IQ-TREE

Efficient software for phylogenomic inference

Stable release 1.6.12 (August 15, 2019)

Download v1.6.12 for macOS

Latest release 2.2.2.6 (May 27, 2023)

Download v2.2.2.6 for macOS

All Downloads | Documentation

IQ-TREE has been developed by 12+ contributors:

From ANU:


James Barbetti


Thomas Wong


Robert Lanfear


Bui Quang Minh


Nhan Ly-Trong


Piyumal Demotte

From international:


Michael Woodhams


Olga Chernomor


Arndt von Haeseler


Dominik Schrempf


Heiko A. Schmidt


Diep Thi Hoang

Past members:

Lam Tung Nguyen

Jana Trifinopoulos

IQ-TREE Intro MOLE 2024

Slides from Bui Quang Minh

(Edited by Blake Fauskee)

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:

• 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)

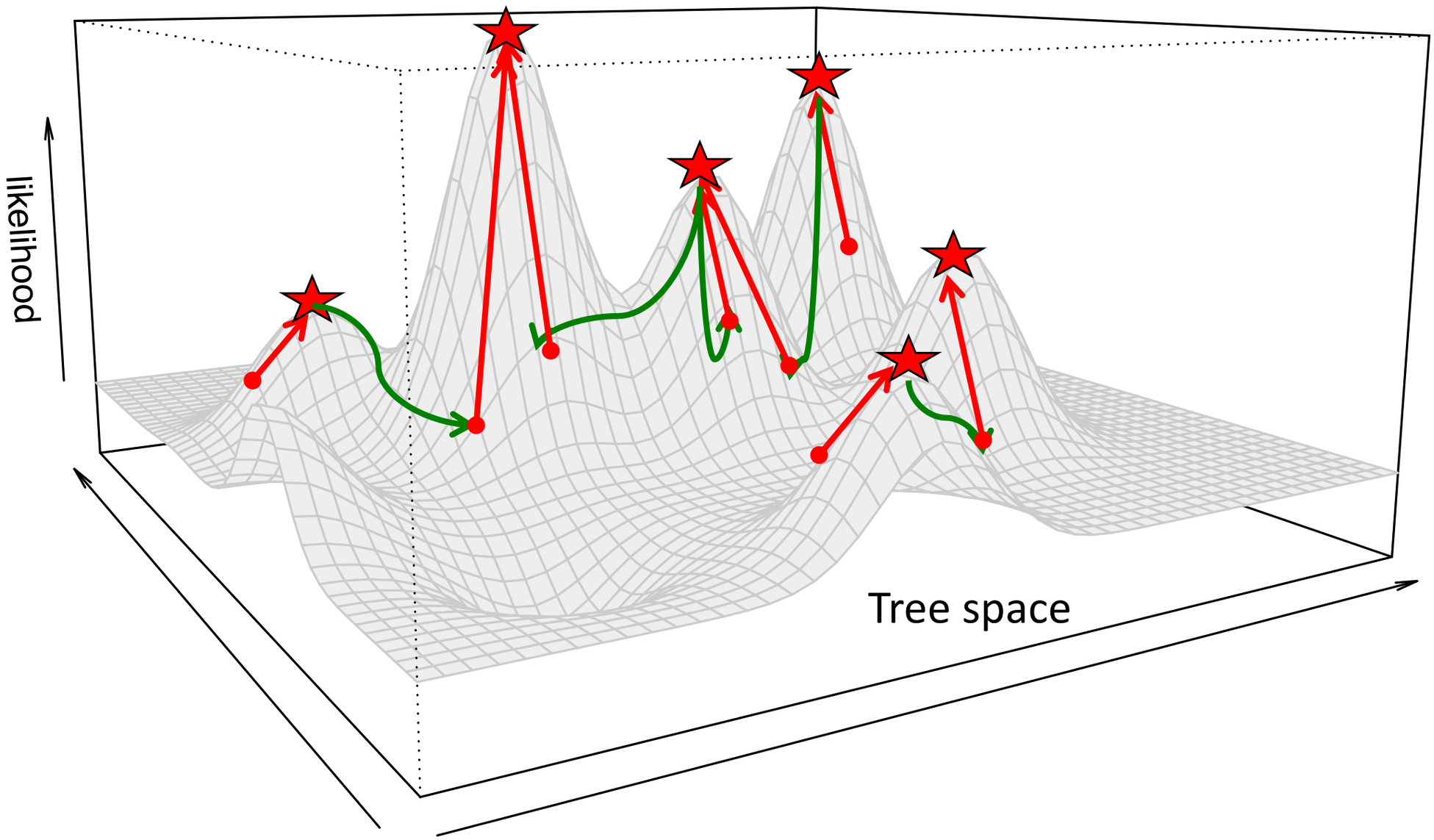
With IQ-TREE we aim to:

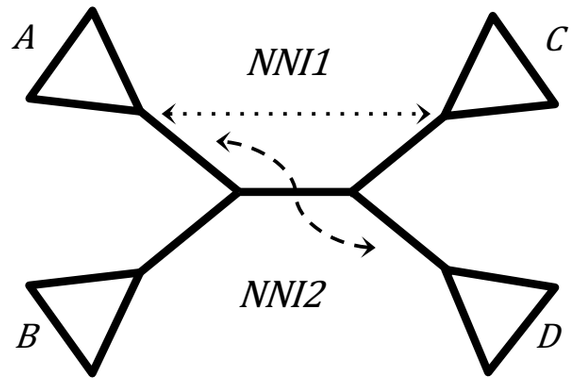
• Analyze ultra-large data sets.
• Provide many (if not most) "useful" models of sequence evolution.

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

IQ-TREE: A new stochastic algorithm





Nearest neighbor interchange

* 100 starting trees (99 parsimony, 1 NJ)
* Keeping a “population” of 20 best trees
* Stop if unsuccessful for 100 consecutive down-hill + up-hill moves

Lam-Tung Nguyen

Heiko Schmidt

Arndt 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



Maximum parsimony
(Population of starting trees)

Aoraki/Mt Cook

Mt Tasman

IQ-TREE: A stochastic tree search algorithm

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



Maximum parsimony

Hill-climbing NNI

Aoraki/Mt Cook

Mt Tasman

IQ-TREE: A stochastic tree search algorithm

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



Maximum parsimony

Hill-climbing NNI

Downhill (random) NNIs

Aoraki/Mt Cook

Mt Tasman

IQ-TREE: A stochastic tree search algorithm

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



Maximum parsimony

Hill-climbing NNI

Downhill (random) NNIs

Aoraki/Mt Cook

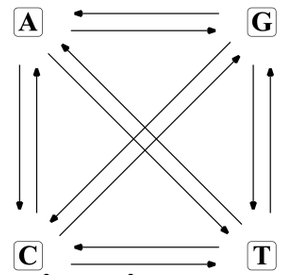
Mt Tasman

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  ----------
```

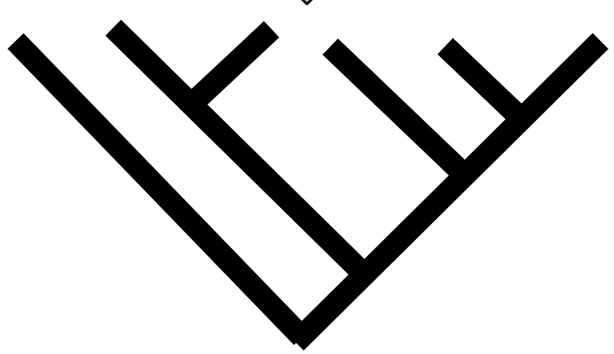
Model selection



Substitution models

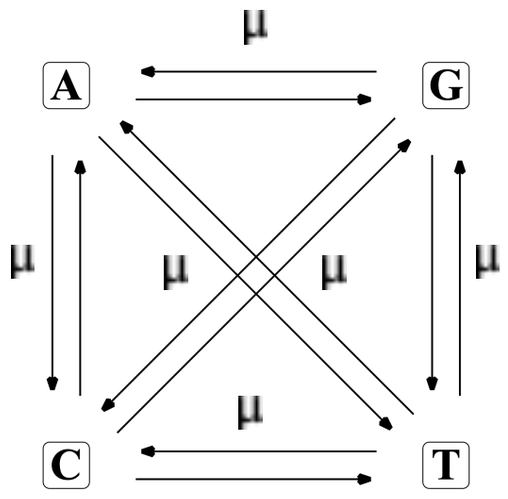
Tree search

Bootstrap



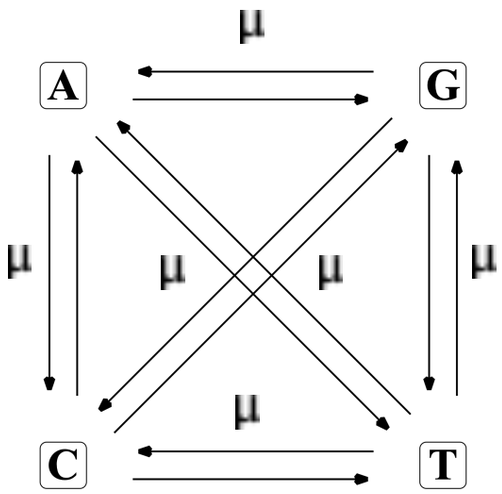
Phylogeny

Models of sequence evolution

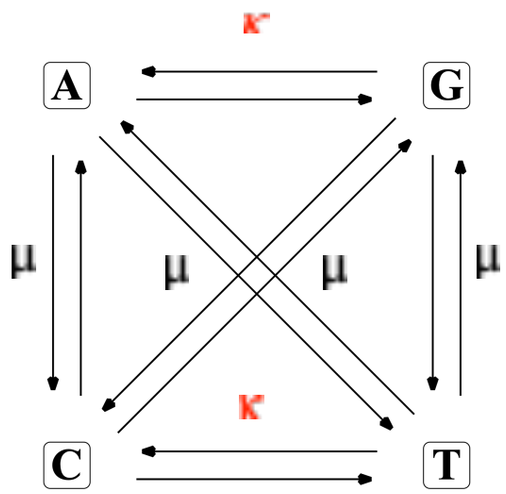


JC
(Jukes & Cantor 1969)

Models of sequence evolution

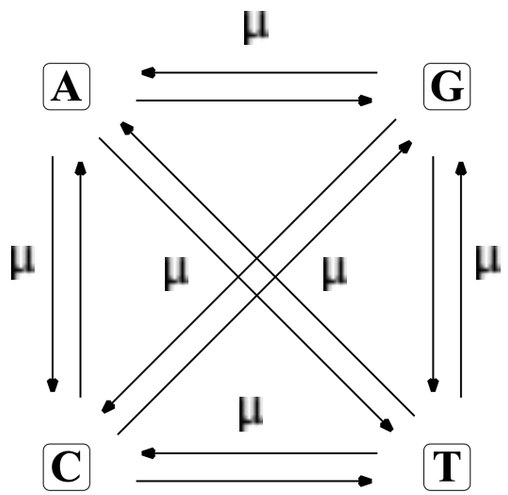


JC
(Jukes & Cantor 1969)

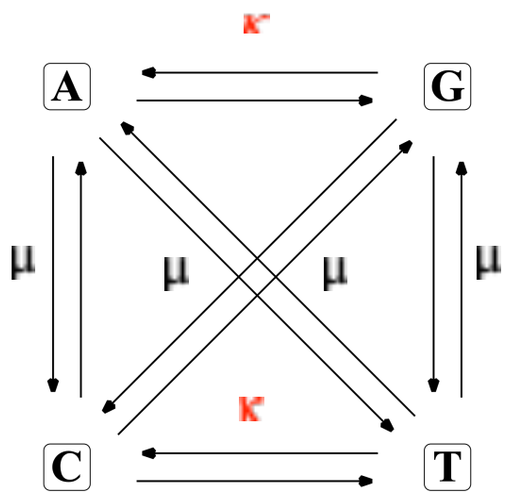


HKY
(Hasegawa, Kishino, Yano 1985)

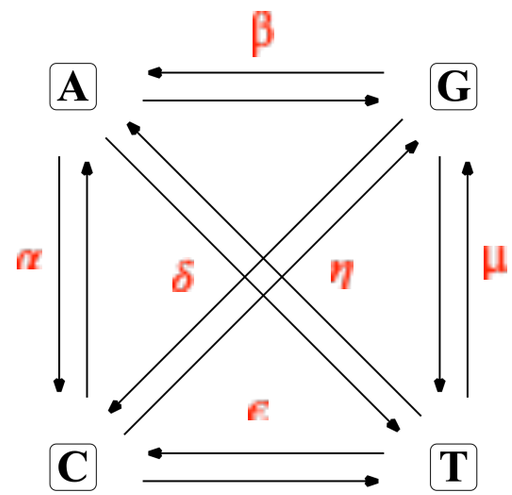
Models of sequence evolution



JC
(Jukes & Cantor 1969)

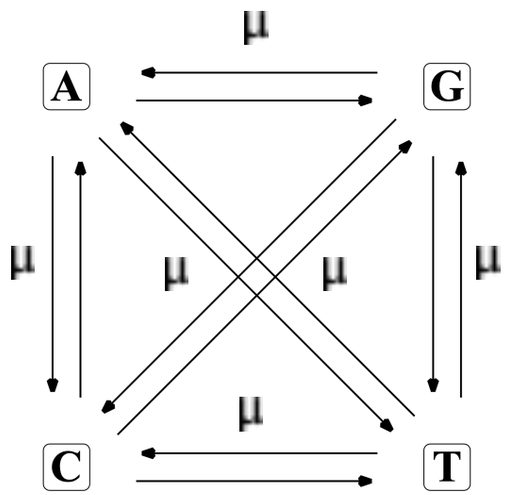


HKY
(Hasegawa, Kishino, Yano 1985)

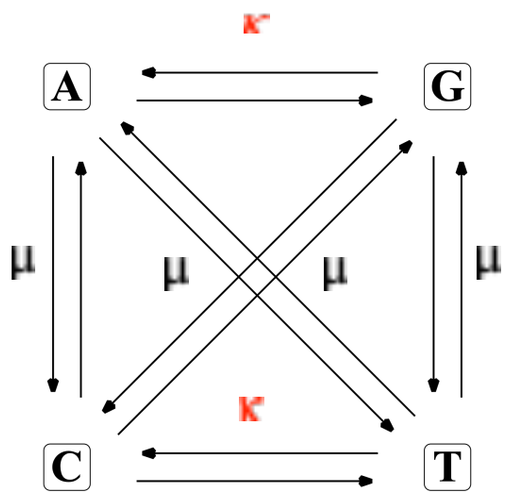


GTR
(General Time Reversible, 1986)

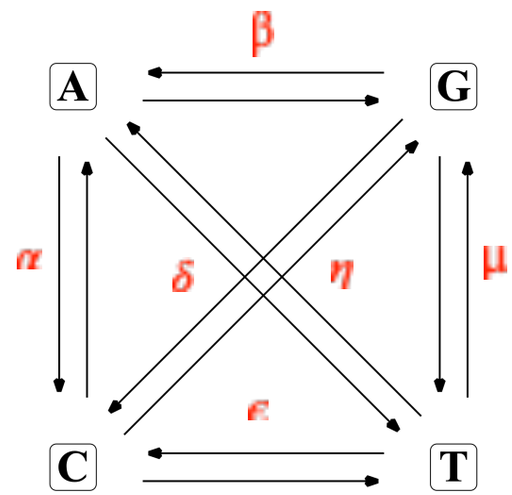
Models of sequence evolution



JC
(Jukes & Cantor 1969)



HKY
(Hasegawa, Kishino, Yano 1985)

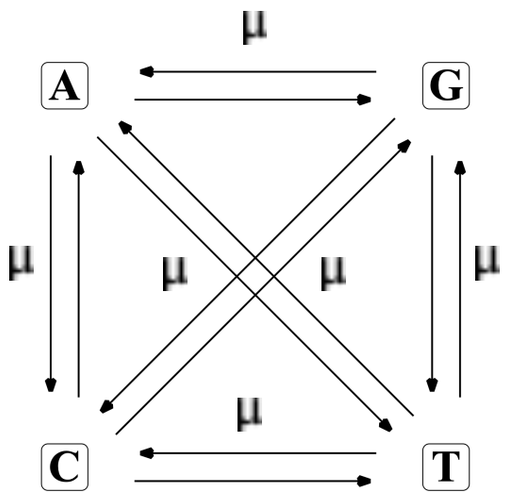


GTR
(General Time Reversible, 1986)

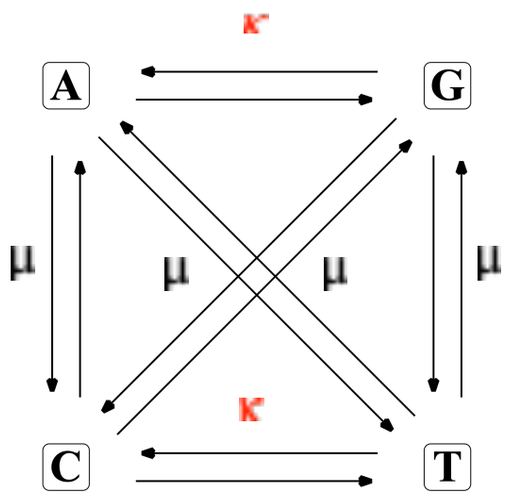
Rate heterogeneity: alignment sites evolved at different rates. Some slow, some fast.

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).

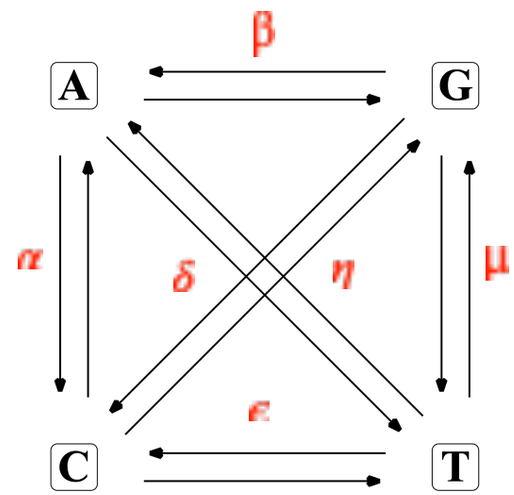
Models of sequence evolution



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

12.1 DNA models

12.1.1 Base substitution rates

IQ-TREE includes all common DNA models (ordered by complexity):

Model	df	Explanation	Code
JC or JC69	0	Equal substitution rates and equal base frequencies (Jukes and Cantor, 1969).	000000
F81	3	Equal rates but unequal base freq. (Felsenstein, 1981).	000000
K80 or K2P	1	Unequal transition/transversion rates and equal base freq. (Kimura, 1980).	010010
HKY or HKY85	4	Unequal transition/transversion rates and unequal base freq. (Hasegawa, Kishino and Yano, 1985).	010010
TN or TN93	5	Like `HKY` but unequal purine/pyrimidine rates (Tamura and Nei, 1993).	010020
TNe	2	Like `TN` but equal base freq.	010020
K81 or K3P	2	Three substitution types model and equal base freq. (Kimura, 1981).	012210
K81u	5	Like `K81` but unequal base freq.	012210
TPM2	2	AC=AT, AG=CT, CG=GT and equal base freq.	010212
TPM2u	5	Like `TPM2` but unequal base freq.	010212
TPM3	2	AC=CG, AG=CT, AT=GT and equal base freq.	012012
TPM3u	5	Like `TPM3` but unequal base freq.	012012
TIM	6	Transition model, AC=GT, AT=CG and unequal base freq.	012230
TIMe	3	Like `TIM` but equal base freq.	012230
TIM2	6	AC=AT, CG=GT and unequal base freq.	010232
TIM2e	3	Like `TIM2` but equal base freq.	010232
TIM3	6	AC=CG, AT=GT and unequal base freq.	012032
TIM3e	3	Like `TIM3` but equal base freq.	012032
TVM	7	Transversion model, AG=CT and unequal base freq.	012314
TVMe	4	Like `TVM` but equal base freq.	012314
SYM	5	Symmetric model with unequal rates but equal base freq. (Zharkikh, 1994).	012345
GTR	8	General time reversible model with unequal rates and unequal base freq. (Tavare, 1986).	012345

Model selection

12.1 DNA models

12.1.1 Base substitution rates

IQ-TREE includes all common DNA models (ordered by complexity):

Model	df	Explanation	Code
JC or JC69	0	Equal substitution rates and equal base frequencies (Jukes and Cantor, 1969).	000000
F81	3	Equal rates but unequal base freq. (Felsenstein, 1981).	000000
K80 or K2P	1	Unequal transition/transversion rates and equal base freq. (Kimura, 1980).	010010
HKY or HKY85	4	Unequal transition/transversion rates and unequal base freq. (Hasegawa, Kishino and Yano, 1985).	010010
TN or TN93	5	Like `HKY` but unequal purine/pyrimidine rates (Tamura and Nei, 1993).	010020
TNe	2	Like `TN` but equal base freq.	010020
K81 or K3P	2	Three substitution types model and equal base freq. (Kimura, 1981).	012210
K81u	5	Like `K81` but unequal base freq.	012210
TPM2	2	AC=AT, AG=CT, CG=GT and equal base freq.	010212
TPM2u	5	Like `TPM2` but unequal base freq.	010212
TPM3	2	AC=CG, AG=CT, AT=GT and equal base freq.	012012
TPM3u	5	Like `TPM3` but unequal base freq.	012012
TIM	6	Transition model, AC=GT, AT=CG and unequal base freq.	012230
TIMe	3	Like `TIM` but equal base freq.	012230
TIM2	6	AC=AT, CG=GT and unequal base freq.	010232
TIM2e	3	Like `TIM2` but equal base freq.	010232
TIM3	6	AC=CG, AT=GT and unequal base freq.	012032
TIM3e	3	Like `TIM3` but equal base freq.	012032
TVM	7	Transversion model, AG=CT and unequal base freq.	012314
TVMe	4	Like `TVM` but equal base freq.	012314
SYM	5	Symmetric model with unequal rates but equal base freq. (Zharkikh, 1994).	012345
GTR	8	General time reversible model with unequal rates and unequal base freq. (Tavare, 1986).	012345

12.6 Rate heterogeneity across sites

IQ-TREE supports all common rate heterogeneity across sites models:

RateType	Explanation
+I	allowing for a proportion of invariable sites.
+G	discrete Gamma model (Yang, 1994) with default 4 rate categories. The number of categories can be changed with e.g. `+G8`.
+GC	continuous Gamma model (Yang, 1994) (for AliSim only).
+I+G	invariable site plus discrete Gamma model (Gu et al., 1995).
+R	FreeRate model (Yang, 1995; Soubrier et al., 2012) that generalizes the `+G` model by relaxing the assumption of Gamma-distributed rates. The number of categories can be specified with e.g. `+R6` (default 4 categories if not specified). The FreeRate model typically fits data better than the `+G` model and is recommended for analysis of large data sets.
+I+R	invariable site plus FreeRate model.

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!

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

1. Akaike information criterion (*A*

2. Bayesian information criterion

where *n* is the number of alignm

Select the model with **smallest *A***

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

ModelFinder: fast model selection for accurate phylogenetic estimates

Subha Kalyaanamoorthy[1,2,6], Bui Quang Minh[3,6], Thomas K F Wong[1,4,6], Arndt von Haeseler[3,5] & Lars S Jermiin[1,4]

Model selection

12.1 DNA models

12.1.1 Base substitution rates

IQ-TREE includes all common DNA models (ordered by complexity):

Model	df	Explanation	Code
JC or JC69	0	Equal substitution rates and equal base frequencies (Jukes and Cantor, 1969).	000000
F81	3	Equal rates but unequal base freq. (Felsenstein, 1981).	000000
K80 or K2P	1	Unequal transition/transversion rates and equal base freq. (Kimura, 1980).	010010
HKY or HKY85	4	Unequal transition/transversion rates and unequal base freq. (Hasegawa, Kishino and Yano, 1985).	010010
TN or TN93	5	Like `HKY` but unequal purine/pyrimidine rates (Tamura and Nei, 1993).	010020
TNe	2	Like `TN` but equal base freq.	010020
K81 or K3P	2	Three substitution types model and equal base freq. (Kimura, 1981).	012210
K81u	5	Like `K81` but unequal base freq.	012210
TPM2	2	AC=AT, AG=CT, CG=GT and equal base freq.	010212
TPM2u	5	Like `TPM2` but unequal base freq.	010212
TPM3	2	AC=CG, AG=CT, AT=GT and equal base freq.	012012
TPM3u	5	Like `TPM3` but unequal base freq.	012012
TIM	6	Transition model, AC=GT, AT=CG and unequal base freq.	012230
TIMe	3	Like `TIM` but equal base freq.	012230
TIM2	6	AC=AT, CG=GT and unequal base freq.	010232
TIM2e	3	Like `TIM2` but equal base freq.	010232
TIM3	6	AC=CG, AT=GT and unequal base freq.	012032
TIM3e	3	Like `TIM3` but equal base freq.	012032
TVM	7	Transversion model, AG=CT and unequal base freq.	012314
TVMe	4	Like `TVM` but equal base freq.	012314
SYM	5	Symmetric model with unequal rates but equal base freq. (Zharkikh, 1994).	012345
GTR	8	General time reversible model with unequal rates and unequal base freq. (Tavare, 1986).	012345

12.6 Rate heterogeneity across sites

IQ-TREE supports all common rate heterogeneity across sites models:

RateType	Explanation
+I	allowing for a proportion of invariable sites.
+G	discrete Gamma model (Yang, 1994) with default 4 rate categories. The number of categories can be changed with e.g. `+G8`.
+GC	continuous Gamma model (Yang, 1994) (for AliSim only).
+I+G	invariable site plus discrete Gamma model (Gu et al., 1995).
+R	FreeRate model (Yang, 1995; Soubrier et al., 2012) that generalizes the `+G` model by relaxing the assumption of Gamma-distributed rates. The number of categories can be specified with e.g. `+R6` (default 4 categories if not specified). The FreeRate model typically fits data better than the `+G` model and is recommended for analysis of large data sets.
+I+R	invariable site plus FreeRate model.

Mixture models

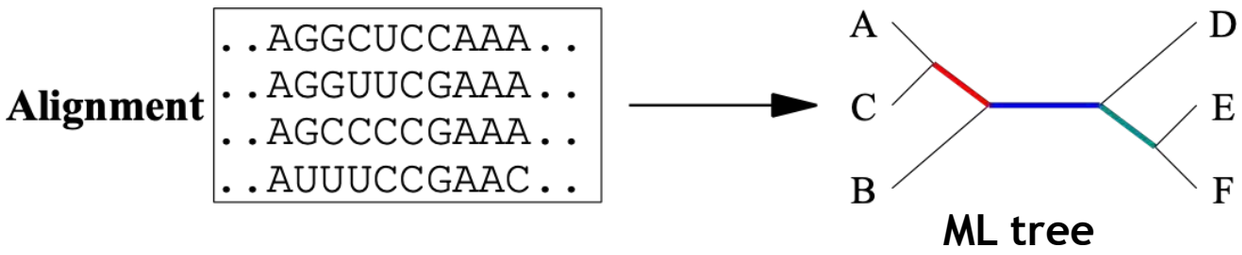
Does not assign alignment sites to a specific model

Rather, assigns each alignment site a probability/ weight of belonging to each mixture class (models)

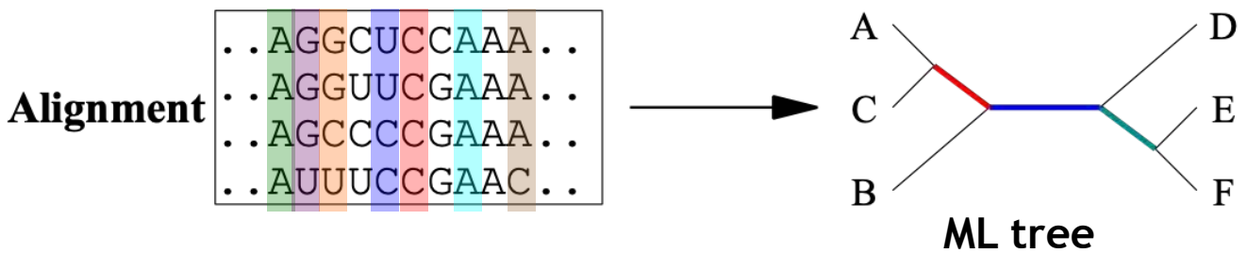
```
iqtree -s example.phy -m "MIX{JC,HKY}"
```

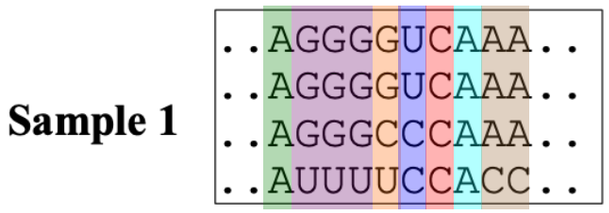
Gamma-distributed site-rate heterogeneity is an example of a mixture model

Bootstrap: How reliable are branches of the tree?

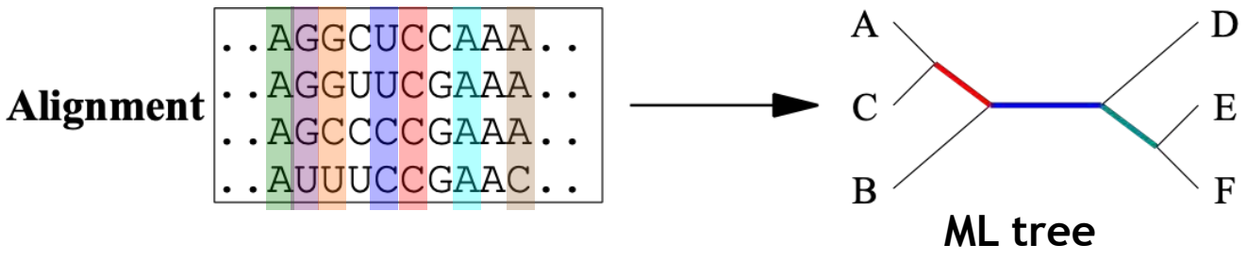


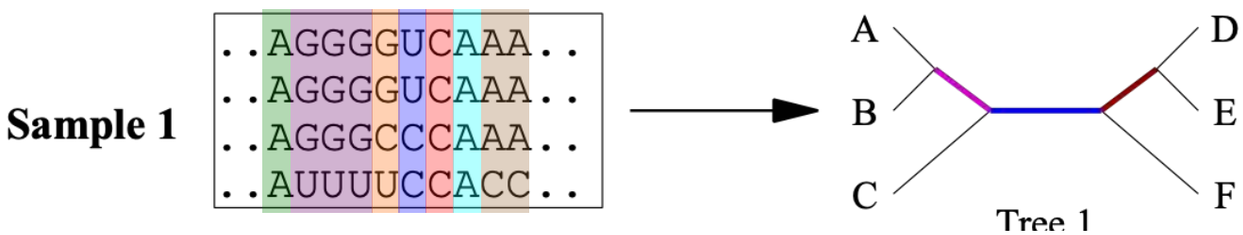
Bootstrap: How reliable are branches of the tree?



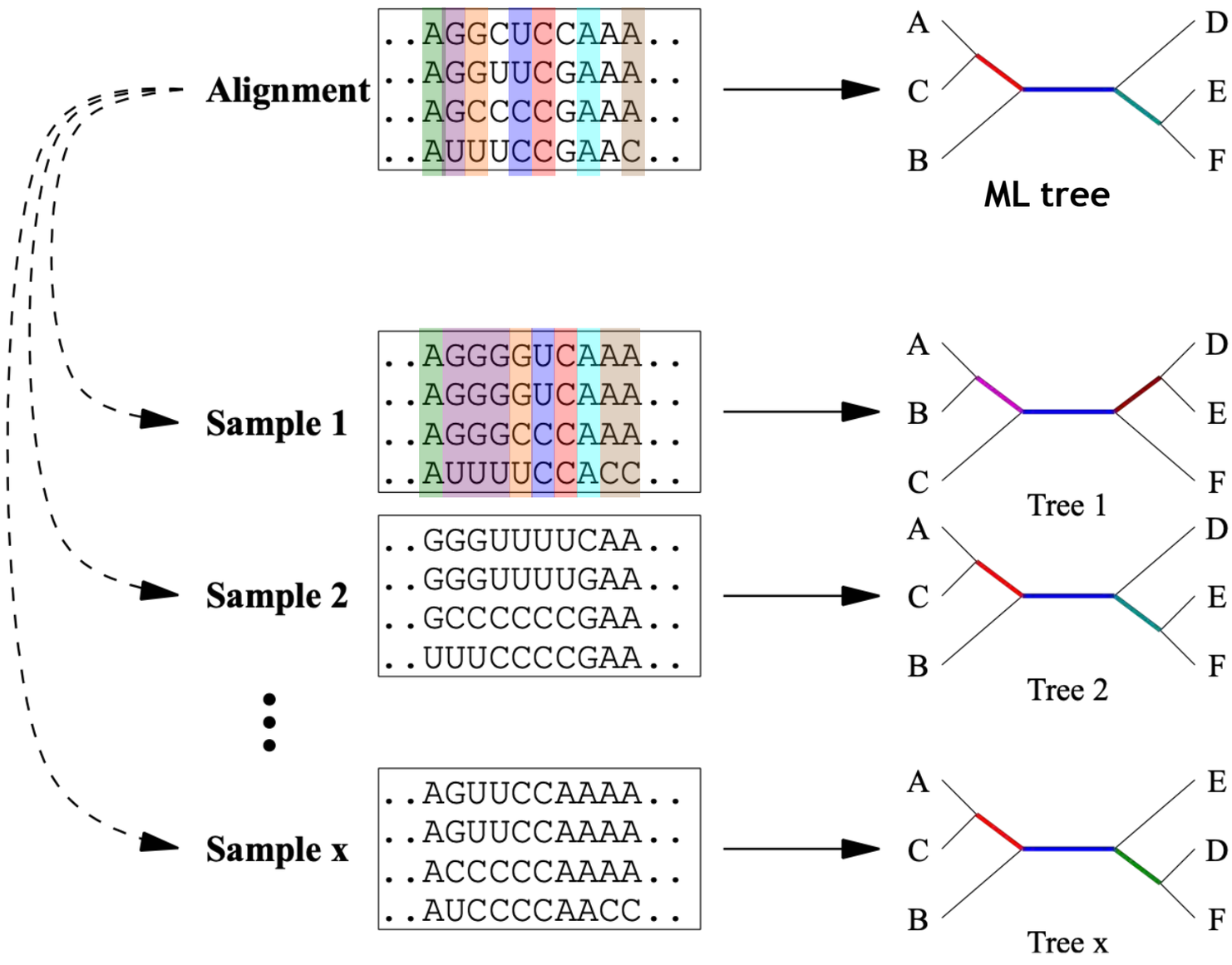


Bootstrap: How reliable are branches of the tree?



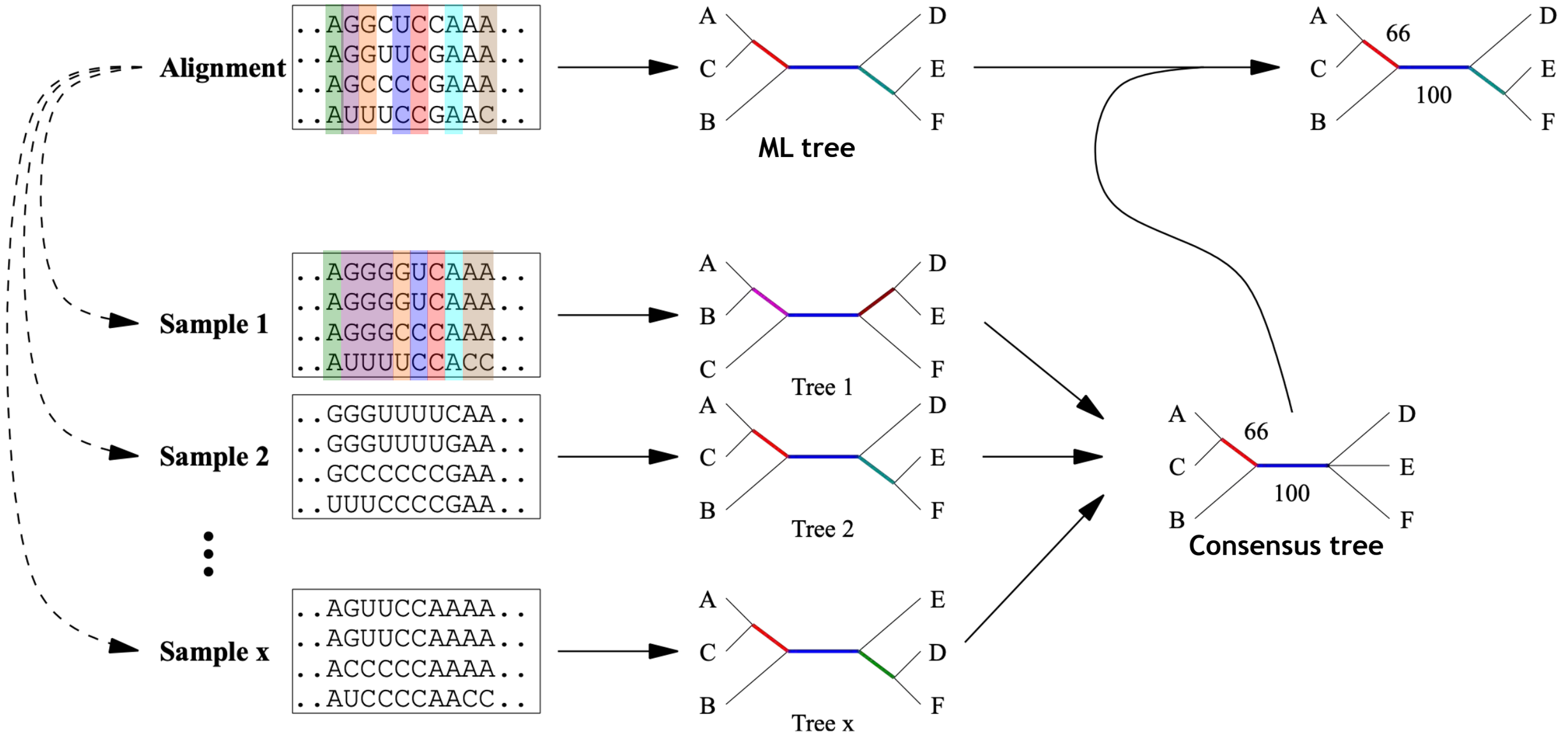


Bootstrap: How reliable are branches of the tree?



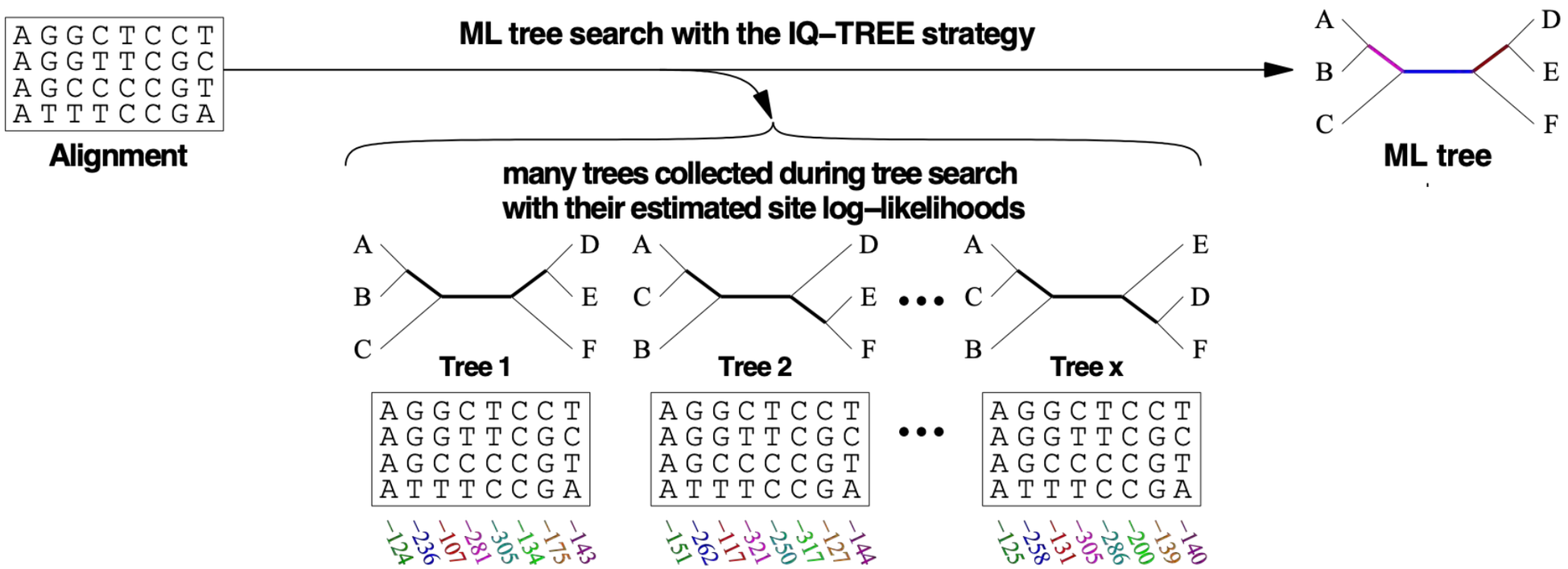
Bootstrap: How reliable are branches of the tree?

Generally time and resource heavy



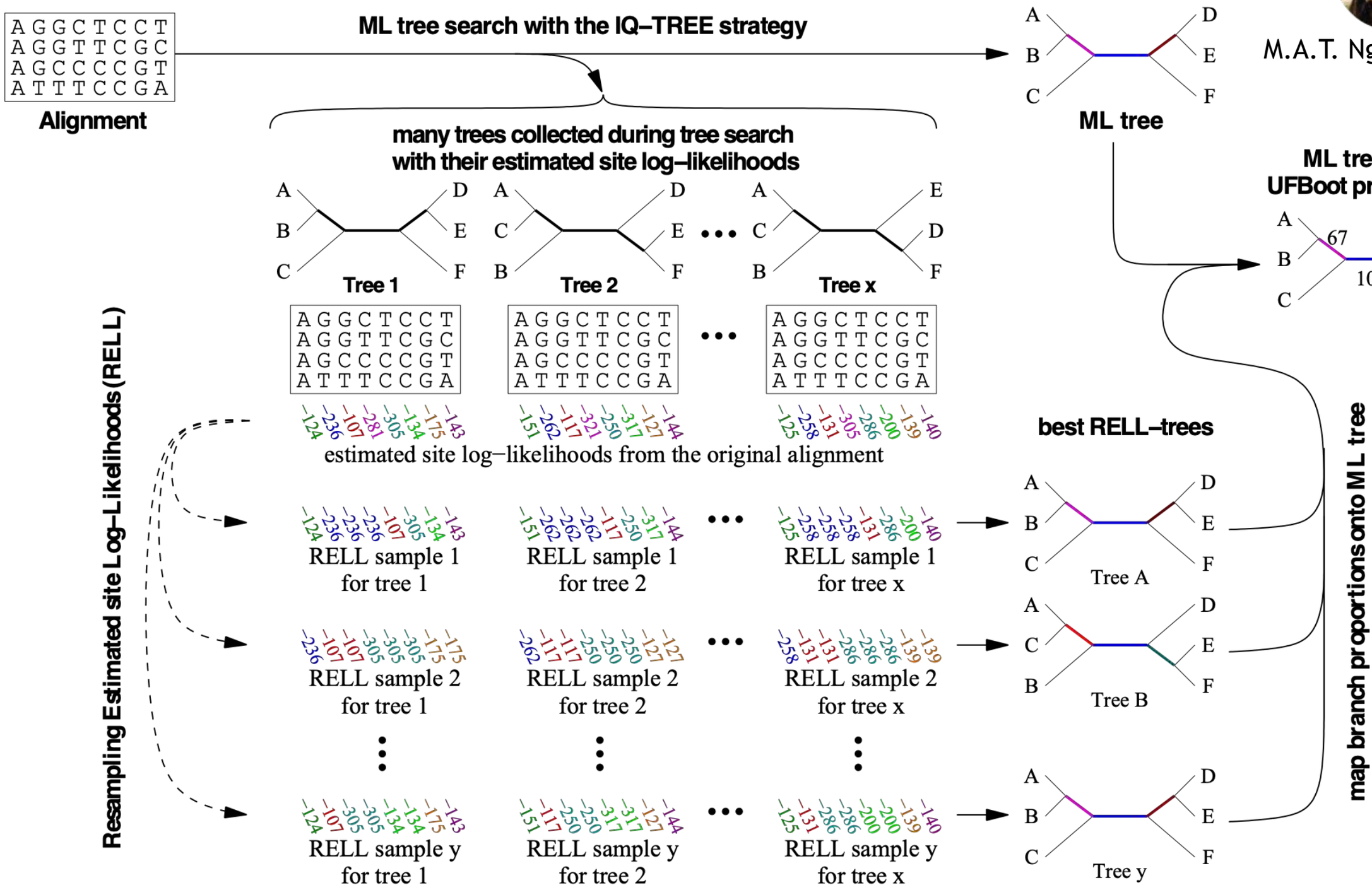
UFBoot: Ultrafast bootstrap approximation


M.A.T. Nguyen, A. von Haese



UFBoot: Ultrafast bootstrap approximation


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

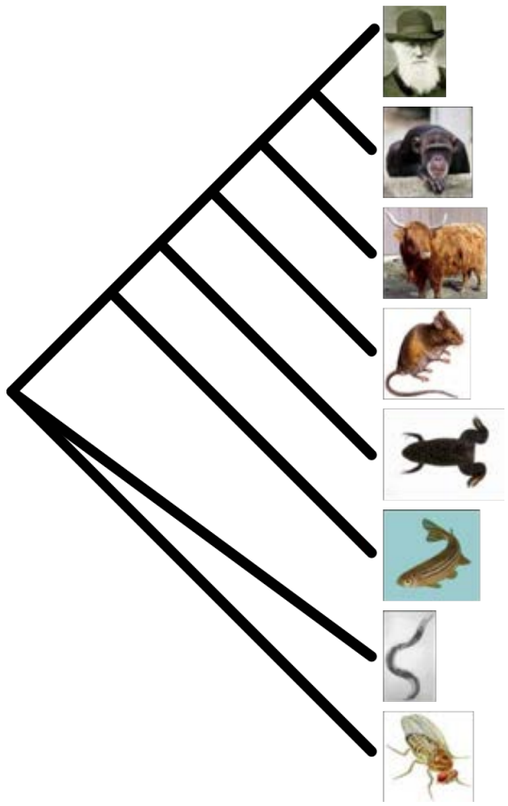


Genome-scale data: Concatenation methods

Supermatrix

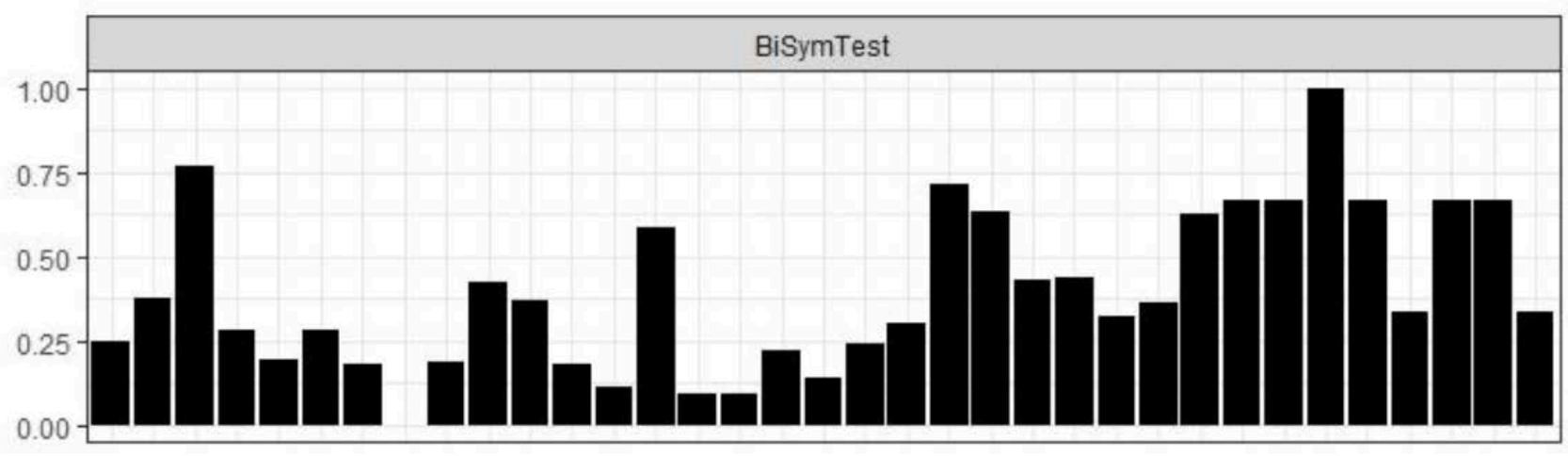
Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG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

Phylogenomic Inference



Species tree of life

“Data-model gap” is increasing!

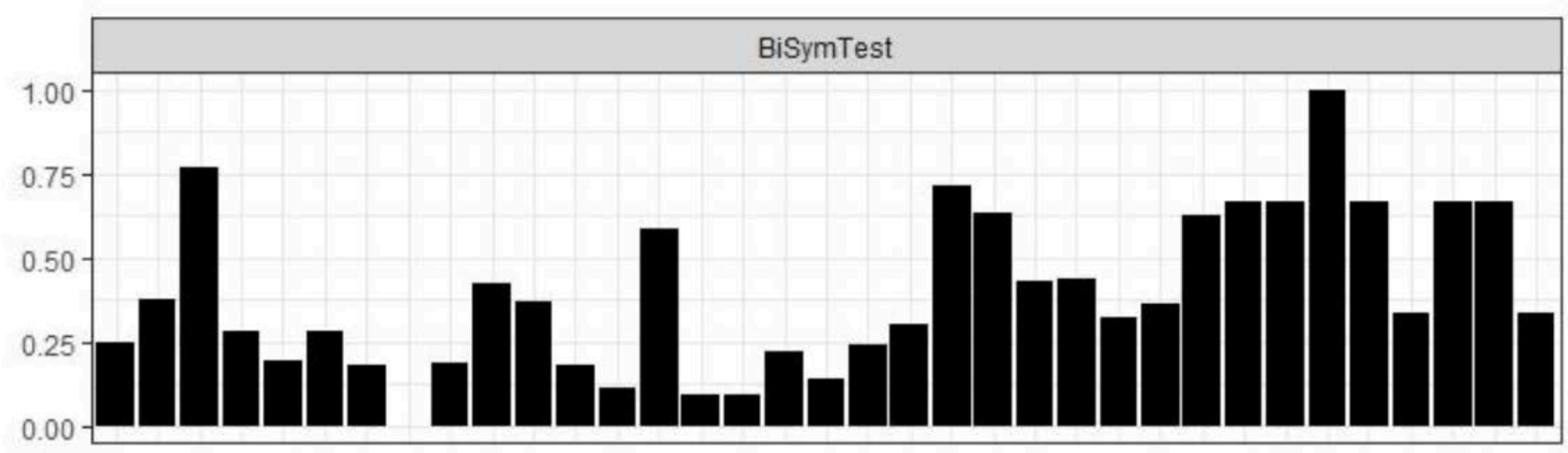


Level of model violations in 35 phylogenomic datasets (https://doi.org/10.1101/460121)

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+R2 … GTR+I+G4

Example partition file (turtle.nex)

```
#nexus
begin sets;
  charset ENSGALG00000000223.macse_DNA_gb = 1-846;
  charset ENSGALG00000001529.macse_DNA_gb = 847-1368;
  charset ENSGALG00000002002.macse_DNA_gb = 1369-2040;
  charset ENSGALG00000002514.macse_DNA_gb = 2041-2772;
  charset ENSGALG00000003337.macse_DNA_gb = 2773-3738;
  charset ENSGALG00000003700.macse_DNA_gb = 3739-4623;
  charset ENSGALG00000003702.macse_DNA_gb = 4624-6168;
  charset ENSGALG00000003907.macse_DNA_gb = 6169-6648;
  charset ENSGALG00000005820.macse_DNA_gb = 6649-7224;
  charset ENSGALG00000005834.macse_DNA_gb = 7225-7920;
  charset ENSGALG00000005902.macse_DNA_gb = 7921-8490;
  charset ENSGALG00000008338.macse_DNA_gb = 8491-9282;
  charset ENSGALG00000008517.macse_DNA_gb = 9283-9822;
  charset ENSGALG00000008916.macse_DNA_gb = 9823-10368;
  charset ENSGALG00000009085.macse_DNA_gb = 10369-11298;
  charset ENSGALG00000009879.macse_DNA_gb = 11299-11895;
  charset ENSGALG00000011323.macse_DNA_gb = 11896-12795;
  charset ENSGALG00000011434.macse_DNA_gb = 12796-13242;
  charset ENSGALG00000011917.macse_DNA_gb = 13243-14223;
  charset ENSGALG00000011966.macse_DNA_gb = 14224-14691;
  charset ENSGALG00000012244.macse_DNA_gb = 14692-15444;
  charset ENSGALG00000012379.macse_DNA_gb = 15445-15963;
  charset ENSGALG00000012568.macse_DNA_gb = 15964-16593;
  charset ENSGALG00000013227.macse_DNA_gb = 16594-17895;
  charset ENSGALG00000014038.macse_DNA_gb = 17896-18456;
  charset ENSGALG00000014648.macse_DNA_gb = 18457-18954;
  charset ENSGALG00000015326.macse_DNA_gb = 18955-19551;
  charset ENSGALG00000015397.macse_DNA_gb = 19552-20145;
  charset ENSGALG00000016241.macse_DNA_gb = 20146-20820;
end;
```

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+R2 … GTR+I+G4

Model of branch lengths	Gene trees
Universally shared	
Proportionally linked	
Unlinked	

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	----------

Model overfitting: Model too complex relative to data
Poor predictive performance

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+I+G4

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.

Relaxed clustering algorithm
(Lanfear et al. 2014):

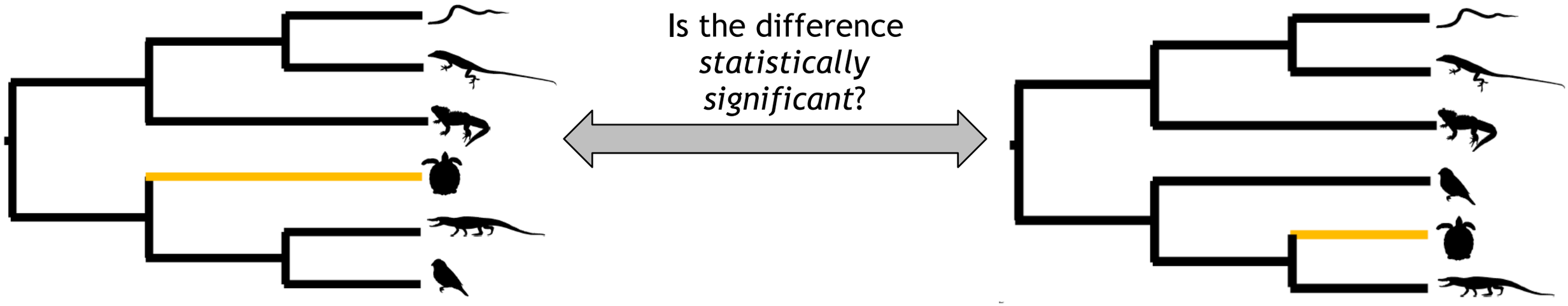
In step 1: only examine the top k% of most "promising" pairs.

Substitution models

HKY

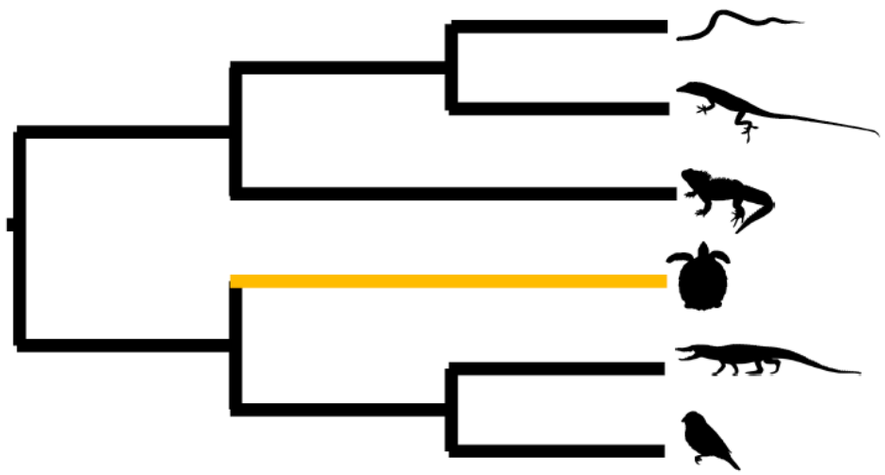
GTR+I+G4

Tree topology tests

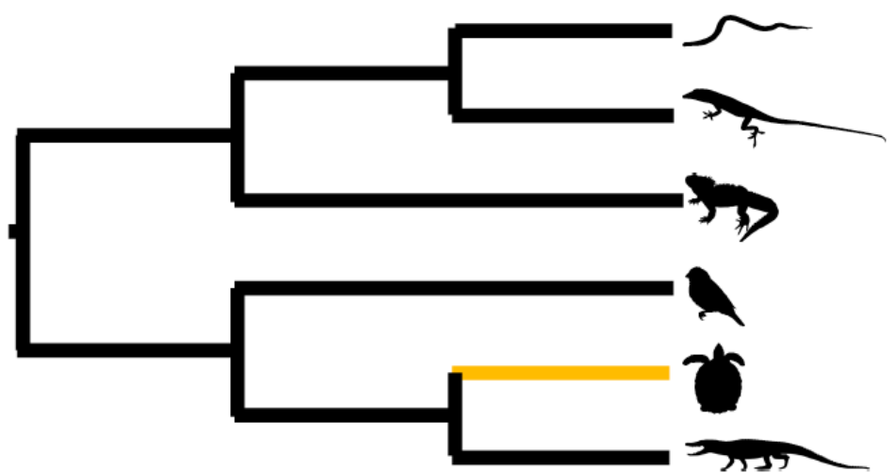


$$\delta = \log\Big(likelihood(T_1)\Big) - \log\Big(likelihood(T_0)\Big)$$

Tree topology tests



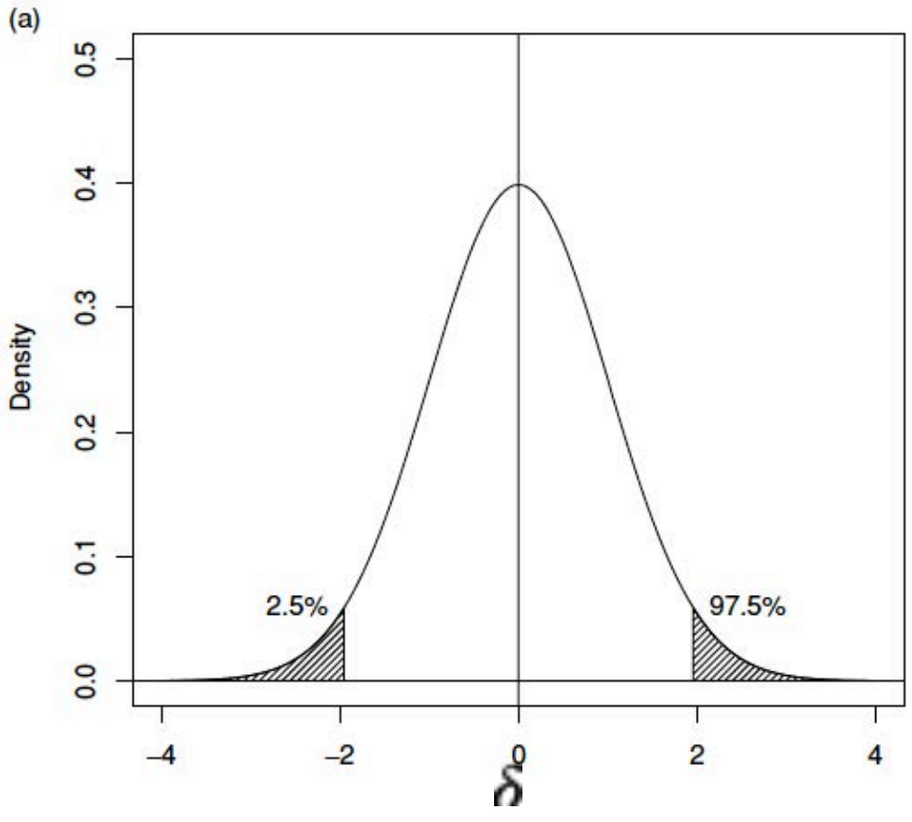
Is the difference *statistically significant*?



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.

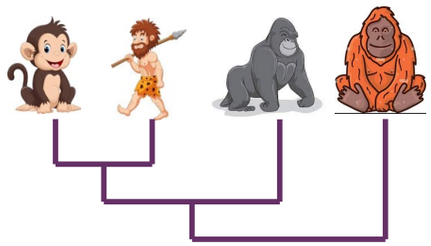


Mixture Across Sites and Trees (MAST) model

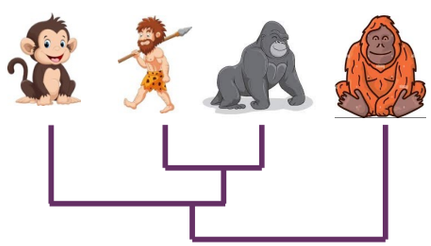
Concatenated alignment

S1:	A	A	–	T	A	A	A	T
S2:	T	A	A	C	C	T	T	T
S3:	T	A	T	A	A	G	T	T
S4:	A	C	–	A	C	A	A	A

L_1^1 L_2^1 L_3^1 L_4^1 L_5^1 L_6^1 L_7^1 L_8^1



L_1^2 L_2^2 L_3^2 L_4^2 L_5^2 L_6^2 L_7^2 L_8^2



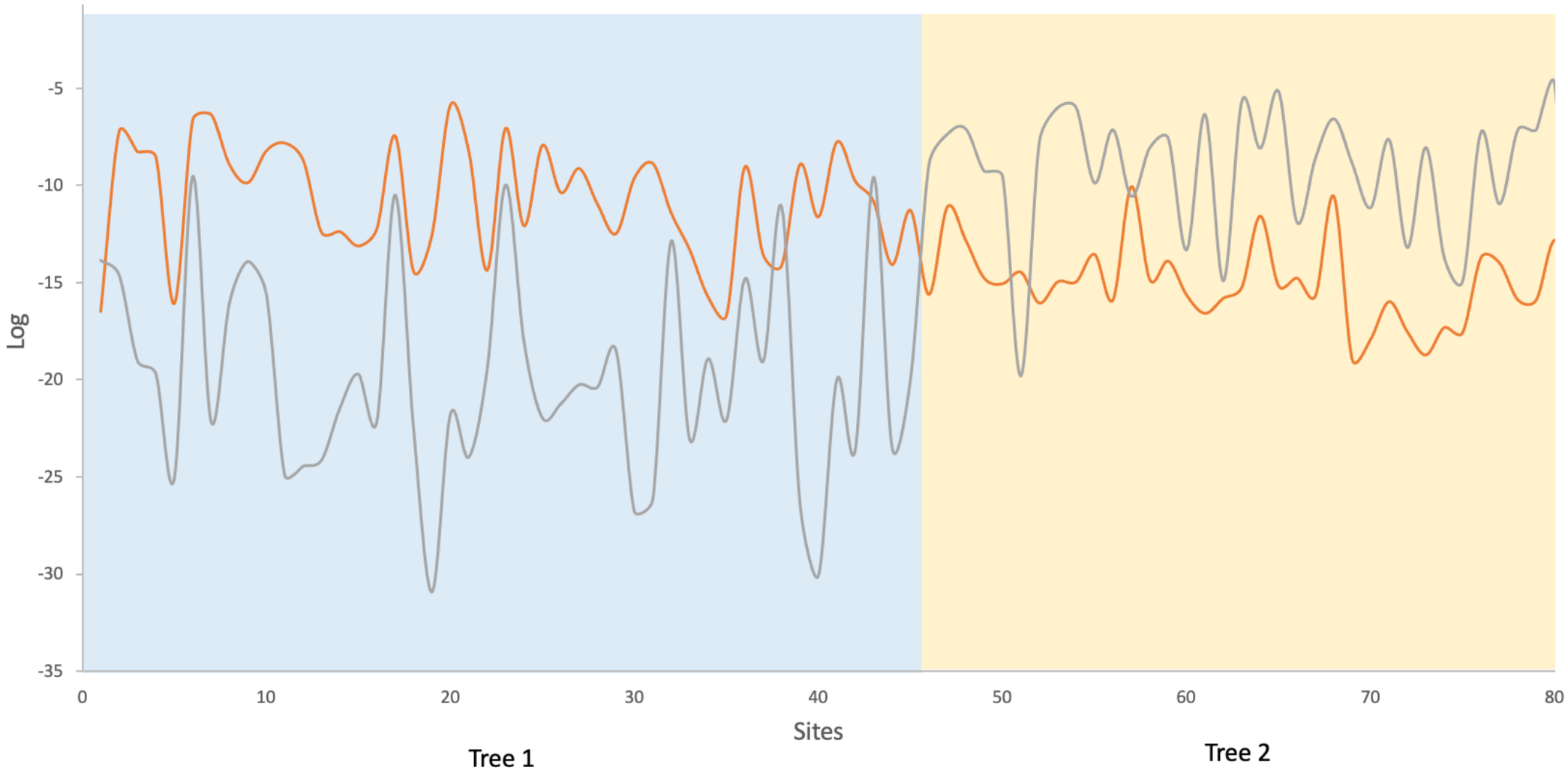
Likelihood for site *i:* $L_i = w_1 L_i^1 + w_1 L_i^2$

where w_j represents the portion of sites belonging to tree j

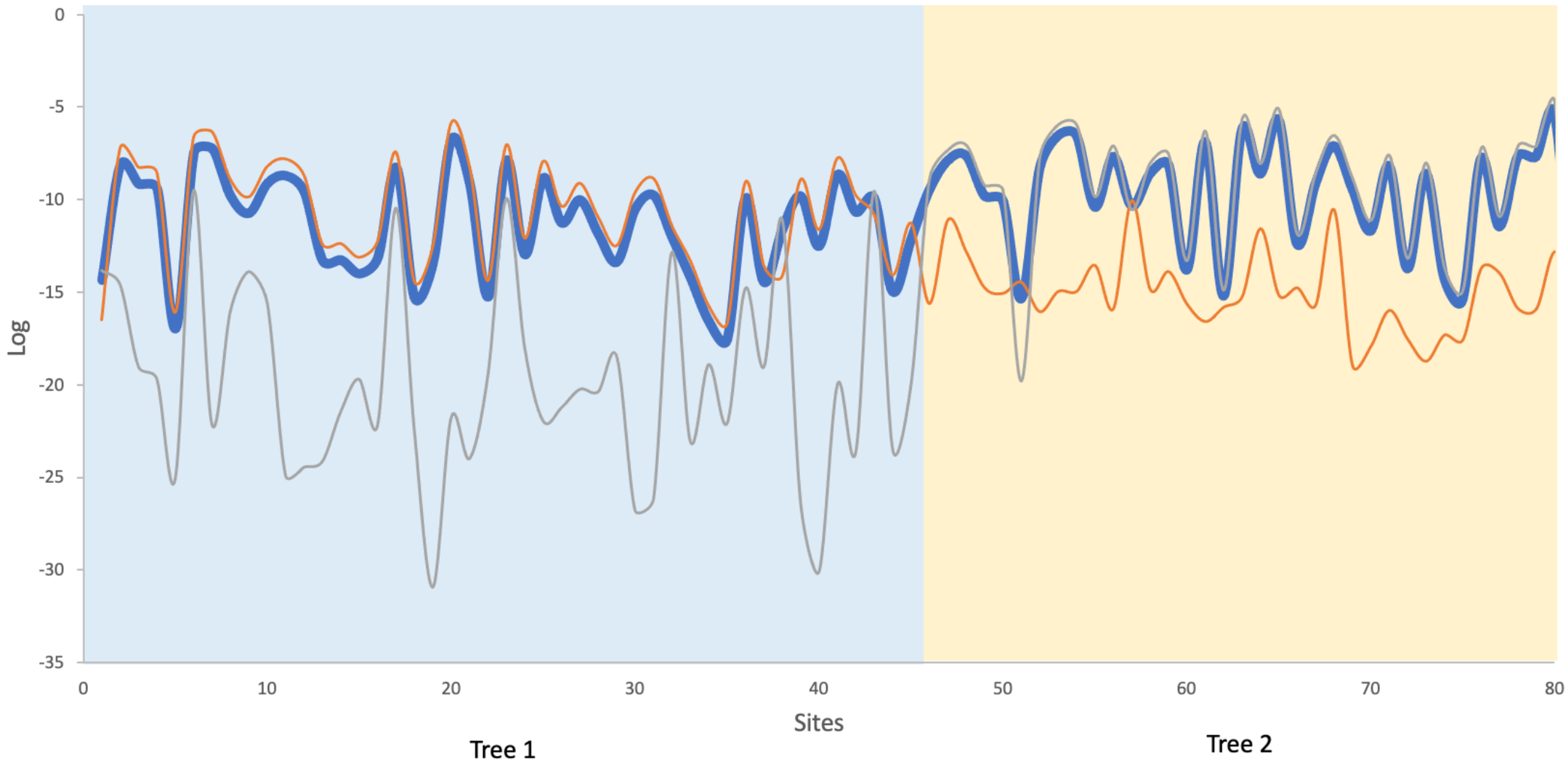
Log-likelihood of the trees: $\sum_i \log(L_i)$

```
iqtree2 -s ALN_FILE -te TREES_FILE -m GTR+G+T
```

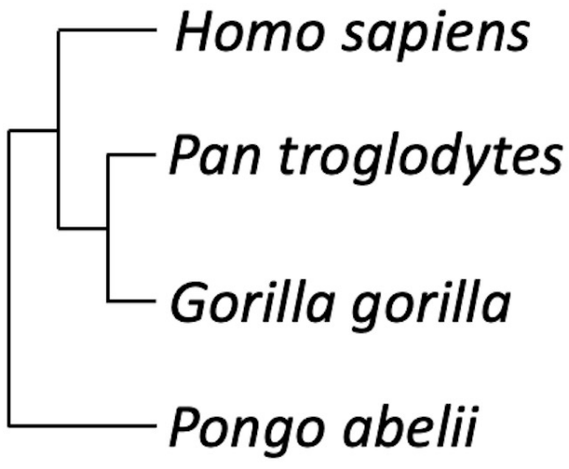
Toy example: Site log-likelihood

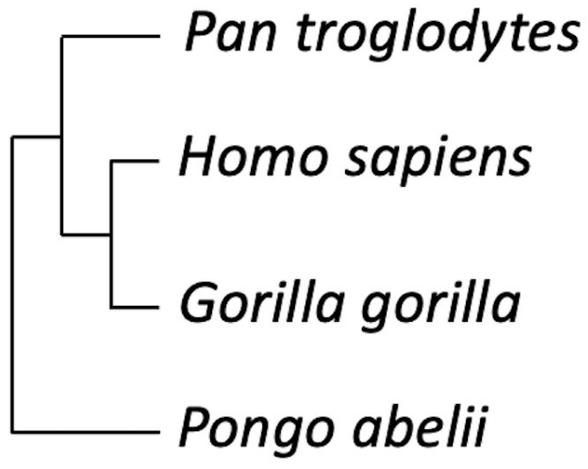


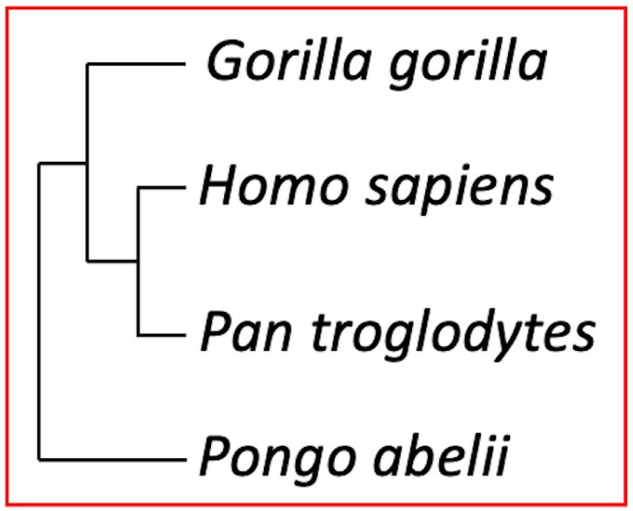
Toy example: Site log-likelihood



The classical example of Human, Chimp, Gorilla



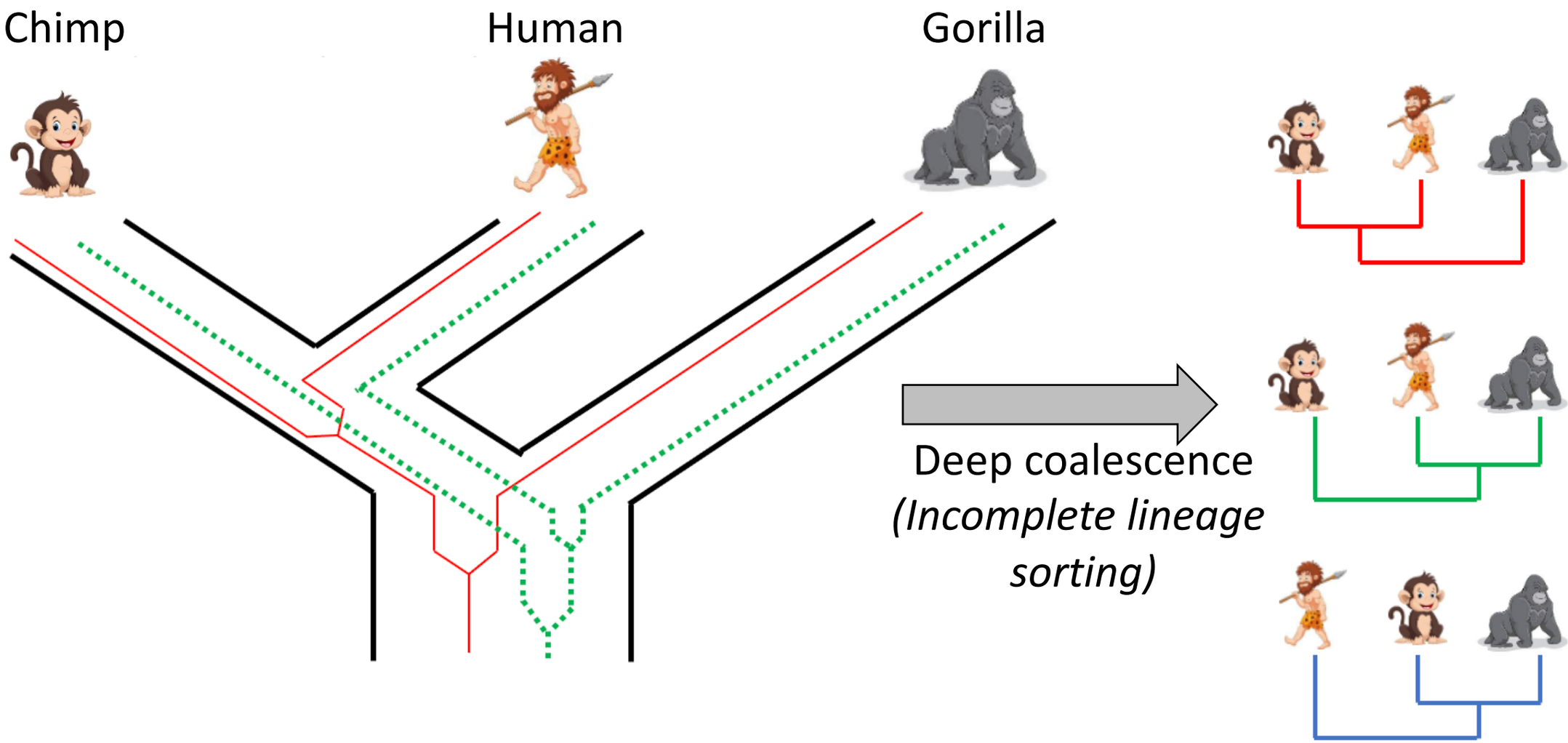




	T_{A1}	T_{A2}	T_{A3}
Gene tree frequencies:	19.8%	20.1%	60.1%
MAST model weights:	**17.9%**	**17.4%**	**64.7%**

Data: 1,595 genes; 1,618,506 bp (Vanderpool et al. 2020)

Gene trees discordance due to deep coalescence

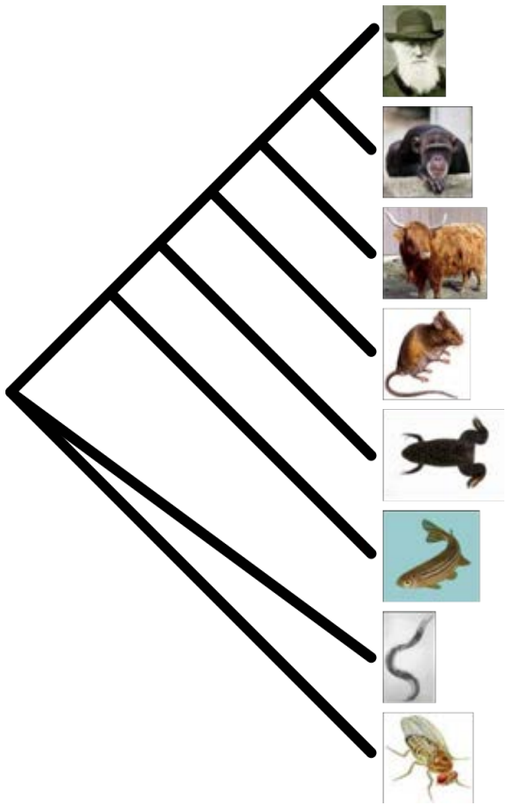


Concatenation methods: Limitation

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

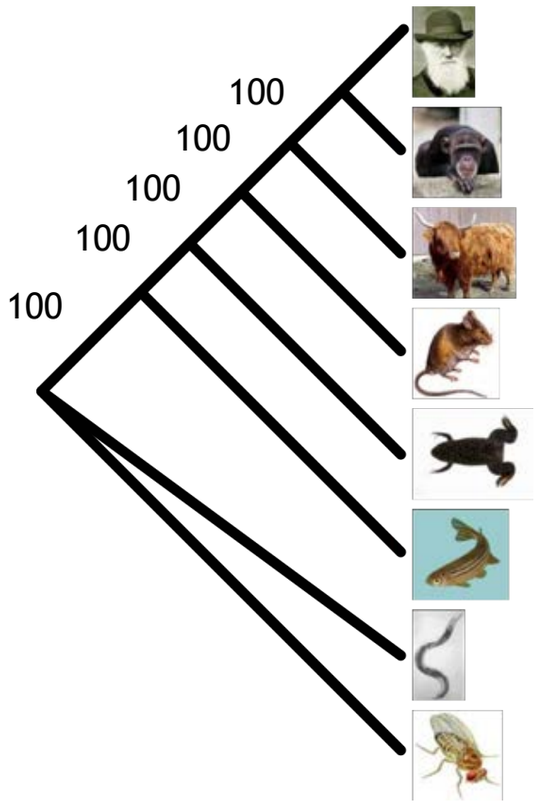
Concatenation methods: Limitation

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

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



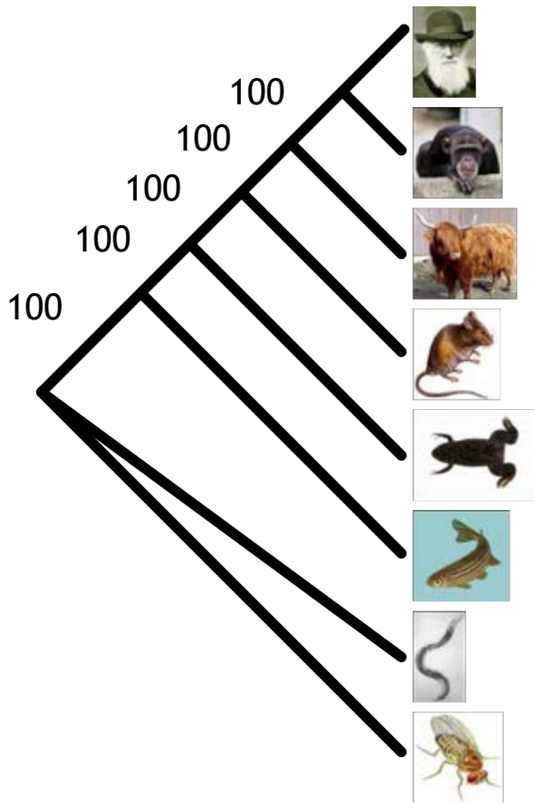
Species tree of life

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 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.

Special Issue

Syst. Biol. 71(4):917–920, 2022

DOI:10.1093/sysbio/syac002
Advance Access publication January 28, 2022

On the Need for New Measures of Phylogenomic Support

ROBERT C. THOMSON[1,*] AND JEREMY M. BROWN[2]

[1]*School of Life Sciences, University of Hawai'i, 2538 McCarthy Mall, Edmondson Hall 216, Honolulu, HI 96822, USA; and* [2]*Department of Biological Sciences and Museum of Natural Science, Louisiana State University, Baton Rouge, LA, USA*
[*]*Correspondence to be sent to: School of Life Sciences, University of Hawai'i, 2538 McCarthy Mall, Edmondson Hall 216, Honolulu, HI 96822, USA; E-mail: thomsonr@hawaii.edu.*

Received 11 November 2021; reviews returned 6 January 2022; accepted 10 January 2022
Associate Editor: Bryan Carstens

Special Issue

Syst. Biol. 71(4):917–920, 2022

DOI:10.1093/sysbio/syac002
Advance Access publication January 28, 2022

On the Need for New Measures of Phylogenomic Support

ROBERT C. THOMSON[1,*] AND JEREMY M. BROWN[2]

[1]*School of Life Sciences, University of Hawai'i, 2538 McCarthy Mall, Edmondson Hall 216, Honolulu, HI 96822, USA;* [2]*Department of Biological*

Syst. Biol. 71(4):921–928, 2022

DOI:10.1093/sysbio/syaa068
Advance Access publication September 11, 2020

An Evolving View of Phylogenetic Support

CHRIS SIMON*

Department of Ecology and Evolutionary Biology, 75 N. Eagleville Road, University of Connecticut, Storrs, CT 06268-3043, USA
**Correspondence to be sent to: Department of Ecology and Evolutionary Biology, 75 N. Eagleville Road, University of Connecticut, Storrs, CT 06268-3043, USA; E-mail: chris.simon@uconn.edu.*

Received 14 February 2020; reviews returned 4 August 2020; accepted 15 August 2020
Associate Editor: Robert Lanfear

Special Issue

Syst. Biol. 71(4):917–920, 2022

DOI:10.1093/sysbio/syac002
Advance Access publication January 28, 2022

On the Need for New Measures of Phylogenomic Support

ROBERT C. THOMSON[1,*] AND JEREMY M. BROWN[2]

[1]*School of Life Sciences, University of Hawai'i at Mānoa, 2538 McCarthy Mall, Edmondson Hall 216, Honolulu, HI 96822, USA;* [2]*Department of Biological*

Syst. Biol. 71(4):921–928, 2022

DOI:10.1093/sysbio/syaa068
Advance Access publication September 11, 2020

An Evolving

Department of Ecology and Evolutionary Biology,
**Correspondence to be sent to: Department of Ecology and Evo*
USA

Received 14 February 2020; r
A

Syst. Biol. 71(4):973–985, 2022

https://doi.org/10.1093/sysbio/syac014
Advance Access publication March 22, 2022

Comparing Likelihood Ratios to Understand Genome-Wide Variation in Phylogenetic Support

GENEVIEVE G. MOUNT[1,2,3,*] AND JEREMY M. BROWN[1]

[1]*Department of Biological Sciences and Museum of Natural Science, Louisiana State University, 202 Life Science Bldg, Baton Rouge, LA 70803, USA;* [2]*Department of Biology, Utah State University, 5305 Old Main Hill, Logan, UT 84322, USA; and* [3]*Museum of Vertebrate Zoology and Department of Integrative Biology, University of California Berkeley, 3101 Valley Life Sciences Building, Berkeley, CA 94720, USA*
**Correspondence to be sent to: Department of Biology, Utah State University, 5305 Old Main Hill, Logan, UT 84322, USA; E-mail: ggmountt@gmail.com.*

Received 26 May 2021; reviews returned 15 February 2022; accepted 22 February 2022
Associate Editor: Lars Jermiin

Special Issue

Syst. Biol. 71(4):917–920, 2022

DOI:10.1093/sysbio/syac002
Advance Access publication January 28, 2022

On the Need for New Measures of Phylogenomic Support

ROBERT C. THOMSON[1,*] AND JEREMY M. BROWN[2]

[1]*School of Life Sciences, University of Hawai'i at Mānoa, 2538 McCarthy Mall, Edmondson Hall 216, Honolulu, HI 96822, USA; and* [2]*Department of Biological*

Syst. Biol. 71(4):921–928, 2022

DOI:10.1093/sysbio/syaa068
Advance Access publication September 11, 2020

An Evolving

Department of Ecology and Evolutionary Biology,
**Correspondence to be sent to: Department of Ecology and Evo*
USA

Received 14 February 2020; r
A

Syst. Biol. 71(4):973–985, 2022

https://doi.org/10.1093/sysbio/syac014
Advance Access publication March 22, 2022

Comparing Likelihood Ratios to Understand Genome-Wide Variation in Phylogenetic Support

GENEVIEVE G. MOUNT[1,2,3,*] AND JEREMY M. BROWN[1]

ty, 202 Life Science Bldg, Baton Rouge, LA 70803, USA;
and [3]*Museum of Vertebrate Zoology and Department of*
ces Building, Berkeley, CA 94720, USA
05 Old Main Hill, Logan, UT 84322, USA;

cepted 22 February 2022

Syst. Biol. 71(4):929–942, 2022

DOI:10.1093/sysbio/syab008
Advance Access publication February 9, 2021

Gene Tree Discord, Simplex Plots, and Statistical Tests under the Coalescent

ELIZABETH S. ALLMAN[1], JONATHAN D. MITCHELL[1,2], AND JOHN A. RHODES[1,*]

[1]*Department of Mathematics and Statistics, University of Alaska Fairbanks, Fairbanks, AK 99709, USA; and* [2] *Unité Bioinformatique Evolutive, C3BI USR 3756, Institut Pasteur & CNRS, Paris, France*
**Correspondence to be sent to: Department of Mathematics and Statistics, University of Alaska Fairbanks, Fairbanks, AK 99709, USA; E-mail: j.rhodes@alaska.edu.*

Received 12 February 2020; reviews returned 31 January 2021; accepted 03 February 2021
Associate Editor: Robert Thomson

New Methods to Calculate Concordance Factors for Phylogenomic Datasets

Bui Quang Minh ,[1,2] Matthew W. Hahn,[3,4] and Robert Lanfear*[,2]

[1]Research School of Computer Science, Australian National University, Canberra, ACT, Australia

[2]Department of Ecology and Evolution, Research School of Biology, Australian National University, Canberra, ACT, Australia

[3]Department of Biology, Indiana University, Bloomington, IN

[4]Department of Computer Science, Indiana University, Bloomington, IN

Correspondence to: *Corresponding author: E-mail: rob.lanfear@anu.edu.au.

Associate editor: Michael Rosenberg

Coalescent/reconciliation methods

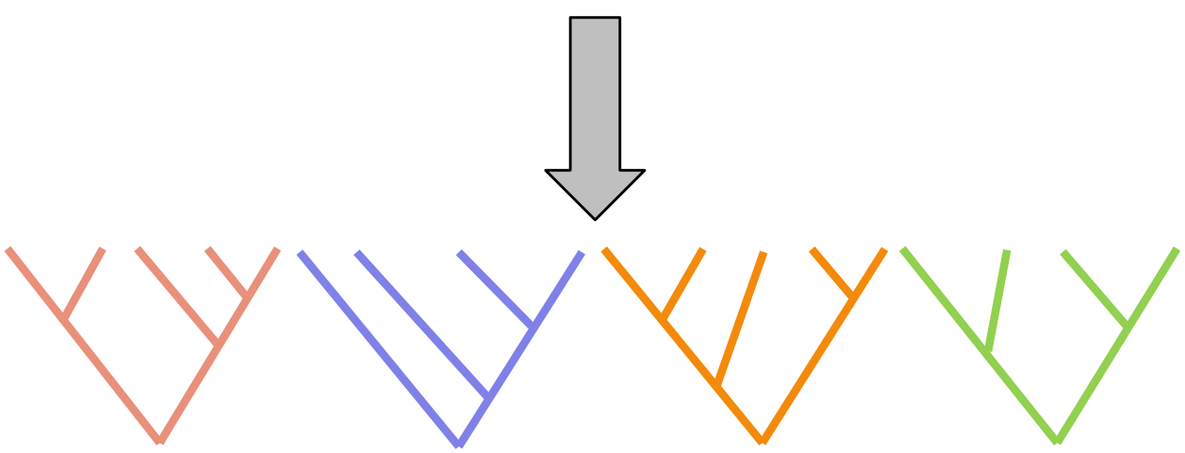
Supermatrix

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	----------

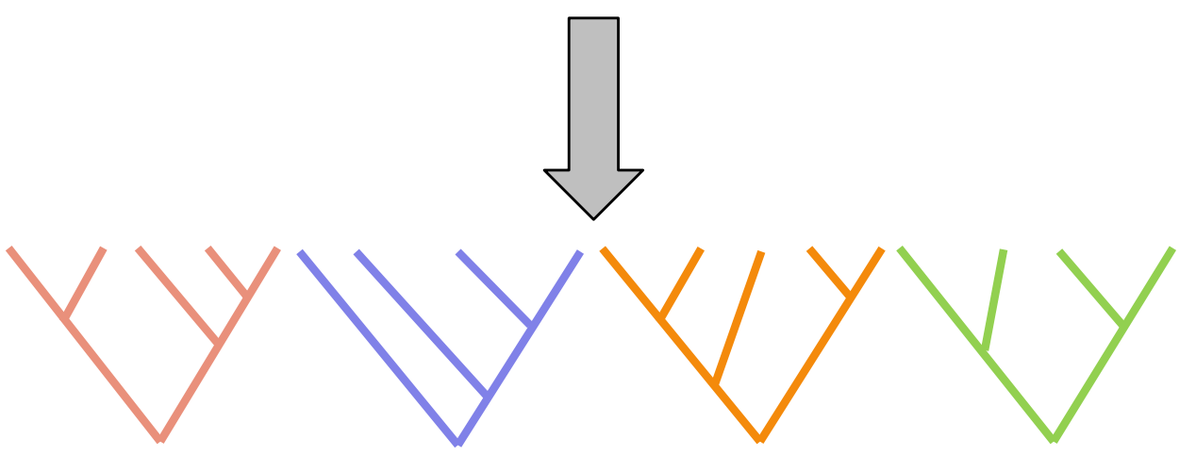


Gene tree 1 Gene tree 2 ..... Gene tree 1,000

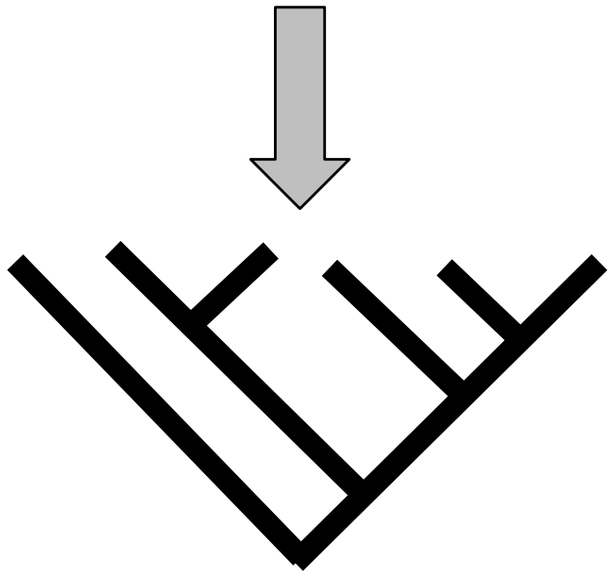
Coalescent/reconciliation methods

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

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

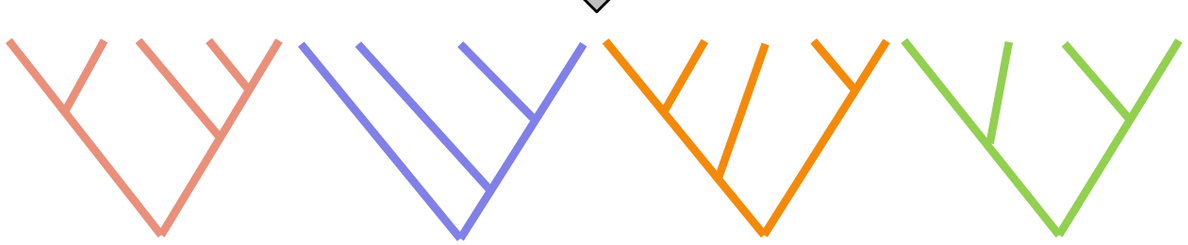
$$\mathbf{gCF(x)} = \frac{\{\, i : T_i \text{ is concordant with x} \,\}}{\{\, i : T_i \text{ is decisive for x} \,\}}$$

i = a gene
T_i = a gene tree
x = an internal branch in the species tree

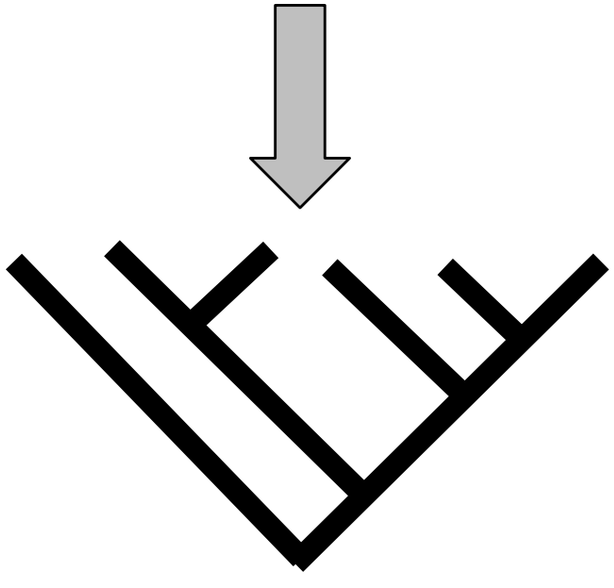
Coalescent/reconciliation methods

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%

Implementation in IQ-TREE fully accounts for missing data

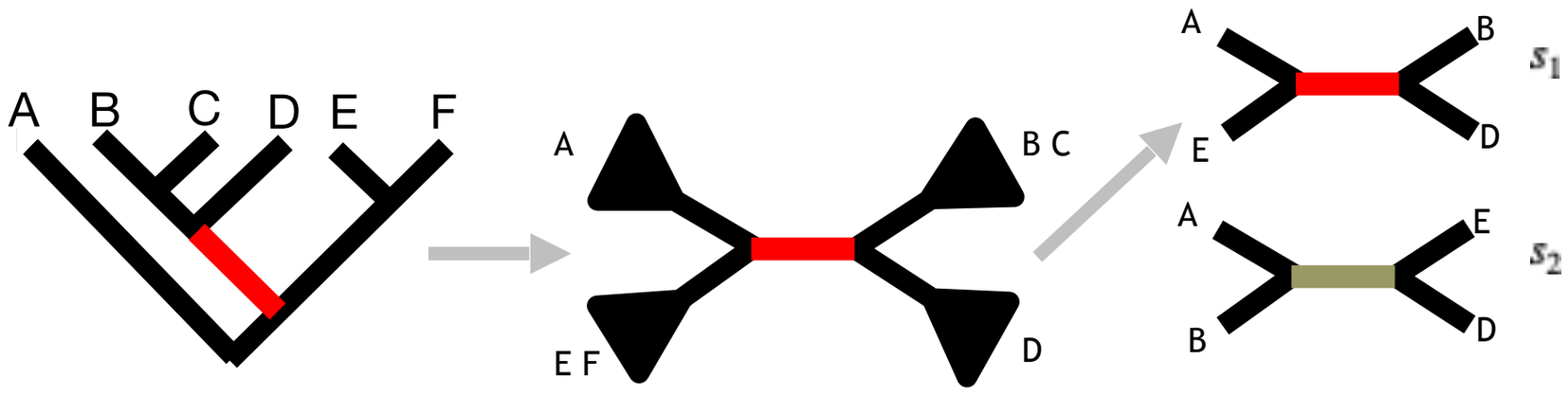
Problem: Uncertainties in gene trees!

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%



$$\text{sCF(x)} = \frac{1}{m}\left[\sum \frac{\{\, j : j \text{ is concordant with x} \,\}}{\{\, j : j \text{ is decisive for x} \,\}}\right]$$

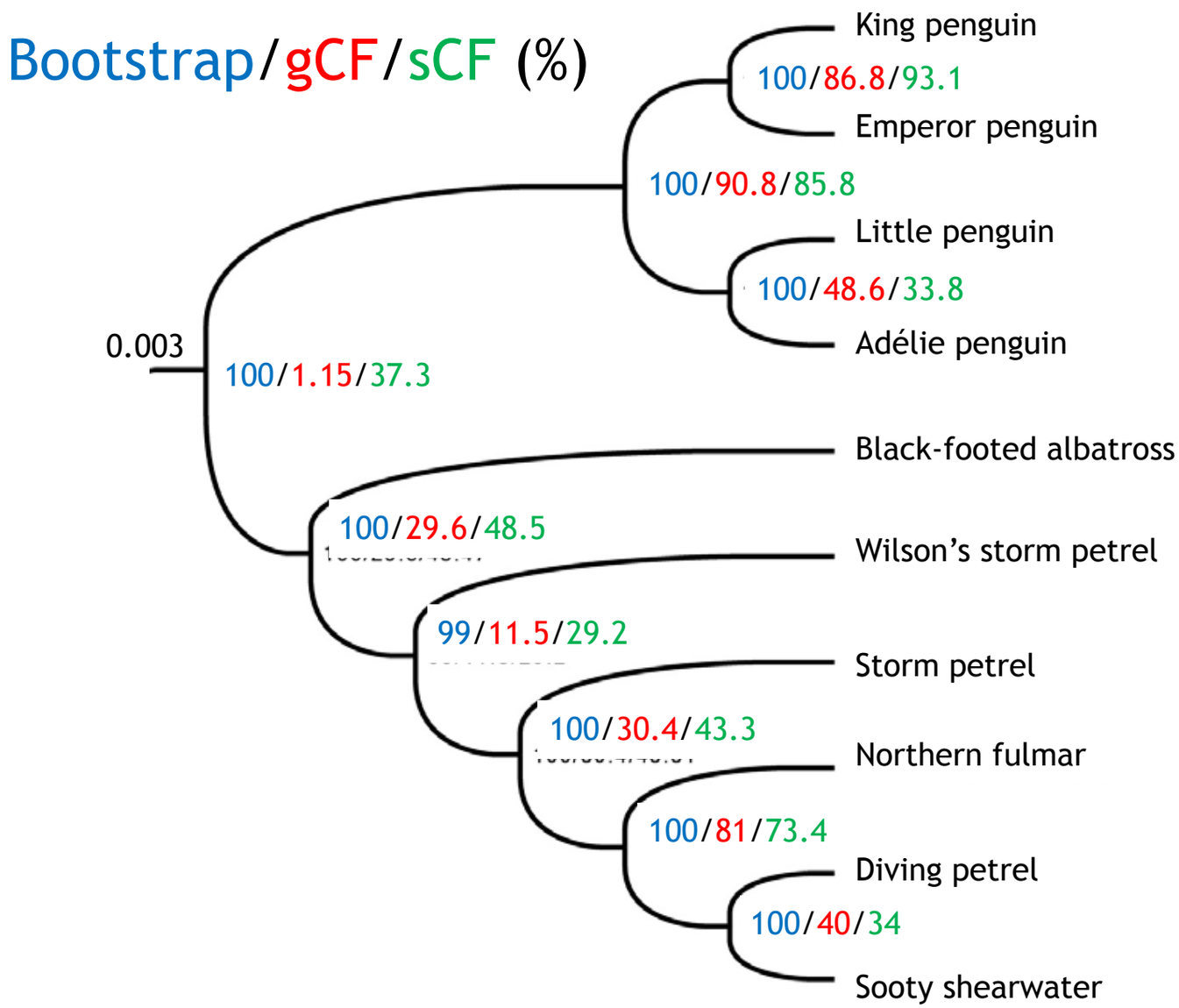
$$qCF(quartet) = \frac{s_1}{s_1 + s_2 + s_3}$$

j = a site
x = an internal branch in the species tree
m = num of quartets to sample

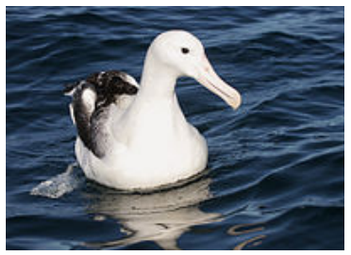
sCF(x) is the ***mean*** qCF(x) over m random quartets

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

88 genes



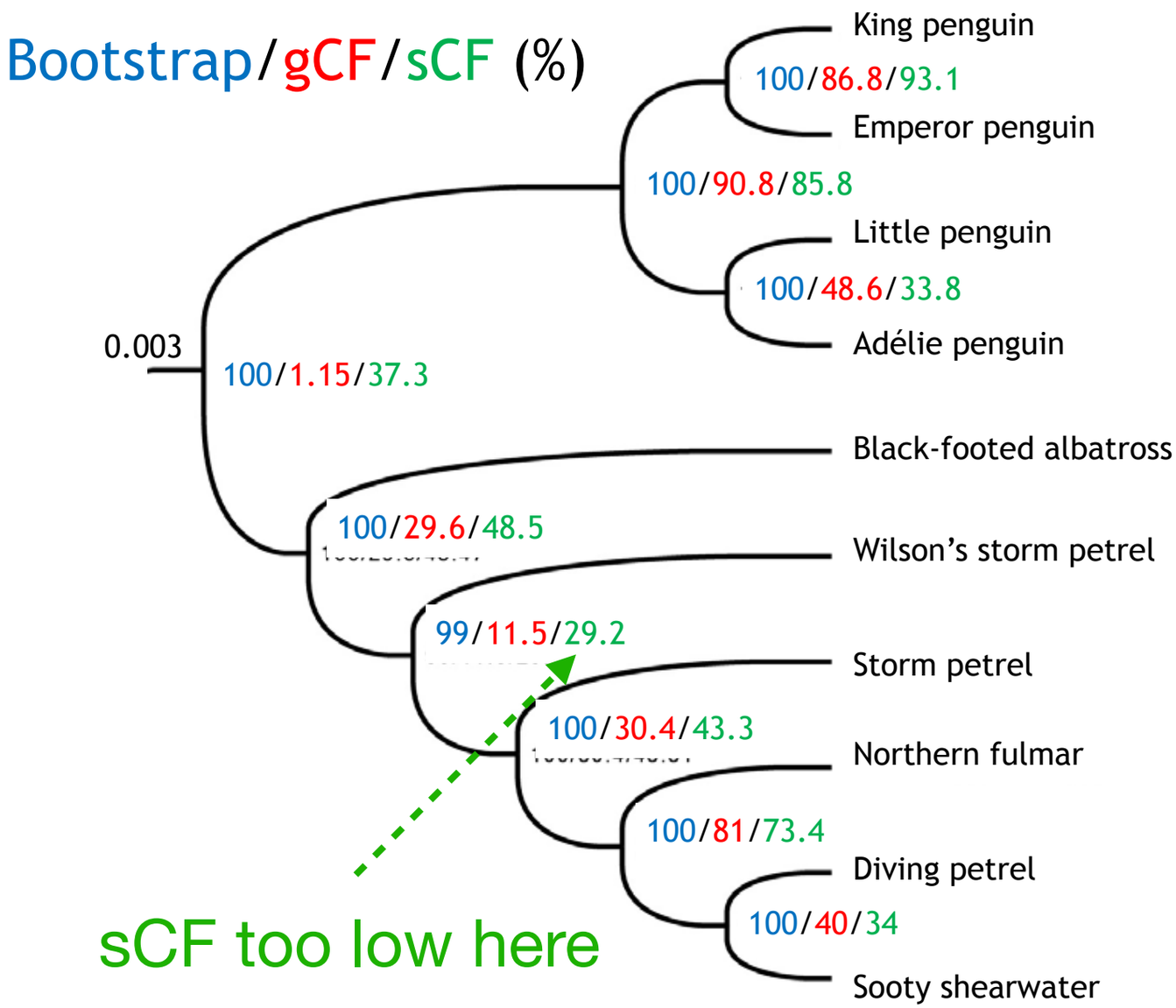
Penguins



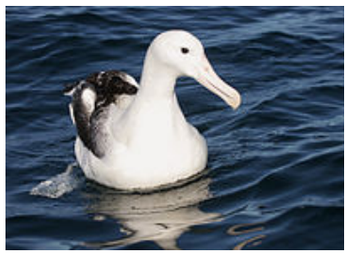
Tubenoses

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

88 genes



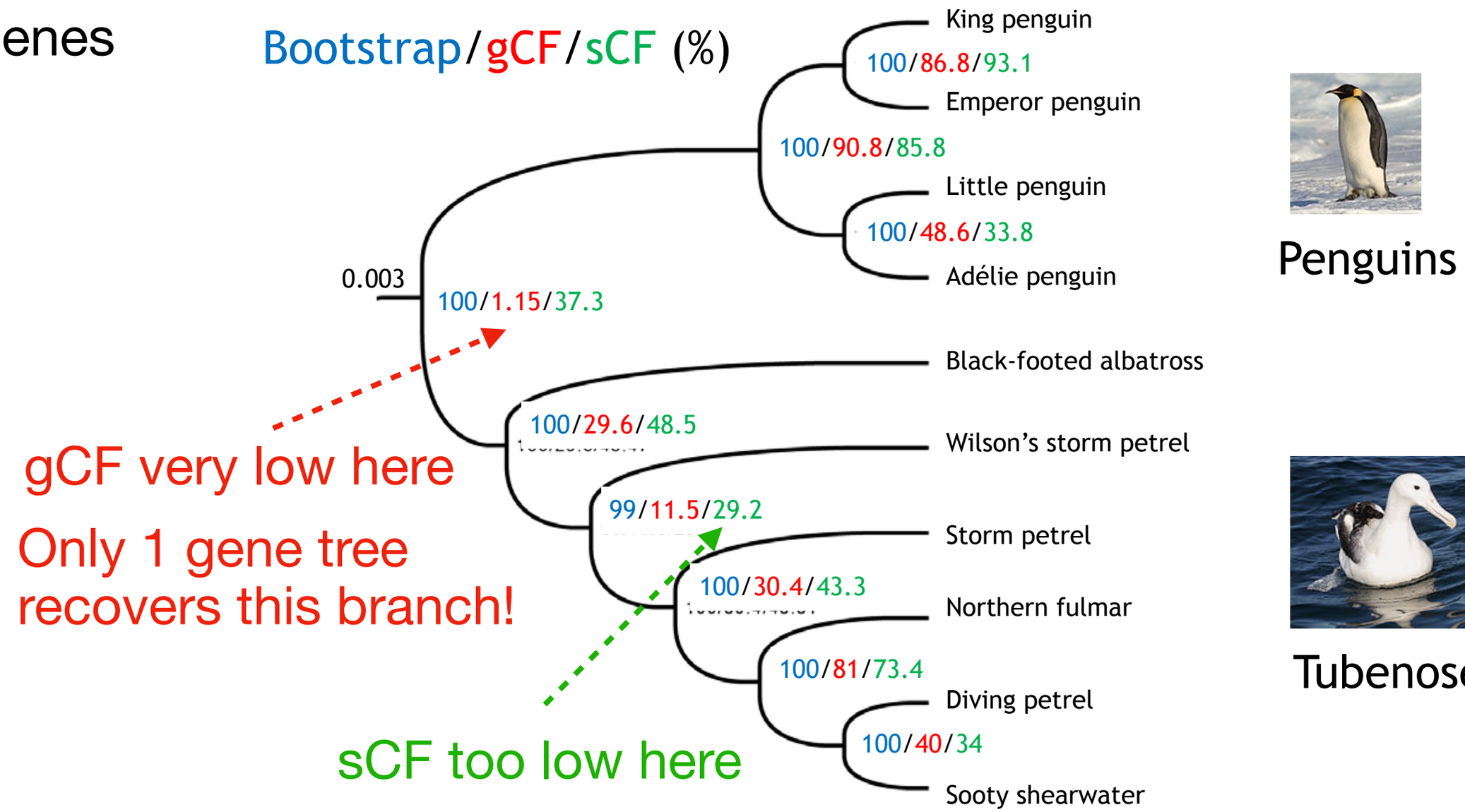
Penguins



Tubenoses

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

88 genes

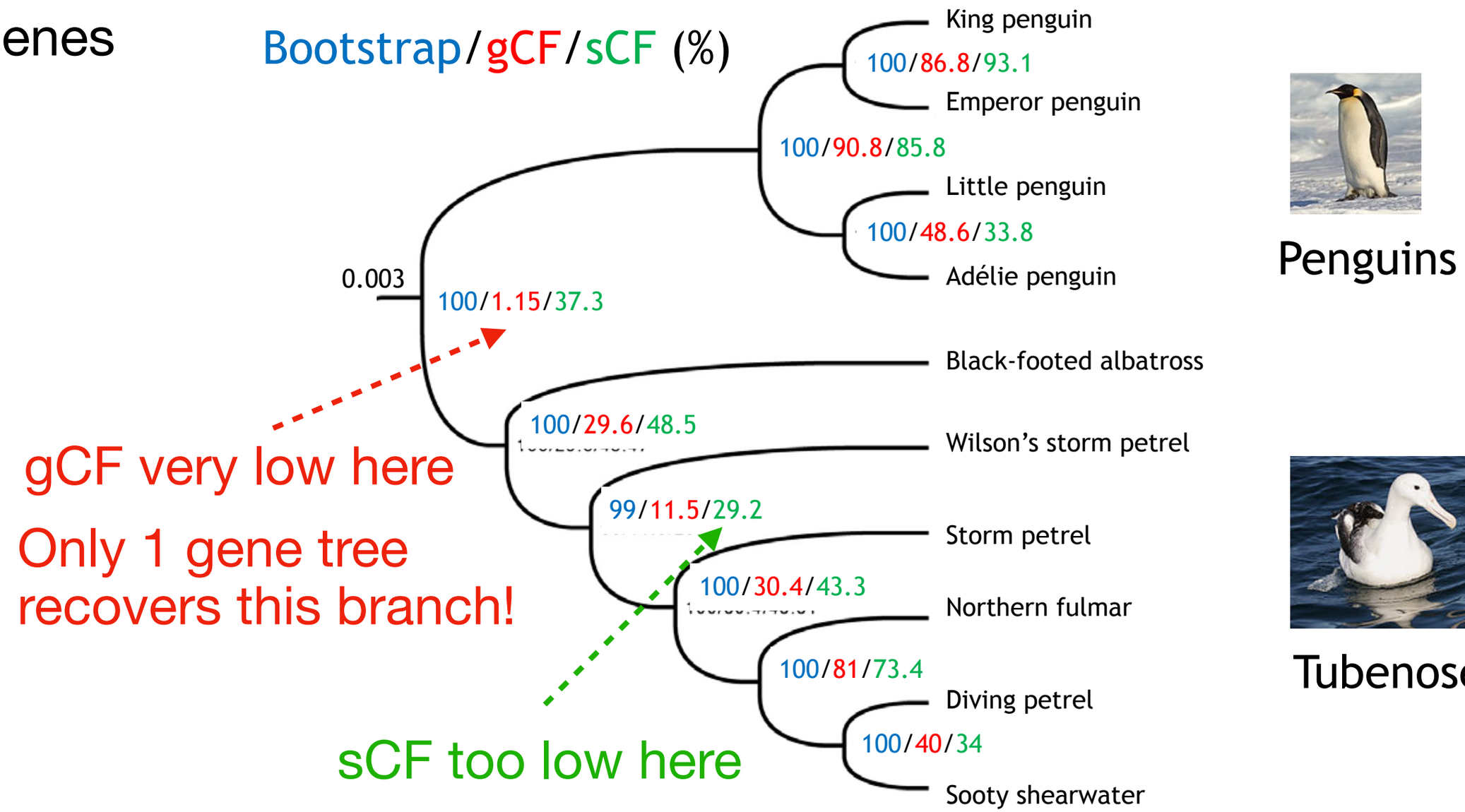


Penguins

Tubenoses

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

88 genes

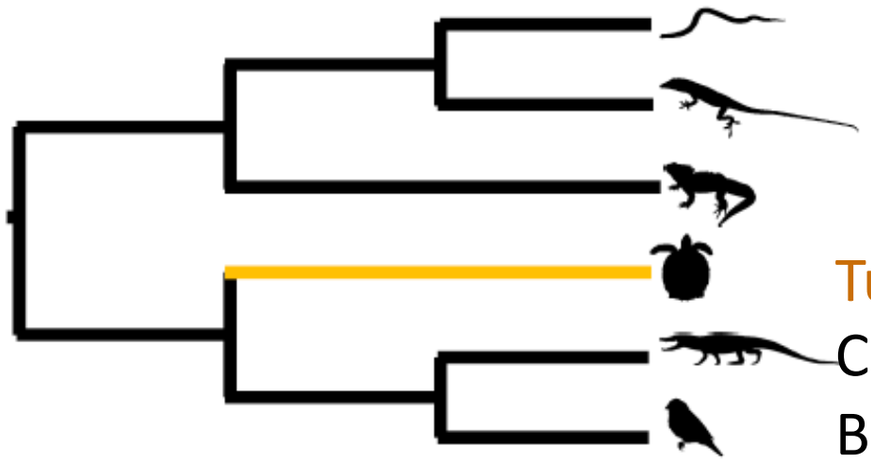


Penguins

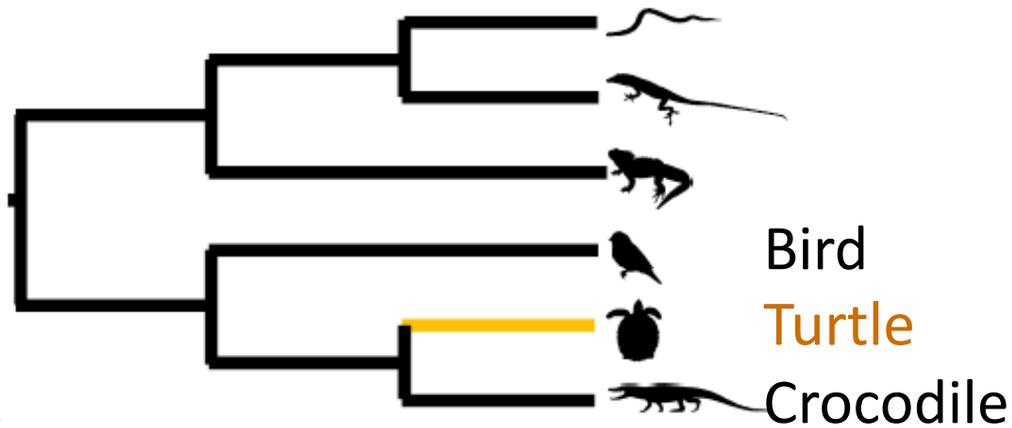
Tubenoses

Generally:
Use Caution when gCF ≈ 0, or sCF ≤ 33.3%, even with 100% BS prop.
Feel good when gCF and sCF ≥ 50%

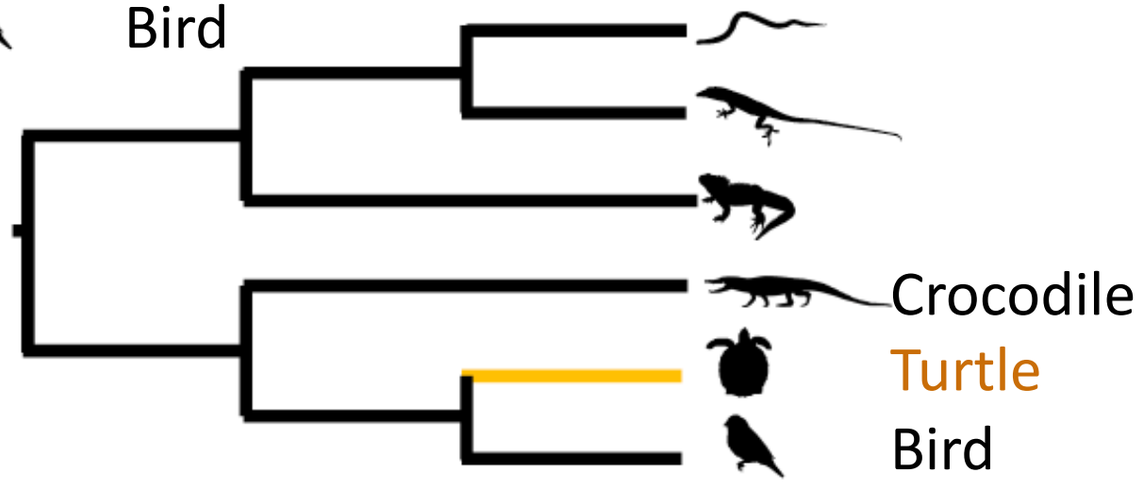
Dataset for IQ-TREE lab: Where is Turtle in the tree?



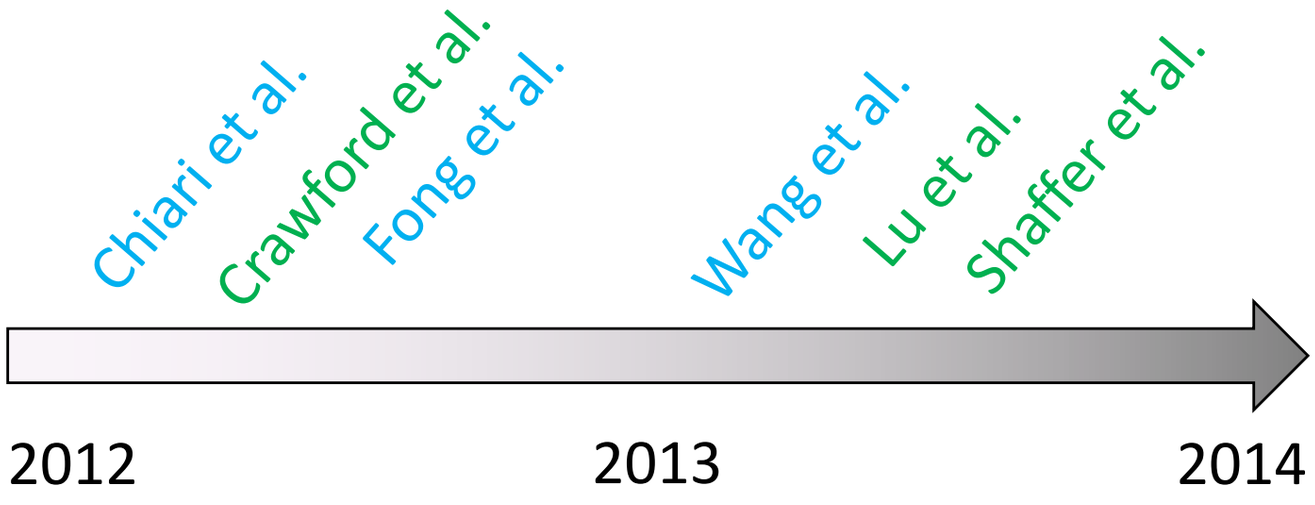
Turtle
Crocodile
Bird



Bird
Turtle
Crocodile



Crocodile
Turtle
Bird



Different studies led to different trees!

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

Thanks Jeremy Brown

IQ-TREE lab

1. Input Data
2. Inferring the first phylogeny
3. Applying a partition model
4. Choosing the best partitioning scheme
5. Tree Topology Tests
6. Tree Mixture Model
7. Identifying the most influential genes
8. Removing influential genes
9. Concordance factors

Link to Lab on course website

Link to “quiz” on course website