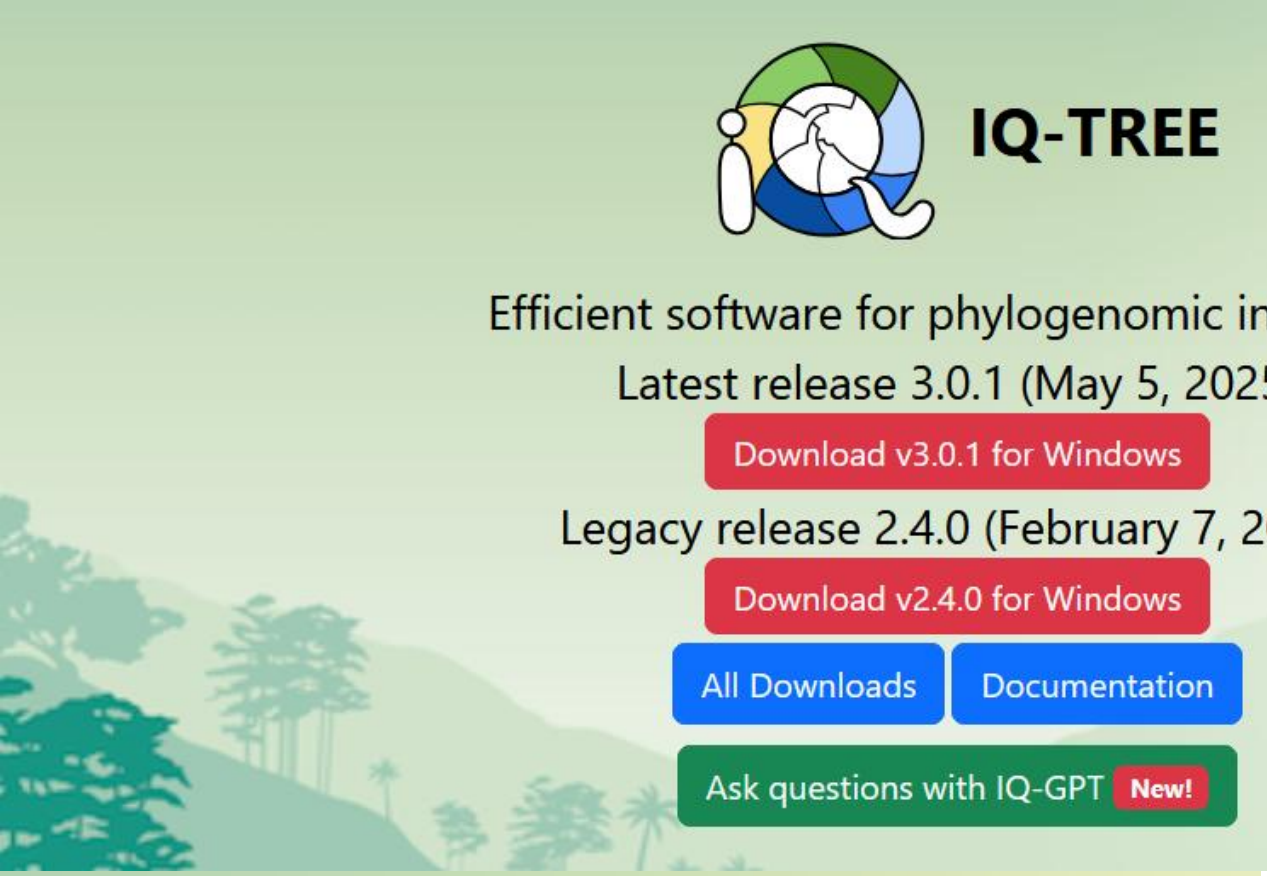




# IQ-TREE introduction for MOLE 2025

Slides by
Bui Quang Minh and
Hanon Solomon McShea

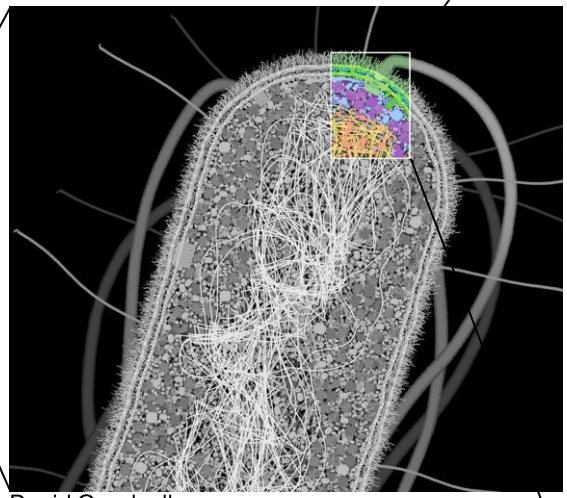
Stanford University →
University of California, San Francisco

# The molecules of molecular evolution

**Bacteria**

Tasha Sturm

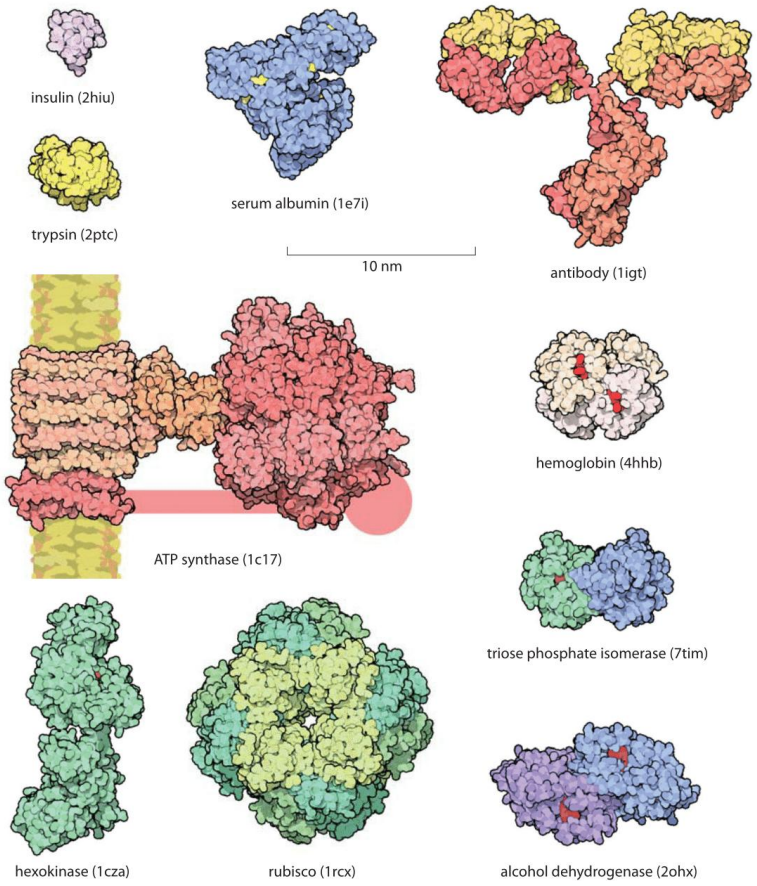
**Bacterial cell**



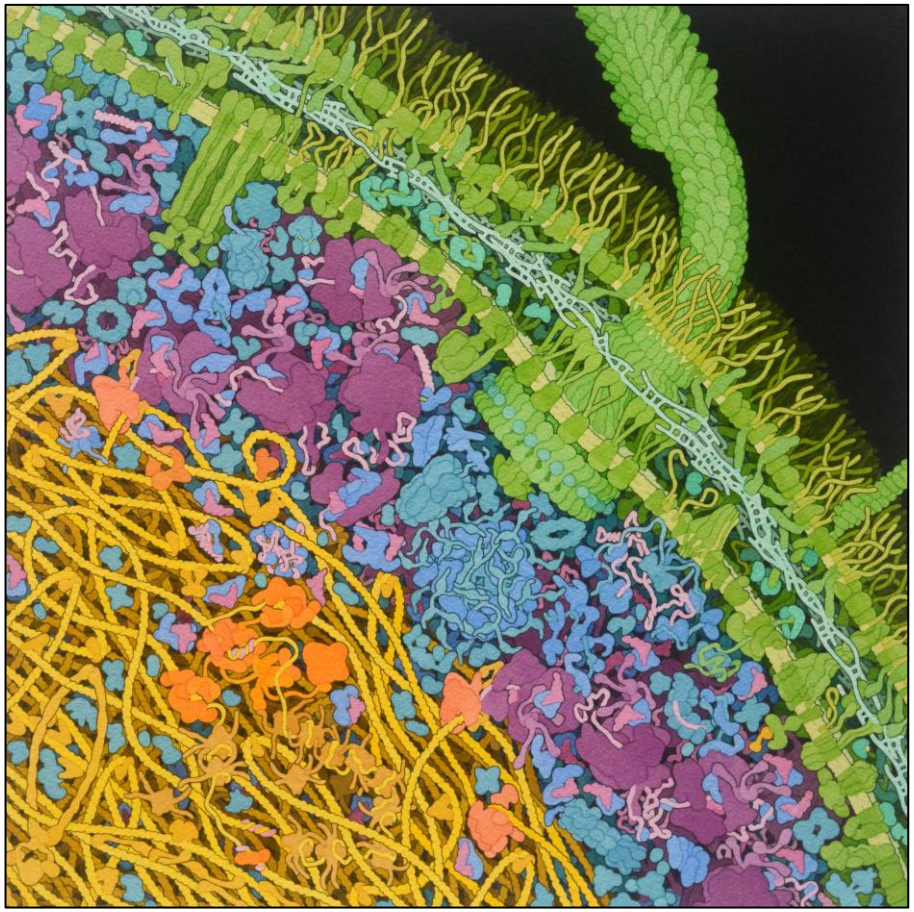
David Goodsell

David Goodsell

# The molecules of molecular evolution

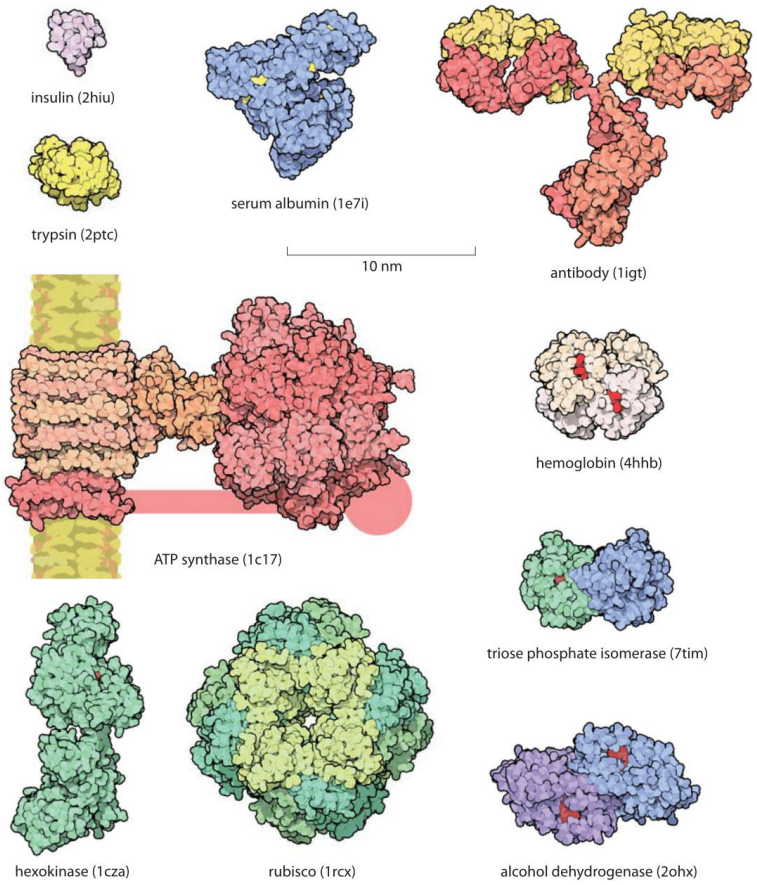



Ron Milo & Rob Phillips, *Cell Biology by the Numbers*

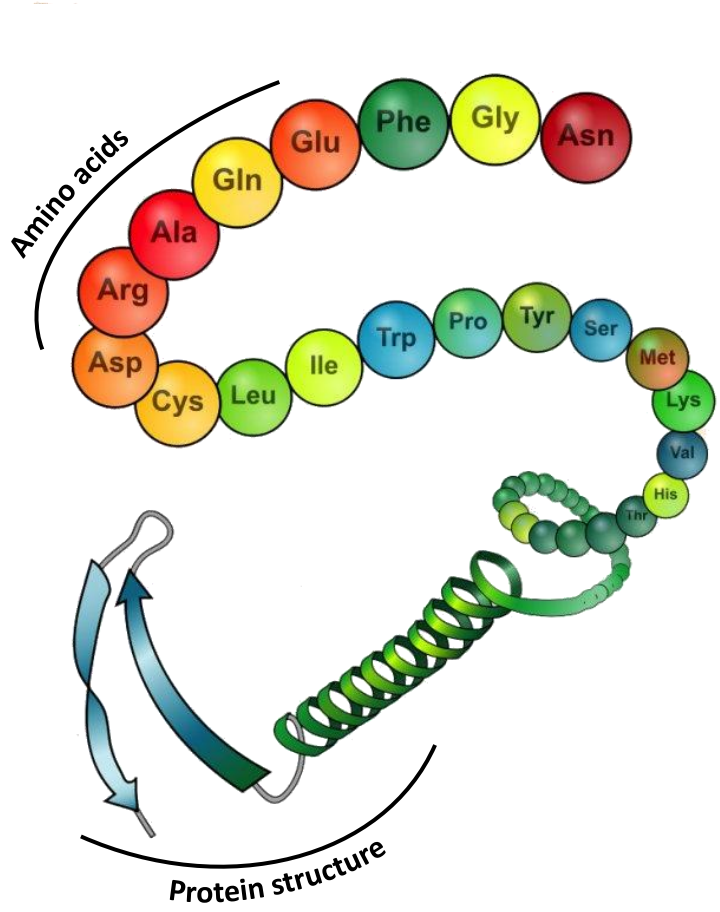


David Goodsell

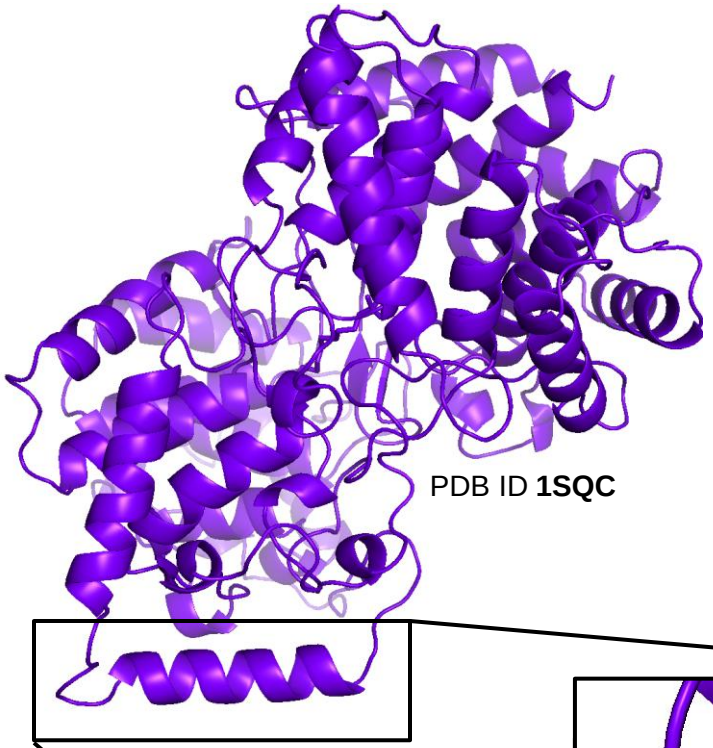
# The molecules of molecular evolution



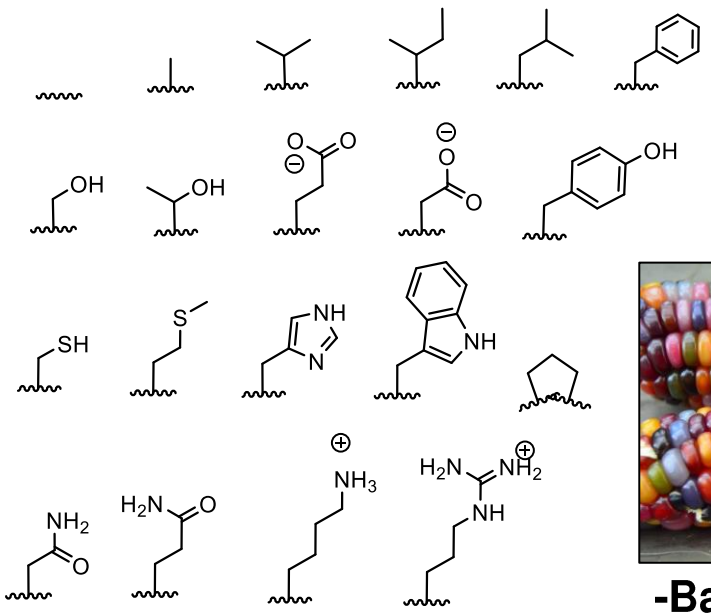

Ron Milo & Rob Phillips, *Cell Biology by the Numbers*



# The molecules of molecular evolution

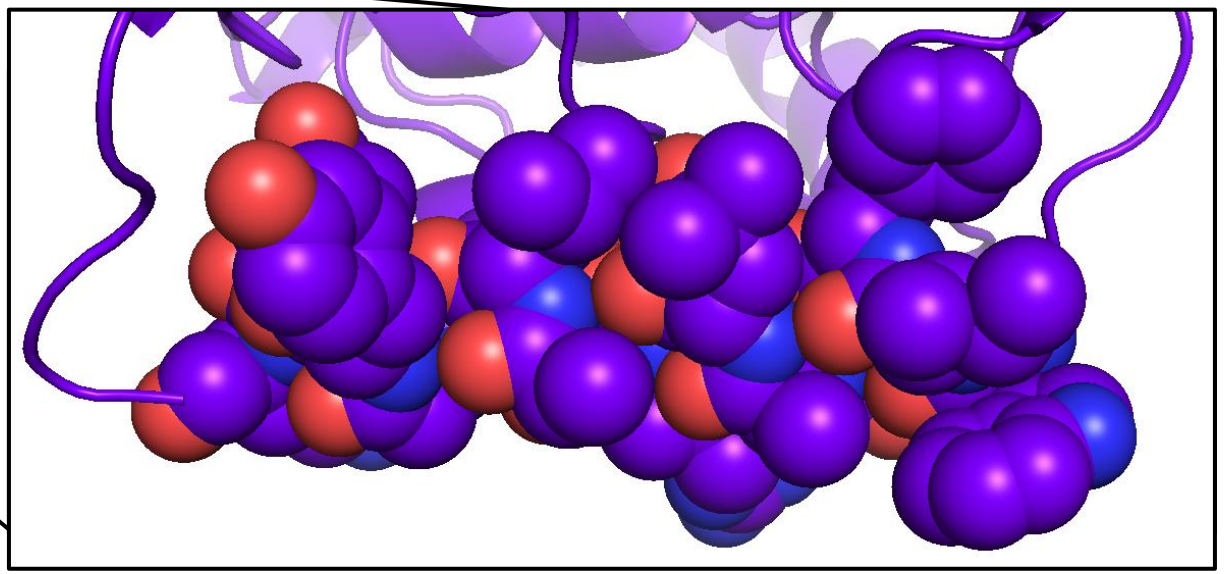


PDB ID **1SQC**

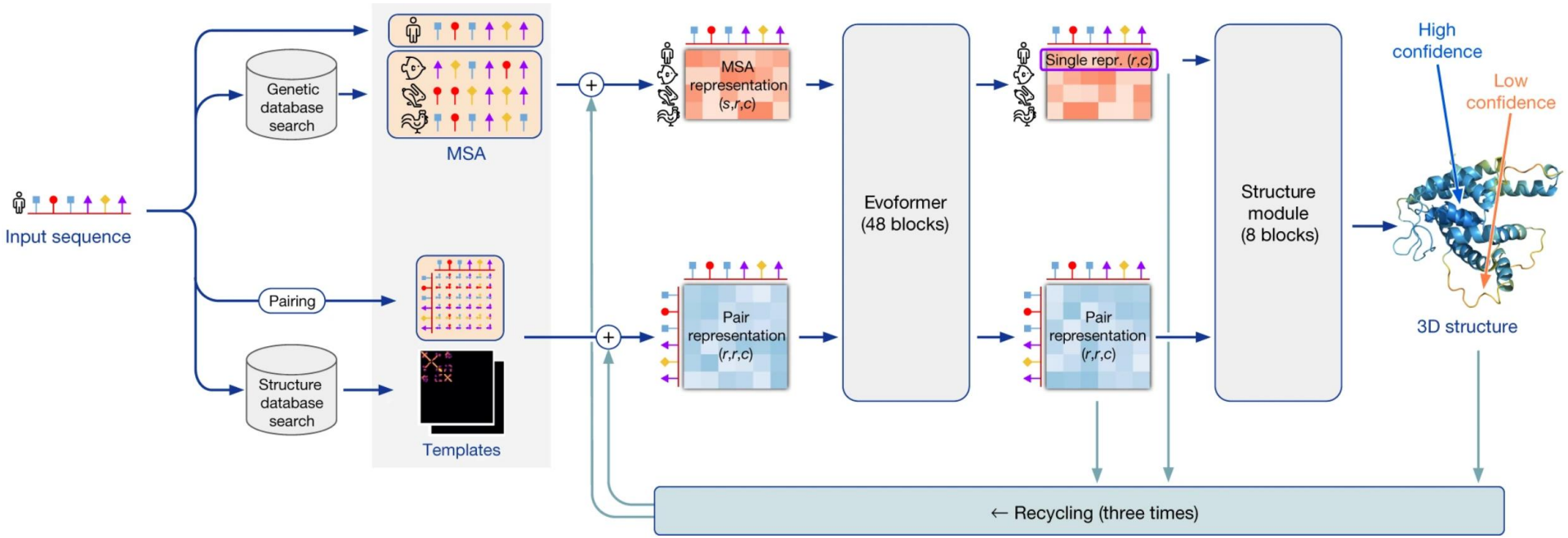



**“A feeling for the organism”**

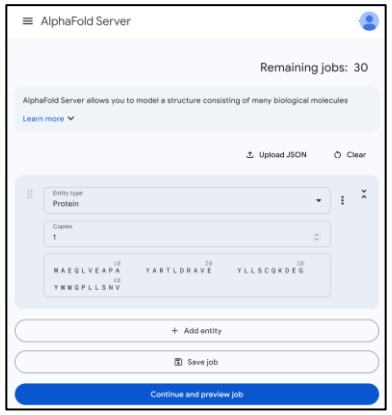
**-Barbara McClintock**



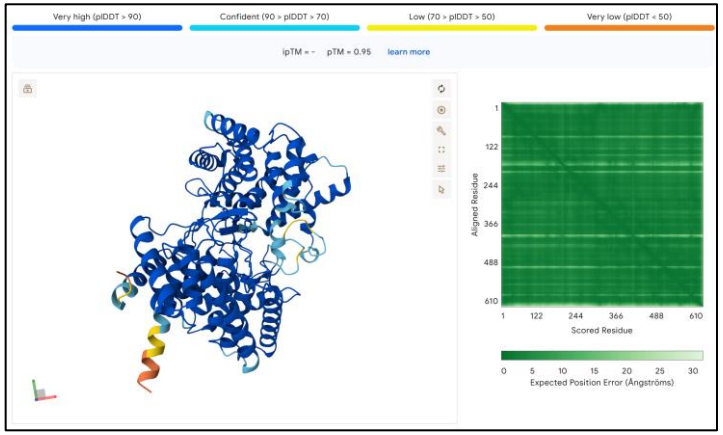
# Predicting protein structures




Jumper et al. 2021




**Alphafold is very easy!**



## Why IQ-TREE?

To rigorously analyze very large data

To implement a diversity of models with appropriate realism for a variety of questions

To, alongside RAxML, PhyML, et al., advance the field of maximum likelihood phylogeny estimation and analysis (just as we have RevBayes, MrBayes, BEAST for Bayesian analysis)

# Typical phylogenetic analysis under maximum likelihood

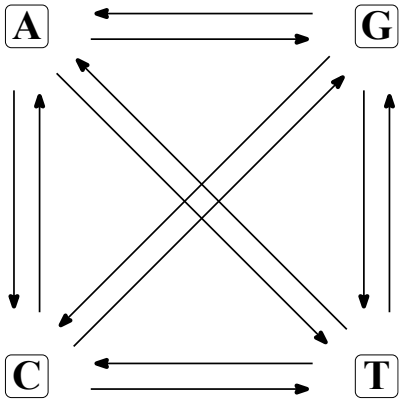
**Multiple sequence alignment**

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

**Model selection**

ModelFinder (2017)
PartitionFinder (2012, 2017)
MixtureFinder (2025)

**Substitution model**

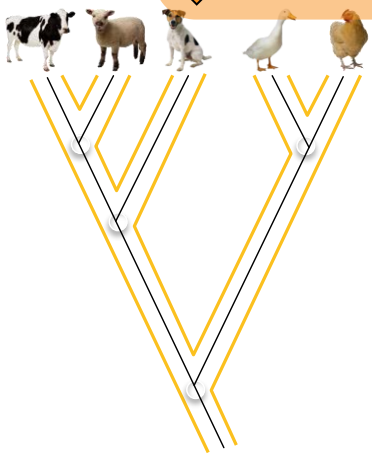



We focused on improving all three steps for large datasets!

IQ-TREE (2015, 2020,2025)

**Tree reconstruction**

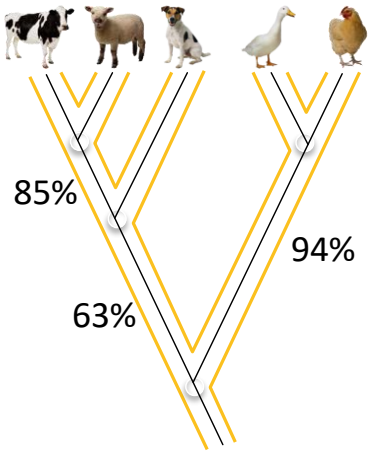
**iqtree3 -s ALN_FILE -B 1000**



**Phylogenetic tree**

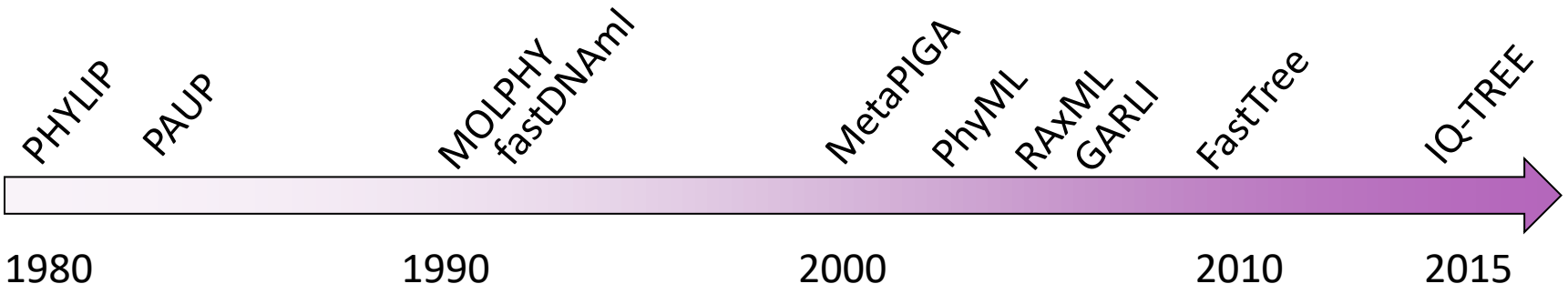
Ultrafast bootstrap (2013, 2018)

**Assessment of branch supports**

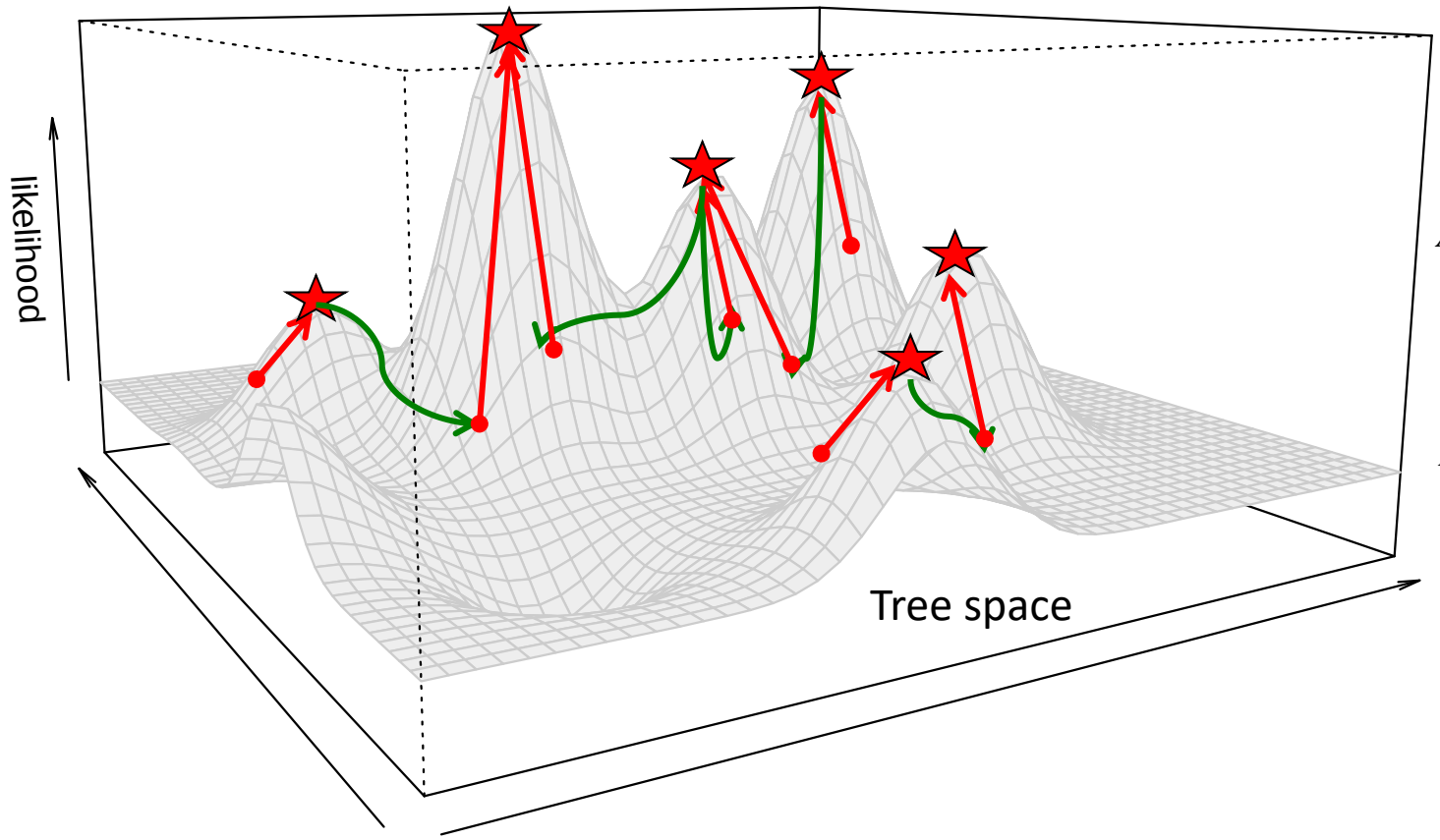


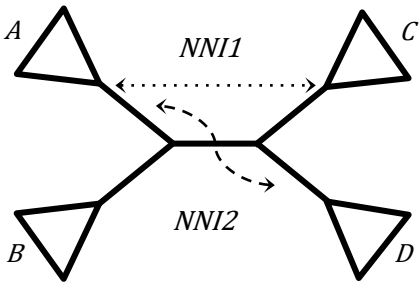

**Tree with branch supports**

# Search heuristics for finding maximum likelihood trees



# IQ-TREE: A new stochastic algorithm




Nearest neighbor interchange

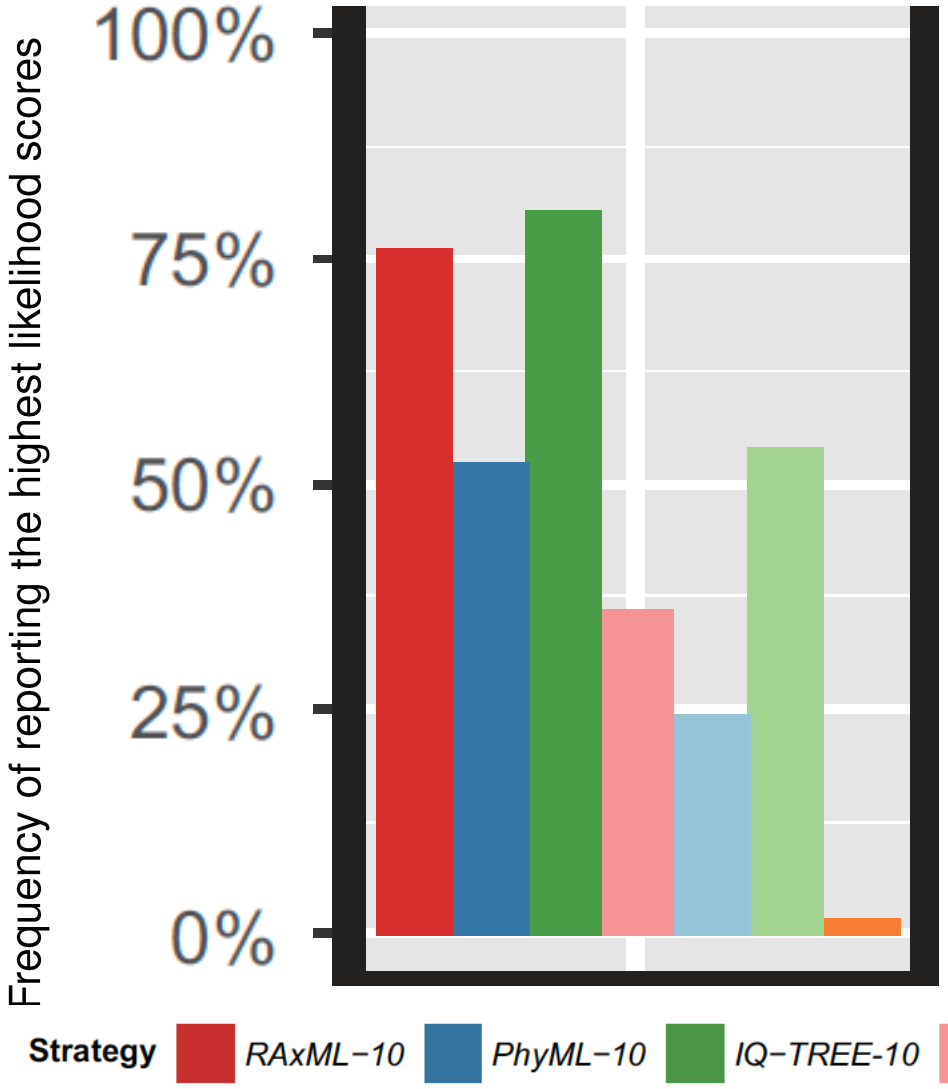
- 100 starting trees (99 parsimony, 1 NJ)
- Keeping a “population” of 20 best trees
- Stop if unsuccessful for 100 consecutive downhill + uphill moves

Lam-Tung Nguyen

Heiko Schmidt

Arndt von Haeseler

# An independent benchmark by Zhou et al. (2018)




**Runtimes:**

FastTree is fastest

PhyML is slow or crashed

IQ-TREE & RAxML are equally fast

# IQ-TREE tree search algorithm

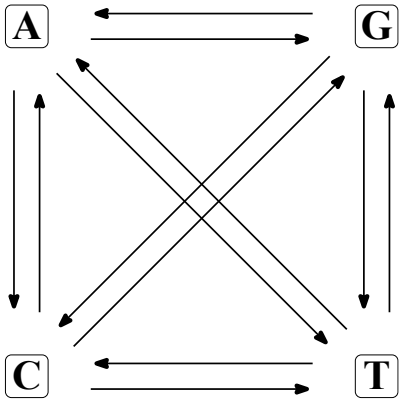
**Multiple sequence alignment**

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

**Model selection**

ModelFinder (2017)
PartitionFinder (2012, 2017)
MixtureFinder (2025)

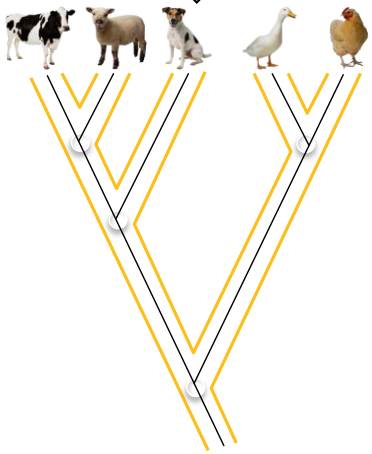
**Substitution model**




IQ-TREE algorithm efficiently explores tree space

IQ-TREE (2015, 2020,2025)

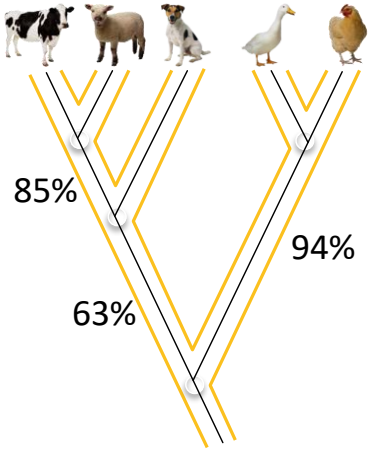
**Tree reconstruction**



**Phylogenetic tree**

Ultrafast bootstrap (2013, 2018)

**Assessment of branch supports**




**Tree with branch supports**

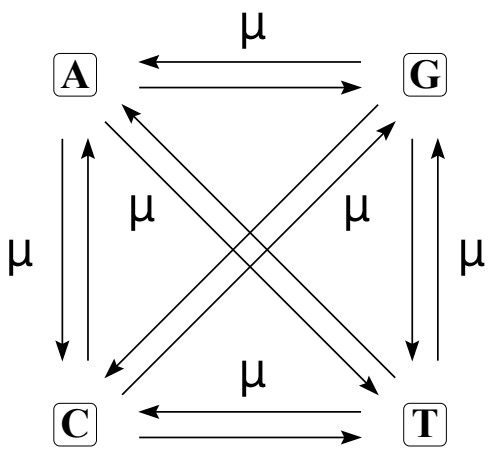
# **Models:** descriptions of sequence evolution

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Human | C | A | A | – | – | – | A | A | T | A | T | T | A | C |
| Chimp | C | A | C | A | – | T | A | C | – | T | T | T | A | C |
| Gorilla | C | A | C | T | A | T | A | A | G | T | T | T | A | C |
| Orangutan | C | A | C | – | – | – | A | C | A | A | A | T | A | C |

∃ empirical and theoretical evolutionary models for nucleic acid sequences, proteins, codons, any trait you can imagine (discrete and continuous)

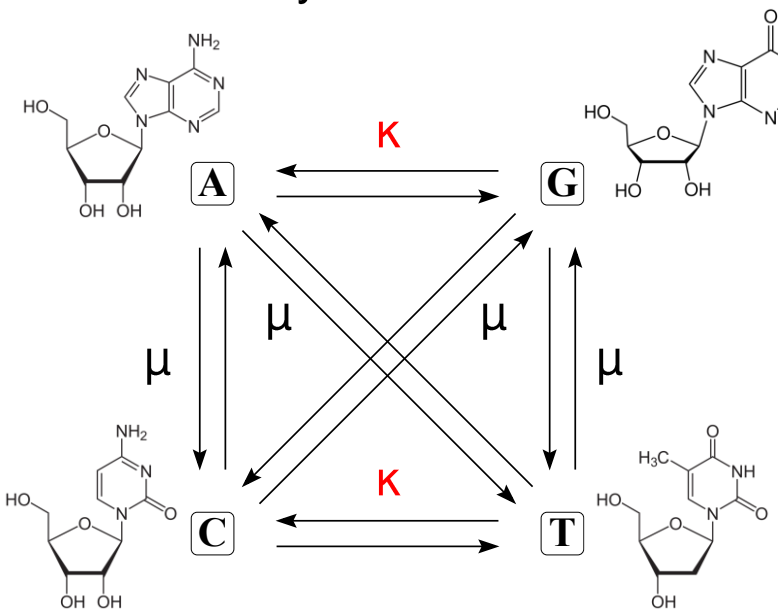
What can we vary in an evolutionary model?

**Exchangeabilities**



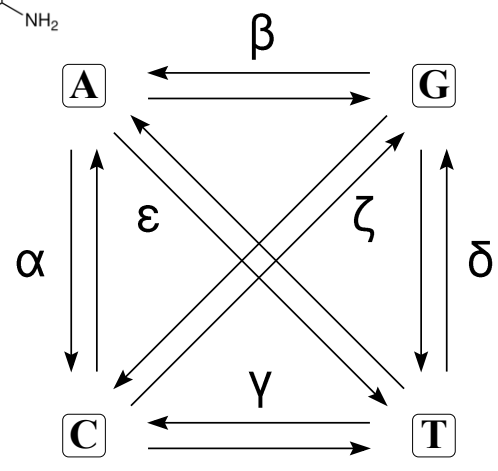

Jukes and Cantor (1969)

**"JC"**




Hasegawa, Kishino, Yano (1985)

**"HKY"**




General Time Reversible (1986)

**"GTR"**

# **Models:** descriptions of sequence evolution

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Human | C | A | A | - | - | - | A | A | T | A | T | T | A | C |
| Chimp | C | A | C | A | - | T | A | C | - | T | T | T | A | C |
| Gorilla | C | A | C | T | A | T | A | A | G | T | T | T | A | C |
| Orangutan | C | A | C | - | - | - | A | C | A | A | A | T | A | C |

What can we vary in an evolutionary model?
**Exchangeabilities**

**Frequencies**

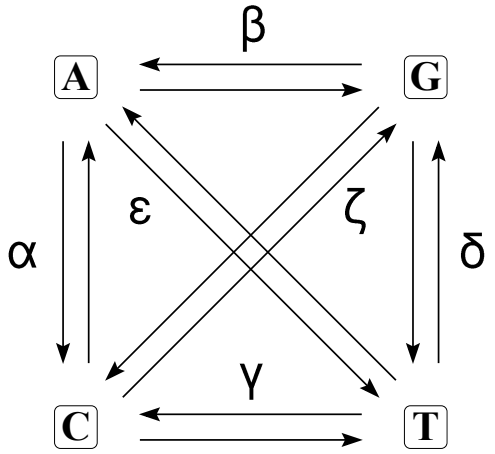
+F = empirical
+FQ uniform
+FO ML optimized
+F{0.079, 0.056, …, 0.034} user-specified

**Rates**

+I some sites are invariant (rate = 0)
+G site rates follow a gamma distribution
+R site rates fall into several rate categories

How model is fit: to whole alignment, to genes, to sites? Deterministically or probabilistically?

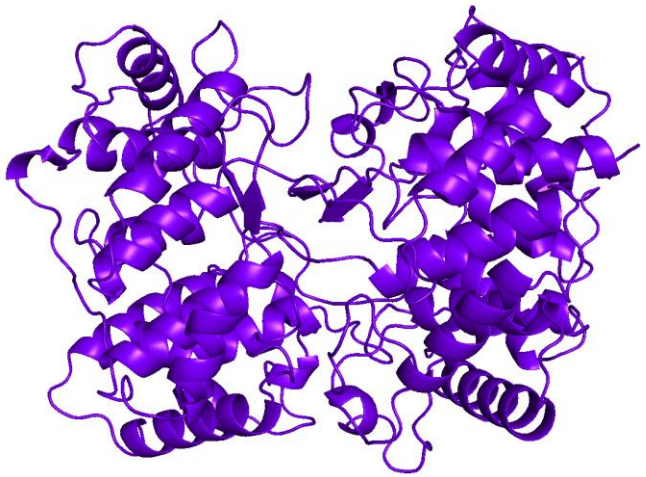
∃ empirical and theoretical evolutionary models for nucleic acid sequences, proteins, codons, any trait you can imagine (discrete and continuous)




General Time Reversible (1986)
**"GTR"**

**You can input OR fit your own in IQ-TREE with Qmaker (Bui et al. 2021) and NQmaker (Dang et al. 2022)!**

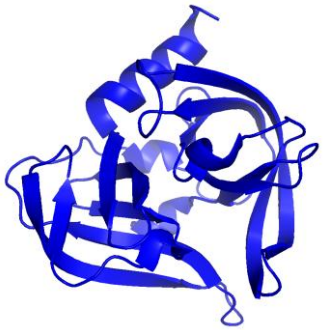
# **Models:** biophysical basis of variation in *Q*



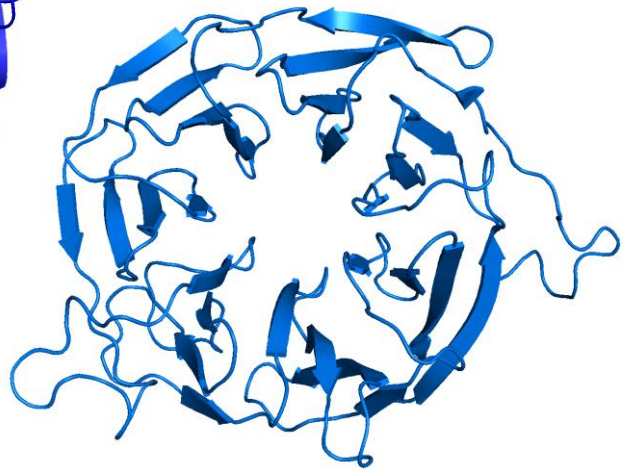
PDB ID **1SQC**

$$Q = \begin{pmatrix} - & r_{1,2}\pi_2 & r_{1,3}\pi_3 & \cdots & r_{1,20}\pi_{20} \\ r_{1,2}\pi_1 & - & r_{2,3}\pi_3 & \cdots & r_{2,20}\pi_{20} \\ r_{1,3}\pi_1 & r_{2,3}\pi_2 & - & \cdots & r_{3,20}\pi_{20} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{1,20}\pi_1 & r_{2,20}\pi_2 & r_{3,20}\pi_3 & \cdots & - \end{pmatrix},$$

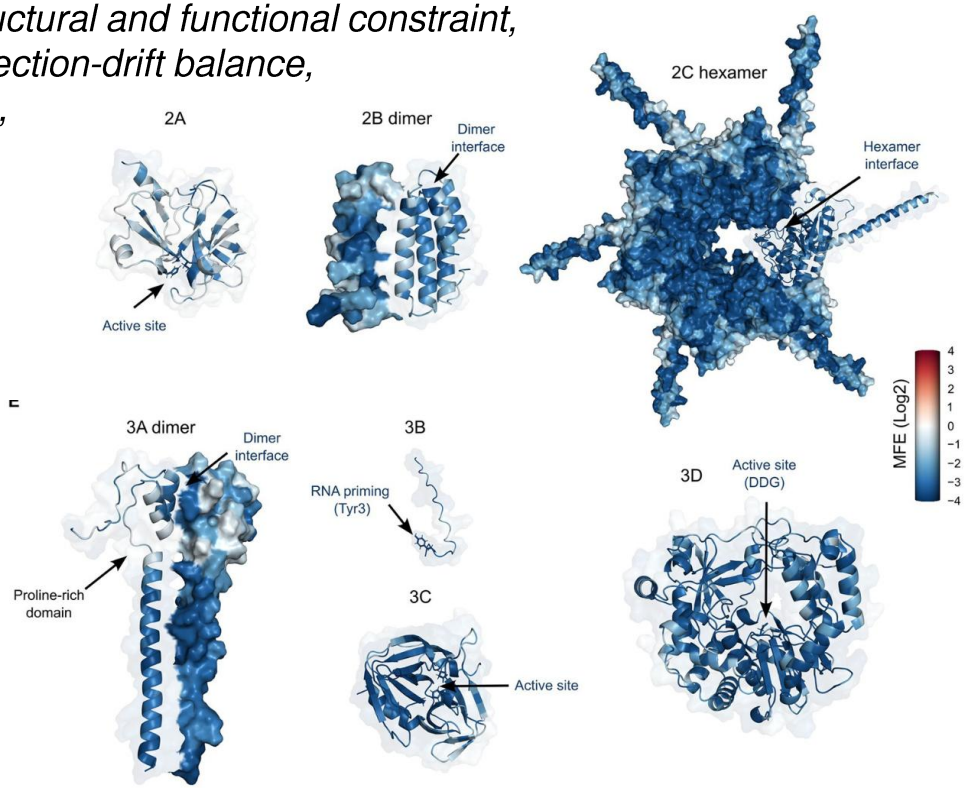
*Where exchangeabilities (r) and frequencies (π) vary with structural and functional constraint, mutation-selection-drift balance, genetic code, biosynthesis, and more*



PDB ID **6FFS**

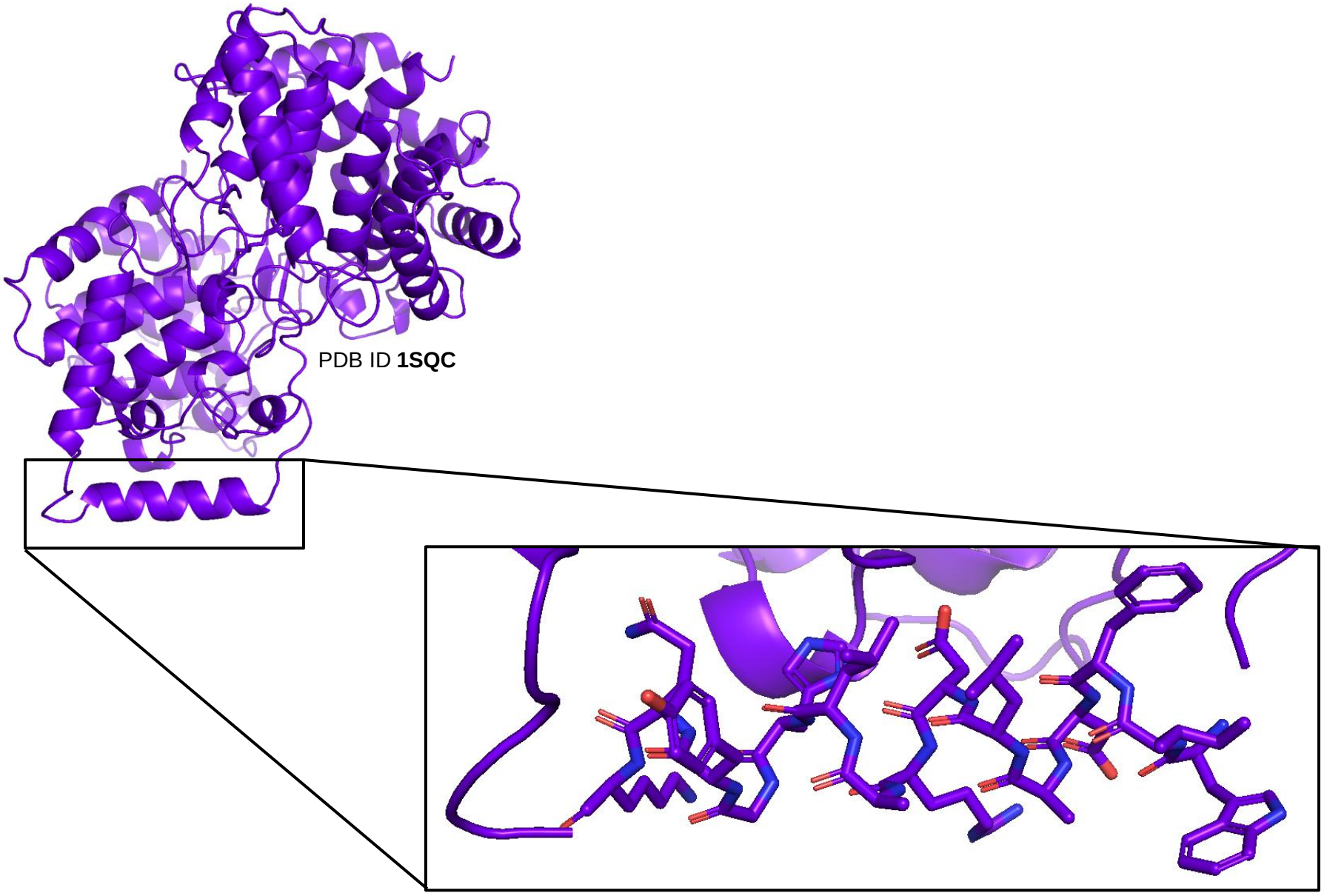


PDB ID **8HMC**




Coxsackievirus B3 proteome from Álvarez-Rodríguez et al. 2024

# Molecular and cellular context matter



# **Models:** those available in IQ-TREE

## NUCLEOTIDE* MODELS

| 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 |

**+Lie Markov models**

## GENERAL MODELS

| Model | Explanation |
|---|---|
| JC2 | Jukes-Cantor type model for binary data. |
| GTR2 | General time reversible model for binary data. |
| MK | Jukes-Cantor type model for morphological data. |
| ORDERED | Allowing exchange of neighboring states only. |

| Model | Explanation |
|---|---|
| C10 to C60 | 10, 20, 30, 40, 50, 60-profile mixture models (Le et al., 2008a) as variants of the CAT model (Lartillot and Philippe, 2004) for ML. Note that these models assume `Poisson` AA replacement and implicitly include a Gamma rate heterogeneity among sites. |
| EX2 | Two-matrix model for exposed/buried AA sites (Le et al., 2008b). |
| EX3 | Three-matrix model for highly exposed/intermediate/buried AA sites (Le et al., 2008b). |
| EHO | Three-matrix model for extended/helix/other sites (Le et al., 2008b). |
| UL2, UL3 | Unsupervised-learning variants of `EX2` and `EX3`, respectively. |
| EX_EHO | Six-matrix model combining `EX2` and `EHO` (Le and Gascuel, 2010). |
| LG4M | Four-matrix model fused with Gamma rate heterogeneity (Le et al., 2012). |
| LG4X | Four-matrix model fused with FreeRate heterogeneity (Le et al., 2012). |
| CF4 | Five-profile mixture model (Wang et al., 2008). |

## PROTEIN MODELS

| Model | Region | Explanation |
|---|---|---|
| Blosum62 | nuclear | BLOcks SUbstitution Matrix (Henikoff and Henikoff, 1992). Note that `BLOSUM62` is not recommended for phylogenetic analysis as it was designed mainly for sequence alignments. |
| cpREV | chloroplast | chloroplast matrix (Adachi et al., 2000). |
| Dayhoff | nuclear | General matrix (Dayhoff et al., 1978). |
| DCMut | nuclear | Revised `Dayhoff` matrix (Kosiol and Goldman, 2005). |
| EAL | nuclear | General matrix. To be used with profile mixture models (for eg. EAL+C60) for reconstructing relationships between eukaryotes and Archaea (Banos et al., 2024). |
| ELM | nuclear | General matrix. To be used with profile mixture models (for eg. ELM+C60) for phylogenetic analysis of proteins encoded by nuclear genomes of eukaryotes (Banos et al., 2024). |
| FLAVI | viral | Flavivirus (Le and Vinh, 2020). |
| FLU | viral | Influenza virus (Dang et al., 2010). |
| GTR20 | general | General time reversible models with 190 rate parameters. *WARNING: Be careful when using this parameter-rich model as parameter estimates might not be stable, especially when not having enough phylogenetic information (e.g. not long enough alignments).* |
| HIVb | viral | HIV between-patient matrix HIV-Bm (Nickle et al., 2007). |
| HIVw | viral | HIV within-patient matrix HIV-Wm (Nickle et al., 2007). |
| JTT | nuclear | General matrix (Jones et al., 1992). |
| JTTDCMut | nuclear | Revised `JTT` matrix (Kosiol and Goldman, 2005). |
| LG | nuclear | General matrix (Le and Gascuel, 2008). |
| mtART | mitochondrial | Mitochondrial Arthropoda (Abascal et al., 2007). |
| mtMAM | mitochondrial | Mitochondrial Mammalia (Yang et al., 1998). |
| mtREV | mitochondrial | Mitochondrial Vertebrate (Adachi and Hasegawa, 1996). |
| mtZOA | mitochondrial | Mitochondrial Metazoa (Animals) (Rota-Stabelli et al., 2009). |
| mtMet | mitochondrial | Mitochondrial Metazoa (Vinh et al., 2017). |
| mtVer | mitochondrial | Mitochondrial Vertebrate (Vinh et al., 2017). |
| mtInv | mitochondrial | Mitochondrial Inverterbrate (Vinh et al., 2017). |
| NQ.bird | nuclear | Non-reversible Q matrix (Dang et al., 2022) estimated for birds (Jarvis et al., 2015). |
| NQ.insect | nuclear | Non-reversible Q matrix (Dang et al., 2022) estimated for insects (Misof et al., 2014). |
| NQ.mammal | nuclear | Non-reversible Q matrix (Dang et al., 2022) estimated for mammals (Wu et al., 2018). |
| NQ.pfam | nuclear | General non-reversible Q matrix (Dang et al., 2022) estimated from Pfam version 31 database (El-Gebali et al., 2018). |
| NQ.plant | nuclear | Non-reversible Q matrix (Dang et al., 2022) estimated for plants (Ran et al., 2018). |
| NQ.yeast | nuclear | Non-reversible Q matrix (Dang et al., 2022) estimated for yeasts (Shen et al., 2018). |
| Poisson | none | Equal amino-acid exchange rates and frequencies. |
| PMB | nuclear | Probability Matrix from Blocks, revised `BLOSUM` matrix (Veerassamy et al., 2004). |
| Q.bird | nuclear | Q matrix (Minh et al., 2021) estimated for birds (Jarvis et al., 2015). |
| Q.insect | nuclear | Q matrix (Minh et al., 2021) estimated for insects (Misof et al., 2014). |
| Q.mammal | nuclear | Q matrix (Minh et al., 2021) estimated for mammals (Wu et al., 2018). |
| Q.pfam | nuclear | General Q matrix (Minh et al., 2021) estimated from Pfam version 31 database (El-Gebali et al., 2018). |
| Q.plant | nuclear | Q matrix (Minh et al., 2021) estimated for plants (Ran et al., 2018). |
| Q.yeast | nuclear | Q matrix (Minh et al., 2021) estimated for yeasts (Shen et al., 2018). |
| rtREV | viral | Retrovirus (Dimmic et al., 2002). |
| VT | nuclear | General 'Variable Time' matrix (Mueller and Vingron, 2000). |
| WAG | nuclear | General matrix (Whelan and Goldman, 2001). |

## CODON MODELS

| Code | Genetic code meaning |
|---|---|
| CODON1 | The Standard Code (same as `-st CODON`) |
| CODON2 | The Vertebrate Mitochondrial Code |
| CODON3 | The Yeast Mitochondrial Code |
| CODON4 | The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code |
| CODON5 | The Invertebrate Mitochondrial Code |
| CODON6 | The Ciliate, Dasycladacean and Hexamita Nuclear Code |
| CODON9 | The Echinoderm and Flatworm Mitochondrial Code |
| CODON10 | The Euplotid Nuclear Code |
| CODON11 | The Bacterial, Archaeal and Plant Plastid Code |
| CODON12 | The Alternative Yeast Nuclear Code |
| CODON13 | The Ascidian Mitochondrial Code |
| CODON14 | The Alternative Flatworm Mitochondrial Code |
| CODON16 | Chlorophycean Mitochondrial Code |
| CODON21 | Trematode Mitochondrial Code |
| CODON22 | Scenedesmus obliquus Mitochondrial Code |
| CODON23 | Thraustochytrium Mitochondrial Code |
| CODON24 | Pterobranchia Mitochondrial Code |
| CODON25 | Candidate Division SR1 and Gracilibacteria Code |

| | |
|---|---|
| MG | Nonsynonymous/synonymous (dn/ds) rate ratio (Muse and Gaut, 1994). |
| MGK | Like `MG` with additional transition/transversion (ts/tv) rate ratio. |
| MG1KTS or MGKAP2 | Like `MG` with a transition rate (Kosiol et al., 2007). |
| MG1KTV or MGKAP3 | Like `MG` with a transversion rate (Kosiol et al., 2007). |
| MG2K or MGKAP4 | Like `MG` with a transition rate and a transversion rate (Kosiol et al., 2007). |
| GY | Nonsynonymous/synonymous and transition/transversion rate ratios (Goldman and Yang, 1994). |
| GY1KTS or GYKAP2 | Like `GY` with a transition rate (Kosiol et al., 2007). |
| GY1KTV or GYKAP3 | Like `GY` with a transversion rate (Kosiol et al., 2007). |

## FREQUENCIES

| FreqType | Explanation |
|---|---|
| +F | Empirical base frequencies. This is the default if the model has unequal base freq. In AliSim, if users neither specify base frequencies nor supply an input alignment, AliSim will generate base frequencies from empirical distributions. |
| +FQ | Equal base frequencies. |
| +FO | Optimized base frequencies by maximum-likelihood. |

## RATES

| 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. |

# **Models:** how to choose?

More complex models always have higher likelihood than simpler models; danger of overfitting

Therefore, penalize by number of parameters

<u>Criteria in ModelFinder and MixtureFinder</u>

Where n=sample size, k=# parameters, L=likelihood

$AIC = 2k - 2\ln(L)$

$AICc = AIC + (2k^2 + 2k)/(n - k - 1)$

**$BIC = \ln(n)k - 2\ln(L)$**

What do you want to learn from this model?

# **Models:** time-(ir)reversibility and (non)stationarity

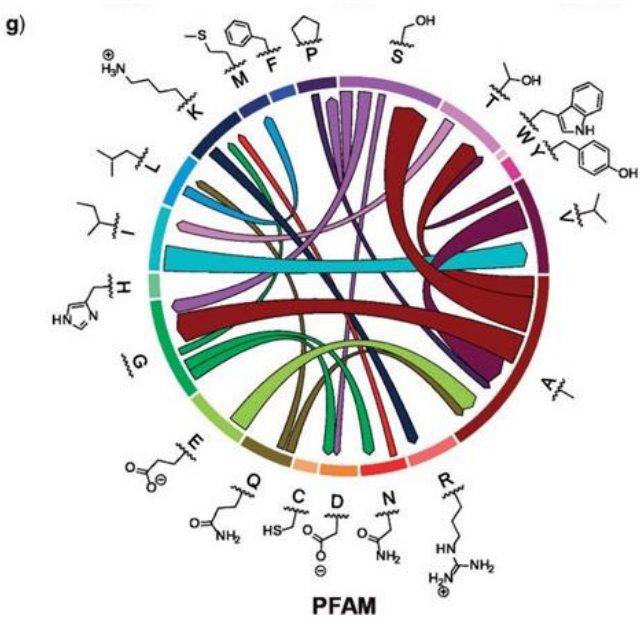
**Time-reversible =** can be described identically if run forward or backward.

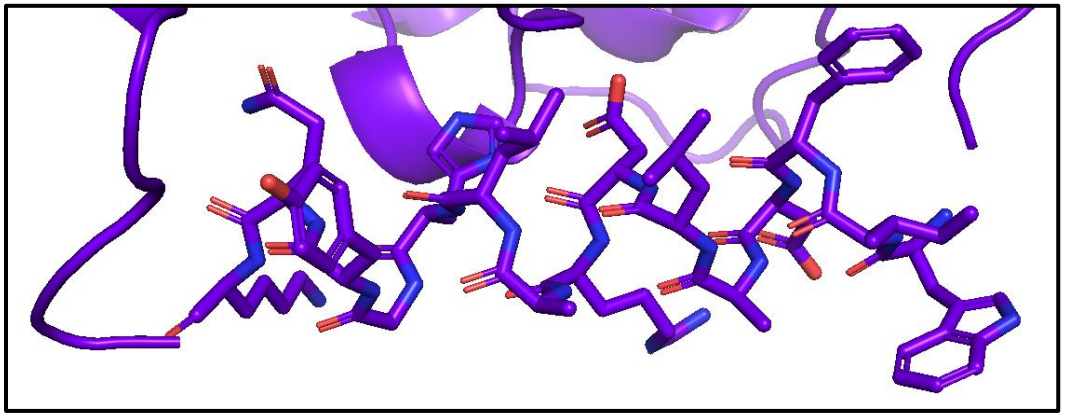
E.g. $r_{A \to C} = r_{C \to A}$

**Stationary =** mean and variance do not change over time.

E.g. trait frequencies and rates are at equilibrium across the tree

Sometimes these assumptions are seriously violated because evolution is often not a reversible or stationary process (e.g. epistasis, mutation bias)



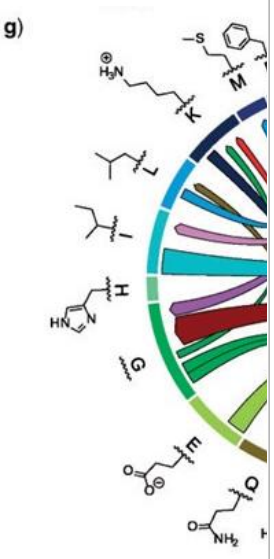
Dang et al. *Syst. Biol.* 2022

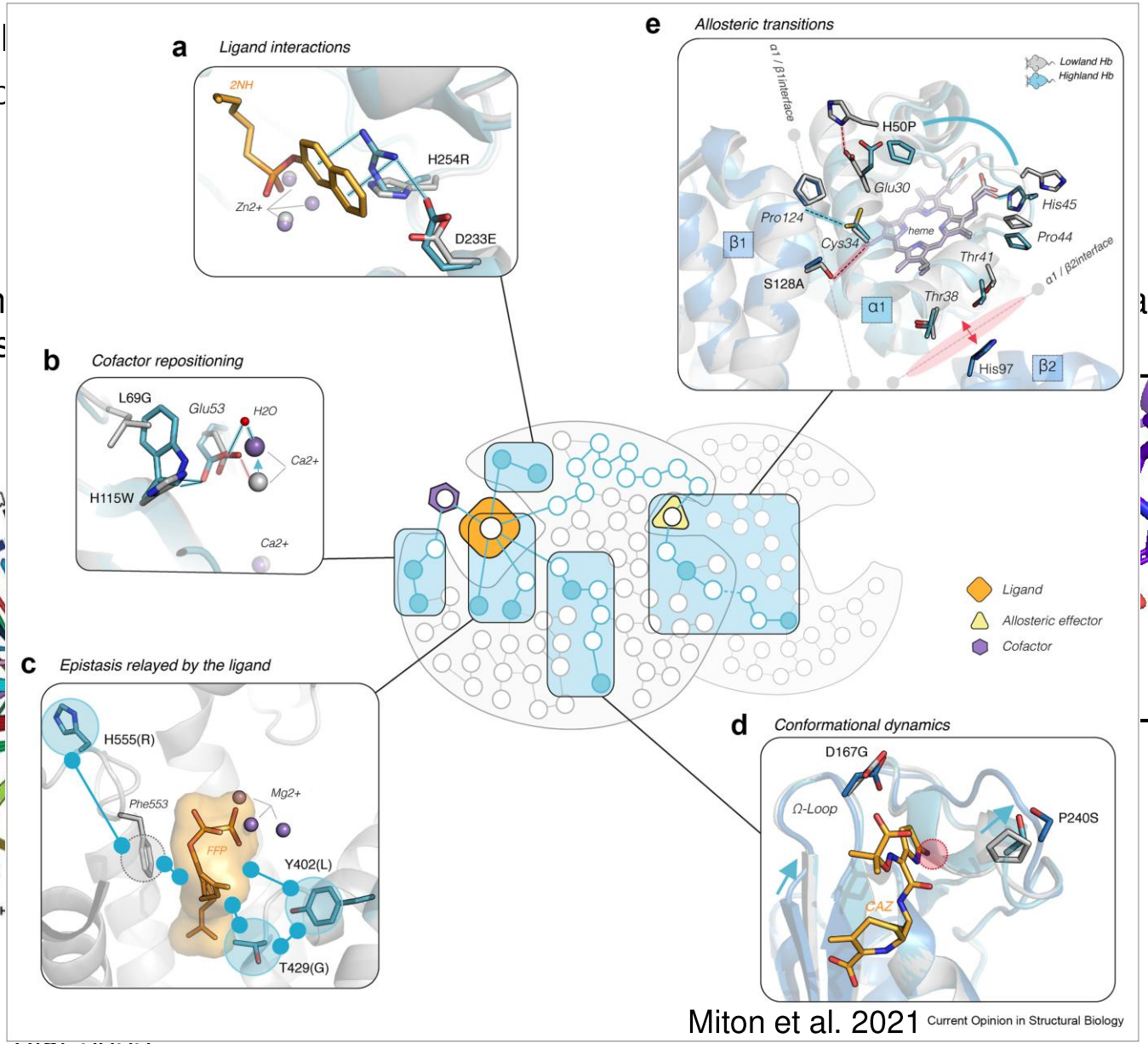
# **Models:** time-(ir)reversibility and (non)stationarity

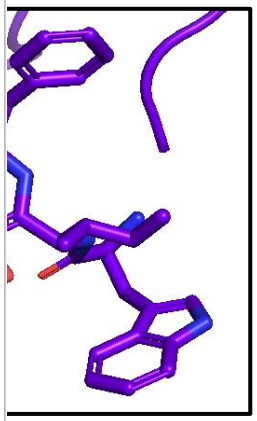
**Time-reversi**
E.g. $r_{A \rightarrow C}$

**Stationary =**
E.g. trait

Sometimes th
reversible or s



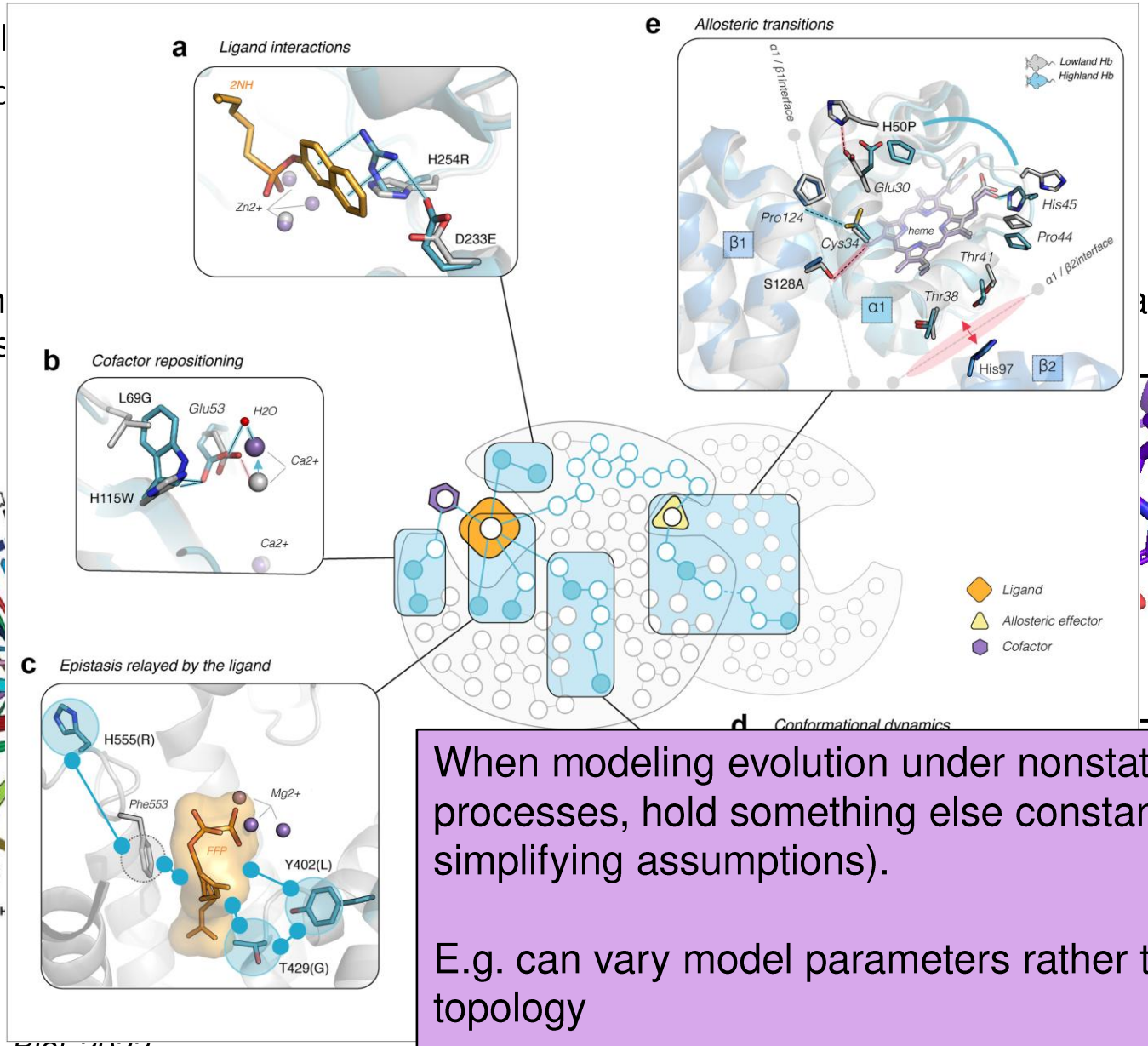
Dang et al. *Syst. Biol.* 2022

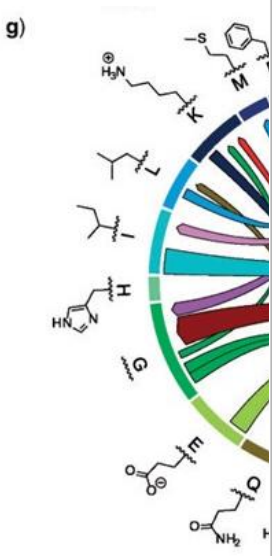
# **Models:** time-(ir)reversibility and (non)stationarity

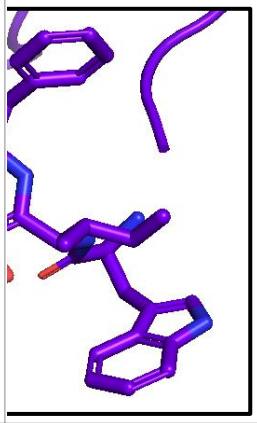
**Time-reversi**
E.g. $r_{A \to C}$

**Stationary =**
E.g. trait

Sometimes th
reversible or s



When modeling evolution under nonstationary processes, hold something else constant (make simplifying assumptions).

E.g. can vary model parameters rather than tree topology

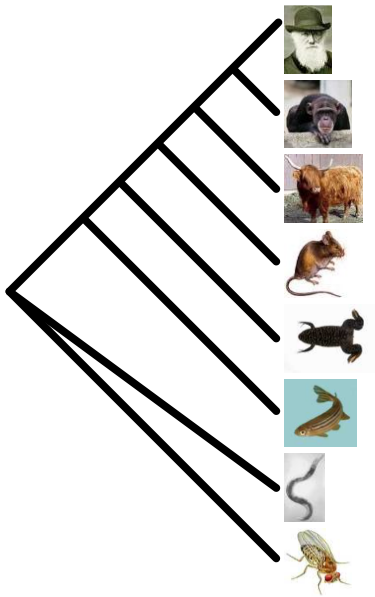
Dang et al. *Syst. Biol.* 2022

# **Models:** concatenation methods for genome-scale data

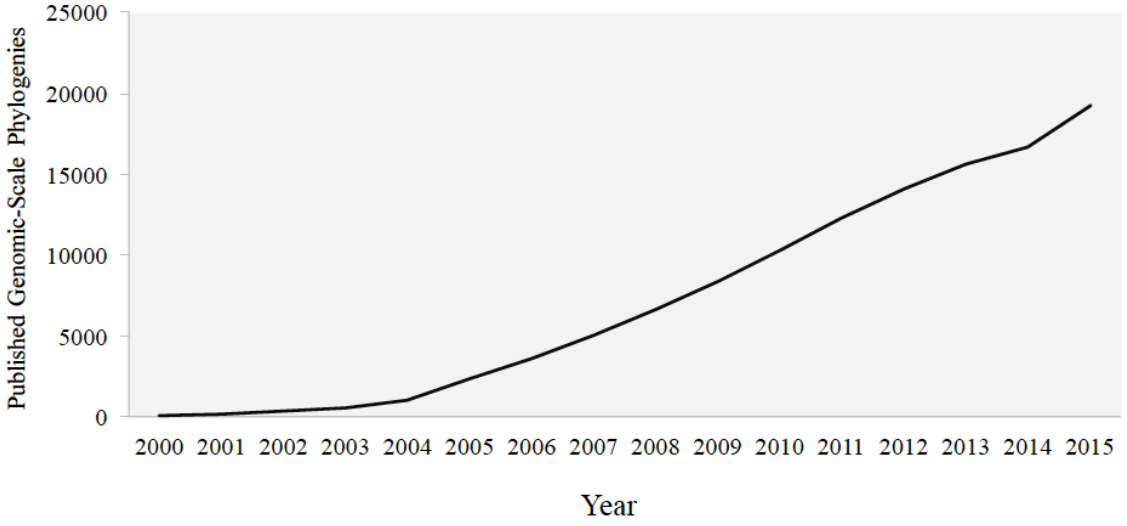
**Supermatrix**

| 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*



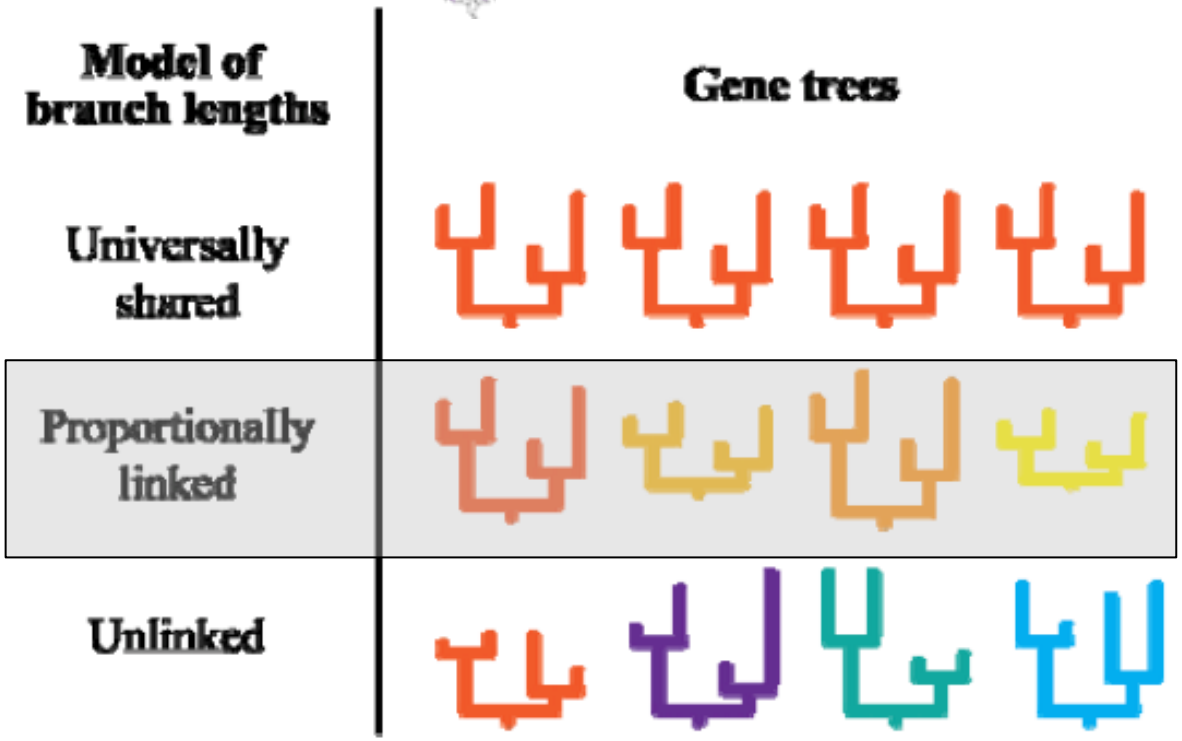

30 days of computation and 280 GB RAM for an insect data set!

# **Models:** partitioning data

**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




**iqtree3 -s ALN_FILE –q PARTITION_FILE**

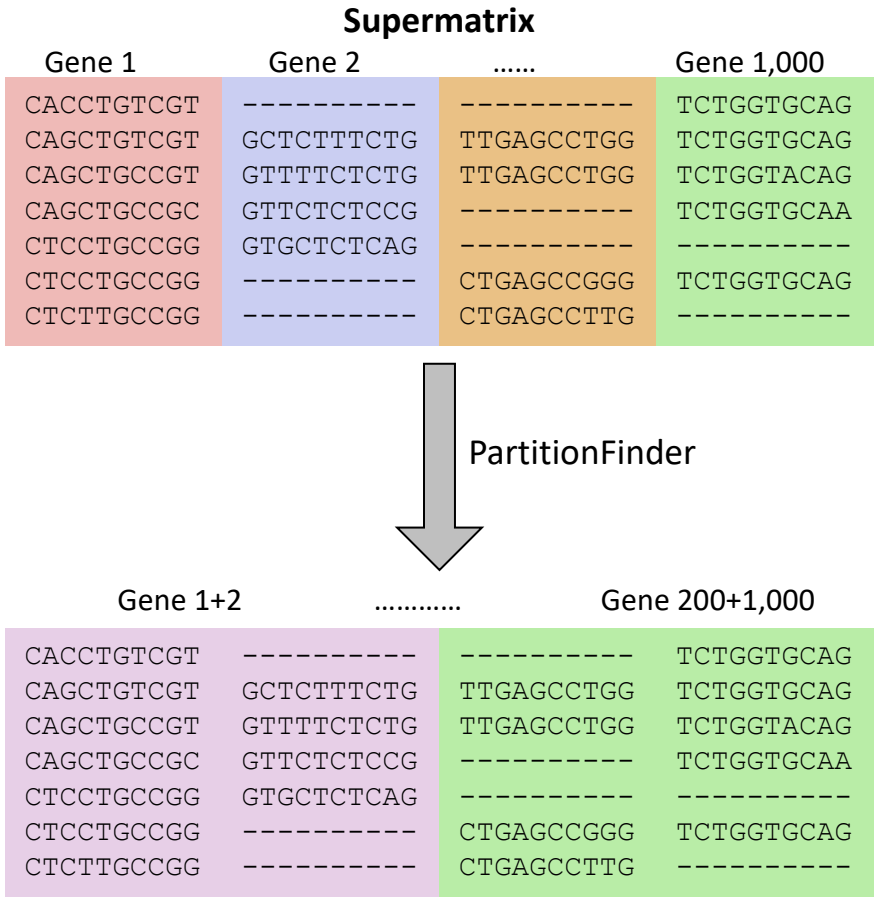
**iqtree3 -s ALN_FILE –p PARTITION_FILE**

**iqtree3 -s ALN_FILE –Q PARTITION_FILE**

# **Models:** 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;
```

# **Models:** how to reduce potential overfitting?




Substitution models: HKY ..... GTR+G

**PartitionFinder algorithm** (Lanfear et al. 2012):

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

iqtree3 … -m MFP+MERGE

**Relaxed clustering algorithm** (Lanfear et al. 2014):

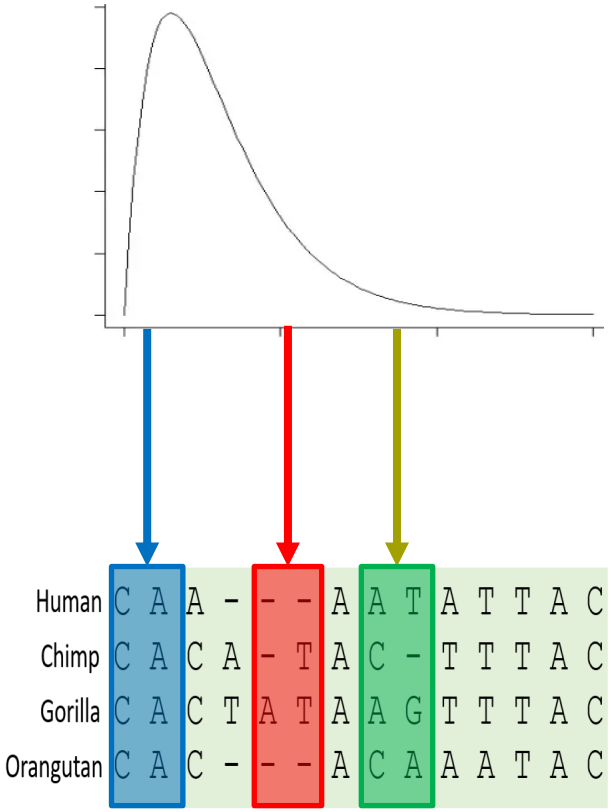
In step 1: only examine the top k% of most "promising" pairs.

iqtree3 … -rcluster 10

# **Models:** Greater realism with mixture models

Heterogeneity across sites

**Rate Model**

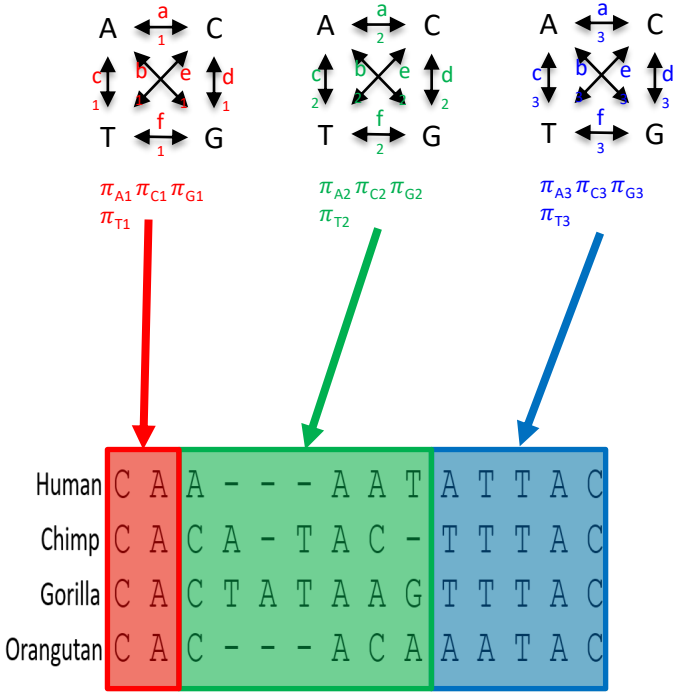



*From Huaiyan Ren*

# **Models:** Greater realism with mixture models

Heterogeneity across sites

Rate Model
**Partition model**



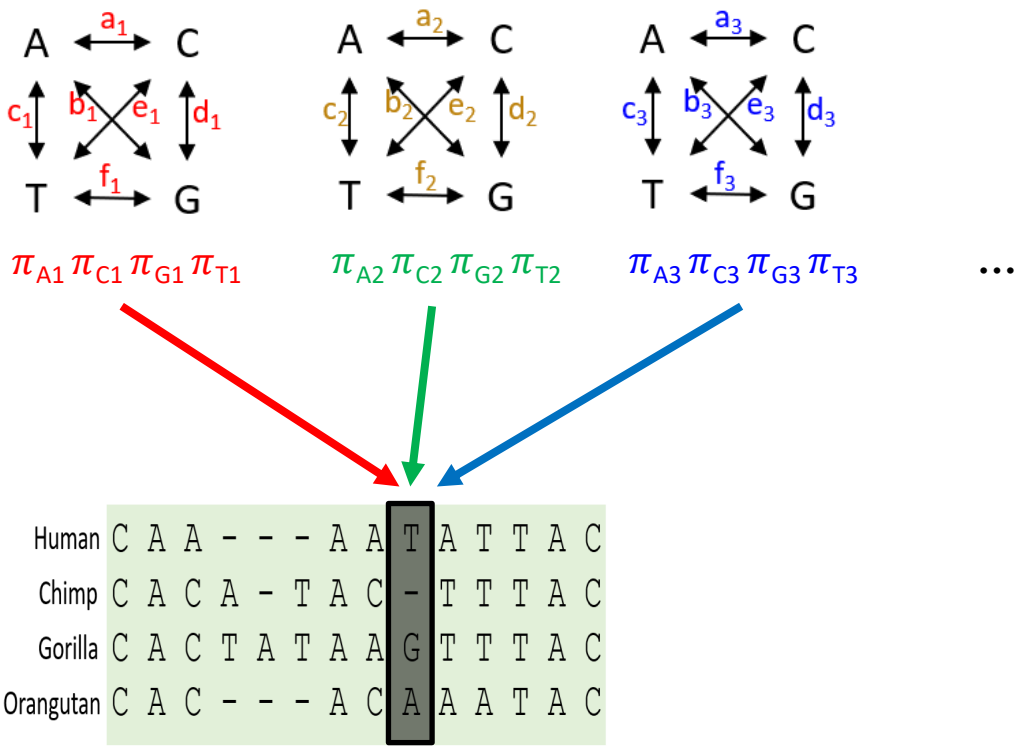

*From Huaiyan Ren*

# **Models:** Greater realism with mixture models

Heterogeneity across sites

Rate Model
Partition model

**Q mixture model**




*From Huaiyan Ren*

# IQ-TREE tree search algorithm

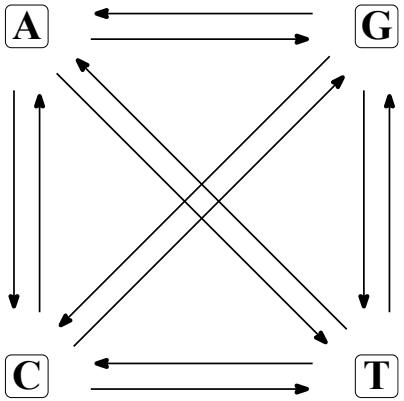
**Multiple sequence alignment**

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

**Model selection**

ModelFinder (2017)
PartitionFinder (2012, 2017)
MixtureFinder (2025)

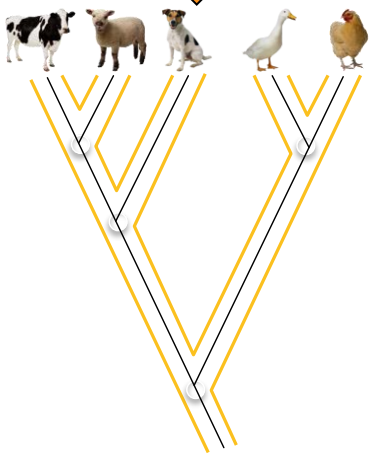
**Substitution model**




IQ-TREE implements a diversity of complex models

IQ-TREE (2015, 2020,2025)

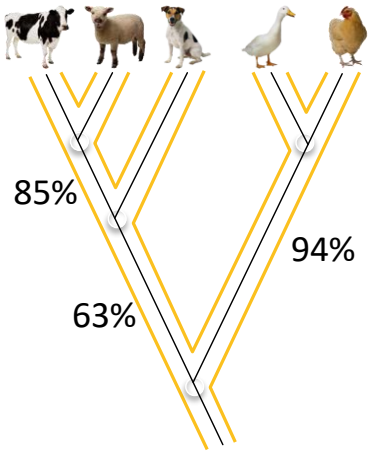
**Tree reconstruction**



**Phylogenetic tree**

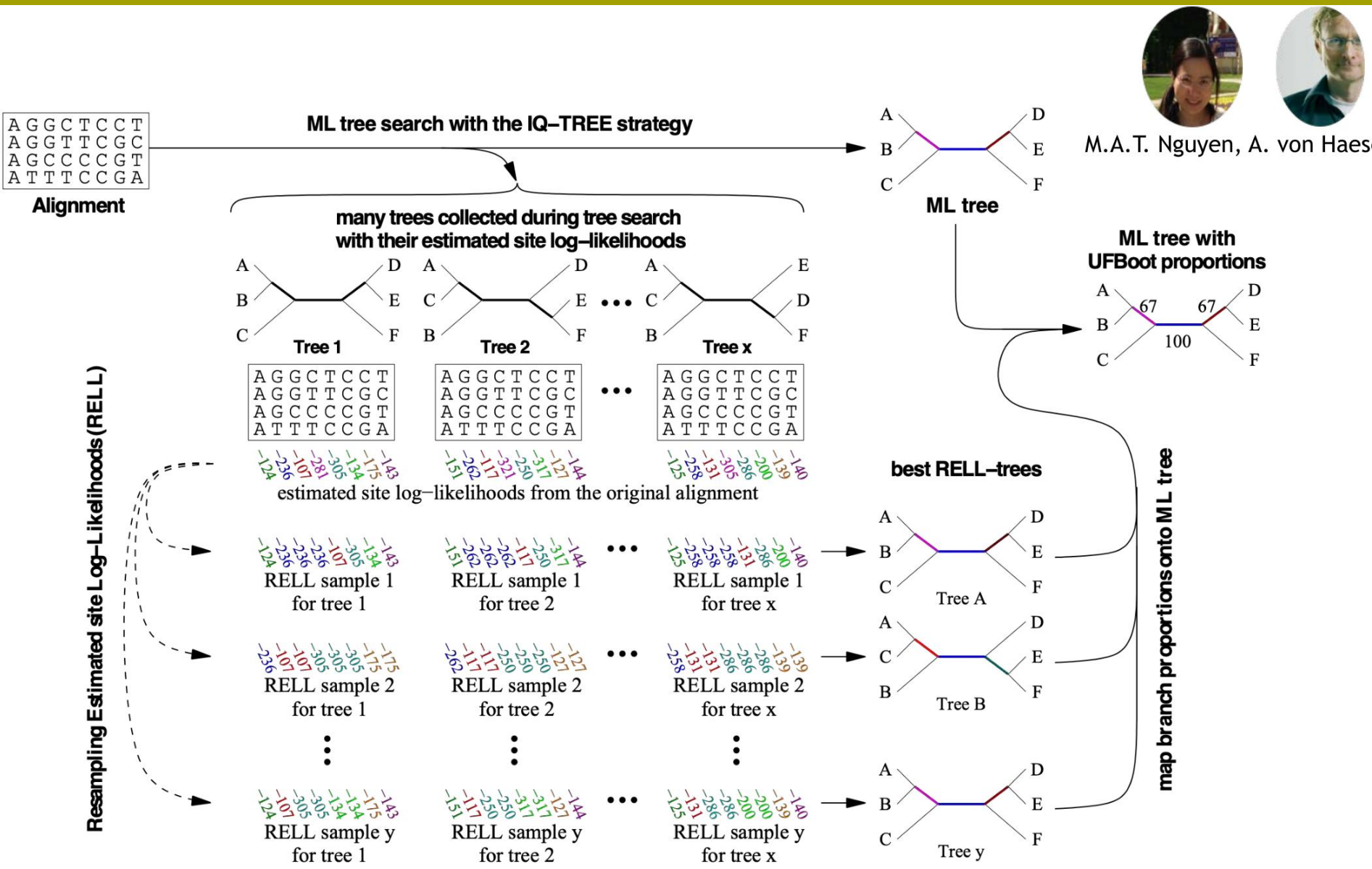
Ultrafast bootstrap (2013, 2018)

**Assessment of branch supports**




**Tree with branch supports**

# **Analysis:** bootstrap → ultrafast bootstrap approximation





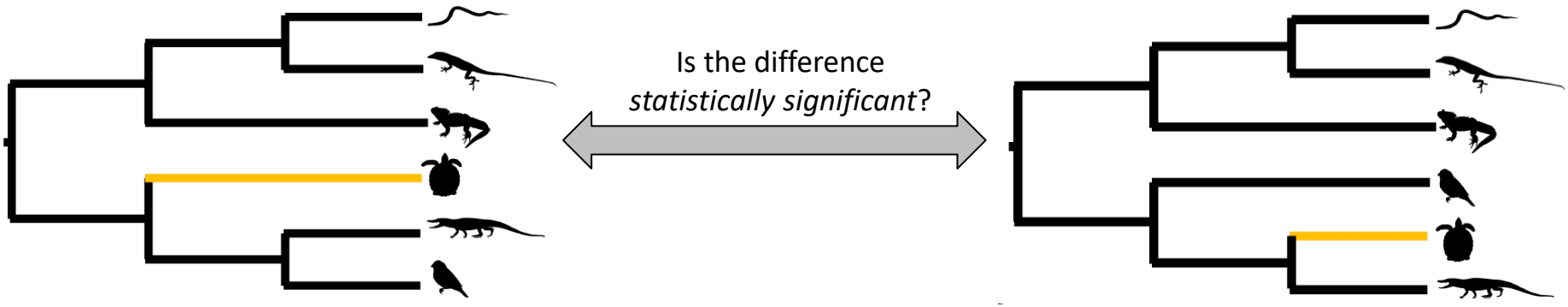



M.A.T. Nguyen, A. von Haesel

**iqtree3 -s ALN_FILE –b 100** for standard nonparametric boot

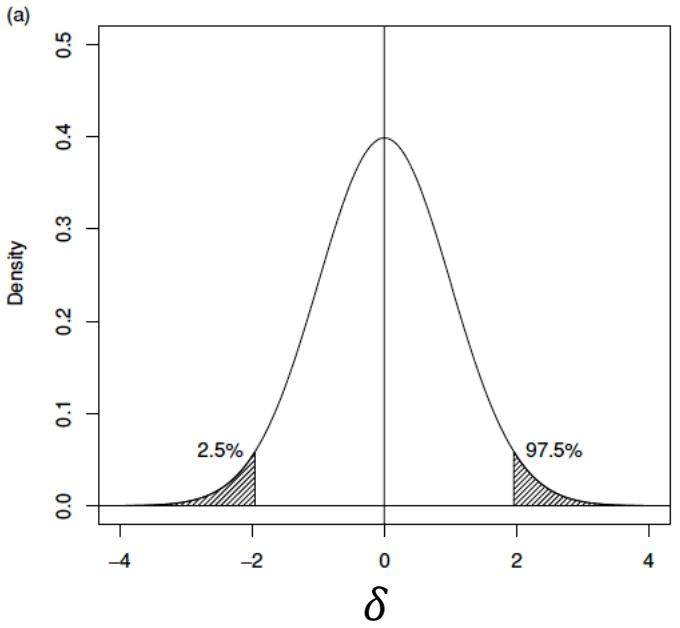
**iqtree3 -s ALN_FILE –B 1000** for ufboot

# **Analysis:** tree topology tests




**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 $\delta$ from many "random" data (e.g. by 10,000 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.



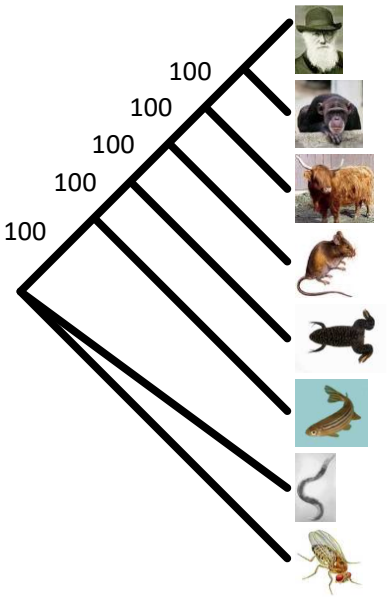

**iqtree3 -s ALN_FILE –p PARTITION_FILE**
**-z TREES_FILE –zb 10000 –au –n 0**

# **Analysis:** limitations of 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*

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

Concatenation assumes a single tree across all loci

Potential *systematic bias*

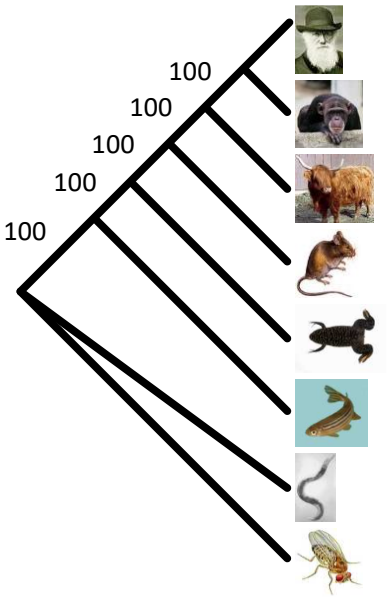
"*When the method of inferring phylogenies is one with undesirable statistical properties such as inconsistency, the bootstrap does not correct for these*" (Felsenstein, 1985)

# **Analysis:** limitations of 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*

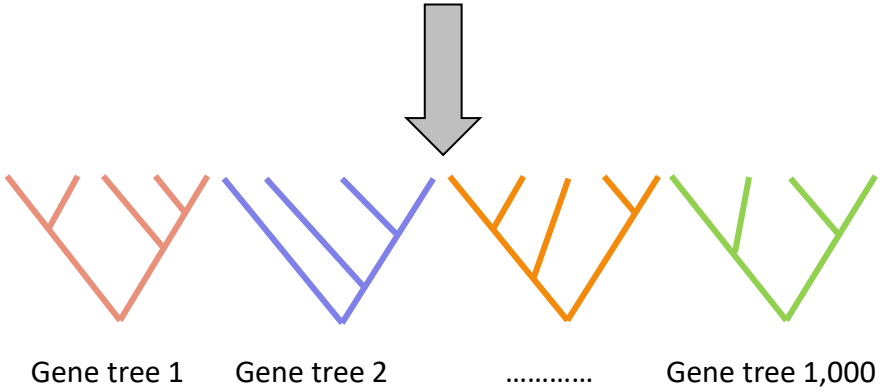
Bootstrap supports and Bayesian posteriors tend to 100% as #genes increases!

Concatenation assumes a single tree across all loci

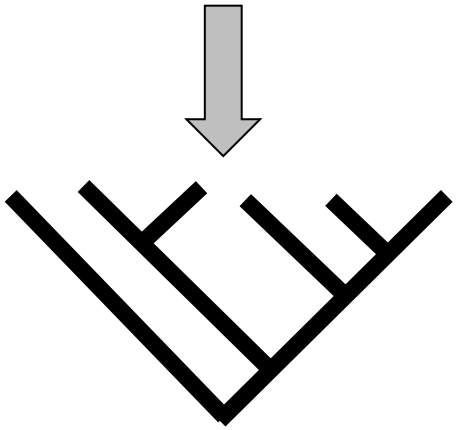
# **Analysis:** coalescent methods

**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 | ---------- |

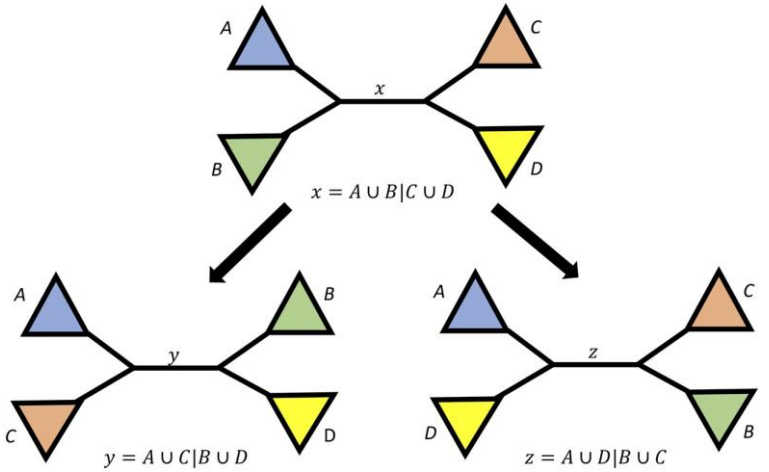


Gene tree 1 Gene tree 2 …… Gene tree 1,000



Species tree

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




Bui et al. *MBE* 2020

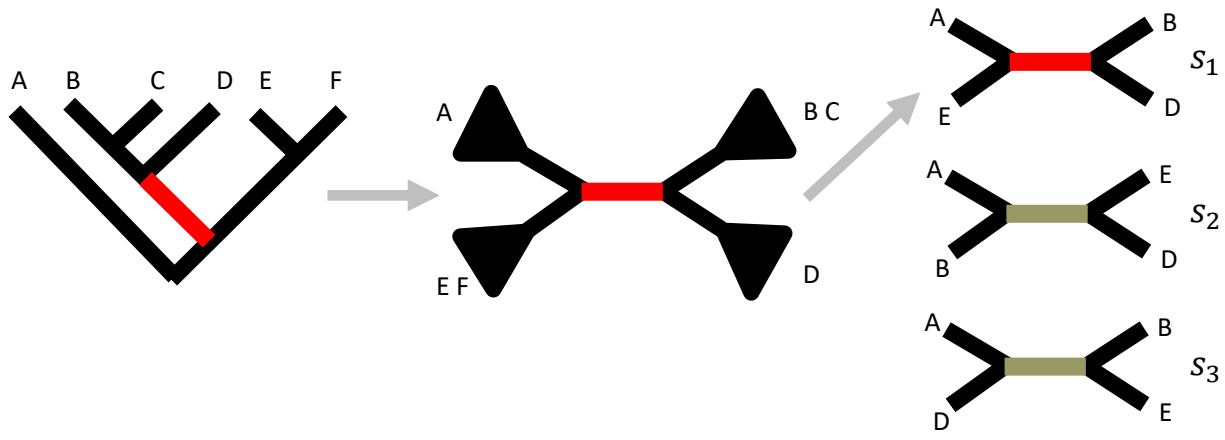
**Problem: Uncertainties in gene trees!**

# **Analysis:** 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%**

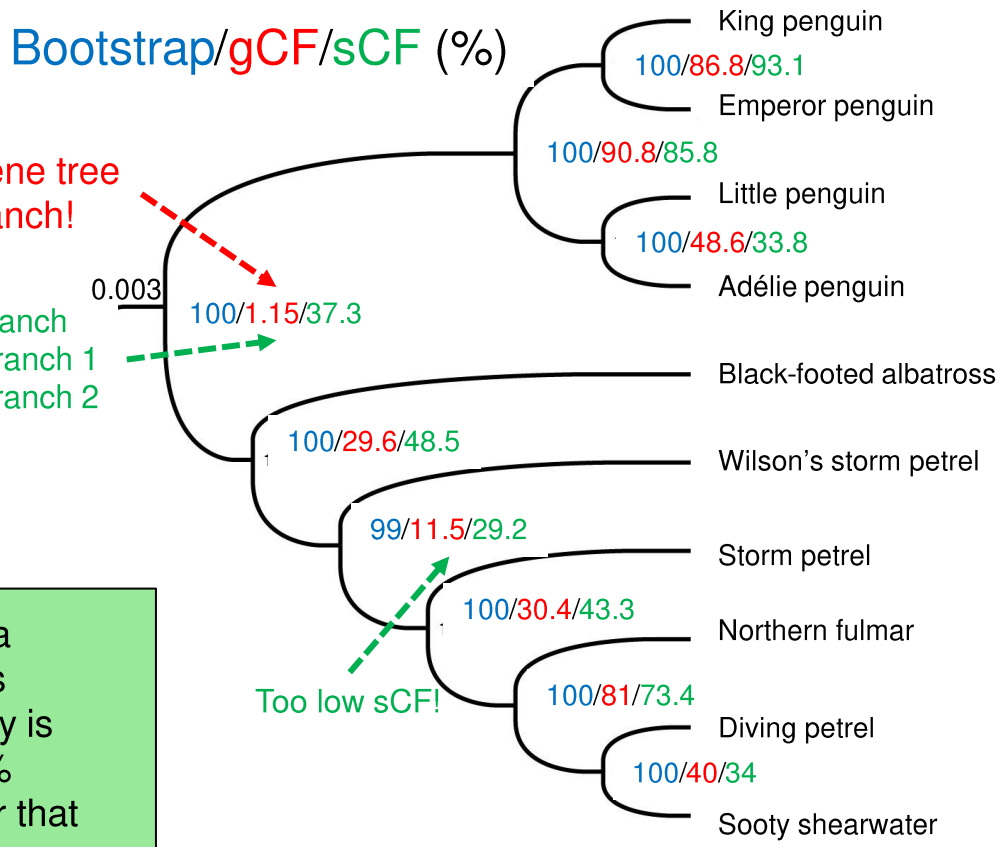



$$qCF(quartet) = \frac{s_1}{s_1 + s_2 + s_3}$$

$$sCF = \overline{qCF(100\ quartets)}$$

**Problem: parsimony**

# **Analysis:** an example from birds




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

Too low sCF!

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

Penguins

Tubenoses

- 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%.

# **Models:** Mixture Across Sites and Trees (MAST) model

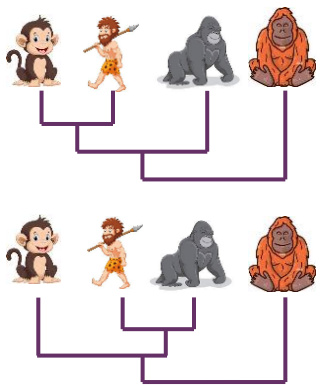
Concatenated alignment + hypothesized trees

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 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 |

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

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

Calculates tree weights

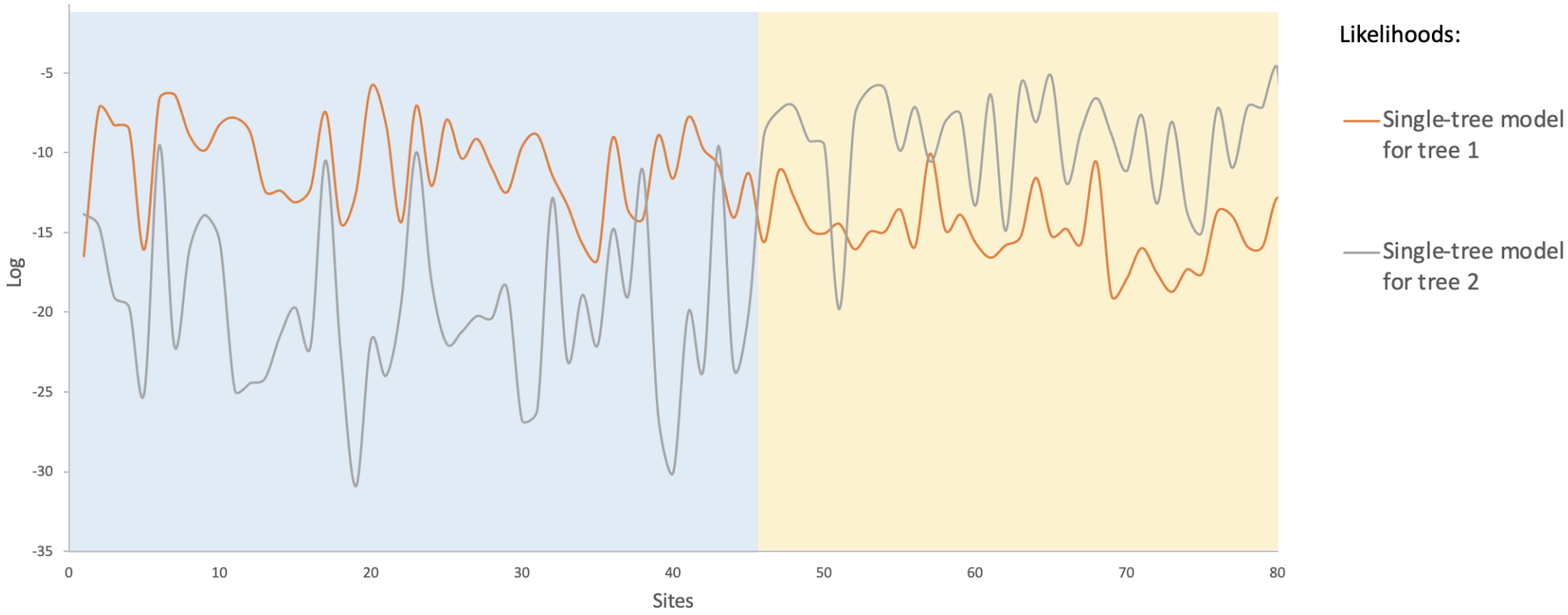



Likelihood for site *i*: $L_i = w_1 L_i^1 + w_1 L_i^2$
where $w_j$ represents the portion of sites belonging to tree *j*

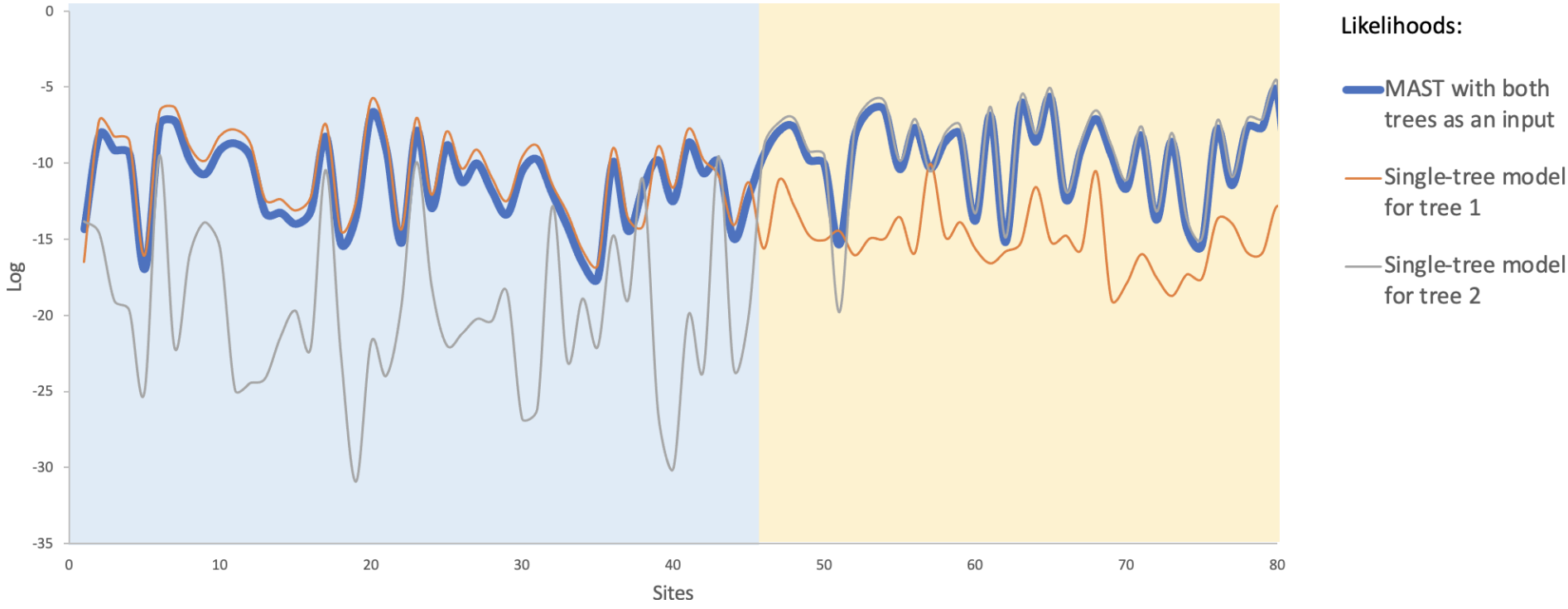
Log-likelihood of the trees: $\sum_i \log(L_i)$

**iqtree3 -s ALN_FILE –te TREES_FILE –m GTR+G+T**

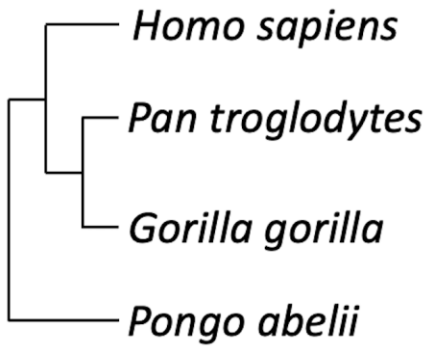
# Toy example: Site log-likelihood



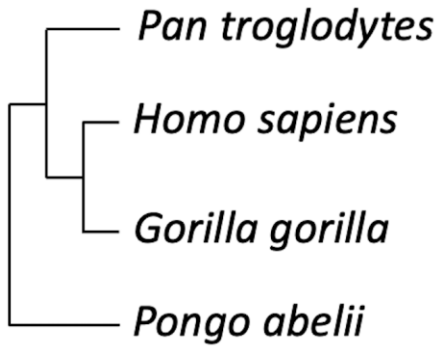
# Toy example: Site log-likelihood



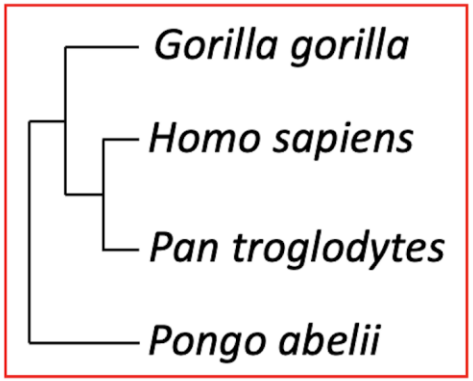
# The classical example of Human, Chimp, Gorilla




$T_{A1}$




$T_{A2}$

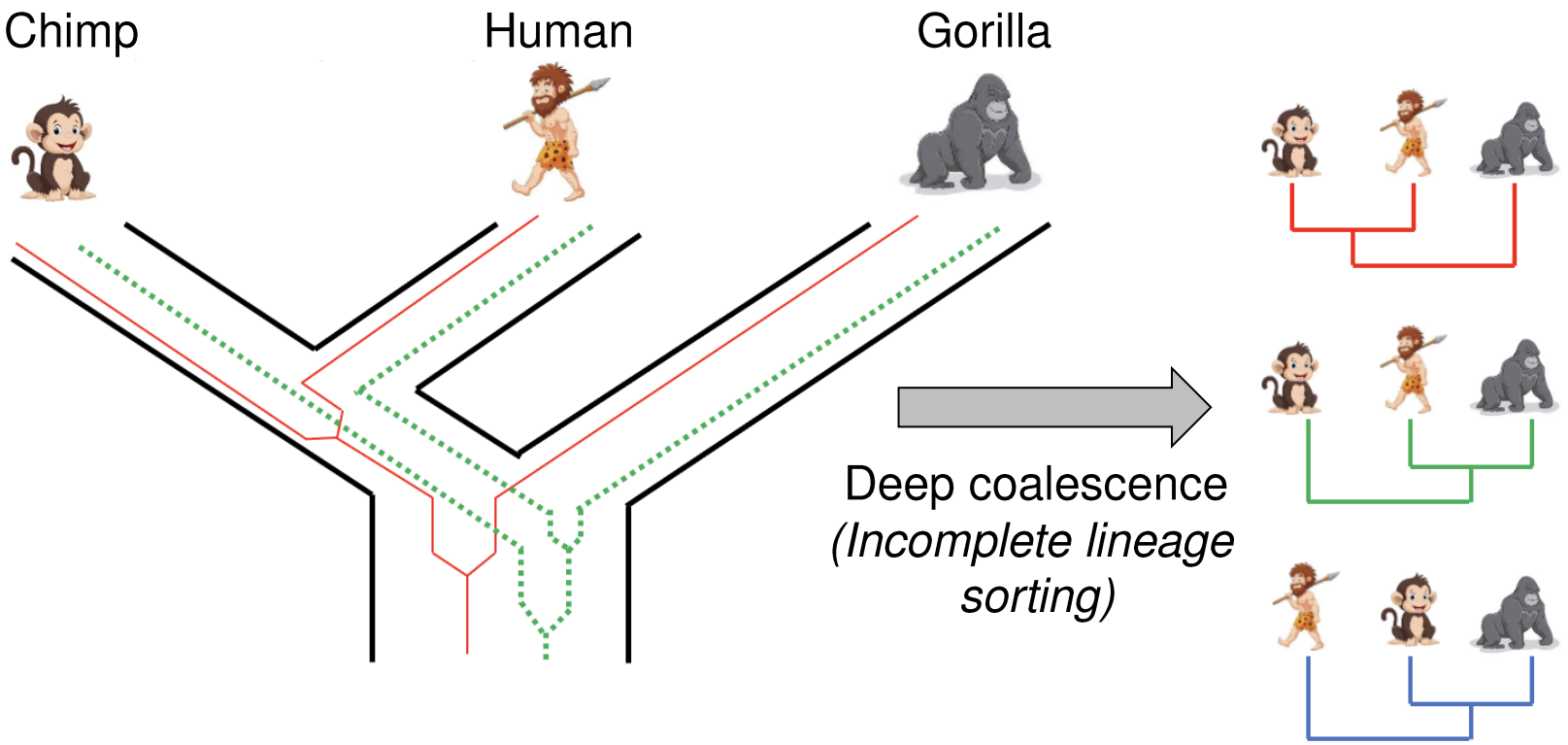



$T_{A3}$

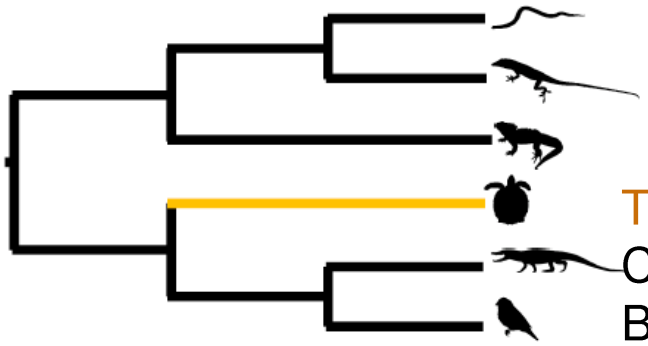
| | | | |
|---|---|---|---|
| Gene tree frequencies: | 19.8% | 20.1% | 60.1% |
| **MAST model weights:** | **17.9%** | **17.4%** | **64.7%** |

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

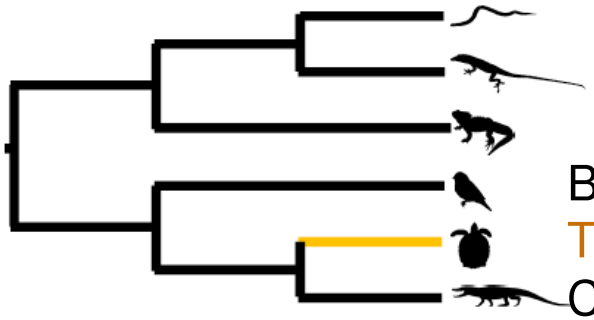
# Gene trees discordance due to deep coalescence



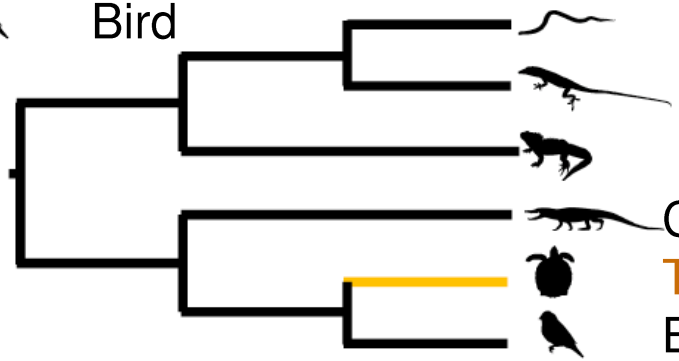
# Dataset for IQ-TREE lab: Where is Turtle in the tree?



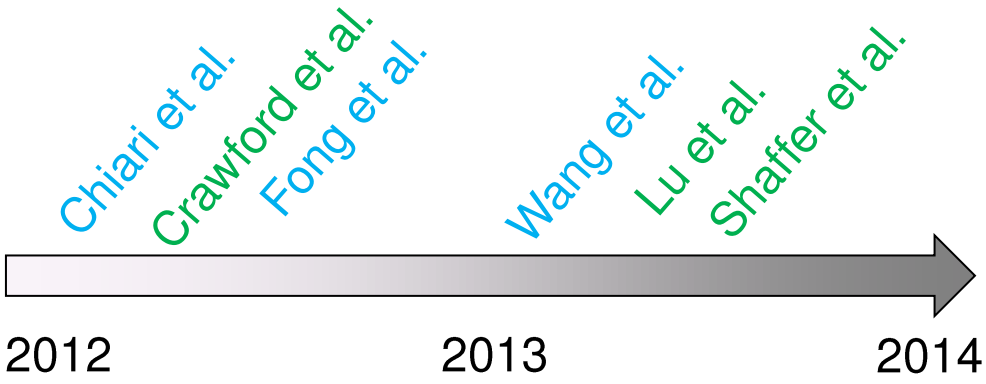
Turtle
Crocodile
Bird



Bird
Turtle
Crocodile



Crocodile
Turtle
Bird

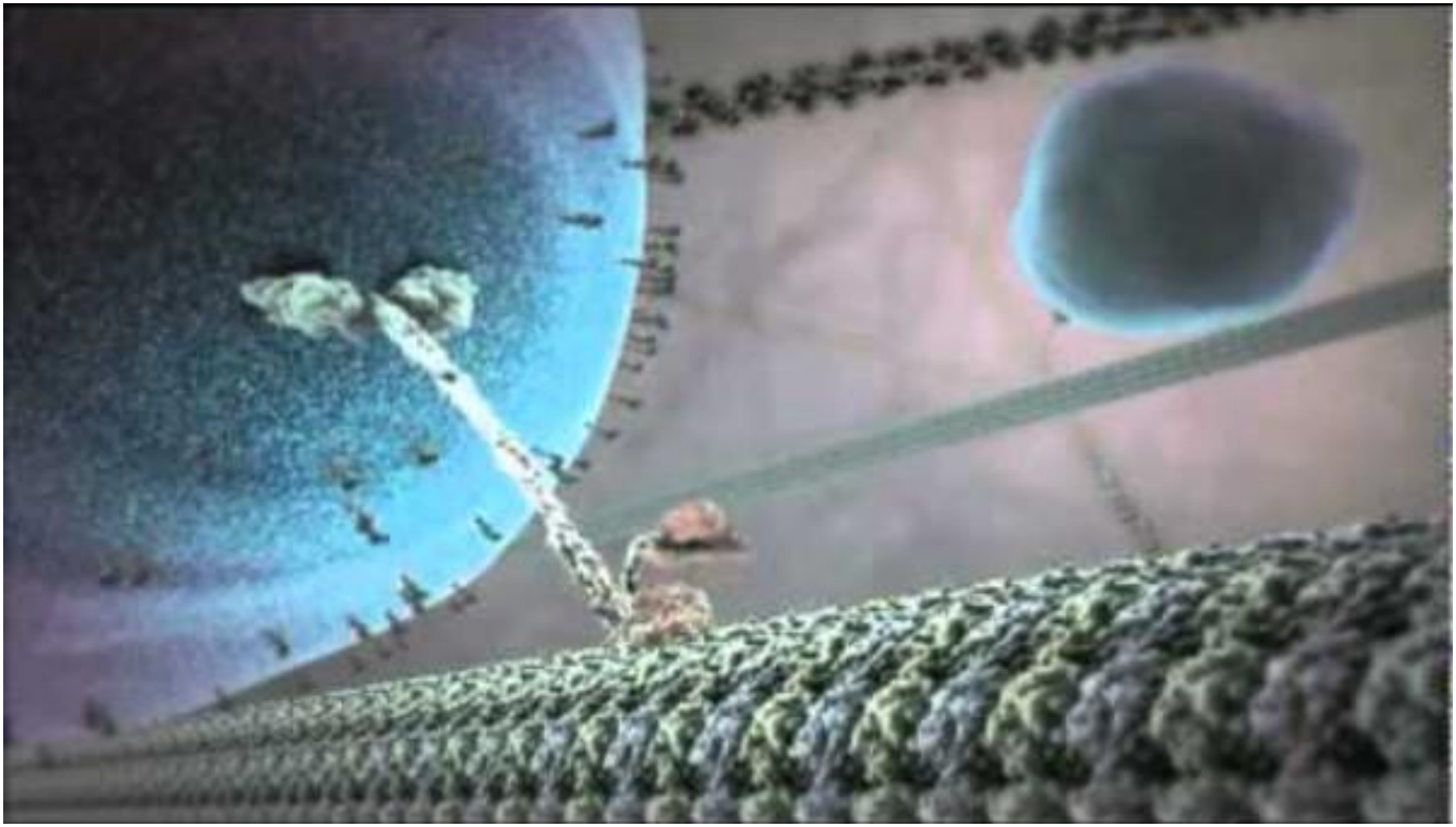



Different studies led to different trees!

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

Thanks Jeremy Brown

# The molecules of MOLE



**The Inner Life of the Cell**
by Bolinsky, Viel, Lue, et al.

# 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. Tree mixture model
7. Identifying most influential genes
8. Removing influential genes
9. Concordance factors
10. Site mixture model

http://www.iqtree.org/workshop/molevol2025

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