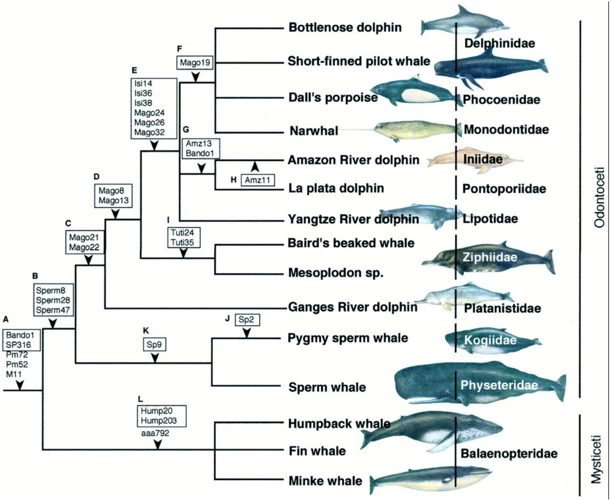
Phylogeny

A phylogenetic tree is a graphical representation depicting the evolution of some set of organisms or taxa.

Often these organisms are distinct species, and the leaves of the tree have one distinct species per leaf.

Mathematically, we can think of a phylogeny as a graph that is a tree where the leaves represent current or most recently observed species (could be fossils), and other vertices represent hypothetical ancestors. Typically, we will use rooted (i.e., directed) binary trees. All of these assumptions can be relaxed. The tree on the next slide is rooted but non-binary.


F
Mago19
E
Isi14
Isi36
Isi38
Mago24
Mago26
Mago32
G
Amz13
Bando1
H
Amz11
D
Mago8
Mago13
C
Mago21
Mago22
I
Tuti24
Tuti35
B
Sperm8
Sperm28
Sperm47
A
Bando1
SP316
Pm72
Pm52
M11
K
Sp9
J
Sp2
L
Hump20
Hump203
aaa792
Bottlenose dolphin
Short-finned pilot whale
Dall's porpoise
Narwhal
Amazon River dolphin
La plata dolphin
Yangtze River dolphin
Baird's beaked whale
Mesoplodon sp.
Ganges River dolphin
Pygmy sperm whale
Sperm whale
Humpback whale
Fin whale
Minke whale
Delphinidae
Phocoenidae
Monodontidae
Iniidae
Pontoporiidae
Lipotidae
Ziphiidae
Platanistidae
Kogiidae
Physeteridae
Balaenopteridae
Odontoceti
Mysticeti

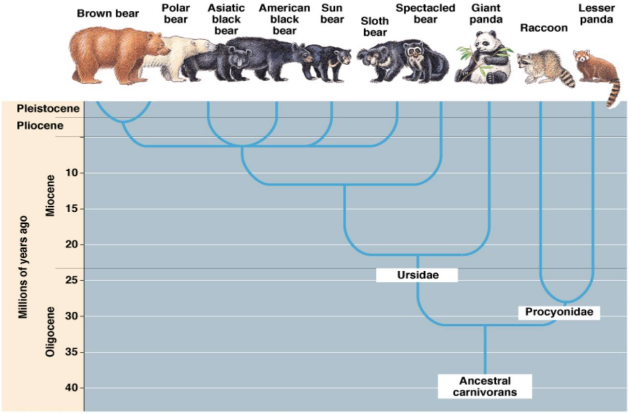
What do you notice about the tree on the previous slide?

Source for previous slide: Nikaido, Matsuno, Hamilton, Okada (2001). Retroposon analysis of major cetacean lineages: The monophyly of toothed whales and the paraphyly of river dolphins. *PNAS* 98 (13) 7384-7389.

According to the phylogeny, are sperm whales more closely related to humpback whales or to bottlenose dolphins?

Do dolphins form a group? What do I mean by a group?

As another example, are Pandas bears?



Phylogenetic relationships of the Ursidae & Procyonidae

https://www.mun.ca/biology/scarr/Phylogeny_of_Ursidae.html

According to the previous slide, giant pandas are more closely related to say, polar bears, than to lesser pandas (red pandas). For some terminology:

The **most recent common ancestor** (MRCA) of a set of taxa (species or lineages) is the most recent vertex (or ancestor) on the graph that is ancestral to all taxa in the set.

A set of species (or lineages) is called **monophyletic** if the MRCA of the set does not include any descendants not in the set.

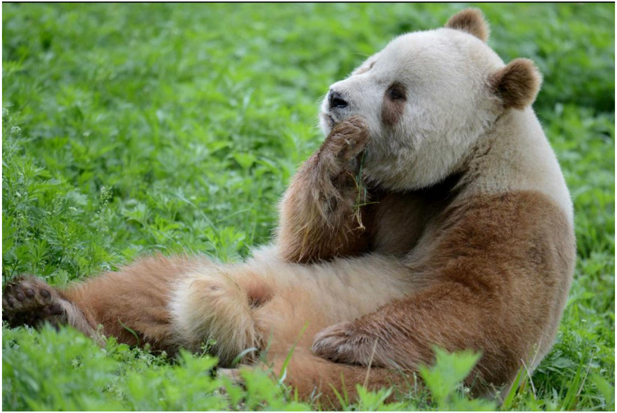
Consequently, what we think of as whales are not a monophyletic group since their MRCA includes dolphins. Bears, however, are monophyletic. A group consisting of Pandas, raccoons, and red pandas, would not be monophyletic, since their MRCA includes all current bear species.

A group A is **paraphyletic** with respect to another group B if some members of A are more closely related to members of B than to other members of A.

Two sets of species, A and B are called **polyphyletic** if there are paraphyletic with respect to each other. I.e., the MRCA of group A includes members of B, and the MRCA of group B includes members of group A.

One application of phylogenetics is to classify species. Some evolutionary biologists have argued that classification of species and higher-order groups (genus, order, family, etc.) should respect phylogenetic trees. If this view is adopted, "bear" is a perfectly good category for a group of species, but "whale" is not.

The concepts of monophylly, paraphyly, and polyphyly can be applied at other levels than the species level. For example, you might want to know if a subset of organisms in a sample is monophyletic, or whether a subspecies is monophyletic. Two subspecies of panda include the *Ailuropoda melanoleuca melanoleuca* (black and white) and *Ailuropoda melanoleuca qinlingensis* (brown)



https://en.wikipedia.org/wiki/Giant_panda#/media/File:Qinling_panda.jpg

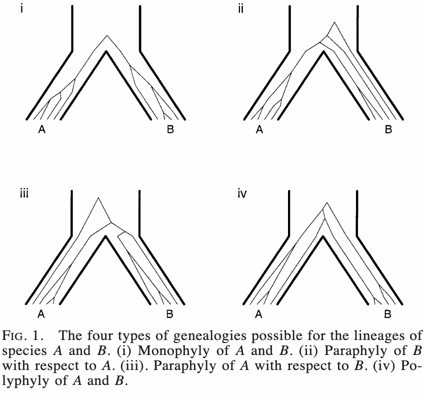


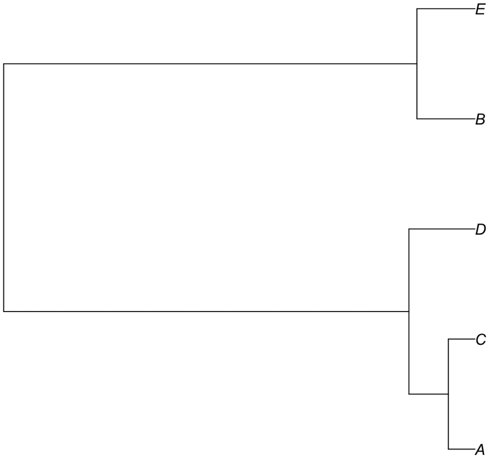
FIG. 1. The four types of genealogies possible for the lineages of species *A* and *B*. (i) Monophyly of *A* and *B*. (ii) Paraphyly of *B* with respect to *A*. (iii). Paraphyly of *A* with respect to *B*. (iv) Polyphyly of *A* and *B*.

https://web.stanford.edu/group/rosenberglab/papers/mono.pdf

The word **taxa** is often used instead of species because the leaves of a tree could represent individual species but could also represent genera, subspecies, or individiuals from the same population. The word taxa avoids worry about species concepts, which can be also be controversial. For example, polar bears and brown bears are typically considered distinct species even though they can interbreed.

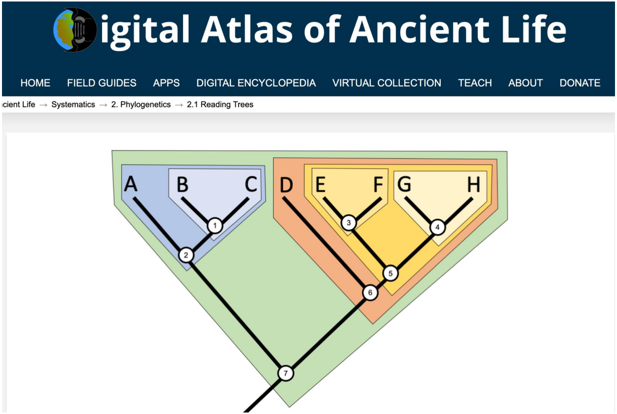
A subset of taxa (leaves) on a tree is also called a **clade** if it is monophyletic. Note that monophyly is relative to the other taxa considered in the study or the tree. Going back to the first phylogeny, if we considered only bottlenose dolphins, Amazon river dolphins, Humpbacks and Fin whales, then dolphins and whales would both be monophyletic, but including more species leads to a more complicated picture.

One way of thinking about a tree is to list all of its clades. Mathematicians at least normally think of a clade as just the subset of extant taxa that is monophyletic.



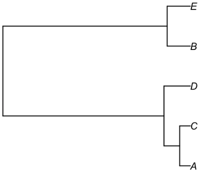
The clades on this tree are $\{A, C\}, \{A, C, D\}, \{B, E\}$ as well as $\{A\}, \{B\}, \{C\}, \{D\}, \{E\}$ and $\{A, B, C, D, E\}$. Typically we think of the non-trivial clades of tree with n leaves as those clades that have sizes strictly greater than 1 and strictly less than n. Clades are also called **clusters**, similar to how statisticians refer to clusters in cluster analysis.

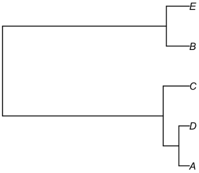
Biologists might refer to clades so as to include their (hypothetical) ancestors up to and including the MRCA, as shown on the next slide.

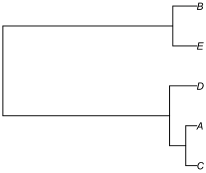

igital Atlas of Ancient Life
HOME
FIELD GUIDES
APPS
DIGITAL ENCYCLOPEDIA
VIRTUAL COLLECTION
TEACH
ABOUT
DONATE
cient Life → Systematics → 2. Phylogenetics → 2.1 Reading Trees
A
B
C
D
E
F
G
H
1
2
3
4
5
6
7

https://www.digitalatlasofancientlife.org/learn/systematics/phylogenetics/reading-trees/

Thinking about a tree as a set of clades is also useful for deciding when two trees are the same. Consider the following three trees. Are they all distinct, or are any of them the same?


E
B
D
C
A


E
B
C
D
A


B
E
D
A
C

It's difficult to tell immediately just by visual inspection which two are the same. The two on the left are the same. The one on the right is different. Why? One answer is that the two on the left have the same clades. The one the right has clade $\{A, D\}$ whereas the two on the left have $\{A, C\}$.

For larger trees (imagine, 50, 100, 1000 leaves), it can be difficult to compare two trees. In graph theory, the problem of determining whether two graphs are identical is the problem of *graph isomorphism*. In phylogenetics, we are also comparing different graphs, but an important difference from typical graphs in graph theory is that the internal (non-leaf) vertices are not labeled.

A typical way to represent trees so that they can be input into computer programs is called the “Newick” form (named after a restaurant in Maine where a bunch of evolutionary biologists decided to standardize the notation). Basically, clusters are grouped using nested parentheses. The trees on the right are

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

And the tree on the right is

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

The bottom tree seems to be written as

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

But that is equivalent to the first representation.

For the bear example, relationships between brown bears, black bears, and pandas can be represented as

```
((brown, black), panda)
```

In this course we'll often simulate trees or simulate DNA data and then infer trees from the simulated data. We might also infer data from empirical DNA sequences. When we do so, computer programs can represent the tree in this Newick format. Unfortunately, the Newick representation of a tree is not unique, and this leads to a lot of complications when trying to analyze results. We often need to check if two trees are identical to determine if a tree inferred from data is correct or is equivalent to an assumed tree.

The following will be some R code for manipulating trees in R using the `APE` package. Try the code yourself to start getting the hang of working with trees and plotting them in R.

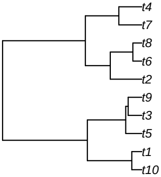
```
rm(list = ls()) # clear the workspace
library(ape)
library(TreeSim) # simulates trees

#set a seed to be able to replicate results
set.seed(2025)

#generate 4 trees with 6 taxa using birth rate 1, extinction r
trees <- sim.bd.taxa(6,4,1,0)

#be able to make a 2x2 array of plots
par(mfrow=c(2,2))

#here trees is a list rather than a vector and needs this
#double bracket notation to access each of the four trees
plot.phylo(trees[[1]])
plot.phylo(trees[[2]])
plot.phylo(trees[[3]])
```

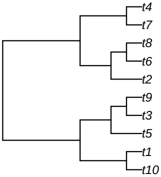
Trees can be plotted in different styles and with different options.

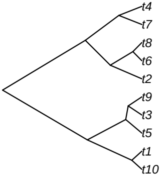
```
plot.phylo {ape} R Documentation
Plot Phylogenies
Description
These functions plot phylogenetic trees.

Usage
## S3 method for class 'phylo'
plot(x, type = "phylogram", use.edge.length = TRUE,
     node.pos = NULL, show.tip.label = TRUE,
     show.node.label = FALSE, edge.color = NULL, edge.width
     = NULL, edge.lty = NULL, node.color = NULL, node.width
     = NULL, node.lty = NULL, font = 3, cex = par("cex"),
     adj = NULL, srt = 0, no.margin = FALSE, root.edge =
     FALSE, label.offset = 0, underscore = FALSE, x.lim =
     NULL, y.lim = NULL, direction = "rightwards", lab4ut =
     NULL, tip.color = par("col"), plot = TRUE, rotate.tree
     = 0, open.angle = 0, node.depth = 1, align.tip.label =
     FALSE, ...)
## S3 method for class 'multiPhylo'
```

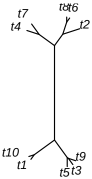
```
plot(x, layout = 1, ...)
Arguments
x
an object of class "phylo" or of class "multiPhylo".

type
a character string specifying the type of phylogeny to be drawn;
it must be one of "phylogram" (the default), "cladogram", "fan",
"unrooted", "radial", "tidy", or any unambiguous abbreviation of
these.

use.edge.length
a logical indicating whether to use the edge lengths of the
phylogeny to draw the branches (the default) or not (if FALSE).
This option has no effect if the object of class "phylo" has
no ‘edge.length’ element.
```

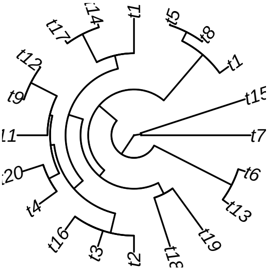
```
x <- sim.bd.taxa(10,1,1,0)
par(mfrow=c(3,2))
plot.phylo(x[[1]],type="phylogram",use.edge.length=TRUE)
plot.phylo(x[[1]],type="phylogram",use.edge.legnth=FALSE)
plot.phylo(x[[1]],type="cladogram")
plot.phylo(x[[1]],type="unrooted")
plot.phylo(x[[1]],type="fan")
plot.phylo(x[[1]],type="fan",use.edge.lengths=FALSE)
```

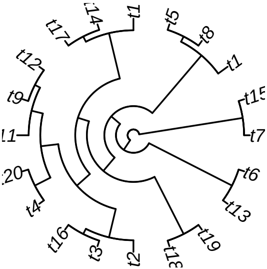

t4
t7
t8
t6
t2
t9
t3
t5
t1
t10


t4
t7
t8
t6
t2
t9
t3
t5
t1
t10


t4
t7
t8
t6
t2
t9
t3
t5
t1
t10


t8
t6
t7
t4
t2
t10
t1
t9
t5
t3

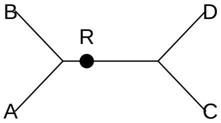

t1
t14
t17
t12
t9
1
20
t4
t16
t3
t2
t18
t19
t13
t6
t7
t15
t1
t8
t5

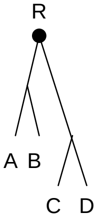

t1
t14
t17
t12
t9
1
20
t4
t16
t3
t2
t18
t19
t13
t6
t7
t15
t1
t8
t5

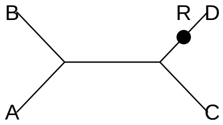
We hadn't talked about unrooted trees before, but I gave an example in the R code. Often software used to infer evolutionary trees only returns an unrooted tree rather than one with a root. For an unrooted tree, the direction of time isn't known but there is still structure in the (unrooted) tree.

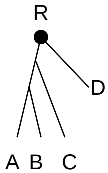
You can get an unrooted tree from a rooted tree by suppressing the root node (vertex) and replacing the two immediate descendants of the root with a single edge. For an unrooted binary tree, all vertices are either degree 1 (leaves) or degree 3 (interior or internal vertices/nodes). For a rooted binary tree, the same is true except that the root has degree 2.

From an unrooted tree, you can get a rooted tree by adding a new vertex on any edge and calling that the root. Therefore, for each n-taxon unrooted, binary tree, there are $(2n - 3)$ rooted trees with the same unrooted topology.


B
R
D
A
C


R
A
B
C
D


B
R
D
A
C


R
D
A
B
C

In interesting counting problem is to determine the number of phylogenetic trees. To start we can consider the number of rooted, binary trees where each leaf gets a distinct label. For 3 taxa, there are three tree topologies (branching patterns)

```
((a,b),c),  ((a,c),b),   ((b,c),a)
```

For 4 taxa, there are 15 tree topologies

```
(((a,b),c),d)     (((b,c),a),d)     ((a,b),(c,d))
(((a,b),d),c)     (((b,c),d),a)     ((a,c),(b,d))
(((a,c),b),d)     (((b,d),a),c)     ((a,d),(b,c))
(((a,c),d),b)     (((b,d),c),a)
(((a,d),b),c)     (((c,d),a),b)
(((a,d),c),b)     (((c,d),b,)a)
```

Of these, 12 are unbalanced, and 3 are balanced, meaning roughly that each ancestor has the same number of left- and right-descendants.

From Felsenstein 1978, Systematic Zoology

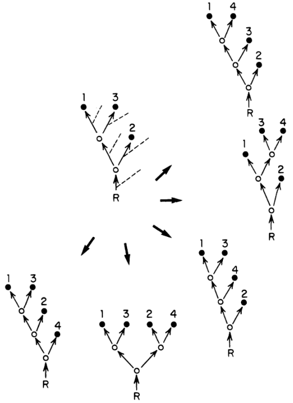


FIG. 2.—All ways in which a fourth species can be added to a given 3-species bifurcating tree so as to result in a bifurcating tree.

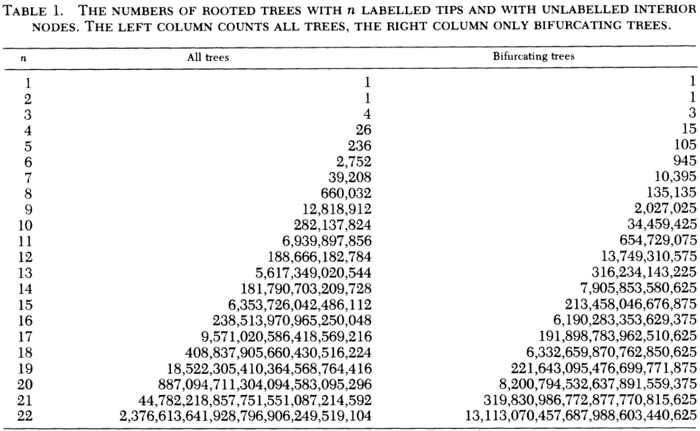


TABLE 1. THE NUMBERS OF ROOTED TREES WITH *n* LABELLED TIPS AND WITH UNLABELLED INTERIOR NODES. THE LEFT COLUMN COUNTS ALL TREES, THE RIGHT COLUMN ONLY BIFURCATING TREES.

n	All trees	Bifurcating trees
1	1	1
2	1	1
3	4	3
4	26	15
5	236	105
6	2,752	945
7	39,208	10,395
8	660,032	135,135
9	12,818,912	2,027,025
10	282,137,824	34,459,425
11	6,939,897,856	654,729,075
12	188,666,182,784	13,749,310,575
13	5,617,349,020,544	316,234,143,225
14	181,790,703,209,728	7,905,853,580,625
15	6,353,726,042,486,112	213,458,046,676,875
16	238,513,970,965,250,048	6,190,283,353,629,375
17	9,571,020,586,418,569,216	191,898,783,962,510,625
18	408,837,905,660,430,516,224	6,332,659,870,762,850,625
19	18,522,305,410,364,568,764,416	221,643,095,476,699,771,875
20	887,094,711,304,094,583,095,296	8,200,794,532,637,891,559,375
21	44,782,218,857,751,551,087,214,592	319,830,986,772,877,770,815,625
22	2,376,613,641,928,796,906,249,519,104	13,113,070,457,687,988,603,440,625

A formula for the number of rooted binary trees with n taxa is

$$n!! = 1 \times 3 \times \cdots \times (2n-3)$$

If the trees are unrooted, the formula is

$$(n-1)!! = 1 \times 3 \times \cdots \times (2n-5)$$

Strange note for statisticians: If Z is standard normal, then

$$E[Z^{2n}] = (2n-3)!!$$

where n is a positive integer.