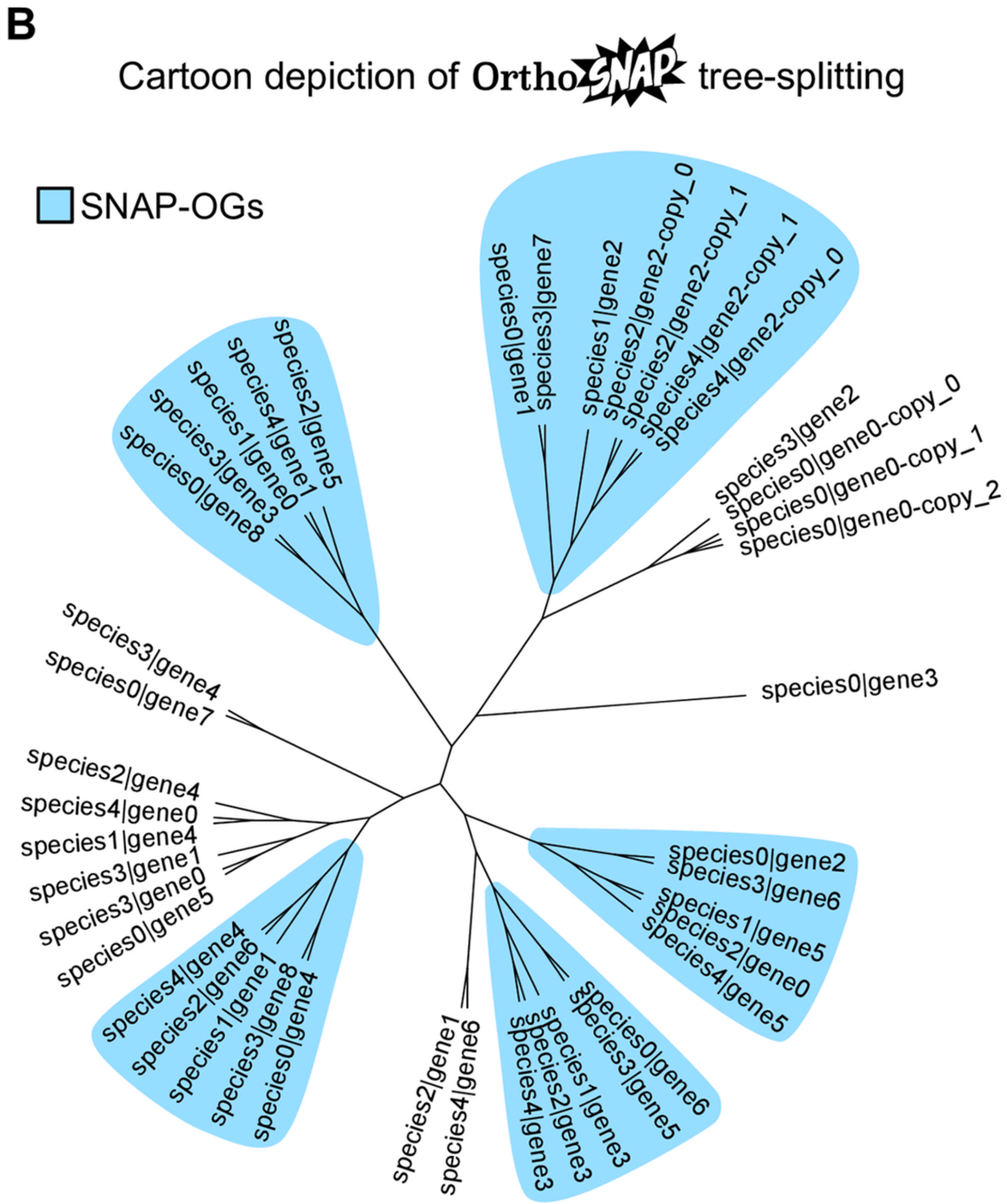
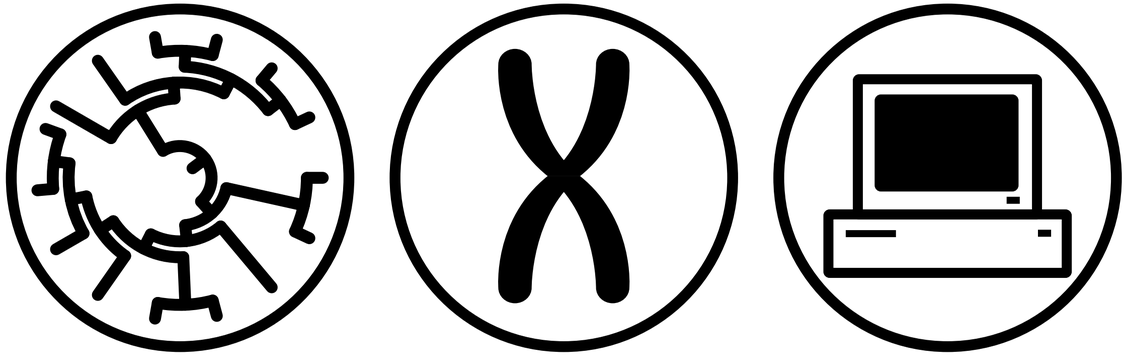
- Not all genes will appear in SC-OGs
- -f input can also be unaligned sequences
  - Enables alignment of SNAP-OGs
- If you still have questions, please feel to ask me to draw a worked example
- You may finish this session early, you can certainly work on previous unfinished practicals
- At 9:30, we will discuss phylogenetic lingo (Gemma & Karin)



Concatenation and partitioning

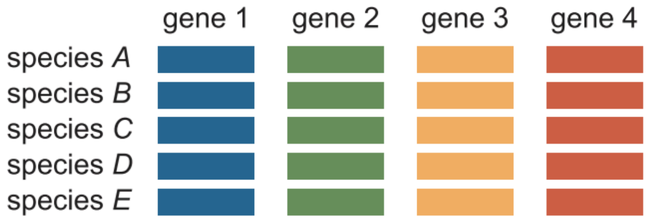
@JLSteenwyk https://jlsteenwyk.com/

Outline



- Major methods in phylogenomics
- Substitution models, in (very) brief
- Methods to concatenate sequences
- Phylogenomic subsampling

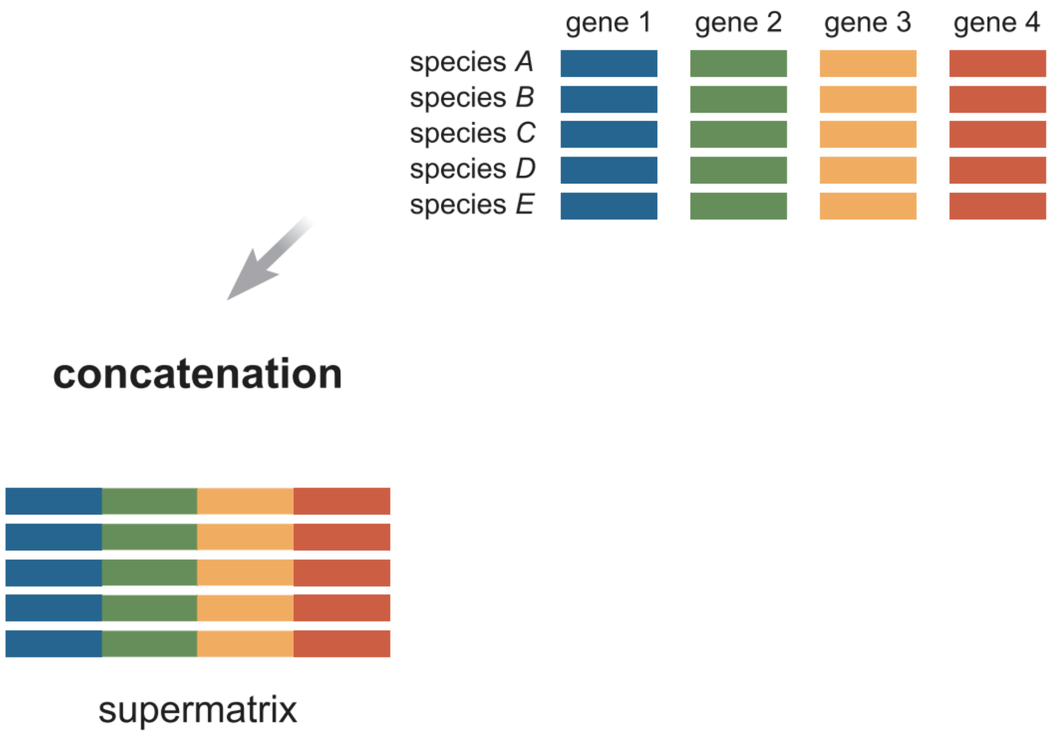
Major methods in phylogenomics



@JLSteenwyk

Liu *et al.* 2015, Ann. N.Y. Acad. Sci.

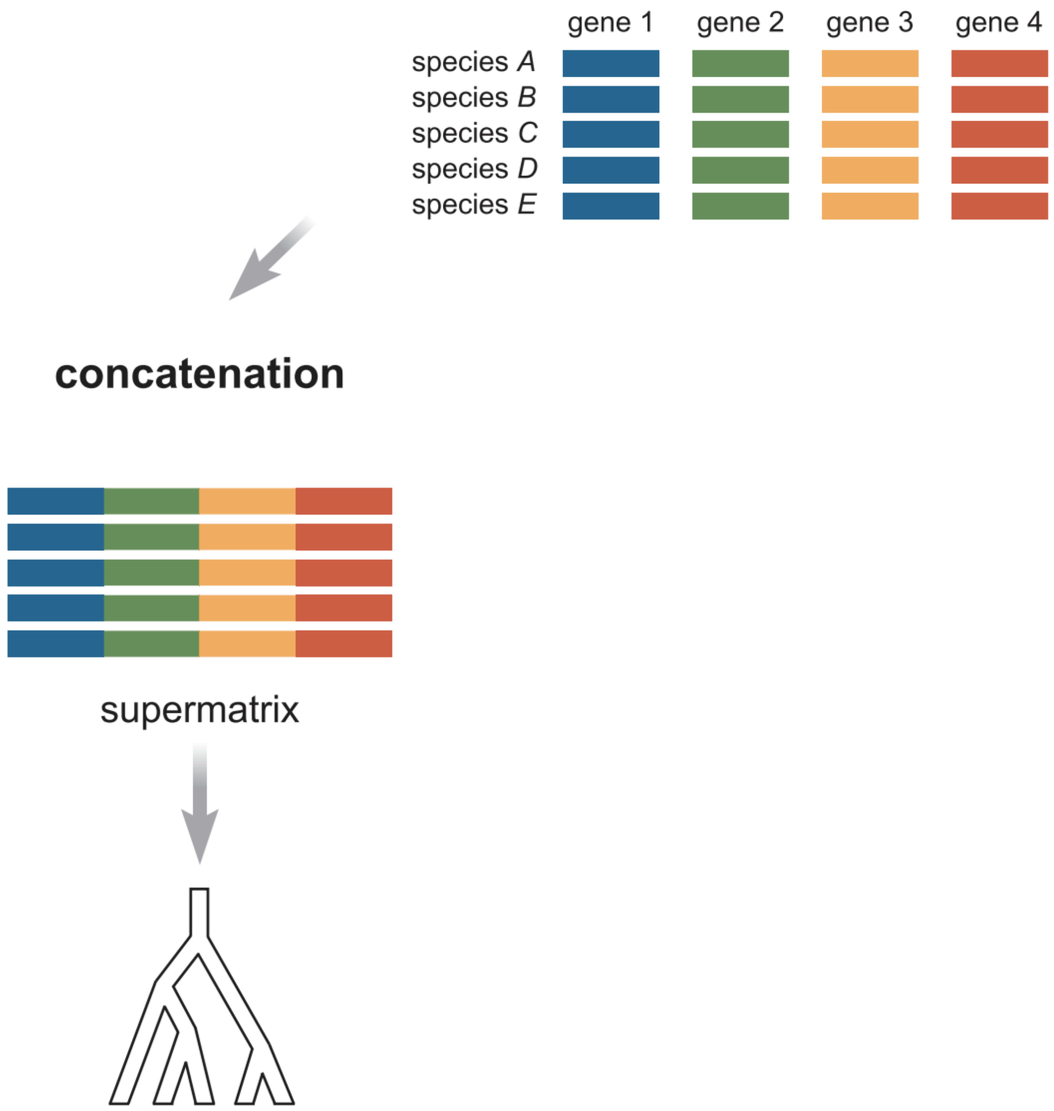
Major methods in phylogenomics



@JLSteenwyk

Liu *et al.* 2015, Ann. N.Y. Acad. Sci.

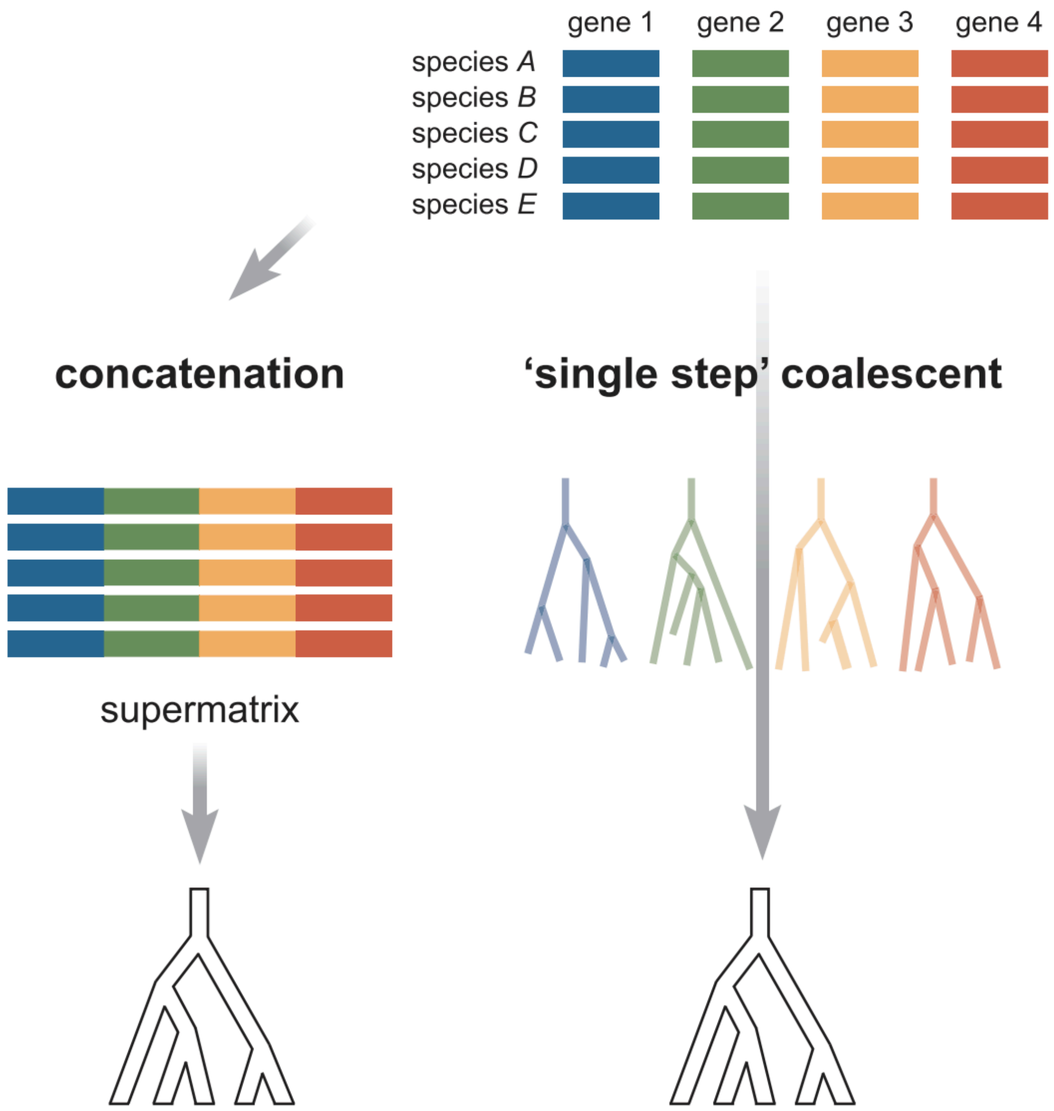
Major methods in phylogenomics



@JLSteenwyk

Liu *et al.* 2015, Ann. N.Y. Acad. Sci.

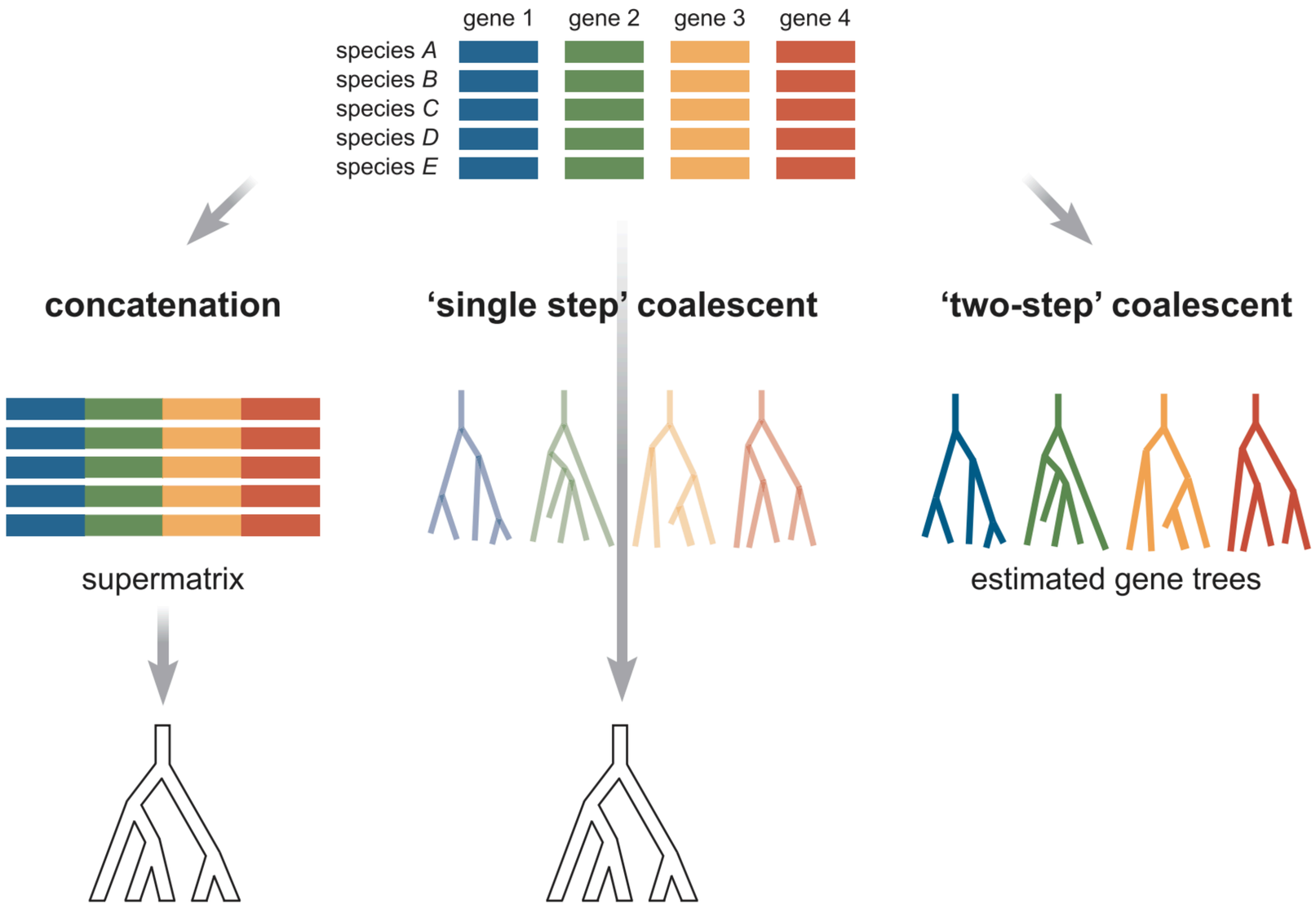
Major methods in phylogenomics



@JLSteenwyk

Liu *et al.* 2015, Ann. N.Y. Acad. Sci.

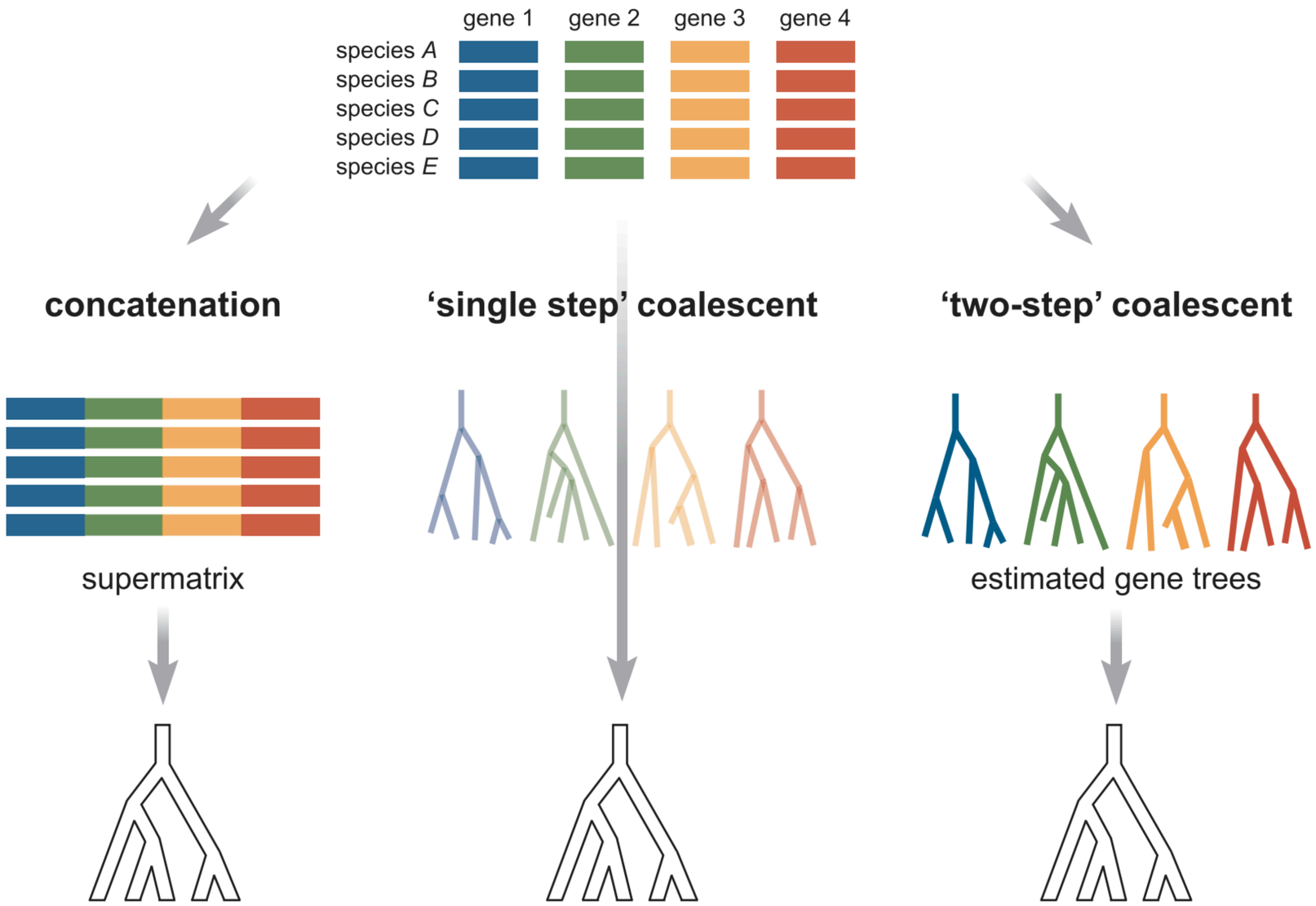
Major methods in phylogenomics



@JLSteenwyk

Liu *et al.* 2015, Ann. N.Y. Acad. Sci.

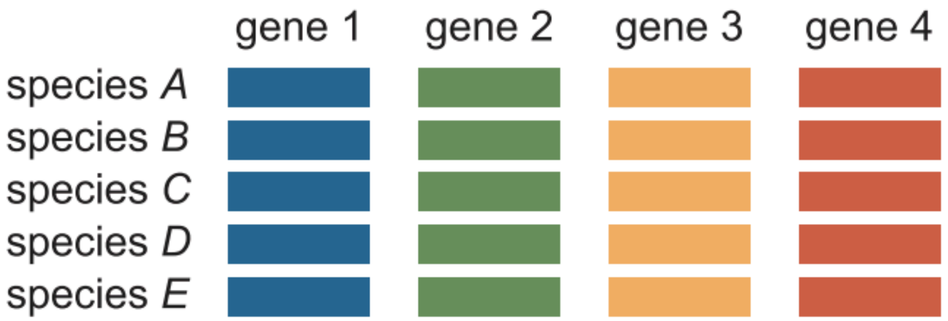
Major methods in phylogenomics



@JLSteenwyk

Liu *et al.* 2015, Ann. N.Y. Acad. Sci.

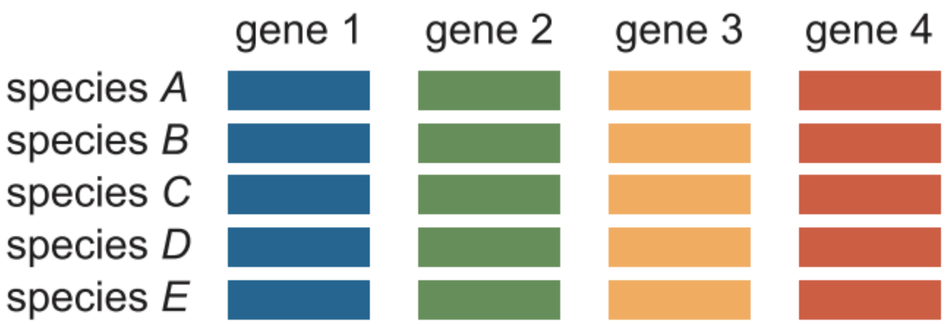
Major methods in phylogenomics



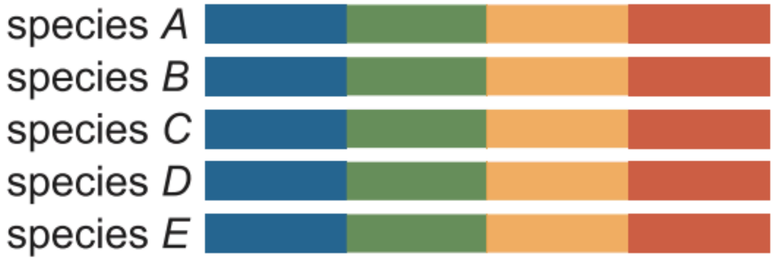
@JLSteenwyk

Liu *et al.* 2015, Ann. N.Y. Acad. Sci.

Major methods in phylogenomics



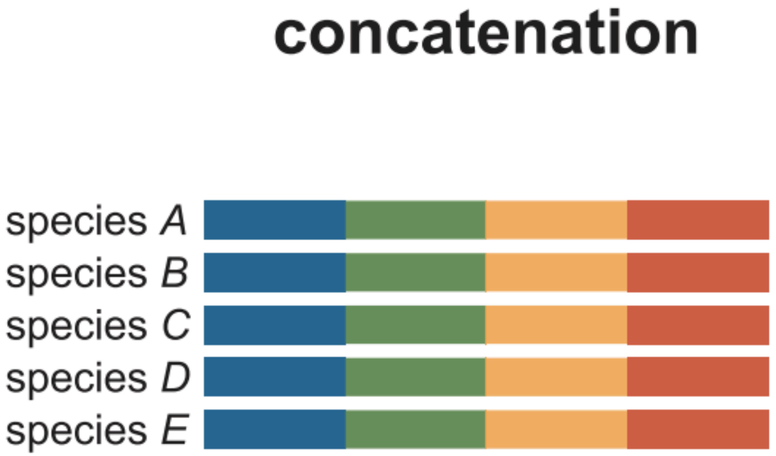
concatenation



@JLSteenwyk

Liu *et al.* 2015, Ann. N.Y. Acad. Sci.

The partition file summarizes gene boundaries



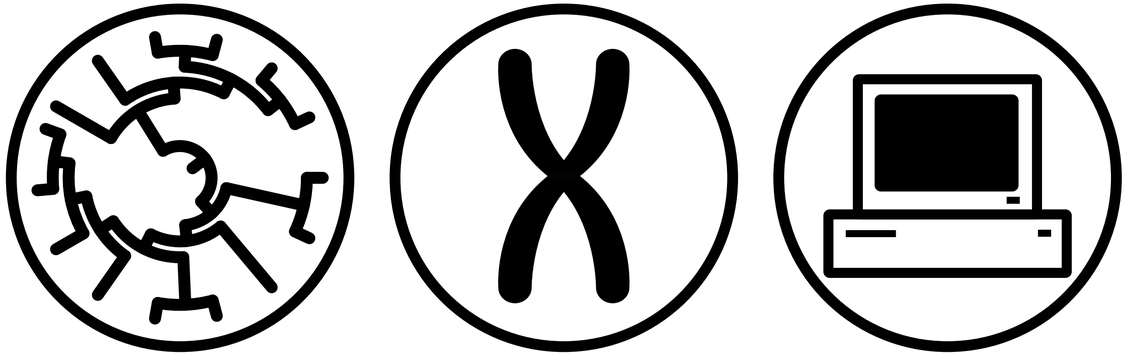
Model, Partition ID = start and stop boundaries

Model, Blue = 1-481
Model, Green = 482-1054
Model, Yellow = 1055-1492
Model, Red = 1493-1918

@JLSteenwyk

Liu *et al.* 2015, Ann. N.Y. Acad. Sci.

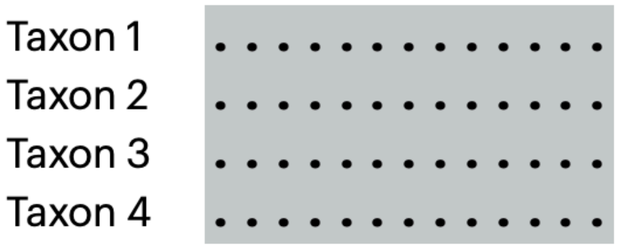
Outline



- Major methods in phylogenomics
- **Substitution models, in (very) brief**
- Methods to concatenate sequences
- Phylogenomic subsampling

Models can be applied to varying portions of the matrix

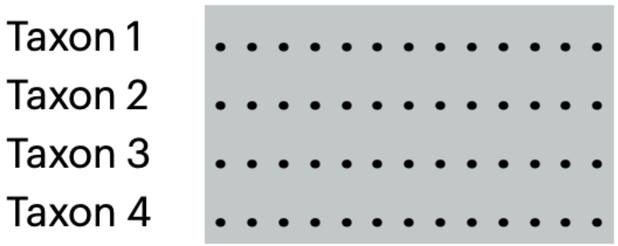
Site-homogeneous model



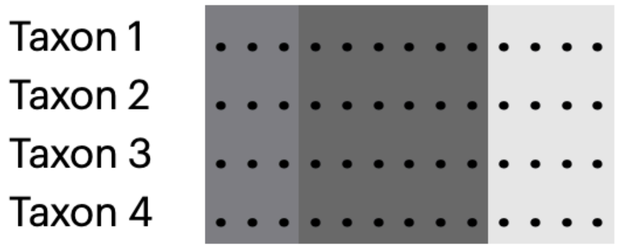
@JLSteenwyk

Models can be applied to varying portions of the matrix

Site-homogeneous model



Site-homogeneous with partitioning



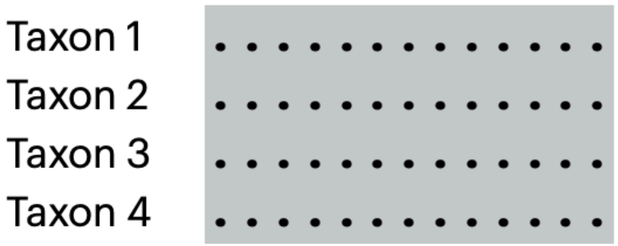
Partitions can be genes
or algorithmically defined

@JLSteenwyk

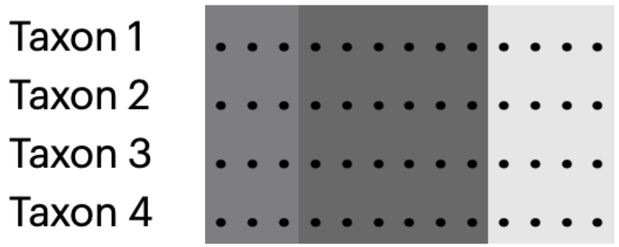
Steenwyk et al. (2023) *Nature Reviews Genetics*

Models can be applied to varying portions of the matrix

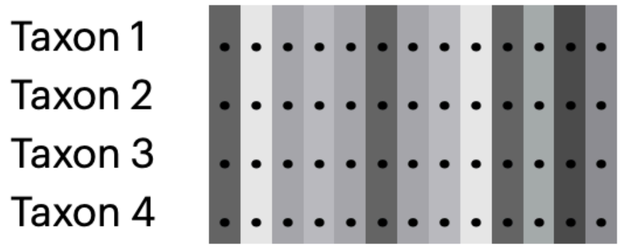
Site-homogeneous model



Site-homogeneous with partitioning



Site-heterogeneous model

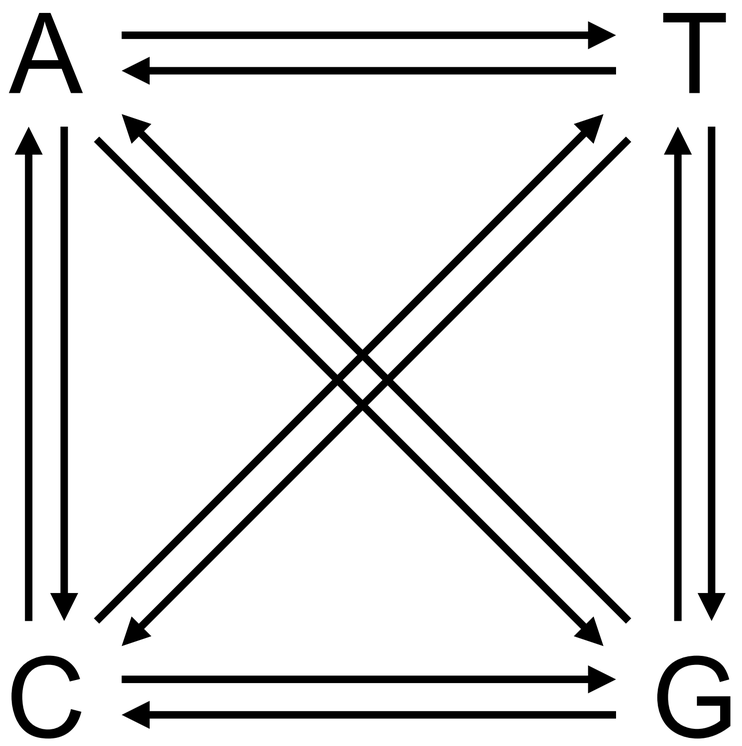


Partitions can be genes
or algorithmically defined

@JLSteenwyk

A (very) brief mention of substitution models

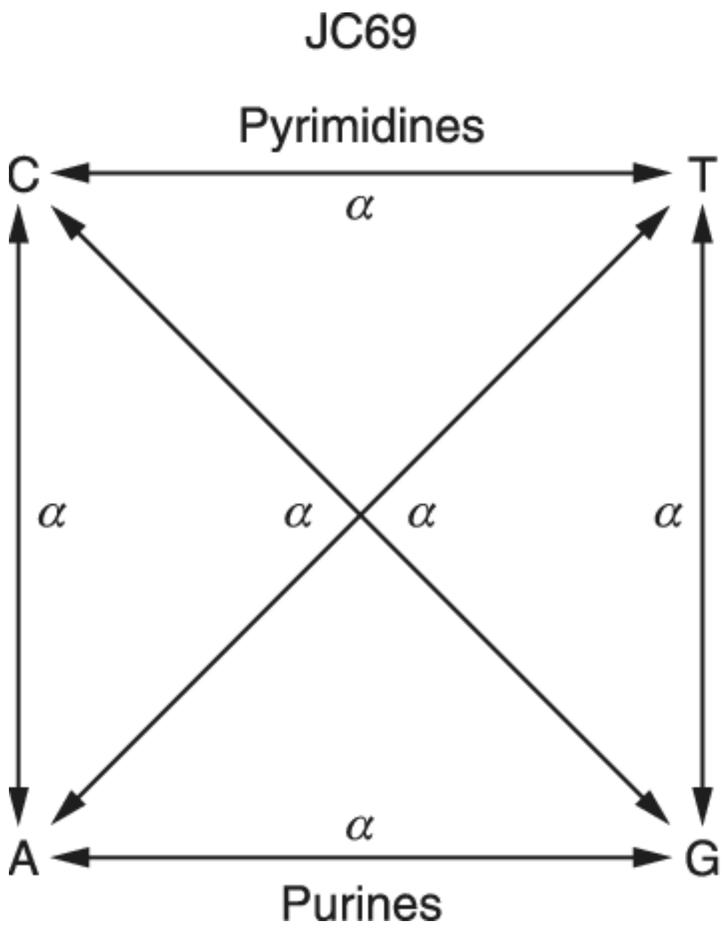
Markov models that describe rates of nucleotide or amino acid substitutions in a locus during evolution



@JLSteenwyk

Steenwyk et al. (2023) *Nature Reviews Genetics*

The simplest model: Jukes Cantor

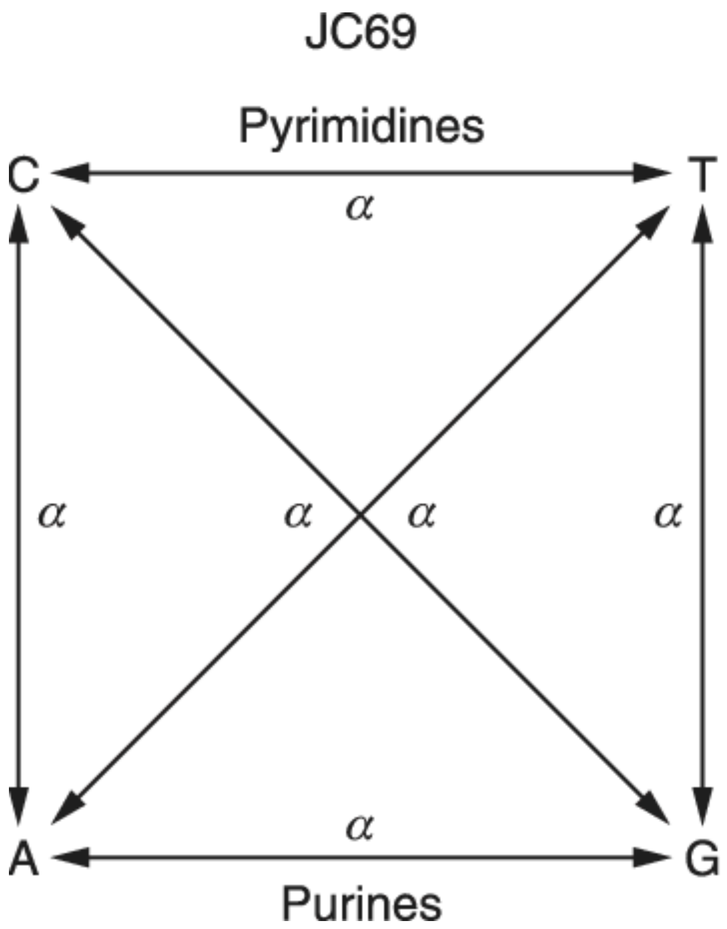


Equal substitution rates
& equal base frequencies

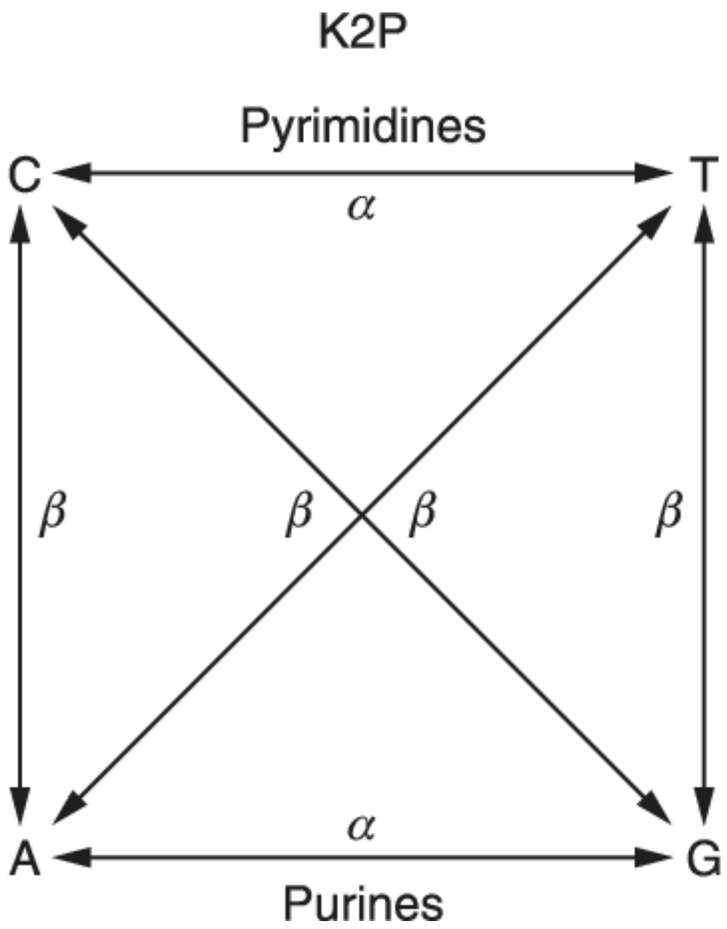
@JLSteenwyk

Jukes and Cantor, 1969; Kimura, 1980

The slightly complex model: K2P



Equal substitution rates
& equal base frequencies



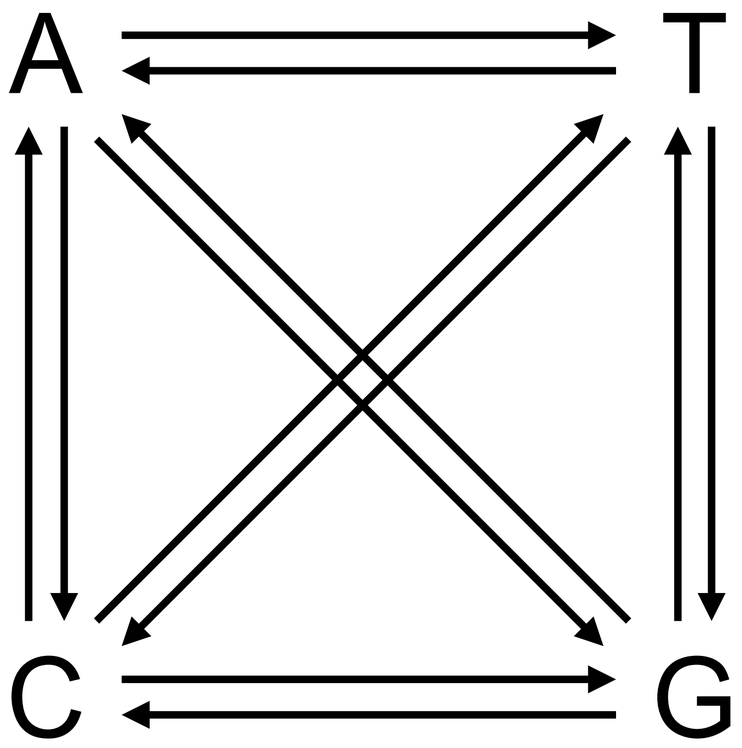
Unequal transition/transversion
rates and equal base freq.

@JLSteenwyk

Jukes and Cantor, 1969; Kimura, 1980

The most complex model: GTR

General time reversible model
with unequal rates and unequal base freq.



(Tavare, 1986

IQ-TREE docs has great explanations

IQ-TREE | Download | News | Web server | Docs | Workshop | About

Protein models

http://www.iqtree.org/doc/Substitution-Models

Amino-acid exchange rate matrices

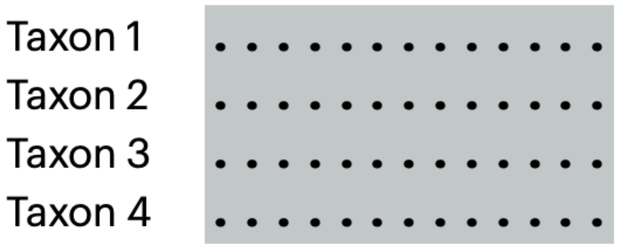
IQ-TREE supports all common empirical amino-acid exchange rate matrices (alphabetical order):

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).
FLAVI	viral	Flavivirus (Le and Vinh, 2020).
FLU	viral	Influenza virus (Dang et al., 2010).
		General time reversible models with 190 rate parameters. *WARNING: Be careful*

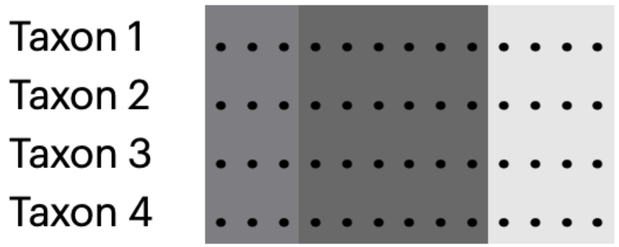
@JLSteenwyk

Models can be applied to varying portions of the matrix

Site-homogeneous model

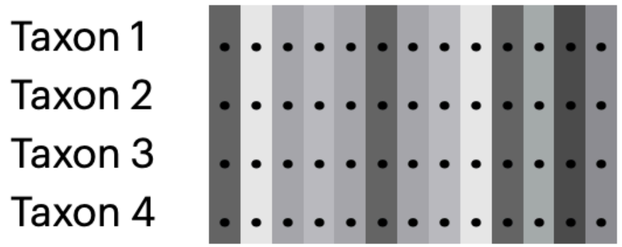


Site-homogeneous with partitioning



Partitions can be genes
or algorithmically defined

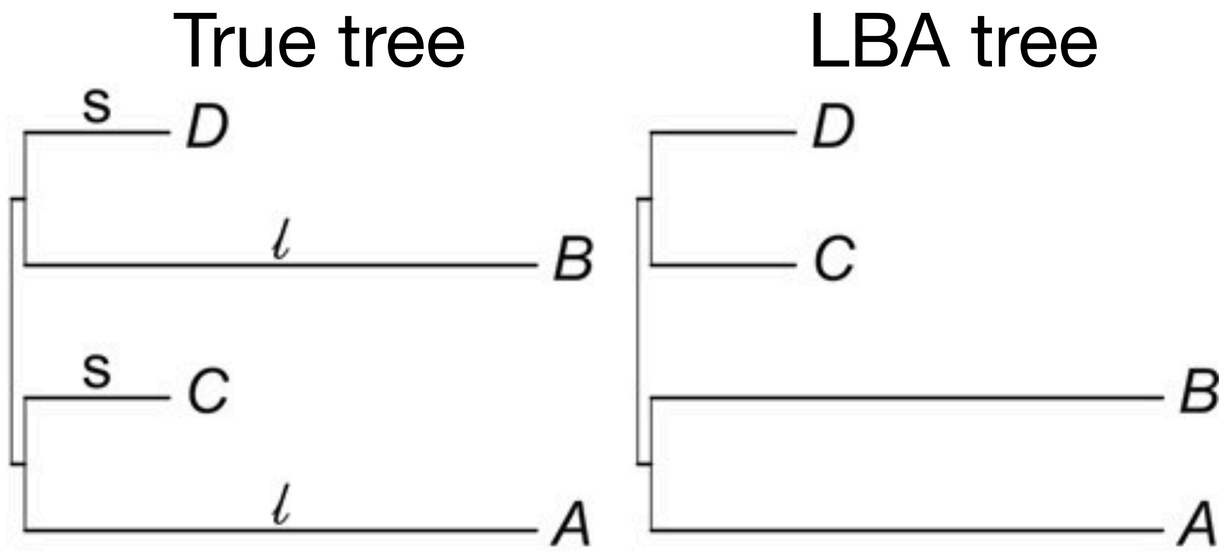
Site-heterogeneous model



@JLSteenwyk

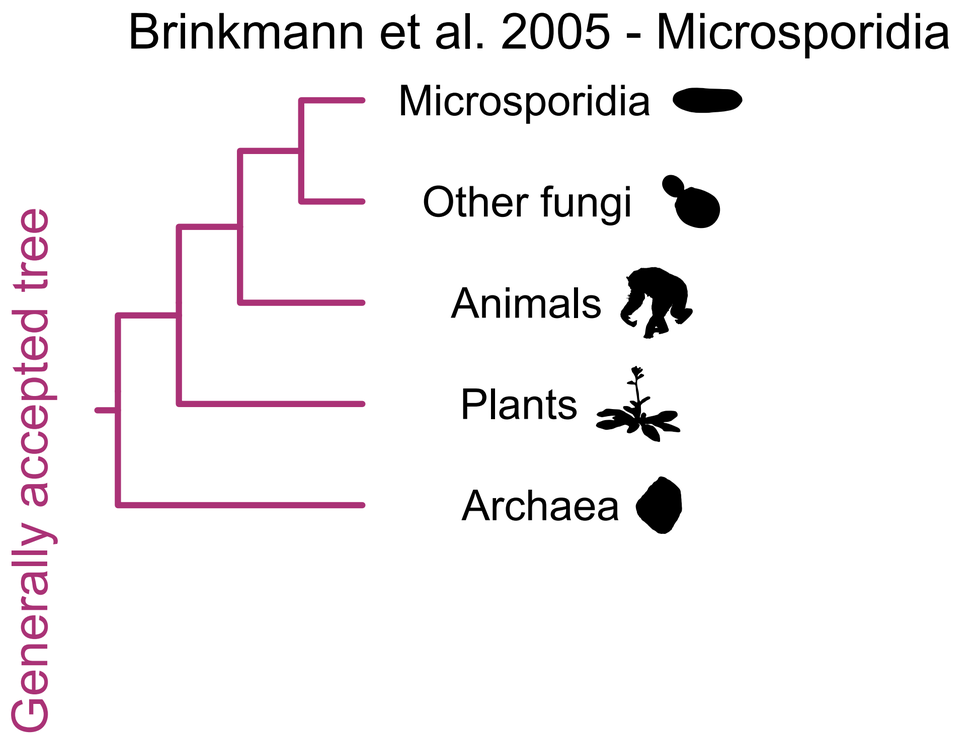
What drives LBA?

when divergent taxa or clades with long branch lengths (i.e., many character changes occurring over time) are inferred as each other's closest relative due to convergent evolution of a given character (e.g., amino acid substitution)

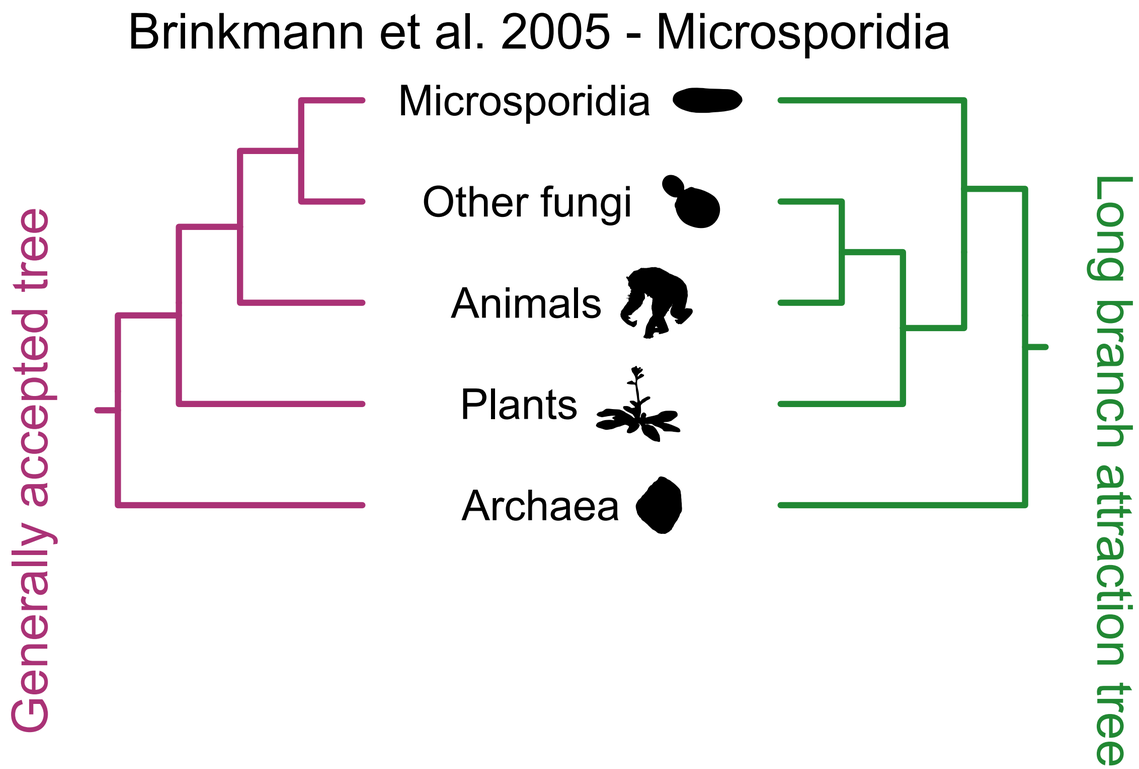


@JLSteenwyk

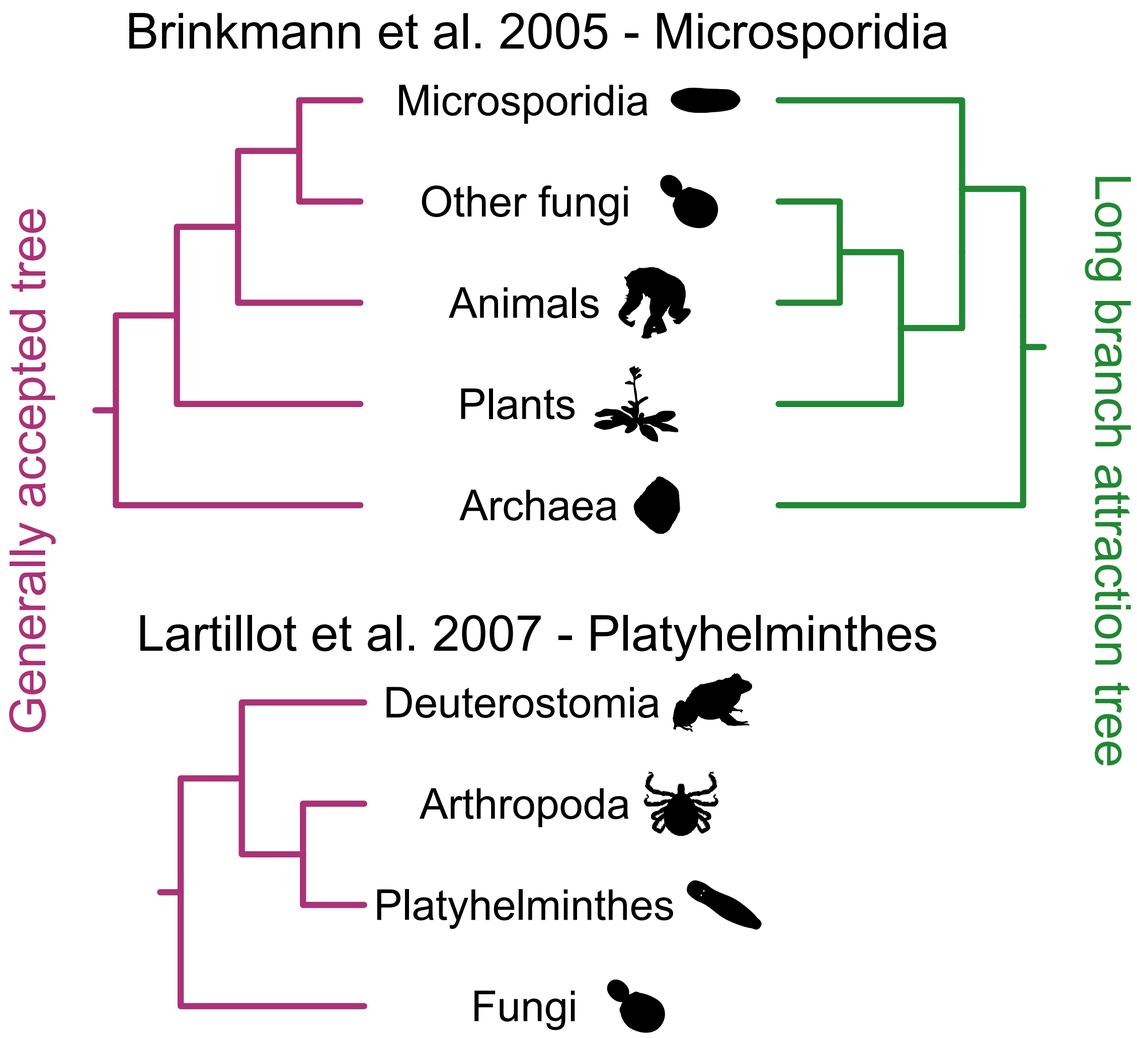
Microsporidia are early diverging fungi



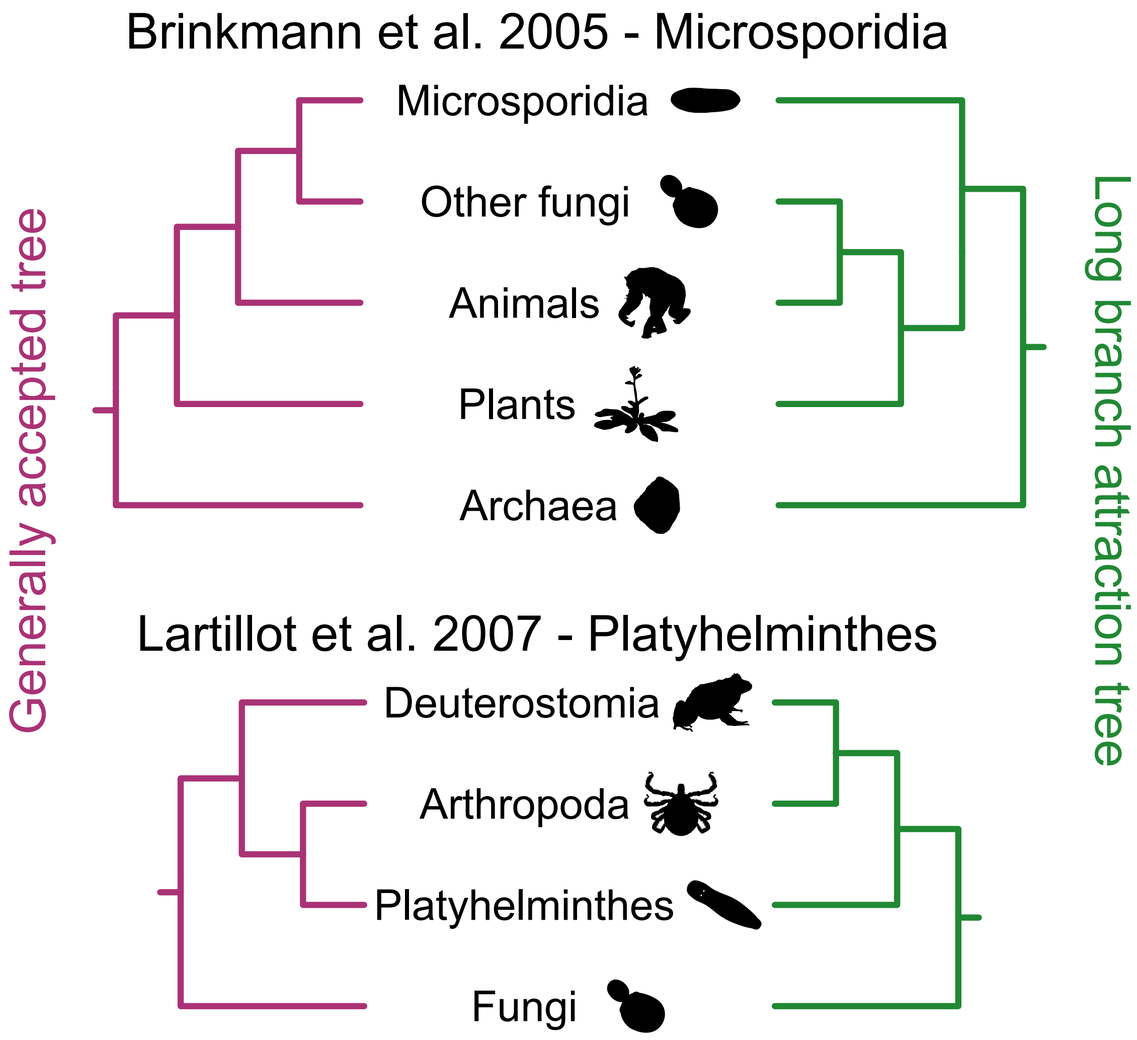
Long-branch attraction can result in an erroneous tree



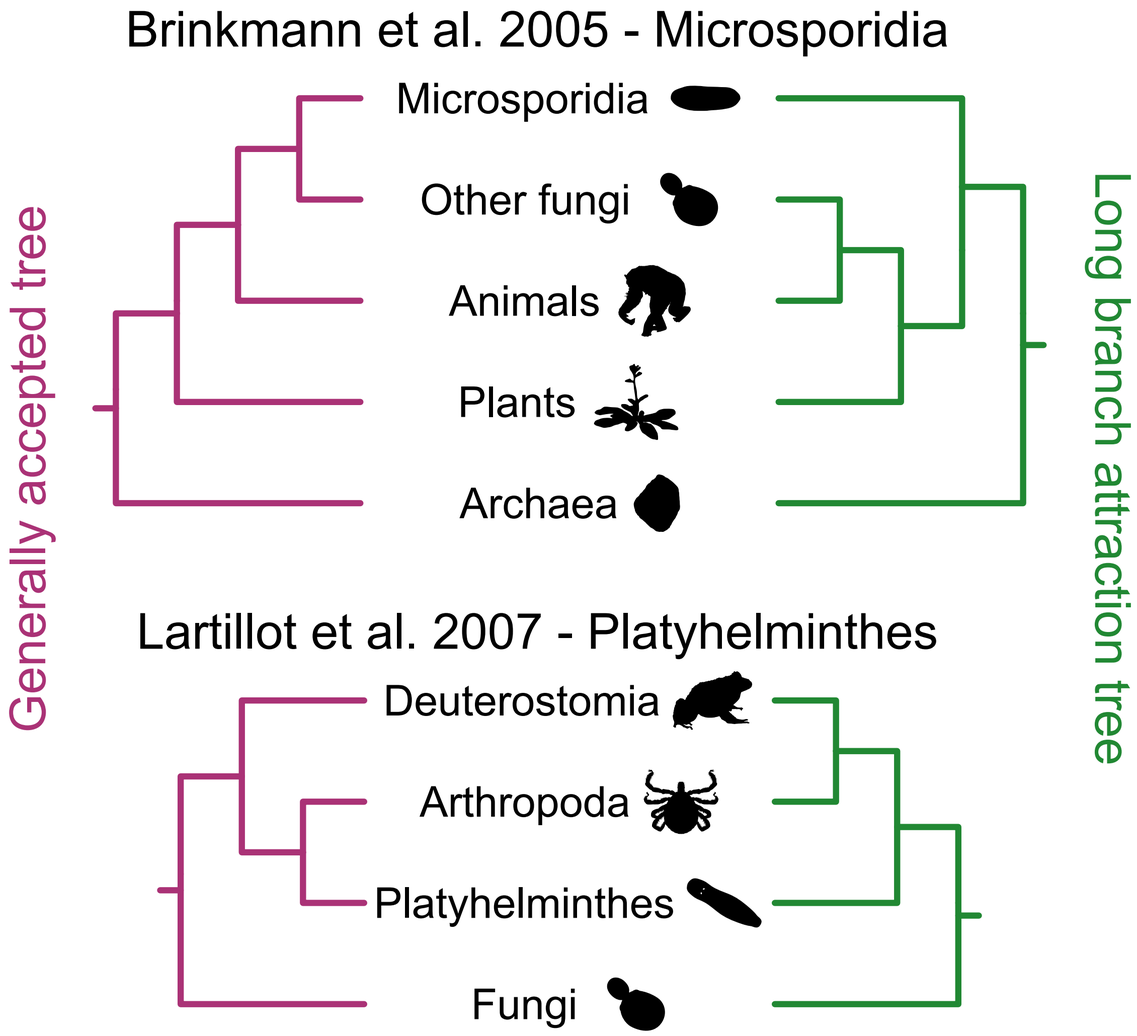
Platyhelminthes are sister to arthropods



LBA can lead to an erroneous tree

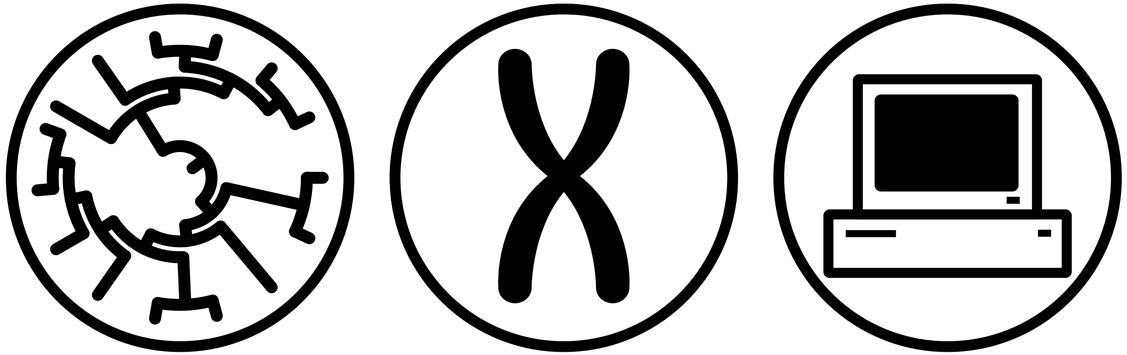


Site-heterogeneous models can overcome LBA



@JLSteenwyk

Outline



- Major methods in phylogenomics
- Substitution models, in (very) brief
- **Methods to concatenate sequences**
- Phylogenomic subsampling

Methods to concatenate sequences

Methods to concatenate sequences

Manual

- That is, by hand

Methods to concatenate sequences

Manual

- That is, by hand….*but why???*

Methods to concatenate sequences

Manual

- That is, by hand….*but why???*

GUI (Graphical User Interface)

- SequenceMatrix

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1096-0031.2010.00329.x

- CONCATENATOR

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1755-0998.2008.02164.x

Methods to concatenate sequences

Manual

- That is, by hand….*but why???*

GUI (Graphical User Interface)

- SequenceMatrix

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1096-0031.2010.00329.x

- CONCATENATOR

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1755-0998.2008.02164.x

Command-line

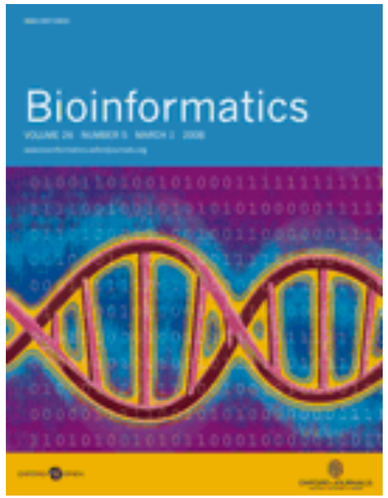
- PhyKIT

https://jlsteenwyk.com/PhyKIT/

- FASconCAT-G

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4243772/

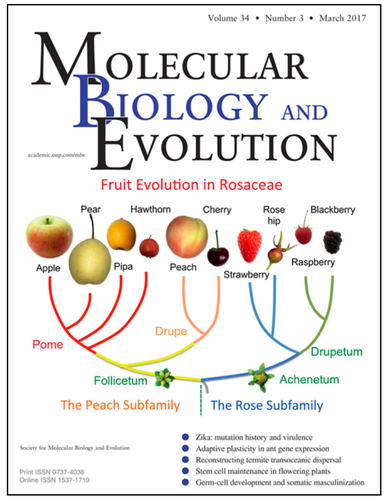
Two main programs used in the practical



PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data FREE

Jacob L Steenwyk ✉, Thomas J Buida, III, Abigail L Labella, Yuanning Li, Xing-Xing Shen, Antonis Rokas ✉

Bioinformatics, Volume 37, Issue 16, August 2021, Pages 2325–2331,
https://doi.org/10.1093/bioinformatics/btab096
Published: 09 February 2021 **Article history** ▾

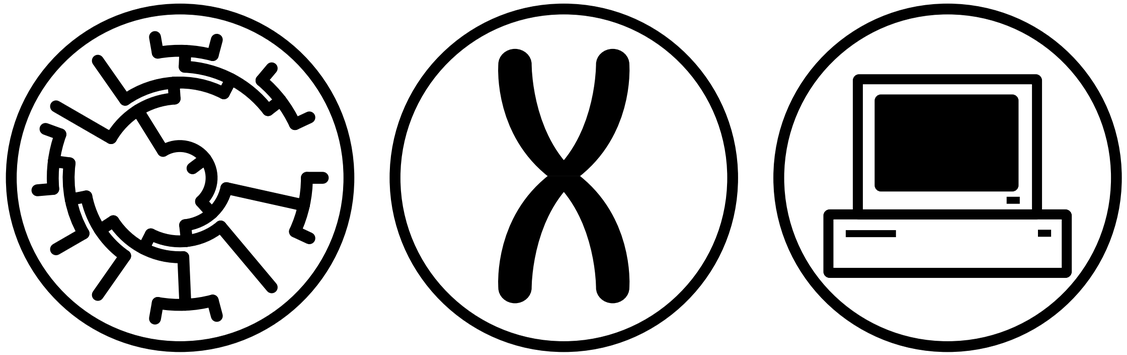


IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era

Bui Quang Minh ✉, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, Robert Lanfear Author Notes

Molecular Biology and Evolution, Volume 37, Issue 5, May 2020, Pages 1530–1534,
https://doi.org/10.1093/molbev/msaa015
Published: 03 February 2020

Outline



- Major methods in phylogenomics
- Substitution models, in (very) brief
- Methods to concatenate sequences
- **Phylogenomic subsampling**

Phylogenomic subsampling, in brief

Phylogenomic subsampling, in brief

1. Unstable bipartitions will be sensitive to gene selection

Phylogenomic subsampling, in brief

1. Unstable bipartitions will be sensitive to gene selection

2. Subsample the full data matrix and reinfer the species tree using fewer (but typically still several dozen to hundreds of genes)

Phylogenomic subsampling, in brief

1. Unstable bipartitions will be sensitive to gene selection

2. Subsample the full data matrix and reinfer the species tree using fewer (but typically still several dozen to hundreds of genes)

3. Compare resulting phylogenies and determine which bipartition are unstable

Phylogenomic subsampling, in brief

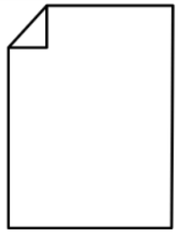
1. Unstable bipartitions will be sensitive to gene selection

2. Subsample the full data matrix and reinfer the species tree using fewer (but typically still several dozen to hundreds of genes)

3. Compare resulting phylogenies and determine which bipartition are unstable

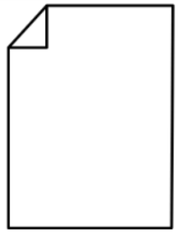
4. Examine potential drivers of incongruence thereafter. Incongruence will be examined in a later lab

Phylogenetic subsampling



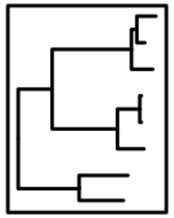
Complete
phylogenomic
data matrix

Phylogenetic subsampling



Complete phylogenomic data matrix

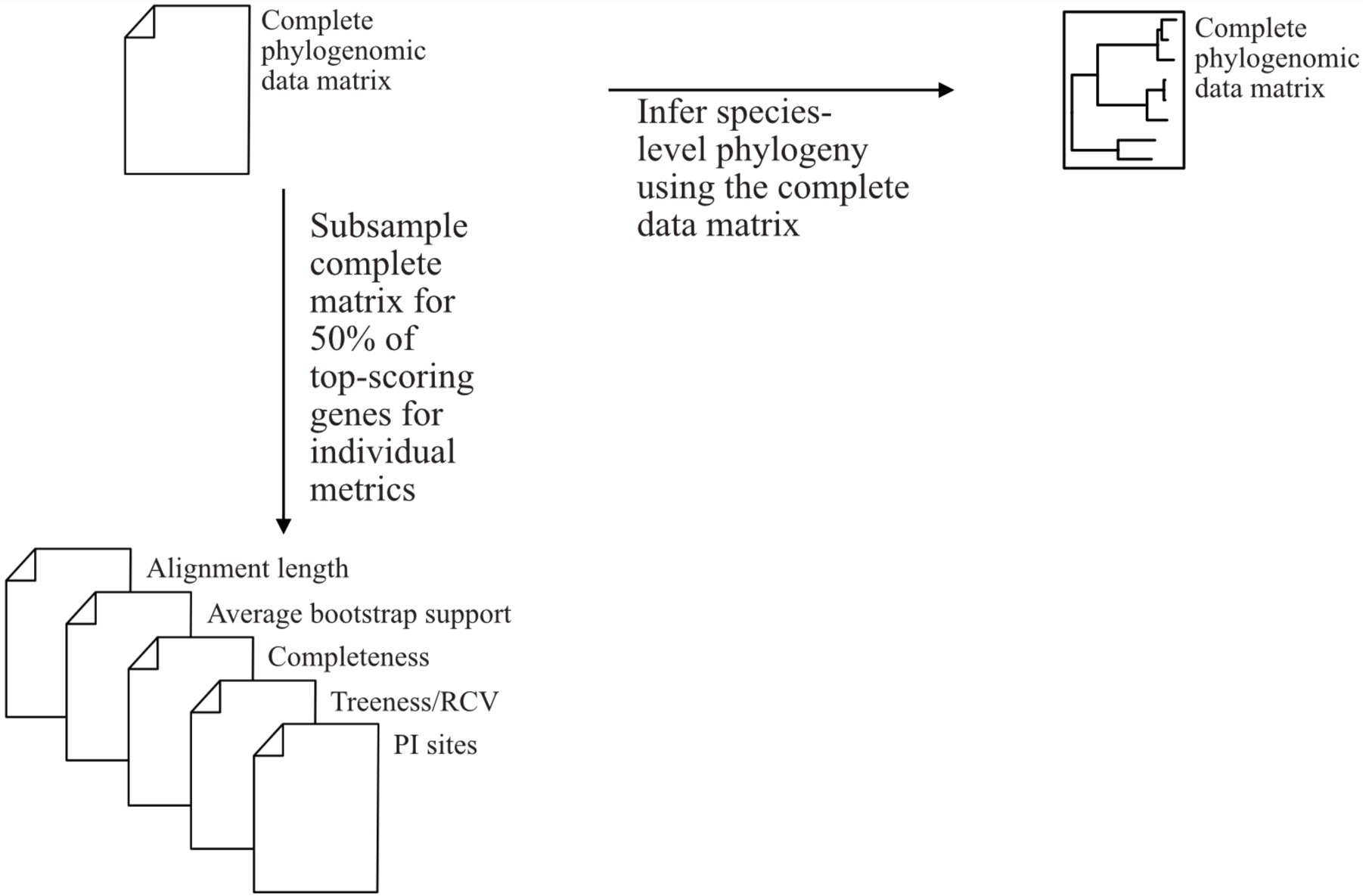
Infer species-level phylogeny using the complete data matrix



Complete phylogenomic data matrix

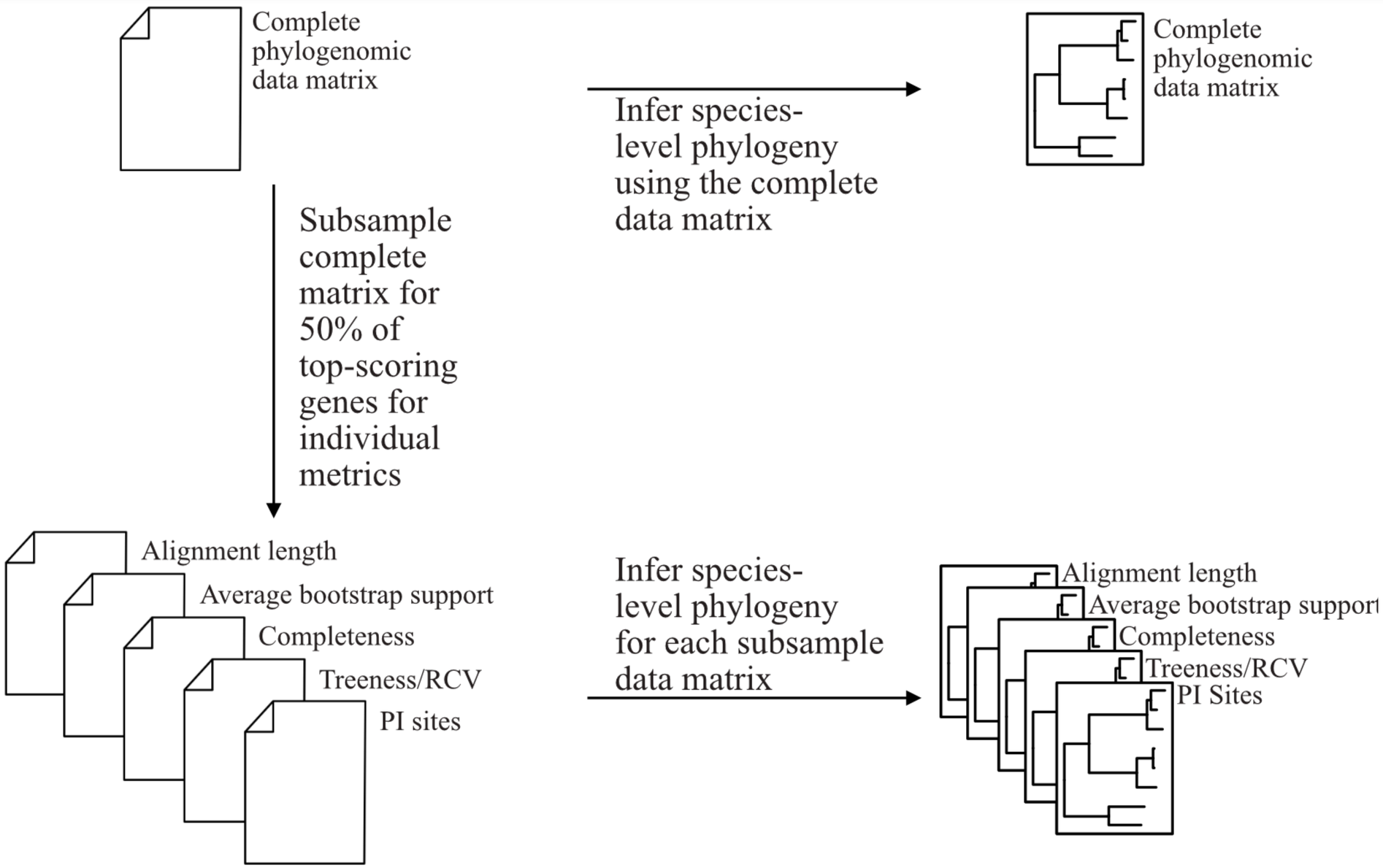
@JLSteenwyk

Phylogenetic subsampling



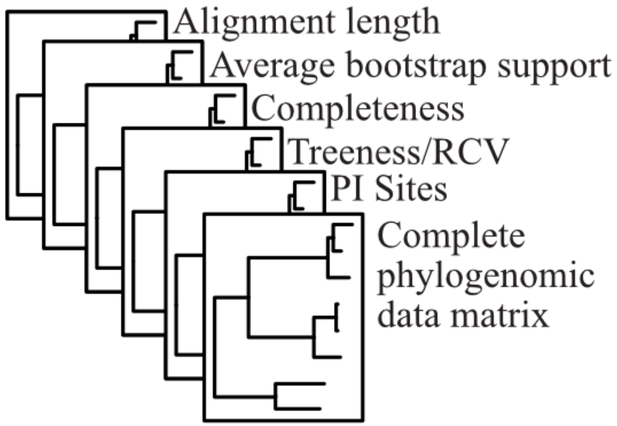
@JLSteenwyk

Phylogenetic subsampling

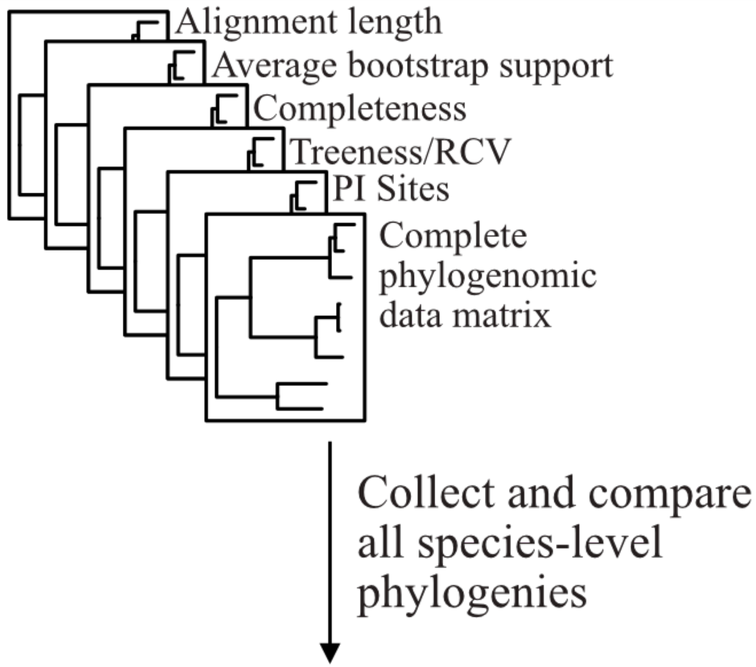


@JLSteenwyk

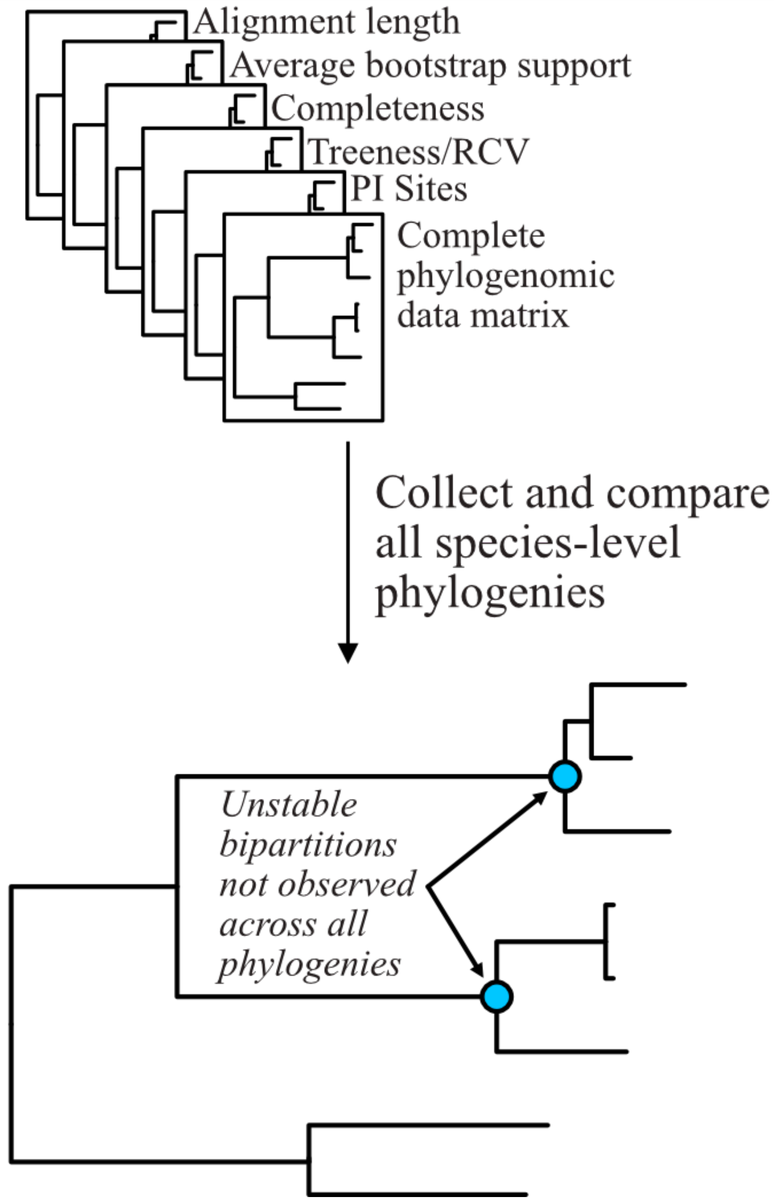
Phylogenetic subsampling



Phylogenetic subsampling



Phylogenetic subsampling



Metrics that capture phylogenetic signal

1. Alignment length
2. Alignment length with no gaps
3. GC content (for NTs)
4. Pairwise identity
5. # of parsimony informative sites
6. # of variable sites
7. Relative composition variability
8. Average bootstrap support value
9. Degree of violation of a molecular clock
10. Evolutionary rate
11. Long branch score
12. Treeness
13. Saturation
14. Treeness / RCV

@JLSteenwyk

Metrics that capture phylogenetic signal

1. **Alignment length**
2. **Alignment length with no gaps**
3. GC content (for NTs)
4. **Pairwise identity**
5. # of parsimony informative sites
6. # of variable sites
7. **Relative composition variability**
8. Average bootstrap support value
9. **Degree of violation of a molecular clock**
10. Evolutionary rate
11. Long branch score
12. Treeness
13. **Saturation**
14. Treeness / RCV

@JLSteenwyk

Alignment length

```
>sp1
ACGTAGCG-TCGATC
>sp2
ACGT-GCGATCGATC
>sp3
ACGTAGC-ATCGATC
>sp4
ACGTAGCGATCGATG
>sp5
AC--AGCGATCGATC
>sp6
ACGTAGCGA---ATC
```

The length of this alignment is 15 sites

Alignment length

```
>sp1
ACGTAGCG-TCGATC
>sp2
ACGT-GCGATCGATC
>sp3
ACGTAGC-ATCGATC
>sp4
ACGTAGCGATCGATG
>sp5
AC--AGCGATCGATC
>sp6
ACGTAGCGA---ATC
```

Higher values are better!

The length of this alignment is 15 sites

Alignment length, no gaps

```
>sp1
ACGTAGCG-TCGATC
>sp2
ACGT-GCGATCGATC
>sp3
ACGTAGC-ATCGATC
>sp4
ACGTAGCGATCGATG
>sp5
AC--AGCGATCGATC
>sp6
ACGTAGCGA---ATC
```

Excluding sites with gaps, the length of this alignment is 7 sites

Alignment length, no gaps

```
>sp1
ACGTAGCG-TCGATC
>sp2
ACGT-GCGATCGATC
>sp3
ACGTAGC-ATCGATC
>sp4
ACGTAGCGATCGATG
>sp5
AC--AGCGATCGATC
>sp6
ACGTAGCGA---ATC
```

Higher values are better!

Excluding sites with gaps, the length of this alignment is 7 sites

Pairwise identity

```
>sp1
ACGTAGCG-TCGATC
>sp2
ACGT-GCGATCGATC
>sp3
ACGTAGC-ATCGATC
>sp4
ACGTAGCGATCGATG
>sp5
AC--AGCGATCGATC
>sp6
ACGTAGCGA---ATC
```

Pairwise identity

```
>sp1
ACGTAGCG-TCGATC
>sp2
ACGT-GCGATCGATC
>sp3
ACGTAGC-ATCGATC
>sp4
ACGTAGCGATCGATG
>sp5
AC--AGCGATCGATC
>sp6
ACGTAGCGA---ATC
```

```
>sp1
ACGTAGCG-TCGATC
>sp2
ACGT-GCGATCGATC
```

13/15 sites are identical between sp1 and sp2. Thus, sp1 and sp2 have a pairwise identity of 0.8667

Pairwise identity

```
>sp1
ACGTAGCG-TCGATC
>sp2
ACGT-GCGATCGATC
>sp3
ACGTAGC-ATCGATC
>sp4
ACGTAGCGATCGATG
>sp5
AC--AGCGATCGATC
>sp6
ACGTAGCGA---ATC
```

→

```
>sp1
ACGTAGCG-TCGATC
>sp2
ACGT-GCGATCGATC
```

13/15 sites are identical between sp1 and sp2. Thus, sp1 and sp2 have a pairwise identity of 0.8667

→

repeat for others and report summary statistics or the pairwise identity of each combination (verbose option)

Relative composition variability

- Average variability in the sequence composition among taxa in an MSA

Relative composition variability

- Average variability in the sequence composition among taxa in an MSA
- Evaluates potential composition biases
    - violate assumptions of site composition homogeneity in standard substitution models

Relative composition variability

- Average variability in the sequence composition among taxa in an MSA
- Evaluates potential composition biases
    - violate assumptions of site composition homogeneity in standard substitution models

$$\sum_{i=1}^{c} \sum_{j=1}^{n} \frac{\left| c_{ij} - \overline{c_i} \right|}{s \times n}$$

Relative composition variability

- Average variability in the sequence composition among taxa in an MSA
- Evaluates potential composition biases
    - violate assumptions of site composition homogeneity in standard substitution models

$$\sum_{i=1}^{c} \sum_{j=1}^{n} \frac{\left| c_{ij} - \overline{c_i} \right|}{s \times n}$$

- *c* is the number of different character states per sequence type
- *n* is the number of taxa in an MSA
- *s* is the number of sites in an MSA

@JLSteenwyk

Relative composition variability

High compositional bias

```
>sp1
AAATTTT
>sp2
AAA-TTA
>sp3
AAATTTT
>sp4
AAATTAT
>sp5
AAATTTT
```

Relative composition variability

High compositional bias

```
>sp1
AAATTTT
>sp2
AAA-TTA
>sp3
AAATTTT
>sp4
AAATTAT
>sp5
AAATTTT
```

RCV = 0.2171

Relative composition variability

High compositional bias

```
>sp1
AAATTTT
>sp2
AAA-TTA
>sp3
AAATTTT
>sp4
AAATTAT
>sp5
AAATTTT
```

RCV = 0.2171

Lower compositional bias

```
>sp1
ACATTGG
>sp2
ACA-TGG
>sp3
ACATTGG
>sp4
ACATTGG
>sp5
ACATTGG
```

@JLSteenwyk

Relative composition variability

High compositional bias

```
>sp1
AAATTTT
>sp2
AAA-TTA
>sp3
AAATTTT
>sp4
AAATTAT
>sp5
AAATTTT
```

RCV = 0.2171

Lower compositional bias

```
>sp1
ACATTGG
>sp2
ACA-TGG
>sp3
ACATTGG
>sp4
ACATTGG
>sp5
ACATTGG
```

RCV = 0.0914

@JLSteenwyk

Relative composition variability

High compositional bias

```
>sp1
AAATTTT
>sp2
AAA-TTA
>sp3
AAATTTT
>sp4
AAATTAT
>sp5
AAATTTT
```

RCV = 0.2171

Lower RCV

values are better

Lower compositional bias

```
>sp1
ACATTGG
>sp2
ACA-TGG
>sp3
ACATTGG
>sp4
ACATTGG
>sp5
ACATTGG
```

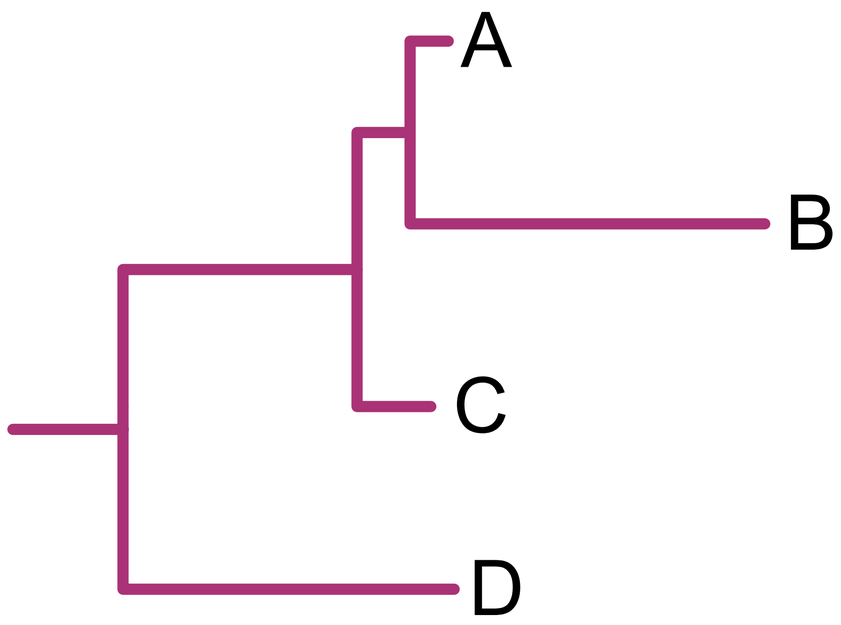
RCV = 0.0914

@JLSteenwyk

Degree of violation of a molecular clock

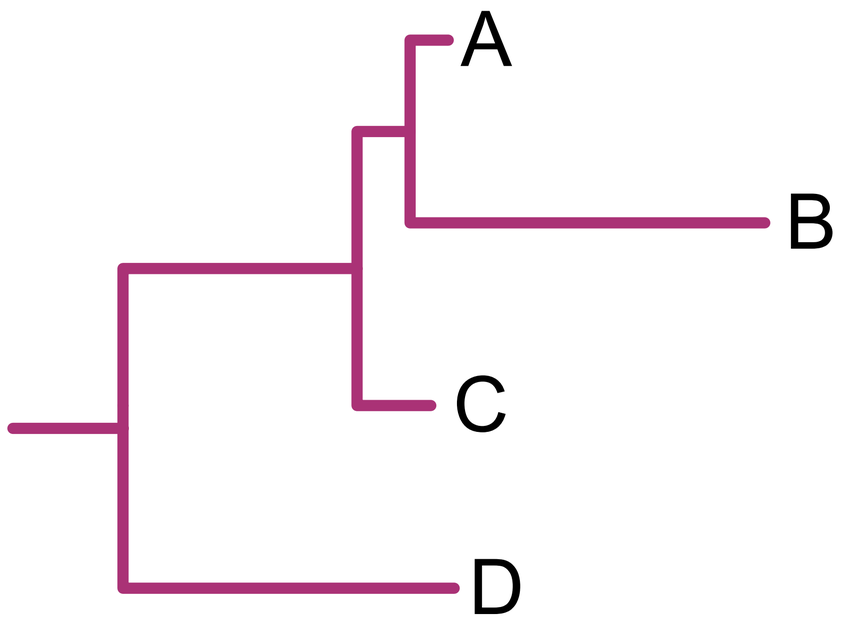
Degree of violation of a molecular clock

Non-clock like



Degree of violation of a molecular clock

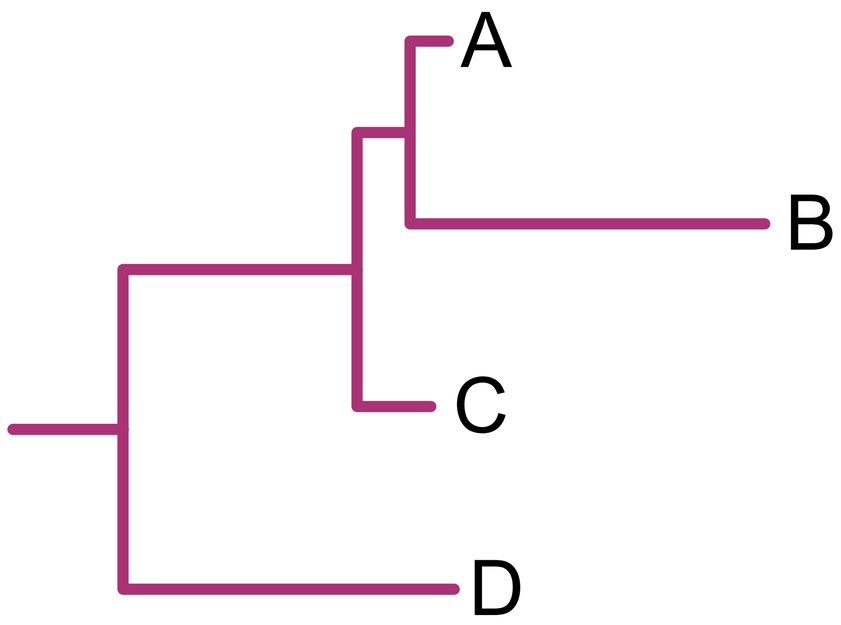
Non-clock like



High DVMC

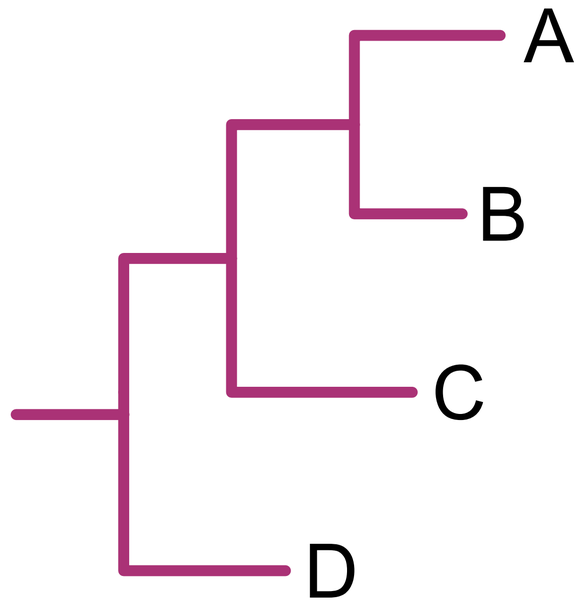
Degree of violation of a molecular clock

Non-clock like



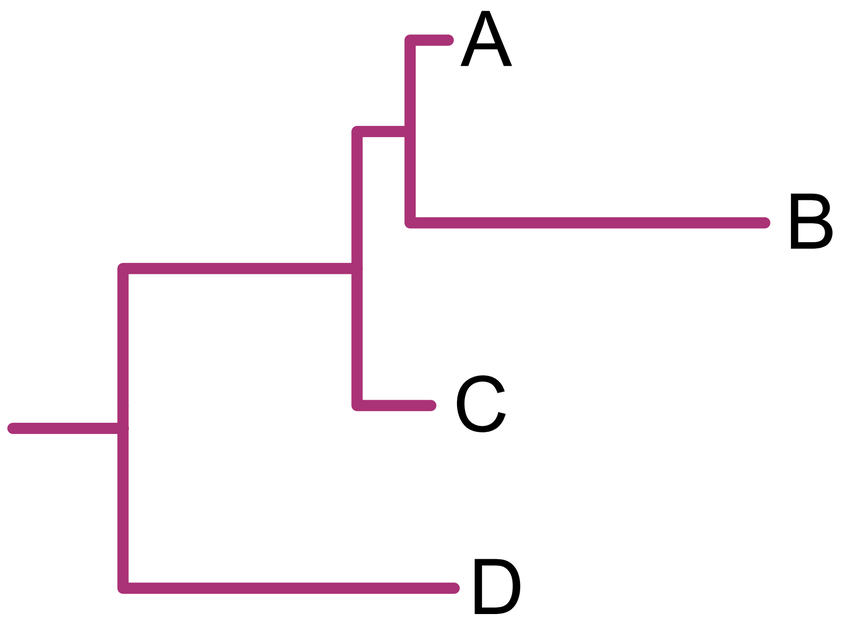
High DVMC

Clock-like



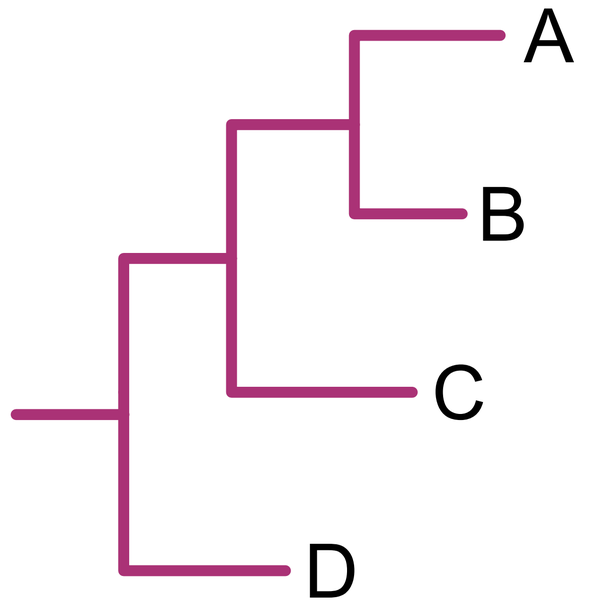
Degree of violation of a molecular clock

Non-clock like



High DVMC

Clock-like

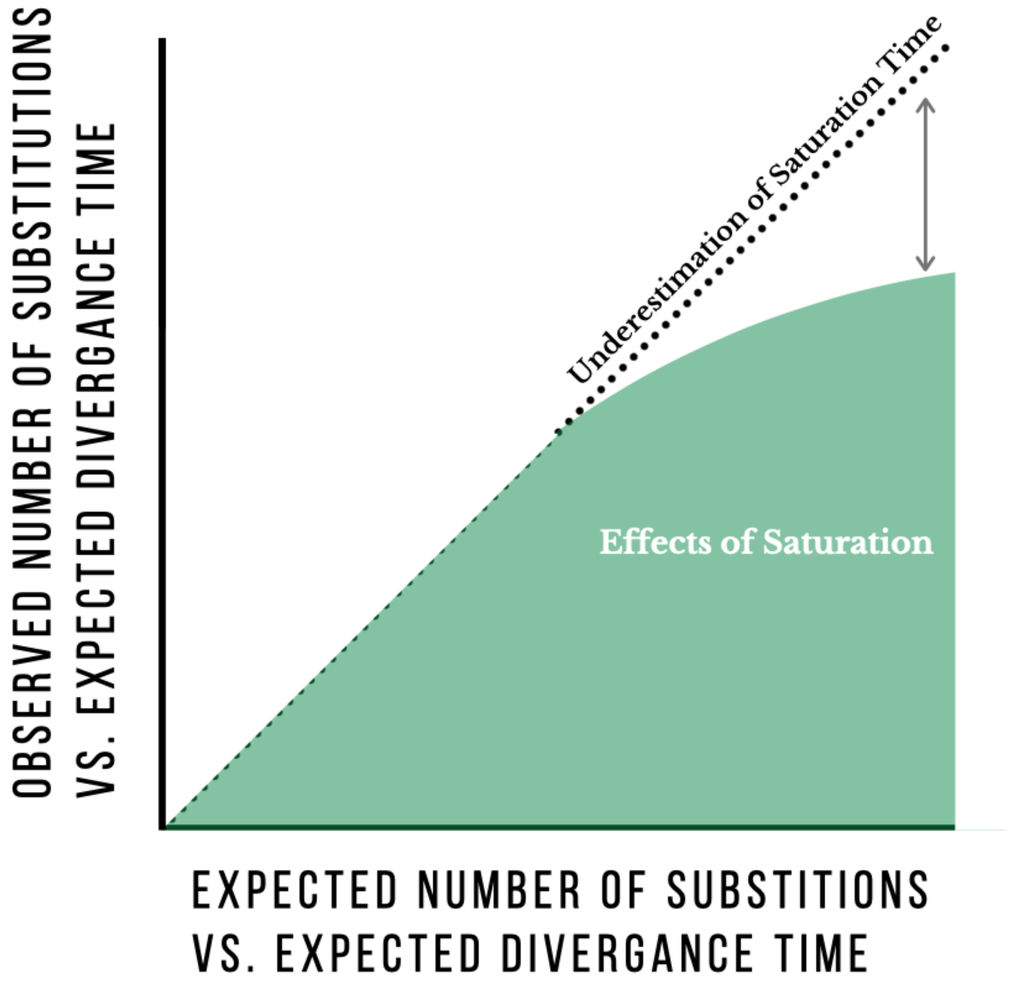


Low DVMC

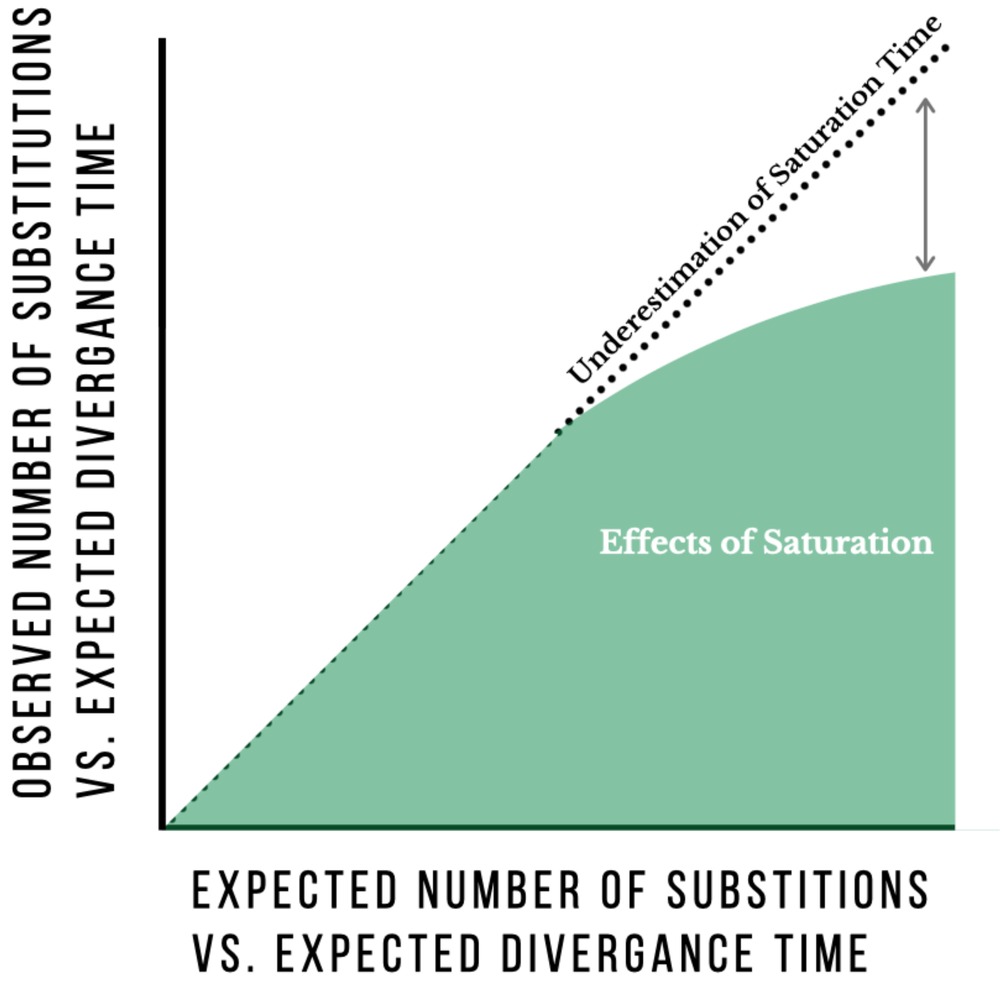
Degree of violation of a molecular clock

Genes with low DVMC may be
useful for divergence time analysis

Saturation by multiple substitutions

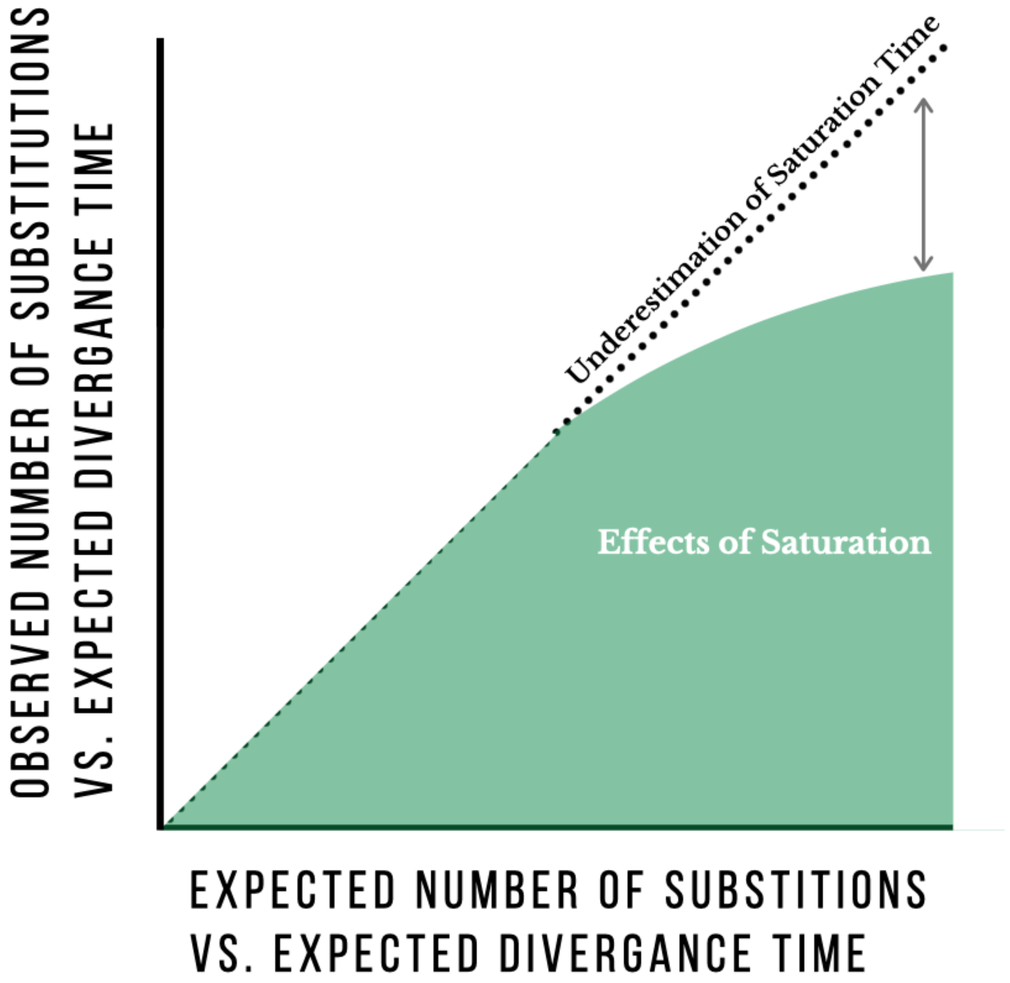


Saturation by multiple substitutions



- X-axis can be approximated using phylogenetic distances
  - Tip-to-tip distances in a tree

Saturation by multiple substitutions



- X-axis can be approximated using phylogenetic distances
  - Tip-to-tip distances in a tree
- Y-axis can be approximated using pairwise identity
  - Distance in an MSA

So many metrics, so many details

1. Alignment length - **higher better**
2. Alignment length with no gaps - **higher better**
3. GC content (for NTs) - **lower better**
4. Pairwise identity - **depends**
5. # of parsimony informative sites - **higher better**
6. # of variable sites - **higher better**
7. Relative composition variability - **lower better**
8. Average bootstrap support value - **higher better**
9. Degree of violation of a molecular clock - **lower better**
10. Evolutionary rate - **depends**
11. Long branch score - **lower better**
12. Treeness - **higher better**
13. Saturation - **higher better**
14. Treeness / RCV - **higher better**

@JLSteenwyk

Where known, PhyKIT documentation will say

PhyKIT

Search docs

- About
- Usage
  - General usage
  - Alignment-based functions
  - Tree-based functions
  - **Alignment- and tree-based functions**
    - Saturation
    - Treeness over RCV
- Tutorials
- Change log
- Other software
- FAQ

https://jlsteenwyk.com/PhyKIT

Saturation

Function names: saturation; sat
Command line interface: pk_saturation; pk_sat

Calculate saturation for a given tree and alignment.

Saturation is defined as sequences in multiple sequence alignments that have undergone numerous substitutions such that the distances between taxa are underestimated.

Data with no saturation will have a value of 1. Completely saturated data will have a value of 0.

Saturation is calculated following Philippe et al., PLoS Biology (2011), doi: 10.1371/journal.pbio.1000602.

```
phykit saturation -a <alignment> -t <tree> [-v/--verbose]
```

Options:
-a/--alignment: an alignment file
-t/--tree: a tree file
-v/--verbose: print out patristic distances and uncorrected distances used to determine saturation

@JLSteenwyk

Concatenation and partitioning

@JLSteenwyk https://jlsteenwyk.com/