Phylogenetics at the QIB

Leonardo de Oliveira Martins, PhD
QIB Head of Phylogenomics

Quadram Institute
Science Health Food Innovation

@leomrtns@mstdn.science

Phylogenetics at the QIB — training material

date	location	time
2023.11.13	**UG55A** (QIB, upper ground floor)	**1400 to 1600**
2023.11.20	**UG44A** (QIB, upper ground floor)	**1400 to 1600**

This repository contains data sets and links you may find useful when following the lectures. Some links will only work if you have access to the NBI/QIB network.

In these lectures I will briefly review phylogenetic concepts, and discuss issues particularly important to those working with microbial genomics. Due to time constraints and the vast availability of further courses and online tutorials, I won't go into details of any method or tool.

I provide below a list of further materials and references for you to follow during or after the lectures.

Material to check before the lectures

The lectures assume some familiarity with the command line and with resources at the QIB. Check the list of resources available at the QIB (internal link) if you need further training. The binfie team also provides a list of external training resources.

In particular, although we'll try to accomodate all levels of expertise, this is not an introductory course on phylogenetics. The following is a list of open resouces I recommend for a overview of modern phylogenetics, in order of relevance for the lectures:

https://github.com/quadram-institute-bioscience/phylo-methods/tree/main/masterclass

Quadram Institute
Science Health Food Innovation

Why? Because phylogenetic analyses rely a lot on

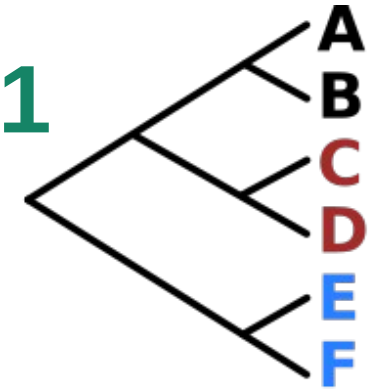
- fixing names, metadata, removing and adding sequences to analyses...

```
EDI017_FG91-AK3909__5337

EDI17_FG91-AK3909__5337

'EDI017_FG91-AK3909 5337'

EDI17-FG91-AK3909
```

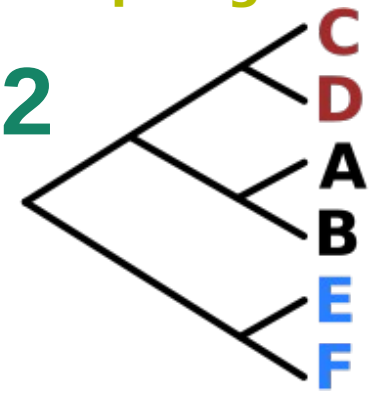
- Checking assumptions, interpreting trees under each assumption etc. **even after obtaining some tree**

1. Which Maximum Likelihood topologies are different?

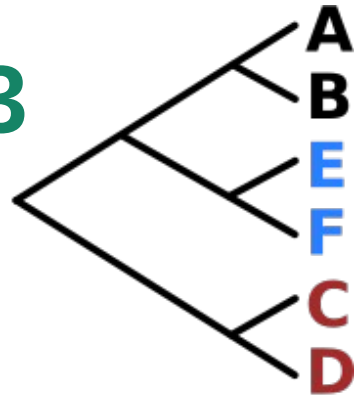
1



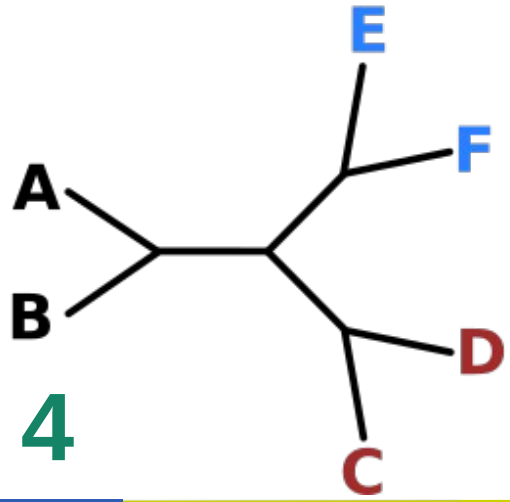
2



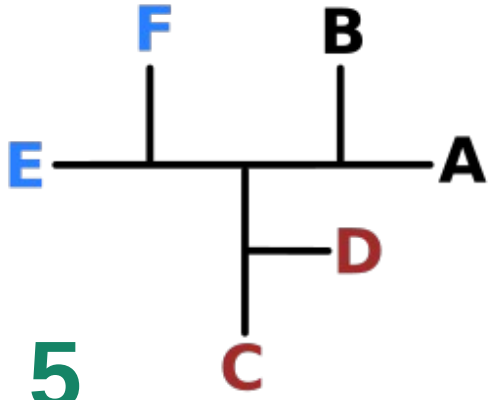
3



4



5



Nomenclature

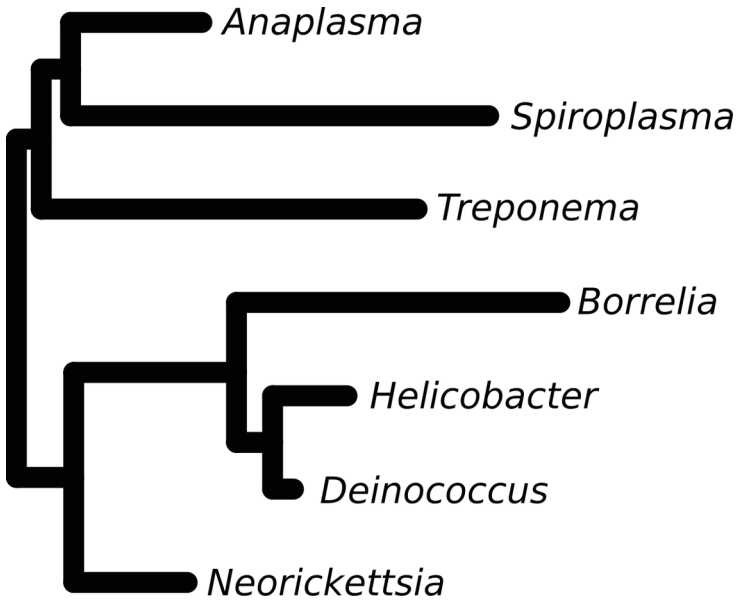
“Toutes choses sont dites déjà; mais comme personne n’écoute, il faut toujours recommencer”
(André Gide)

Quadram Institute
Science Health Food Innovation

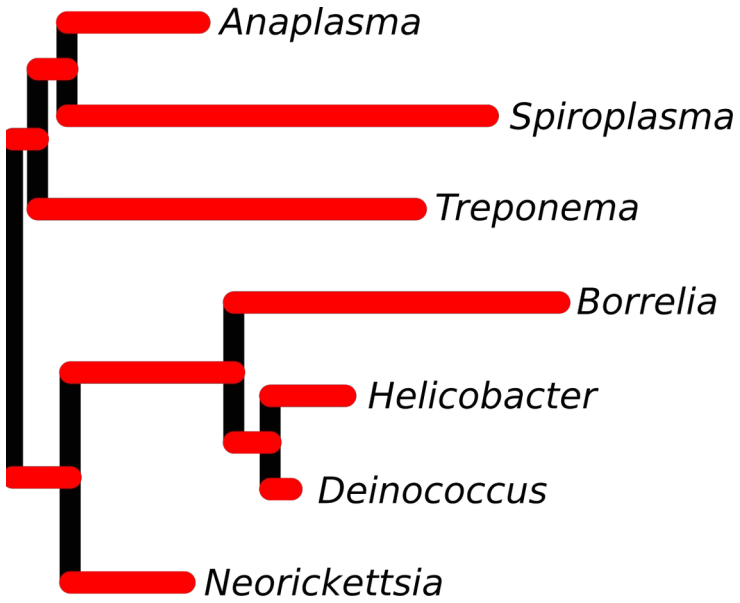
2. Which phylogenetic terms do you know?

1. branch
2. node
3. root
4. tip
5. iqtree
6. clade
7. bipartition or split

Tree terms

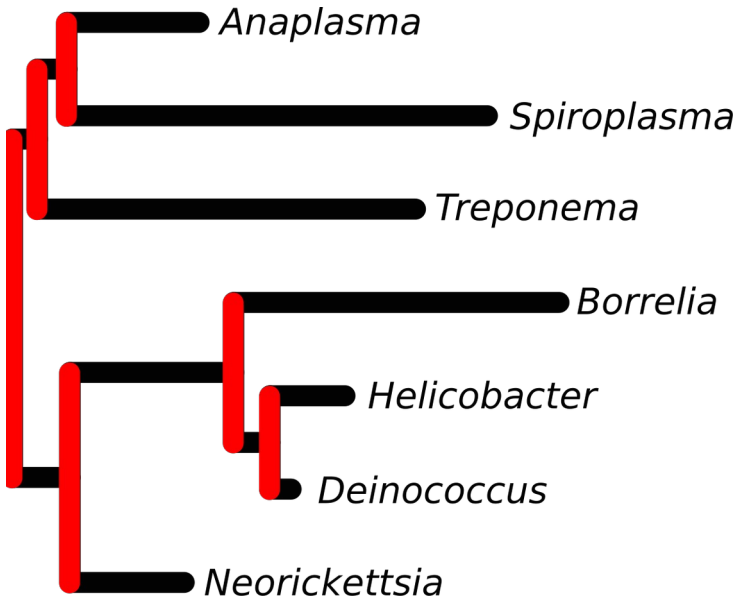

Anaplasma
Spiroplasma
Treponema
Borrelia
Helicobacter
Deinococcus
Neorickettsia

Tree terms



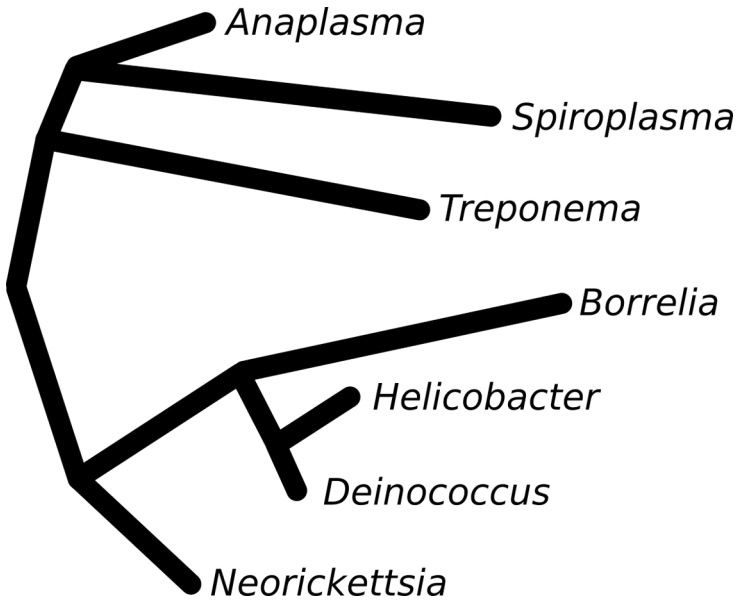
- branch (or edge)
- x-axis has information about branch lengths

Tree terms



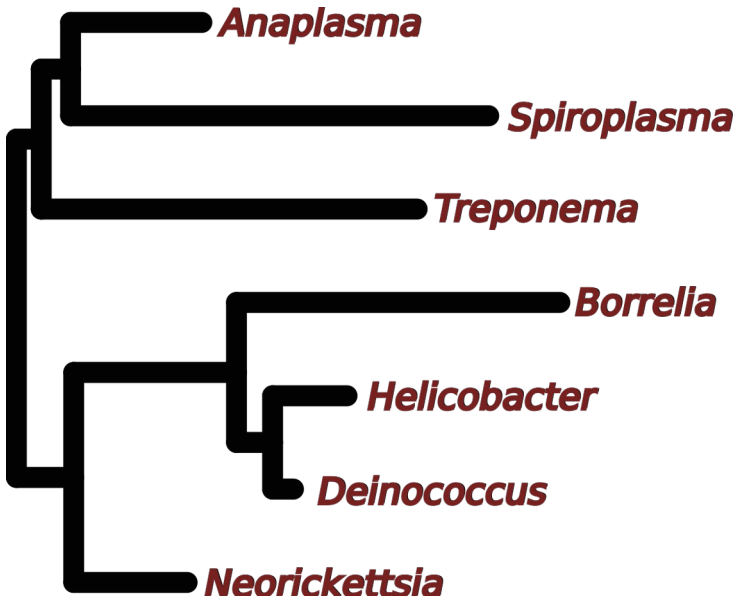
- on the other hand, y-axis has NO information

Tree terms



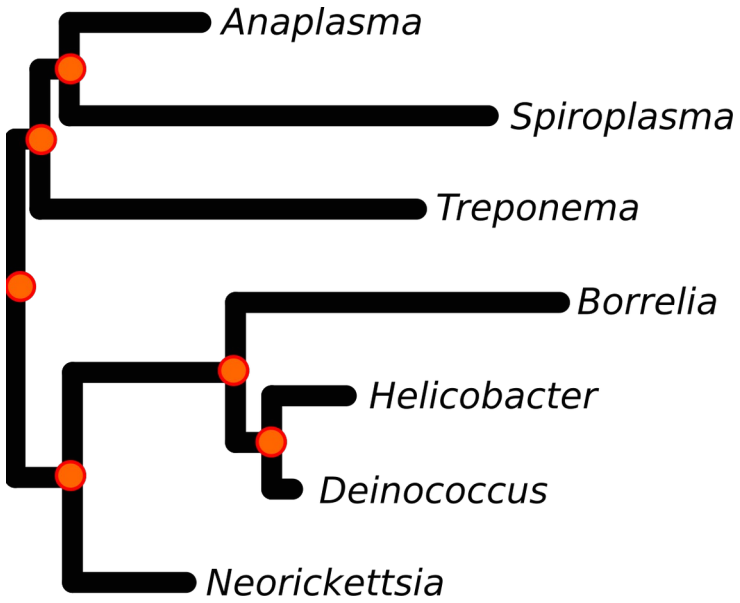
- This tree has same information, but it’s harder to see branch lengths
- However, so-called cladograms are used without length information

Tree terms



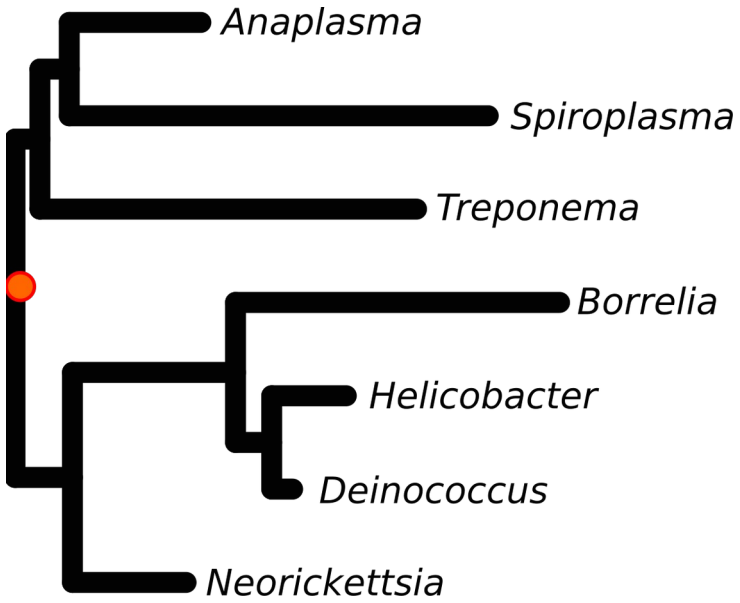
- Tips, leaves, or external nodes
- Usually represent our data (“extant information”)

Tree terms



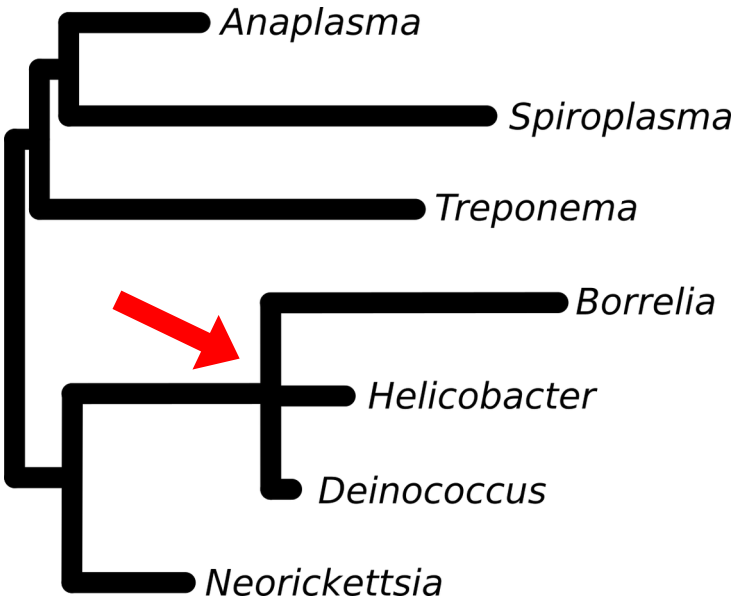
- Internal nodes, represent “entities” inferred by phylogenetics
- “nodes” are also known as “vertices”
- parental (ancestral) nodes have descendant nodes
- Most Recent Common Ancestor (MRCA)

Tree terms



- Root (root node)
- define a time directionality: root in the past, descendant nodes towards the present

Tree terms



- polytomy or multifurcation
- otherwise, a dichotomy or bifurcation

→ Most phylogenetic reconstruction methods assume **binary trees**, i.e. without any multifurcation.

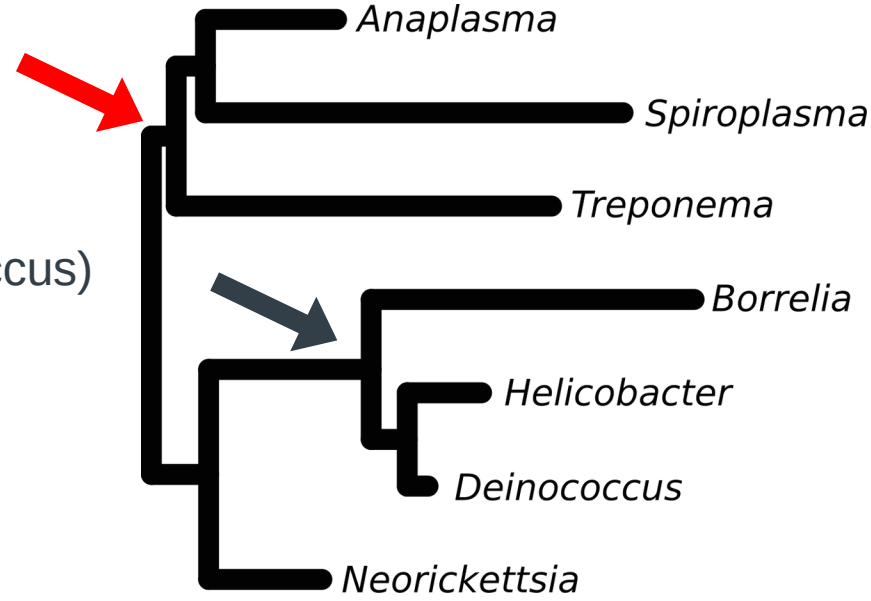
→ If you see a multifurcation, it usually means one of two things:

- a very short branch length, which is, for all effects, same as zero
- a summary from several trees (e.g. bootstrap replicates), where there was no agreement for that clade or split

Tree terms

(Anaplasma, Spiroplasma, Treponema)

(Borrelia, Helicobacter, Deinococcus)



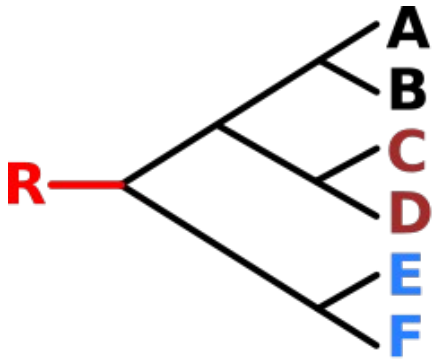
- clades (rooted trees)
- splits or bipartitions (unrooted trees)

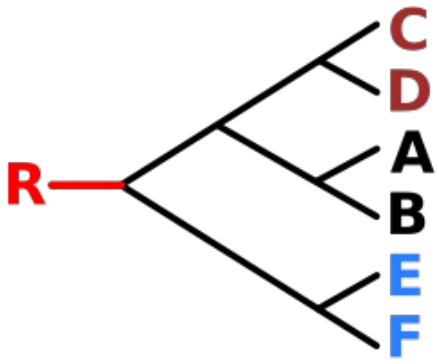
Rooted vs unrooted

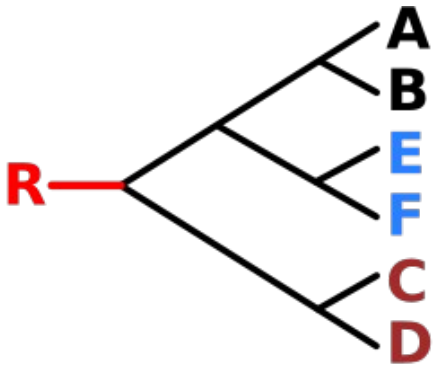
newick format vs visualisation

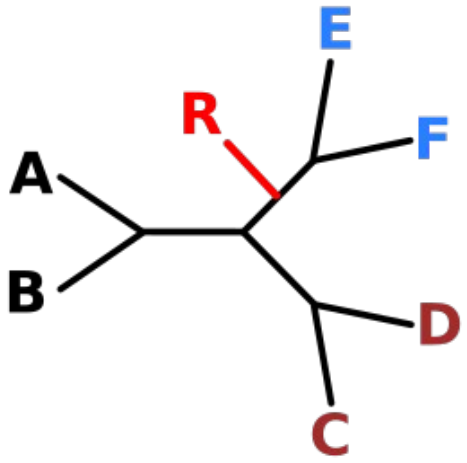
Quadram Institute
Science Health Food Innovation

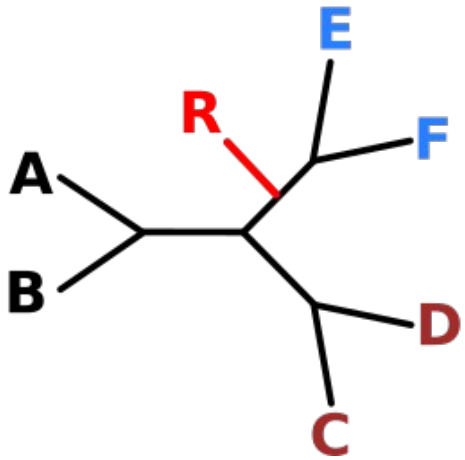
Rooted trees are like unrooted trees with a special leaf

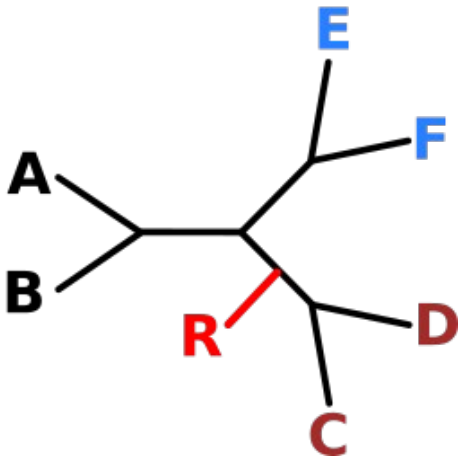






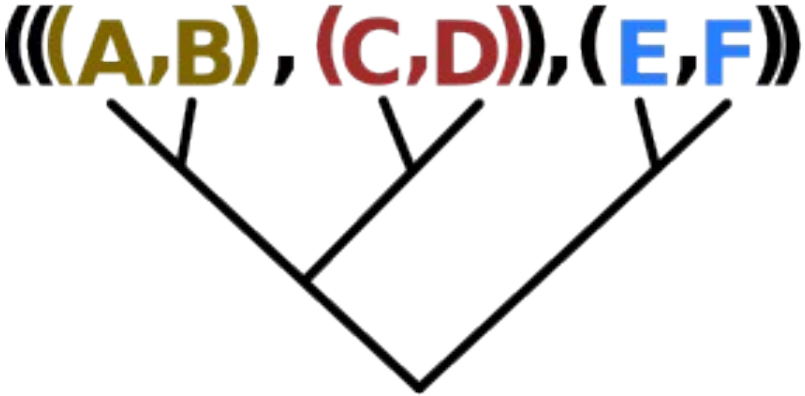


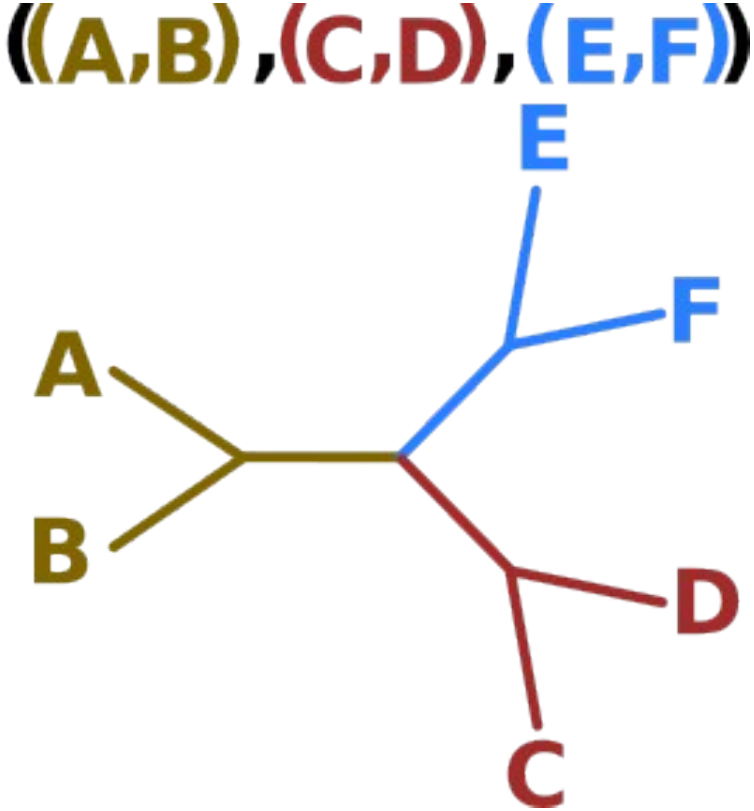




Quadram Institute
Science Health Food Innovation

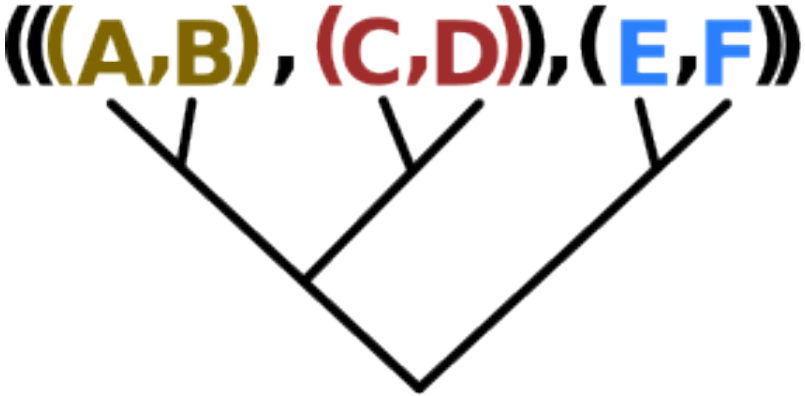
Newick representation of trees

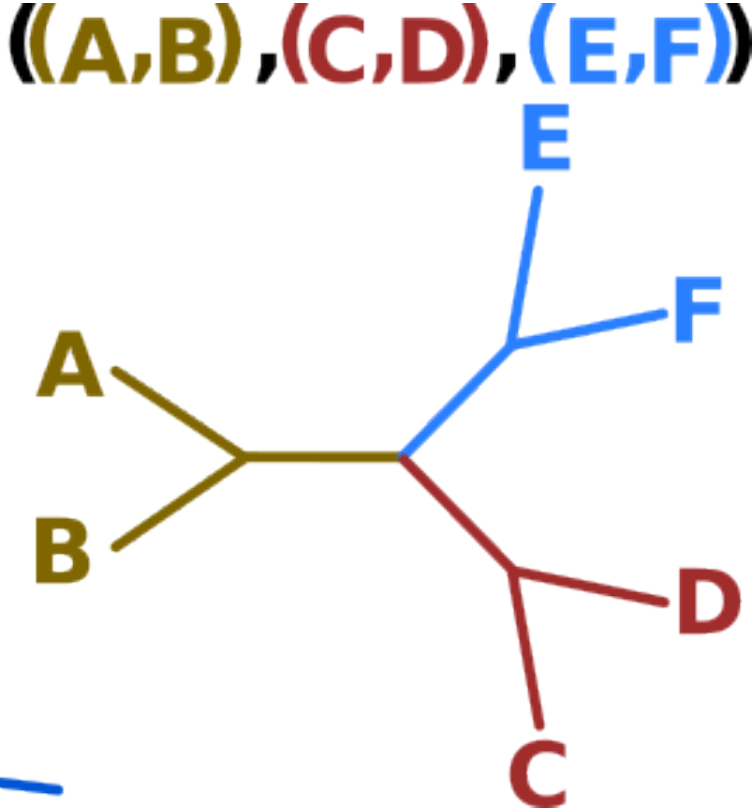


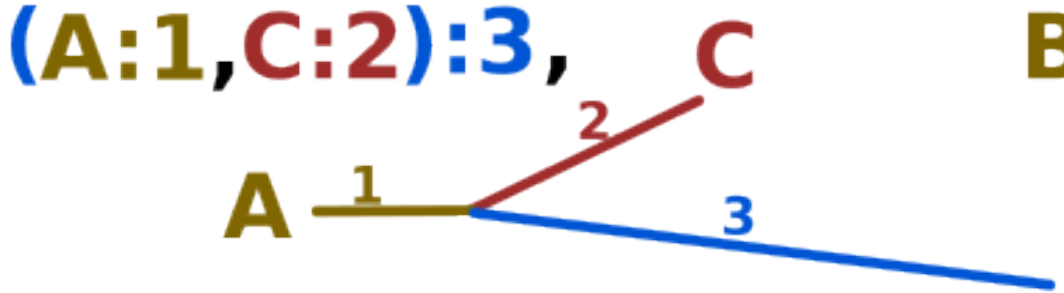


- rooted: always dicotomy (binary)
- unrooted: basal tricotomy (multifurcation)

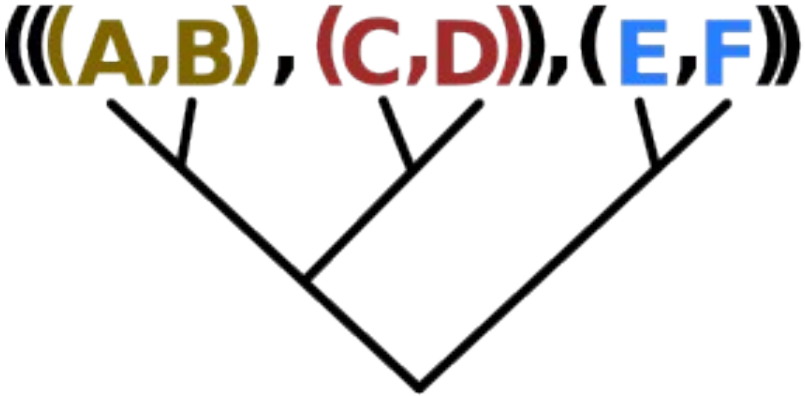
Newick representation of trees

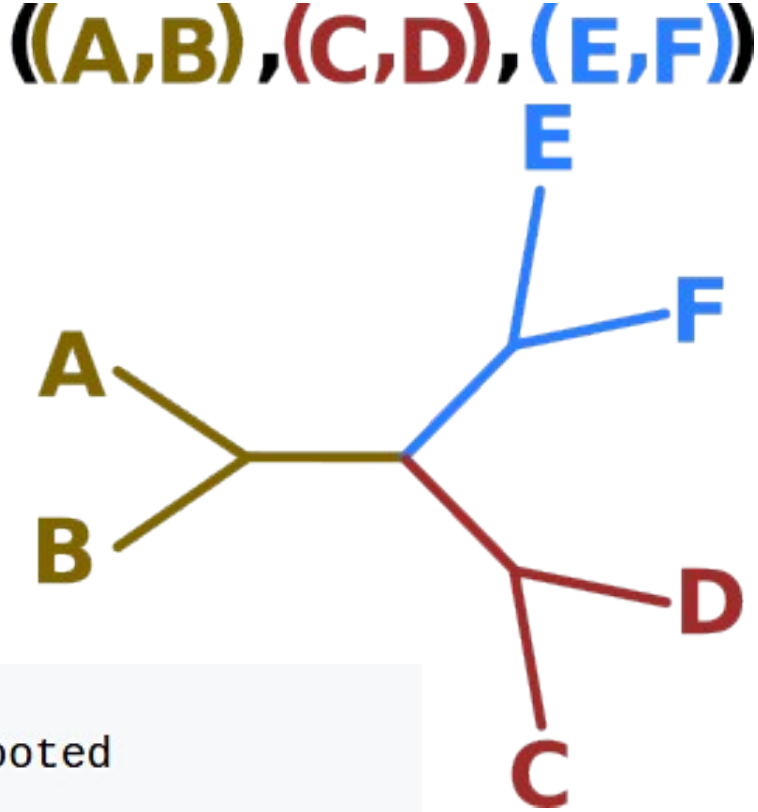






Newick representation of trees






```
(A,B,C)              ## this tree is unrooted
[U] ((A,B),C)        ## this tree is also unrooted
((A:1,B:1):0,C:1) ## also unrooted?
((A,B,C),D)          ## for all effects, this tree is rooted
```

→ Keep in mind that most phylogenetic reconstruction methods return **unrooted** trees (likelihood, parsimony, minimum evolution). Even if newick tree looks rooted, and tree visualisation software shows it rooted.

→ Exceptions are neighbour joining and UPGMA, which are *clustering* methods.

→ Options to find the root:

- use a **known** outgroup
- midpoint, root-to-tip regression, or another molecular clock approach.
- use a non-reversible model (hard) or a relaxed clock model (more common)

Curiosity: the nexus format

— used for both sequences and trees

— comments in square brackets allowed for many extensions to the newick format

Quadram Institute
Science Health Food Innovation

“nexus” file format is used for both sequences and trees

```
#NEXUS
Begin TAXA;
  Dimensions ntax=4;
  TaxLabels SpaceDog SpaceCat SpaceOrc SpaceElf;
End;

Begin data;
  Dimensions nchar=15;
  Format datatype=dna missing=? gap=- matchchar=.;
  Matrix
    [ comments in brackets ]
    SpaceDog   atgctagctagctcg
    SpaceCat   ......??...-.a.
    SpaceOrc   ...t.......-.g. [ same as atgttagctag-tgg ]
    SpaceElf   ...t.......-.a.
  ;
End;

BEGIN TREES;

  Tree tree1 = [newick format] (((SpaceDog,SpaceCat),SpaceOrc,SpaceElf));
END;
```

example from wikipedia

Used by `figtree` to annotate the tree

```
#NEXUS
begin taxa;
  dimensions ntax=297;
  taxlabels
  'England/ALDP-18A770B/2021'[&global="global",lineage="AY.4"]
  'England/ALDP-19CAA17/2021'[&global="global",lineage="AY.4"]
  'England/ALDP-19CACF3/2021'[&global="global",lineage="AY.4"];
end;

begin trees;
  tree tree_1 = [&R] ((((((((((((('England/QEUH-2E6571F/2021'[&lineage="AY.4",scor
pio_call="Delta (AY.4-like)"]:6.8E-5,'England/ALDP-2989E15/2021'[&lineage="AY.4",s
corpio_call="Delta (AY.4-like)"]:1.0E-6)[&lineage="AY.4",scorpio_call="Delta (AY.4
-like)"]:3.3E-5,(((((((((('England/ALDP-2CD1BC9/2021'[&lineage="AY.4",scorpio_call
="Delta (AY.4-like)"]:3.3E-5,'England/QEUH-2AD42C3/2021'[&lineage="AY.4",scorpio_c
all="Delta (AY.4-like)"]:3.2E-5)[&lineage="AY.4",scorpio_call="Delta (AY.4-like)"]
...
end;

begin figtree;
  set appearance.backgroundColorAttribute="Default";
  set appearance.backgroundColour=#ffffff;
  set appearance.branchColorAttribute="scorpio_call";
  set appearance.branchColorGradient=false;
  set appearance.branchLineWidth=2.0;
  set appearance.branchMinLineWidth=0.0;
```

Bayesian software outputs posterior probabilities of trees

all trees

```
#NEXUS

Begin trees;
 Translate
  1  schizosaccharomyces_pombe,
  2  saccharomyces_cerevisiae,
  3  drosophila_yakuba,
  4  monosiga_brevicollis,
  5  caenorhabditis_elegans,
  6  schistosoma_mansoni,
  7  drosophila_melanogaster,
  8  gallus_gallus,
  9  homo_sapiens;

tree tree_0  =  ((((((6,5),(2,1)),(7,3)),4),8),9);
tree tree_1  =  ((((((2,1),5),6),(7,3)),4),9),8);
tree tree_2  =  (((((7,3),(2,1)),(6,5)),4),(9,8));
tree tree_3  =  ((((((7,3),(2,1)),(6,5)),4),9),8);
tree tree_4  =  ((((((2,1),5),6),(7,3)),4),8),9);
tree tree_5  =  (((((6,5),(2,1)),(7,3)),4),(9,8));
tree tree_6  =  (((((7,3),(2,1)),(6,5)),4),(9,8));
tree tree_7  =  (((((7,3),(2,1)),(6,5)),4),(9,8));
      ...
tree tree_999  =  ((((((6,5),(2,1)),(7,3)),4),8),9);

End;
```

summary

```
#NEXUS

Begin trees;
 Translate
  1  schizosaccharomyces_pombe,
  2  saccharomyces_cerevisiae,
  3  drosophila_yakuba,
  4  monosiga_brevicollis,
  5  caenorhabditis_elegans,
  6  schistosoma_mansoni,
  7  drosophila_melanogaster,
  8  gallus_gallus,
  9  homo_sapiens;

tree tree_3   [p = 0.1748, P = 0.1748] =  ((((((7,3),(2,1)),(6,5)),4),9),8);
tree tree_0   [p = 0.1658, P = 0.3407] =  ((((((6,5),(2,1)),(7,3)),4),8),9);
tree tree_2   [p = 0.1618, P = 0.5025] =  (((((7,3),(2,1)),(6,5)),4),(9,8));
tree tree_6   [p = 0.1499, P = 0.6523] =  ((((((6,5),(2,1)),(7,3)),4),9),8);
tree tree_7   [p = 0.1479, P = 0.8002] =  ((((((7,3),(2,1)),(6,5)),4),8),9);
tree tree_5   [p = 0.1469, P = 0.9471] =  (((((6,5),(2,1)),(7,3)),4),(9,8));
tree tree_4   [p = 0.0110, P = 0.9580] =  ((((((2,1),5),6),(7,3)),4),8),9);
tree tree_9   [p = 0.0100, P = 0.9680] =  ((((((2,1),6),5),(7,3)),4),8),9);
tree tree_1   [p = 0.0090, P = 0.9770] =  ((((((2,1),5),6),(7,3)),4),9),8);
tree tree_10  [p = 0.0080, P = 0.9850] =  (((((2,1),5),6),(7,3)),4),(9,8));
tree tree_11  [p = 0.0070, P = 0.9920] =  (((((2,1),6),5),(7,3)),4),(9,8));
tree tree_8   [p = 0.0060, P = 0.9980] =  ((((((2,1),6),5),(7,3)),4),9),8);
tree tree_12  [p = 0.0020, P = 1.0000] =  (((((7,3),(2,1)),(6,5)),(9,8)),4);

End;
```

Quadram Institute
Science • Health • Food • Innovation

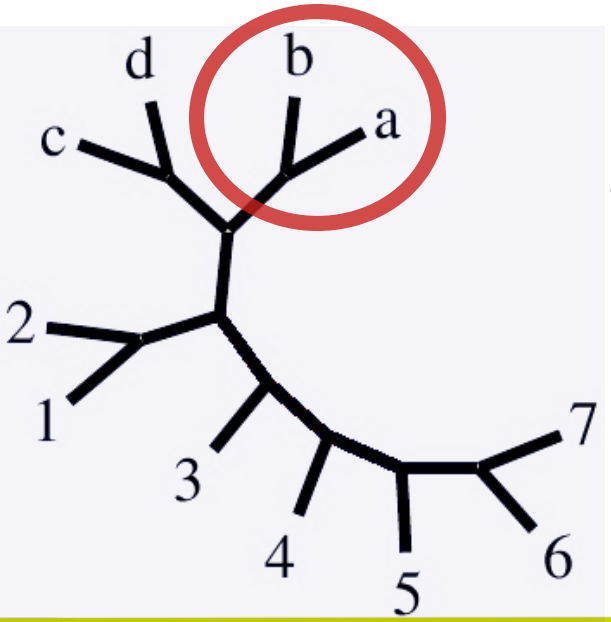
Distance between trees

RF distance is the most common

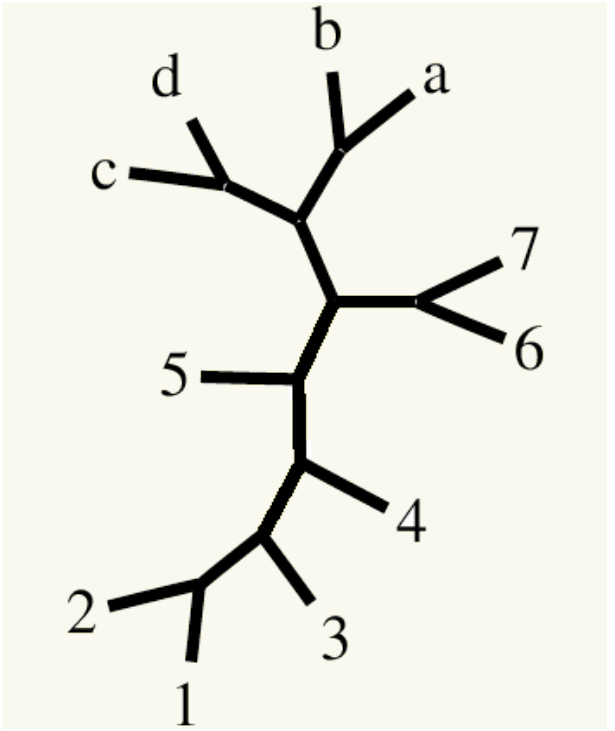
Quadram Institute
Science Health Food Innovation

Robinson-Foulds distance (a.k.a. symmetric difference)

1. get all splits from both trees

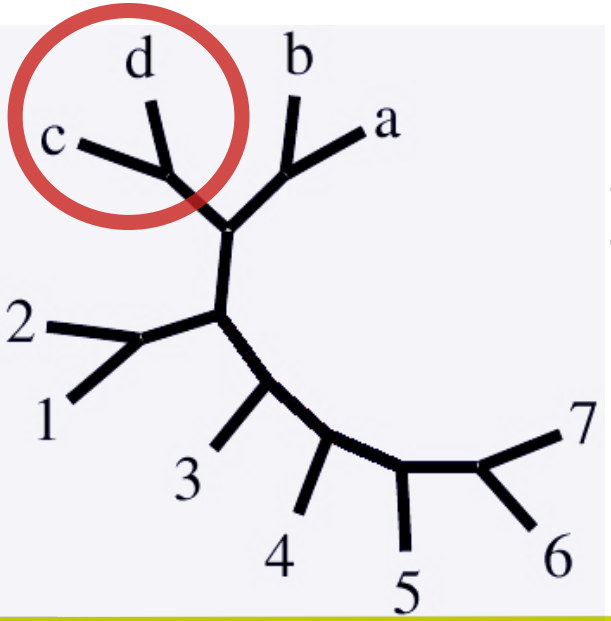


ab | cd1234567

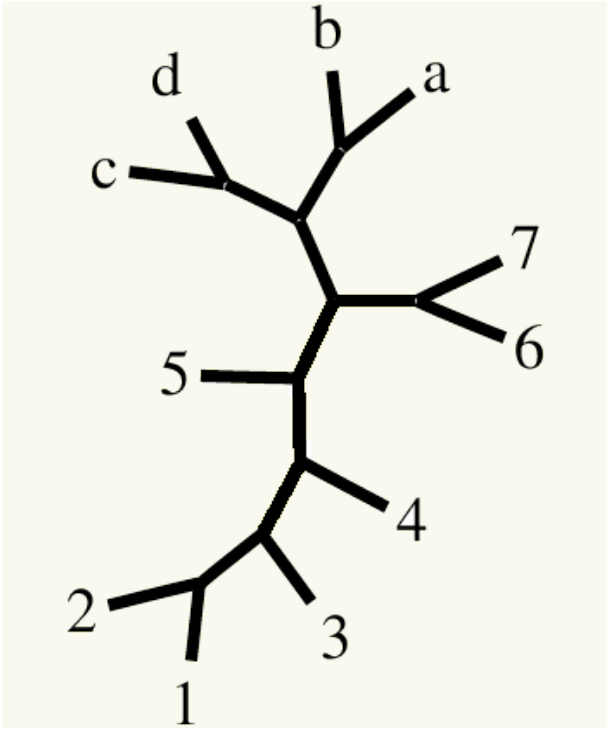


Robinson-Foulds distance (a.k.a. symmetric difference)

1. get all splits from both trees

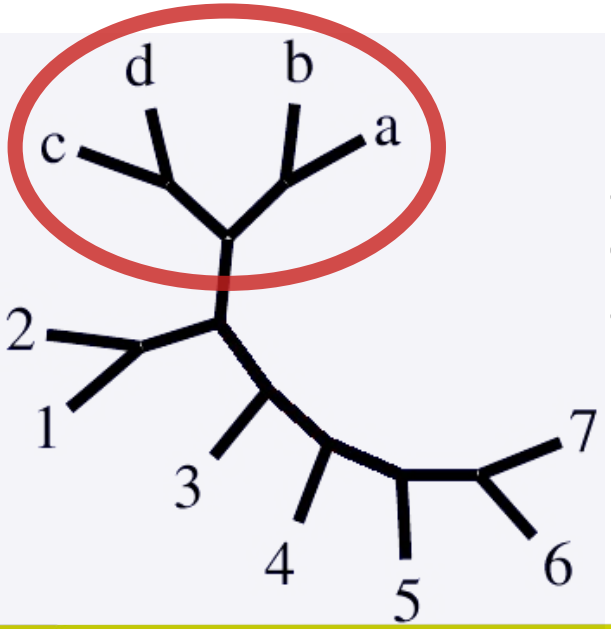


ab | cd1234567
cd | ab1234567

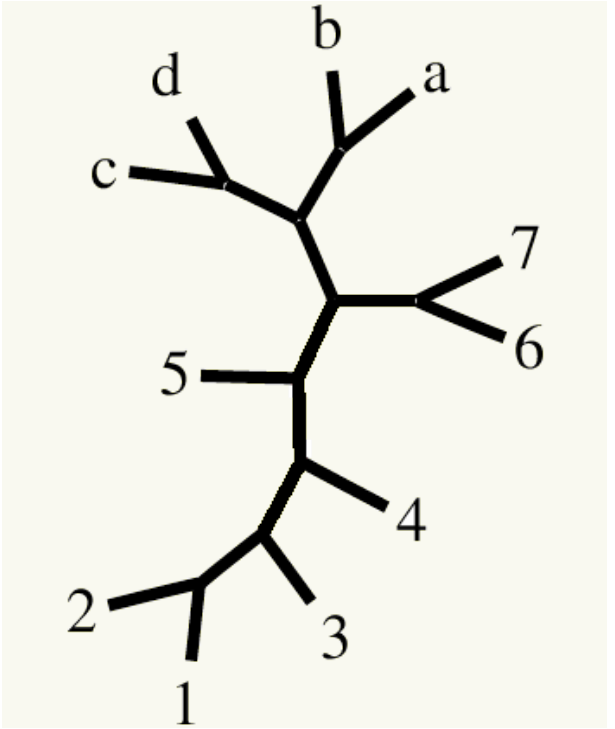


Robinson-Foulds distance (a.k.a. symmetric difference)

1. get all splits from both trees

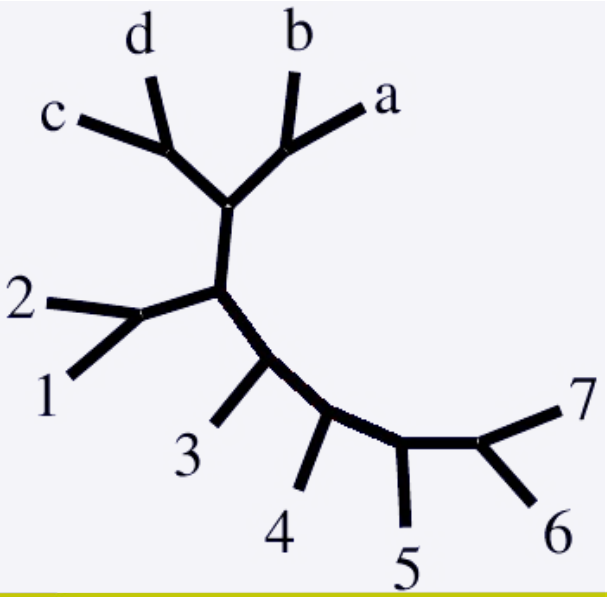


ab | cd1234567
cd | ab1234567
abcd | 1234567

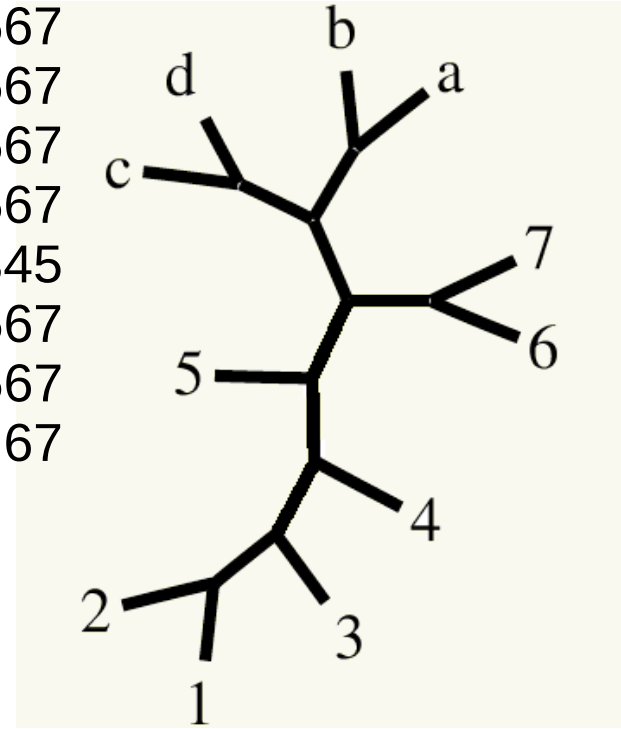


Robinson-Foulds distance (a.k.a. symmetric difference)

1. get all splits from both trees



ab | cd1234567
cd | ab1234567
abcd | 1234567
12 | abcd34567
abcd12 | 34567
abcd123 | 4567
abcd1234 | 567
abcd12345 | 67

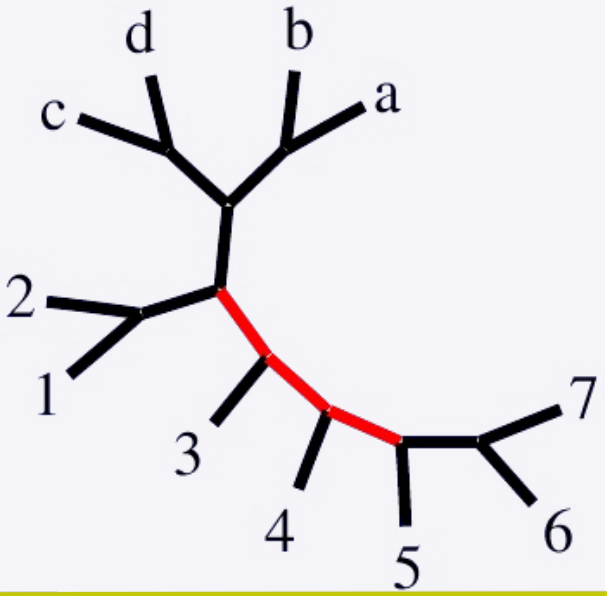


ab | cd1234567
cd | ab1234567
abcd | 1234567
12 | abcd34567
67 | abcd12345
123 | abcd4567
1234 | abcd567
12345 | abcd67

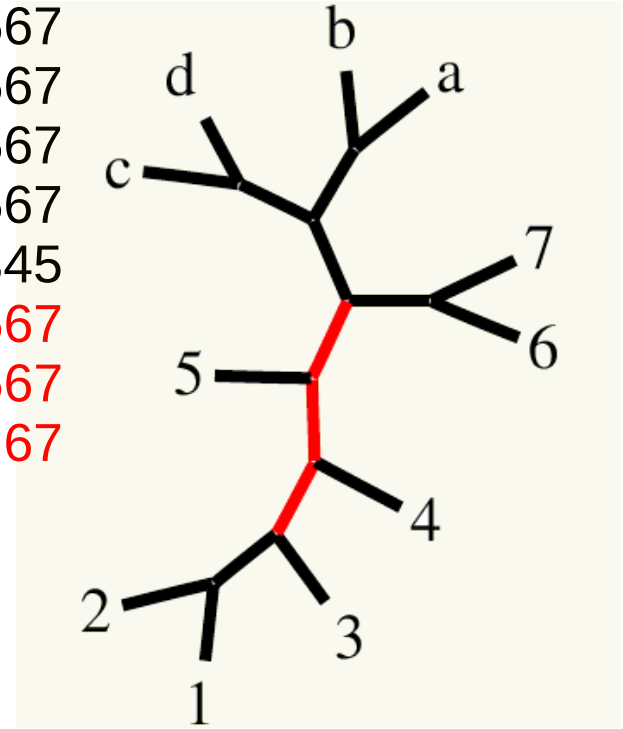
Quadram Institute
Science Health Food Innovation

Robinson-Foulds distance (a.k.a. symmetric difference)

1. get all splits from both trees
2. How many are different between trees



ab | cd1234567
cd | ab1234567
abcd | 1234567
12 | abcd34567
abcd12 | 34567
abcd123 | 4567
abcd1234 | 567
abcd12345 | 67



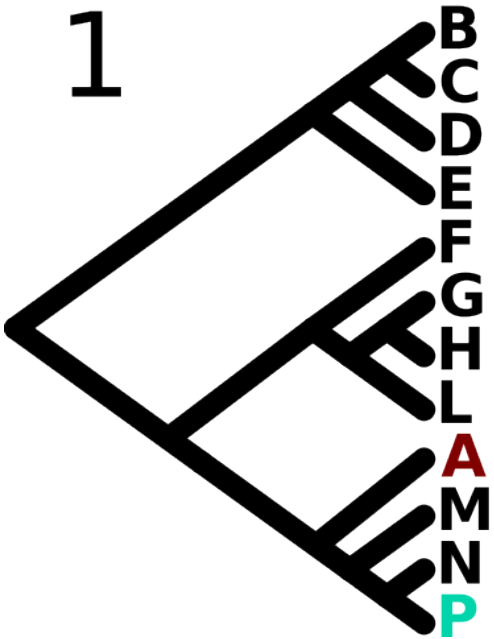
ab | cd1234567
cd | ab1234567
abcd | 1234567
12 | abcd34567
67 | abcd12345
123 | abcd4567
1234 | abcd567
12345 | abcd67

6 are different, 10 are the same

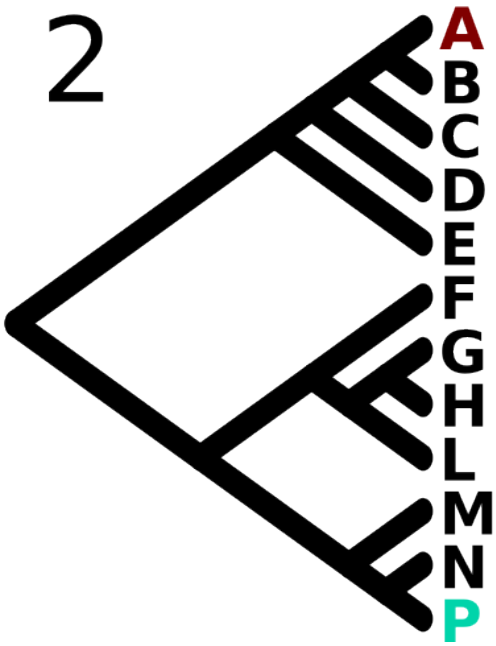
Quadram Institute
Science Health Food Innovation

3. Which tree is more similar to the second tree?

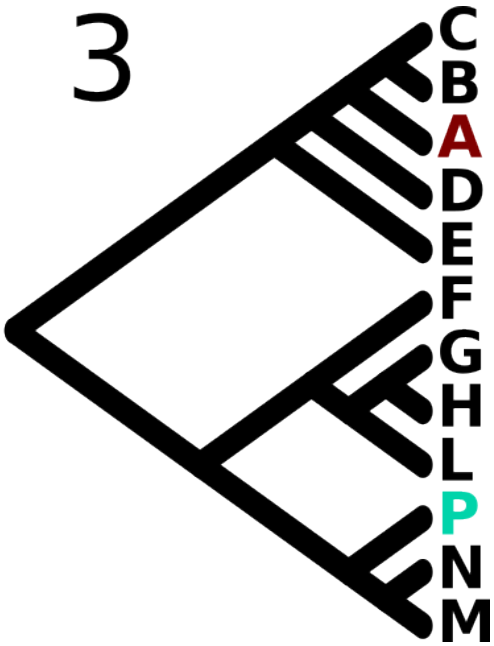
1



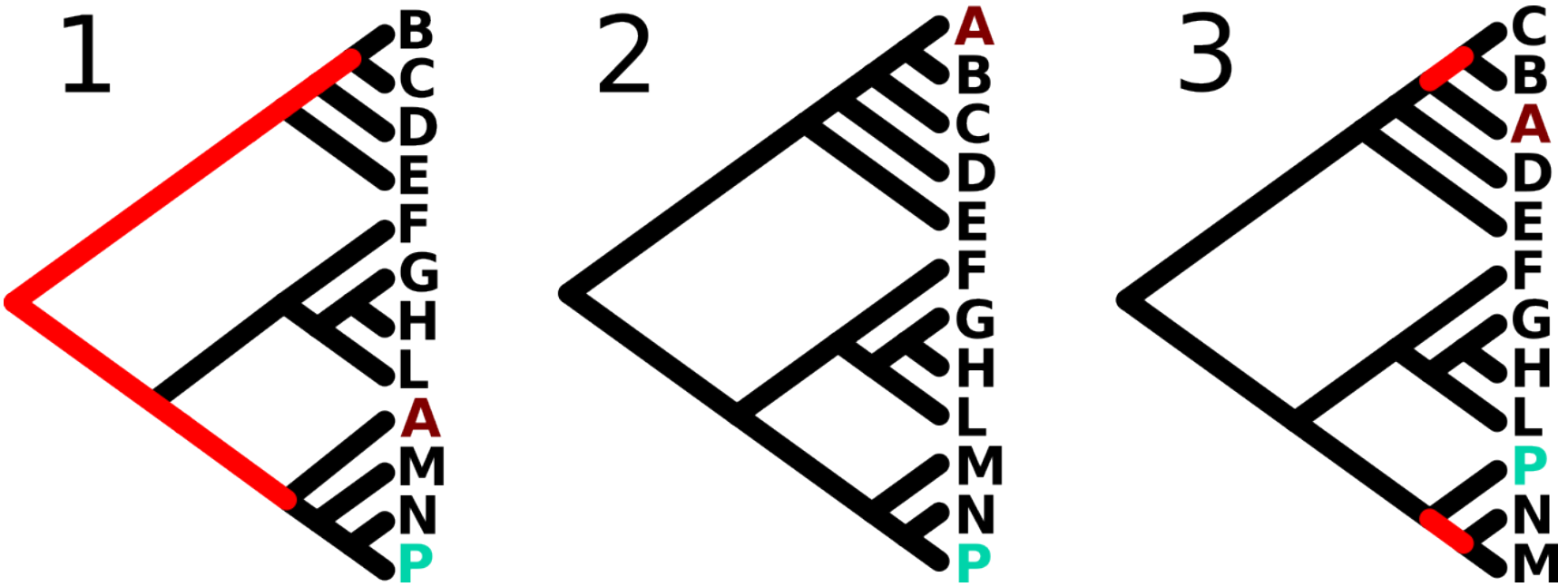
2



3



3. Which tree is more similar to the second tree?



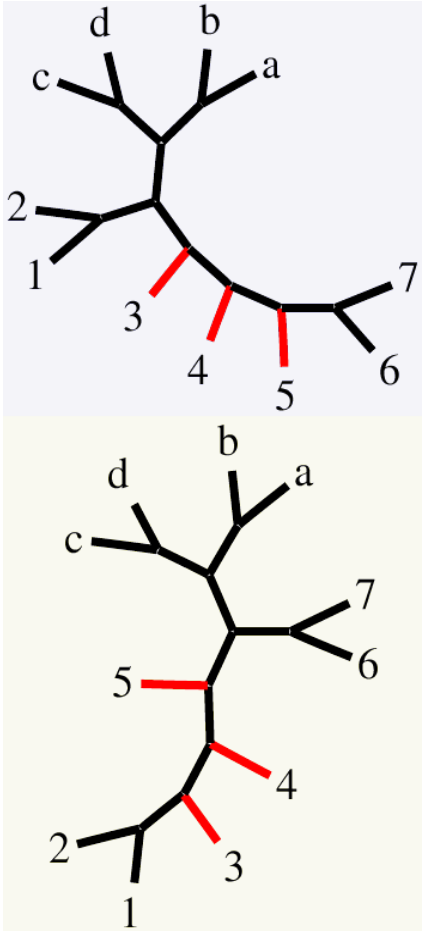
Robinson-Foulds counts all distinct bipartitions

Distance between trees

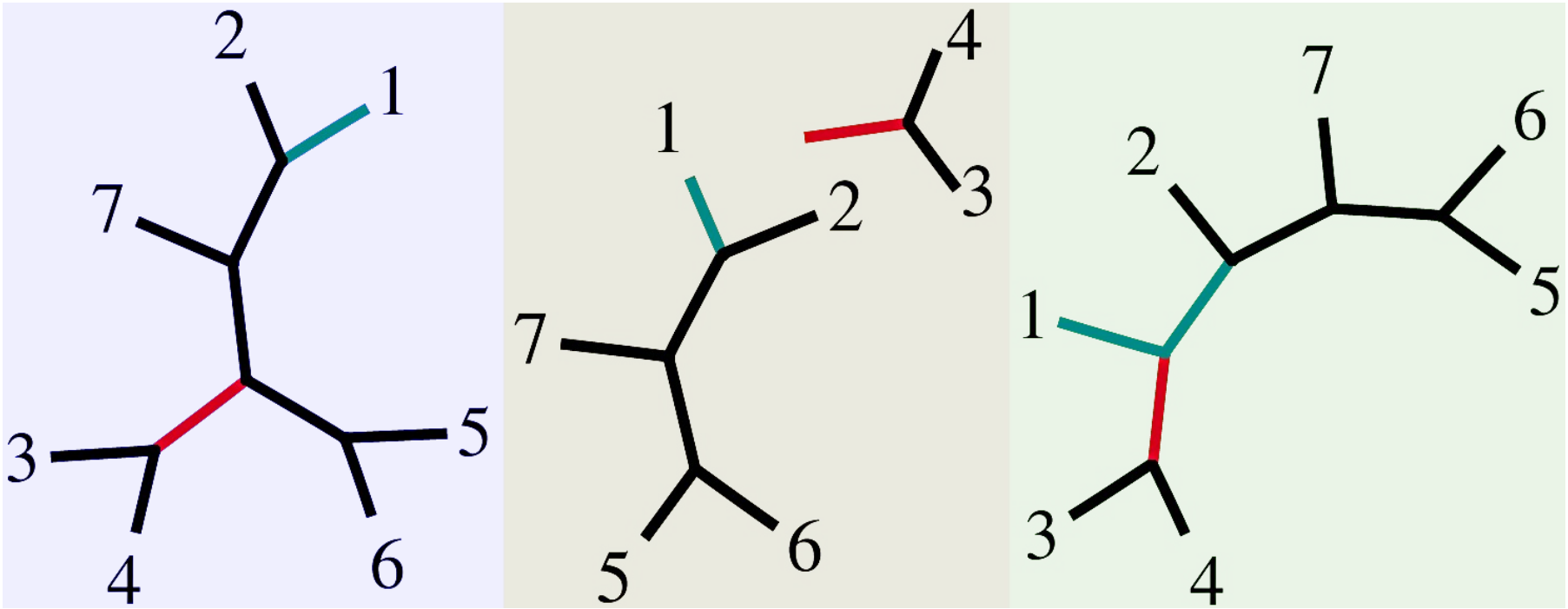
Other distances, all implemented in R

Quadram Institute
Science Health Food Innovation

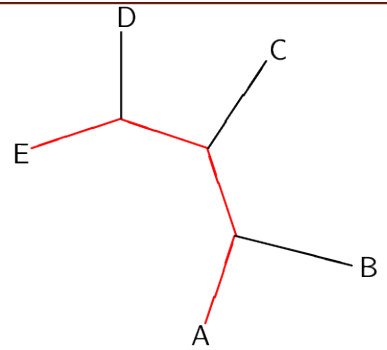
Maximum Agreement Subtree (MAST)

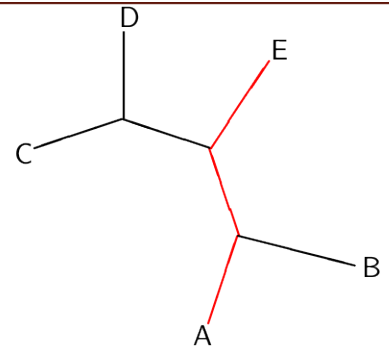


Subtree Prune-Regraft (SPR) distance



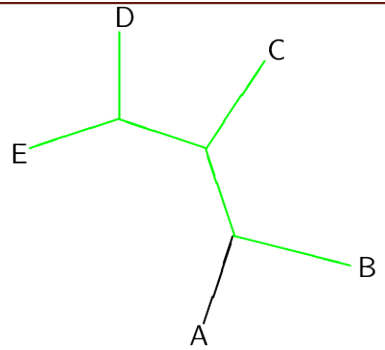
Quartet distance, Path difference

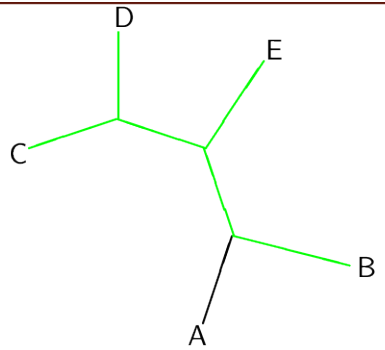




Path difference (number of speciations between trees)

- path from A to E is one edge longer in one tree than the other
- (...)
- the overall difference is 6

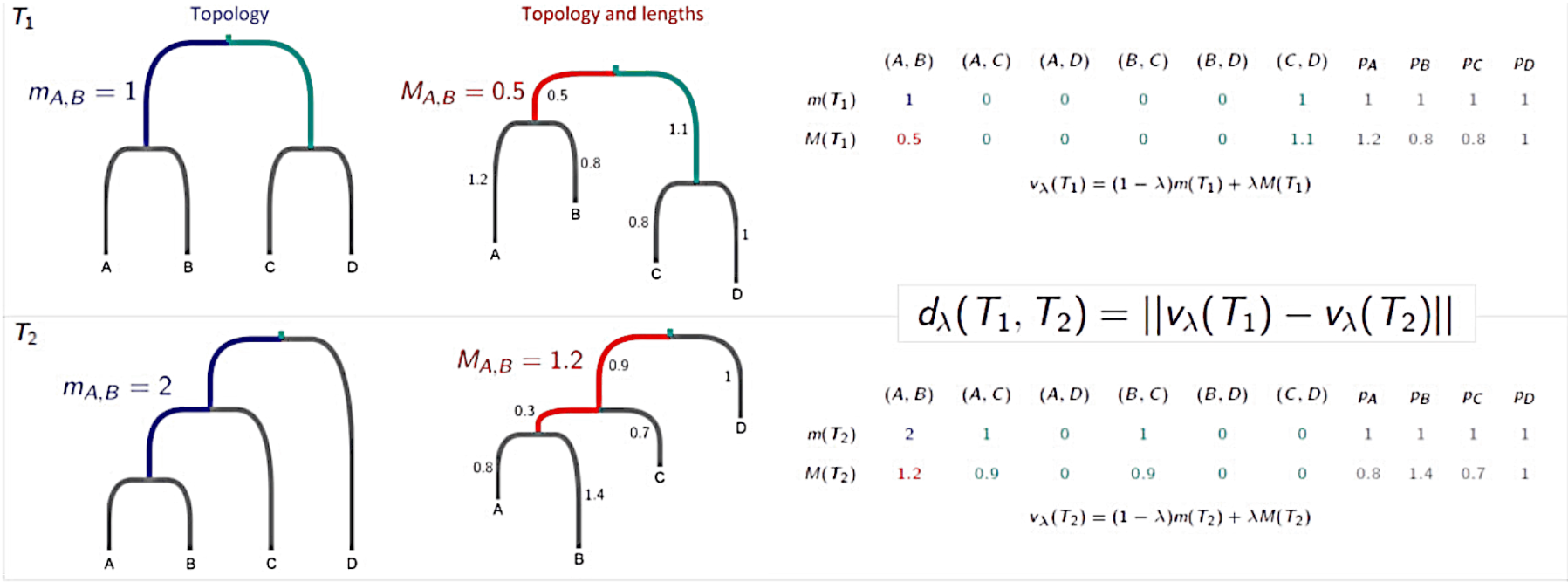




Quartet distance

- AC|DE and AE|CD
- BC|DE and BE|CD
- 4 quartets are different

Kendall–Colijn distance



	(A,B)	(A,C)	(A,D)	(B,C)	(B,D)	(C,D)	p_A	p_B	p_C	p_D
$m(T_1)$	1	0	0	0	0	1	1	1	1	1
$M(T_1)$	0.5	0	0	0	0	1.1	1.2	0.8	0.8	1

$$v_\lambda(T_1) = (1-\lambda)m(T_1) + \lambda M(T_1)$$

$$d_\lambda(T_1, T_2) = ||v_\lambda(T_1) - v_\lambda(T_2)||$$

	(A,B)	(A,C)	(A,D)	(B,C)	(B,D)	(C,D)	p_A	p_B	p_C	p_D
$m(T_2)$	2	1	0	1	0	0	1	1	1	1
$M(T_2)$	1.2	0.9	0	0.9	0	0	0.8	1.4	0.7	1

$$v_\lambda(T_2) = (1-\lambda)m(T_2) + \lambda M(T_2)$$

FIG. 1. A tree is characterized by the vectors m and M, which are calculated as shown. These are used to calculate the distance between two trees for any $\lambda \in [0, 1]$. Here, $d_0(T_1, T_2) = 2$ and $d_1(T_1, T_2) = 1.96$.

rooted trees only!

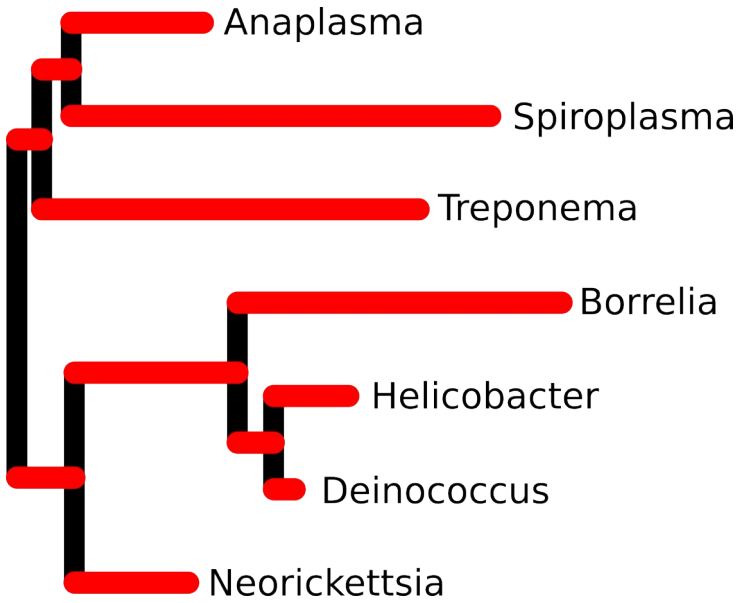
doi:10.1093/molbev/msw124

Quadram Institute
Science Health Food Innovation

Tree branch lengths

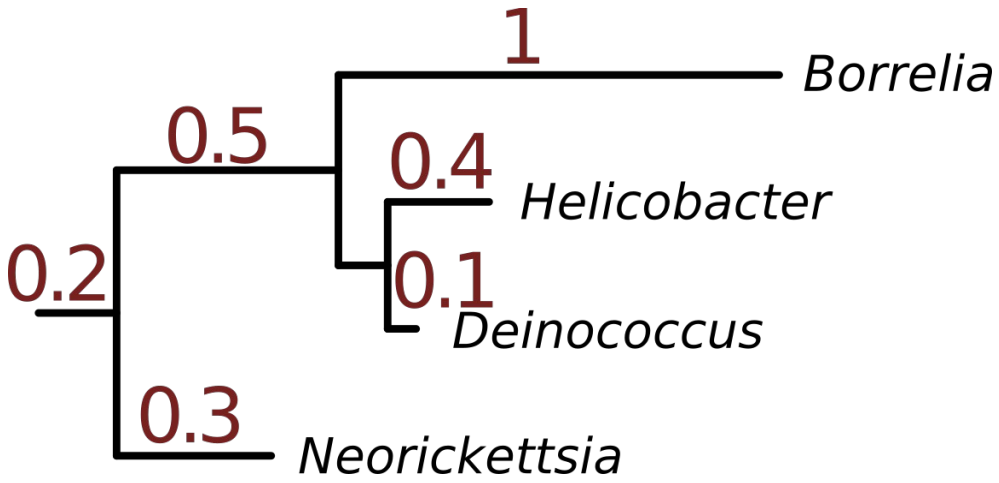
Quadram Institute
Science Health Food Innovation

Branch lengths



→ unit of a phylogenetic tree is “expected substitutions per site”

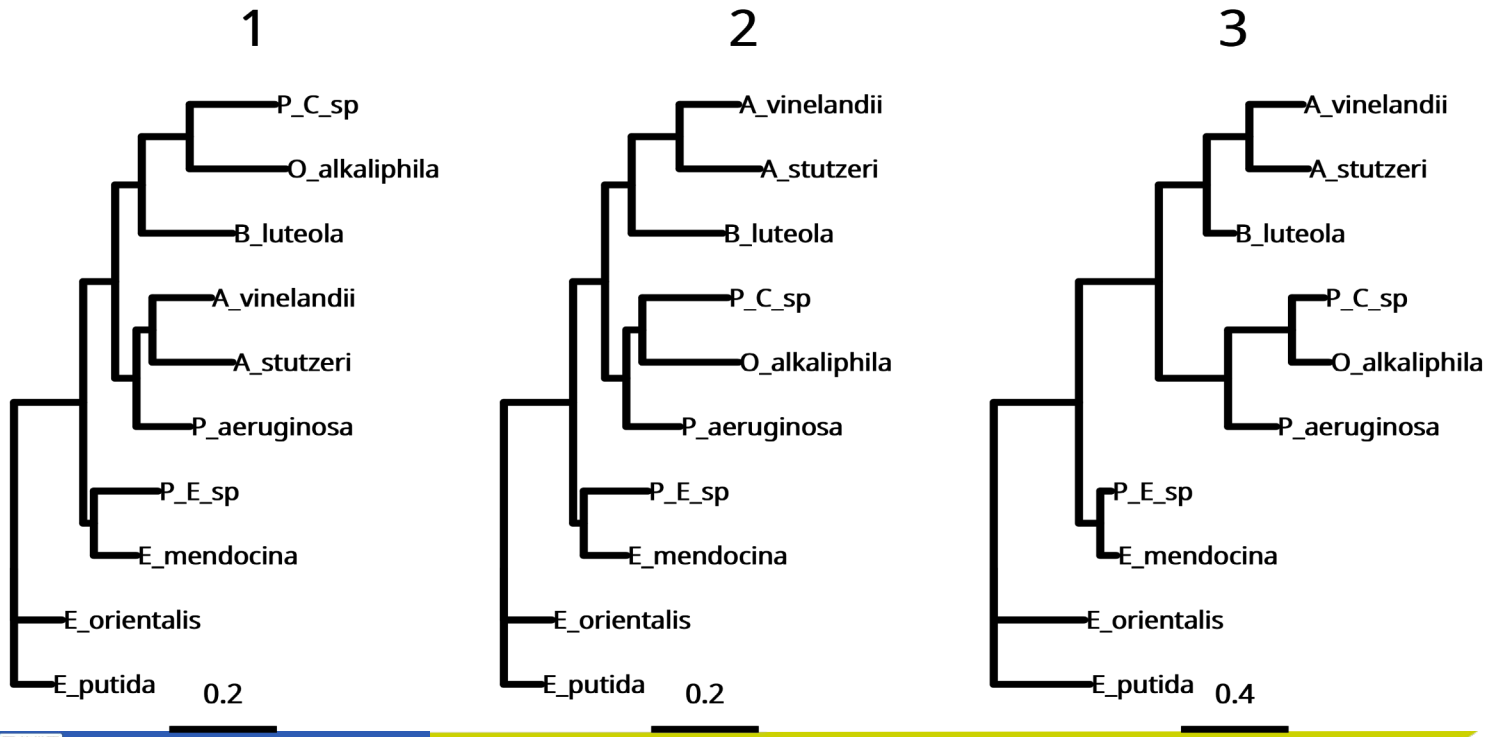
Branch lenghts



→ unit of a phylogenetic tree is “expected substitutions per site”

→ if alignment has 1000 DNA sites, we expect 500 substitutions (SNPs) between *Helicobacter* and *Deinococcus*

3. Which tree is more similar to tree 2?



3. Which tree is more similar to tree 2?

SPR.dist()

	1	2
2	1	
3	1	0

wRF.dist()

	1	2
2	0.30	
3	2.33	2.15

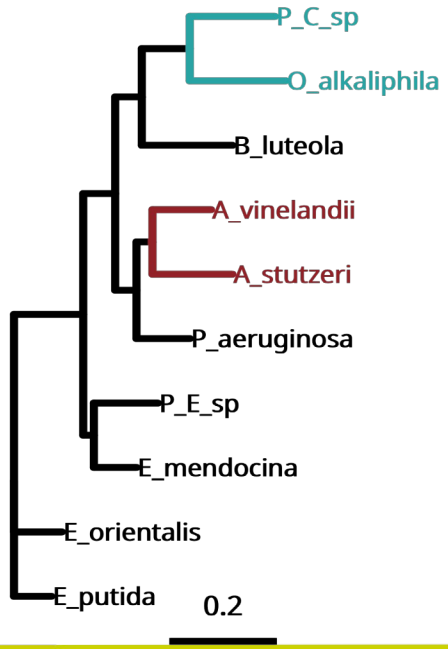
RF.dist()

	1	2
2	4	
3	4	0

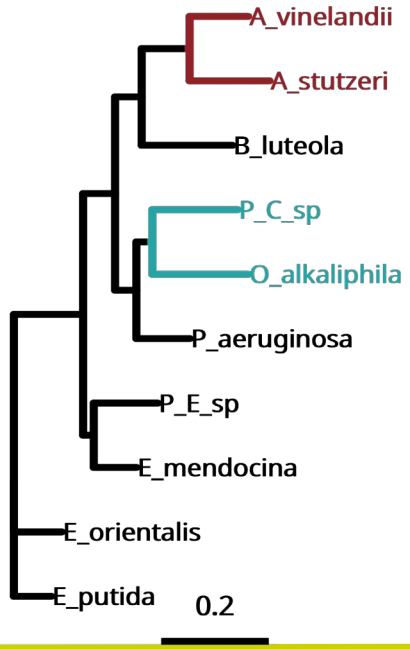
KF.dist()

	1	2
2	0.1240967	
3	0.6493843	0.6264982

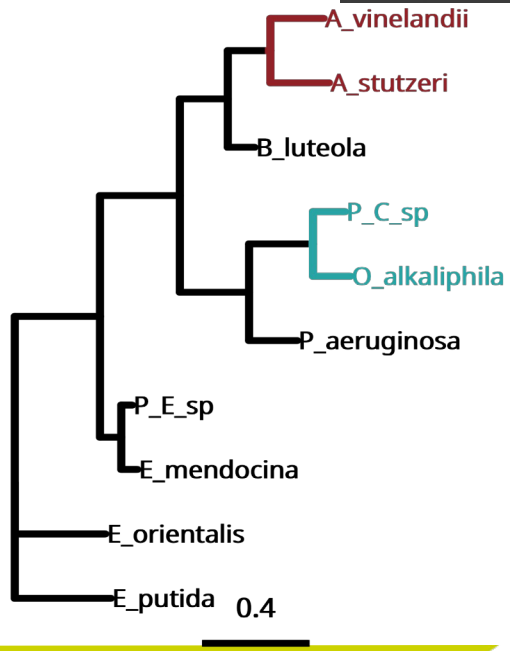
1



2

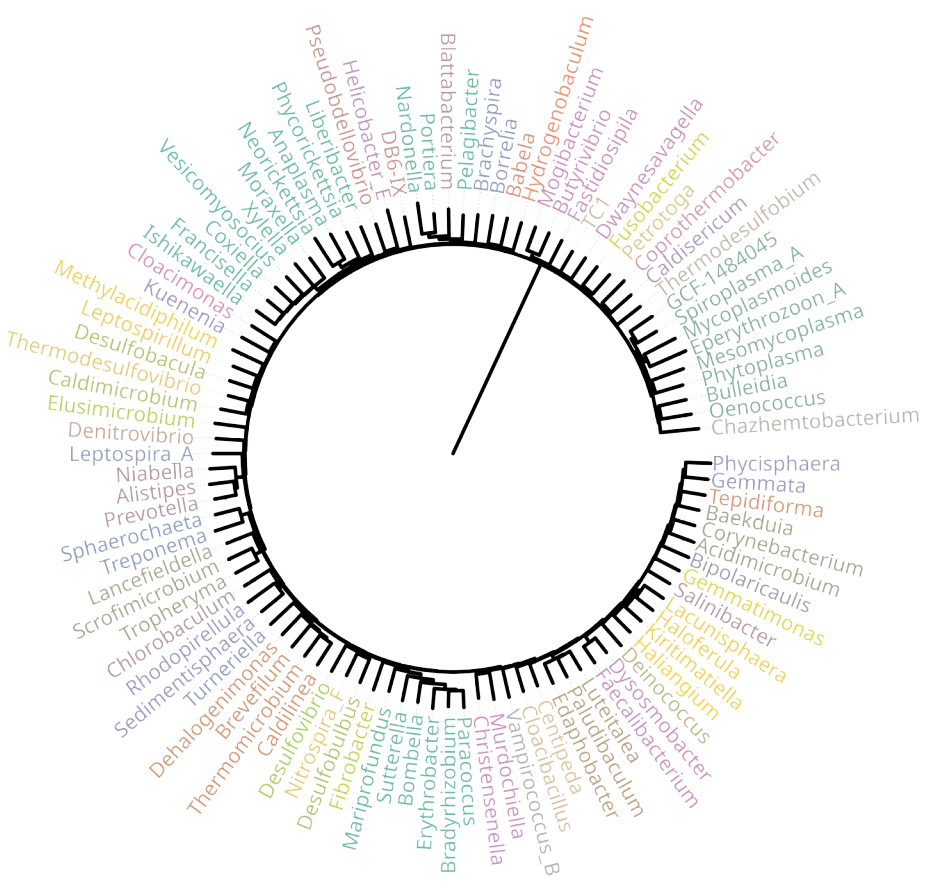


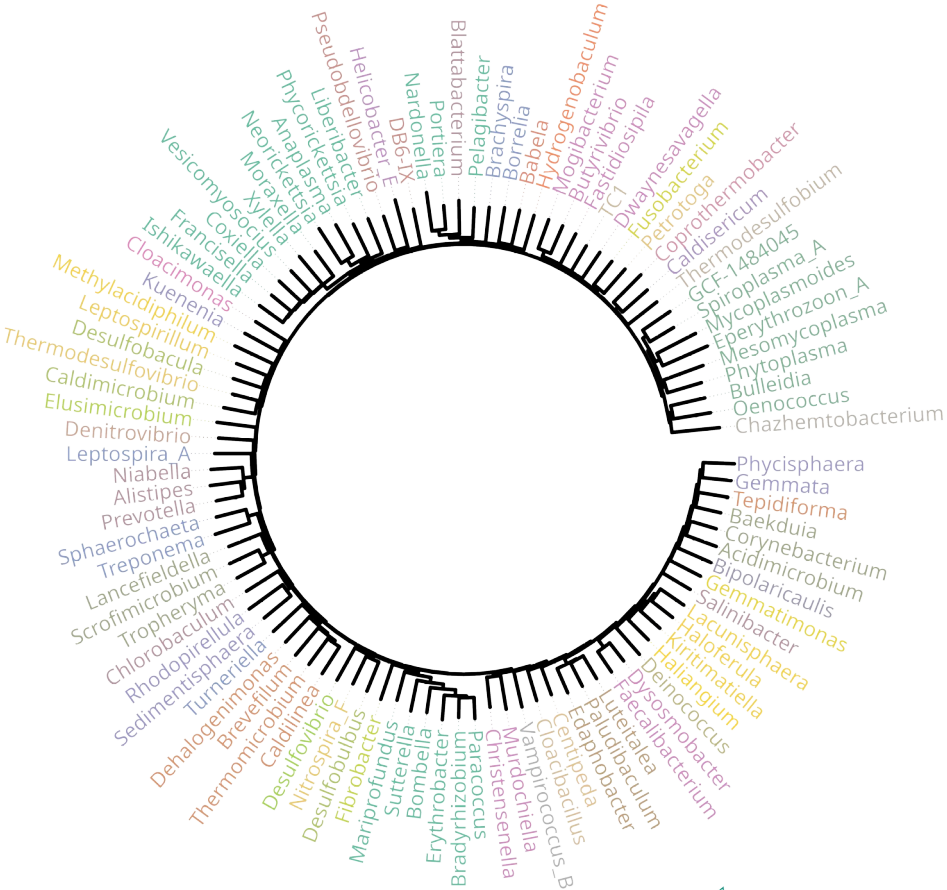
3



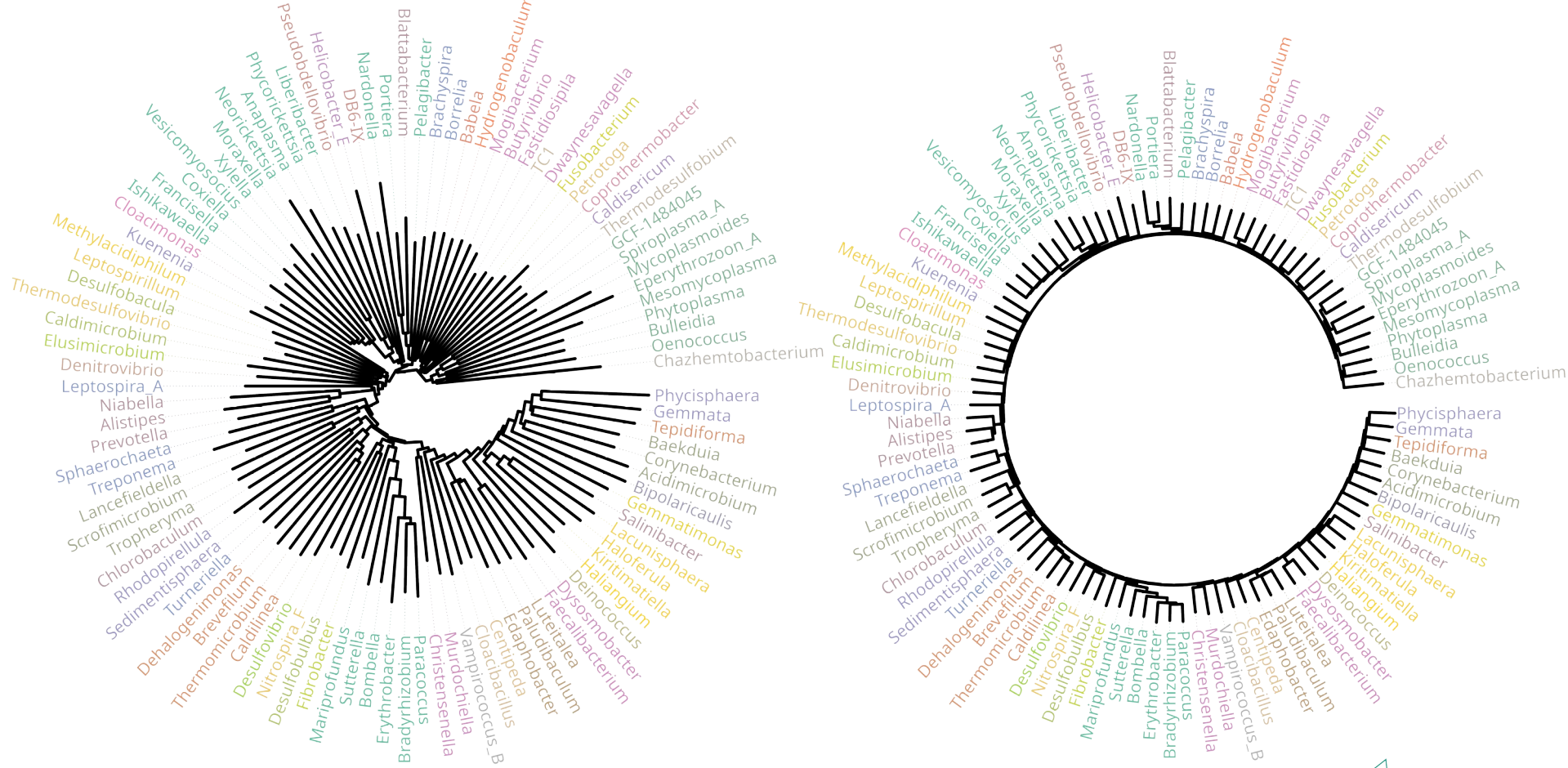
Quadram Institute
Science & Health & Food & Innovation

What’s wrong with these plots?





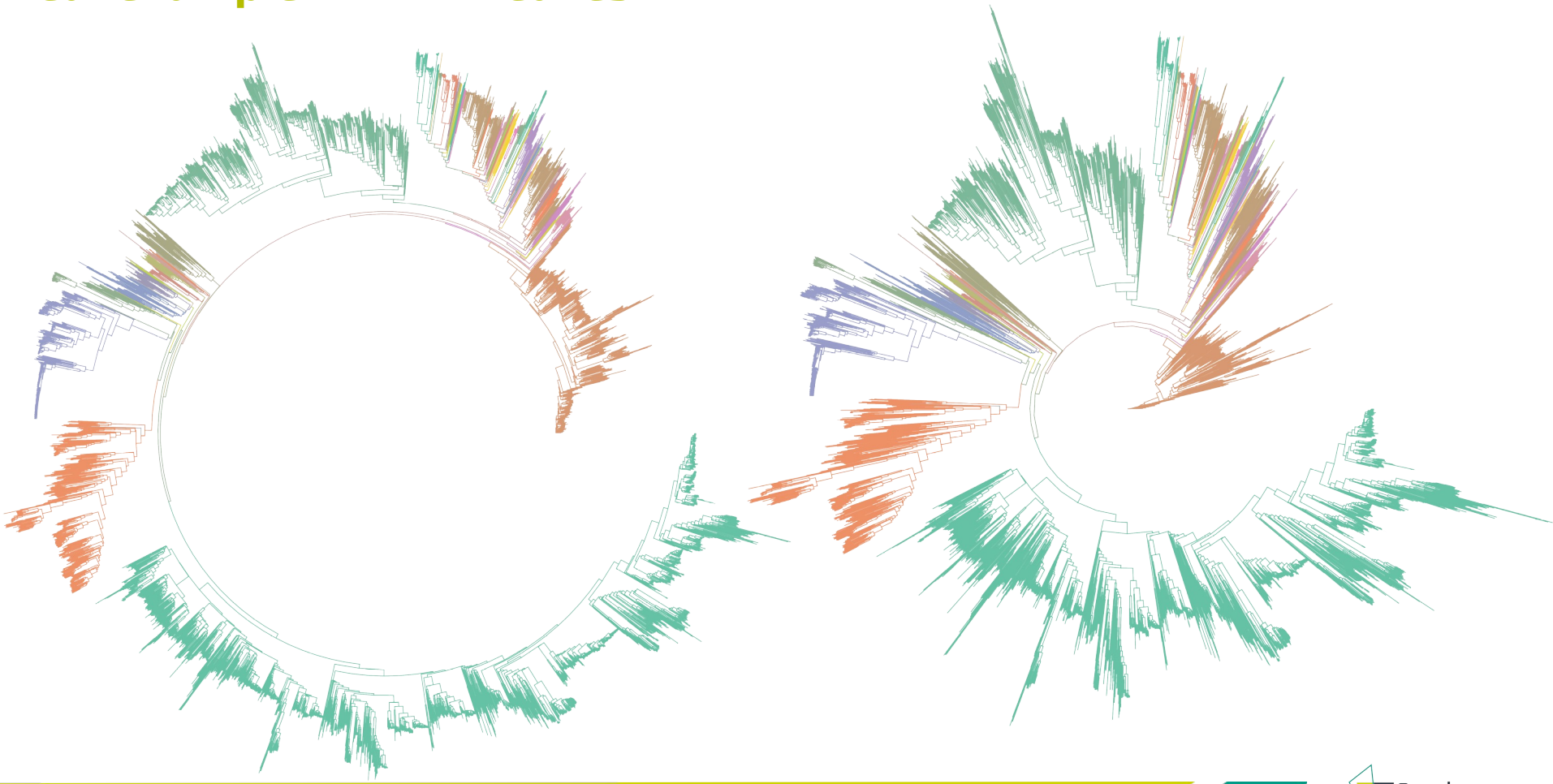
What's wrong with these plots? Wasting space with the root!



ggtree() + xlim(0.49,NA)

Quadram Institute
Science • Health • Food • Innovation

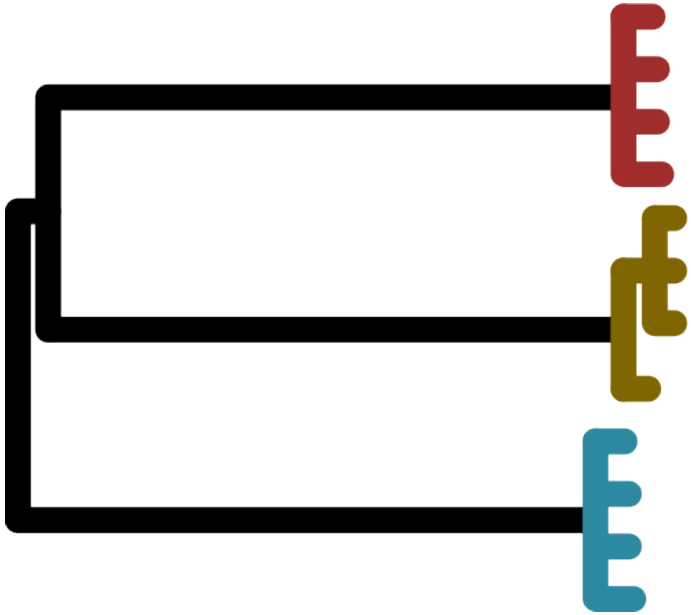
Real example with 4k leaves

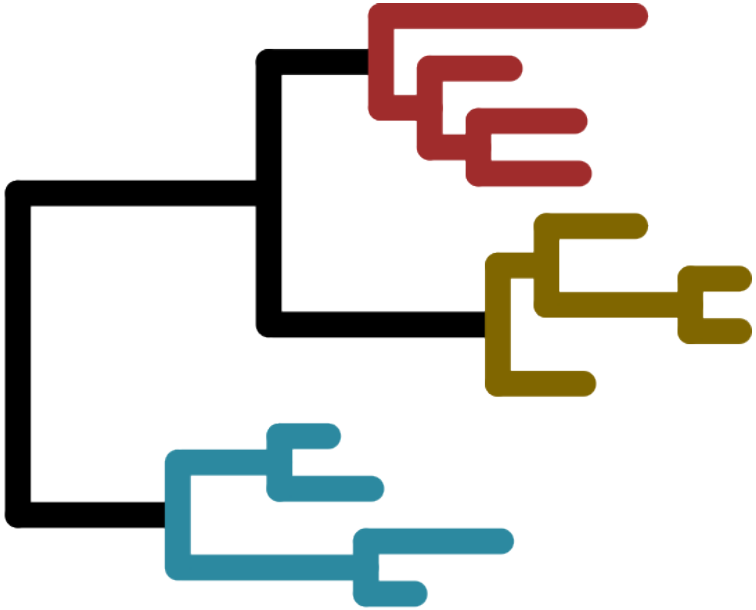


Phylogenetics is not classification

Quadram Institute
Science Health Food Innovation

A good classification dendrogram is not a good phylogeny





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