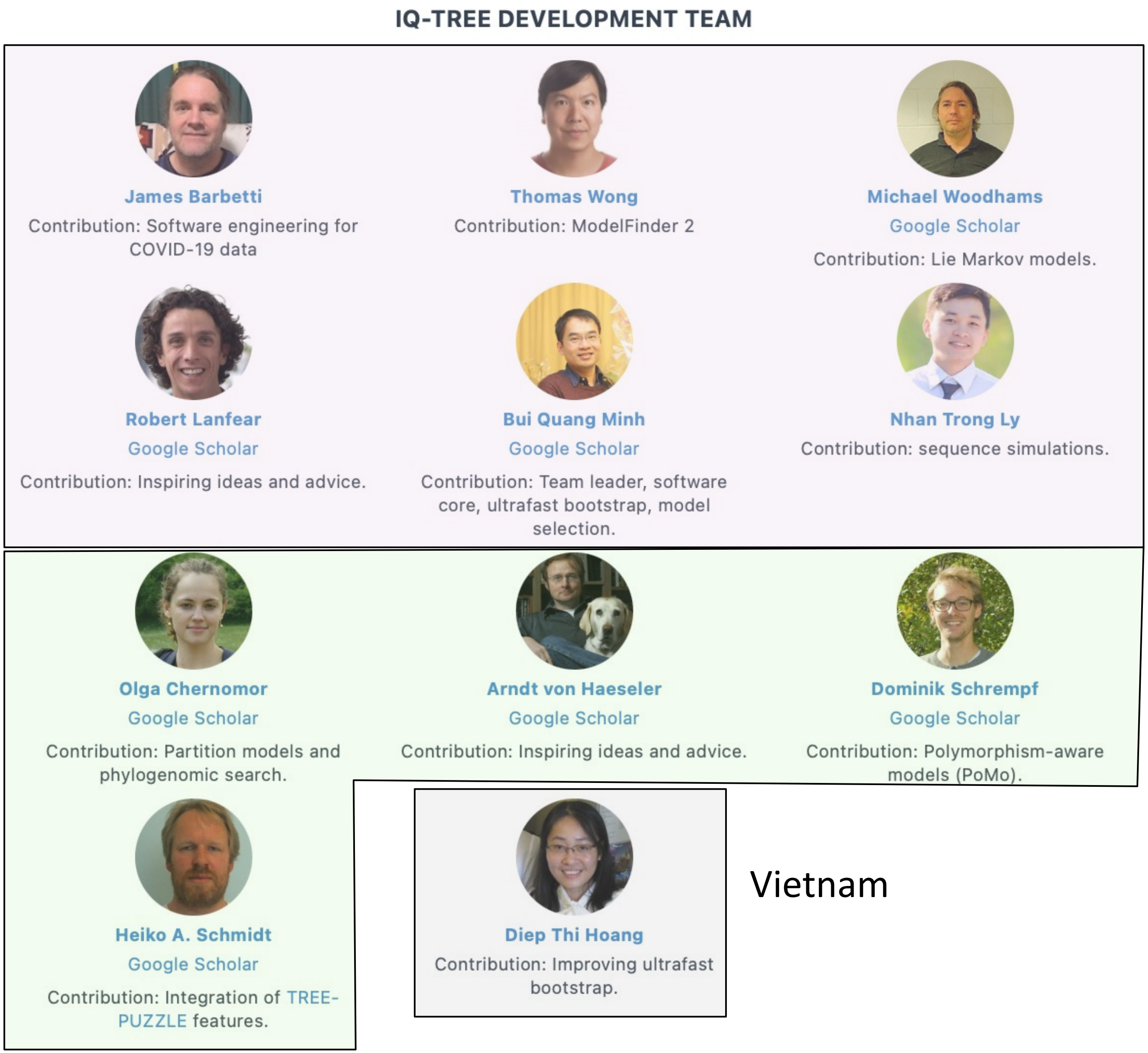
IQ-TREE

http://www.iqtree.org

Methods and Practice

Minh Bui
Australian National University

Workshop on Molecular Evolution
Woods Hole, June 2022



IQ-TREE DEVELOPMENT TEAM

Australia

James Barbetti

Contribution: Software engineering for COVID-19 data

Thomas Wong

Contribution: ModelFinder 2

Michael Woodhams

Google Scholar

Contribution: Lie Markov models.

Robert Lanfear

Google Scholar

Contribution: Inspiring ideas and advice.

Bui Quang Minh

Google Scholar

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

Nhan Trong Ly

Contribution: sequence simulations.

Austria

Olga Chernomor

Google Scholar

Contribution: Partition models and phylogenomic search.

Arndt von Haeseler

Google Scholar

Contribution: Inspiring ideas and advice.

Dominik Schrempf

Google Scholar

Contribution: Polymorphism-aware models (PoMo).

Heiko A. Schmidt

Google Scholar

Contribution: Integration of TREE-PUZZLE features.

Vietnam

Diep Thi Hoang

Contribution: Improving ultrafast bootstrap.

Thanks to plenty of users for feedback and bug reports!

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.

Typical phylogenetic analysis under maximum likelihood

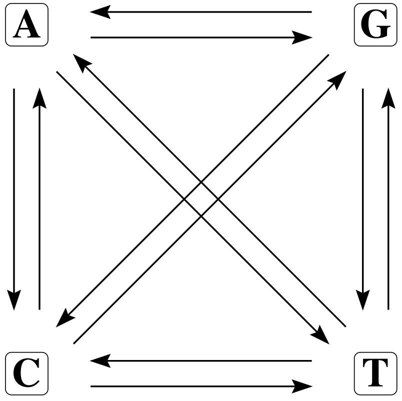
Multiple sequence alignment

```
ACGGGAT--C--C----CATTAC
ACGGGAT--C--C----CACTAC
CCGGGATAGCTTC----CATTAC
ACCCCCTATC--CACTGGATTAC
ACGACATATC--CACTGGATTCC
```

Model selection

ModelFinder (2017)

Substitution model

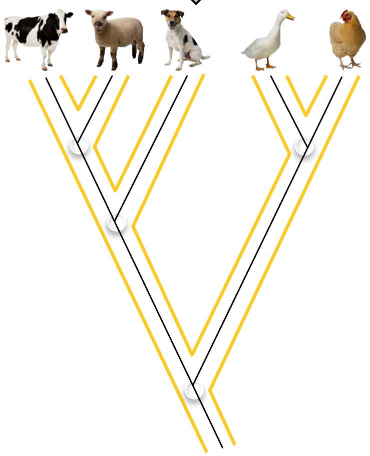


My work focused on improving all three steps for large datasets!

IQ-TREE (2015, 2020)

Tree reconstruction

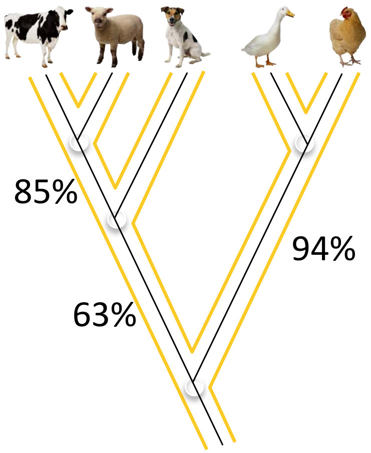
```
iqtree2 -s ALN_FILE -B 1000
```



Phylogenetic tree

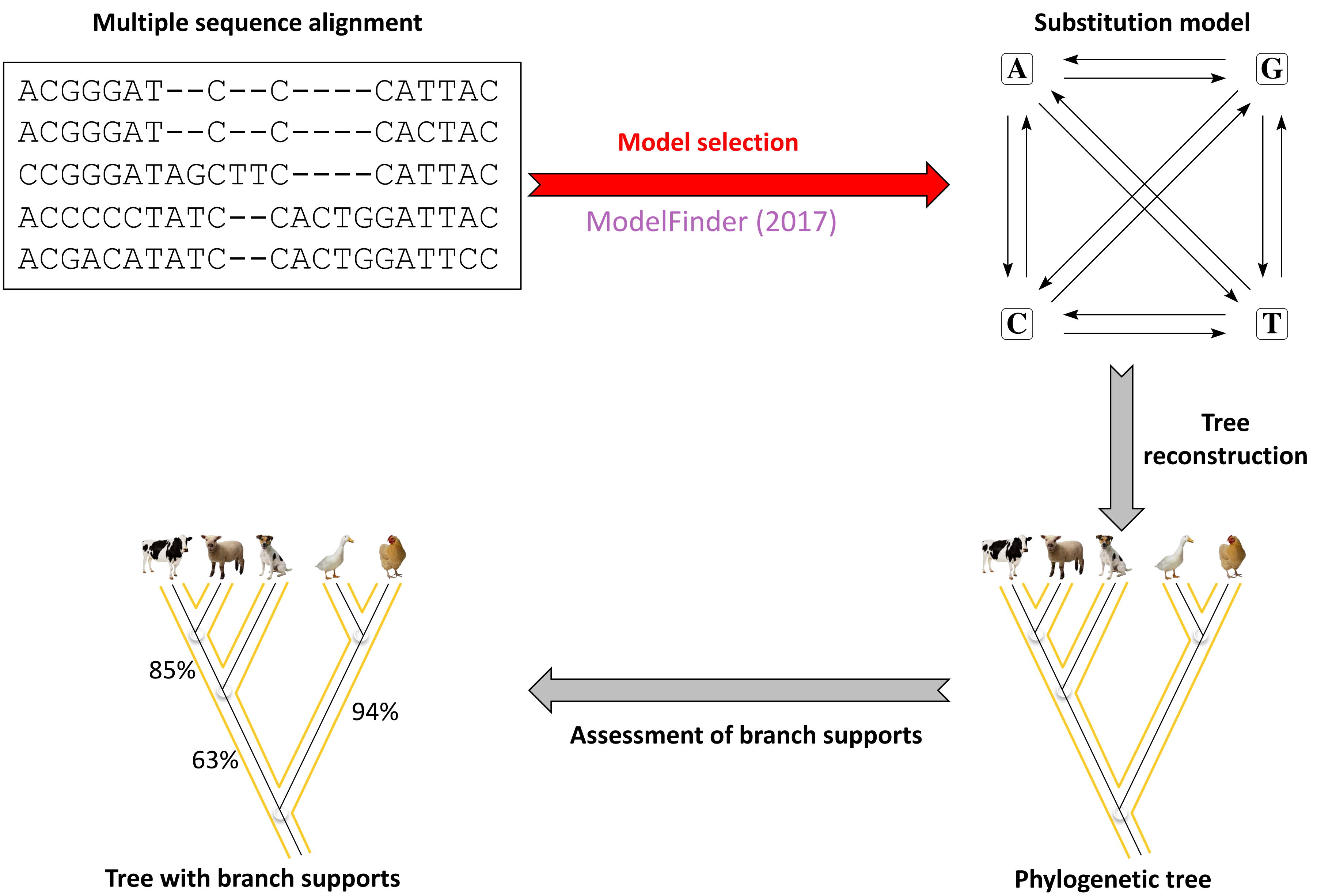
Ultrafast bootstrap (2013, 2018)

Assessment of branch supports



Tree with branch supports

Step 1: Model selection



Multiple sequence alignment

```
ACGGGAT--C--C----CATTAC
ACGGGAT--C--C----CACTAC
CCGGGATAGCTTC----CATTAC
ACCCCCTATC--CACTGGATTAC
ACGACATATC--CACTGGATTCC
```

Model selection

ModelFinder (2017)

Substitution model

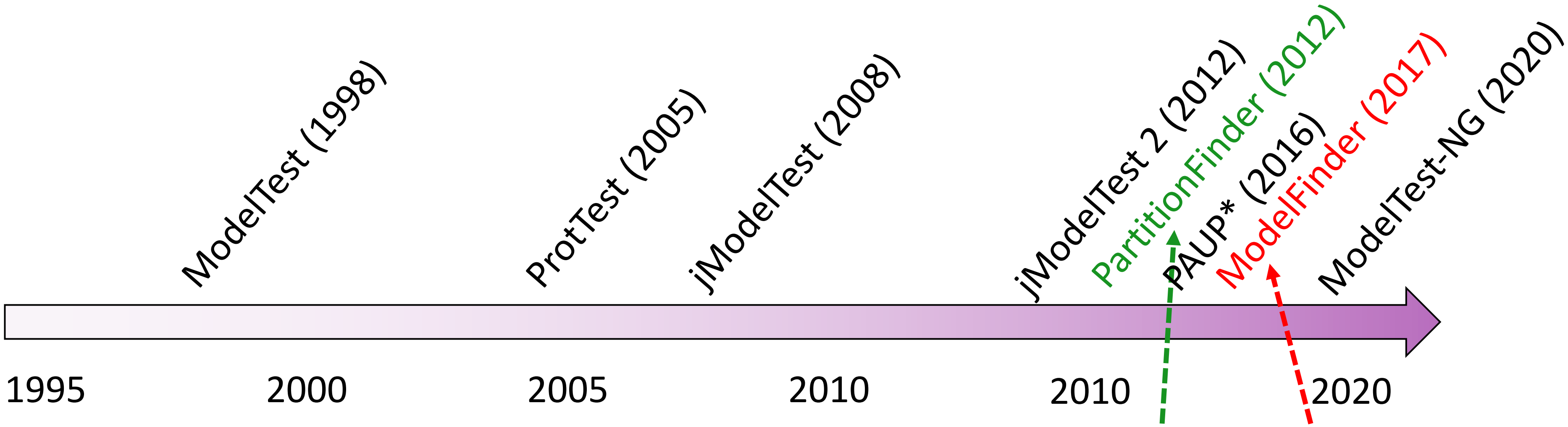
Tree reconstruction

Phylogenetic tree

Assessment of branch supports

Tree with branch supports

Model selection approaches



Robert Lanfear
(ANU)

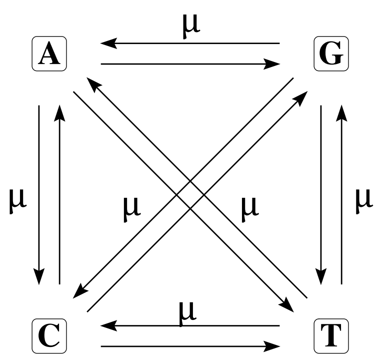
Lars Jermiin
(ANU & CSIRO)

Thomas Wong
(ANU)

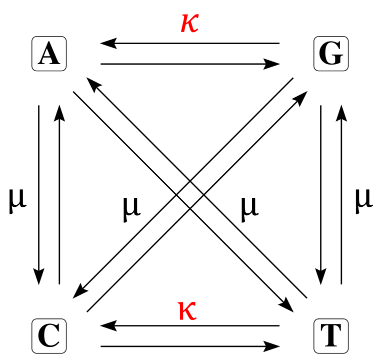
- (j)Modeltest / ProtTest: slow and limited on models.
- PartitionFinder: better models for genomic data but still slow.
- ModelFinder: >10x faster and more realistic models.
- Current work: ModelFinder 2 = ModelFinder + PartitionFinder

(https://www.nature.com/articles/nmeth.4285)

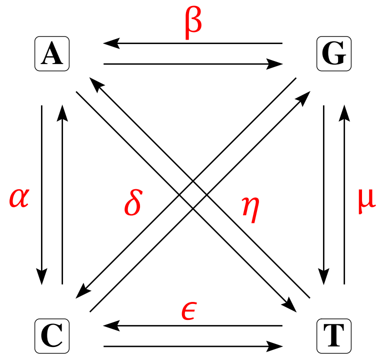
Models of sequence evolution



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”

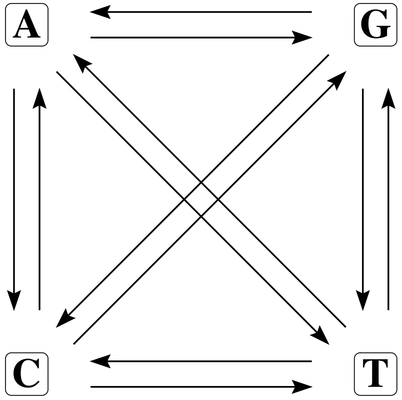
Step 2: Tree reconstruction

Multiple sequence alignment

```
ACGGGAT--C--C----CATTAC
ACGGGAT--C--C----CACTAC
CCGGGATAGCTTC----CATTAC
ACCCCCTATC--CACTGGATTAC
ACGACATATC--CACTGGATTCC
```

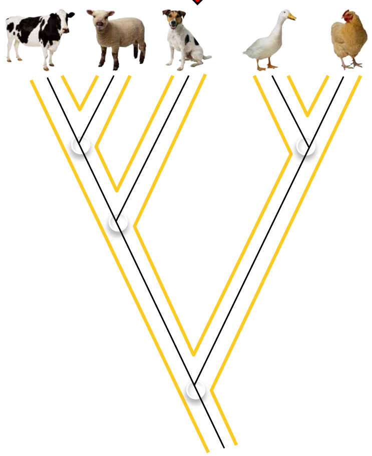
Model selection

Substitution model



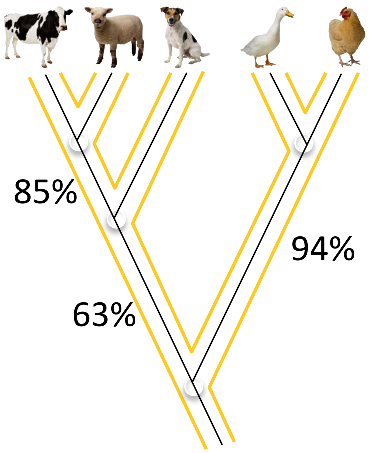
IQ-TREE (2015, 2020)

Tree reconstruction



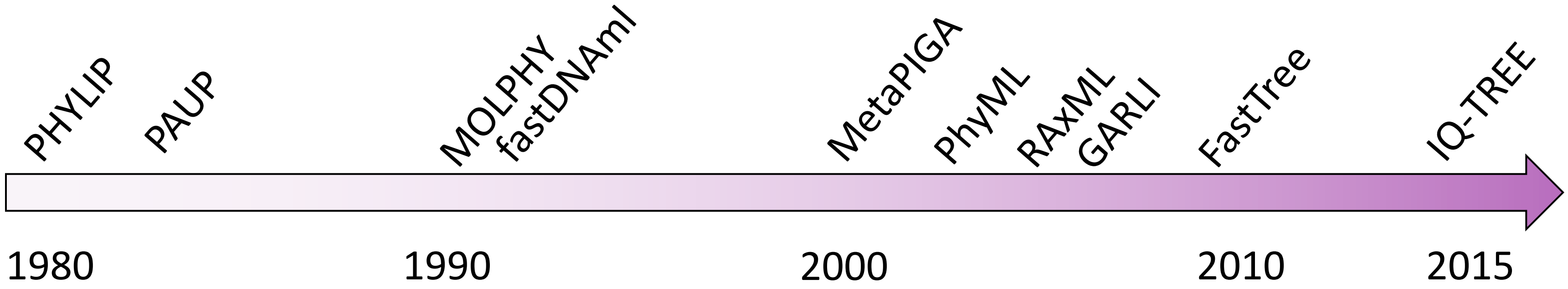
Phylogenetic tree

Assessment of branch supports

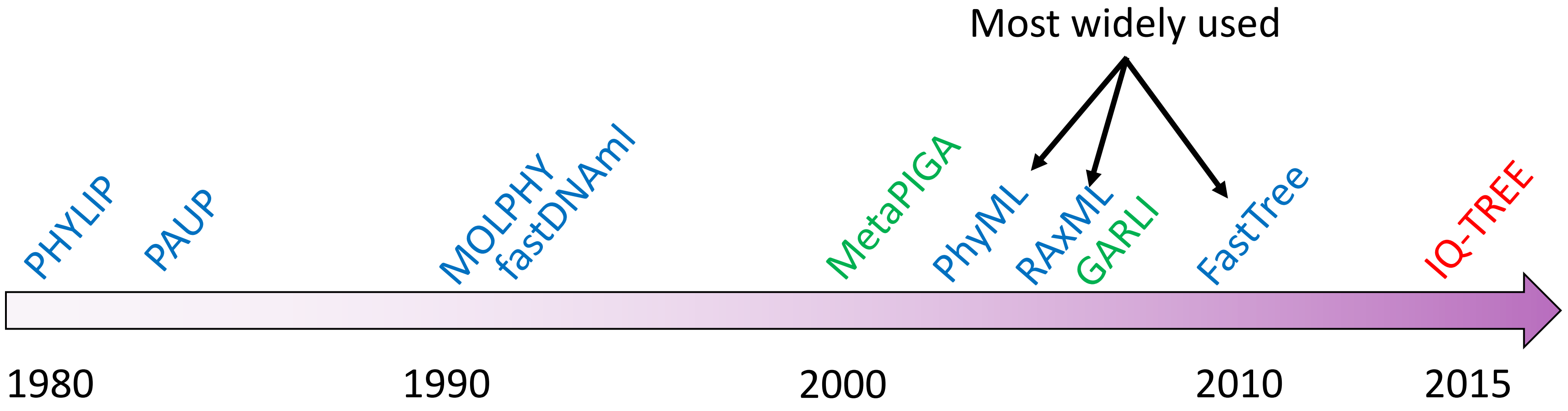


Tree with branch supports

Search heuristics for finding maximum likelihood trees



Search heuristics for finding maximum likelihood trees

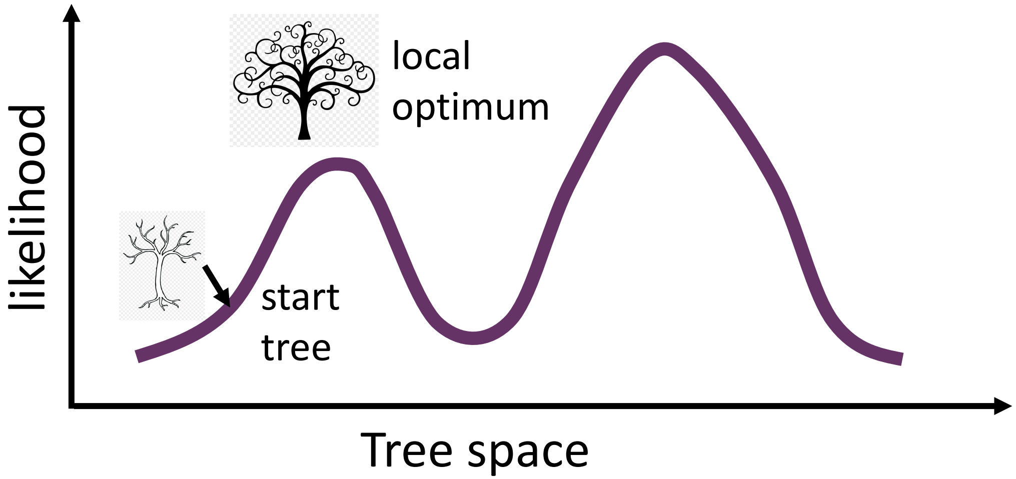


Most widely used

PHYLIP
PAUP
MOLPHY
fastDNAml
MetaPIGA
PhyML
RAxML
GARLI
FastTree
IQ-TREE

1980
1990
2000
2010
2015

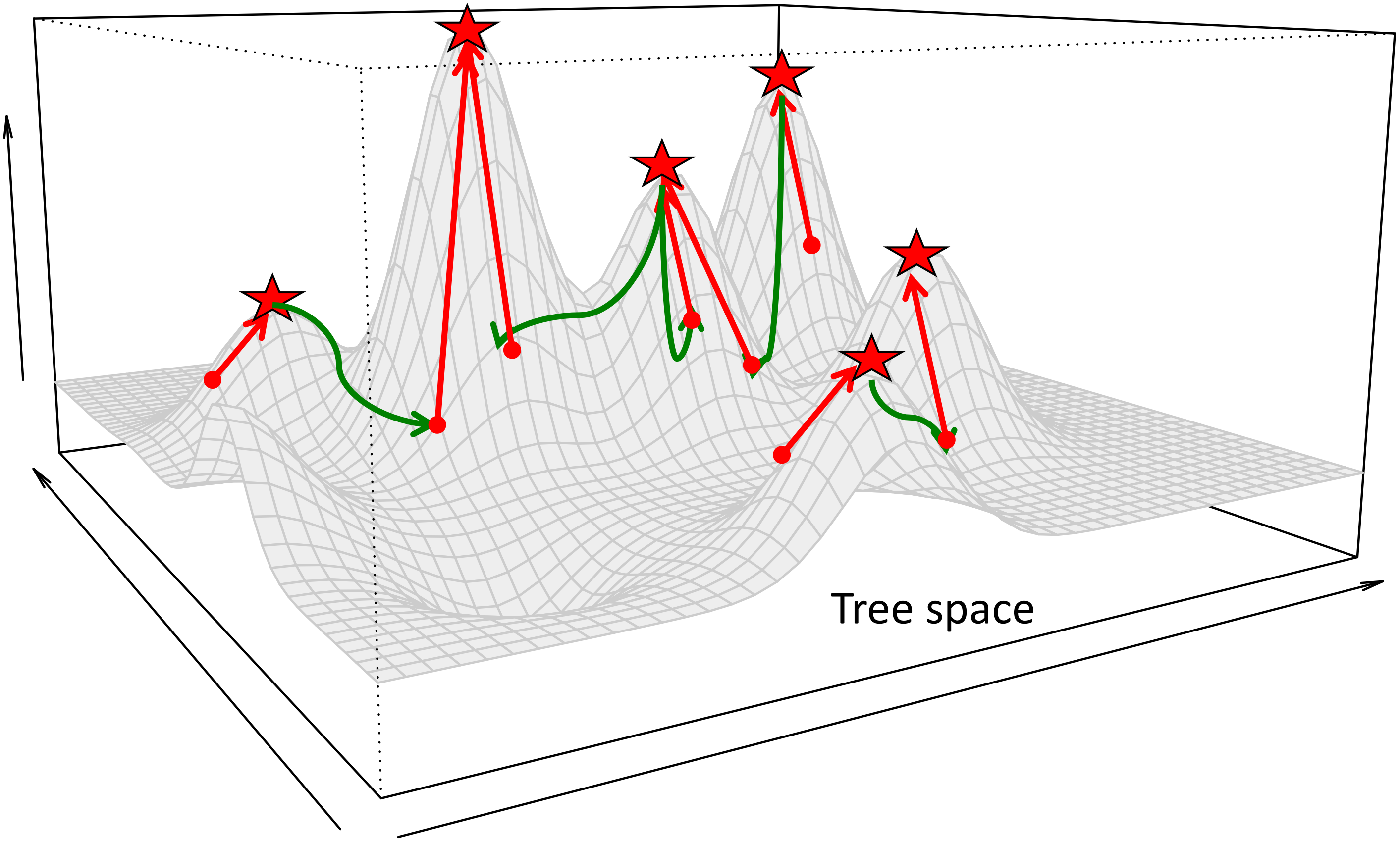
1. Hill-climbing / greedy algorithms: Fast but local optimum
2. Genetic algorithm: Slow but escaping local optima
3. IQ-TREE: Fast and escaping local optima

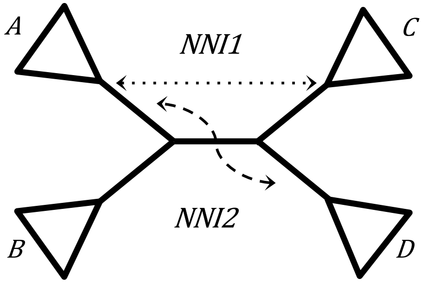


Tree search algorithms in RAxML and IQ-TREE

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 new stochastic algorithm





Nearest neighbor interchange

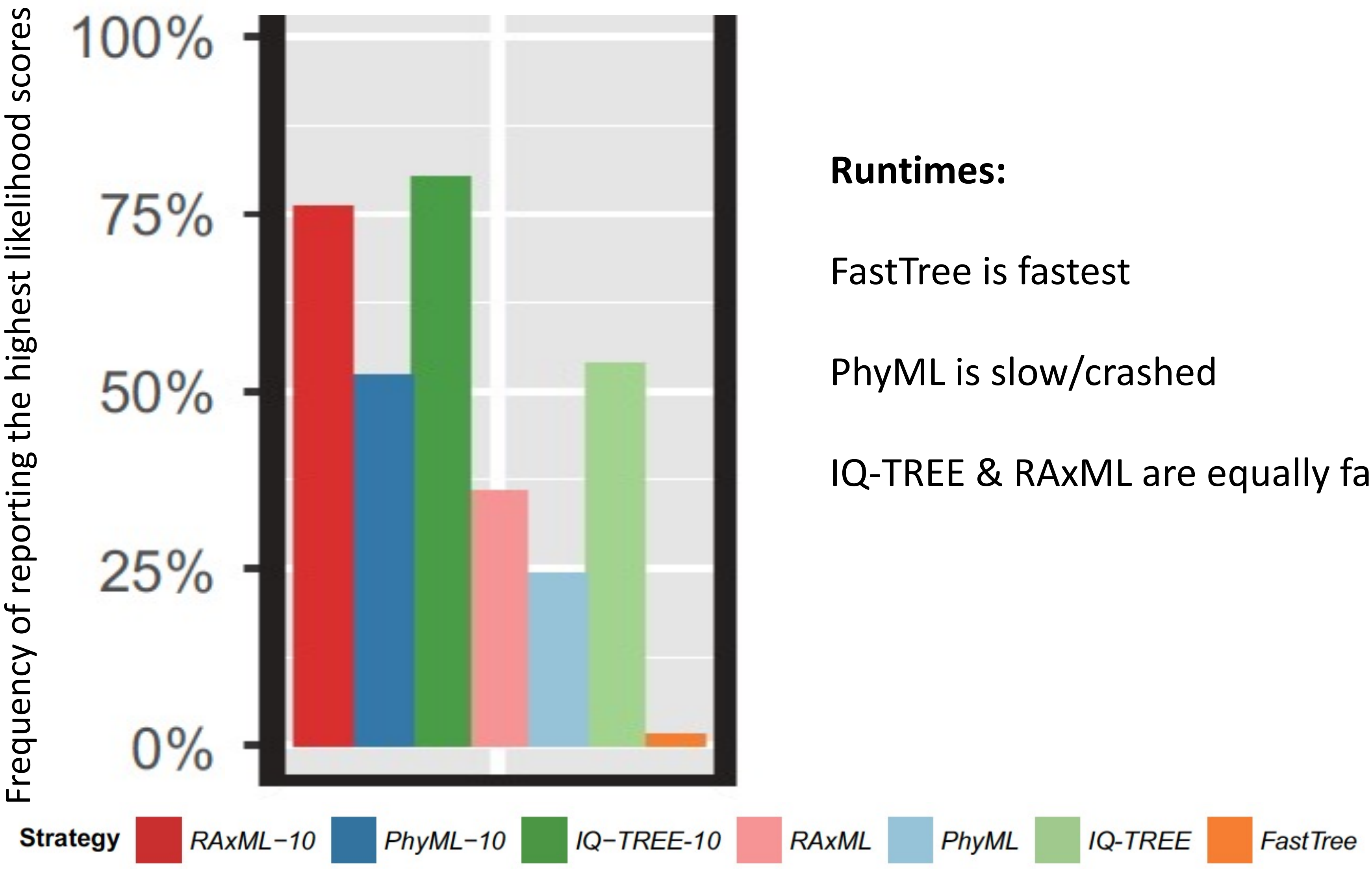
Metaheuristics:
Random restart, Iterated local search, Evolution strategy

Lam-Tung Nguyen

Heiko Schmidt

Arndt von Haeseler

An independent benchmark by Zhou et al. (2018)



Runtimes:

FastTree is fastest

PhyML is slow/crashed

IQ-TREE & RAxML are equally fast

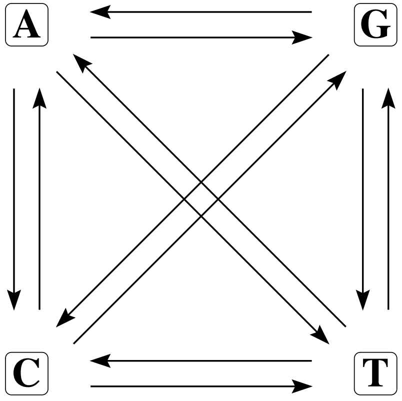
Step 2: Summary

Multiple sequence alignment

```
ACGGGAT--C--C----CATTAC
ACGGGAT--C--C----CACTAC
CCGGGATAGCTTC----CATTAC
ACCCCCTATC--CACTGGATTAC
ACGACATATC--CACTGGATTCC
```

Model selection

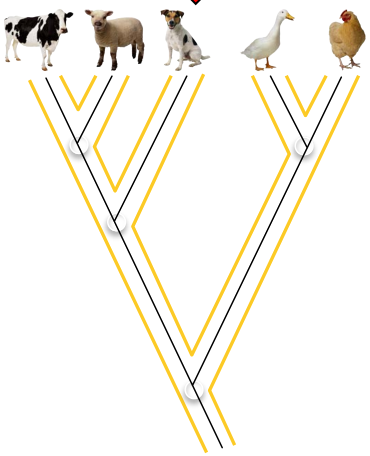
Substitution model



- IQ-TREE algorithm efficiently explores tree space

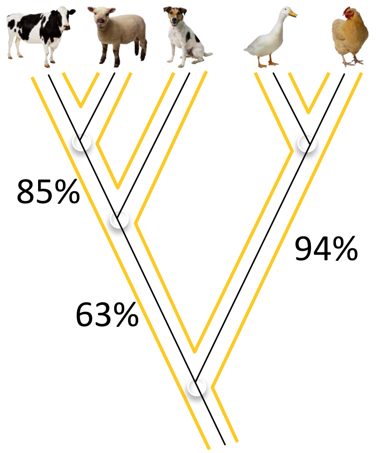
IQ-TREE (2015, 2020)

Tree reconstruction



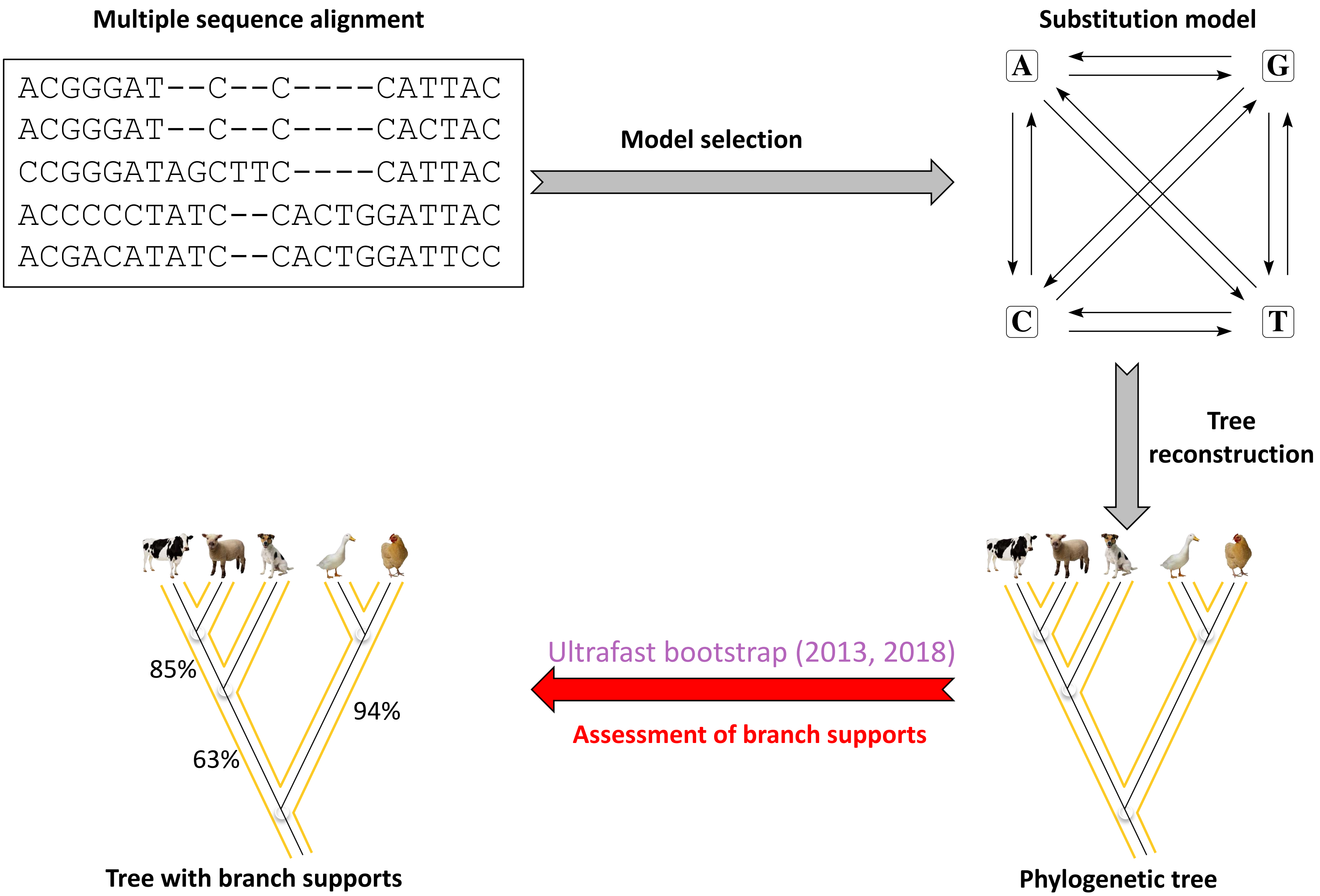
Phylogenetic tree

Assessment of branch supports



Tree with branch supports

Step 3: Ultrafast bootstrap



Multiple sequence alignment

```
ACGGGAT--C--C----CATTAC
ACGGGAT--C--C----CACTAC
CCGGGATAGCTTC----CATTAC
ACCCCCTATC--CACTGGATTAC
ACGACATATC--CACTGGATTCC
```

Model selection

Substitution model

Tree reconstruction

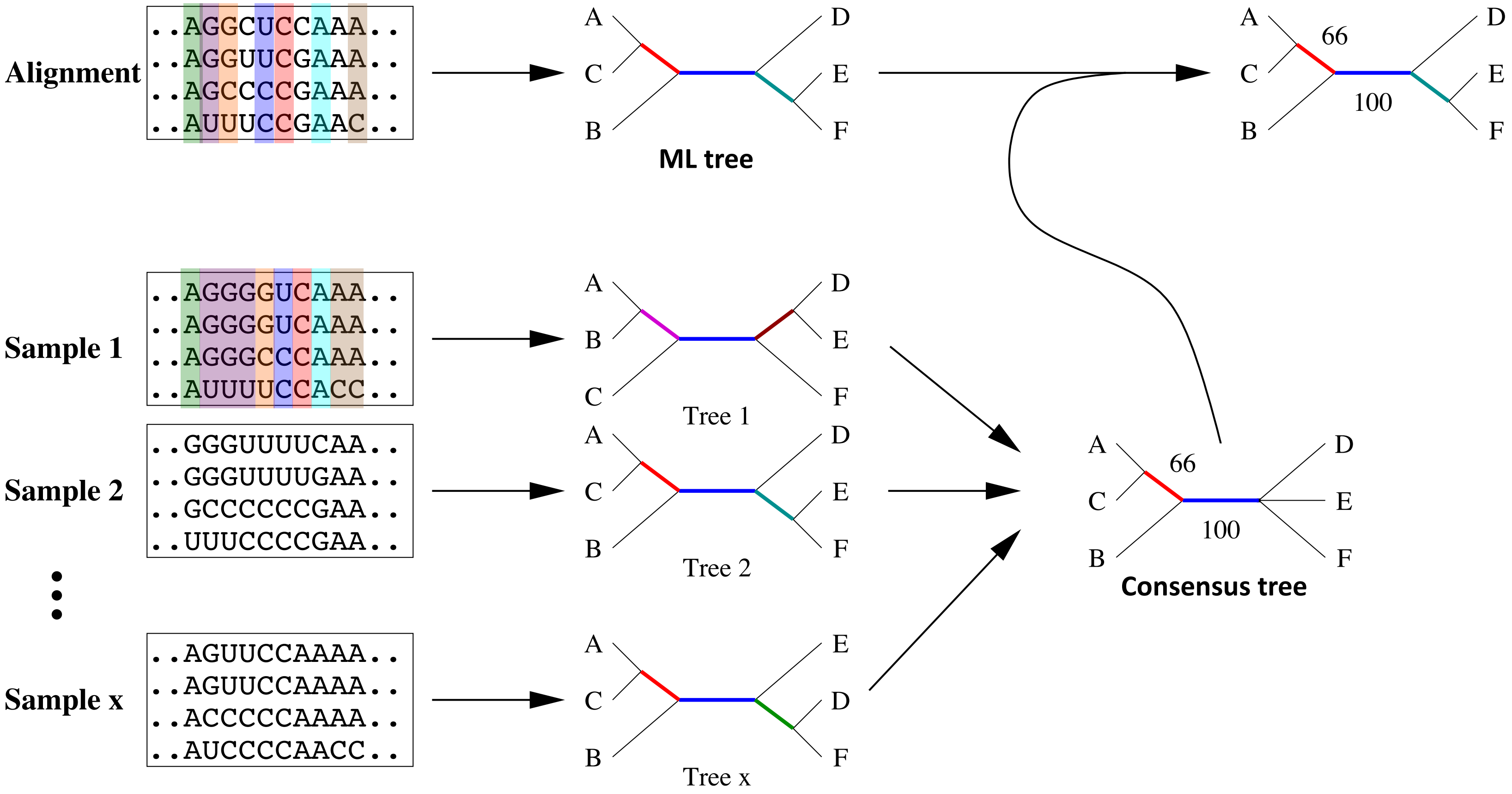
Phylogenetic tree

Ultrafast bootstrap (2013, 2018)

Assessment of branch supports

Tree with branch supports

Bootstrap: How reliable are branches of the tree?



Alignment

```
..AGGCUCCAAA..
..AGGUUCGAAA..
..AGCCCCGAAA..
..AUUUCCGAAC..
```

Sample 1

```
..AGGGGUCAAA..
..AGGGGUCAAA..
..AGGGCCCAAA..
..AUUUUCCACC..
```

Sample 2

```
..GGGUUUUCAA..
..GGGUUUUGAA..
..GCCCCCCGAA..
..UUUCCCCGAA..
```

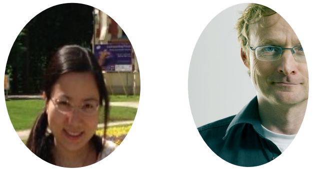
⋮

Sample x

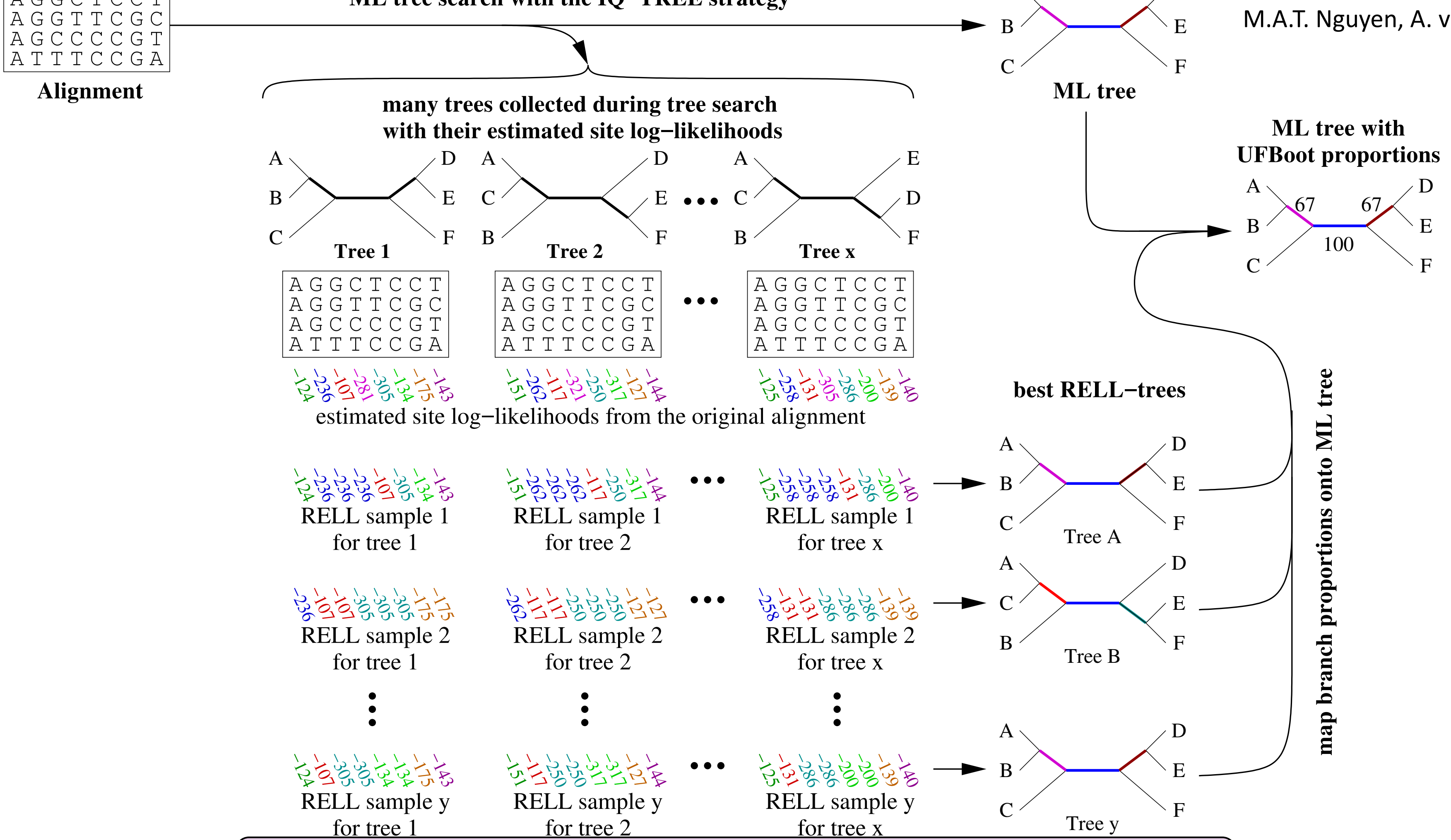
```
..AGUUCCAAAA..
..AGUUCCAAAA..
..ACCCCCAAAA..
..AUCCCCAACC..
```

Bootstrap analysis is extremely time-consuming!

UFBoot: Ultrafast bootstrap approximation



M.A.T. Nguyen, A. von Haeseler



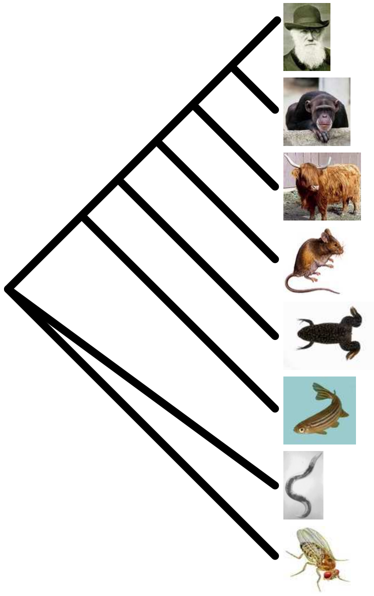
Use UFBoot >= 95% instead of 70% !

Genome-scale data: Concatenation methods

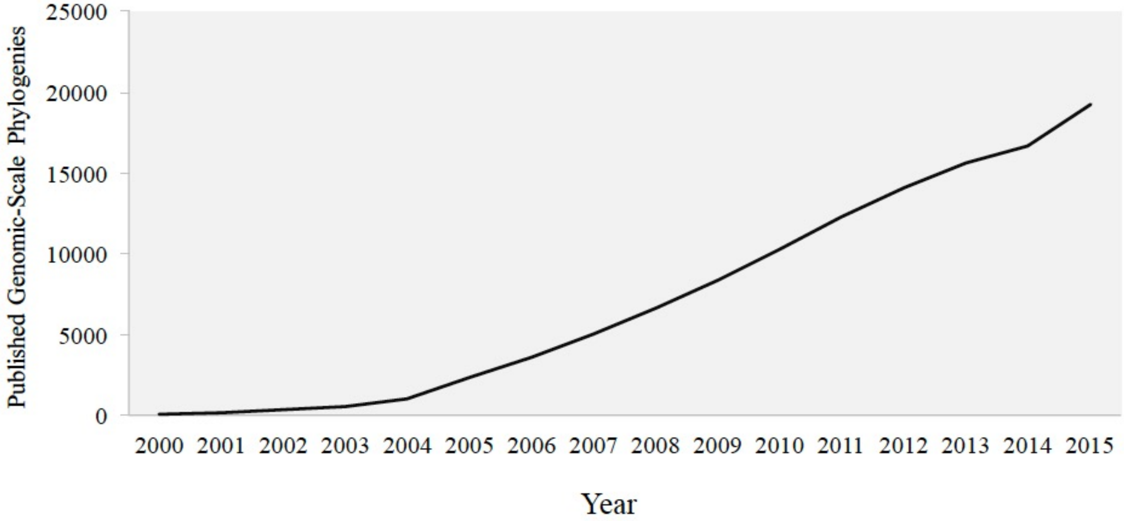
Supermatrix

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



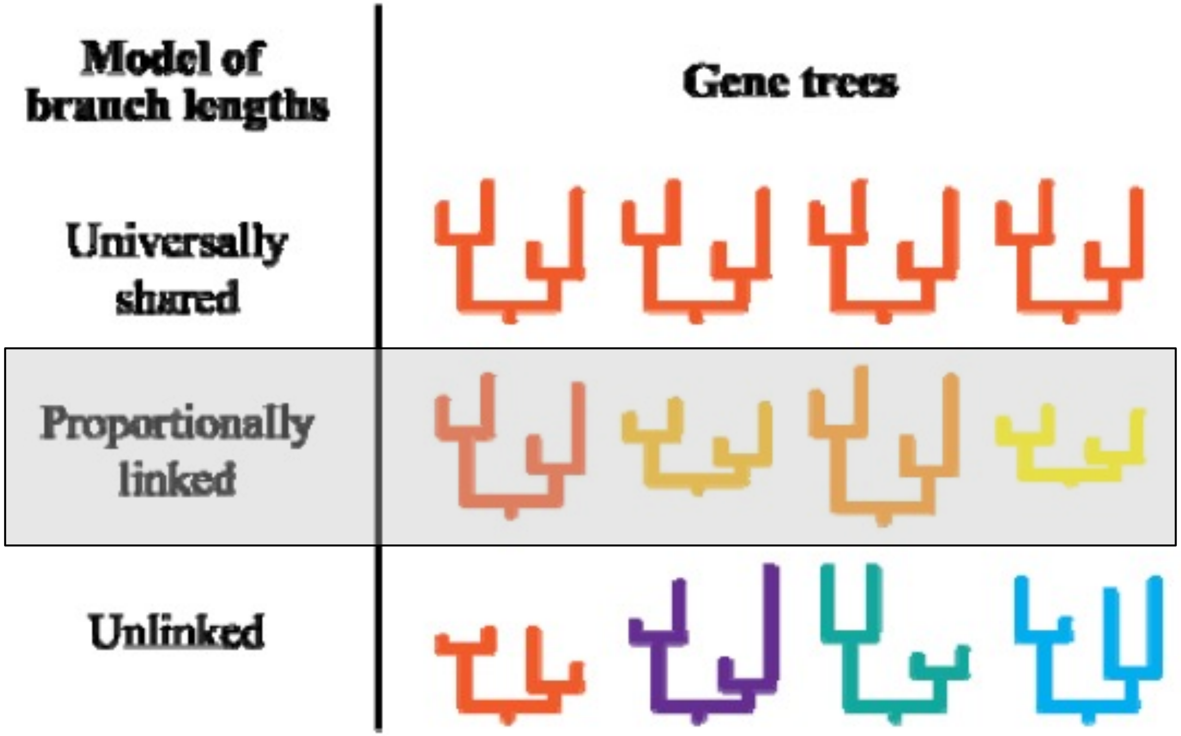
30 days of computation and 280 GB RAM for an insect data set!

Partition model

Supermatrix

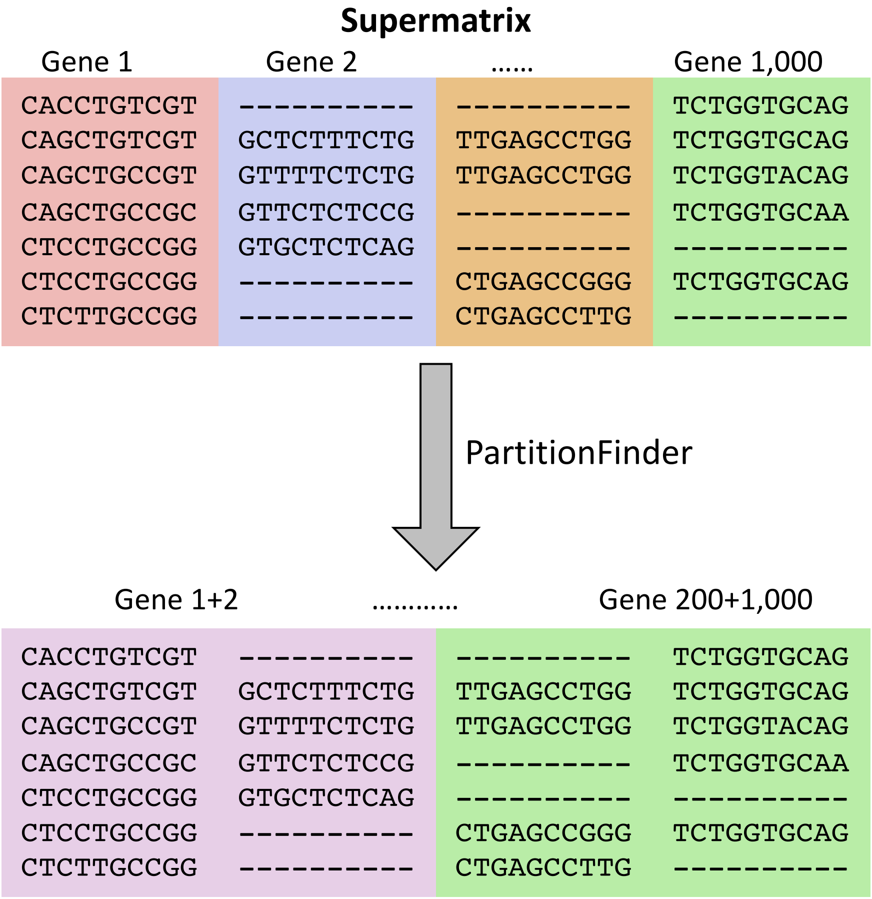
Gene 1	Gene 2	......	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG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

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



Recommended for typical analysis (Duchene et al. 2020)

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	----------

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

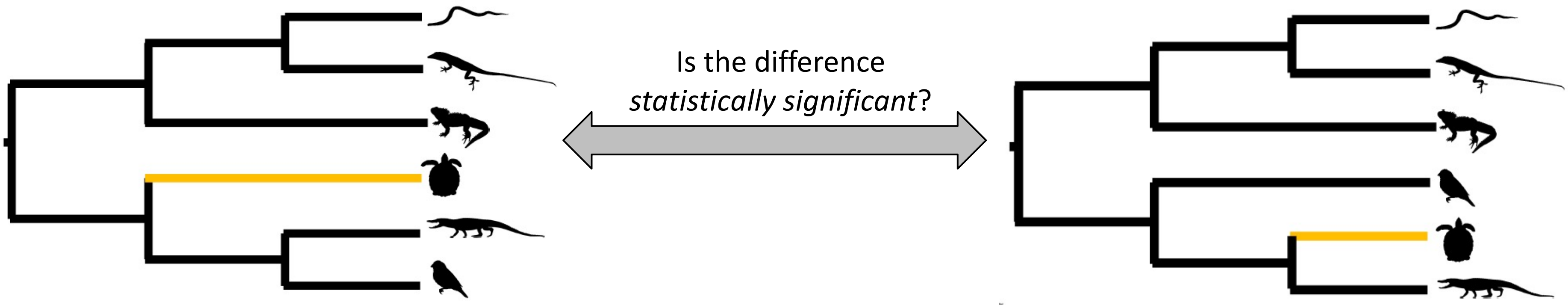
PartitionFinder algorithm (Lanfear et al. 2012):

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.

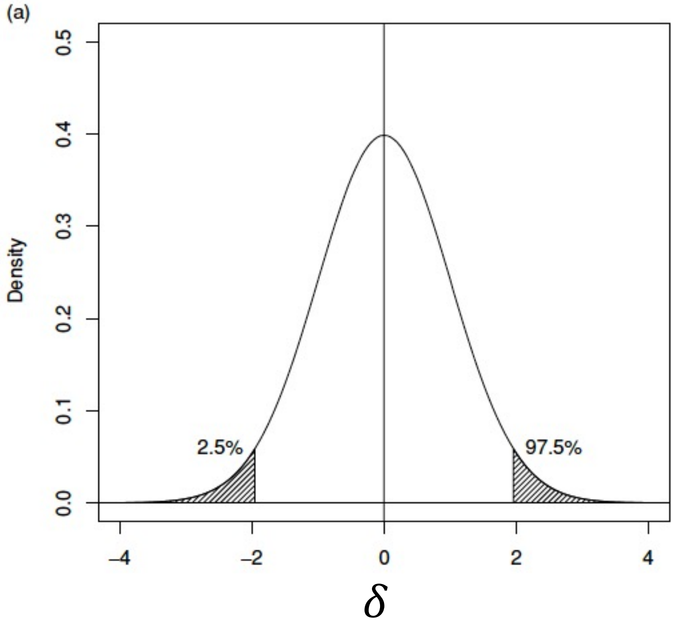
Tree topology tests



Is the difference *statistically significant*?

Testing two trees (Kishino & Hasegawa, 1989):

1. Statistic: $\delta = \log\big(likelihood(T_1)\big) - \log\big(likelihood(T_0)\big)$.
2. Generate distribution of δ 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.
5. If p-value >= 0.05: NO! they are not.

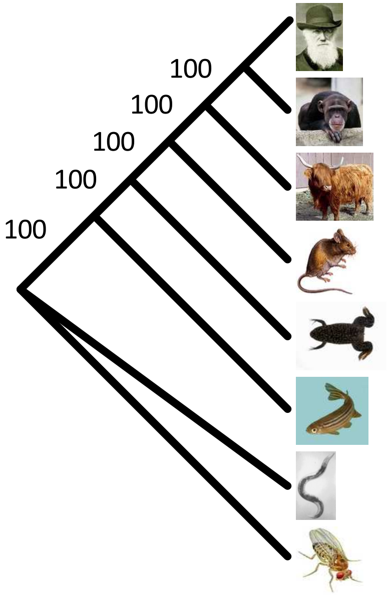


Concatenation methods: Limitation

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



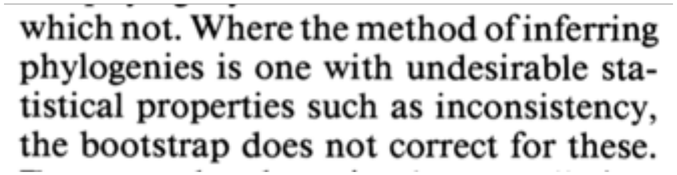
Species tree of life

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

Concatenation assumes a single tree across all loci

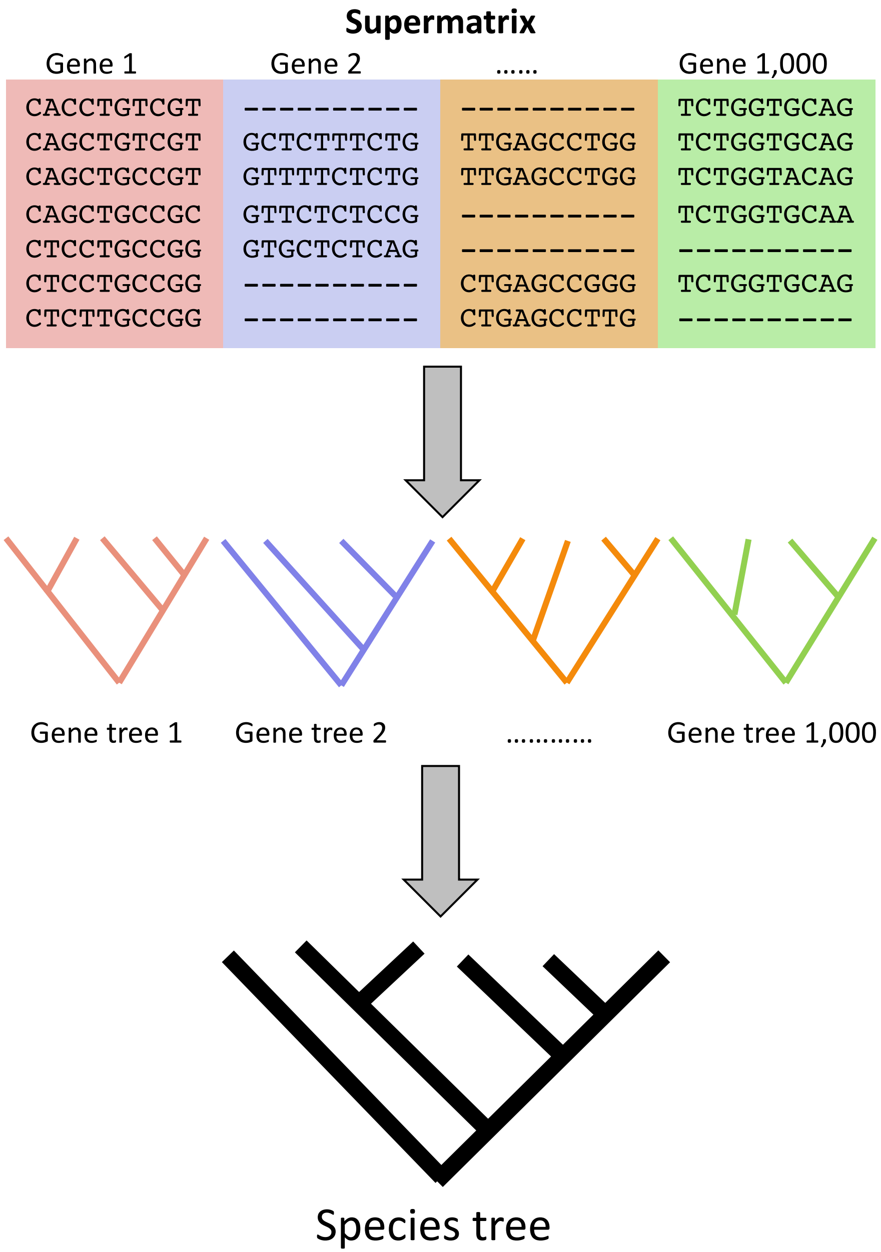
Potential *systematic bias*

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	----------

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

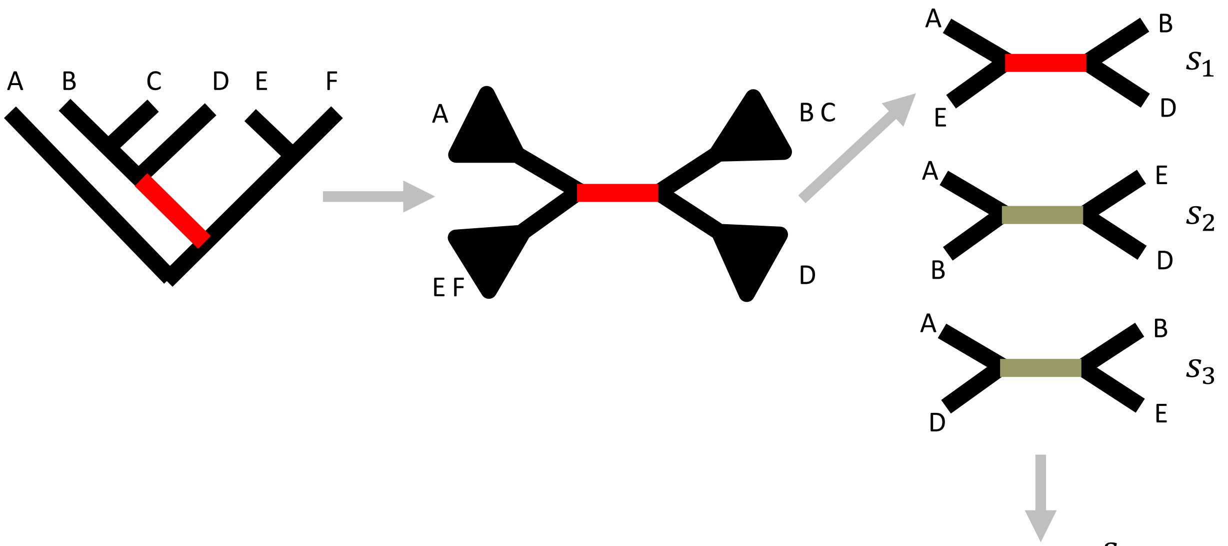
Problem: Uncertainties in gene trees!

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%



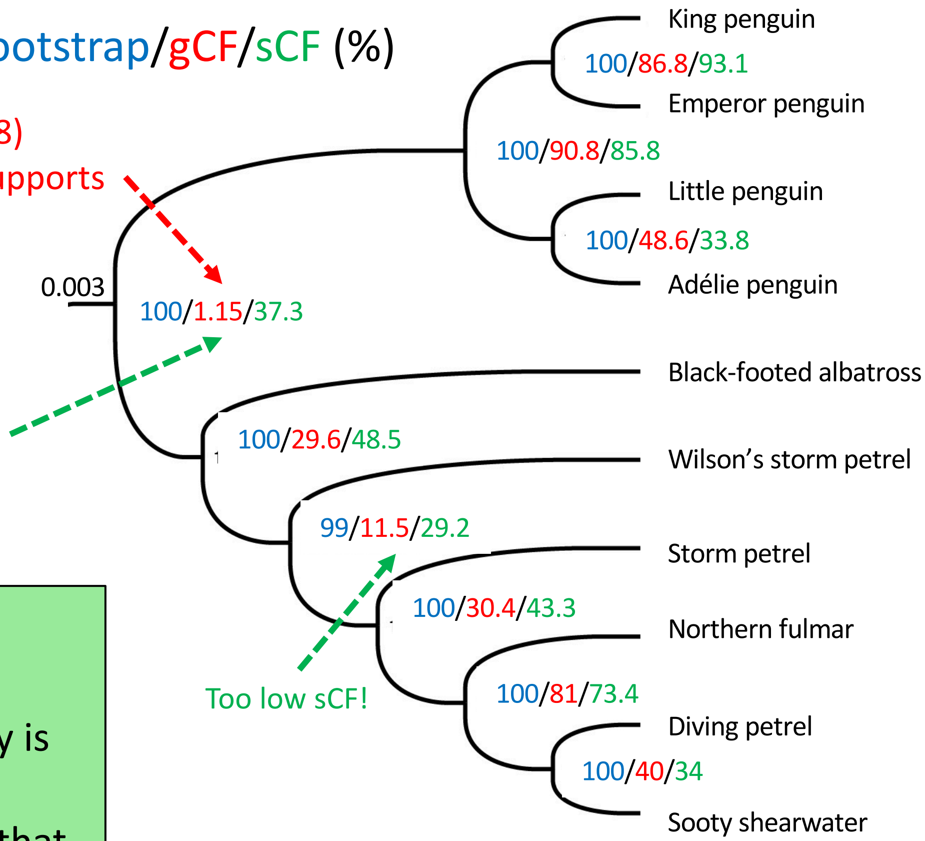
$$sCF = \overline{qCF(100\ quartets)} \leftarrow qCF(quartet) = \frac{s_1}{s_1 + s_2 + s_3}$$

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

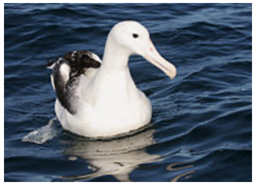
Bootstrap/gCF/sCF (%)

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

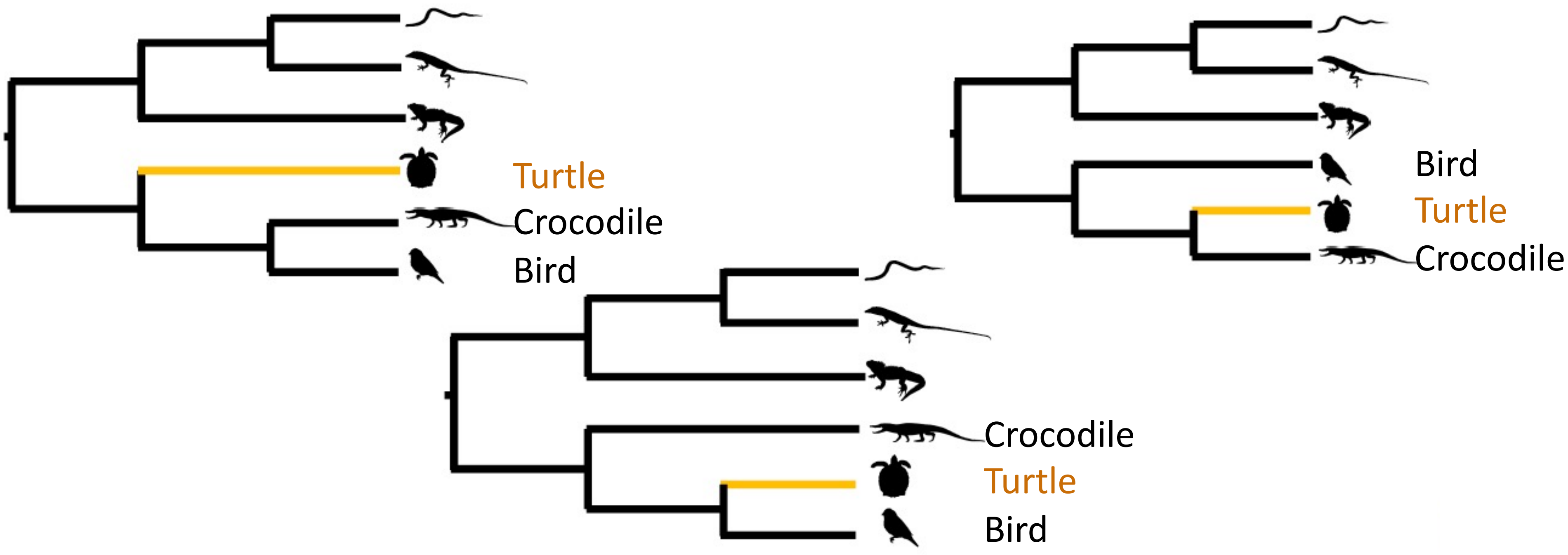


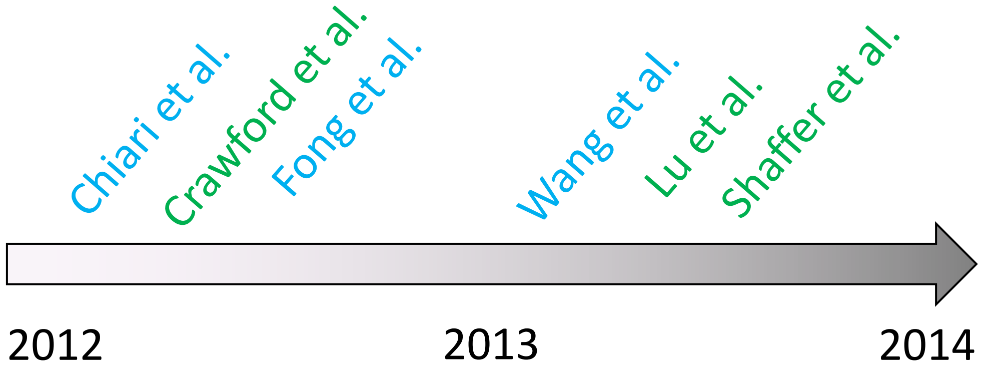
Tubenoses

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

Dataset for IQ-TREE lab: 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

IQ-TREE lab

1. Input data
2. Inferring the first phylogeny
3. Applying partition model
4. Choosing the best partitioning scheme
5. Tree topology tests
6. Identifying most influential genes
7. Removing influential genes
8. Concordance factors (*advanced)

http://www.iqtree.org/workshop/molevol2022

Fill out your answers in a Google form (shared via Slack)