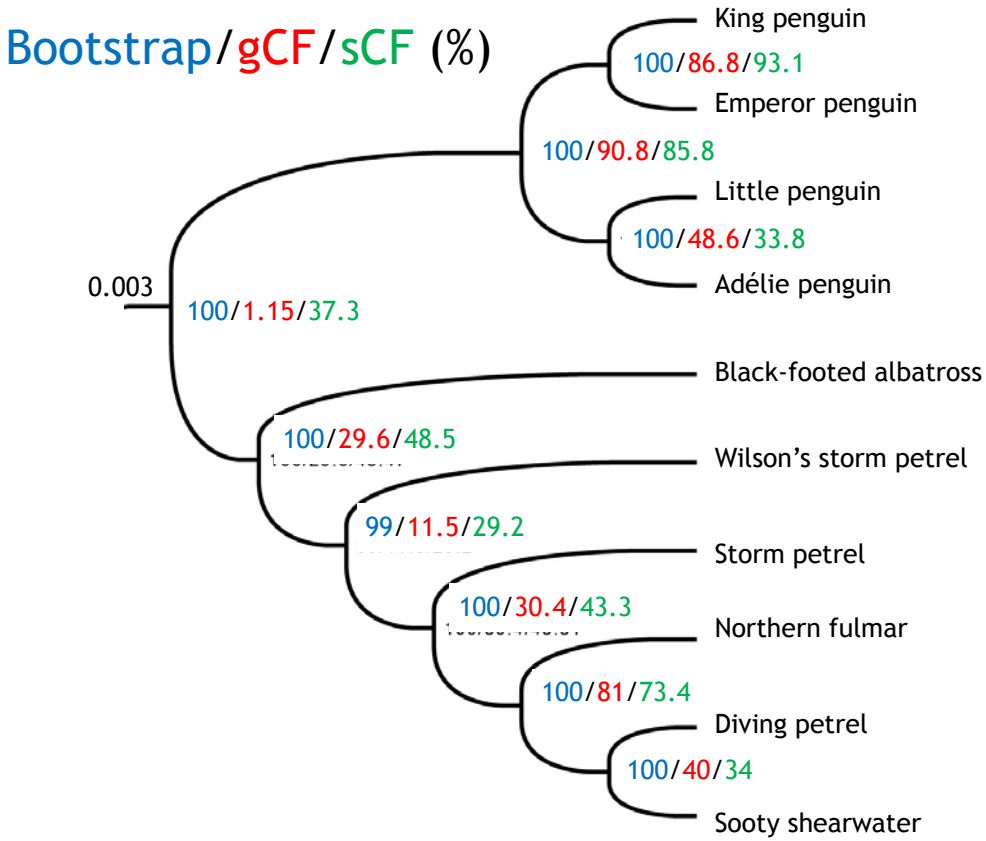
Site Concordance Factor (sCF):
How often a branch is "supported" by alignment sites?
33.3% ≦ sCF ≤ 100%



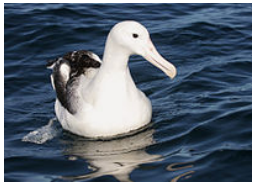
$$qCF(quartet) = \frac{s_1}{s_1 + s_2 + s_3}$$

$$sCF = qCF(100\ quartets)$$

An example birds data set (Reddy et al., 2017)

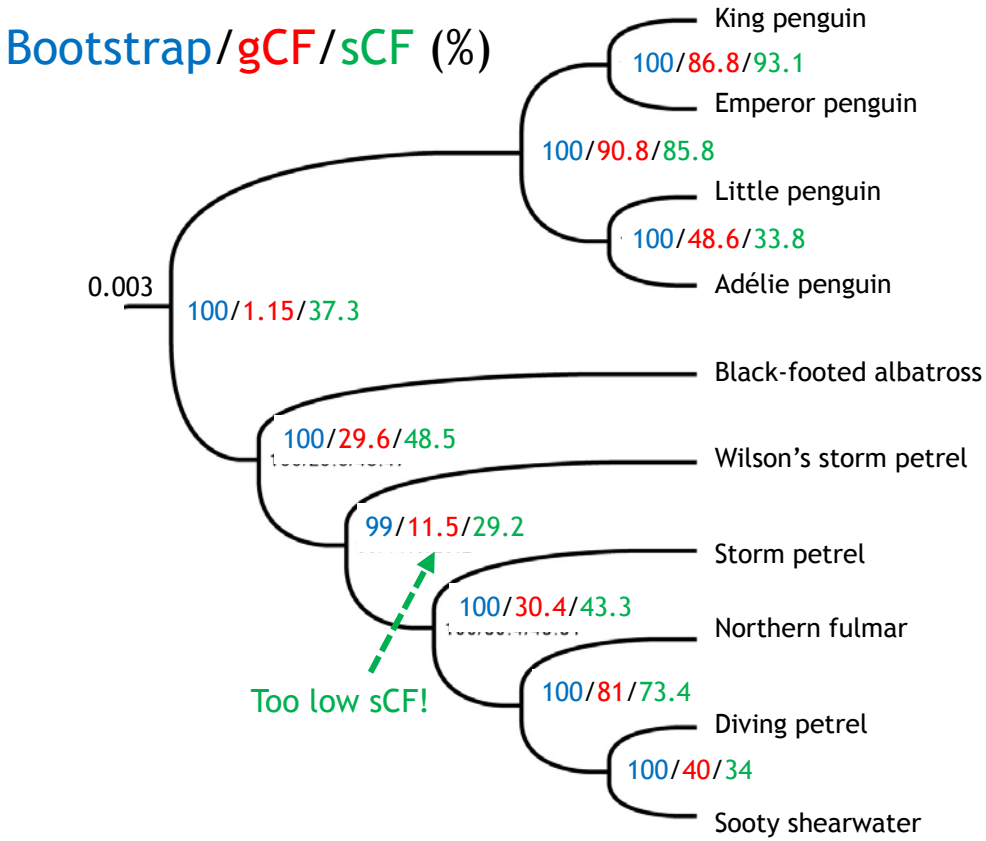


Penguins

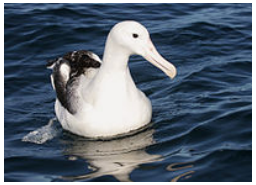


Tubenoses

An example birds data set (Reddy et al., 2017)

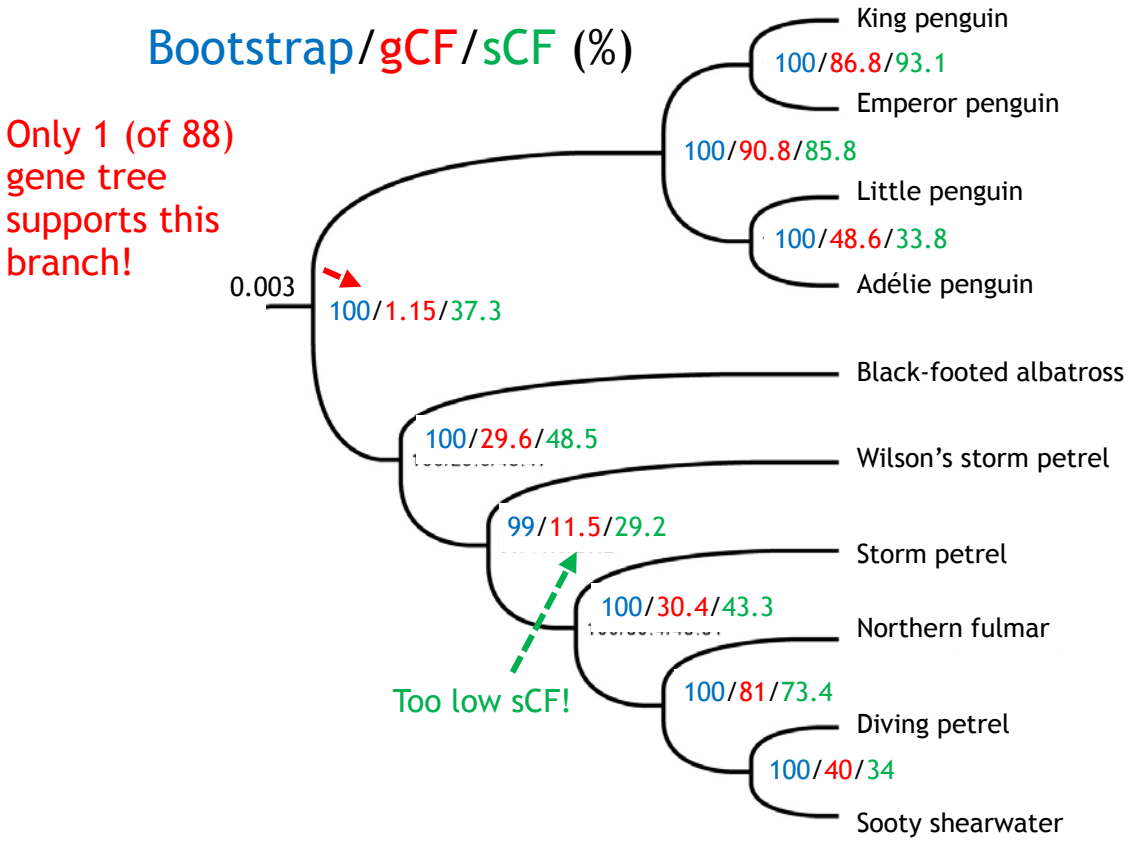


Penguins

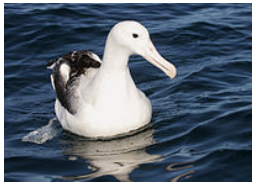


Tubenoses

An example birds data set (Reddy et al., 2017)



Penguins



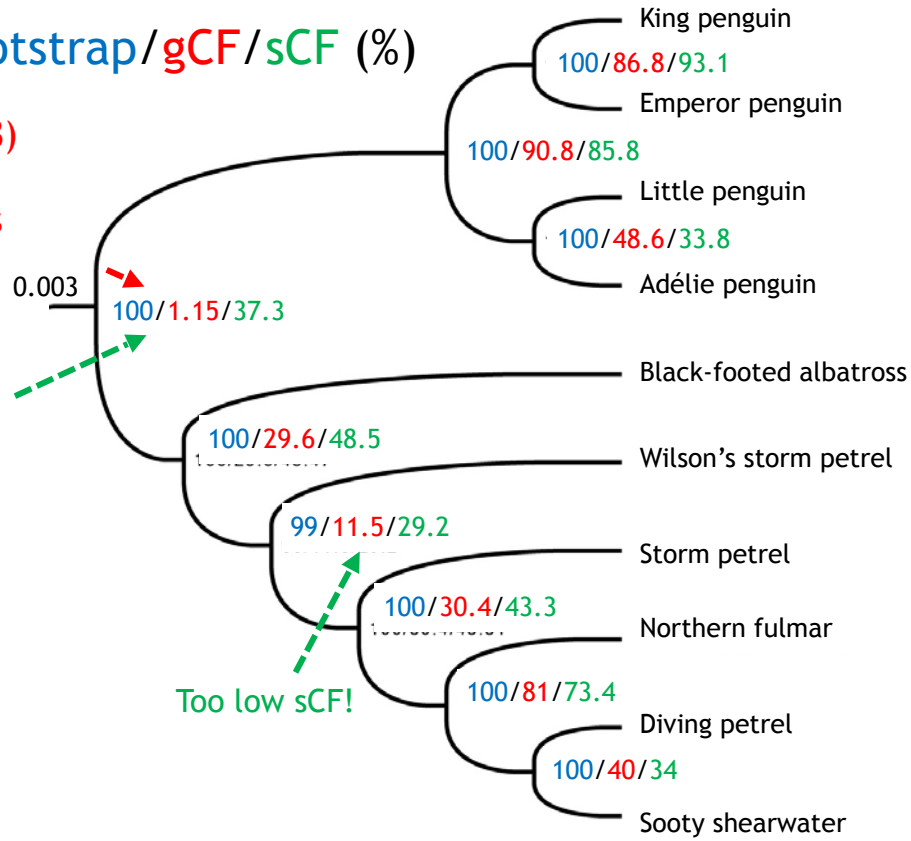
Tubenoses

An example birds data set (Reddy et al., 2017)

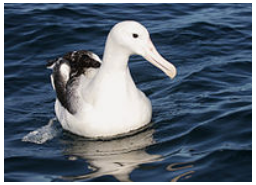
Bootstrap/gCF/sCF (%)

Only 1 (of 88) gene tree supports this branch!

- 131 sites support this branch
- 105 sites support NNI branch 1
- 114 sites support NNI branch 2



Penguins



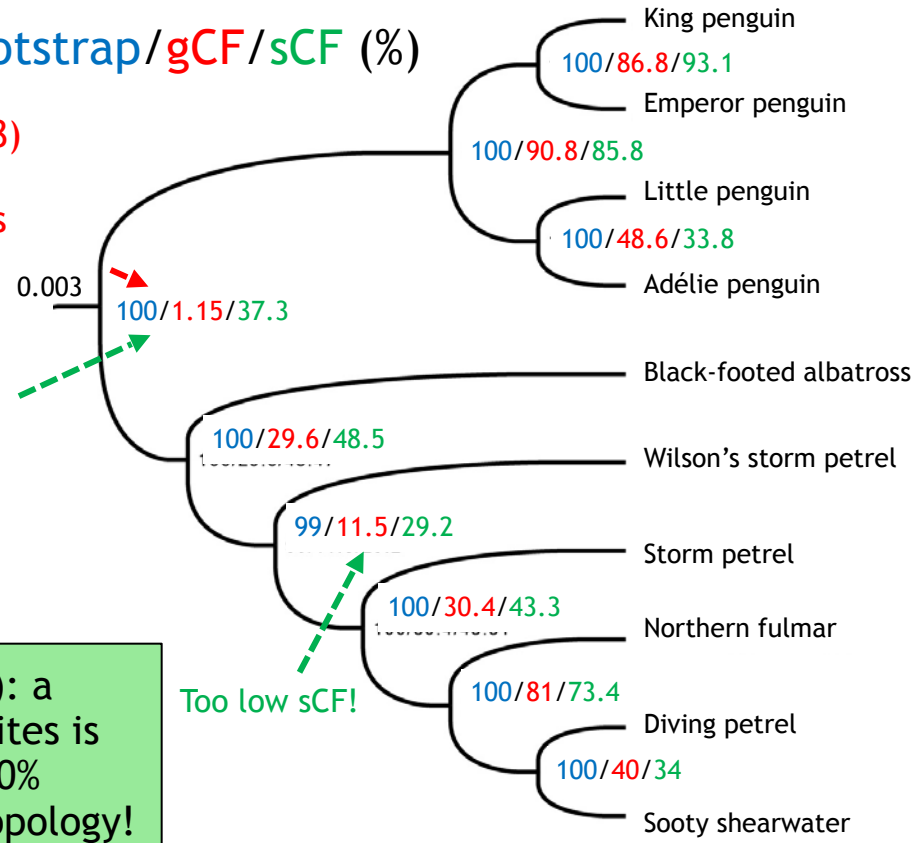
Tubenoses

An example birds data set (Reddy et al., 2017)

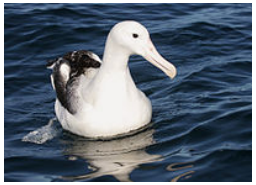
Bootstrap/gCF/sCF (%)

Only 1 (of 88) gene tree supports this branch!

- 131 sites support this branch
- 105 sites support NNI branch 1
- 114 sites support NNI branch 2



Penguins



Tubenoses

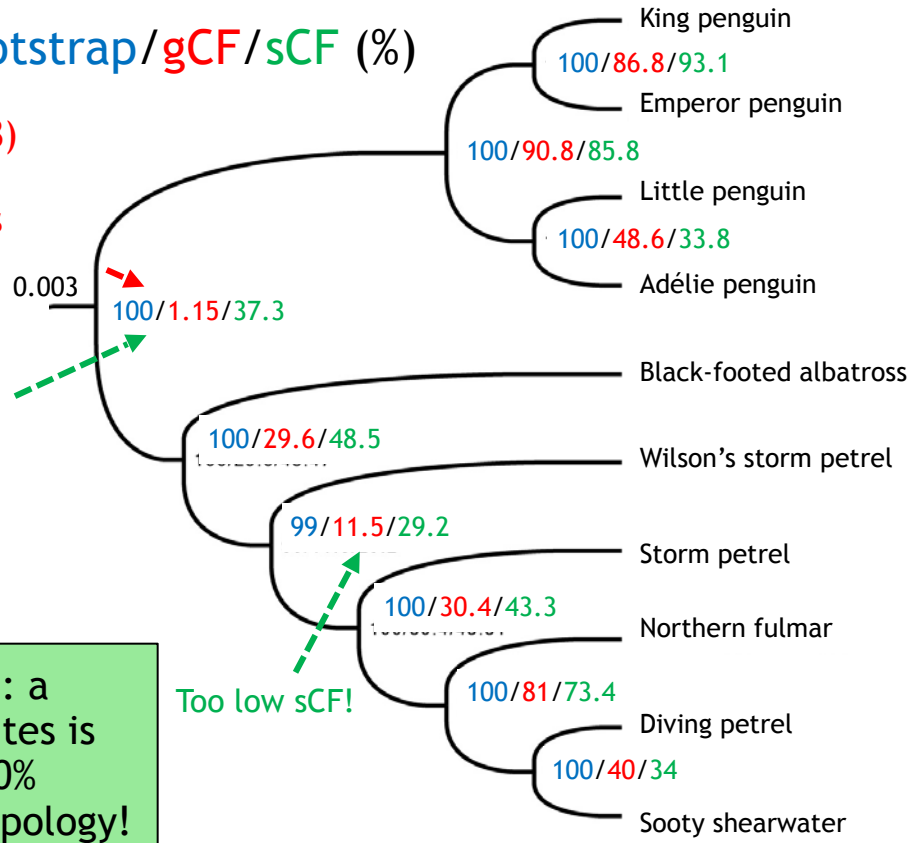
Felsenstein (1985): a difference of 20 sites is enough to give 100% support for one topology!

An example birds data set (Reddy et al., 2017)

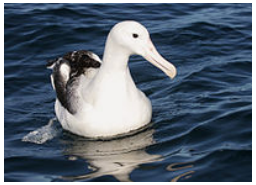
Bootstrap/gCF/sCF (%)

Only 1 (of 88) gene tree supports this branch!

- 131 sites support this branch
- 105 sites support NNI branch 1
- 114 sites support NNI branch 2



Penguins



Tubenoses

Felsenstein (1985): a difference of 20 sites is enough to give 100% support for one topology!

- gCF and sCF are useful when bootstrap supports reach 100%.
- CAUTION when gCF ~ 0% or sCF ~ 33%, even if BS ~ 100%.
- GREAT when gCF and sCF > 50%.

Exercises

1. Input data
2. Inferring the first phylogeny
3. Applying partition model
4. Choosing the best partitioning scheme
5. Tree topology tests
6. Concordance factors
7. Resampling partitions and sites
8. Identifying most influential genes
9. Wrapping up

http://www.iqtree.org/workshop/molevol2019

Suggestion: Use your laptop if possible,
as cluster nodes are quite slow!