Phylogenetics at the QIB 002

Leonardo de Oliveira Martins, PhD
QIB Head of Phylogenomics

Quadram Institute
Science Health Food Innovation

@leomrtns@mstdn.science

Motivation for phylogenetics

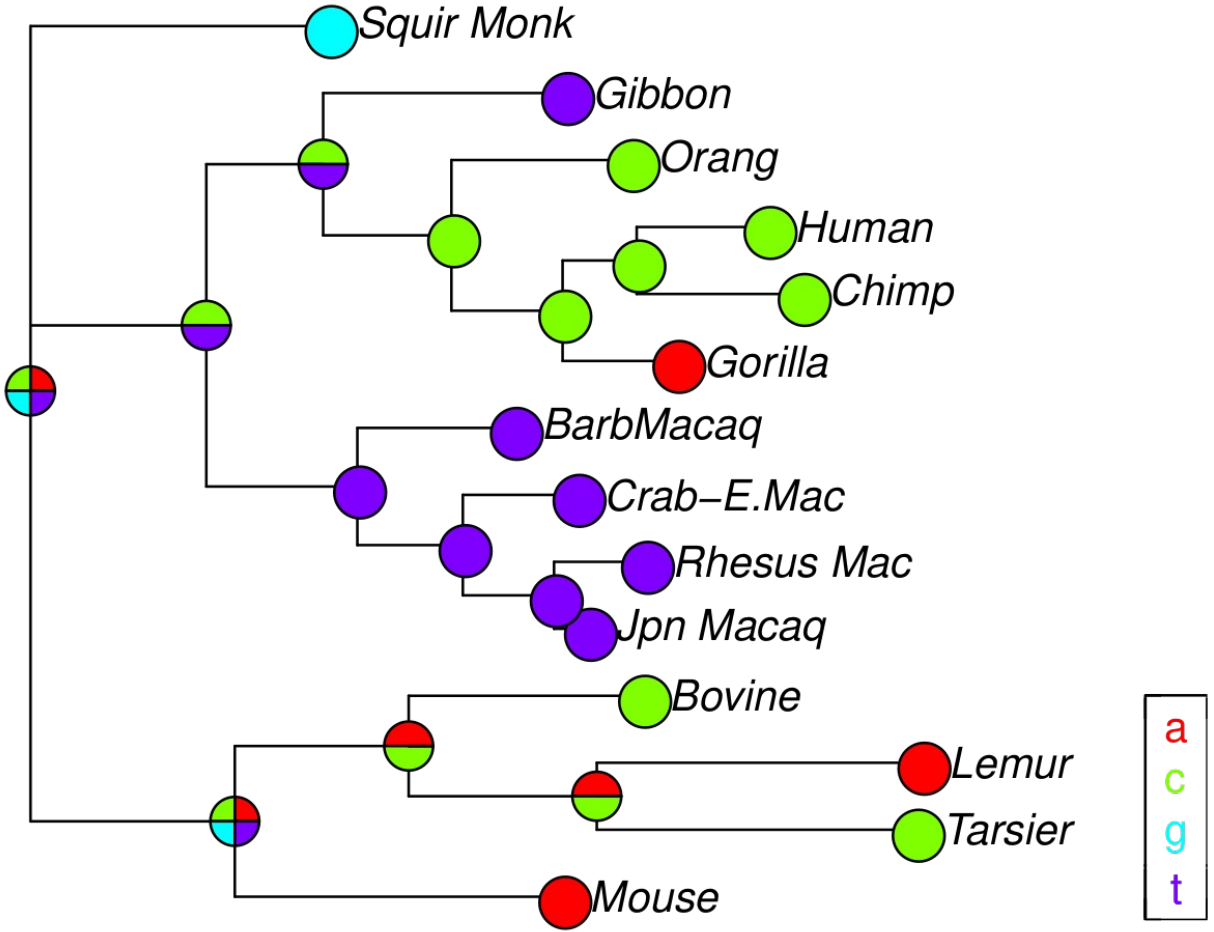
Ancestral State Reconstruction

Quadram Institute
Science Health Food Innovation

1. I know the “state” for these living organisms

2. I know the evolutionary tree

Can I estimate the states in the past?



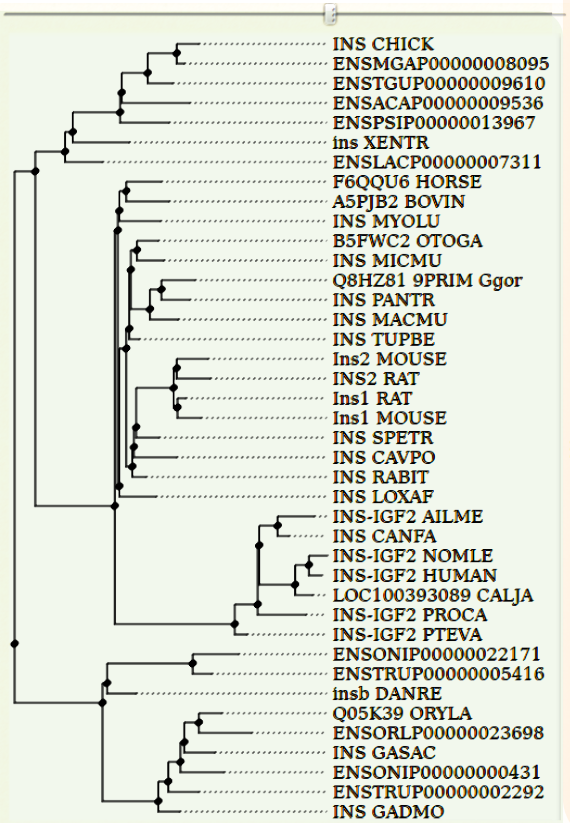
tree from phangorn’s vignette “Ancestral sequence reconstruction with phangorn”

Alignment

Multiple sequence alignments

Quadram Institute
Science Health Food Innovation

Alignment: homology (common ancestry) at every site



Phylogeny: evolution of the homologous sites along a tree

https://selectome.org/family/ENSGT00390000015440.1.Euteleostomi

Quadram Institute
Science • Health • Food • Innovation

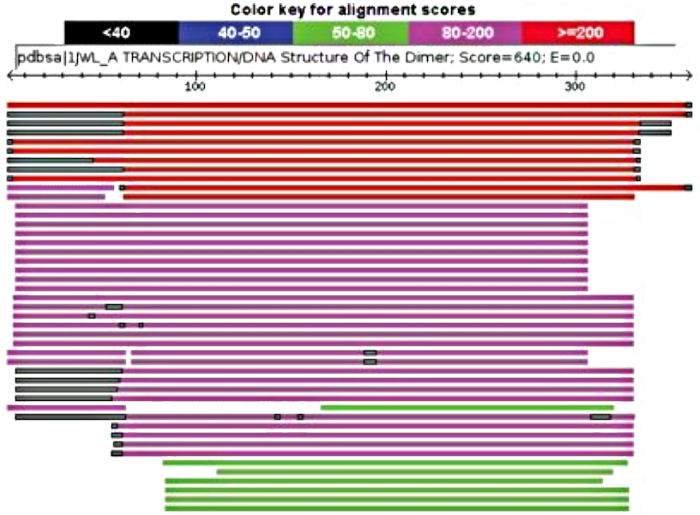
Composed of substitutions and indels (insertions/deletions)

Quadram Institute
Science & Health & Food & Innovation

Alignment = multiple sequence alignment (MSA)

→ similar, but not the same as read mapping

Alignment = multiple sequence alignment (MSA)

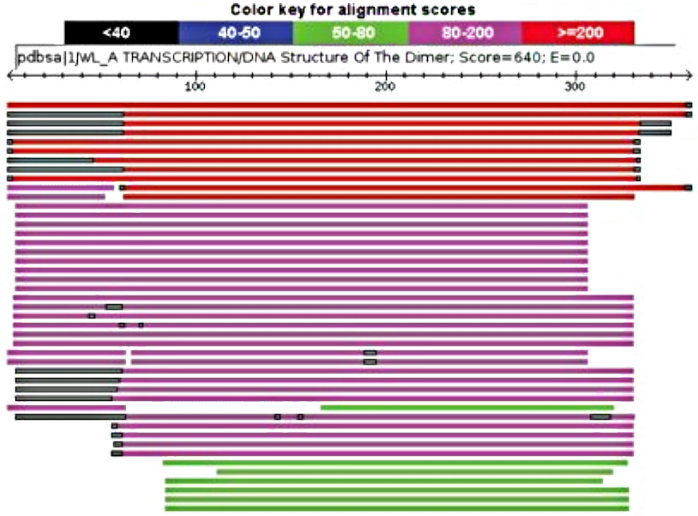


Sequences producing significant alignments:	Score (bits)	E value	Source DB	NCBI Entrez	Cath Prot/Chain
pdbsa|1LBG_A TRANSCRIPTION/DNA Lactose Operon Repressor Bo...	688	0	SDB	NCBI	CATH Prot
pdbsa|1LBH_A TRANSCRIPTION REGULATION Intact Lactose Opero...	688	0	SDB	NCBI	CATH Prot
pdbsa|1JYF_A TRANSCRIPTION Structure Of The Dimeric Lac Re...	669	0	SDB	NCBI	CATH Prot
pdbsa|1JYE_A TRANSCRIPTION Structure Of A Dimeric Lac Repr...	666	0	SDB	NCBI	CATH Prot
pdbsa|1JWL_A TRANSCRIPTION/DNA Structure Of The Dimeric La...	640	0	SDB	NCBI	CATH Prot
pdbsa|1EFA_A TRANSCRIPTION/DNA Crystal Structure Of The La...	640	0	SDB	NCBI	CATH Prot
pdbsa|1EFA_C TRANSCRIPTION/DNA Crystal Structure Of The La...	640	0	SDB	NCBI	CATH Prot
pdbsa|1JWL_C TRANSCRIPTION/DNA Structure Of The Dimeric La...	640	0	SDB	NCBI	CATH Prot
pdbsa|1EFA_B TRANSCRIPTION/DNA Crystal Structure Of The La...	640	0	SDB	NCBI	CATH Prot
pdbsa|1TLF_A TRANSCRIPTION REGULATION Unprecedented Quater...	575	1e-164	SDB	NCBI	CATH Prot

→ similar, but not the same as database search (e.g. BLAST)

https://doi.org/10.1093/nar/gkn237

Alignment = multiple sequence alignment (MSA)



Sequences producing significant alignments:	Score (bits)	E value	Source DB	NCBI Entrez	Cath Prot/Chain
pdbsa|1LBG_A TRANSCRIPTION/DNA Lactose Operon Repressor Bo...	688	0	SDB	NCBI	CATH Prot
pdbsa|1LBH_A TRANSCRIPTION REGULATION Intact Lactose Opero...	688	0	SDB	NCBI	CATH Prot
pdbsa|1JYF_A TRANSCRIPTION Structure Of The Dimeric Lac Re...	669	0	SDB	NCBI	CATH Prot
pdbsa|1JYE_A TRANSCRIPTION Structure Of A Dimeric Lac Repr...	666	0	SDB	NCBI	CATH Prot
pdbsa|1JWL_A TRANSCRIPTION/DNA Structure Of The Dimeric La...	640	0	SDB	NCBI	CATH Prot
pdbsa|1EFA_A TRANSCRIPTION/DNA Crystal Structure Of The La...	640	0	SDB	NCBI	CATH Prot
pdbsa|1EFA_C TRANSCRIPTION/DNA Crystal Structure Of The La...	640	0	SDB	NCBI	CATH Prot
pdbsa|1JWL_C TRANSCRIPTION/DNA Structure Of The Dimeric La...	640	0	SDB	NCBI	CATH Prot
pdbsa|1EFA_B TRANSCRIPTION/DNA Crystal Structure Of The La...	640	0	SDB	NCBI	CATH Prot
pdbsa|1TLF_A TRANSCRIPTION REGULATION Unprecedented Quater...	575	1e-164	SDB	NCBI	CATH Prot

DB search → Smith-Waterman (local)

	(empty string)	A	AA	AAC	AACC	AACCT	AACCTG
(empty string)	0	1	2	3	4	5	6
C	1	1	2	2	3	4	5
CC	2	2	2	2	2	3	4
CCC	3	3	3	2	2	3	4
CCCG	4	4	4	3	3	3	3

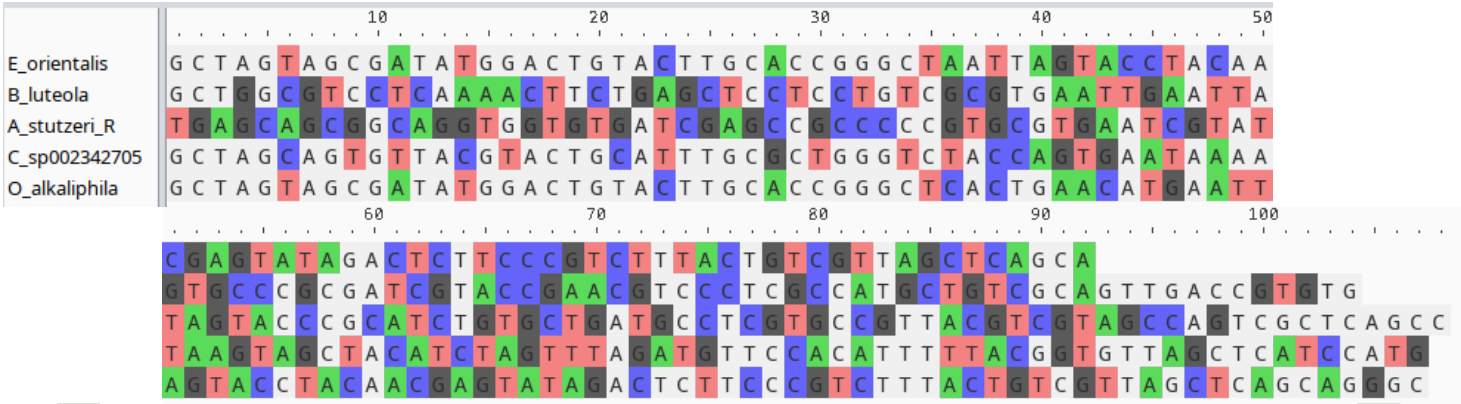
AACCTG
--CCCG

MSA → Needleman-Wunsch (global)

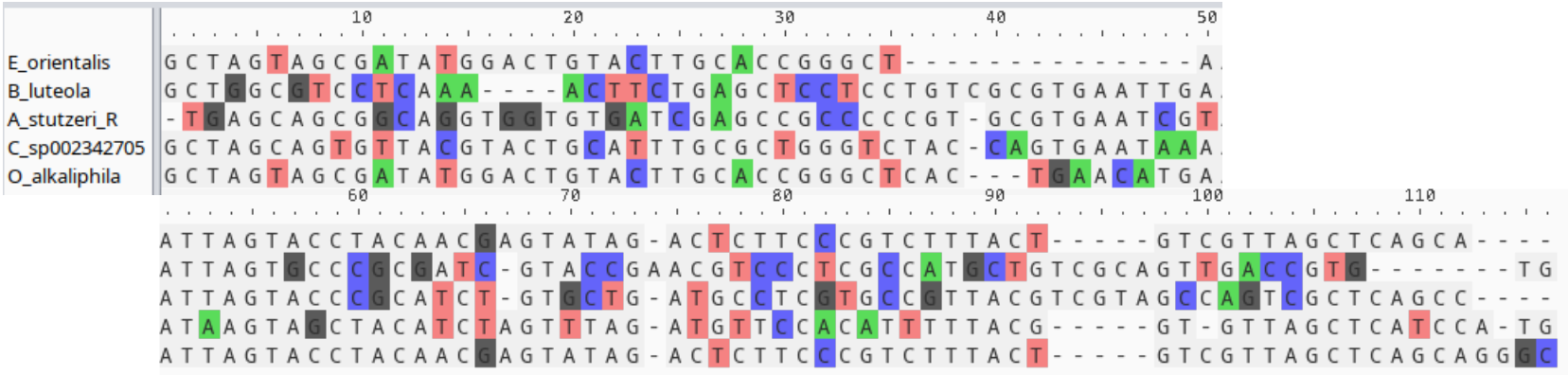
https://doi.org/10.1093/nar/gkn237

Alignment optimises homology by adding indels (–)

unaligned



aligned

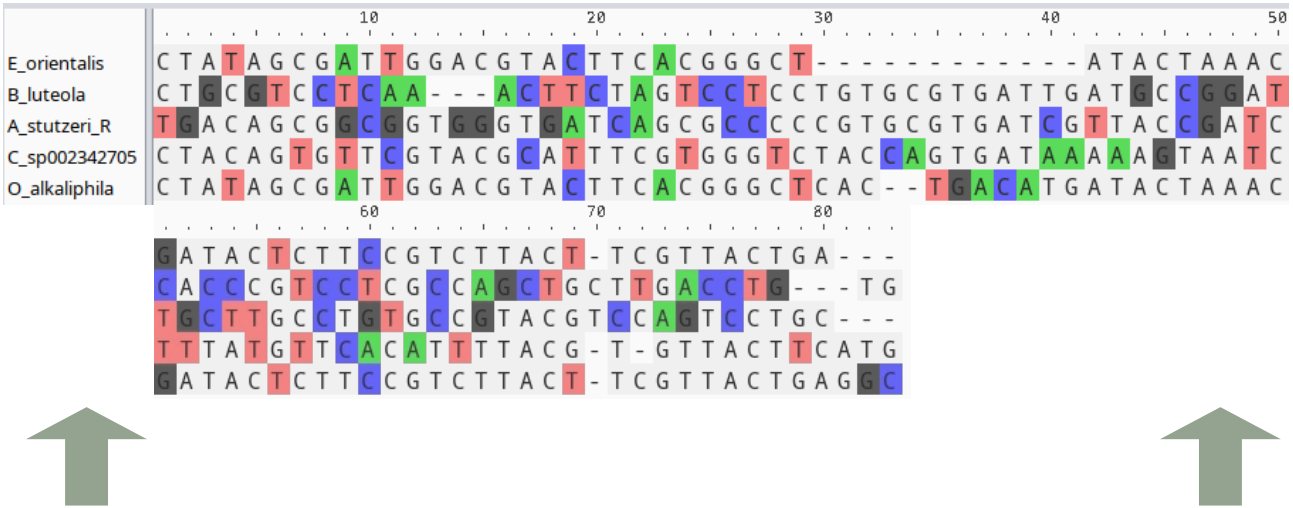


grey represent strict consensus over rows

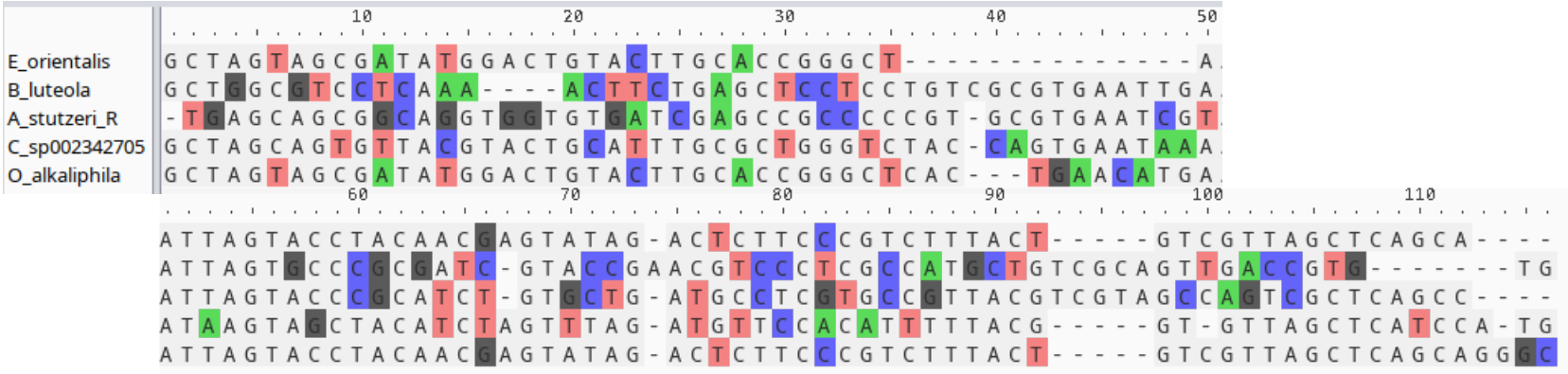
Quadram Institute
Science & Health & Food & Innovation

Sometimes we remove invariant sites

Only SNPs



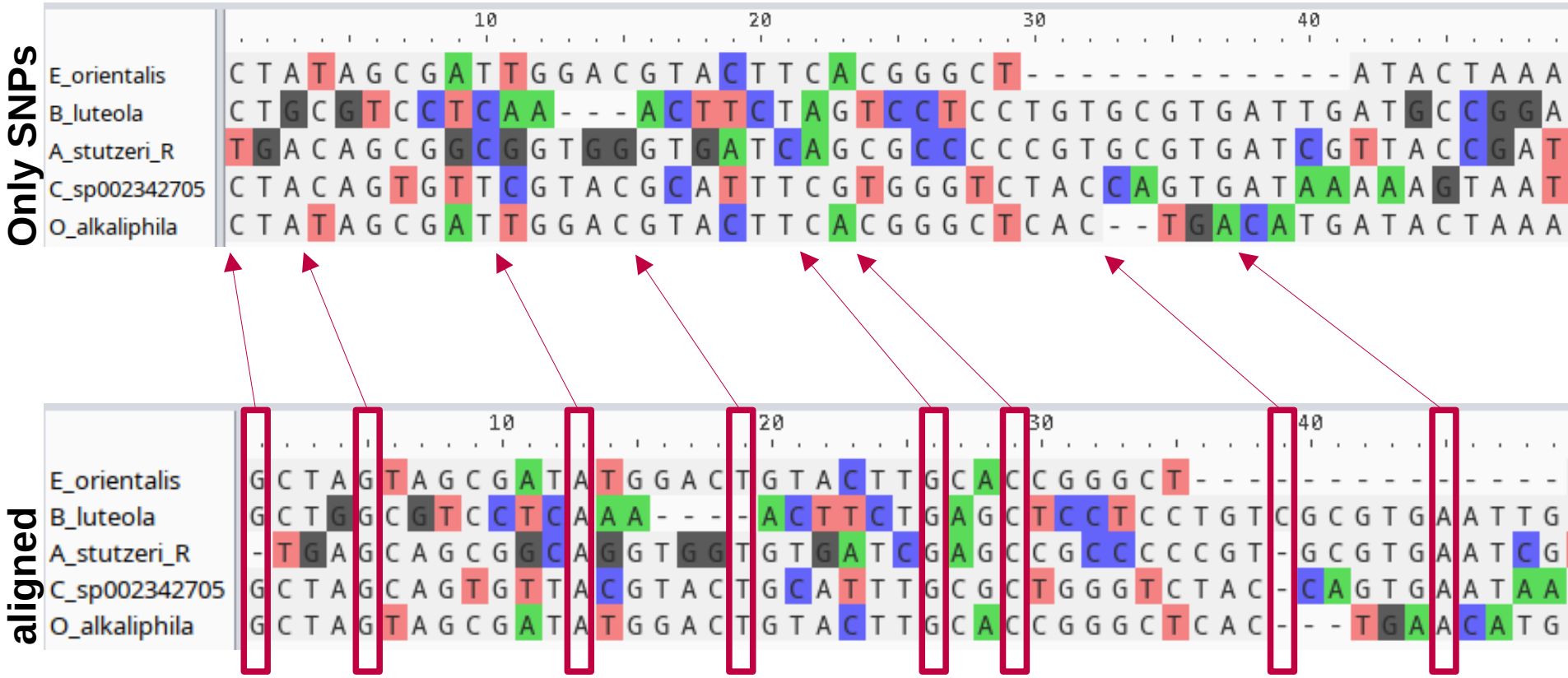
aligned



grey represent strict consensus over rows

Quadram Institute
Science & Health & Food & Innovation

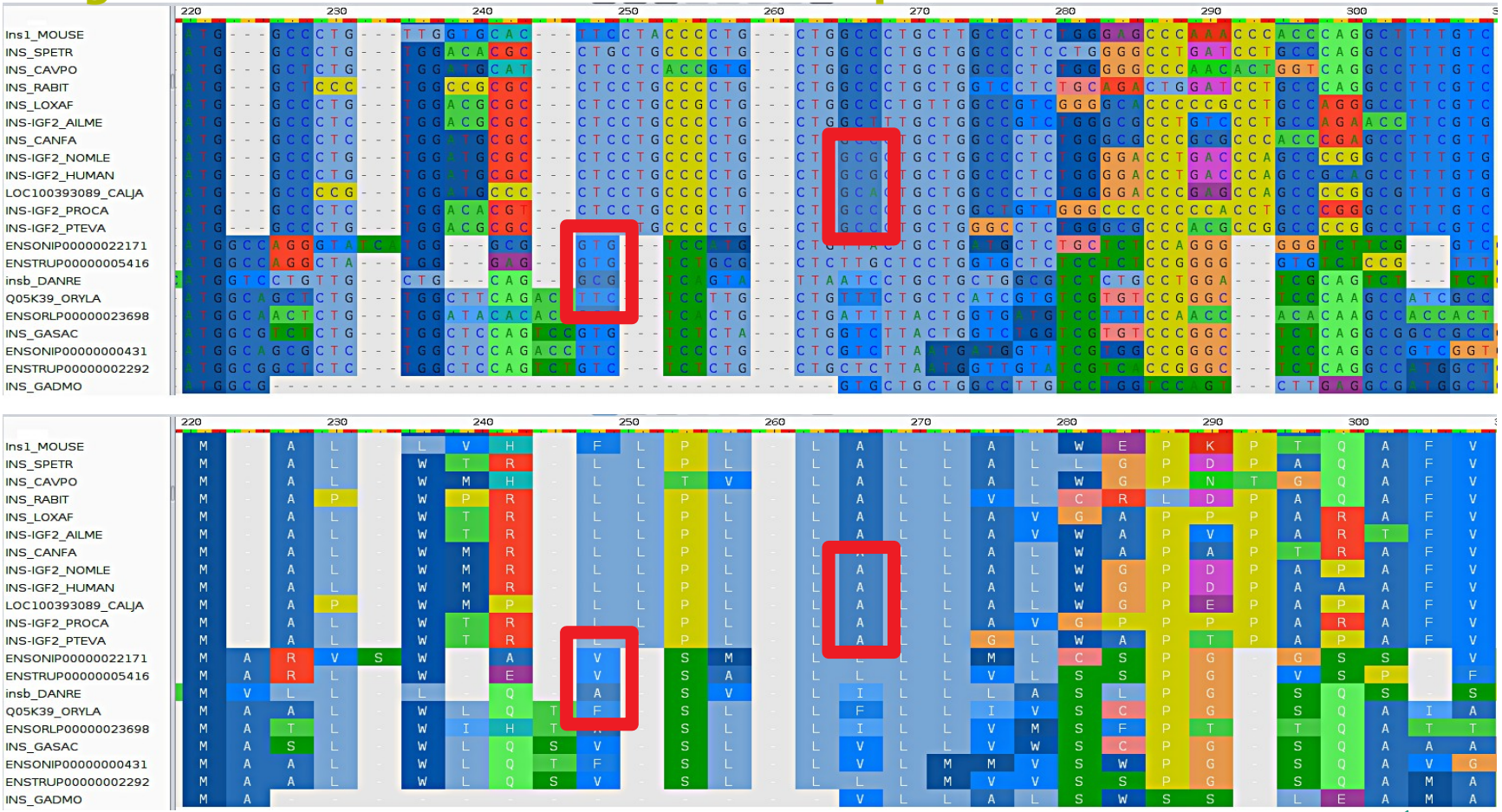
Sometimes we remove invariant sites



grey represent strict consensus over rows

Quadram Institute
Science & Health & Food & Innovation

Coding DNA → codons or amino acid sequences



https://selectome.org/euteleostomi/family/ENSGT00390000015440.1.Euteleostomi

Quadram Institute
Science & Health & Food & Innovation

Many sequence formats besides FASTA

```
 Known biosequence formats:
ID  Name              Read  Write  Int'leaf  Features  Sequence  Suffix  Content-type
 1  IG|Stanford       yes    yes       --        --       yes    .ig     biosequence/ig
 2  GenBank|gb        yes    yes       --       yes       yes    .gb     biosequence/genbank
 3  NBRF              yes    yes       --        --       yes    .nbrf   biosequence/nbrf
 4  EMBL|em           yes    yes       --       yes       yes    .embl   biosequence/embl
 5  GCG               yes    yes       --        --       yes    .gcg    biosequence/gcg
 6  DNAStrider        yes    yes       --        --       yes    .strider  biosequence/strider
 7  Fitch              --     --       --        --       yes    .fitch  biosequence/fitch
 8  Pearson|Fasta|fa   yes    yes       --        --       yes    .fasta  biosequence/fasta
 9  Zuker              --     --       --        --       yes    .zuker  biosequence/zuker
10  Olsen              --     --      yes        --       yes    .olsen  biosequence/olsen
11  Phylip3.2         yes    yes      yes        --       yes    .phylip2  biosequence/phylip2
12  Phylip|Phylip4    yes    yes      yes        --       yes    .phylip  biosequence/phylip
13  Plain|Raw         yes    yes       --        --       yes    .seq    biosequence/plain
14  PIR|CODATA        yes    yes       --        --       yes    .pir    biosequence/codata
15  MSF               yes    yes      yes        --       yes    .msf    biosequence/msf
16  ASN.1              --     --       --        --       yes    .asn    biosequence/asn1
17  PAUP|NEXUS        yes    yes      yes        --       yes    .nexus  biosequence/nexus
18  Pretty             --    yes      yes        --       yes    .pretty  biosequence/pretty
19  XML               yes    yes       --       yes       yes    .xml    biosequence/xml
20  BLAST             yes     --      yes        --       yes    .blast  biosequence/blast
21  SCF               yes     --       --        --       yes    .scf    biosequence/scf
22  Clustal           yes    yes      yes        --       yes    .aln    biosequence/clustal
23  FlatFeat|FFF      yes    yes       --       yes        --    .fff    biosequence/fff
24  GFF               yes    yes       --       yes        --    .gff    biosequence/gff
25  ACEDB             yes    yes       --        --       yes    .ace    biosequence/acedb
  (Int'leaf = interleaved format; Features = documentation/features are parsed)
```

Screenshot from *readseq* help page

Quadram Institute
Science & Health & Food & Innovation

phylip format for aligned sequences

Here is a hypothetical example of interleaved format:

```
  5    42
Turkey    AAGCTNGGGC ATTTCAGGGT
Salmo gairAAGCCTTGGC AGTGCAGGGT
H. SapiensACCGGTTGGC CGTTCAGGGT
Chimp     AAACCCTTGC CGTTACGCTT
Gorilla   AAACCCTTGC CGGTACGCTT

GAGCCCGGGC AATACAGGGT AT
GAGCCGTGGC CGGGCACGGT AT
ACAGGTTGGC CGTTCAGGGT AA
AAACCGAGGC CGGGACACTC AT
AAACCATTGC CGGTACGCTT AA
```

while in sequential format the same sequences would be:

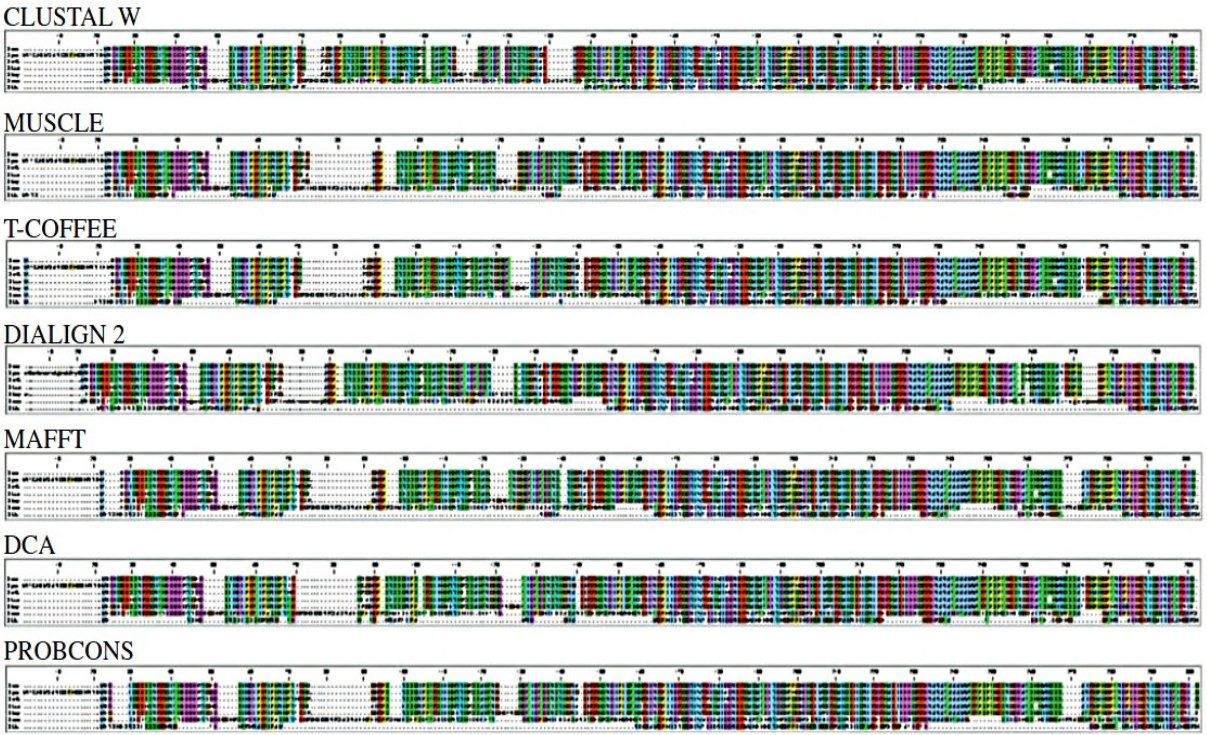
```
  5    42
Turkey    AAGCTNGGGC ATTTCAGGGT
GAGCCCGGGC AATACAGGGT AT
Salmo gairAAGCCTTGGC AGTGCAGGGT
GAGCCGTGGC CGGGCACGGT AT
H. SapiensACCGGTTGGC CGTTCAGGGT
ACAGGTTGGC CGTTCAGGGT AA
Chimp     AAACCCTTGC CGTTACGCTT
AAACCGAGGC CGGGACACTC AT
Gorilla   AAACCCTTGC CGGTACGCTT
AAACCATTGC CGGTACGCTT AA
```

Note, of course, that a portion of a sequence like this: **300 AAGCGTGAAC GTTGTACTAA TRCAG**

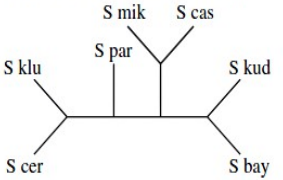
is perfectly legal, assuming that the species name has gone before, and is filled out to full length by blanks.

The above digits and blanks will be ignored,

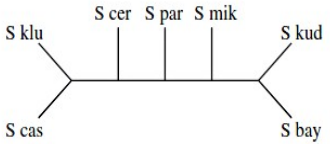
Caveat I: different programs may lead to distinct trees



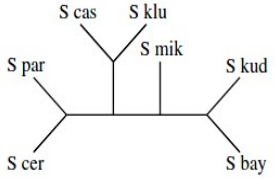
CLUSTAL/DIALIGN (0.24)



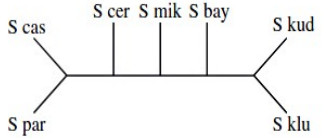
MUSCLE (0.25)



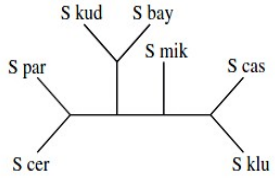
MAFFT (0.18)



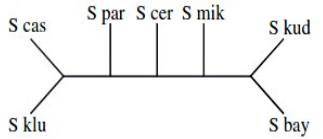
DCA (0.12)



T-COFFEE (0.30)



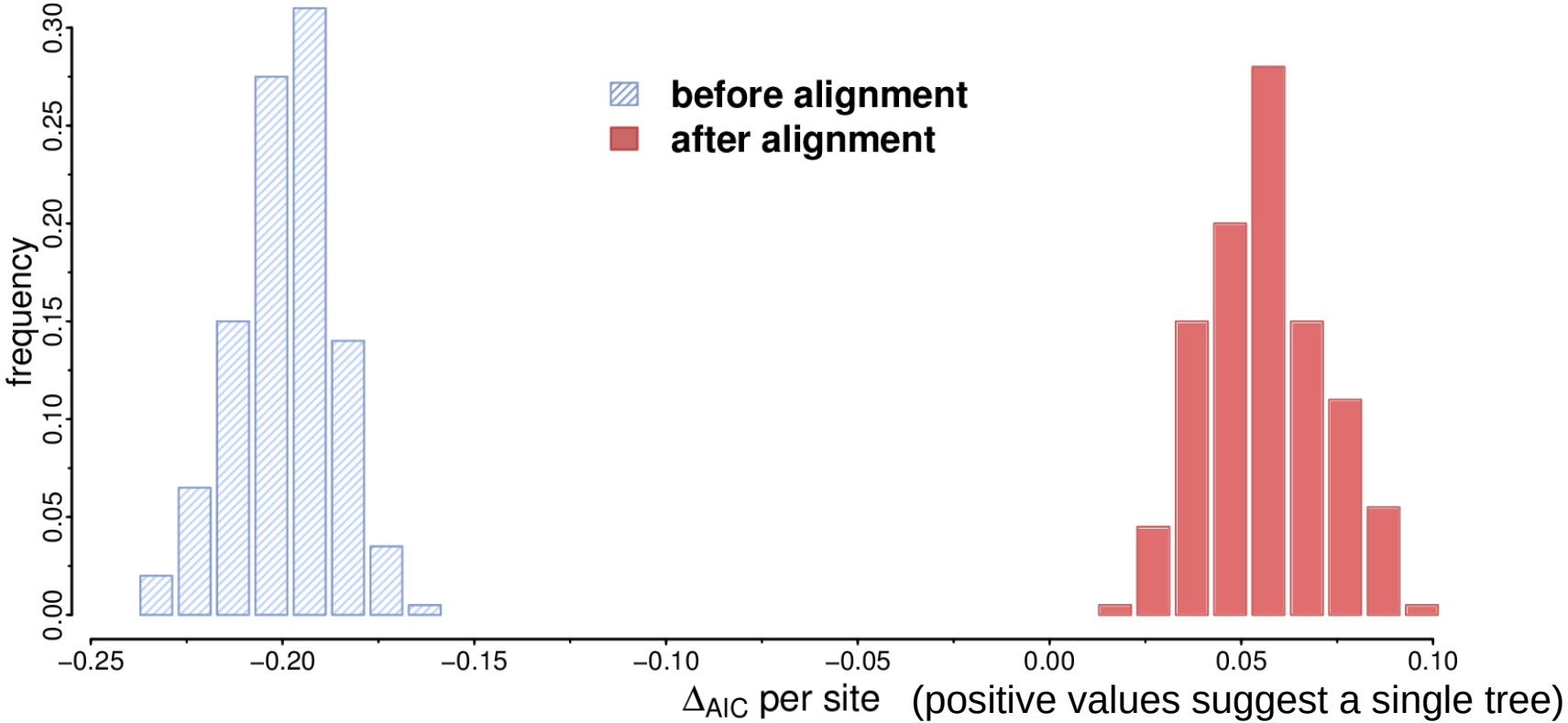
PROBCONS (0.05)



https://doi.org/10.1126/science.1151532

Quadram Institute
Science Health Food Innovation

Caveat II: programs can align anything, even random sequences



doi: 10.1093/sysbio/syu041

Quadram Institute
Science Health Food Innovation

- Make sure that genes / contigs you want to align are homologous
- Alignment algorithms or sequences/regions can affect the tree
- Should you add more sequences before or after the alignment?
- coding regions can be translated or codon-aligned
- software: `trimal`, `guidance`

SNP-sites

for example

- snippy's "core SNP genome"
- gubbins output alignment

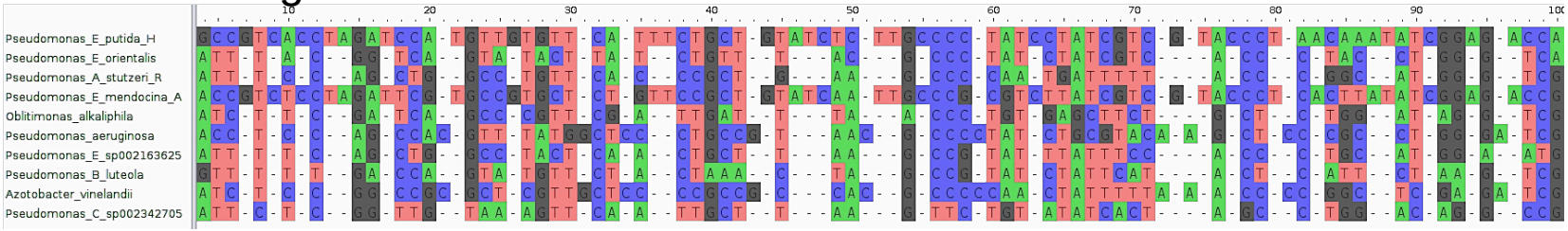
Quadram Institute
Science Health Food Innovation

Did you try the github data sets?

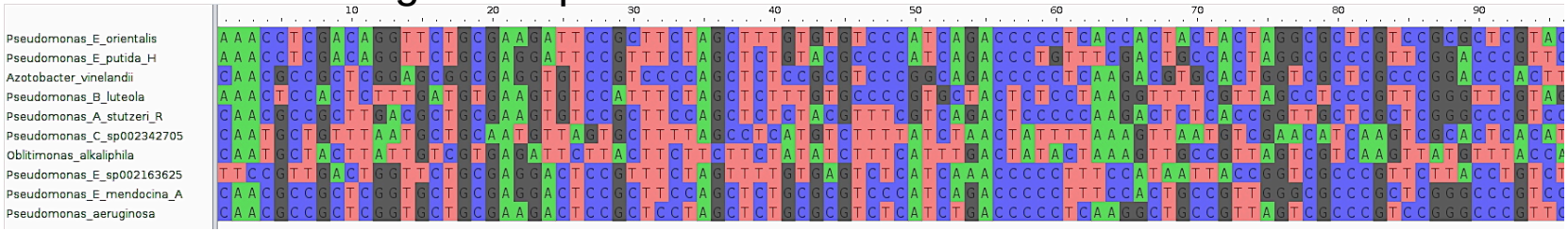
- Data set 1 (SNPs) only
- Data set 2 (concatenate alignments) only
- Both data sets
- I did not play with the data

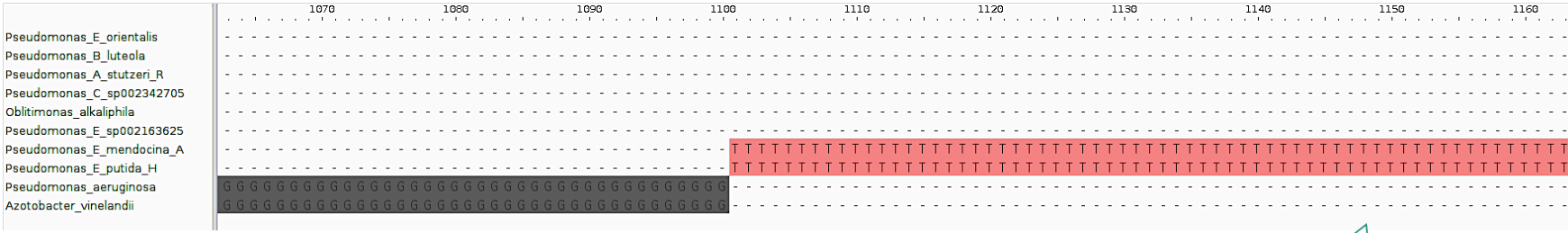
Data set 1 from github material

How it was shared on github: columns shuffled

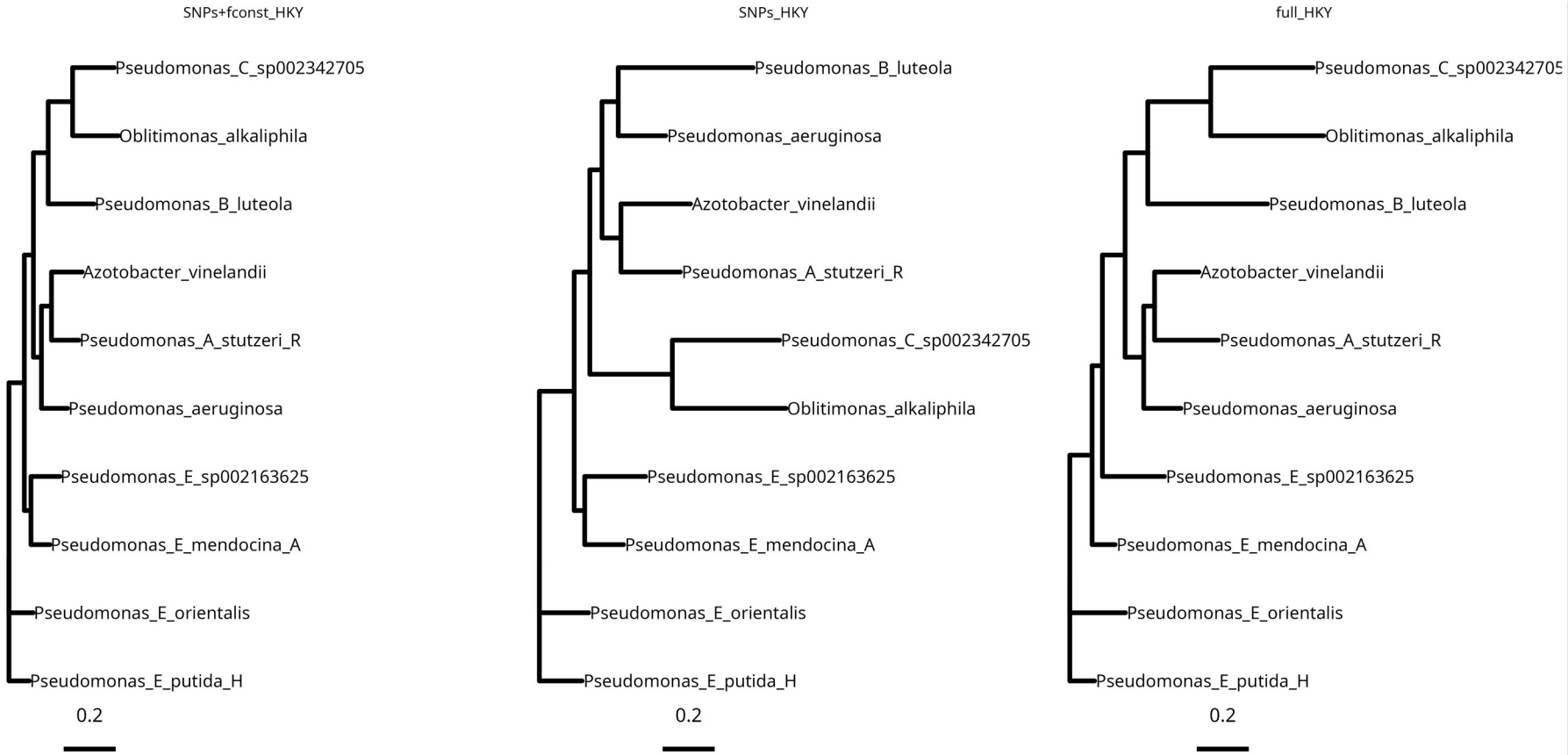


How it was created: real alignment plus “invariant” columns

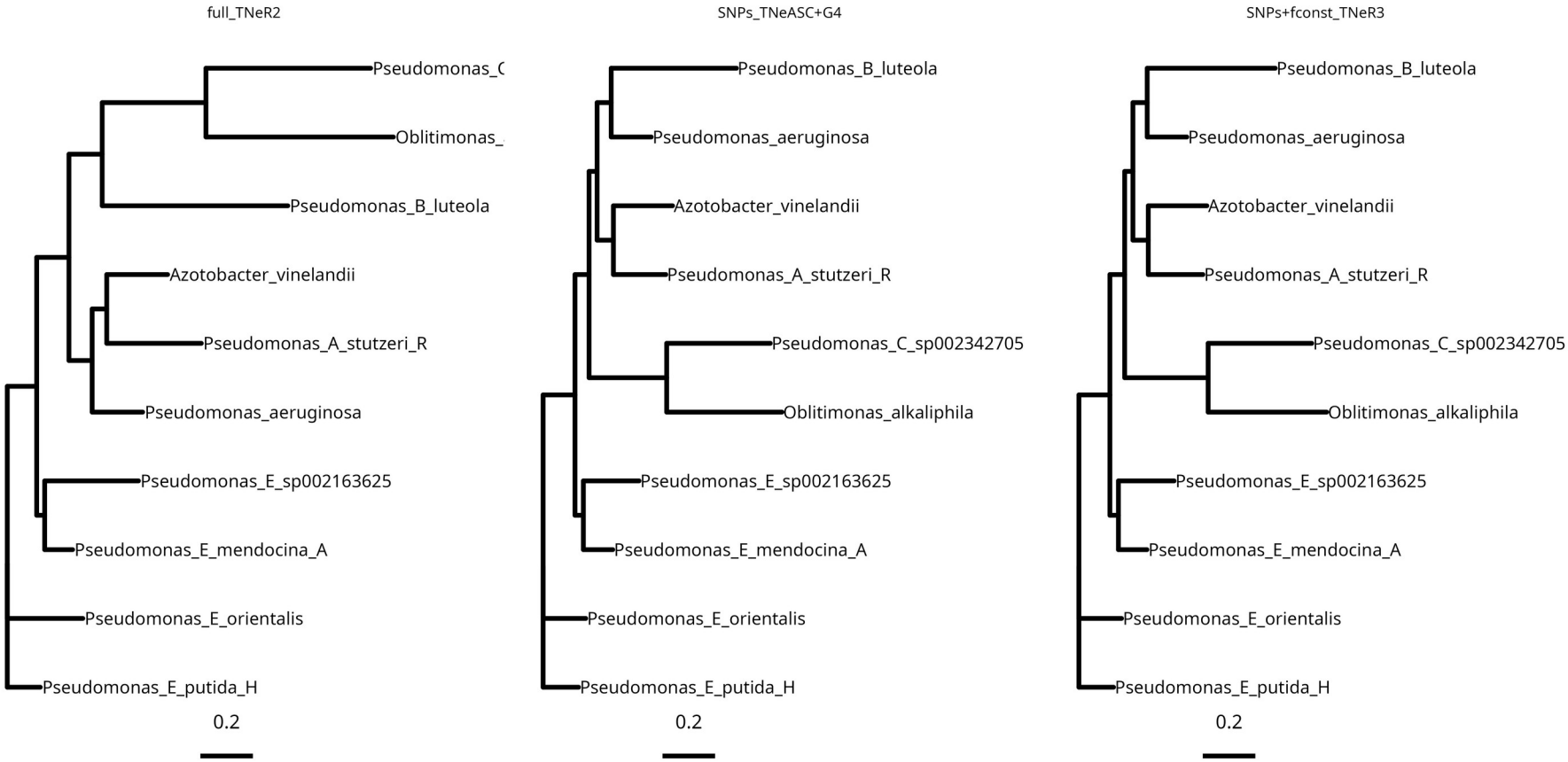




ML tree under fixed (HKY) model



ML tree under best model



Are the trees different?

	SNPs+fconst_HKY	SNPs_HKY	full_HKY	full_TNeR2	SNPs_TNeASC+G4
SNPs_HKY	1				
full_HKY	1	2			
full_TNeR2	0	1	1		
SNPs_TNeASC+G4	1	0	2	1	
SNPs+fconst_TNeR3	1	0	2	1	0

SPR

	SNPs+fconst_HKY	SNPs_HKY	full_HKY	full_TNeR2	SNPs_TNeASC+G4
SNPs_HKY	4				
full_HKY	2	6			
full_TNeR2	0	4	2		
SNPs_TNeASC+G4	4	0	6	4	
SNPs+fconst_TNeR3	4	0	6	4	0

RF

	SNPs+fconst_HKY	SNPs_HKY	full_HKY	full_TNeR2	SNPs_TNeASC+G4
SNPs_HKY	2.16060480				
full_HKY	1.65863479	0.92129548			
full_TNeR2	3.10375600	1.67082095	1.57537866		
SNPs_TNeASC+G4	1.79993846	0.36453405	0.72916240	1.71250523	
SNPs+fconst_TNeR3	1.84060299	0.34925395	0.74870724	1.64916454	0.07588229

wRF

file `08.snp_only_all_models.iqtree`

Quadram Institute
Science & Health & Food & Innovation

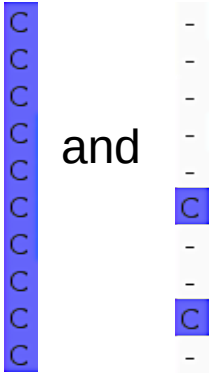
Are the trees significantly different? Based on data (=alignment)

```
See 08.snp_only_all_models.trees for trees with branch lengths.

Tree      logL    deltaL  bp-RELL    p-KH     p-SH    p-WKH    p-WSH       c-ELW       p-AU
--------------------------------------------------------------------------------------------
  1 -7566.973604  2.2582   0.162 +  0.377 +  0.589 +  0.377 +  0.804 +     0.173 +    0.492 +
  2  -7573.24419  8.5288  0.0564 +  0.169 +  0.195 +  0.128 +  0.201 +    0.0512 +    0.111 +
  3 -7566.973589  2.2582   0.172 +  0.377 +  0.589 +  0.377 +  0.716 +     0.175 +    0.491 +
  4 -7564.715398       0   0.304 +  0.501 +      1 +  0.501 +  0.756 +     0.301 +    0.564 +
  5 -7564.715444 4.6282e-05   0.306 +  0.499 +  0.854 +  0.499 +  0.852 +       0.3 +    0.578 +

deltaL  : logL difference from the maximal logl in the set.
bp-RELL : bootstrap proportion using RELL method (Kishino et al. 1990).
p-KH    : p-value of one sided Kishino-Hasegawa test (1989).
p-SH    : p-value of Shimodaira-Hasegawa test (2000).
p-WKH   : p-value of weighted KH test.
p-WSH   : p-value of weighted SH test.
c-ELW   : Expected Likelihood Weight (Strimmer & Rambaut 2002).
p-AU    : p-value of approximately unbiased (AU) test (Shimodaira, 2002).

Plus signs denote the 95% confidence sets.
Minus signs denote significant exclusion.
All tests performed 50000 resamplings using the RELL method.
```

RELL = resampling of estimated log-likelihoods

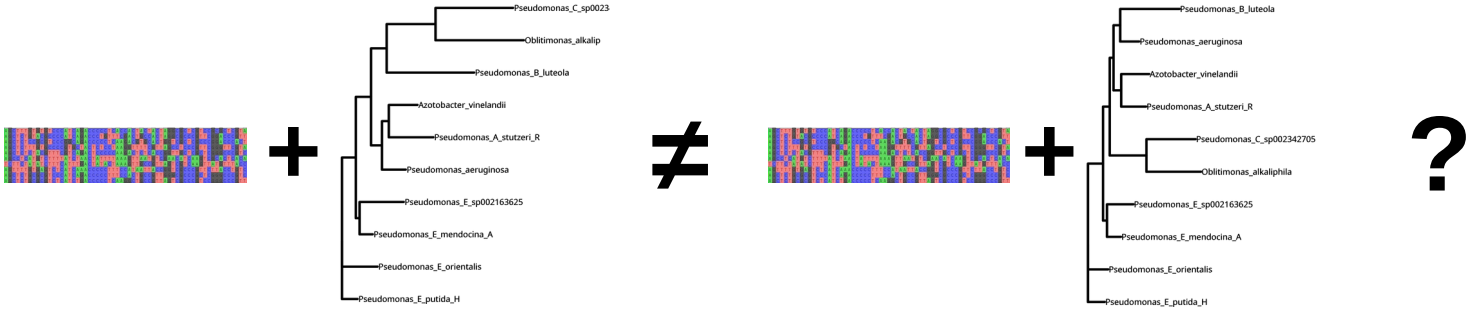
file `08.snp_only_all_models.iqtree`

- We hope that only SNPs have split (i.e. phylogenetic) information



- However likelihood models incorporate uncertainty (N or indel);i.e. and are different

- Tree likelihood tests: given the same alignment, are the trees significantly different?



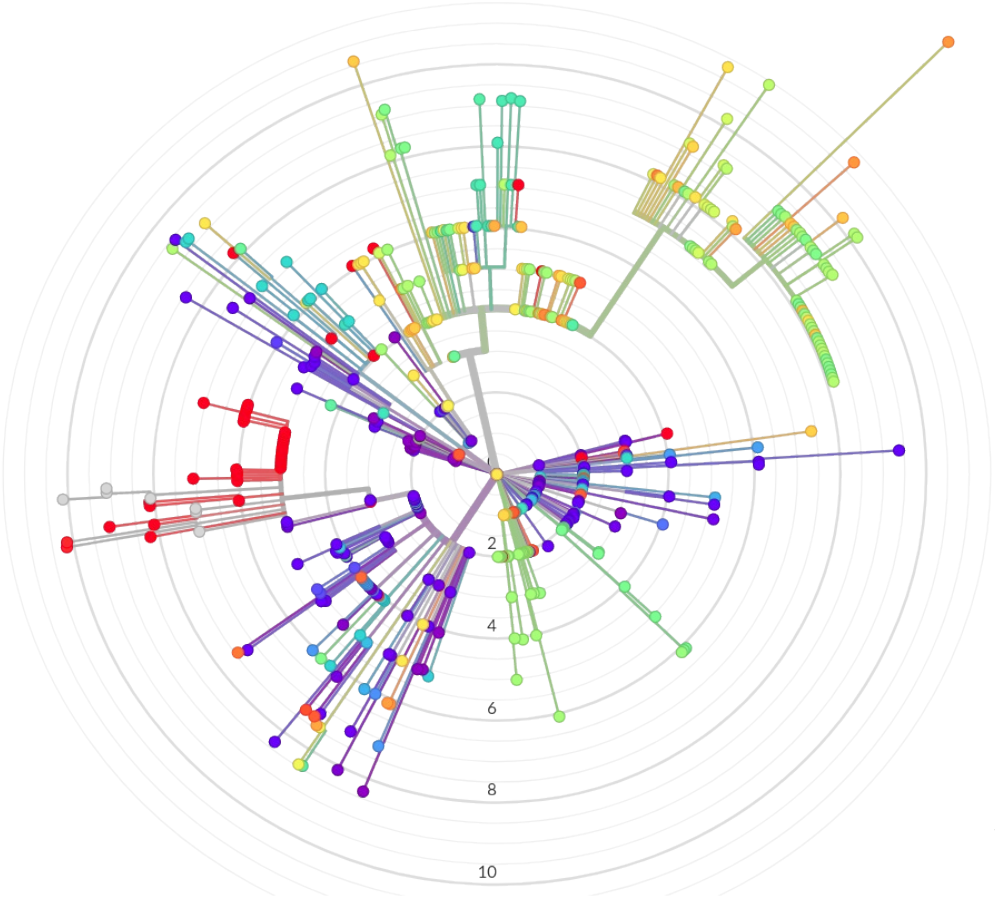
Quadram Institute
Science Health Food Innovation

The molecular clock

strict and relaxed

Quadram Institute
Science Health Food Innovation

If sampling times are known, we can try to fit the tree tips

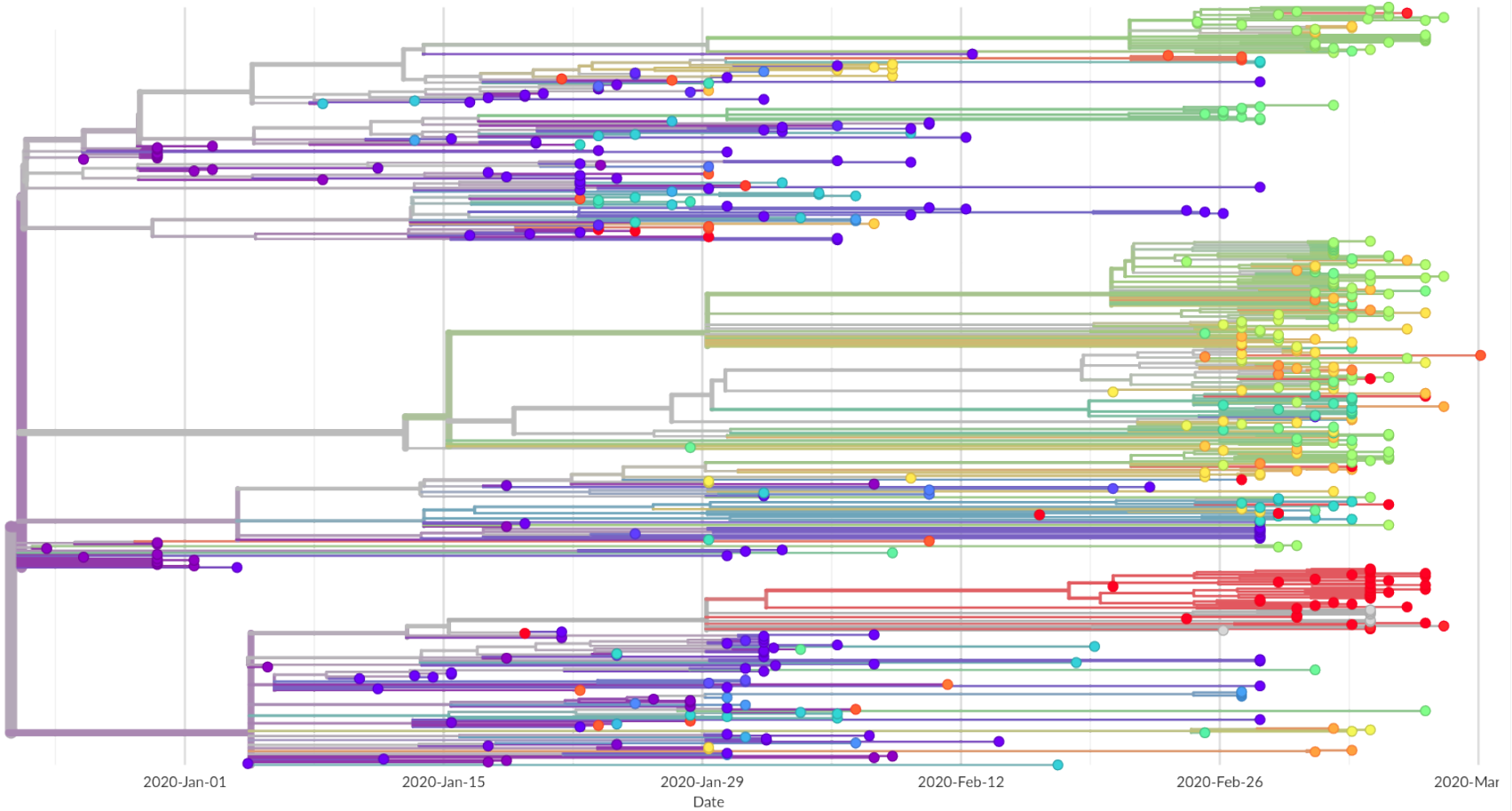


input data: ML (unrooted) tree

+

sampling times for each sequence

If sampling times are known, we can try to fit the tree tips



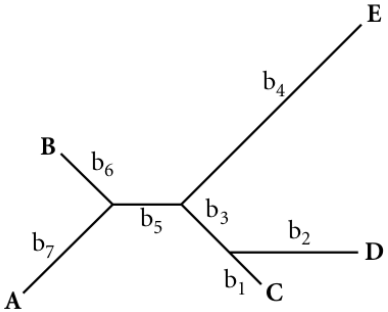
from https://nextstrain.org/

Quadram Institute
Science & Health & Food & Innovation

Curiosity: strict clock models used to assume ultrametricity

A

Nonclocklike phylogenetic tree
***n* taxa = 5**

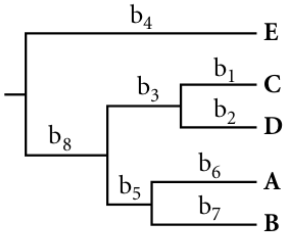


unrooted tree
$2n - 3$ independent branches

All b_1, b_2, b_3, b_4, b_5, b_6, and b_7 need to be estimated

B

Clocklike phylogenetic tree
***n* taxa = 5**



rooted tree
$n - 1$ independent branches

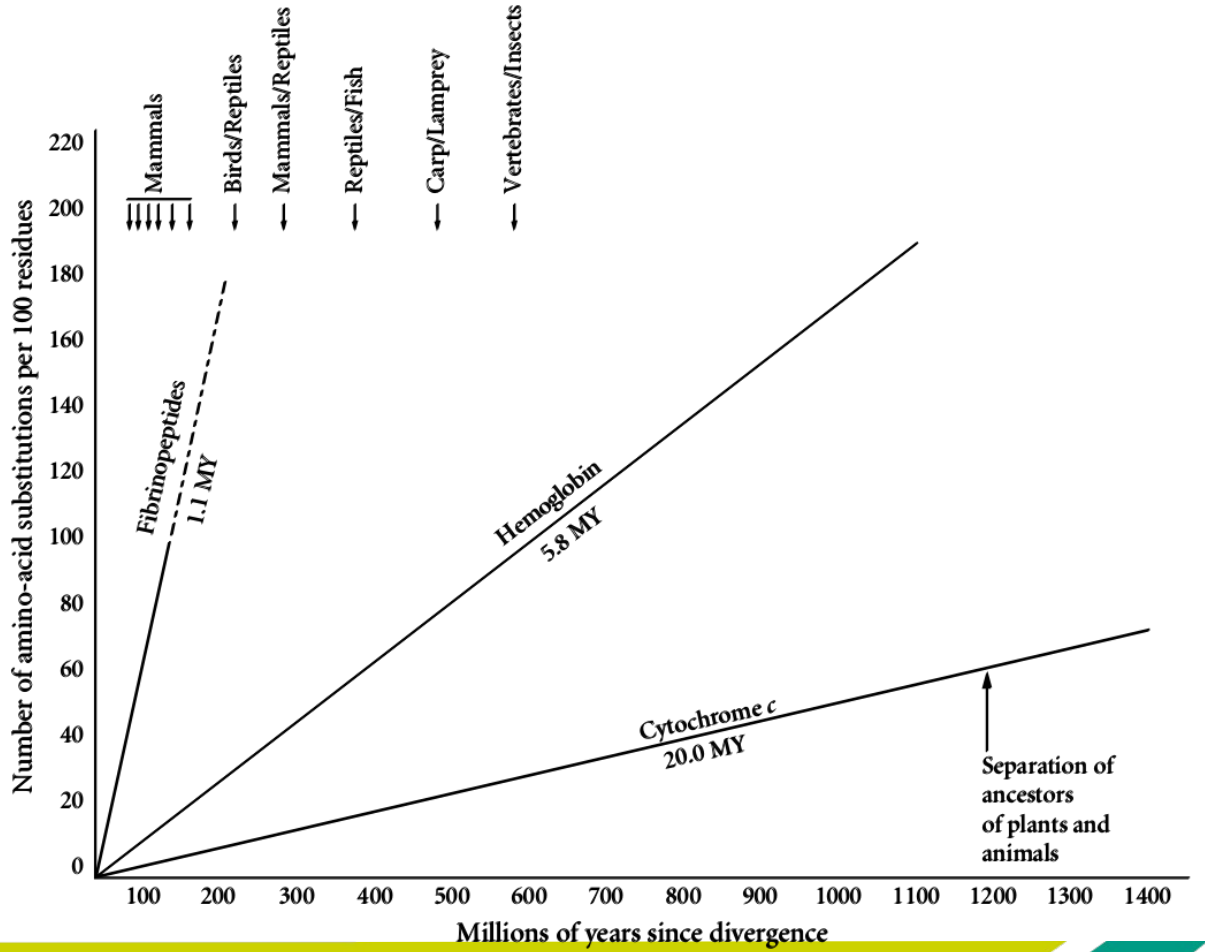
Only b_1, b_3, b_4, and b_6, for example, need to be estimated, because under the molecular clock:

$b_2 = b_1$
$b_5 = b_1 + b_3 - b_6$
$b_7 = b_6$
$b_8 = b_4 - b_5 - b_6$

Number of free parameters in clock and nonclock trees. Under the free rates model (= nonclock), all the branches need to be estimated ($2n - 3$). Under the molecular clock, only $n - 1$ branches have to be estimated. The difference in the number of parameters among a nonclock and a clock model is $n - 2$.

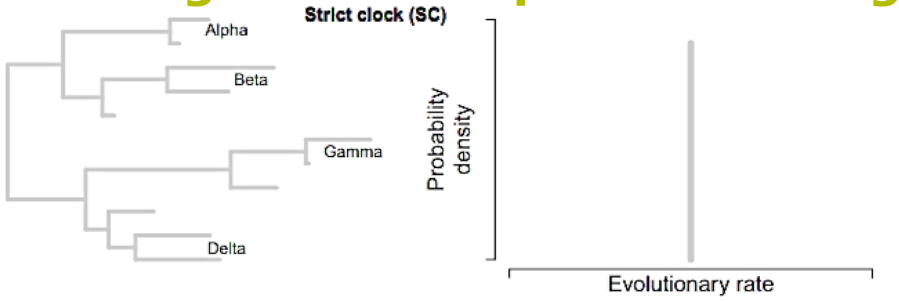
https://doi.org/10.1017/CBO9780511819049.012

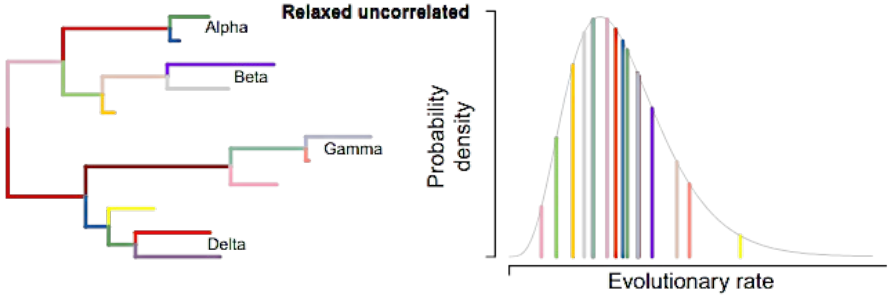
The (strict) molecular clock: each molecule appears to change at a constant rate

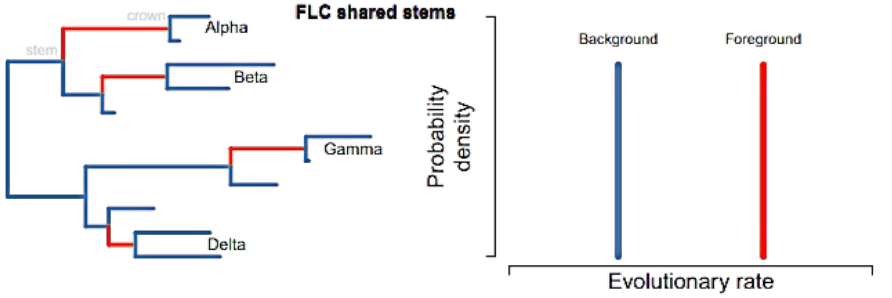


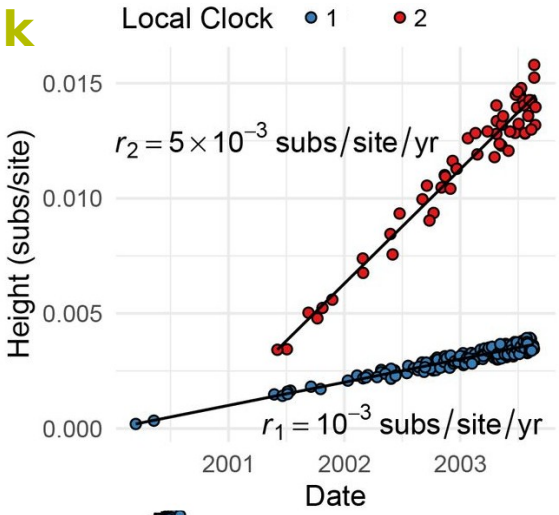
https://doi.org/10.1017/CBO9780511819049.012

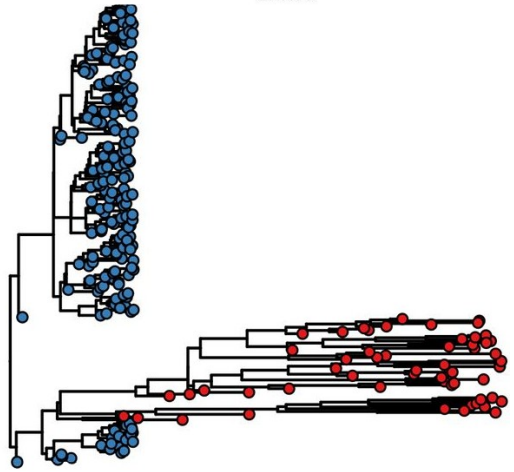
Relaxing the assumption of a single clock











doi:10.1093/molbev/msad212

doi:10.1101/2023.07.13.548947

Quadram Institute
Science & Health & Food & Innovation

Which of these phylogenetic methods have you heard of?

- Distance
- NJ or UPGMA
- Bayesian
- Parsimony
- bootstrap

For which of these phylogenetic methods do you know a software implementing them?

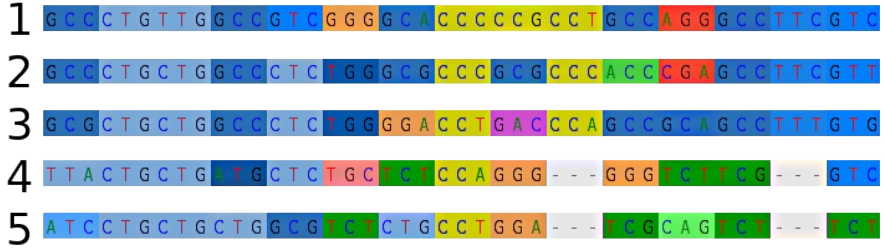
- Distance
- NJ or UPGMA
- Bayesian
- Parsimony
- bootstrap

Tree inference — Distance

- Clustering and minimum evolution

Quadram Institute
Science Health Food Innovation

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & & \\ 2 & 0 & & & \\ 3 & 5 & 0 & & \\ 4 & 7 & 8 & 0 & \\ 7 & 8 & 7 & 6 & 0 \end{pmatrix} \end{array}$$



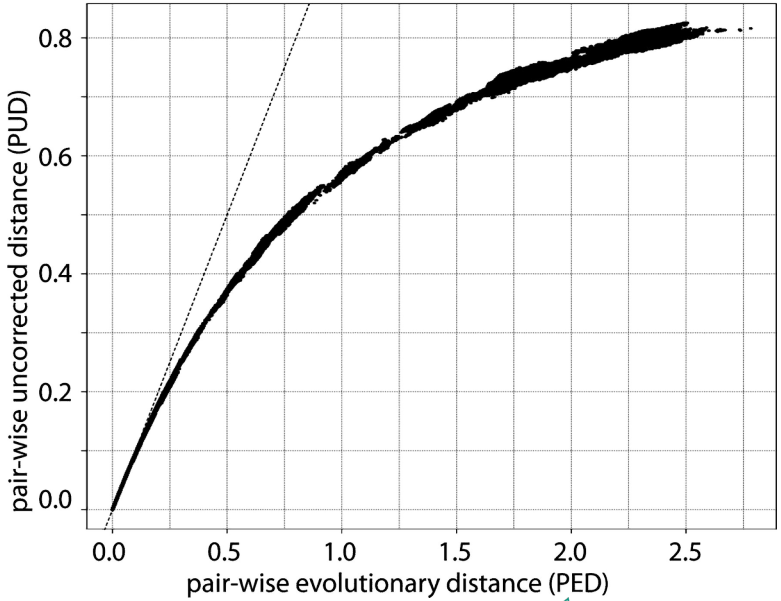
- based on pairwises similarity between sequences (distance matrix)

- clustering algorithms (*Neighbour Joining*, *UPGMA*) or "minimum evolution"

Caveats:

- assumes a homogeneous infinite sites model →

- single number representing differences



https://doi.org/10.1128/jvi.07173-11

Quadram Institute
Science & Health & Food & Innovation

Hierarchical clustering:

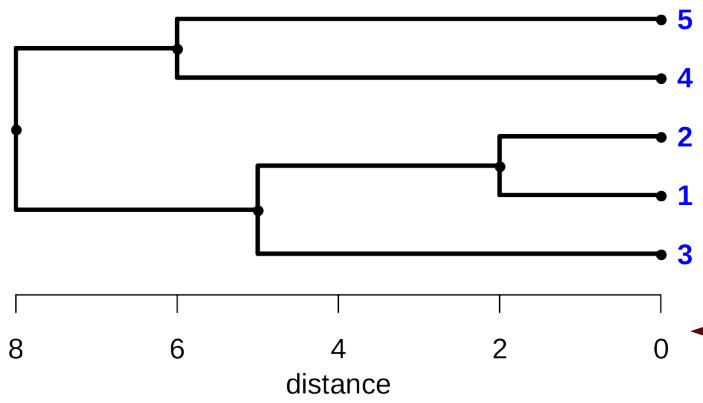
1. find smallest distance
2. merge groups
3. update distances

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ 1 & 0 & & & & \\ 2 & \mathbf{2} & 0 & & & \\ 3 & 3 & 5 & 0 & & \\ 4 & 4 & 7 & 8 & 0 & \\ 5 & 7 & 8 & 7 & 6 & 0 \end{array} \Rightarrow (1,2)$$

$$\begin{array}{c|cccc} & (1,2) & 3 & 4 & 5 \\ (1,2) & 0 & & & \\ 3 & \mathbf{5} & 0 & & \\ 4 & 7 & 8 & 0 & \\ 5 & 8 & 7 & 6 & 0 \end{array} \Rightarrow ((1,2),3)$$

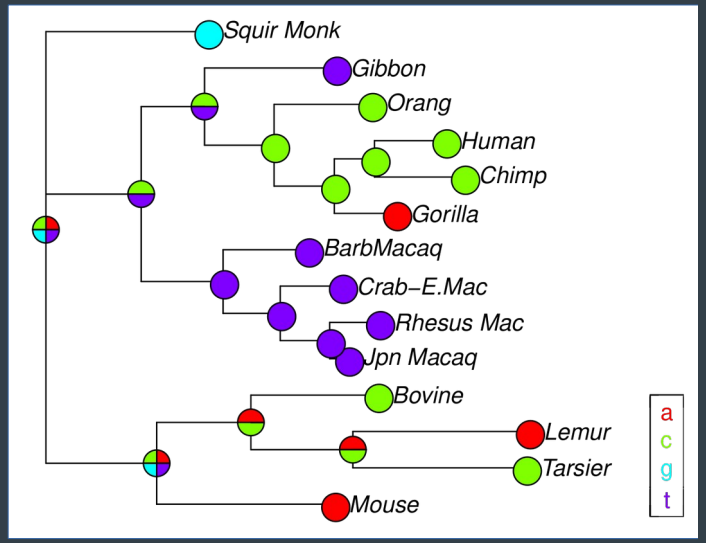
$$\begin{array}{c|ccc} & ((1,2),3) & 4 & 5 \\ ((1,2),3) & 0 & & \\ 4 & 8 & 0 & \\ 5 & 8 & \mathbf{6} & 0 \end{array} \Rightarrow (4,5)$$

$$\begin{array}{c|cc} & ((1,2),3) & (4,5) \\ ((1,2),3) & 0 & \\ (4,5) & 8 & 0 \end{array} \Rightarrow (((1,2),3),(4,5))$$



Tree inference — Parsimony

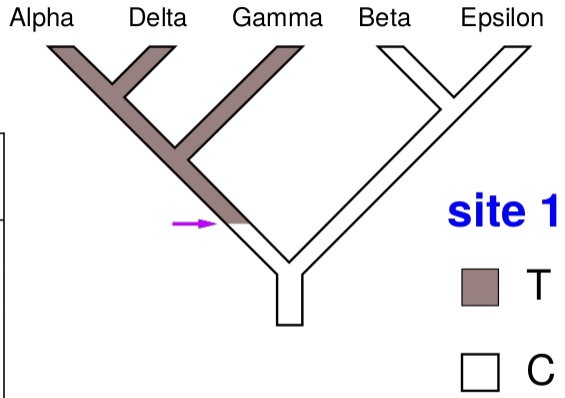
- Least amount of changes

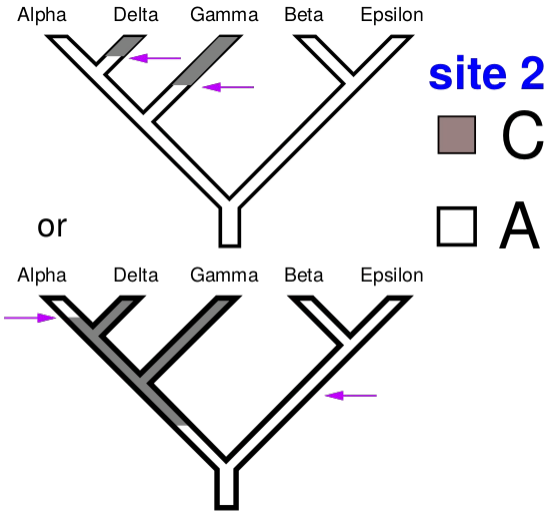


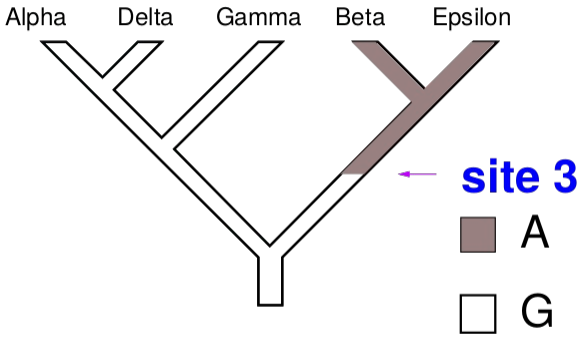
Quadram Institute
Science Health Food Innovation

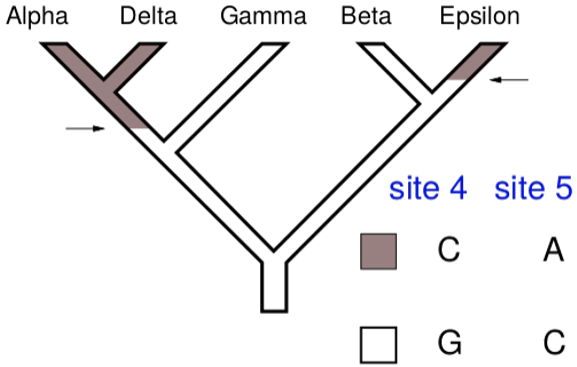
Parsimony score of each site

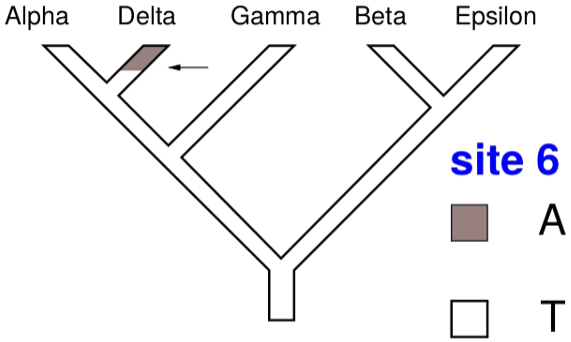
Species	Characters 1	2	3	4	5	6
Alpha	T	A	G	C	A	T
Beta	C	A	A	G	C	T
Gamma	T	C	G	G	C	T
Delta	T	C	G	C	A	A
Epsilon	C	A	A	C	A	T







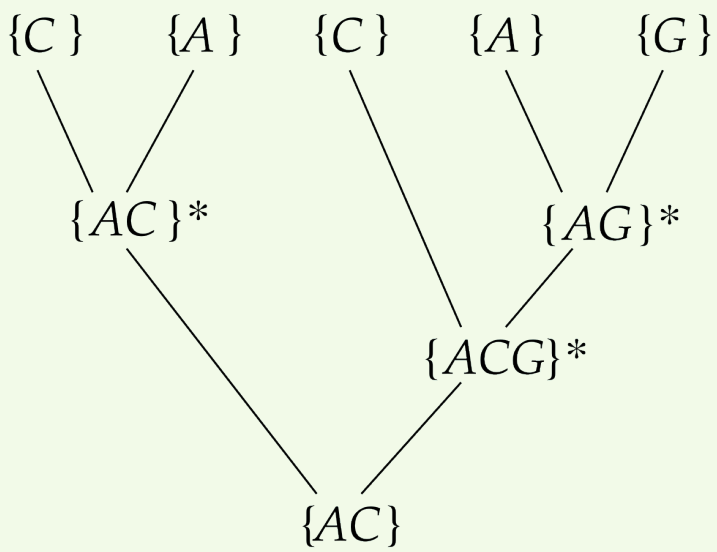




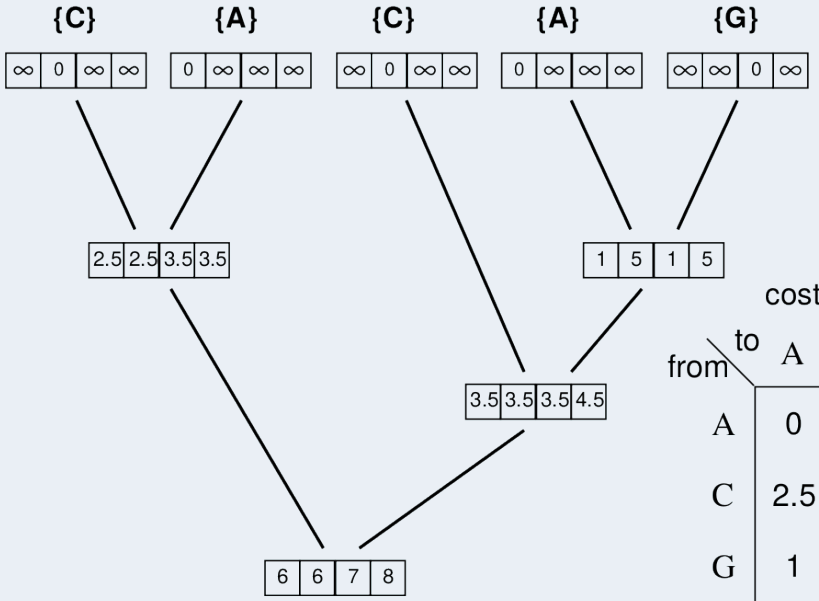
figures from https://evolution.gs.washington.edu/

Algorithms to calculate the parsimony score of a site

Fitch algorithm



Sankoff algorithm



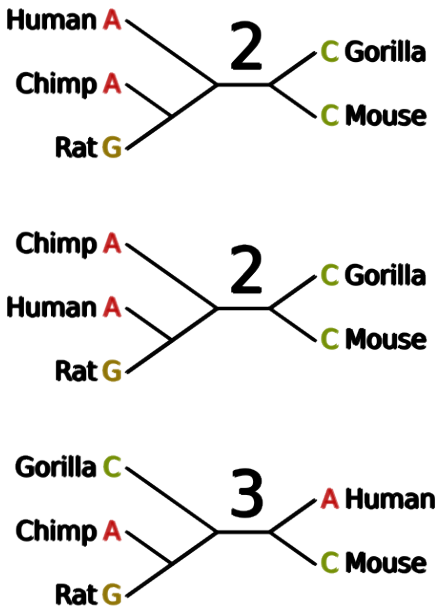
cost matrix:

from \ to	A	C	G	T
A	0	2.5	1	2.5
C	2.5	0	2.5	1
G	1	2.5	0	2.5
T	2.5	1	2.5	0

figures from https://evolution.gs.washington.edu/

Quadram Institute
Science Health Food Innovation

Finding tree with lowest parsimony score



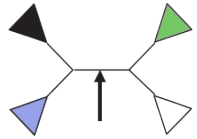
1. calculate parsimony score of tree

2. change tree

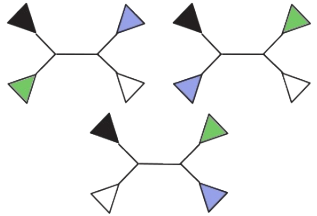
3. repeat

Nearest Neighbour Interchange

Step 1: Take original topology

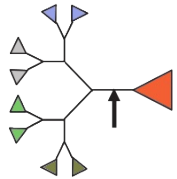


Step 2: choose one of three possible rearrangements

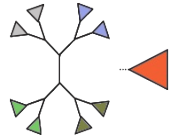


Subtree Pruning and Regrafting

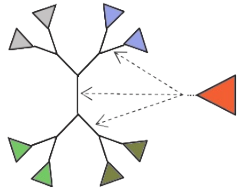
Step 1: Take original topology



Step 2: Break branch under consideration



Step 3: Attach subtree to any branch in other subtree



doi:10.1007/978-1-60327-159-2

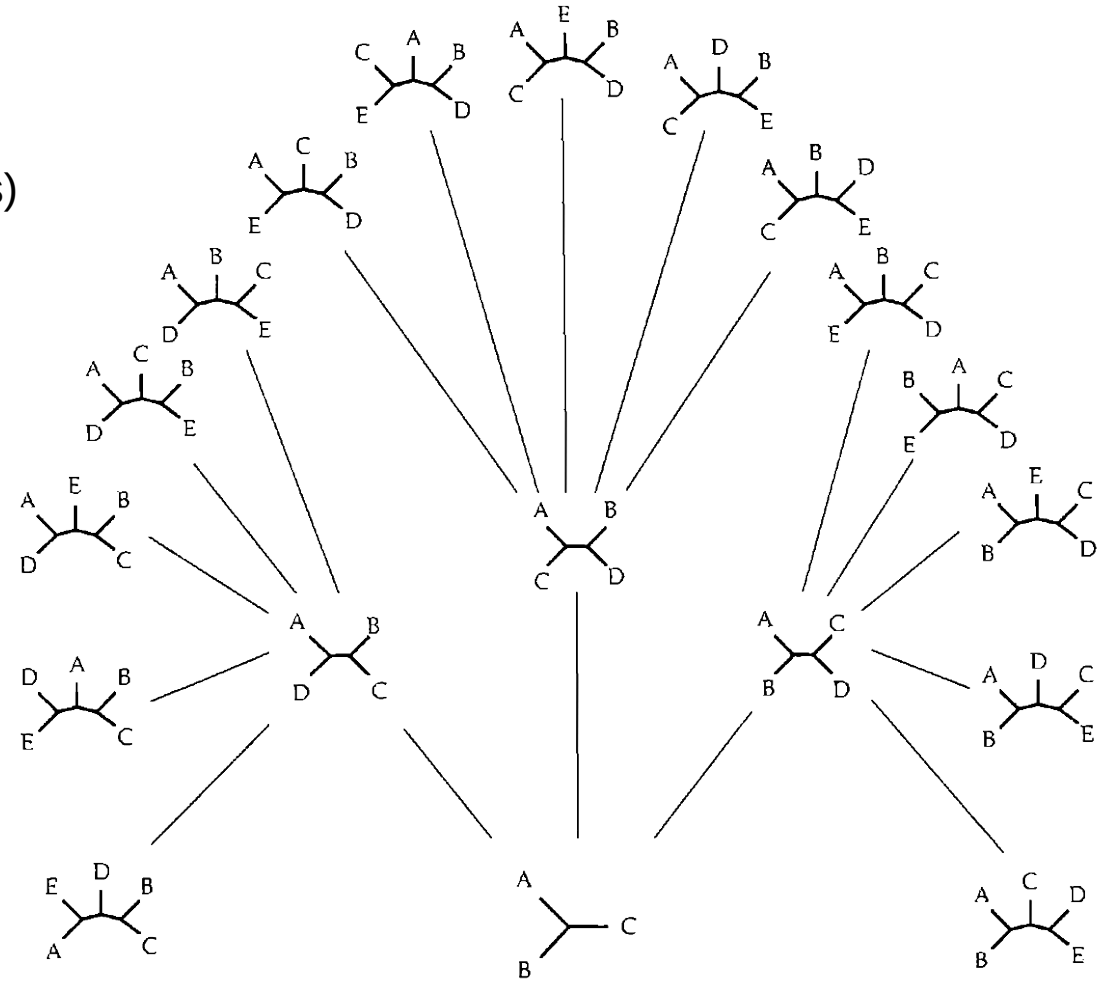
Caveats:

- heuristic (impossible to look at all trees)
- no branch lengths
- many equally good trees
- evolution is not parsimonious

Parsimony renaissance:

10.1186/s12862-018-1131-3 (MPboost)

10.1038/s41588-021-00862-7 (UShER)



from Inferring Phylogenies; Joseph Felsenstein 2004 ISBN: 9780878931774

Quadram Institute
Science Health Food Innovation

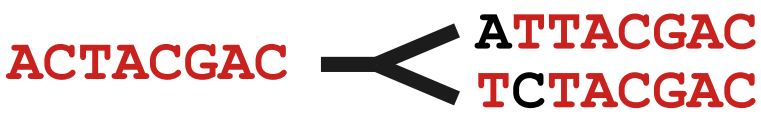
Tree inference — Likelihood

- statistical models of evolution
- main component of Bayesian phylogenetics

Quadram Institute
Science Health Food Innovation

Why likelihood?

ancestral extant



https://montypython.fandom.com/

Only two differences are observed between the two present-day sequences, so that the proportion of different sites is $\hat{p} = 2/8 = 0.25$

Why likelihood?

ancestral

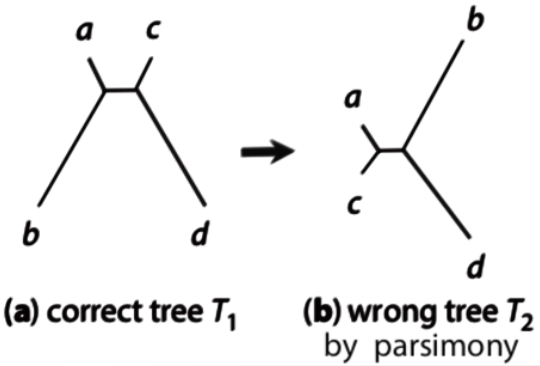
TTCAAGAC

extant

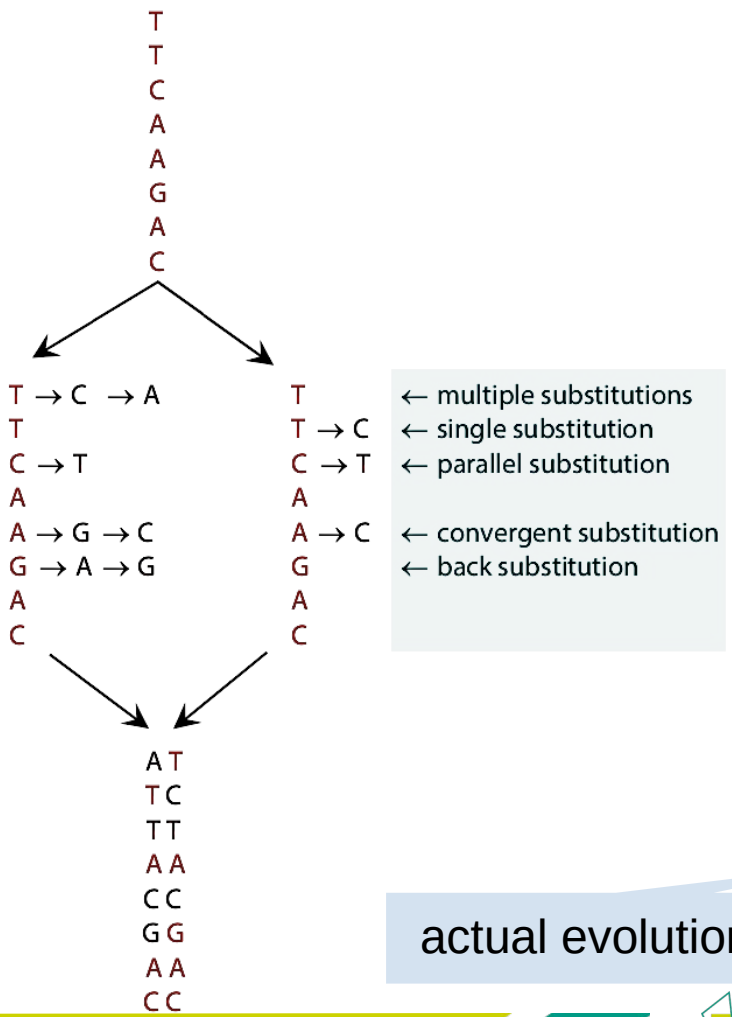
ATTACGAC

TCTACGAC

Only two differences are observed between the two present-day sequences, so that the proportion of different sites is $\hat{p} = 2/8 = 0.25$, while in fact as many as 10 substitutions (seven on the left lineage and three on the right lineage) occurred so that the true distance is 10/8 = 1.25 substitutions per site.



(a) correct tree T_1 **(b) wrong tree T_2** by parsimony



actual evolution

from http://abacus.gene.ucl.ac.uk/MESA/

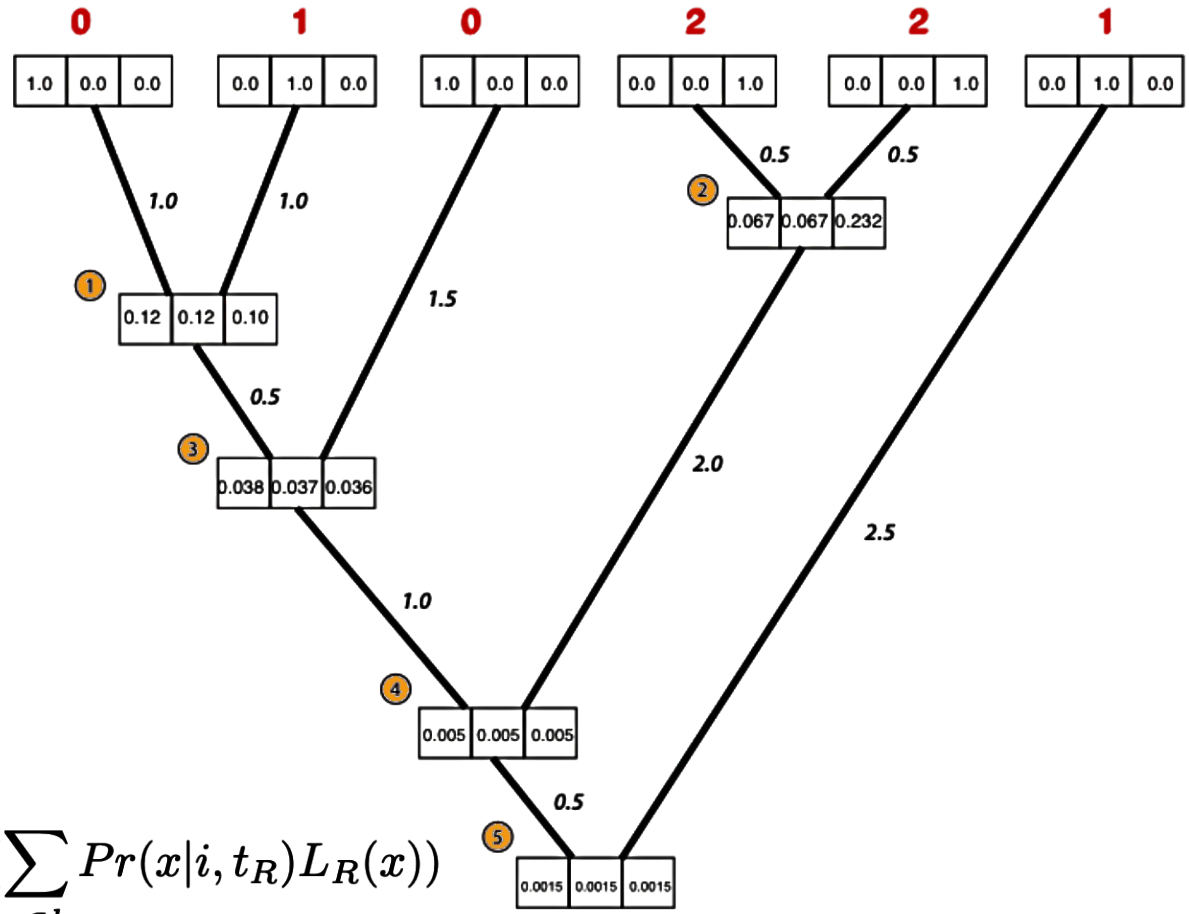
Quadram Institute
Science Health Food Innovation

1. “probabilities” of all states (0,1,2) in the past

2. calculate these conditionals in specific order {1,2,3,4,5}

3. change tree and branch lengths

4. repeat

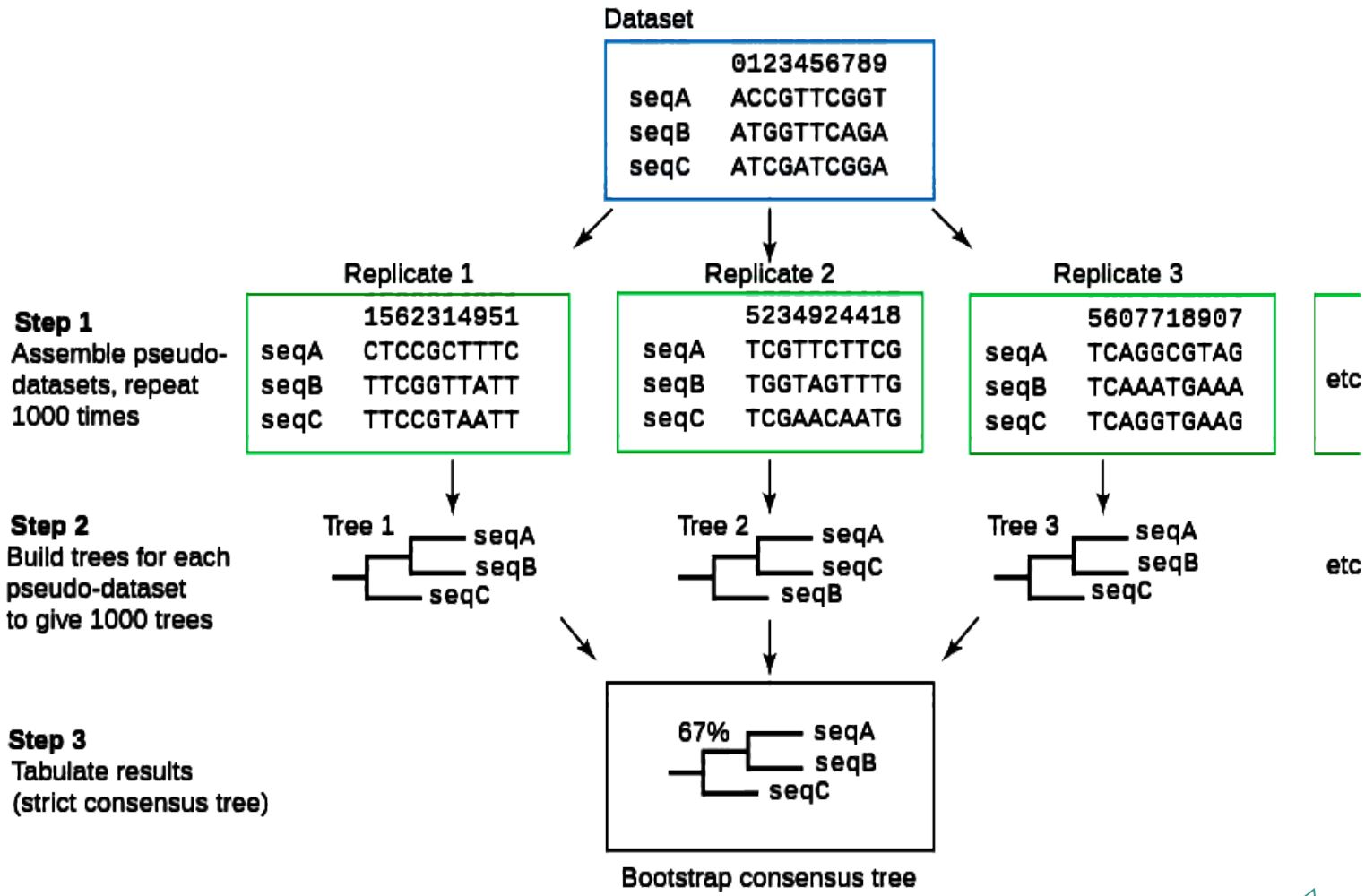


$$L_P(i) = (\sum_{x \in k} Pr(x|i, t_L) L_L(x)) \cdot (\sum_{x \in k} Pr(x|i, t_R) L_R(x))$$

Bootstrap and posterior

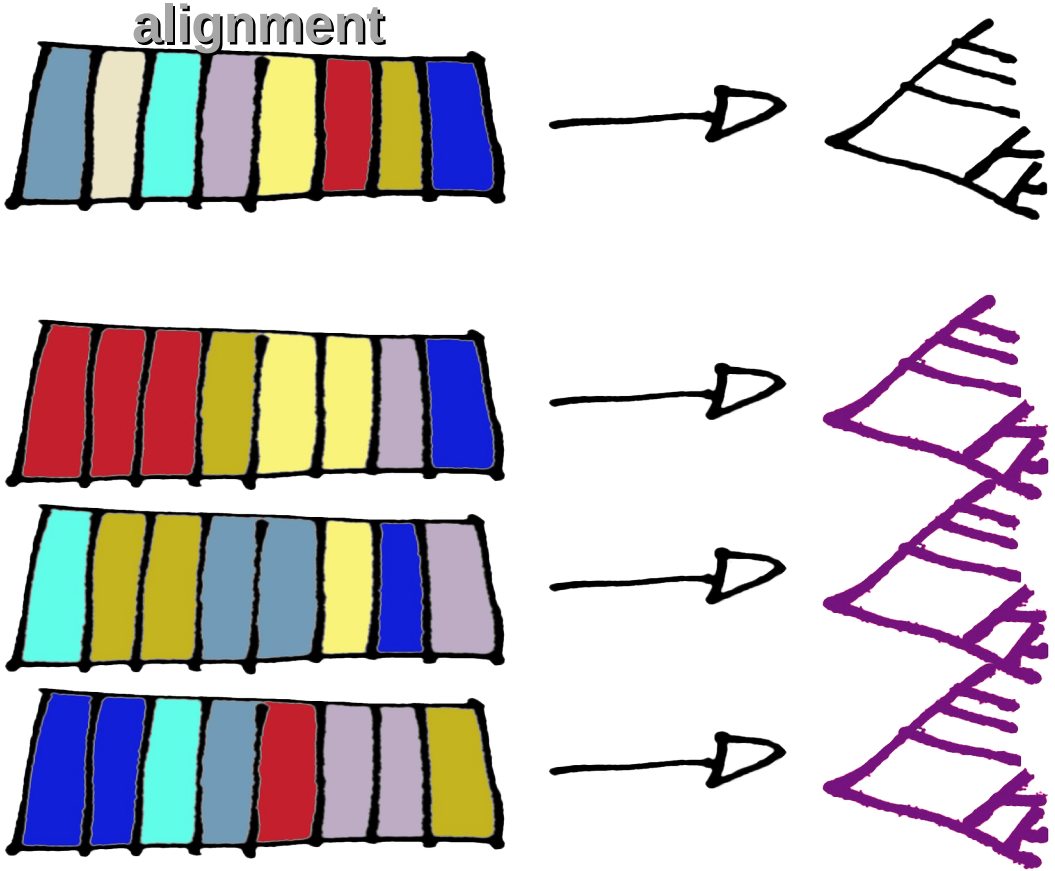
- tree uncertainty

Quadram Institute
Science Health Food Innovation



doi:10.1016/S0168-9525(03)00112-4

Phylogenetic Bootstraps (“Transfer” and “Felsenstein”)

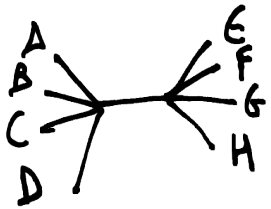
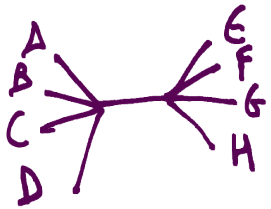
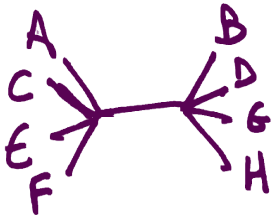
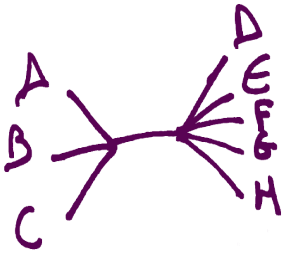
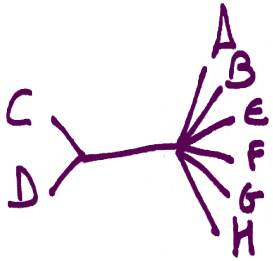


“reference tree” (estimated from original data)

bootstrap replicates (trees estimated from bootstrapped data)

with replacement → replicates have same size as original data

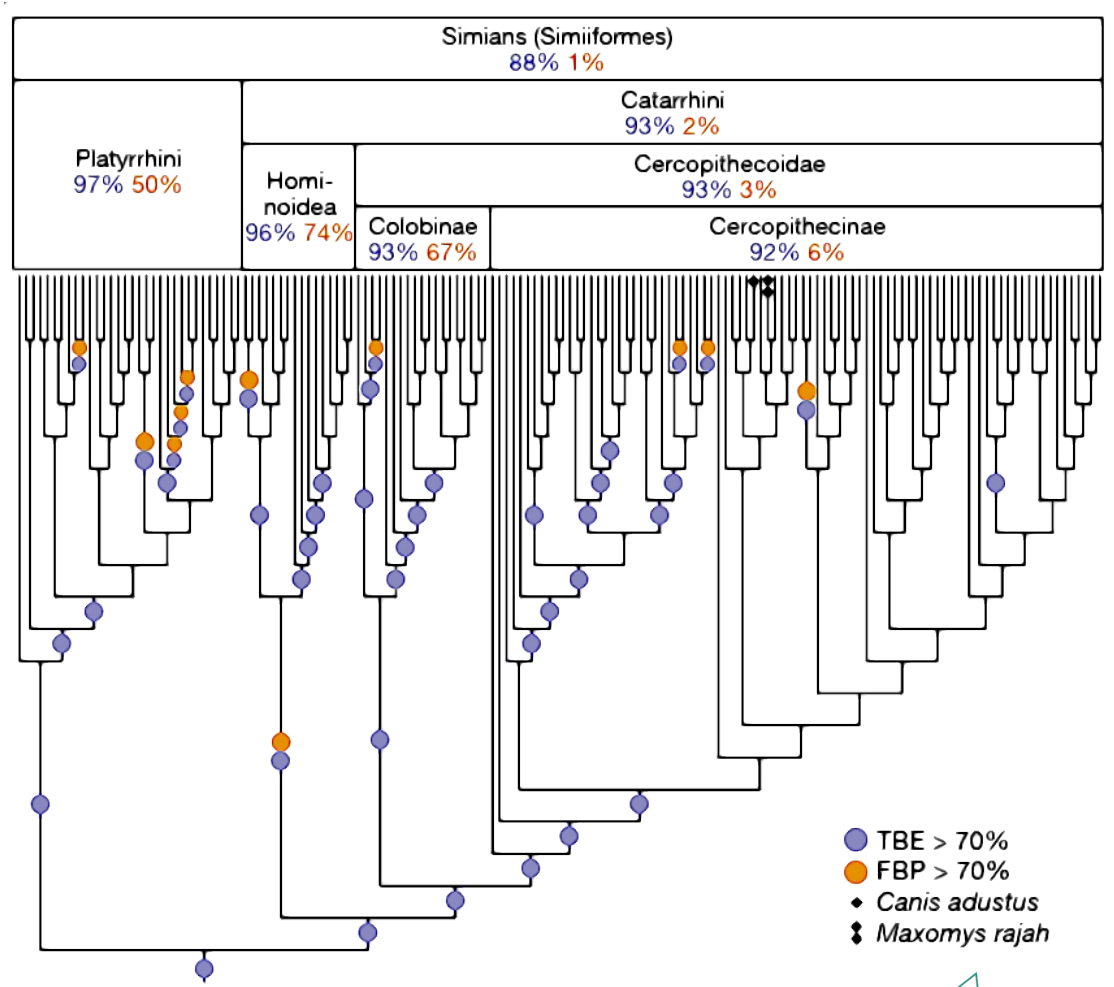
Transfer Bootstrap Expectation (TBE)

ref tree	replicate1	replicate2	replicate3	replicate4
				
FBP	1	0	0	0
transfer index	**0**	**3**	**1**	**2**

Interpretation

If TBE of branch is 95%, it means that on average 5% of leaves have to be removed from bootstrap samples.

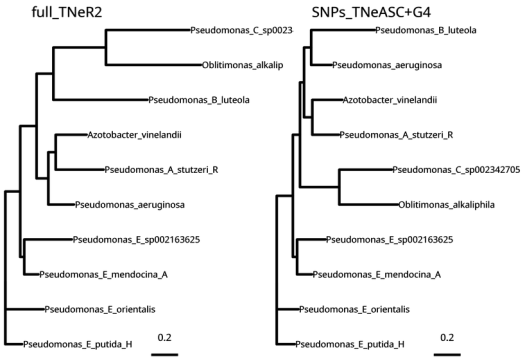
If FBP of branch is 95%, it means that 5% of the bootstrap replicate trees didn't have this split.



doi:10.1038/s41586-018-0043-0

Fig. 3 | FBP and TBE bootstrap supports on the simian clade.

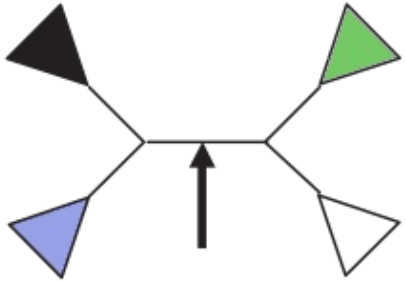
Revisiting data set 1 (full alignment) with RAxML

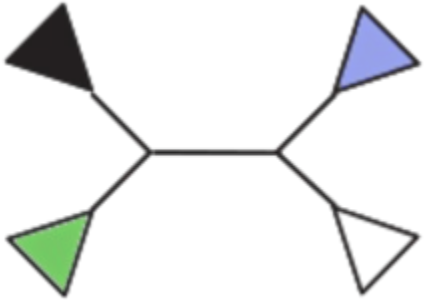


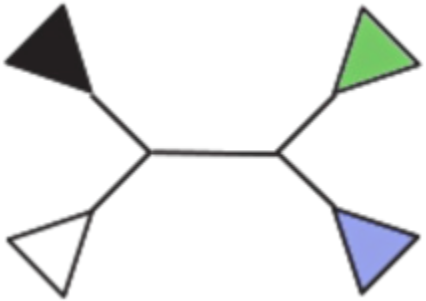
Quadram Institute
Science Health Food Innovation

Alternative

UFBoot (ultra fast bootstrap): at every internal branch, bootstraps around the two other competing NNI trees.







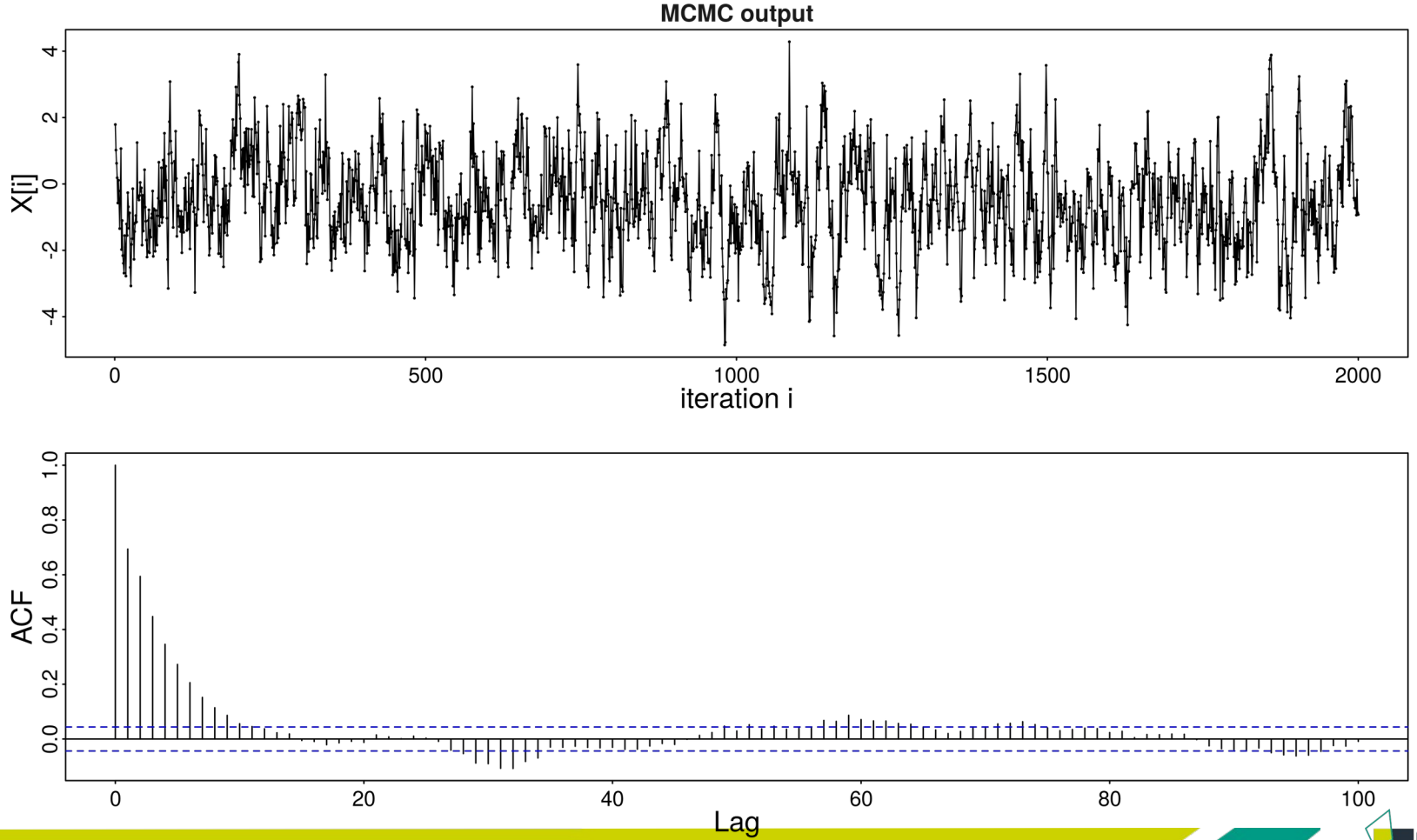
see also “concordance factors”:
doi:10.1093/molbev/msaa106 (IQTREE)
doi:10.1093/molbev/msw040 (RAxML)

curiosity: MCMC convergence

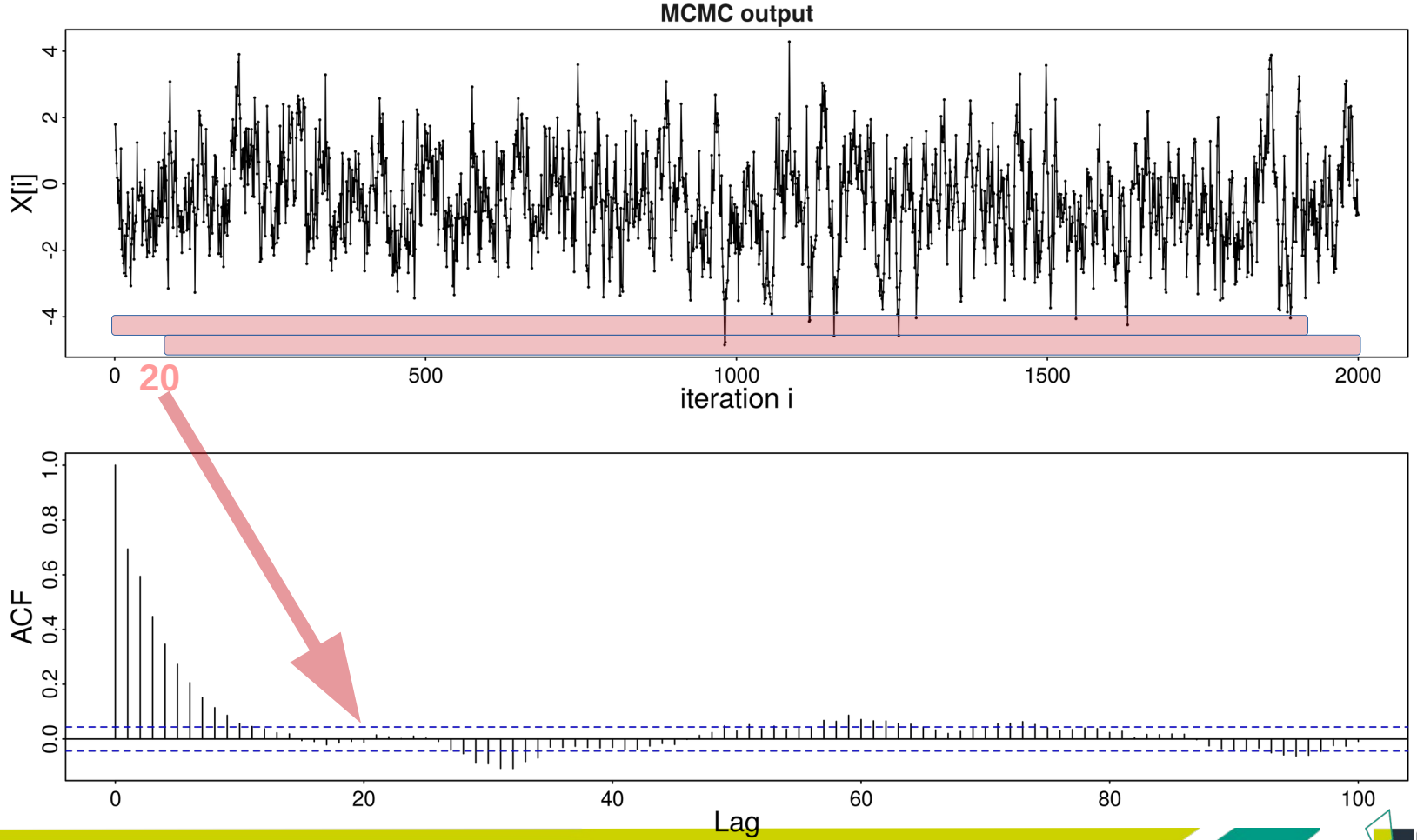
Posterior samples of trees or parameters in Bayesian Phylogenetics

Quadram Institute
Science Health Food Innovation

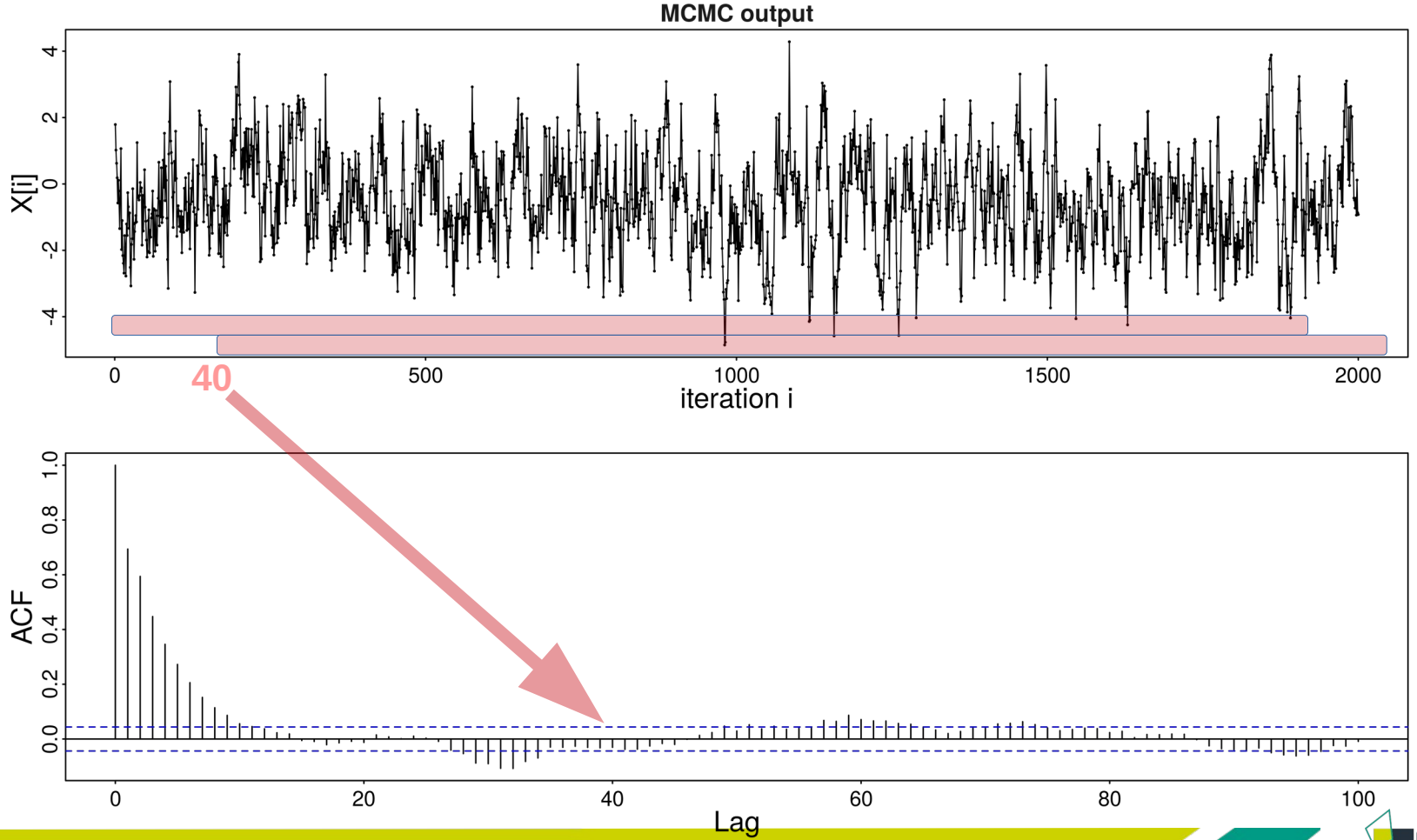
Markov chain Monte Carlo samples are correlated



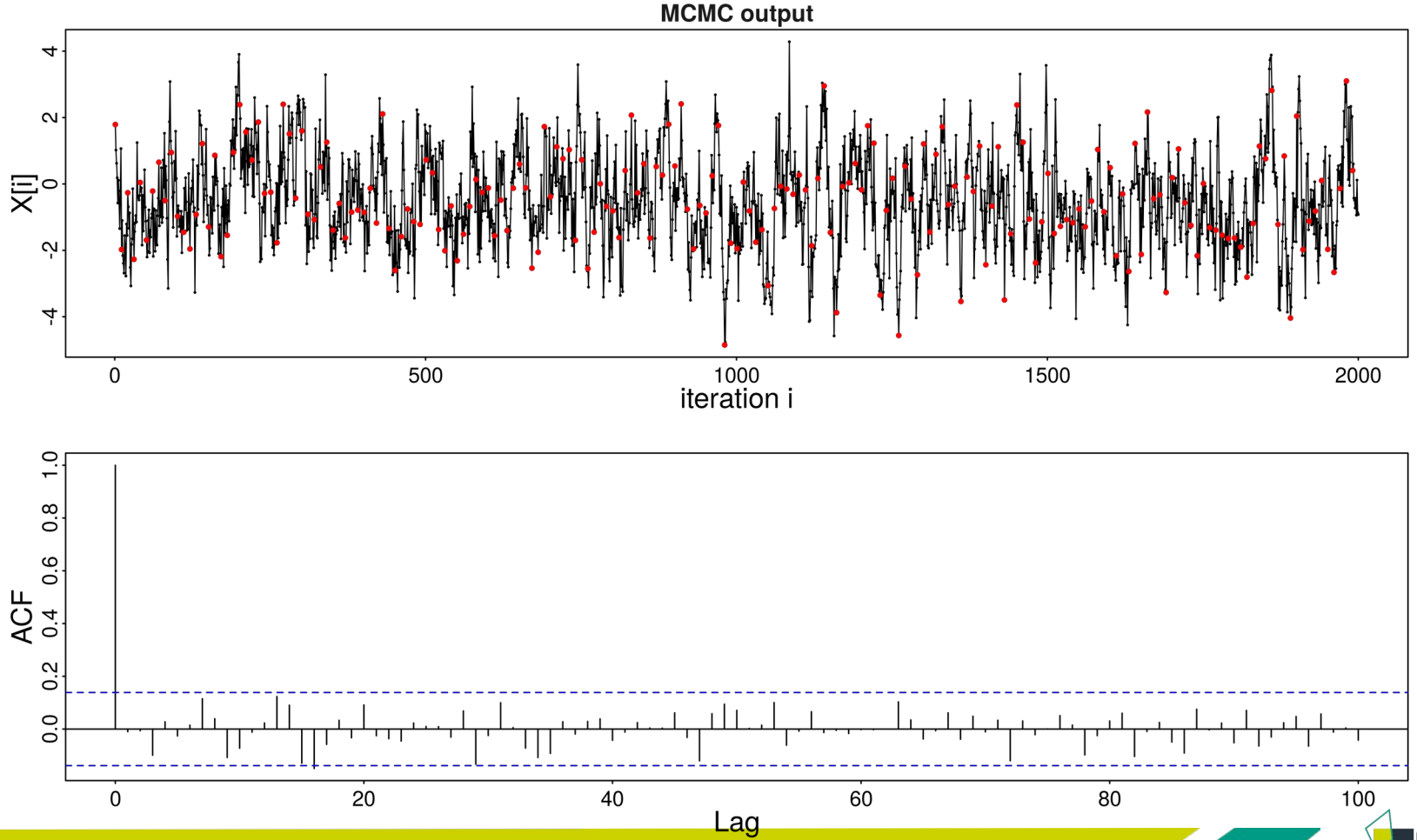
What is the autocorrelation function (ACF)?



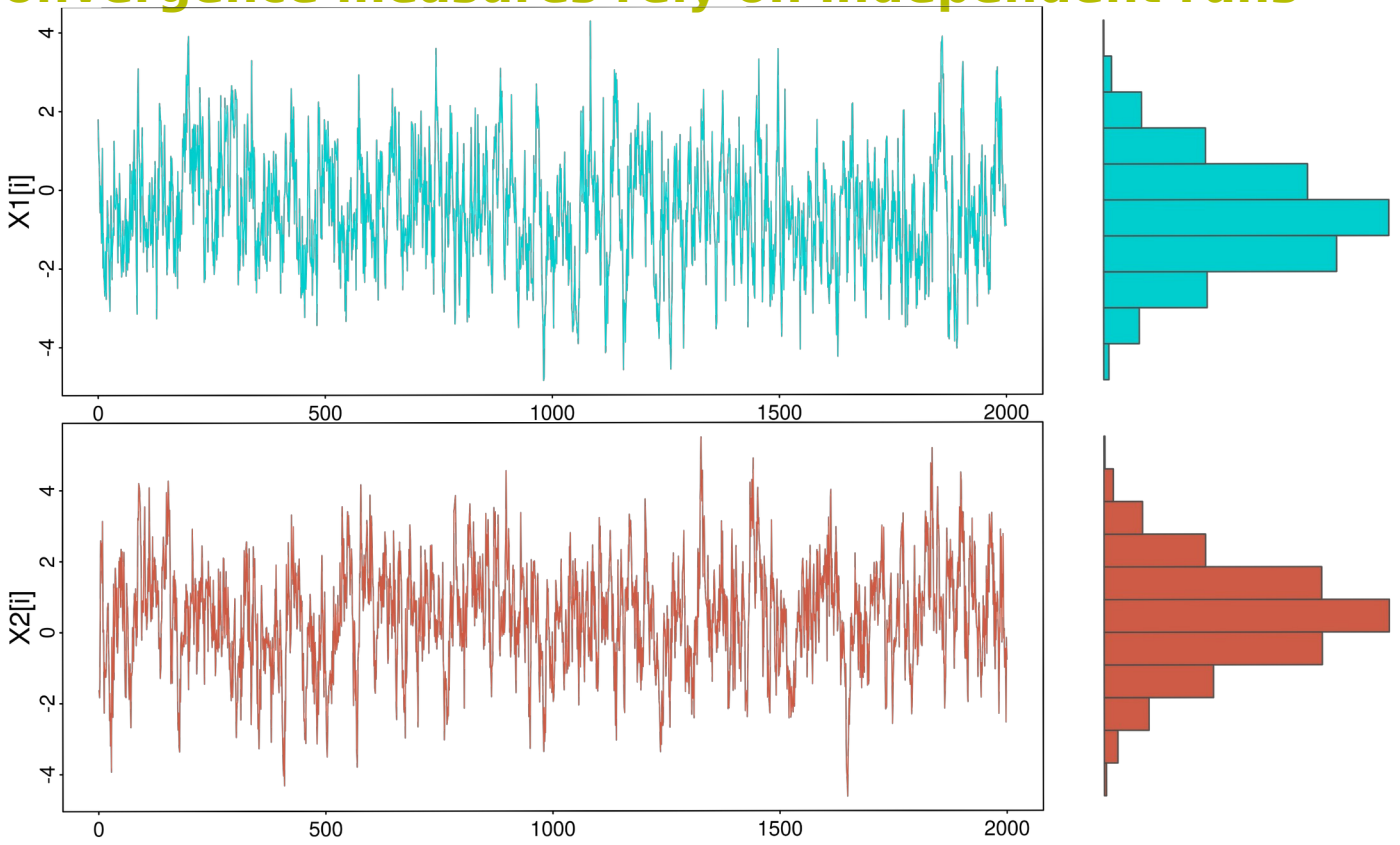
What is the autocorrelation function (ACF)?



Thinning (shown) and burn-in (not shown) reduce correlation



Other convergence measures rely on independent runs



ACF, scaled regeneration quantile plot, potential scale reduction factor (PSRF)

software: RWTY and AWTY

Quadram Institute
Science Health Food Innovation

Maximum Likelihood Inference

IQTREE2 line by line

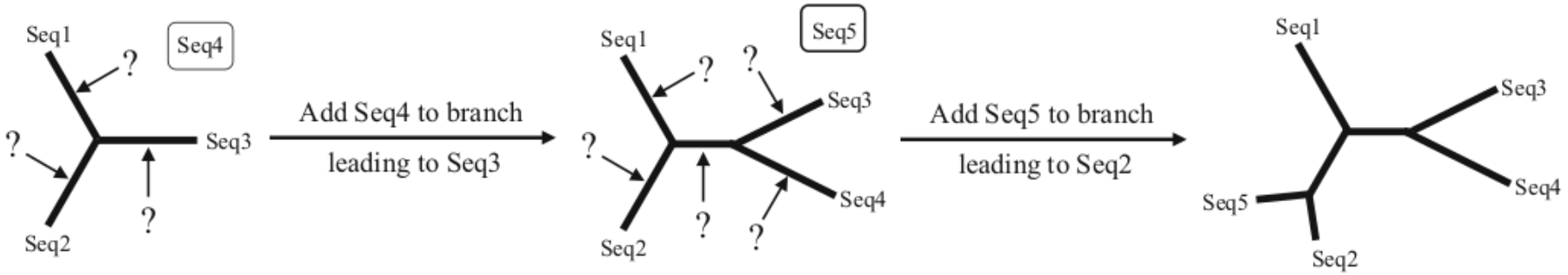
Quadram Institute
Science Health Food Innovation

```
IQ-TREE multicore version 2.2.2.7 COVID-edition for Linux 64-bit built Jun  7 2023
Developed by Bui Quang Minh, James Barbetti, Nguyen Lam Tung,
Olga Chernomor, Heiko Schmidt, Dominik Schrempf, Michael Woodhams, Ly Trong Nhan.

Host:    N114312.nbi.ac.uk (AVX2, FMA3, 15 GB RAM)
Command: iqtree2 --ninit 1000 -t PARS -s 08.full_align.aln --prefix o
Seed:    994063 (Using SPRNG - Scalable Parallel Random Number Generator)
Time:    Wed Nov  8 13:51:34 2023
Kernel:  AVX+FMA - 1 threads (8 CPU cores detected)

HINT: Use -nt option to specify number of threads because your CPU has 8 cores!
HINT: -nt AUTO will automatically determine the best number of threads to use.

Reading alignment file 08.full_align.aln ... Fasta format detected
Reading fasta file: done in 0.000737228 secs using 62.67% CPU
Alignment most likely contains DNA/RNA sequences
Constructing alignment: done in 0.00205722 secs using 70% CPU
Alignment has 10 sequences with 1455 columns, 615 distinct patterns
564 parsimony-informative, 235 singleton sites, 656 constant sites
                         Gap/Ambiguity  Composition  p-value
Analyzing sequences: done in 3.0332e-05 secs
   1  Pseudomonas_E_putida_H      20.69%    passed     15.91%
   2  Pseudomonas_E_orientalis    45.09%    passed     60.31%
   3  Pseudomonas_A_stutzeri_R    45.09%    failed      0.01%
   4  Pseudomonas_E_mendocina_A   20.69%    passed     49.21%
   5  Oblitimonas_alkaliphila     45.09%    failed      0.00%
   6  Pseudomonas_aeruginosa      24.40%    failed      0.00%
   7  Pseudomonas_E_sp002163625   45.09%    passed     23.84%
   8  Pseudomonas_B_luteola       45.09%    failed      0.00%
   9  Azotobacter_vinelandii      24.40%    failed      0.00%
  10  Pseudomonas_C_sp002342705   45.09%    failed      0.00%
****  TOTAL                       36.07%  6 sequences failed composition chi2 test (p-value<5%; df=3)
Checking for duplicate sequences: done in 4.8551e-05 secs
```

Quadram Institute
Science • Health • Food • Innovation

```
Creating fast initial parsimony tree by random order stepwise addition...
0.001 seconds, parsimony score: 1923 (based on 799 sites)
Perform fast likelihood tree search using GTR+I+G model...
Estimate model parameters (epsilon = 5.000)
Perform nearest neighbor interchange...
Optimizing NNI: done in 0.00406809 secs using 99.95% CPU
Estimate model parameters (epsilon = 1.000)
1. Initial log-likelihood: -8801.377
2. Current log-likelihood: -8799.214
3. Current log-likelihood: -8797.933
Optimal log-likelihood: -8797.173
Rate parameters:  A-C: 0.61481  A-G: 2.99951  A-T: 0.94860  C-G: 0.58165  C-T: 4.12668  G-T: 1.00000
Base frequencies:  A: 0.203  C: 0.310  G: 0.243  T: 0.244
Proportion of invariable sites: 0.002
Gamma shape alpha: 6.268
Parameters optimization took 3 rounds (0.014 sec)
Time for fast ML tree search: 0.102 seconds

NOTE: ModelFinder requires 2 MB RAM!
ModelFinder will test up to 484 DNA models (sample size: 1455) ...
 No. Model         -LnL         df  AIC          AICc         BIC
  1  GTR+F         8805.128     25  17660.256    17661.166    17792.325
  2  GTR+F+I       8805.126     26  17662.252    17663.235    17799.604
  3  GTR+F+G4      8796.062     26  17644.124    17645.107    17781.476
  4  GTR+F+I+G4    8796.199     27  17646.399    17647.458    17789.033
  5  GTR+F+R2      8767.610     27  17589.221    17590.280    17731.855
  6  GTR+F+R3      8767.525     29  17593.051    17594.272    17746.251
```

```
Creating fast initial parsimony tree by random order stepwise addition...
0.001 seconds, parsimony score: 1923 (based on 799 sites)
Perform fast likelihood tree search using GTR+I+G model...
Estimate model parameters (epsilon = 5.000)
```




```
Gamma shape alpha: 6.268
Parameters optimization took 3 rounds (0.014 sec)
Time for fast ML tree search: 0.102 seconds

NOTE: ModelFinder requires 2 MB RAM!
ModelFinder will test up to 484 DNA models (sample size: 1455) ...
 No. Model         -LnL         df  AIC          AICc         BIC
  1  GTR+F         8805.128     25  17660.256    17661.166    17792.325
  2  GTR+F+I       8805.126     26  17662.252    17663.235    17799.604
  3  GTR+F+G4      8796.062     26  17644.124    17645.107    17781.476
  4  GTR+F+I+G4    8796.199     27  17646.399    17647.458    17789.033
  5  GTR+F+R2      8767.610     27  17589.221    17590.280    17731.855
  6  GTR+F+R3      8767.525     29  17593.051    17594.272    17746.251
```

doi:10.1007/978-1-4939-6622-6_14

```
Creating fast initial parsimony tree by random o
0.001 seconds, parsimony score: 1923 (based on 7
Perform fast likelihood tree search using GTR+I+
Estimate model parameters (epsilon = 5.000)
Perform nearest neighbor interchange...
Optimizing NNI: done in 0.00406809 secs using 99
Estimate model parameters (epsilon = 1.000)
1. Initial log-likelihood: -8801.377
2. Current log-likelihood: -8799.214
3. Current log-likelihood: -8797.933
Optimal log-likelihood: -8797.173
Rate parameters:  A-C: 0.61481  A-G: 2.99951  A-
Base frequencies:  A: 0.203  C: 0.310  G: 0.243
Proportion of invariable sites: 0.002
Gamma shape alpha: 6.268
Parameters optimization took 3 rounds (0.014 sec
Time for fast ML tree search: 0.102 seconds

NOTE: ModelFinder requires 2 MB RAM!
ModelFinder will test up to 484 DNA models (samp
 No. Model         -LnL         df  AIC
  1  GTR+F         8805.128     25  17660.256
  2  GTR+F+I       8805.126     26  17662.252
  3  GTR+F+G4      8796.062     26  17644.124
  4  GTR+F+I+G4    8796.199     27  17646.399
  5  GTR+F+R2      8767.610     27  17589.221
  6  GTR+F+R3      8767.525     29  17593.051
```

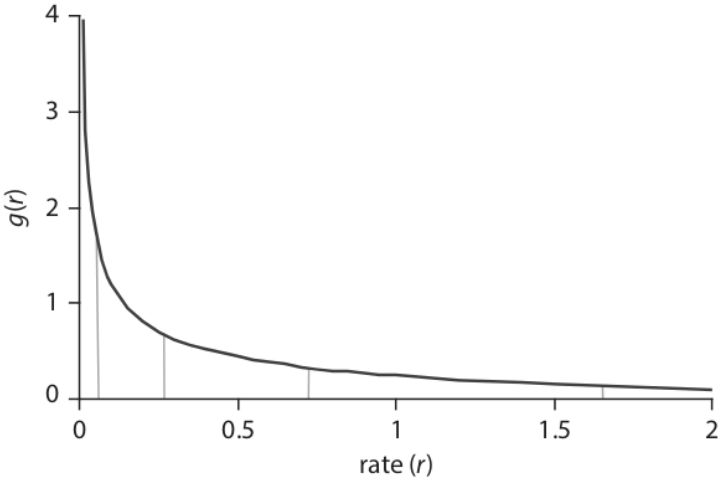


Fig. 4.9 The discrete gamma model of variable rates across sites uses K equal-probability categories to approximate the continuous gamma distribution, with the mean rate in each category used to represent all rates in that category. Shown here is the use of $K = 5$ categories to approximate the gamma density $g(r)$ with $\alpha = 0.5$ (and mean 1). The four vertical lines are at $r = 0.06418,\ 0.27500,\ 0.70833,$ and 1.64237. These are the 20%, 40%, 60%, and 80% percentiles of the distribution and cut the density into five categories, each of proportion 1/5. The mean rates in the five categories are 0.02121, 0.15549, 0.46708, 1.10712, and 3.24910.

1.00000

from http://abacus.gene.ucl.ac.uk/MESA/

```
Akaike Information Criterion:           TIM3e+R2
Corrected Akaike Information Criterion: TIM3e+R2
Bayesian Information Criterion:         TNe+R2
Best-fit model: TNe+R2 chosen according to BIC

All model information printed to o.model.gz
CPU time for ModelFinder: 4.589 seconds (0h:0m:4s)
Wall-clock time for ModelFinder: 4.685 seconds (0h:0m:4s)

NOTE: 0 MB RAM (0 GB) is required!
Estimate model parameters (epsilon = 0.100)
1. Initial log-likelihood: -8752.796
Optimal log-likelihood: -8752.794
Rate parameters:  A-C: 1.00000  A-G: 4.15899  A-T: 1.00000  C-G: 1.00000  C-T: 7.21407  G-T: 1.00000
Base frequencies:  A: 0.250  C: 0.250  G: 0.250  T: 0.250
Site proportion and rates:  (0.950,0.622) (0.050,8.129)
Parameters optimization took 1 rounds (0.004 sec)
Wrote distance file to...
Computing ML distances based on estimated model parameters...
Calculating distance matrix: done in 0.000293082 secs using 97.58% CPU
Computing ML distances took 0.000417 sec (of wall-clock time) 0.000413 sec (of CPU time)
Setting up auxiliary I and S matrices: done in 5.2725e-05 secs using 94.83% CPU
Constructing RapidNJ tree: done in 0.000119629 secs using 96.97% CPU
Computing RapidNJ tree took 0.000281 sec (of wall-clock time) 0.000276 sec (of CPU time)
Log-likelihood of RapidNJ tree: -8759.295
--------------------------------------------------------------------
|             INITIALIZING CANDIDATE TREE SET                      |
--------------------------------------------------------------------
```

```
Akaike Information Criterion:           TIM3e+R2
Corrected Akaike Information Criterion: TIM3e+R2
Bayesian Information Criterion:         TNe+R2
Best-fit model: TNe+R2

All model information p
CPU time for ModelFinde
Wall-clock time for Mod

NOTE: 0 MB RAM (0 GB) i
Estimate model paramete
1. Initial log-likeliho
Optimal log-likelihood:
Rate parameters:  A-C:
Base frequencies:  A: 0
Site proportion and rat
Parameters optimization
Wrote distance file to.
Computing ML distances
Calculating distance ma
Computing ML distances
Setting up auxiliary I
Constructing RapidNJ tr
Computing RapidNJ tree
Log-likelihood of Rapid
-----------------------
|             INITIALIZ
-----------------------
```

Model	df	Explanation	Code
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
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
GTR	8	General time reversible model with unequal rates and unequal base freq. (Tavare, 1986).	012345

The last column `Code` is a 6-digit code defining the equality constraints for 6 *relative* substitution rates: A-C, A-G, A-T, C-G, C-T and G-T.

```
Akaike Information Criterion:           TIM3e+R2
Corrected Akaike Information Criterion: TIM3e+R2
Bayesian Information Criterion:         TNe+R2
Best-fit
All mode
CPU time
Wall-clo
NOTE: 0
Estimate
1. Initi
Optimal
Rate par
Base fre
Site pro
Paramete
Wrote di
Computin
Calculat
Computin
Setting
Construc
Computin
Log-like
--------
|
--------
```

Table 1.1 Substitution rate matrices for commonly used Markov models of nucleotide substitution

	p	From	To T	To C	To A	To G
JC69 (Jukes and Cantor 1969)	1	T	·	λ	λ	λ
		C	λ	·	λ	λ
		A	λ	λ	·	λ
		G	λ	λ	λ	·
TN93 (Tamura and Nei 1993)	6	T	·	$\alpha_1\pi_C$	$\beta\pi_A$	$\beta\pi_G$
		C	$\alpha_1\pi_T$	·	$\beta\pi_A$	$\beta\pi_G$
		A	$\beta\pi_T$	$\beta\pi_C$	·	$\alpha_2\pi_G$
		G	$\beta\pi_T$	$\beta\pi_C$	$\alpha_2\pi_A$	·
GTR (REV) (Tavaré 1986; Yang 1994b; Zharkikh 1994)	9	T	·	$a\pi_C$	$b\pi_A$	$c\pi_G$
		C	$a\pi_T$	·	$d\pi_A$	$e\pi_G$
		A	$b\pi_T$	$d\pi_C$	·	$f\pi_G$
		G	$c\pi_T$	$e\pi_C$	$f\pi_A$	·

from http://abacus.gene.ucl.ac.uk/MESA/

```
Akaike Information Criterion:           TIM3e+R2
Corrected Akaike Information Criterion: TIM3e+R2
Bayesian Information Criterion:         TNe+R2
Best-fit model: TNe+R2 chosen according to BIC

All model information printed to o.model.gz
CPU time for ModelFinder: 4.589 seconds (0h:0m:4s)
Wall-clock time for ModelFinder: 4.685 seconds (0h:0m:4s)

NOTE: 0 MB RAM (0 GB) is required!
Estimate model parameters (epsilon = 0.100)
1. Initial log-likelihood: -8752.796
Optimal log-likelihood: -8752.794
Rate parameters:  A-C: 1.00000  A-G: 4.15899  A-T: 1.00000  C-G: 1.00000  C-T: 7.21407  G-T: 1.00000
Base frequencies:  A: 0.250  C: 0.250  G: 0.250  T: 0.250
Site proportion and rates:  (0.950,0.622) (0.050,8.129)
Parameters optimization took 1 rounds (0.004 sec)
```

Model	df	Explanation	Code
TNe	2	Like `TN` but equal base freq.	010020

The last column `Code` is a 6-digit code defining the equality constraints for 6 *relative* substitution rates: A-C, A-G, A-T, C-G, C-T and G-T.

```
Log-likelihood of RapidNJ tree: -8759.295
--------------------------------------------------------------------
|             INITIALIZING CANDIDATE TREE SET                      |
--------------------------------------------------------------------
```

Quadram Institute
Science & Health & Food & Innovation

```
--------------------------------------------------------------------
|             INITIALIZING CANDIDATE TREE SET                      |
--------------------------------------------------------------------
Generating 998 parsimony trees... 0.777 second
Computing log-likelihood of 198 initial trees ... 0.169 seconds
Current best score: -8752.794

Do NNI search on 20 best initial trees
Optimizing NNI: done in 0.00232697 secs using 95.32% CPU
Estimate model parameters (epsilon = 0.100)
BETTER TREE FOUND at iteration 1: -8752.794
Optimizing NNI: done in 0.00581815 secs using 99.96% CPU
Optimizing NNI: done in 0.0056042 secs using 99.98% CPU
Optimizing NNI: done in 0.00809975 secs using 99.93% CPU
Optimizing NNI: done in 0.00765216 secs using 99.96% CPU
Optimizing NNI: done in 0.0058694 secs using 40.98% CPU
```

```
Iteration 10 / LogL: -8753.261 / Time: 0h:0m:1s
```

```
Optimizing NNI: done in 0.00756817 secs using 99.98% CPU
Optimizing NNI: done in 0.00564065 secs using 99.97% CPU
Iteration 20 / LogL: -8754.235 / Time: 0h:0m:1s
Finish initializing candidate tree set (2)
Current best tree score: -8752.794 / CPU time: 1.101
Number of iterations: 20
--------------------------------------------------------------------
|               OPTIMIZING CANDIDATE TREE SET                      |
--------------------------------------------------------------------
Optimizing NNI: done in 0.0116323 secs using 99.99% CPU
Optimizing NNI: done in 0.00910824 secs using 99.99% CPU
```

```
Iteration 100 / LogL: -8753.231 / Time: 0h:0m:1s (0h:0m:0s left)
Optimizing NNI: done in 0.0105482 secs using 62.7% CPU
Optimizing NNI: done in 0.0139313 secs using 99.82% CPU
TREE SEARCH COMPLETED AFTER 102 ITERATIONS / Time: 0h:0m:1s

--------------------------------------------------------------------
|                    FINALIZING TREE SEARCH                        |
--------------------------------------------------------------------
Performs final model parameters optimization
Estimate model parameters (epsilon = 0.010)
1. Initial log-likelihood: -8752.794
Optimal log-likelihood: -8752.794
Rate parameters:  A-C: 1.00000  A-G: 4.16143  A-T: 1.00000  C-G: 1.00000  C-T: 7.22268  G-T: 1.00000
Base frequencies:  A: 0.250  C: 0.250  G: 0.250  T: 0.250
Site proportion and rates:  (0.950,0.621) (0.050,8.155)
Parameters optimization took 1 rounds (0.004 sec)
BEST SCORE FOUND : -8752.794
Total tree length: 4.635

Total number of iterations: 102
CPU time used for tree search: 1.673 sec (0h:0m:1s)
Wall-clock time used for tree search: 1.900 sec (0h:0m:1s)
Total CPU time used: 1.700 sec (0h:0m:1s)
Total wall-clock time used: 1.925 sec (0h:0m:1s)

Analysis results written to:
  IQ-TREE report:                o.iqtree
  Maximum-likelihood tree:       o.treefile
  Likelihood distances:          o.mldist
  Screen log file:               o.log
```

Identical sequences are identified and sometimes removed during inference

```
NOTE: NZ_CP041754.1 is identical to NC_005773.3 but kept for subsequent analysis
NOTE: NZ_CP026676.1 is identical to NC_022738.1 but kept for subsequent analysis
NOTE: NZ_CP014343.1 is identical to NZ_CP023299.1 but kept for subsequent analysis
NOTE: NZ_CP008863.1 is identical to NZ_LT608330.1 but kept for subsequent analysis
NOTE: NZ_CP061848.1 is identical to NC_019905.1 but kept for subsequent analysis
NOTE: NZ_CP045359.1 is identical to NZ_CP045349.1 but kept for subsequent analysis
NOTE: NZ_CP030750.1 is identical to NZ_CP011789.1 but kept for subsequent analysis
NOTE: NZ_CP011110.1 is identical to NZ_CP010892.1 but kept for subsequent analysis
NOTE: NZ_LR130527.1 is identical to NC_017548.1 but kept for subsequent analysis
NOTE: NZ_CP033439.1 is identical to NZ_CP070467.1 but kept for subsequent analysis
```

```
Checking for duplicate sequences: done in 0.00459436 secs using 89.39% CPU
Identifying sites to remove: done in 0.00152092 secs using 99.87% CPU
NOTE: 327 identical sequences (see below) will be ignored for subsequent analysis
NOTE: NC_017911.1 (identical to NZ_CP017296.1) is ignored but added at the end
```

```
NOTE: NZ_CP027721.1 (identical to NZ_CP027722.1) is ignored but added at the end
NOTE: NZ_CP031659.1 (identical to NZ_CP013989.1) is ignored but added at the end
NOTE: NZ_CP029088.1 (identical to NZ_CP013989.1) is ignored but added at the end
NOTE: NZ_CP010555.1 (identical to NZ_CP013989.1) is ignored but added at the end
NOTE: NZ_CP050323.1 (identical to NZ_CP013989.1) is ignored but added at the end
NOTE: NZ_CP065866.1 (identical to NZ_CP013989.1) is ignored but added at the end
NOTE: NZ_CP058332.1 (identical to NZ_CP013989.1) is ignored but added at the end
NOTE: NZ_LR739068.1 (identical to NZ_CP013989.1) is ignored but added at the end
NOTE: NZ_CP046069.1 (identical to NZ_CP013989.1) is ignored but added at the end
NOTE: NZ_CP025053.1 (identical to NZ_CP013989.1) is ignored but added at the end
```

File {prefix}.iqtree has some more information and important notices

```
Input data: 700 sequences with 445 nucleotide sites
Number of constant sites: 188 (= 42.2472% of all sites)
Number of invariant (constant or ambiguous constant) sites: 188 (= 42.2472% of all sites)
Number of parsimony informative sites: 246
Number of distinct site patterns: 266
```

```
Total tree length (sum of branch lengths): 11.2056
Sum of internal branch lengths: 6.7741 (60.4527% of tree length)

WARNING: 117 near-zero internal branches (<0.0022) should be treated with caution
         Such branches are denoted by '**' in the figure below

NOTE: Tree is UNROOTED although outgroup taxon 'NZ_CP017296.1' is drawn at root
```

Thank you

UKRI Biotechnology and Biological Sciences Research Council

Quadram Institute
Norwich Research Park
Norfolk NR4 7UQ

@leomrtns@mstdn.science

Quadram Institute
Science Health Food Innovation