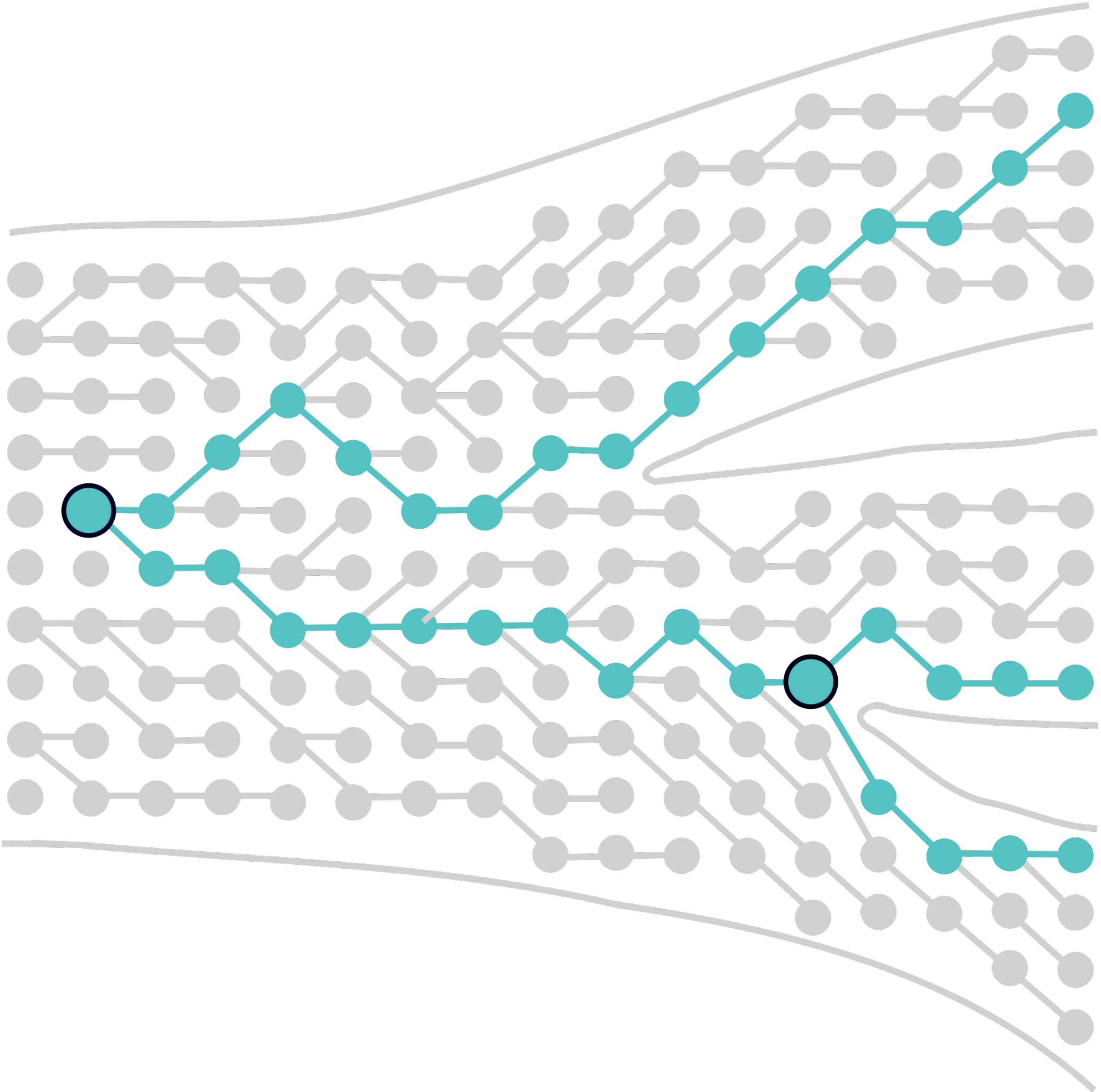
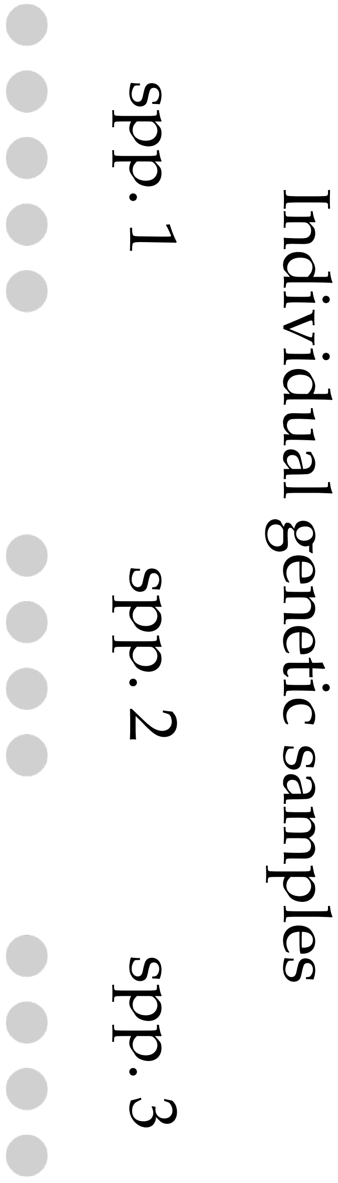
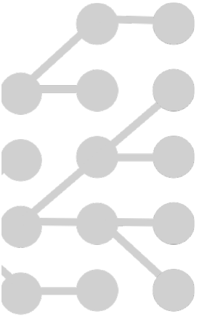


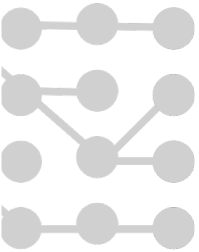
RADseq phylogenetics

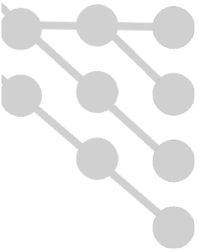
How do we actually build a phylogeny?



How do we actually build a phylogeny?



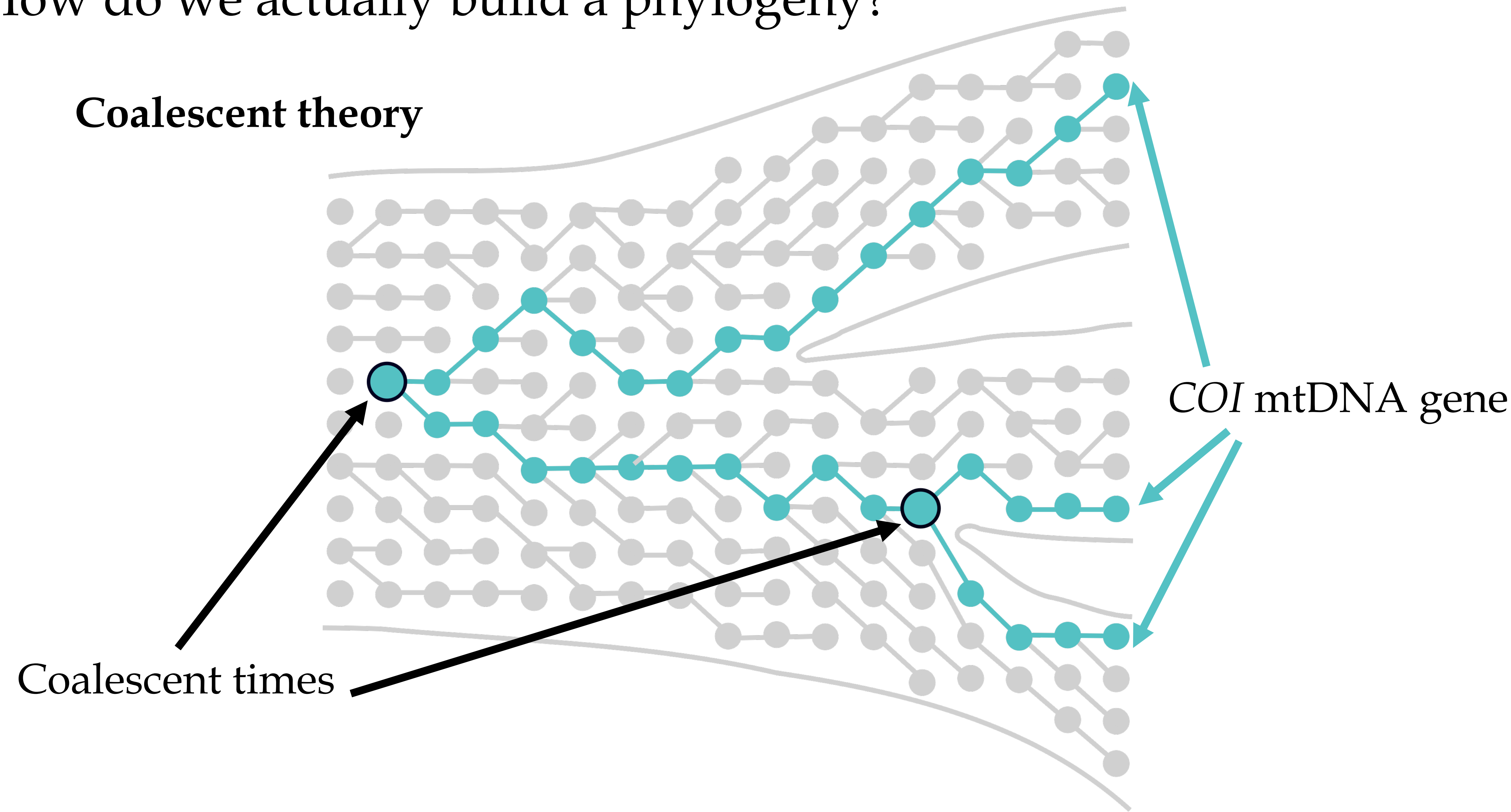




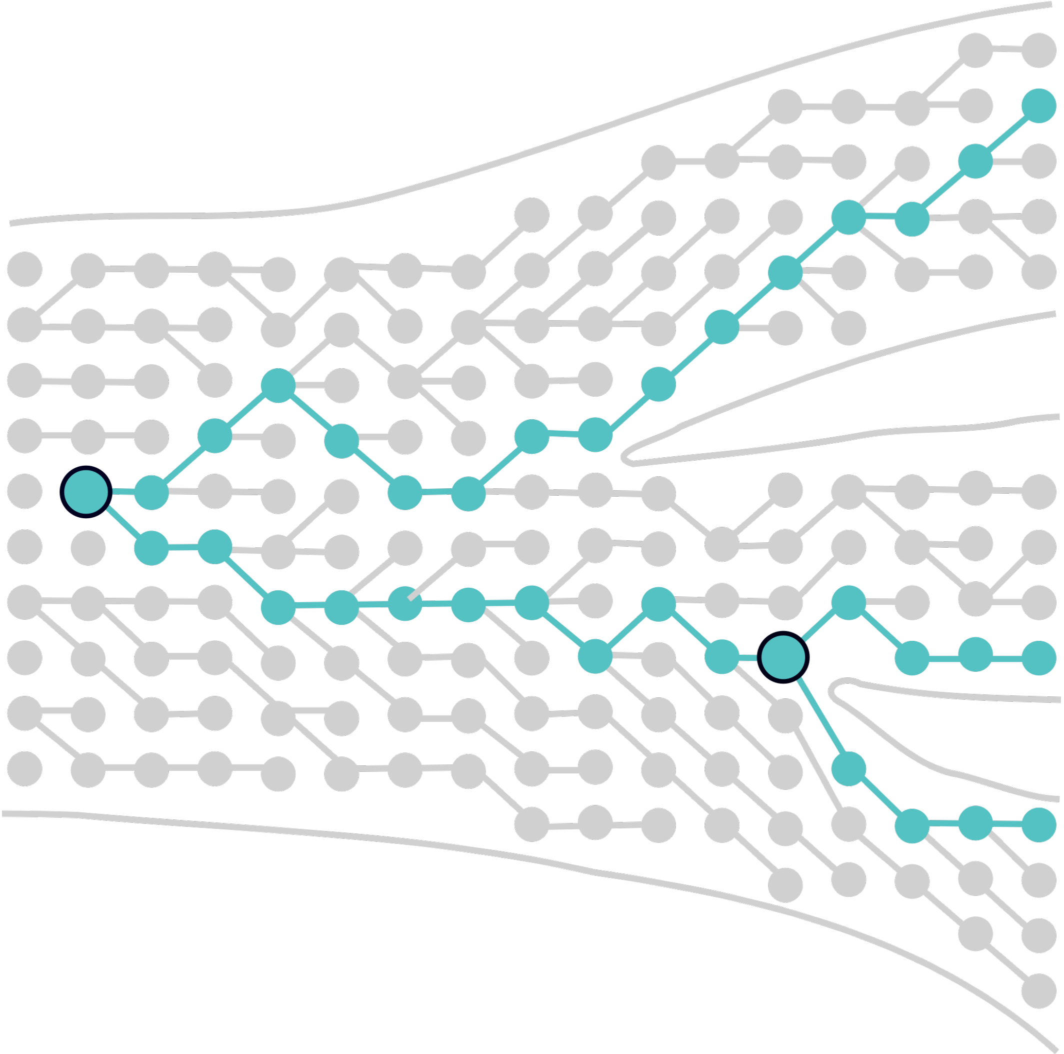
How do we actually build a phylogeny?

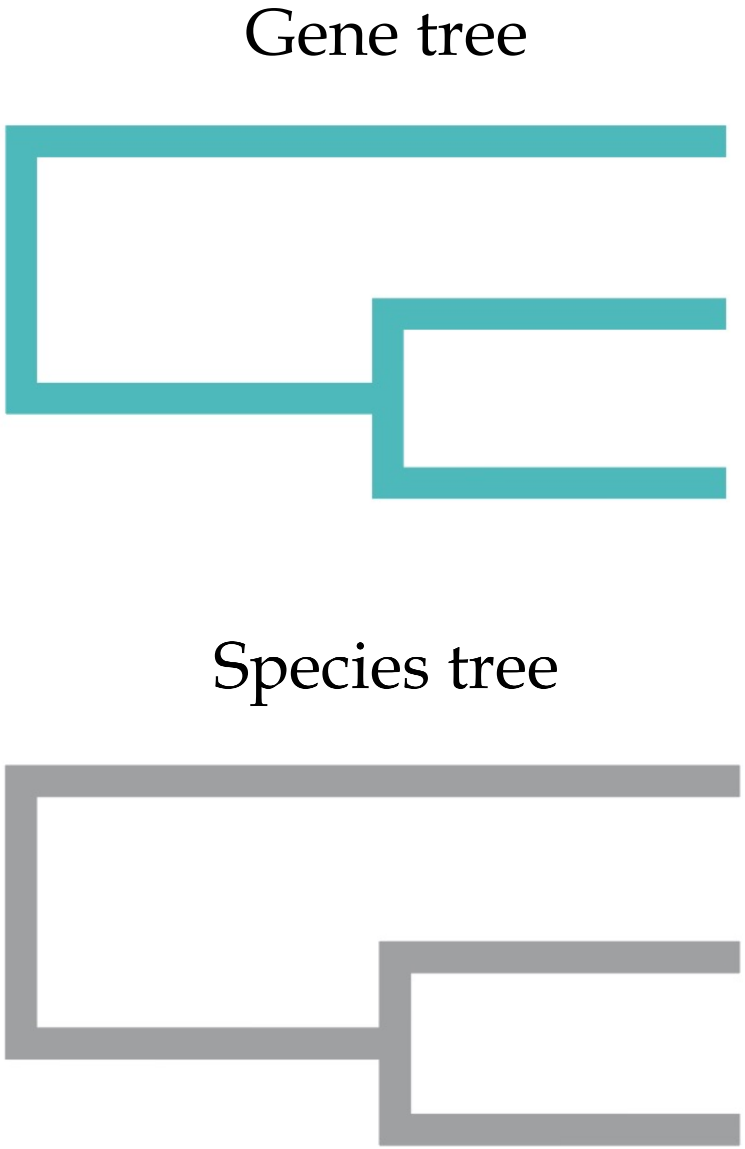
How do we actually build a phylogeny?

Coalescent theory



How do we actually build a phylogeny?





Types of trees

- **Gene trees** – phylogeny depicting the evolutionary history of a gene or gene family in a specific group of organisms
- **Species trees** – phylogeny depicting the evolutionary history of a species or group of species (should use multiple genes for a best estimate)
- **Consensus trees** – a summary technique for depicting statistical support of a single tree
- **Supertrees** – different phylogenies that are connected manually by nodes

From R. D. Tarvin

Tree reconstruction: **maximum likelihood**

- L = Pr (Data | Hypothesis)

- L = Pr (Alignment,Model of Evolution | Tree)

likelihood of a hypothesis is the probability of observing a set of data given a particular hypothesis

- In phylogenetics, the data is the sequence data; the hypothesis is the tree topology and the model of seq evolution

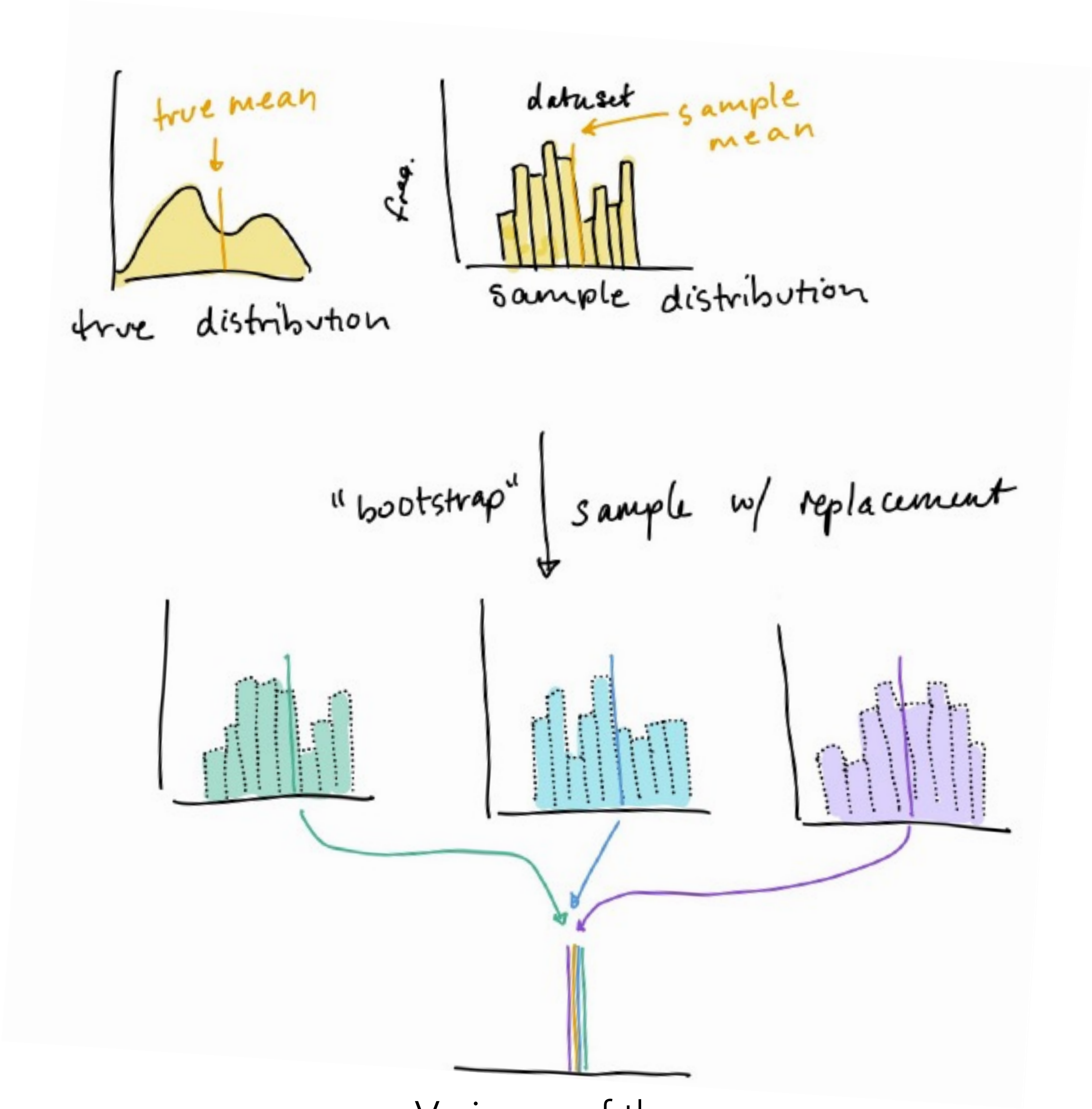
- Maximum likelihood is an optimality criterion that uses probabilistic models

- ML is a general statistical method of estimation, and not limited to phylogenetics

Slide: R. D. Tarvin

Assessing support

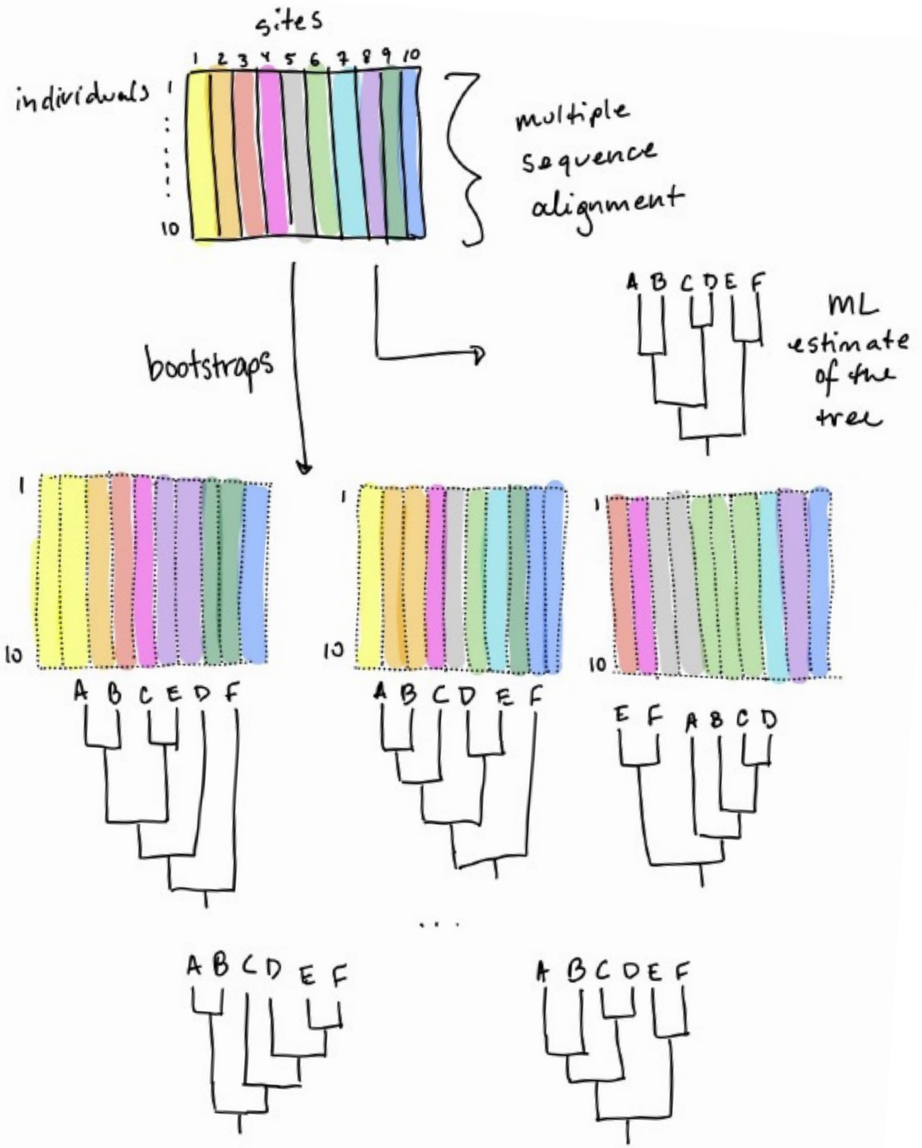
Bootstrapping is non-parametric sampling with replacement



Variance of the mean
(allows us to evaluate accuracy of our estimate)

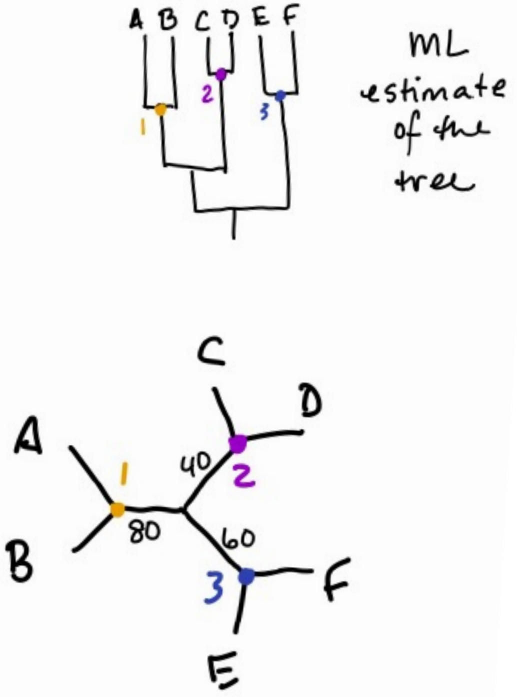
Slide: R. D. Tarvin

What does this look like for trees?



times each partition found (out of 5 bootstraps)

1: AB|CDEF	4
2: CD|ABEF	2
3: ABCD|EF	3



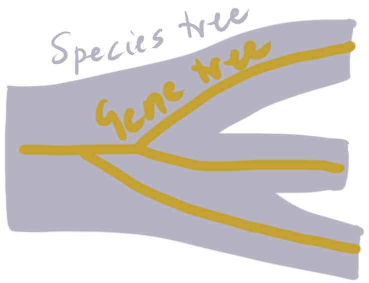
Slide: R. D. Tarvin

Assessing support: bootstrapping

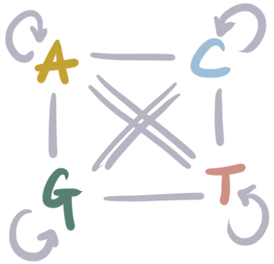
- If your data approximate a good sampling of the population, resampling with replacement should give you consistent tree estimates
- Bootstrap support values are assigned at each node: if the node appears in 100 of 100 bootstrap replicates, then the bootstrap support (BS) value is **100%.** If the node appears in 75 of 100, BS is 75%
    - 50–75% indicates poor support
    - 75–90%: good support
    - >90%: great support

Slide: R. D. Tarvin

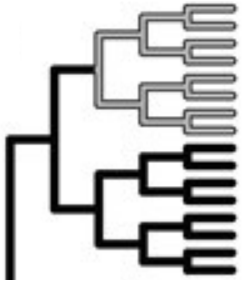
Things to take into account when estimating phylogenies



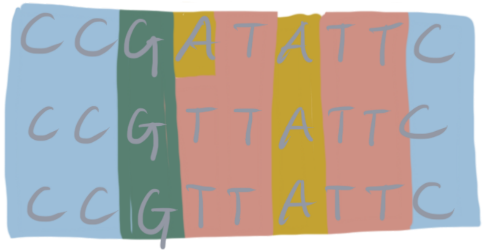
- Gene trees conflict with species trees (i.e., the genes you select may not be accurately representing interspecies relationships)



- Substitution models (modeling sequence evolution)
  - Linked sites, partitioning the data, estimating substitution models for each locus

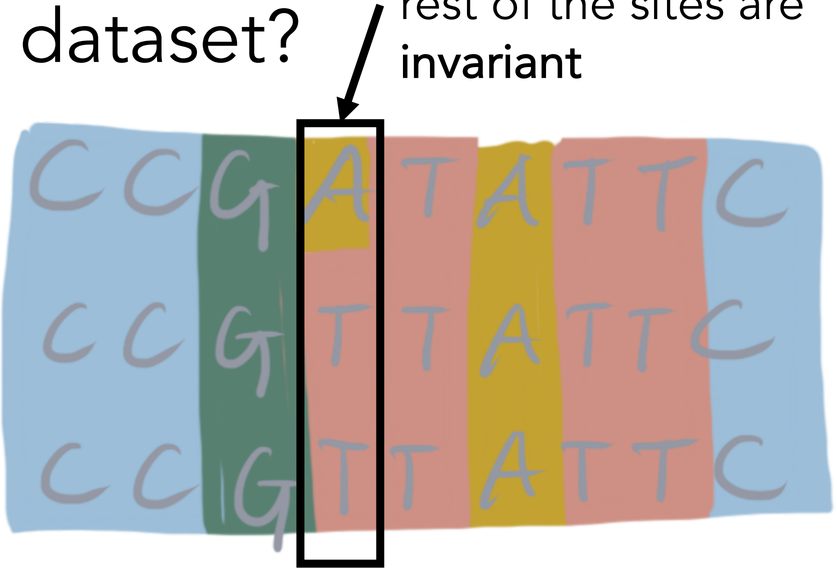


- Amounts, sources, and patterns of missing data
- Informative sites (invariant vs SNPs)



Phylogenetic information

- How much information is contained in our dataset?



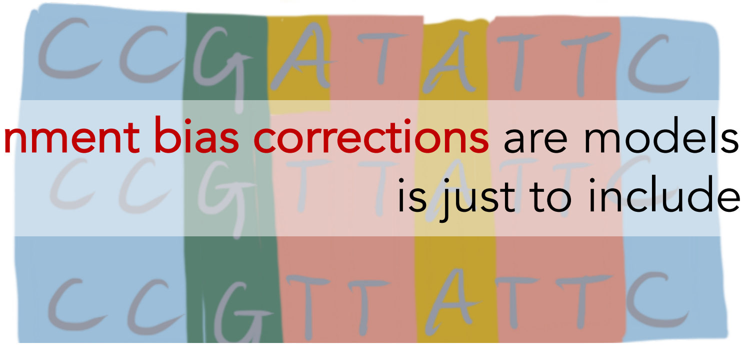
site 4 is a **SNP**, the rest of the sites are **invariant**

- Single nucleotide polymorphisms (SNPs)
- Parsimony-informative sites:
  - Sites with different numbers of steps on trees
  - Can be used to discriminate among alternative trees under the parsimony criterion
  - Two taxa much have one state and two others must have a different state differs from SNPs in this way)

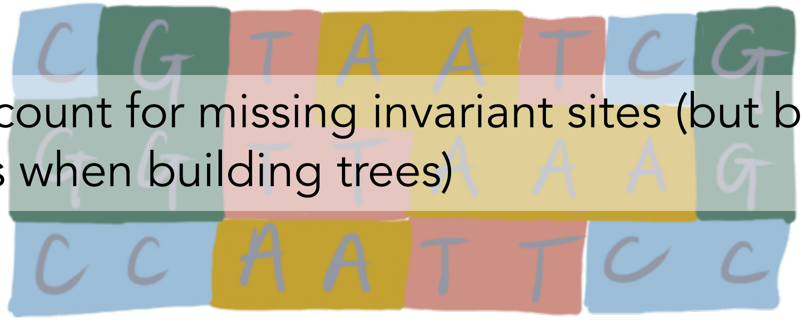
Phylogenetic information

- Which sites do we use to reconstruct phylogenies?
  - Variable sites (SNPs) can mislead phylogenetic inference

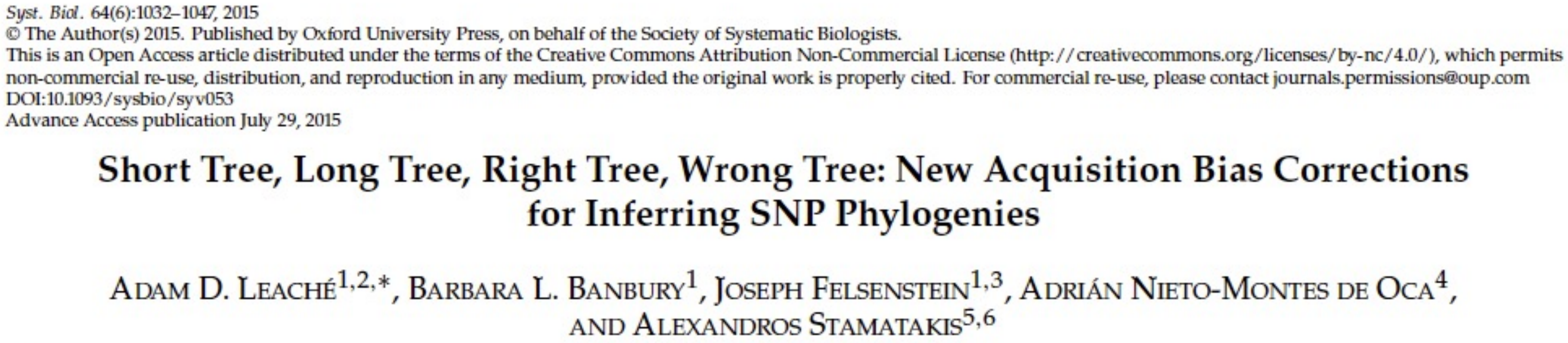
All sites (SNPs and invariant)



SNPs only



Ascertainment bias corrections are models that account for missing invariant sites (but best option is just to include all sites when building trees)



Syst. Biol. 64(6):1032–1047, 2015

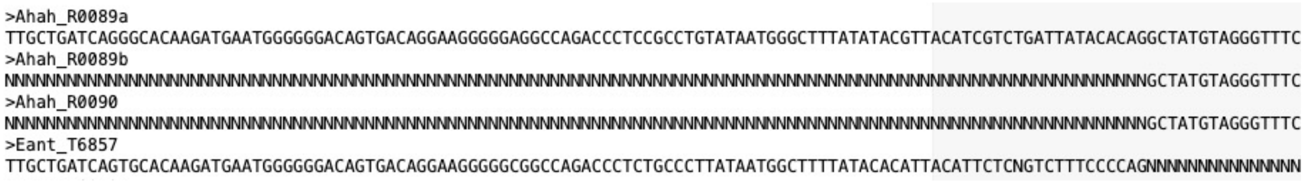
DOI:10.1093/sysbio/syv053
Advance Access publication July 29, 2015

Short Tree, Long Tree, Right Tree, Wrong Tree: New Acquisition Bias Corrections for Inferring SNP Phylogenies

Adam D. Leaché[1,2,*], Barbara L. Banbury[1], Joseph Felsenstein[1,3], Adrián Nieto-Montes de Oca[4], and Alexandros Stamatakis[5,6]

Input files used for standard phylogenetics

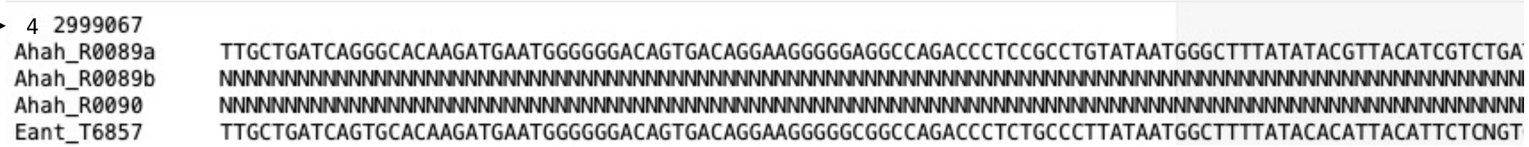
fasta file



```
>Ahah_R0089a
TTGCTGATCAGGGCACAAGATGAATGGGGGGACAGTGACAGGAAGGGGGAGGCCAGACCCTCCGCCTGTATAATGGGCTTTATATACGTTACATCGTCTGATTATACACAGGCTATGTAGGGTTTC
>Ahah_R0089b
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGCTATGTAGGGTTTC
>Ahah_R0090
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGCTATGTAGGGTTTC
>Eant_T6857
TTGCTGATCAGTGCACAAGATGAATGGGGGGACAGTGACAGGAAGGGGGCGGCCAGACCCTCTGCCCTTATAATGGCTTTTATACACATTACATTCTCNGTCTTTCCCCAGNNNNNNNNNNNNNNN
```

Phylip file

[# individuals] [# sites] →



```
 4 2999067
Ahah_R0089a      TTGCTGATCAGGGCACAAGATGAATGGGGGGACAGTGACAGGAAGGGGGAGGCCAGACCCTCCGCCTGTATAATGGGCTTTATATACGTTACATCGTCTGA
Ahah_R0089b      NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
Ahah_R0090       NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
Eant_T6857       TTGCTGATCAGTGCACAAGATGAATGGGGGGACAGTGACAGGAAGGGGGCGGCCAGACCCTCTGCCCTTATAATGGCTTTTATACACATTACATTCTCNGT
```

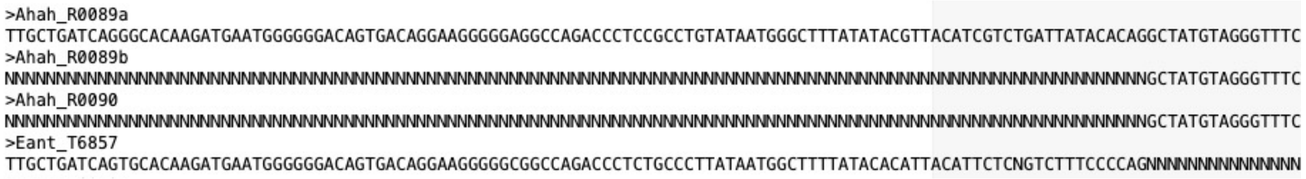
Nexus file

data “block” (same as phylip file format) →

```
#NEXUS

BEGIN DATA;
DIMENSIONS NTAX=4 NCHAR=50;
FORMAT DATATYPE=DNA GAP=- MISSING=?;
MATRIX

Eant_T6857      CGGTCCTGACGTGCAAATCGGTCGTCTGACCTGGTTCCACCTTGCTGATC
Etri_T6842      CGGTCCTGACGTGCAAATCGGTCGTCTGACCTGGTTCCACCTTGCTGATC
Eant_T6859a     CGGTCCTGACGTGCAAATCGGTCGTCTGACCTGGTTCCACCTTGCTGATC
Ebou_R0153      CGGTCCTGACGTGCAAATCGGTCGTCTGACCTGGTTCCACCTTGCTGATC
;

END;
```

Input files used for standard phylogenetics

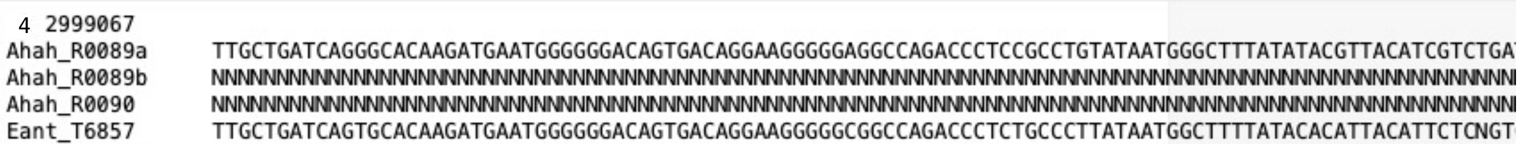
fasta file



```
>Ahah_R0089a
TTGCTGATCAGGGCACAAGATGAATGGGGGGACAGTGACAGGAAGGGGGAGGCCAGACCCTCCGCCTGTATAATGGGCTTTATATACGTTACATCGTCTGATTATACACAGGCTATGTAGGGTTTC
>Ahah_R0089b
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGCTATGTAGGGTTTC
>Ahah_R0090
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGCTATGTAGGGTTTC
>Eant_T6857
TTGCTGATCAGTGCACAAGATGAATGGGGGGACAGTGACAGGAAGGGGGCGGCCAGACCCTCTGCCCTTATAATGGCTTTTATACACATTACATTCTCNGTCTTTCCCCAGNNNNNNNNNNNNNNN
```

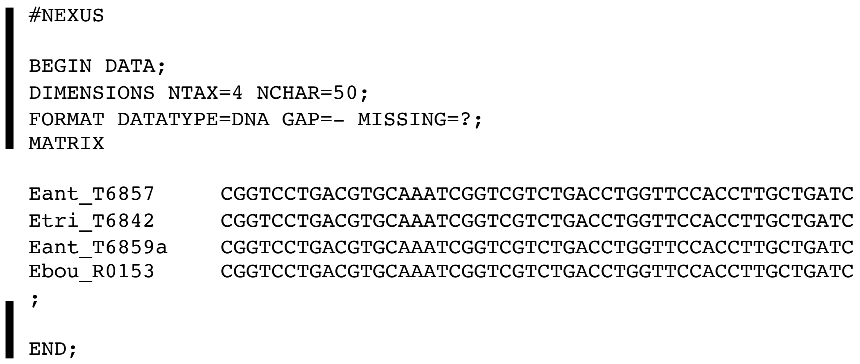
Phylip file

[# individuals] [# sites] →



```
 4 2999067
Ahah_R0089a        TTGCTGATCAGGGCACAAGATGAATGGGGGGACAGTGACAGGAAGGGGGAGGCCAGACCCTCCGCCTGTATAATGGGCTTTATATACGTTACATCGTCTGA
Ahah_R0089b        NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
Ahah_R0090         NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
Eant_T6857         TTGCTGATCAGTGCACAAGATGAATGGGGGGACAGTGACAGGAAGGGGGCGGCCAGACCCTCTGCCCTTATAATGGCTTTTATACACATTACATTCTCNGT
```

Nexus file

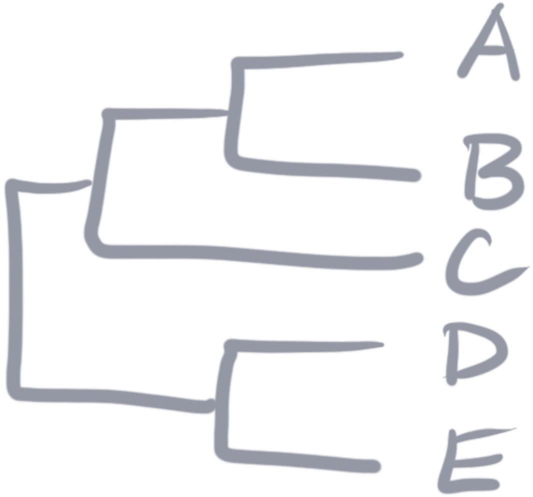


```
#NEXUS

BEGIN DATA;
DIMENSIONS NTAX=4 NCHAR=50;
FORMAT DATATYPE=DNA GAP=- MISSING=?;
MATRIX

Eant_T6857      CGGTCCTGACGTGCAAATCGGTCGTCTGACCTGGTTCCACCTTGCTGATC
Etri_T6842      CGGTCCTGACGTGCAAATCGGTCGTCTGACCTGGTTCCACCTTGCTGATC
Eant_T6859a     CGGTCCTGACGTGCAAATCGGTCGTCTGACCTGGTTCCACCTTGCTGATC
Ebou_R0153      CGGTCCTGACGTGCAAATCGGTCGTCTGACCTGGTTCCACCTTGCTGATC
;

END;
```

called a "block" because it's nested between these lines
(this is *absolutely* required for Nexus files)

Output files

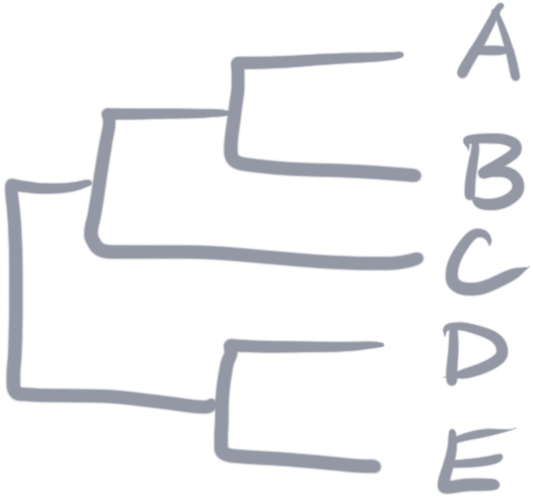
Brackets and commas are used to describe topology



(A,B)

Output files

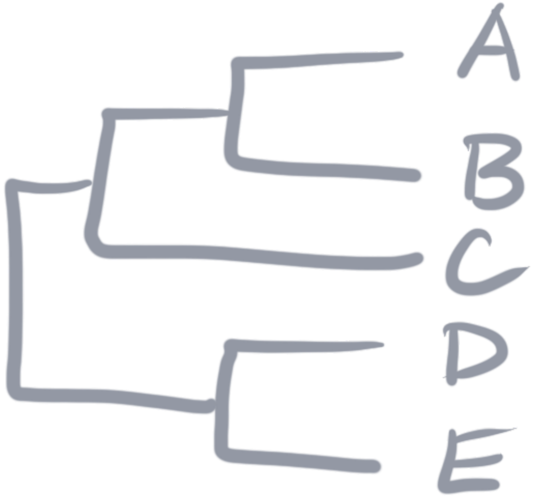
Brackets and commas are used to describe topology



((A,B),C)

Output files

Brackets and commas are used to describe topology

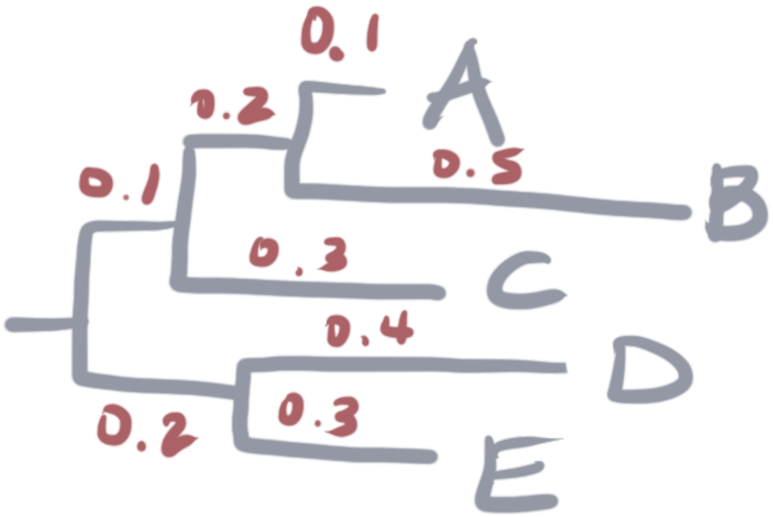


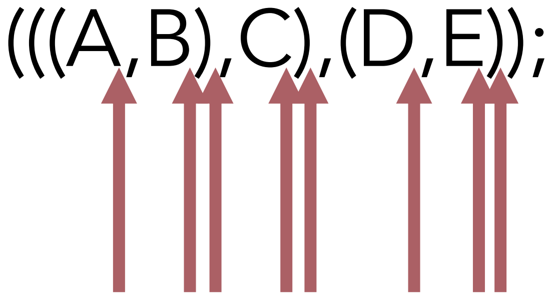
((A,B),C),(D,E);

Semicolon indicates the end of the tree

Output files

Incorporating **branch lengths** with colons

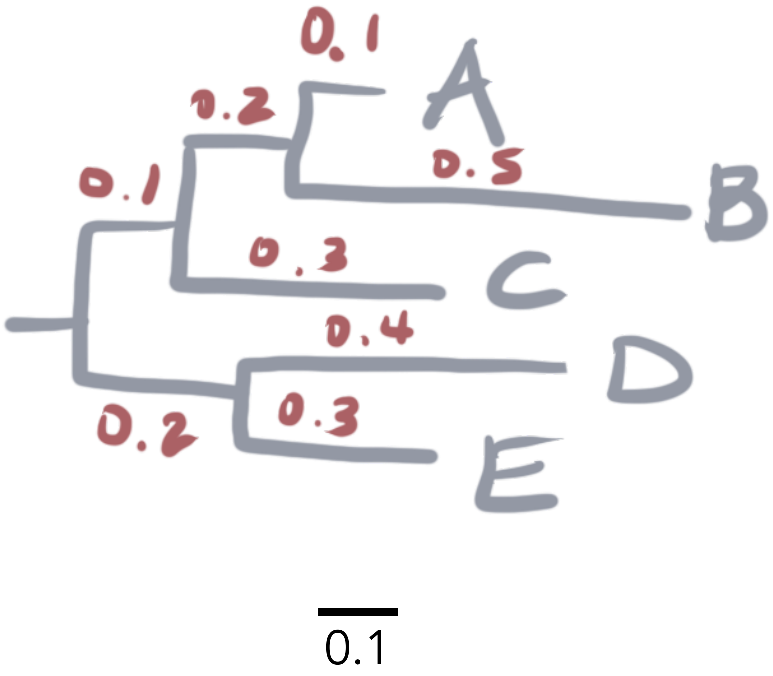




(((A,B),C),(D,E));

Output files

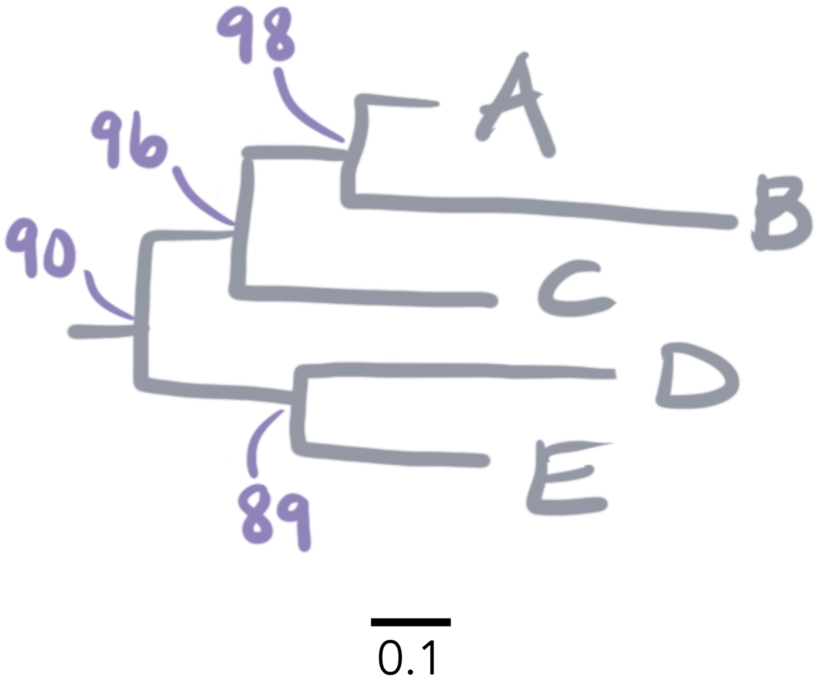
Incorporating **branch lengths** with colons



(((A:0.1,B:0.5):0.2,C:0.3):0.1,(D:0.4,E:0.3):0.2);

Output files

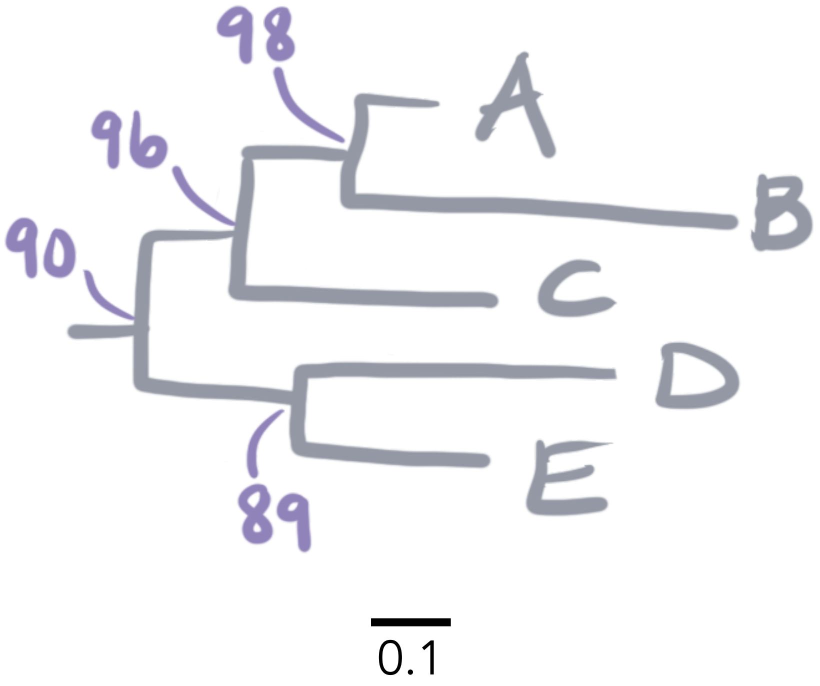
Incorporating **bootstrap support values** after any sets of brackets (nodes) and *before* BLs



(((A:0.1,B:0.5):0.2,C:0.3):0.1,(D:0.4,E:0.3):0.2);

Output files

Incorporating **bootstrap support values** after any sets of brackets (nodes) and *before* BLs



(((A:0.1,B:0.5)**98**:0.2,C:0.3)**96**:0.1,(D:0.4,E:0.3)**89**:0.2)**90**;

Tying everything together

Remember our Nexus file? We can have trees within Nexus files too (not just data)

Nexus file for sequence data

```
#NEXUS

BEGIN DATA;
DIMENSIONS NTAX=4 NCHAR=50;
FORMAT DATATYPE=DNA GAP=- MISSING=?;
MATRIX

Eant_T6857     CGGTCCTGACGTGCAAATCGGTCGTCTGACCTGGTTCCACCTTGCTGATC
Etri_T6842     CGGTCCTGACGTGCAAATCGGTCGTCTGACCTGGTTCCACCTTGCTGATC
Eant_T6859a    CGGTCCTGACGTGCAAATCGGTCGTCTGACCTGGTTCCACCTTGCTGATC
Ebou_R0153     CGGTCCTGACGTGCAAATCGGTCGTCTGACCTGGTTCCACCTTGCTGATC
;

END;
```

Nexus file for a tree (“tree block”)

```
#NEXUS

BEGIN TREES;
     TRANSLATE
       1 ‘Eant_T6857’,
       2 ‘
       3 ‘
       4 ‘
       ;
     tree epitree = [&r]((1:1.0E-6,2:2.0E-6)100:3.01E-4,(3:1.9E-5,4:2.3E-5)100:3.84E-4);
END;
```

Nexus file with data and a tree

```
#NEXUS

BEGIN DATA;
DIMENSIONS NTAX=4 NCHAR=50;
FORMAT DATATYPE=DNA GAP=- MISSING=?;
MATRIX

Eant_T6857     CGGTCCTGACGTGCAAATCGGTCGTCTGACCTGGTTCCACCTTGCTGATC
Etri_T6842     CGGTCCTGACGTGCAAATCGGTCGTCTGACCTGGTTCCACCTTGCTGATC
Eant_T6859a    CGGTCCTGACGTGCAAATCGGTCGTCTGACCTGGTTCCACCTTGCTGATC
Ebou_R0153     CGGTCCTGACGTGCAAATCGGTCGTCTGACCTGGTTCCACCTTGCTGATC
;

END;

BEGIN TREES;
     TRANSLATE
       1 ‘Eant_T6857’,
       2 ‘Etri_6842’,
       3 ‘Eant_T6859a’,
       4 ‘Ebou_R0153’
       ;
     tree epitree = [&r]((1:1.0E-6,2:2.0E-6)100:3.01E-4,(3:1.9E-5,4:2.3E-5)100:3.84E-4);
END;
```