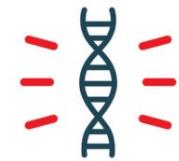


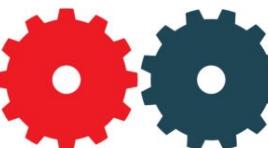
BIOINFORMATICS



BIOLOGY



COMPUTER SCIENCE



INFORMATION ENGINEERING



$$f(x)$$

MATHEMATICS



STATISTICS

We will draw from our understanding of DNA, RNA, and proteins and how they contribute to cellular processes

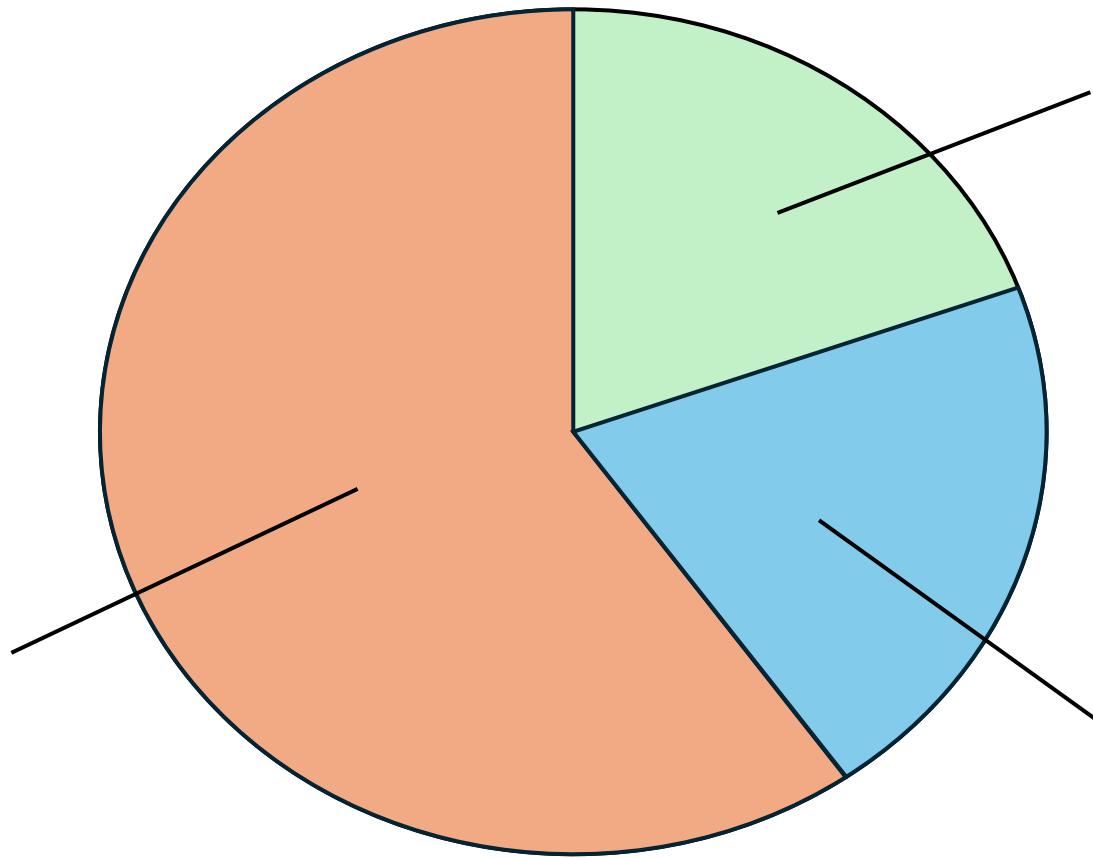
We will learn to use coding and command line programs to analyze biological data

We will learn how to store, process, and share large datasets

We will employ algorithms that rely heavily on mathematics

We will learn to use statistics to help us understand the patterns and trends in our analyses

1000 total points



Weekly project write-ups:

- 1 per week for 6 weeks (week 3-8)
- 100 points each = 600 total points
- A larger applied analysis of a real biological dataset
- Assigned Thursday of every week

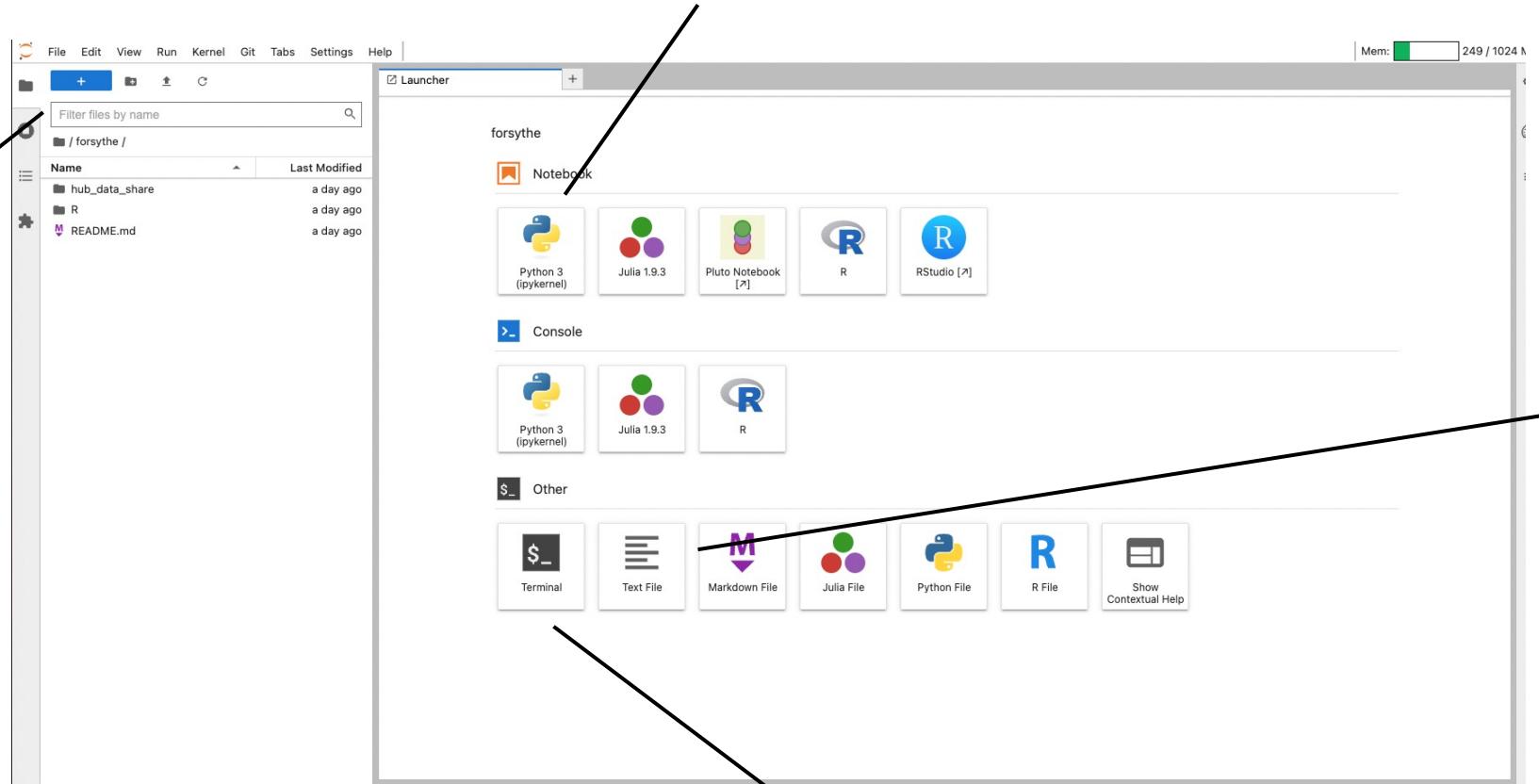
Tutorial assignments:

- 1 per week
- 20 points each = 200 total points
- Will accompany in-class tutorials/exercises
- Assigned Tuesday of each week

Final project:

- Choose from projects covered in week 9
- 200 total points
- Will yield a publication-quality analysis and figure(s)

Opens a python notebook for testing python code



Opens a new
'launcher' window,
which can be used
to open different
types of tools

Opens a text
file for creating
scripts (e.g.
python scripts)

Opens a command line terminal
for navigating the file system or
running linux/unix commands

Indicates a unix command being run from the command line. Do not include the “\$” in your command

The cp command

Two ‘arguments’ for the cp command. Arguments are separated by spaces

```
$ cp original_file.txt new_copy.txt
```

```
$ cp <your original file> <name of copy file>
```

Text surround by <> means this is info you need to fill in yourself. Do not include the “<” and “>” in your command

BioInformatics with python™



UNIX



LINUX

How do computers work?

Computer: an electronic device for storing and processing **data**

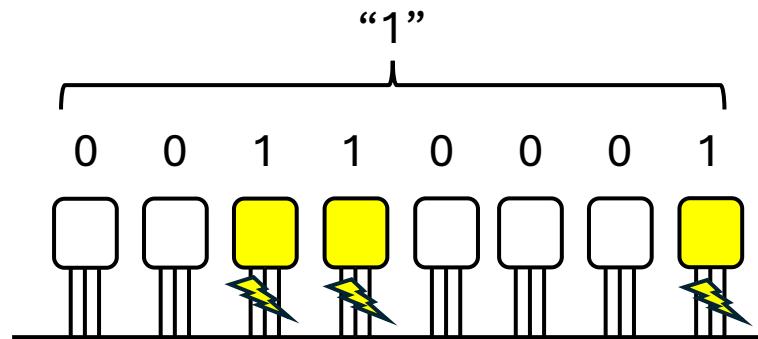
BINARY CODE CONVERTER			
A	01000001	S	01010011
B	01000010	T	01010100
C	01000011	U	01010101
D	01000100	V	01010110
E	01000101	W	01010111
F	01000110	X	01011000
G	01000111	Y	01011001
H	01001000	Z	01011010
I	01001001	0	00110000
J	01001010	1	00110001
K	01001011	2	00110010
L	01001100	3	00110011
M	01001101	4	00110100
N	01001110	5	00110101
O	01001111	6	00110110
P	01010000	7	00110111
Q	01010001	8	00111000
R	01010010	9	00111001

↑
1 bit

8 bits = 1 byte

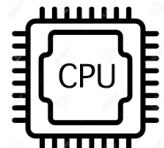
Why is computer data stored in binary?

- Data are stored by electrically manipulating physical components called **transistors**



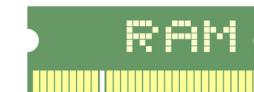
<https://onlinetexttools.com/convert-text-to-binary>

Hardware components of a simplified computer



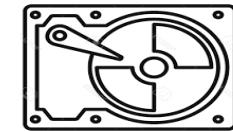
Central processing unit

- Contains 1 or many 'cores'
- Executes processes (e.g. calculate 1+2)



RAM / 'Memory'

- Temporary storage during processing
- (e.g. 00110001 + 00110010 = 00110011)



Harddrive/'Disk'/'Storage'

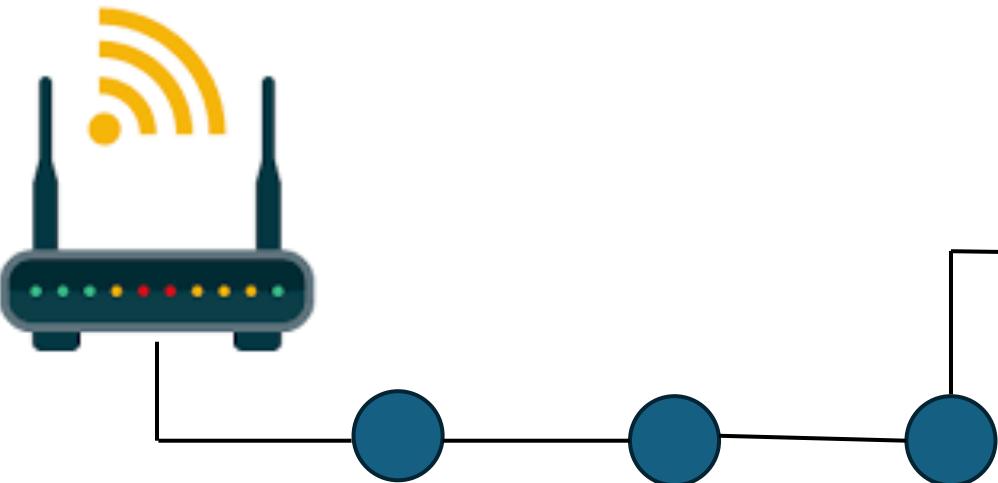
- Data is saved/'written' to disk for long-term storage
- e.g. 00110011 (i.e. "3")

Remote servers and ‘the cloud’

All data need to be physically stored somewhere:

- The data on the internet are physically stored a series of interconnected servers
- When you visit a website, your devices pings a message to a server somewhere in the world and requests access to the data stored there

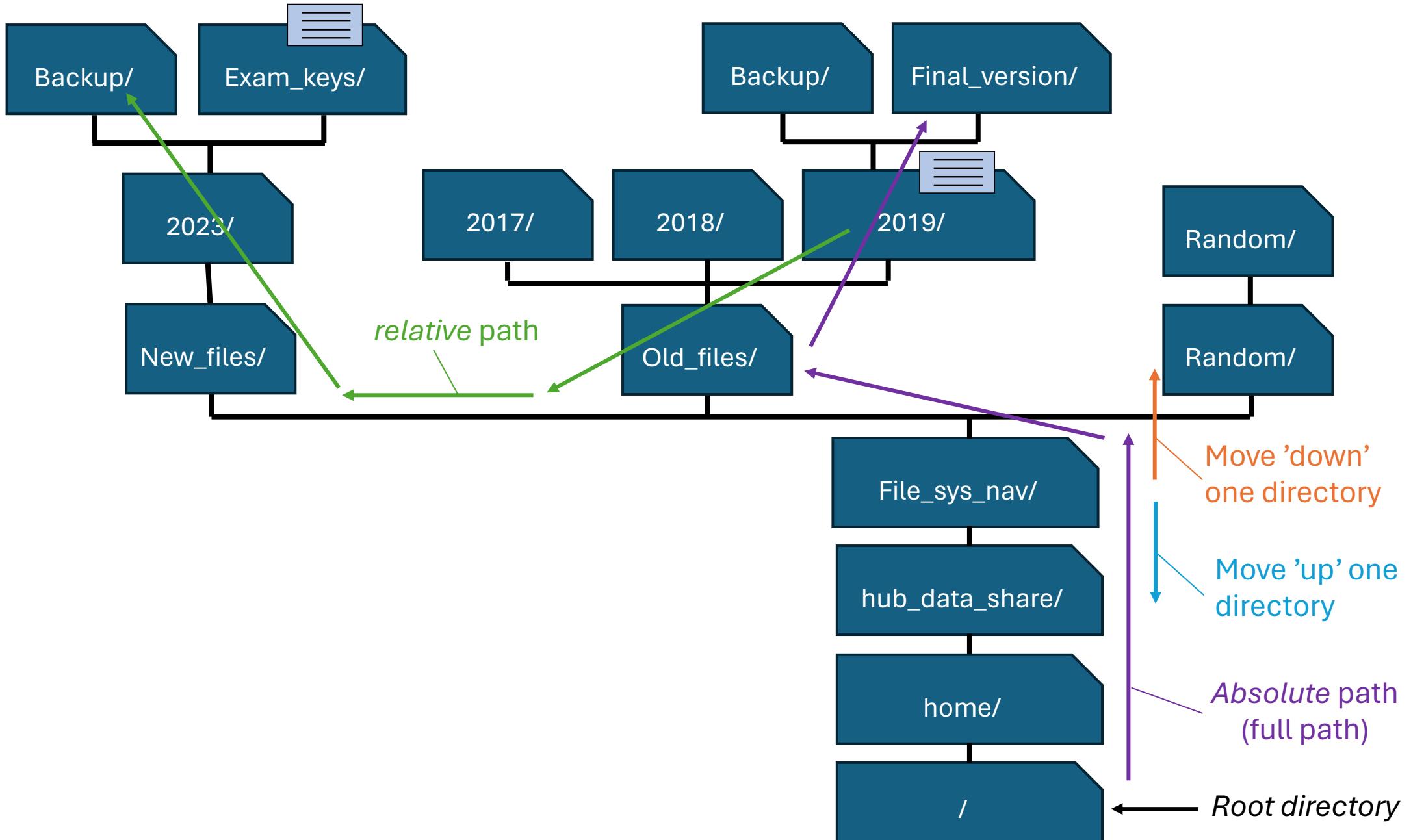
Personal computer:
your ‘local machine’

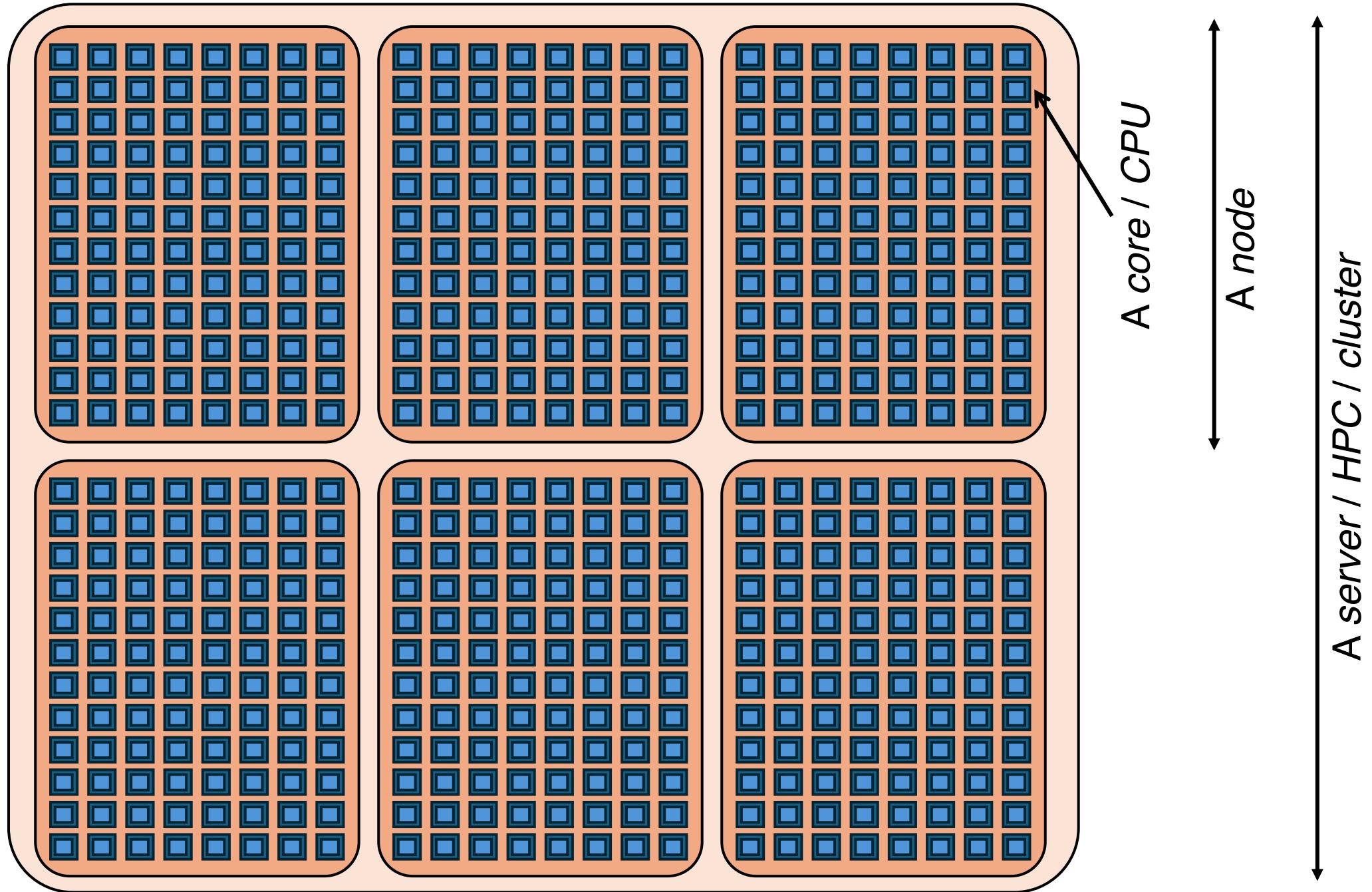


Server (‘remote machine’): big computer with hundreds of CPUs



An example file system

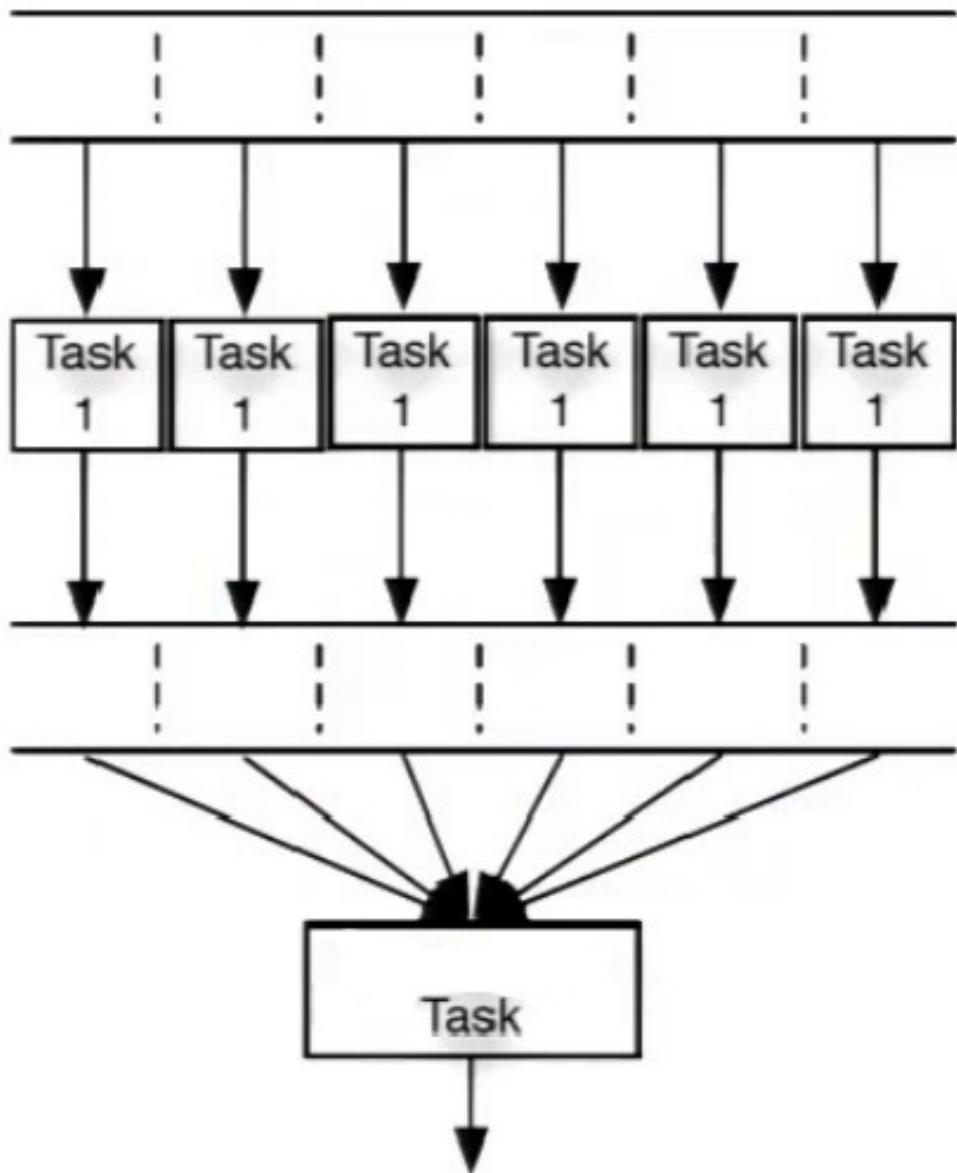




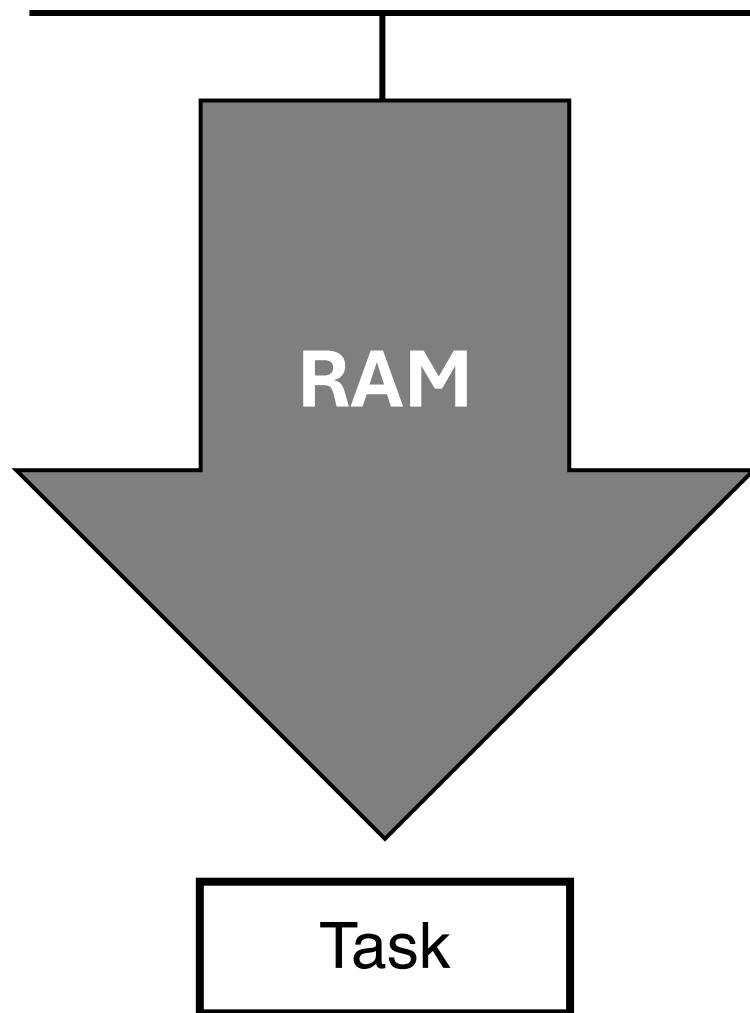


A node

Parallel computing



Memory-intensive computing



Text editing with *vim*

Insert mode: type text
into your document

```
#!/usr/bin/python3

#print statement
print("Hello, world!")

~
~
```

Press “i” →

-- INSERT --

Command mode: enter “:wq”
command to write and save

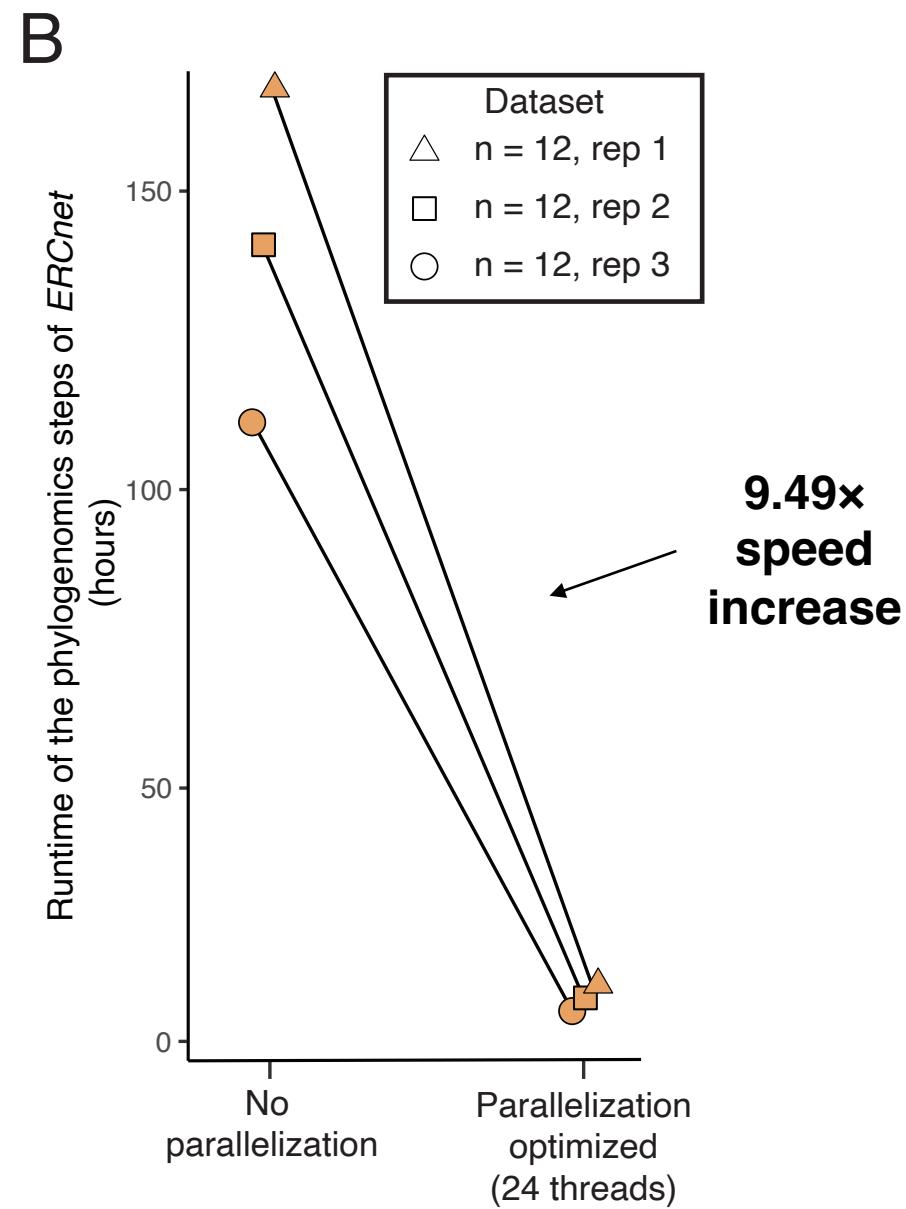
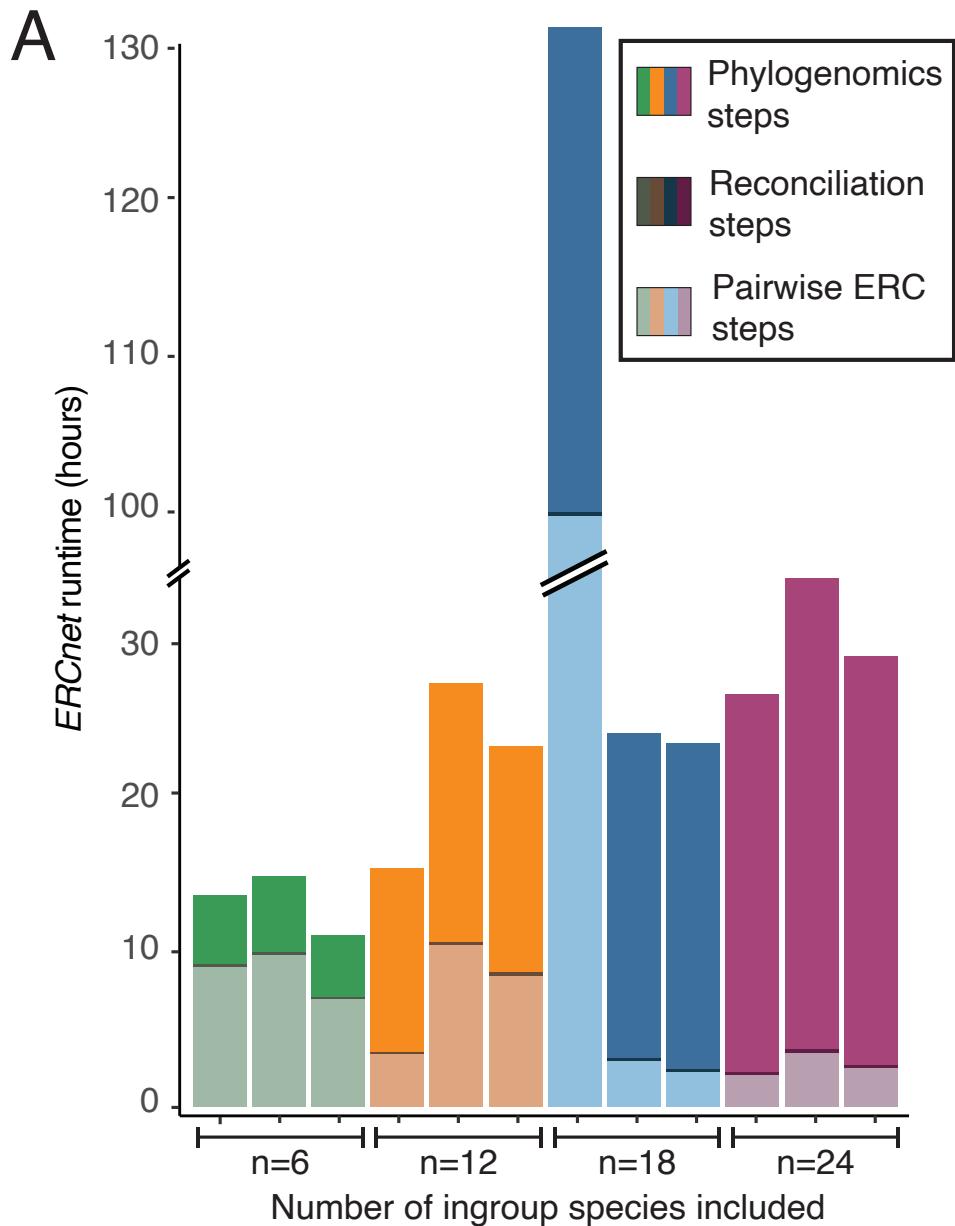
```
#!/usr/bin/python3

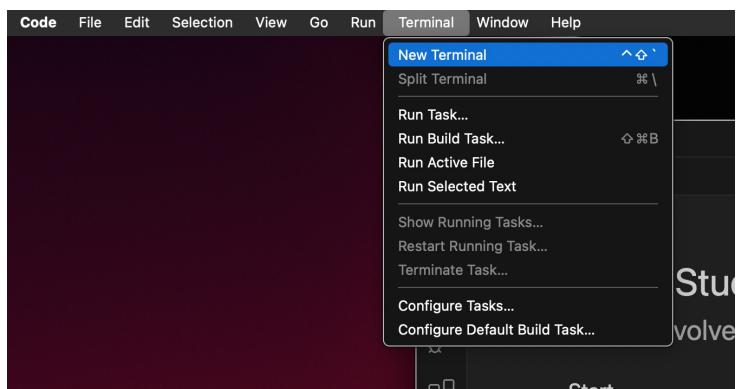
#print statement
print("Hello, world!")

~
~
```

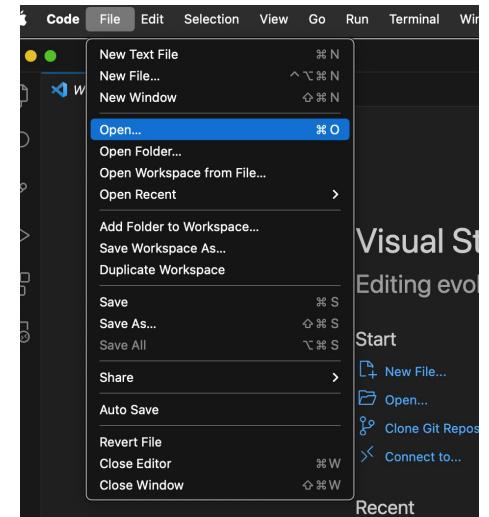
Press “Esc” →

:wq

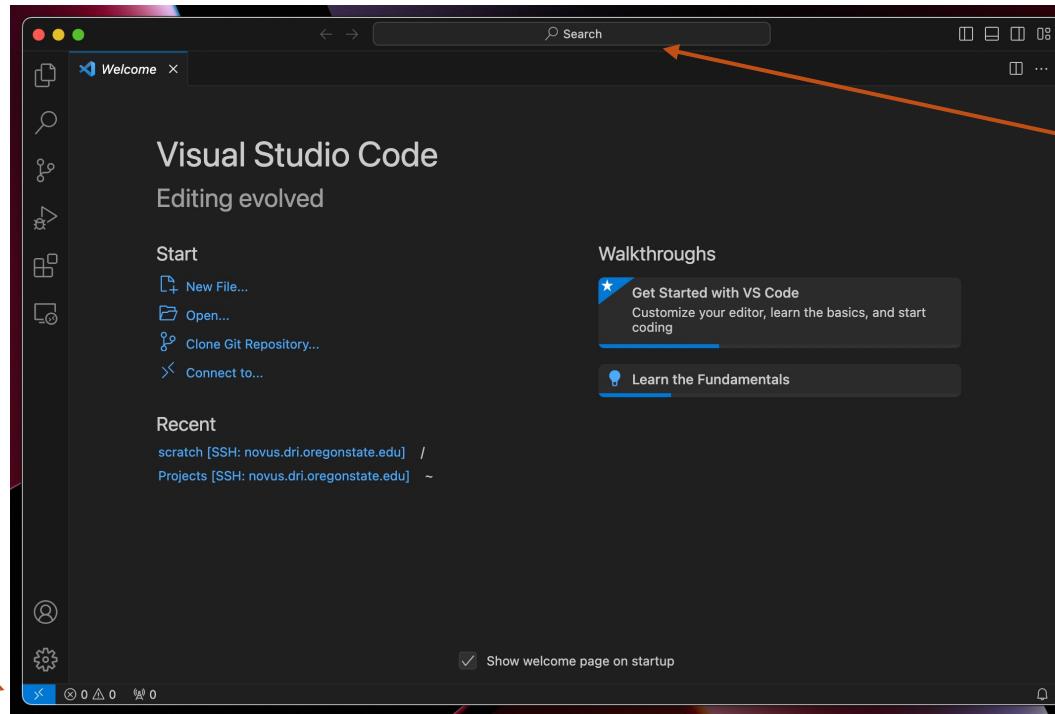




Open a terminal



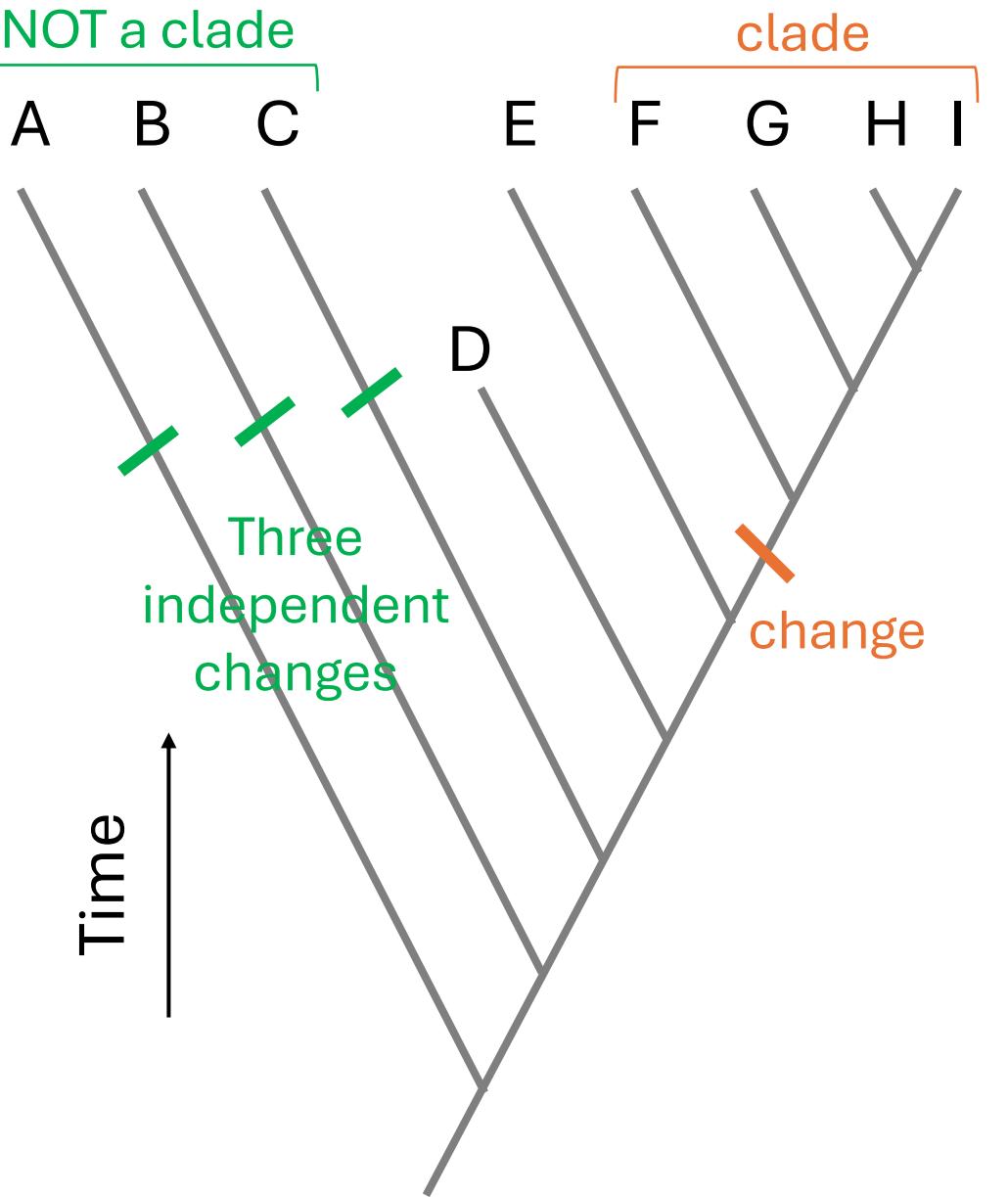
Edit a text file



Connect to HPC via SSH plugin

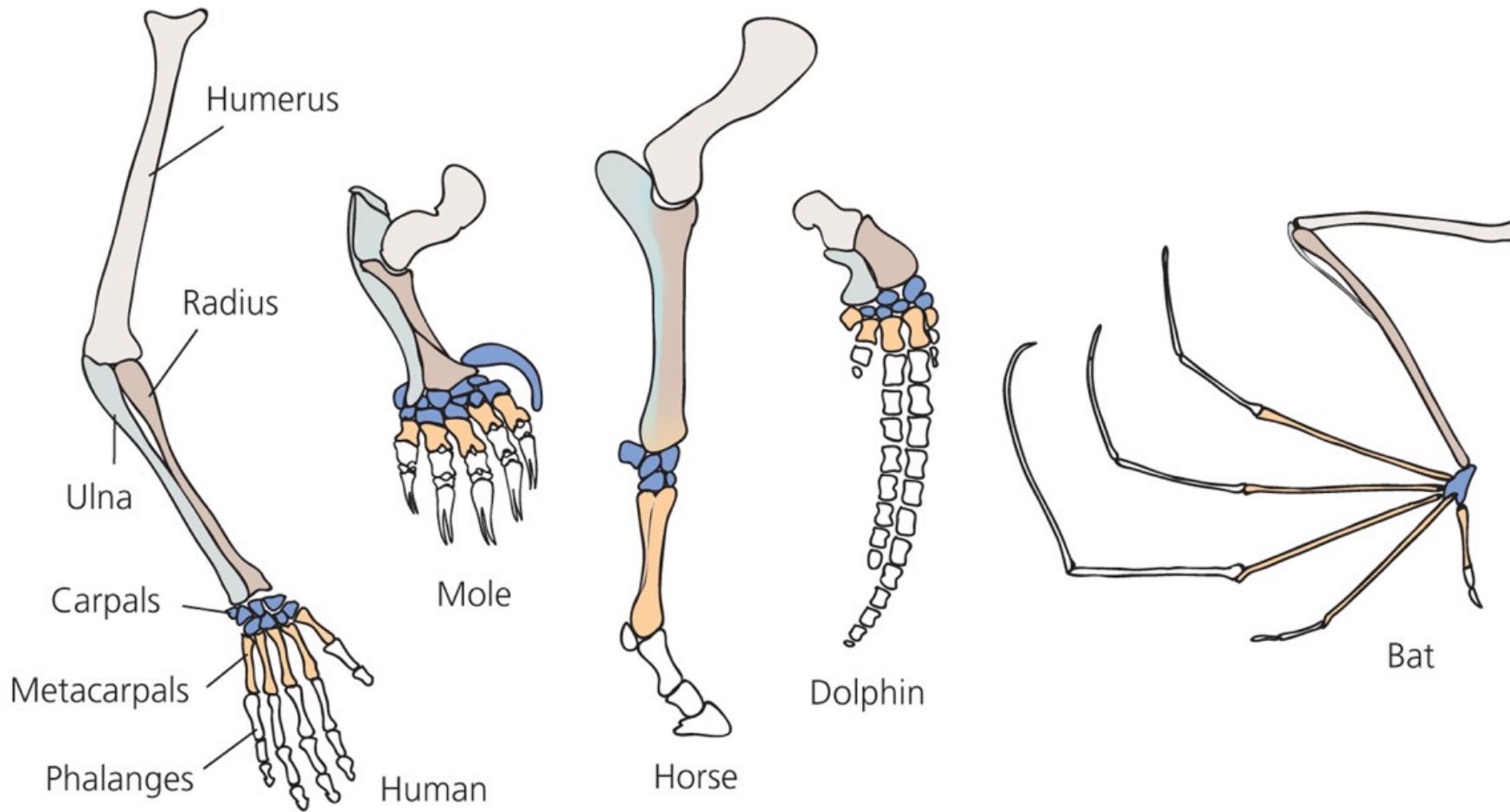
Enter info (e.g. ONID password, SSH commands, folder names, etc...)

Intro to phylogenetic trees



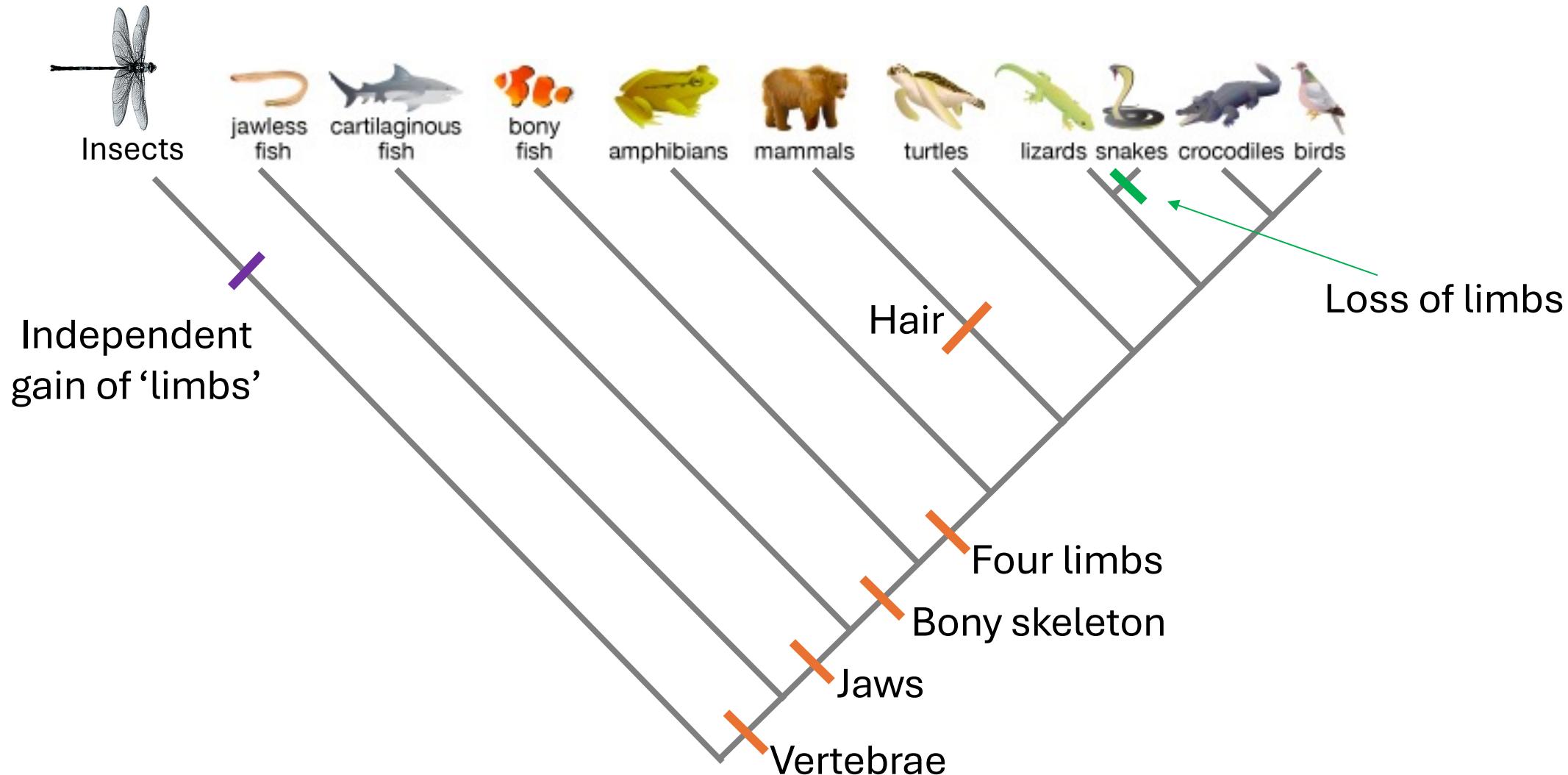
- The tips of the tree represent extant species
 - Extinct species are depicted by truncated branches (e.g. Species D)
- The nodes represent ancestral species that lived in the past
- Branches represent the passage of time
 - Evolutionary changes occur during the time between nodes
- A clade is a group of species that share a common ancestor
 - Clade = monophyletic

Homology



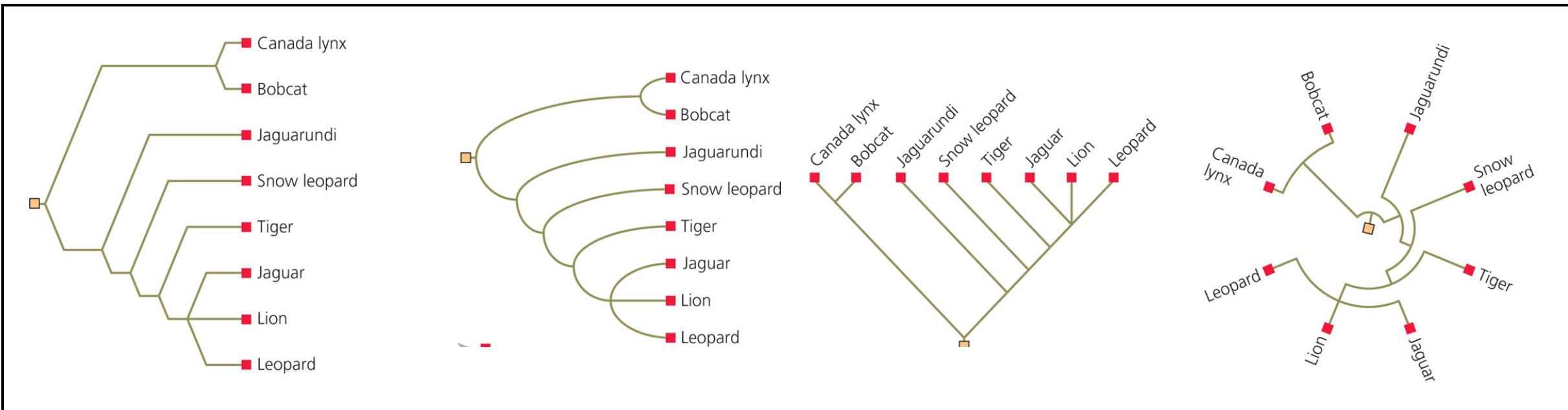
Homology: traits that are shared due to common ancestry

'Tree thinking': framework for understanding when traits evolved



Homology: traits that are shared due to common ancestry

Different ways of visualizing the exact same tree

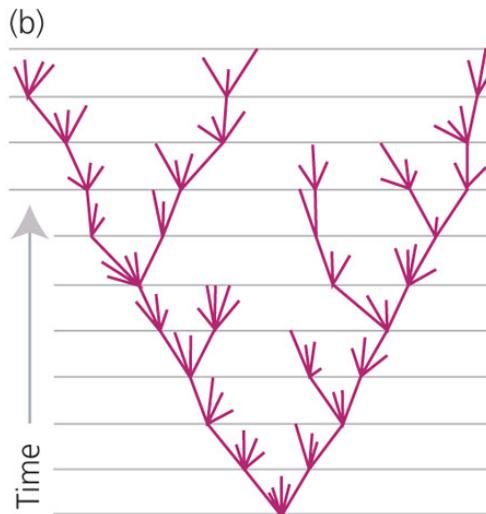
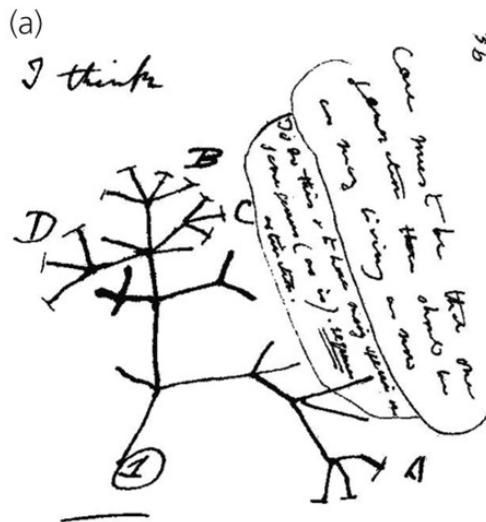


Newick (text format):

(Canada_lynx, Bobcat),(Jaguarundi,(Snow_leopard,(Tiger,(Jaguar, Lion, Leopard)))

Binary format:

```
00101000 01000011 01100001 01101110 01100001 01100100 01100001 01011111 01101100 01111001 01101110 01111000 00101100 00100000  
01000010 01101111 01100010 01100011 01100001 01110100 00101001 00101100 00101000 01001010 01100001 01100111 01110101 01100001  
01110010 01110101 01101110 01100100 01101001 00101100 00101000 01010011 01101110 01101111 01110111 01011111 01101100 01100101  
01101111 01110000 01100001 01110010 01100100 00101100 00101000 01010100 01101001 01100111 01100101 01110010 00101100 00101000  
01001010 01100001 01100111 01110101 01100001 01110010 00101100 00100000 01001100 01101001 01101111 01101110 00101100 00100000  
01001100 01100101 01101111 01110000 01100001 01110010 01100100 00101001 00101001
```



Phylogeny

- summary of a group's evolutionary history
- usually constructed from characters in living taxa
- useful for testing hypotheses about evolutionary history

Character

- Any feature that is scored for each taxon (morphological features, DNA base pairs, protein structure, etc.)
- They should be independent and homologous.
- Can be either ancestral or derived
 - Ancestral – retained from common ancestor
 - Derived – newly-evolved character

Synapomorphy: a shared, derived character

- shared = 2+ taxa have one character state
- derived = novel with respect to ancestral state
- Synapomorphies define clades

Homoplasy: “noise” in the data due to convergent evolution and reversals

- complicates phylogenetic reconstruction
- common in DNA sequence data (only four states: A,T,C,G)
- Solution: use many informative characters

Parsimony:

- Logical argument that the simplest explanation is the best explanation
- Simplest = fewest transitions

Examples of synapomorphies

Synapomorphy

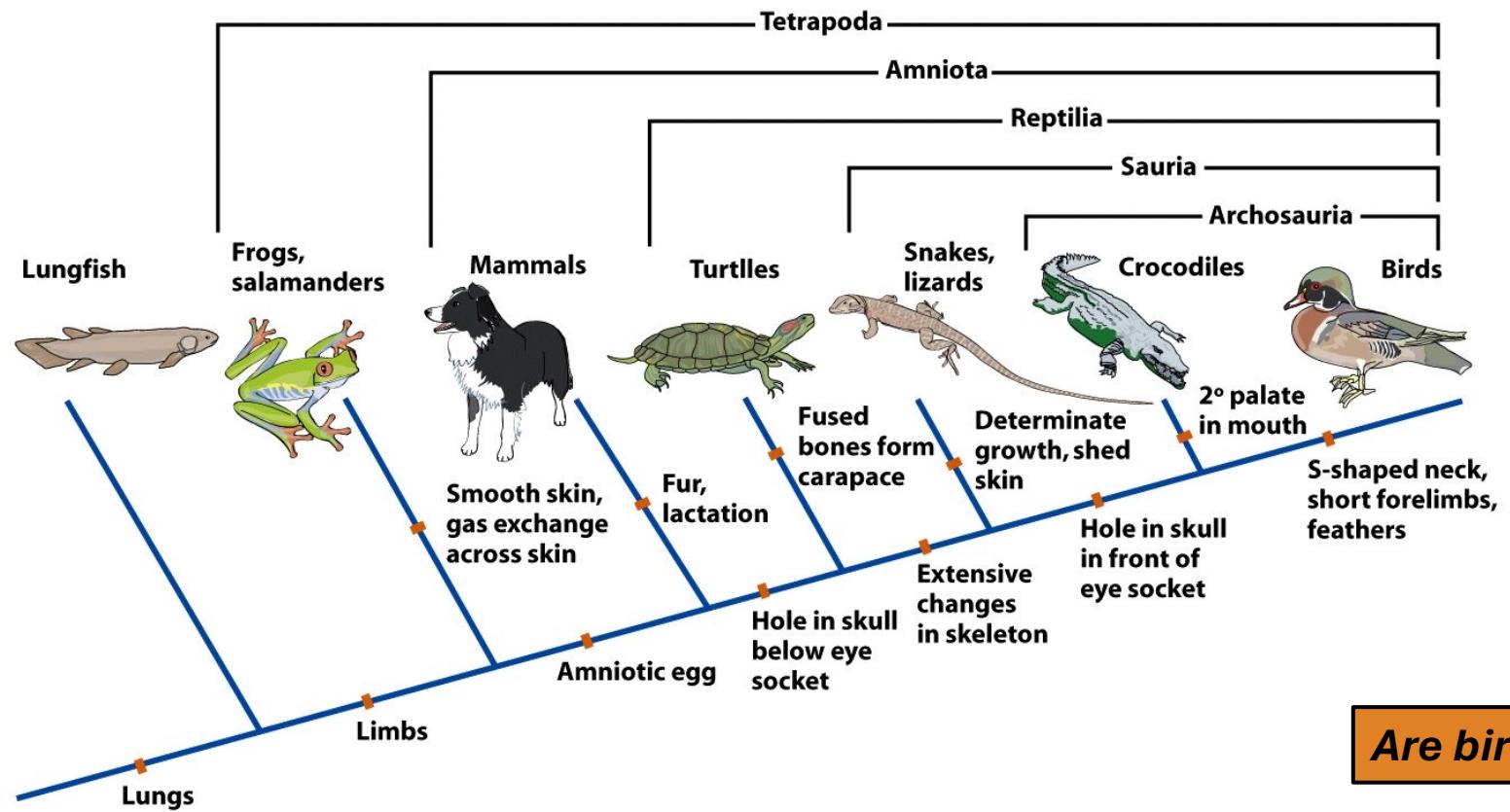
feathers

hole below eye socket

Clade that it defines

Birds

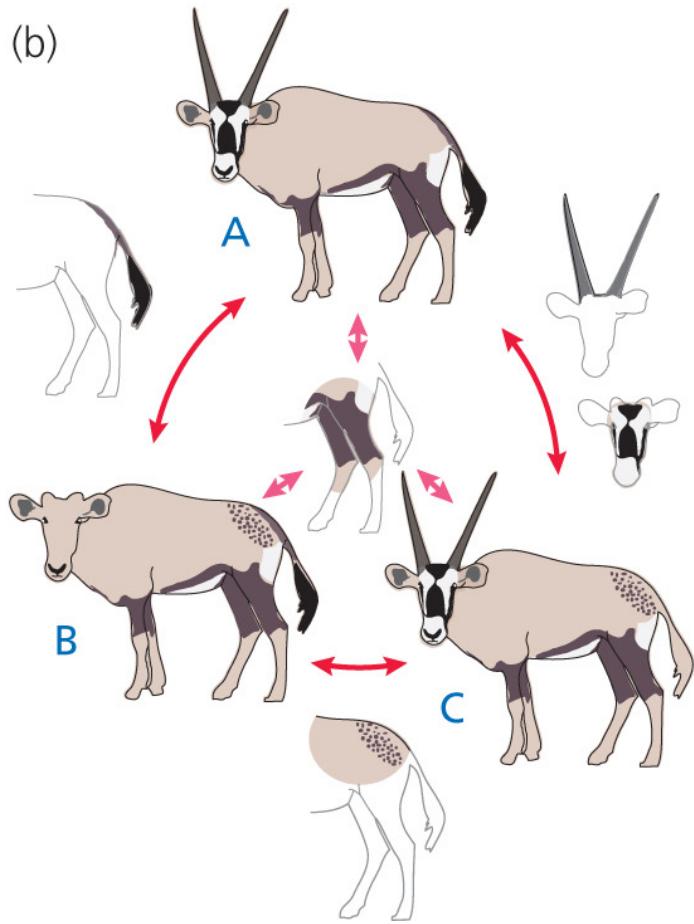
Reptilia



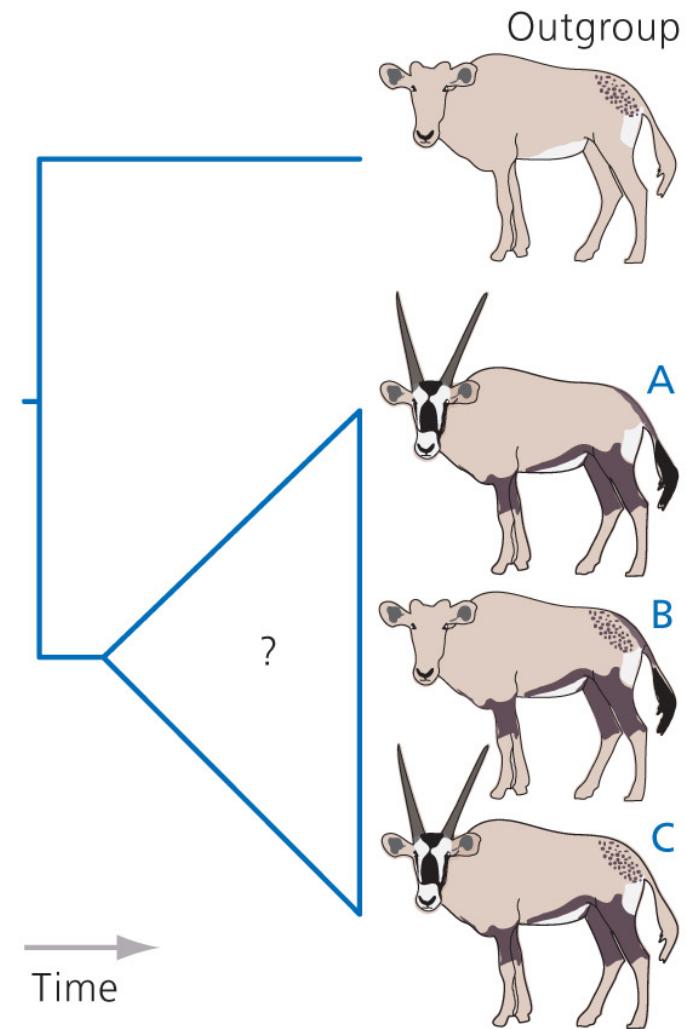
Are birds reptiles?

Extant species

(b)



We usually don't know which traits are derived and ancestral



Calculating a parsimony score for a tree based on characters

Character matrix

	Leg color	Spotted butt	Tail color	Horns	Face pattern
Hypothesis 1					
					
					
					
					

Parsimony score

1

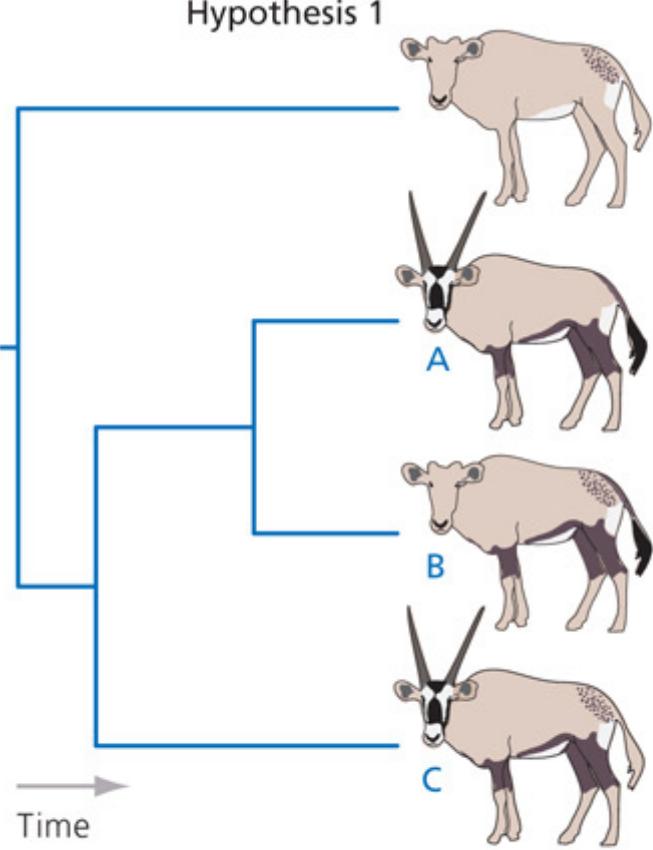
1

1

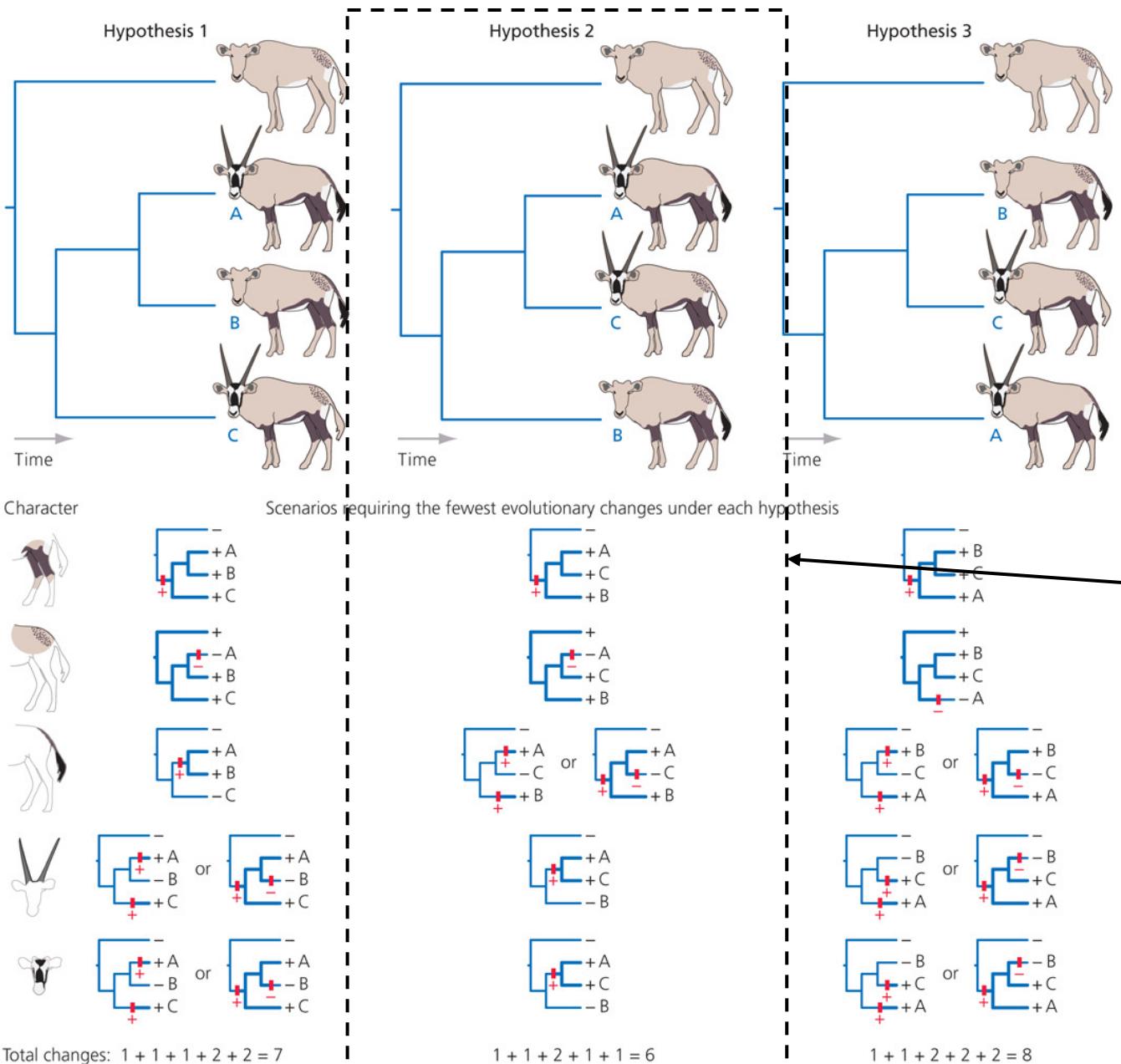
2

2

Total: 7



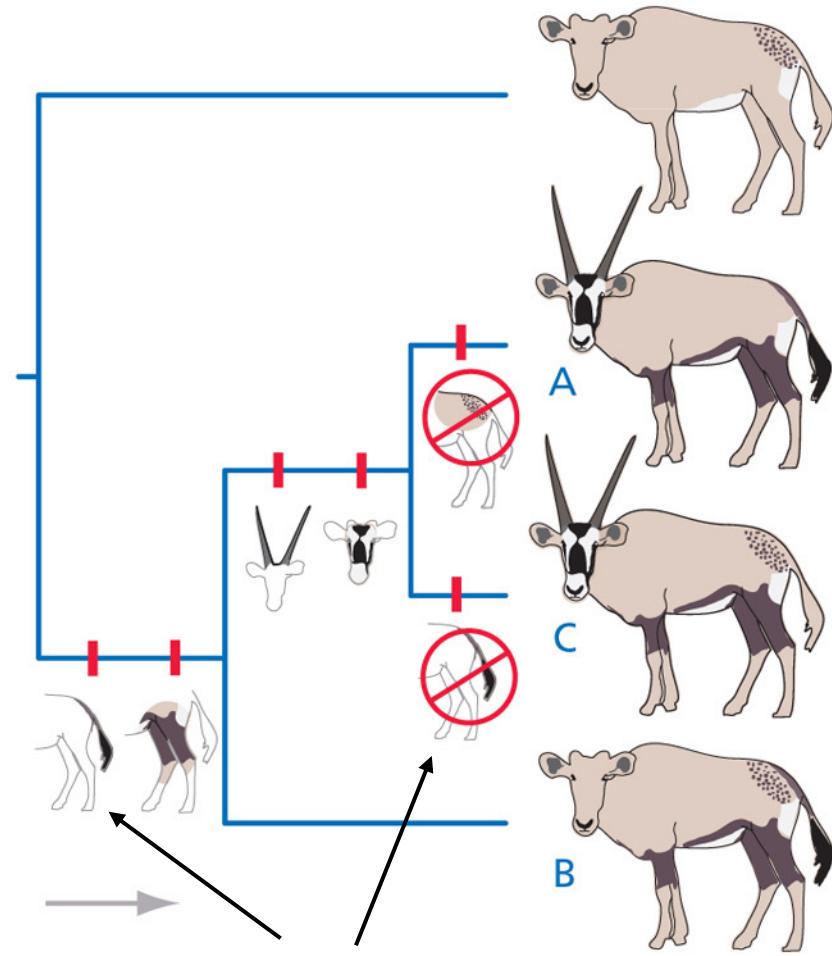
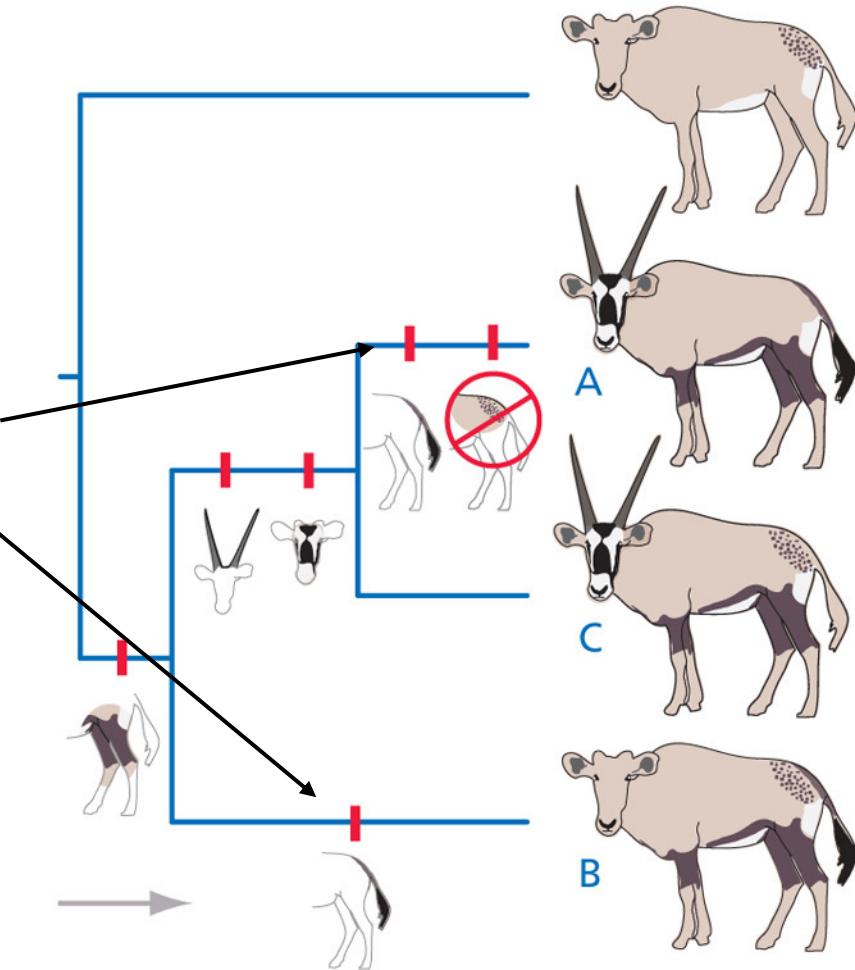
Parsimony-based tree search



The most parsimonious tree is favored
• Most parsimonious = fewest transitions

Two equally parsimonious histories

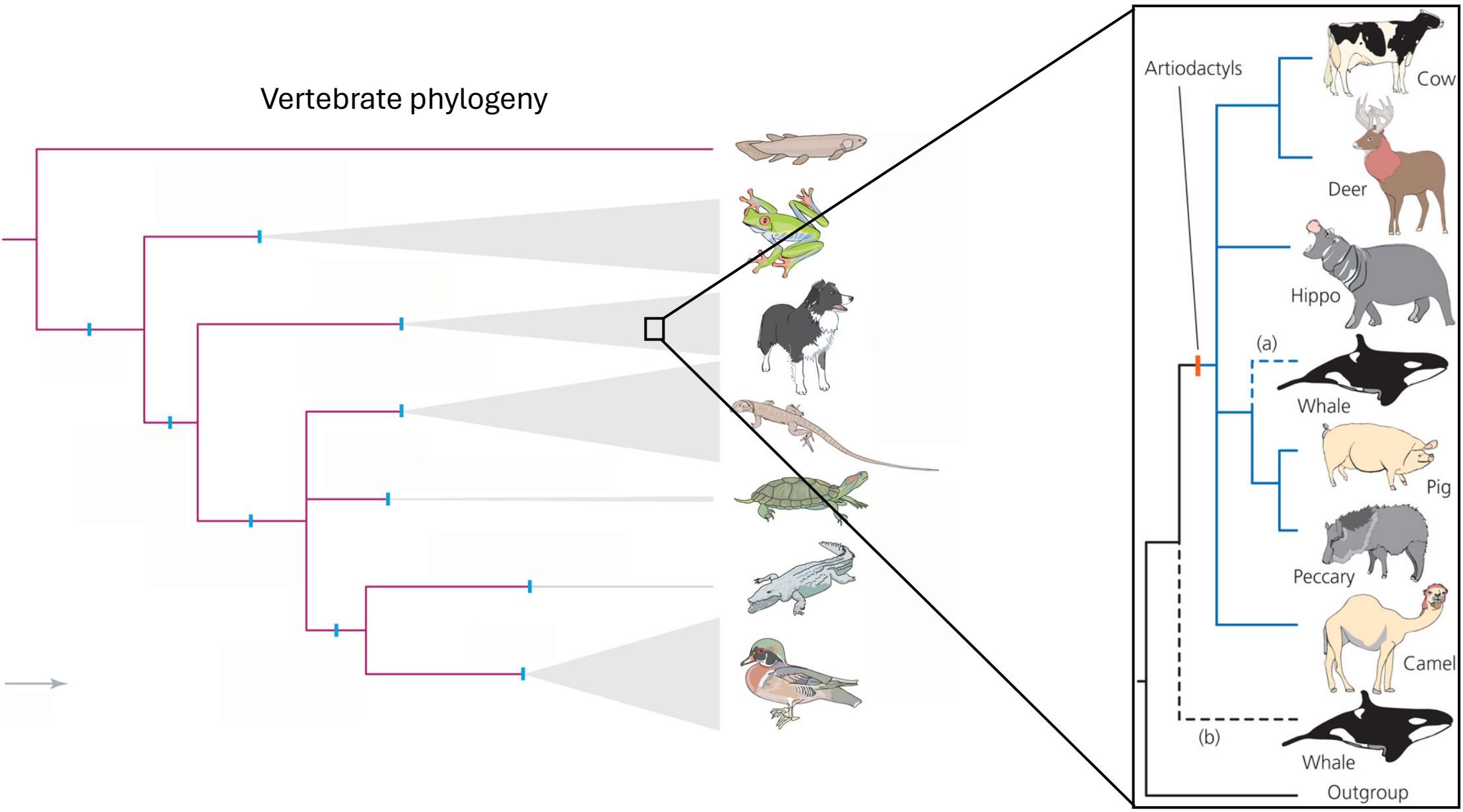
The same transition occurred in two independent lineages (convergent evolution)



Homoplasy:

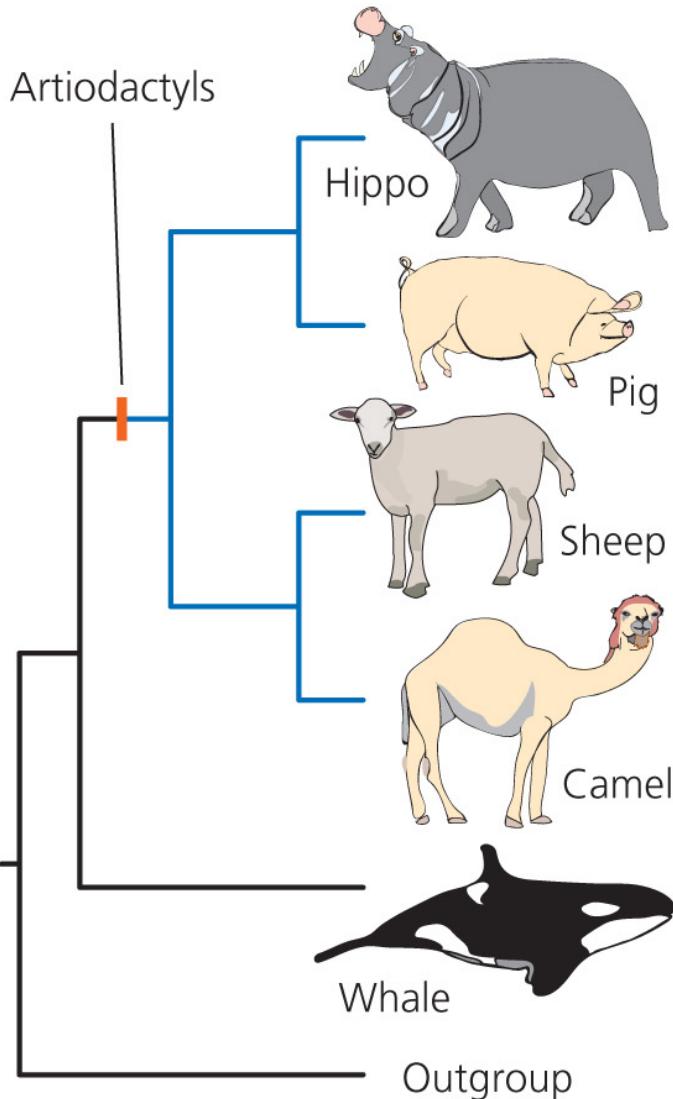
- When characters experienced convergence or back-transitions
- Creates major problems for phylogenetic inference (especially with parsimony)

What are whales???

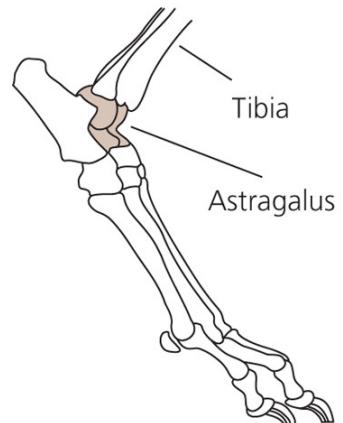


Parsimony phylogenetic analysis using morphological traits

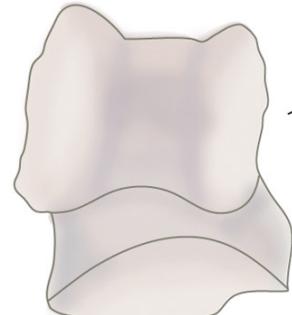
Topology favored by parsimony-based analysis of morphological characters



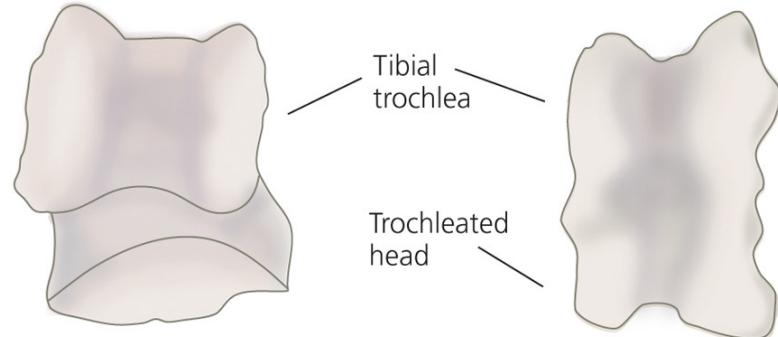
a) Dog



(b) Tapir—a nonartiodactyl ungulate



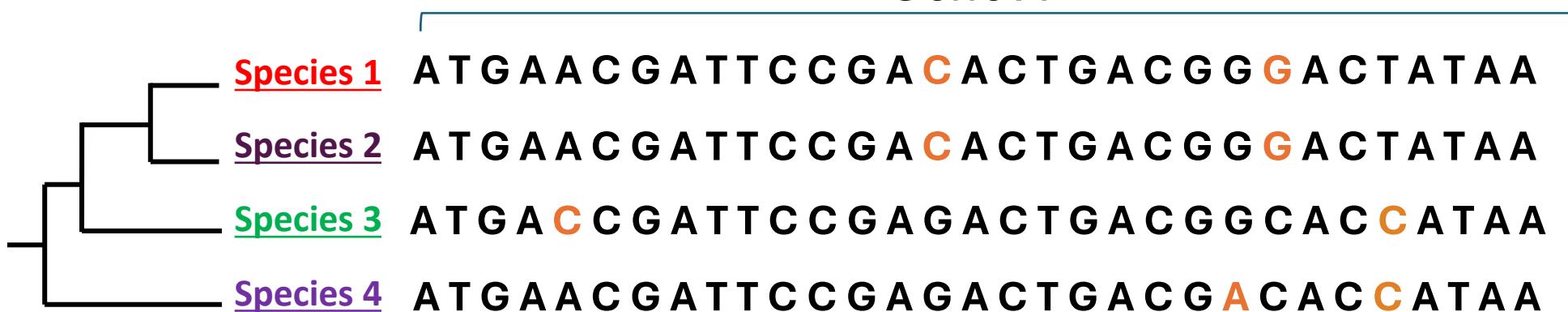
(c) Deer—an artiodactyl ungulate



- Bones in the hind leg are traditionally a useful character for defining the Artiodactyl group

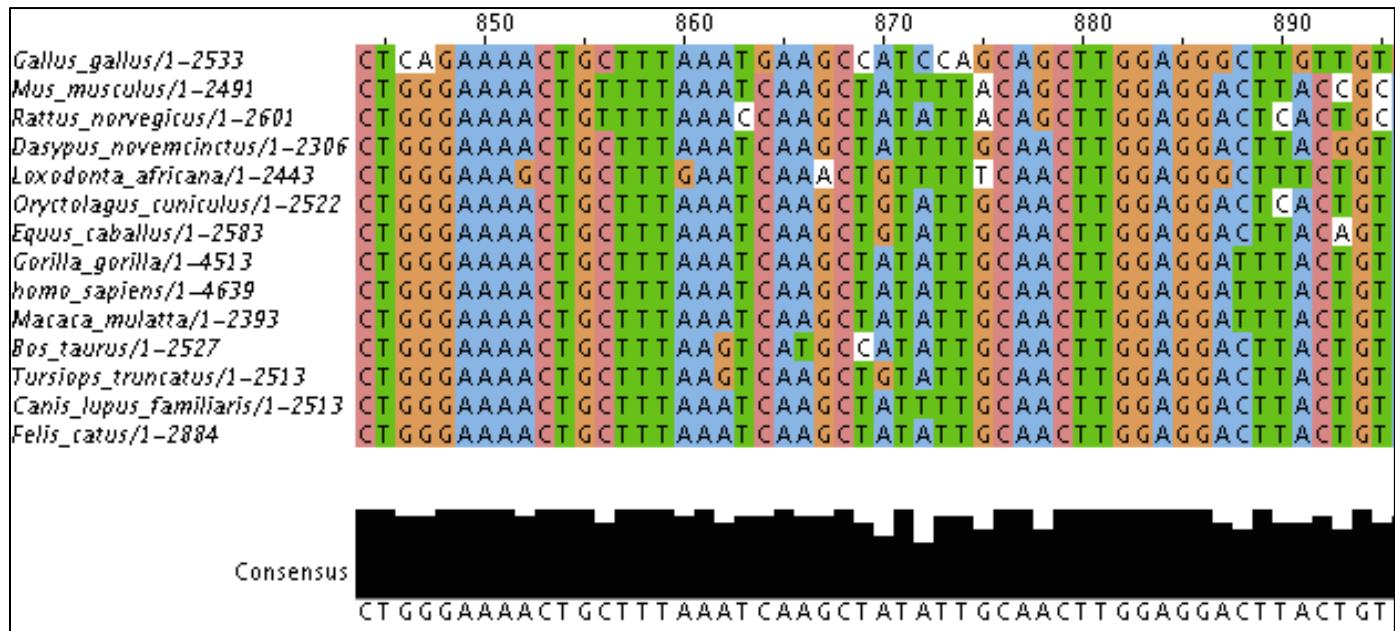
Molecular homology

Gene A



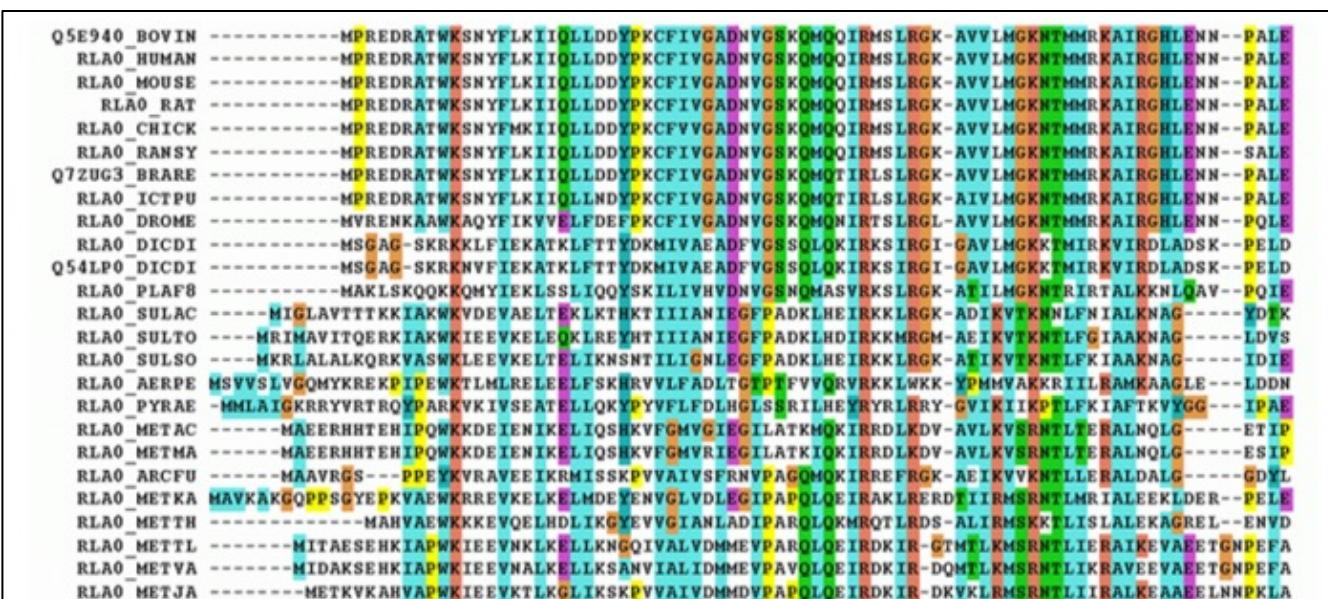
- The DNA sequences are *nearly* identical from these four different species
 - This reflects the shared ancestry of the species and that *Gene A* is a homologous gene shared by these four species
- The differences in the DNA sequences are due to **mutations**.
 - Species 1 and 2 share two mutations, meaning the mutation occurred in their common ancestor but not in the ancestor of Species 3 and 4

Molecular (DNA or Amino Acid sequence) phylogenetics



DNA sequences can be used to make a character matrix

- DNA samples are sequenced from multiple different species
- Sequences are ‘aligned’ so that each column (i.e. site) in the alignment represents a homologous character



Protein (amino acid) sequences can be used to make a character matrix

- Protein’s can be ‘sequenced’ with mass spectrometry (rare)
- Protein sequence can be inferred from DNA sequence once it’s been sequenced (common)

Multiple sequence alignment (MSA)

Nucleotide sequence before alignment

50

60

70

whale: ... GGG CCA ATC CCT TAC CCT ATT CTT ACA CAA AAC ...

cow: ... **G G G C C C A T C C C T A A C** A G C C T C C C A C A A A A C ...

After alignment

50

60

70

whale: ... GGG CCA ATC CCT TAC CCT ATT CTT ACA CAA AAC

cow: ... **GGG CCC ATC CCT AAC** --- **AGC CTC C CA CAA AAC**.

Encoded amino acid sequence before alignment

whale: ... Gly Pro Ile Pro Tyr Pro Ile Leu Thr Gln Asn ...

cow: ... Gly Pro Ile Pro Asn Ser Leu Pro Gln Asn ...

| After alignment

whale: ... Gly Pro Ile Pro Tyr Pro Ile Leu Thr Gln Asn ...

cow: ... Gly Pro Ile Pro Asn — Ser Leu Pro Gln A

Sequences are
different lengths

Multiple sequence alignment:

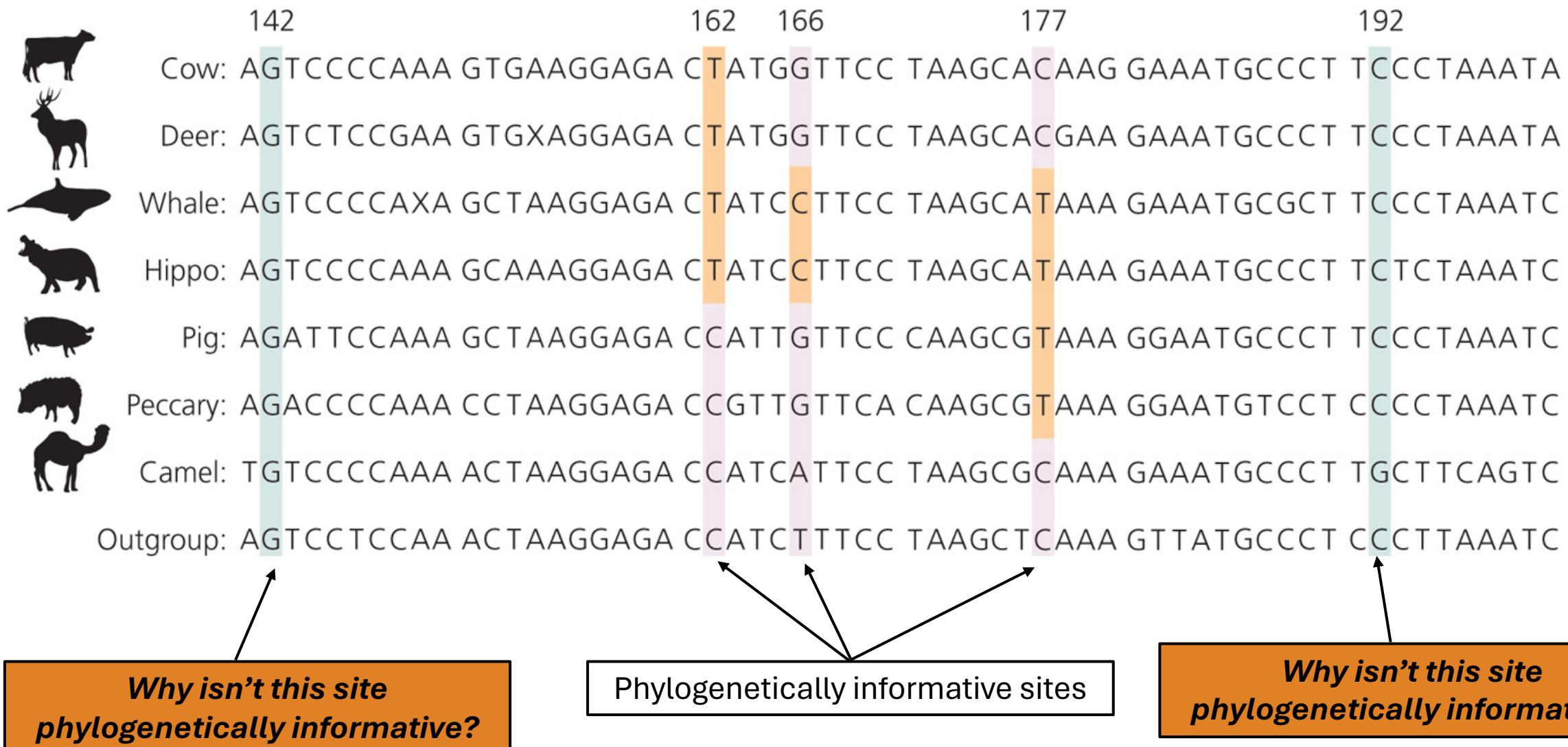
- ‘lining up’ sequences so that columns are most likely to represent homologous sites
 - Often involves introducing gaps

Indel

- Insertion or deletion

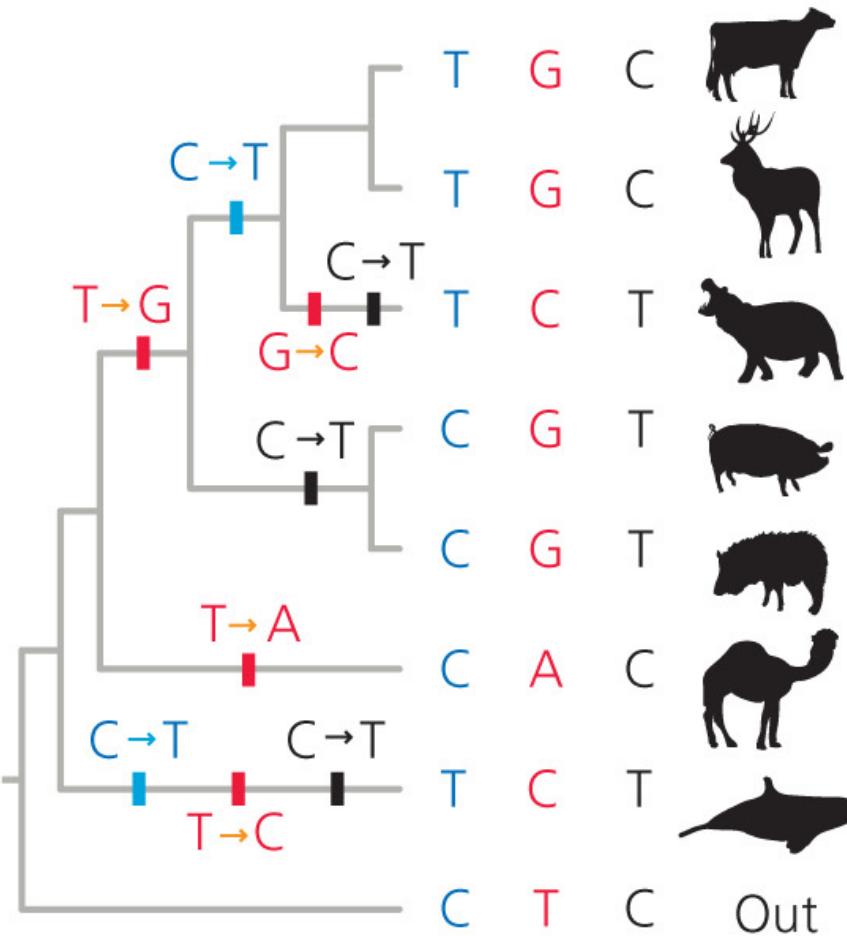
How could we learn whether it's an insertion or deletion?

Using DNA alignments to perform phylogenetic inference

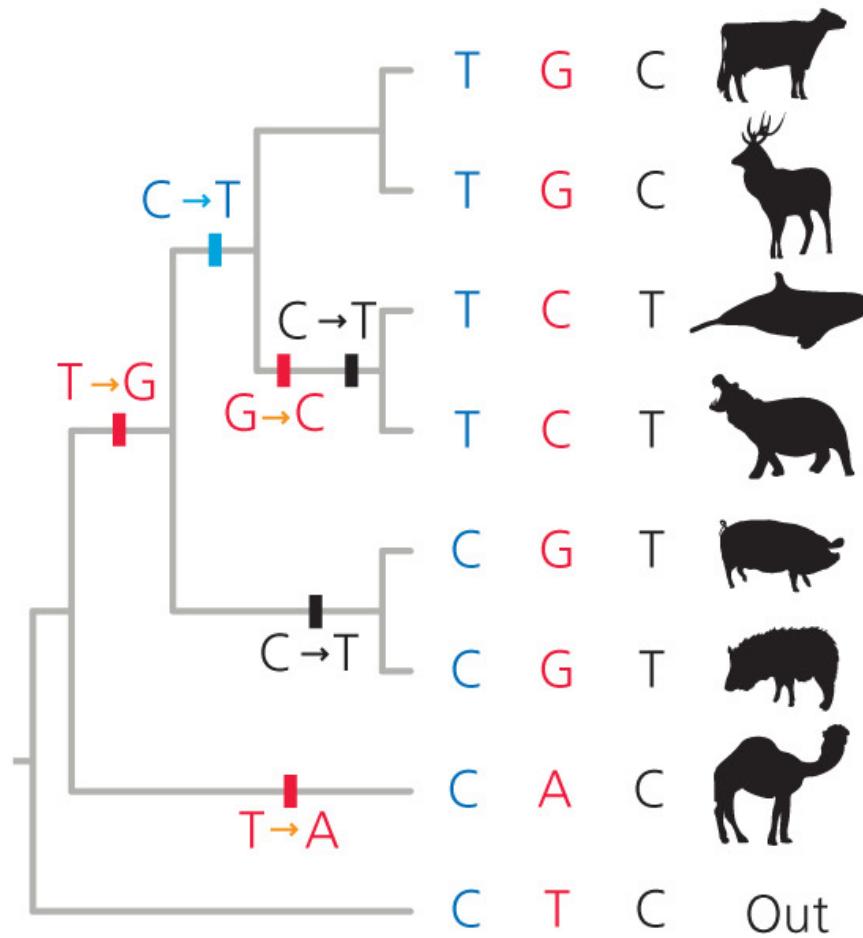


Using DNA alignments to perform phylogenetic inference

Hypothesis 1: whales are sister to all other artiodactyls



Hypothesis 2: whales are sister to hippos (within artiodactyls)



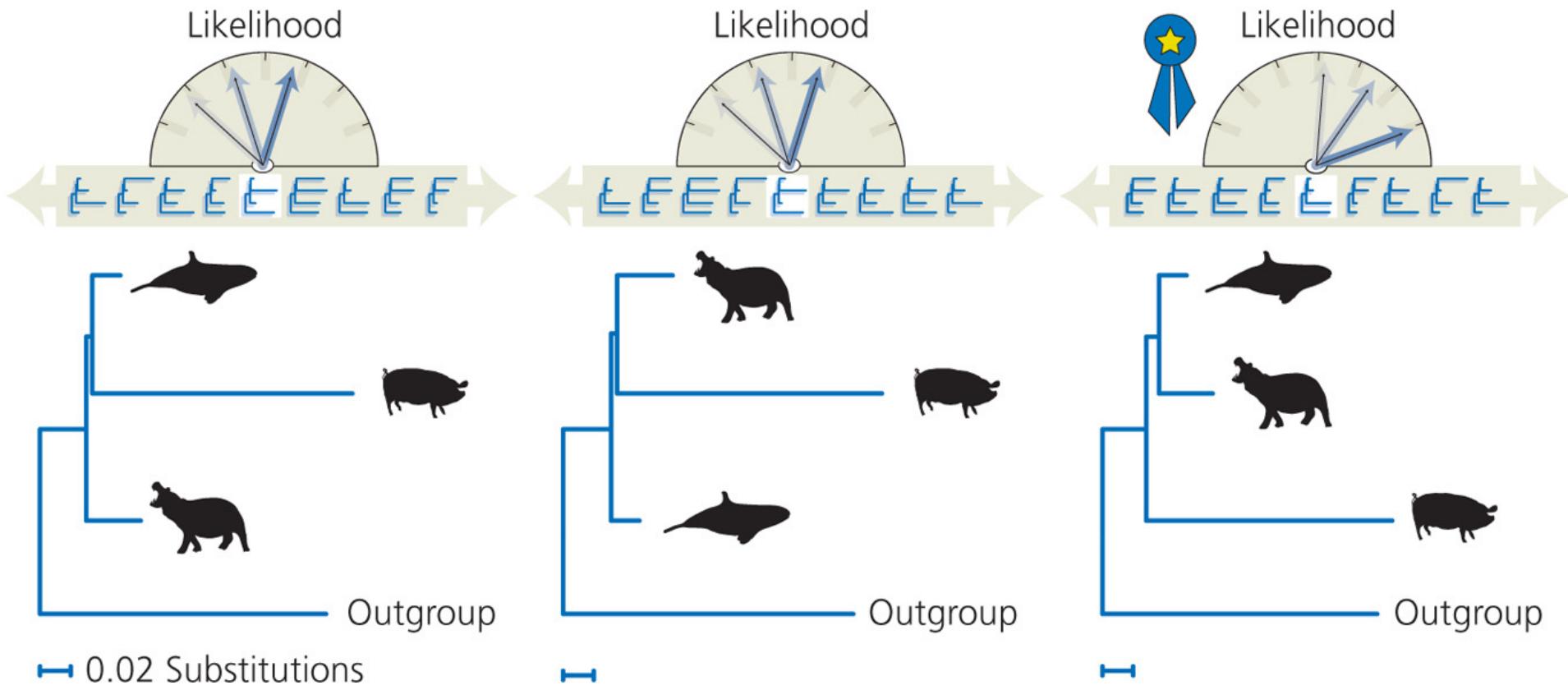
Which topology is favored by parsimony analysis of these three phylogenetically informative sites?

Is there homoplasy at these sites?

Maximum likelihood (ML) phylogenetic inference

Likelihood score:

- A statistic that quantifies the probability of the observed alignment, given a particular tree



Maximum likelihood phylogenetic inference:

- Calculate the likelihood score for many trees and choose the one with the highest (i.e. maximum) likelihood score

Maximum likelihood (ML) phylogenetic inference

The user provides multiple sequence alignment:

- MSAs are the input data for ML inference

The data

 AGTCCCCAXA GCTAAGGAGA ... TCCCTAAATC

 AGTCCCCAAA GCAAAGGAGA ... TCTCTAAATC

 AGATTCCAAA GCTAAGGAGA ... TCCCTAAATC

Outgroup AGTCCTCCAA ACTAAGGAGA ... CCCTTAAATC

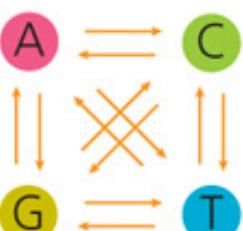
The algorithm chooses a random tree

- Topology and branch lengths



ML uses models of DNA sequence evolution:

- a set of numbers giving the probability of each possible substitution
- Models can account for homoplasy!



Calculating the likelihood score for a tree

Likelihood score of a tree

= Probability of the observed alignment given the tree and the model of evolution

$$L(\text{tree}) = P(\text{AGTCCCCAXA GCTAAGGAGA ... TCCCTAAATC} | \text{AGTCCCCAAA GCAAAGGAGA ... TCTCTAAATC}, \text{AGATTCCAAA GCTAAGGAGA ... TCCCTAAATC}, \text{AGTCCTCCAA ACTAAGGAGA ... CCCTTAAATC})$$

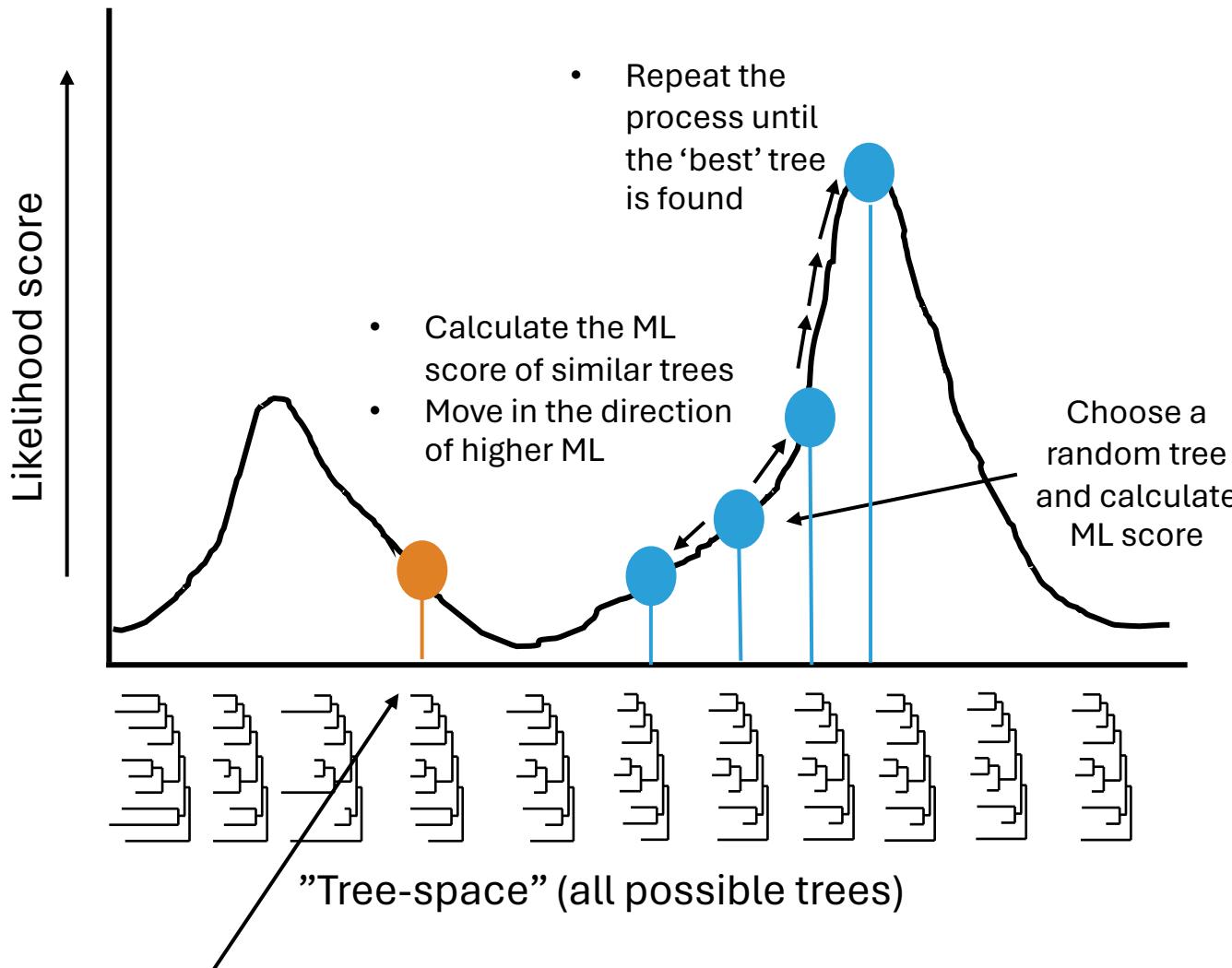
$$= P(\text{A A A} | \text{tree}) \times P(\text{G G G} | \text{tree}) \times \dots$$

Probability of site 1 (given the tree)

Probability of site 2 (given the tree)

= Likelihood score

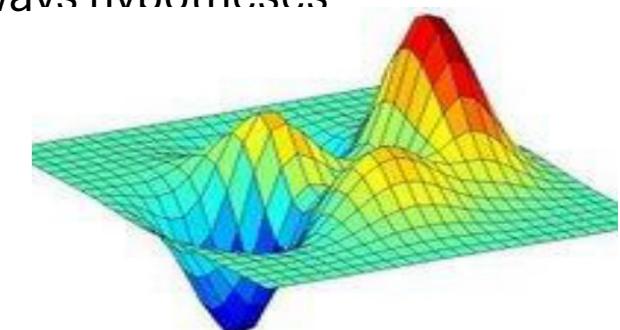
Maximum likelihood (ML) phylogenetic inference



Would the algorithm select the 'globally optimal tree' if it randomly started here?

Maximum likelihood inference:

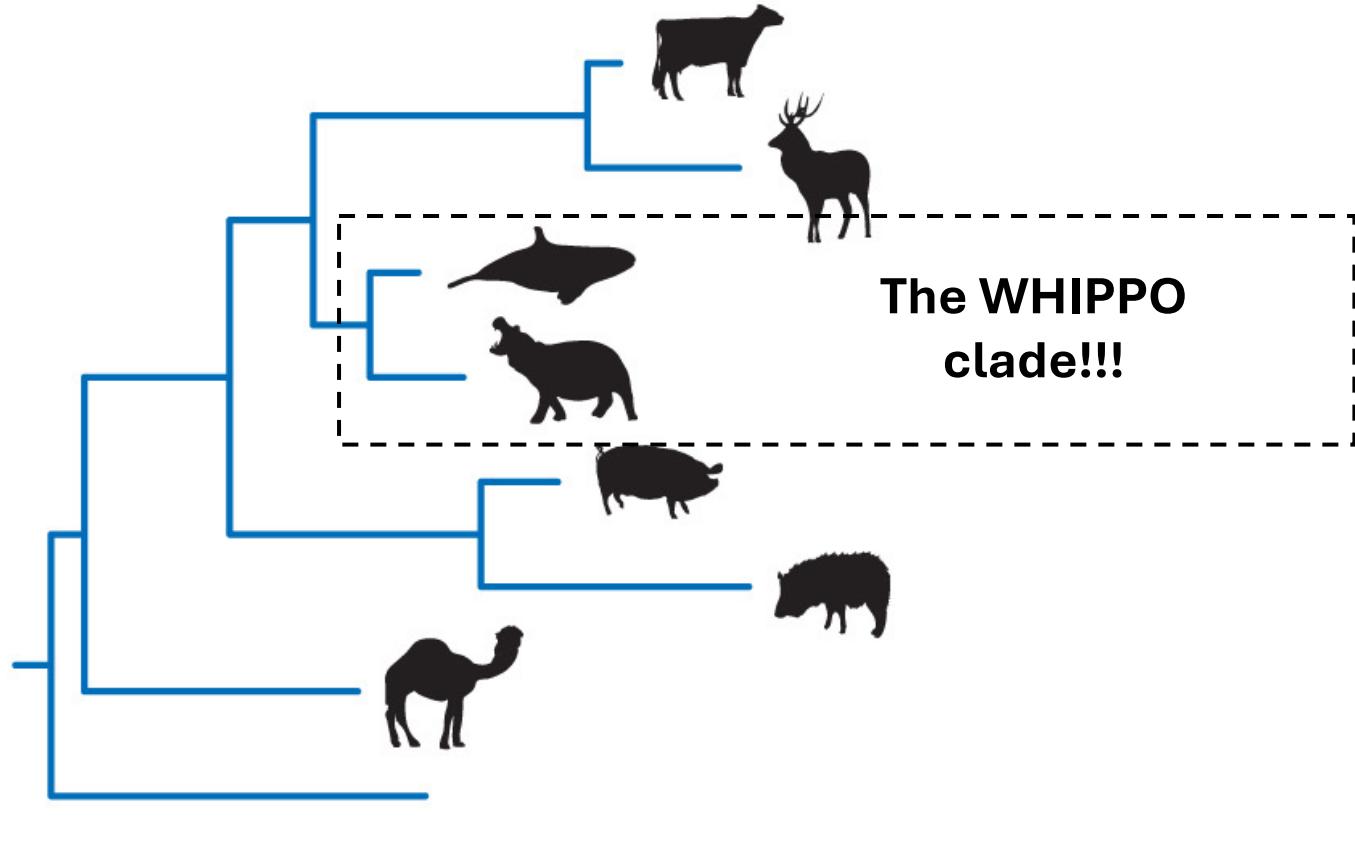
- Sometime called 'tree search'
- Iterative (occurs through multiple repeated steps)
- 'Guess', 'check', 'repeat'
- The number of possible trees is usually **LARGE**
- Requires powerful computers to search many possible trees
- Even with large computers, it's impossible to test every tree
 - This is one of the reasons trees are always hypotheses



Tree-space represented in 3-dimensions

ML trees support the Whippo clade

The current leading hypothesis

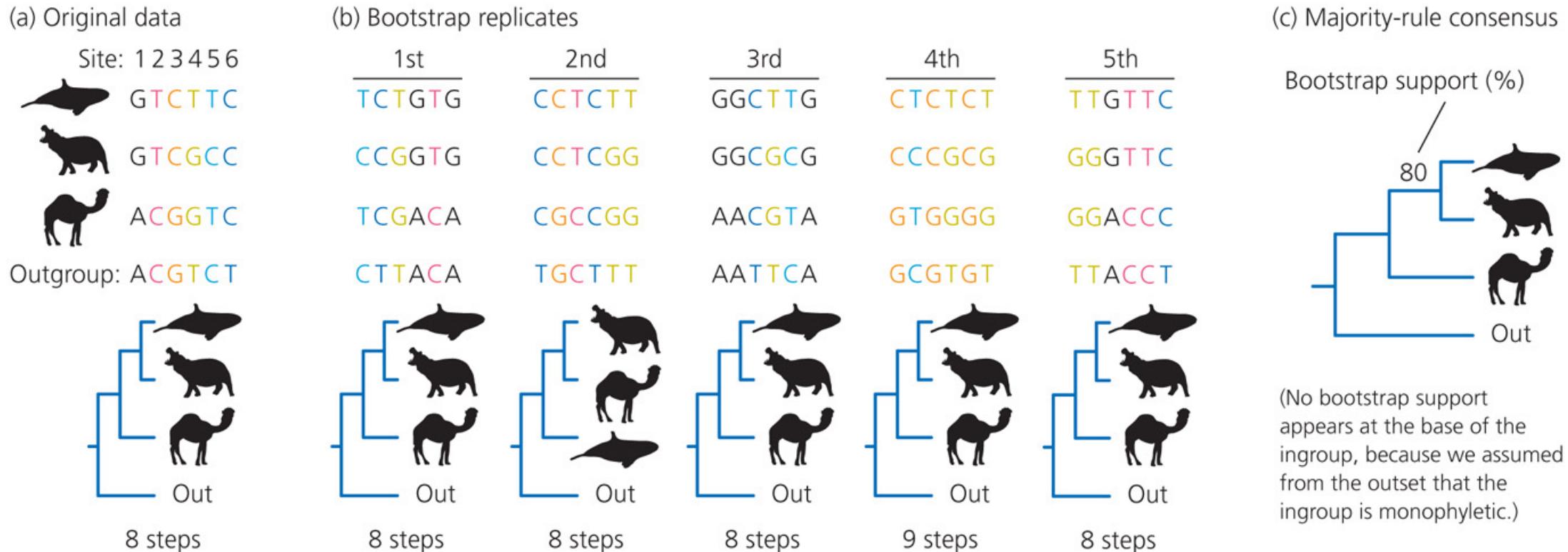


Notes about interpreting phylogenies:

- Phylogenies ALWAYS represent hypotheses
- We will never know the answer
- We can only use phylogenetics to favor some hypotheses over others (not prove or disprove)

0.07 (substitutions per site in alignment)

Assessing our confidence in our chosen tree with bootstrap resampling of the original data

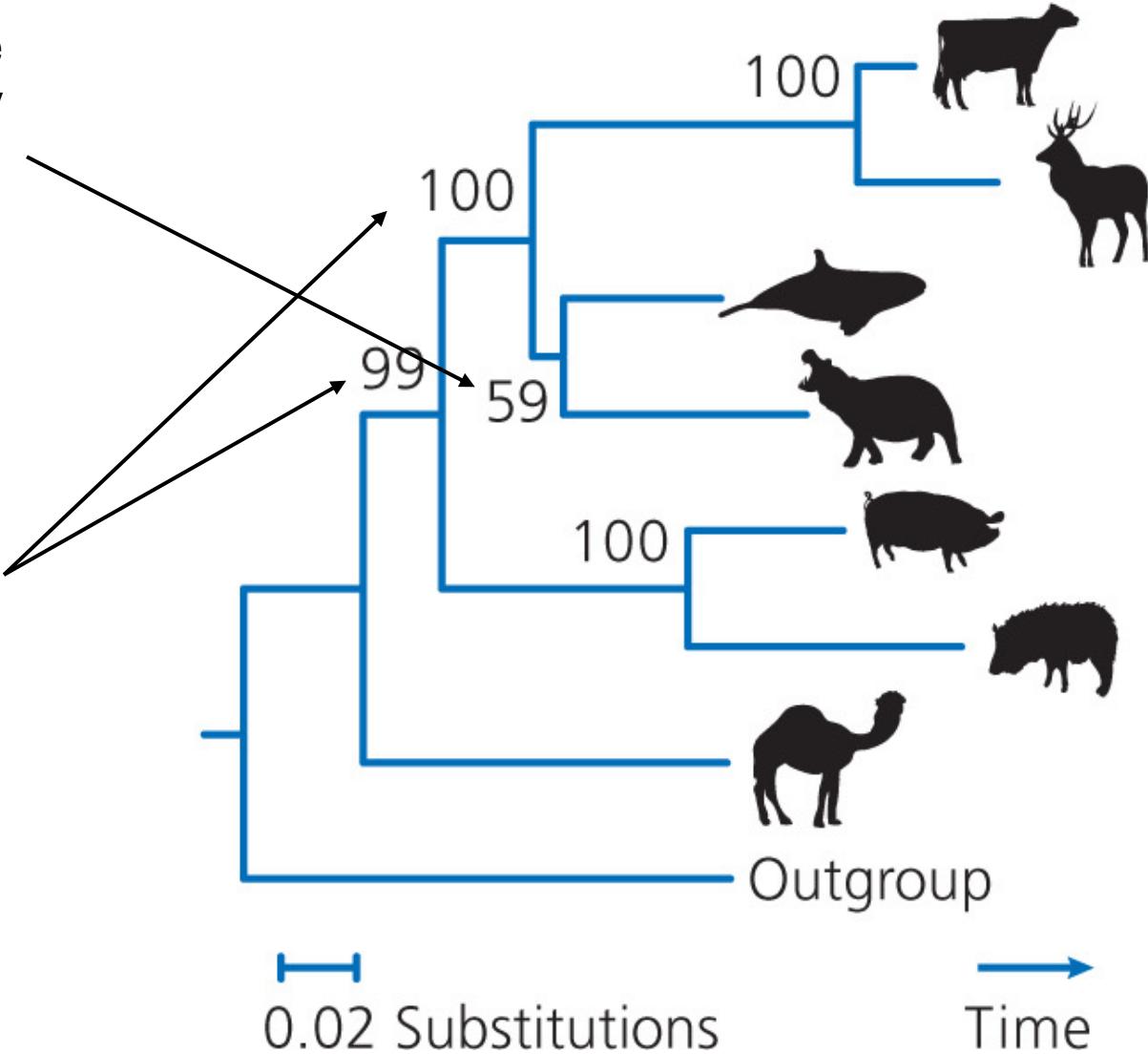


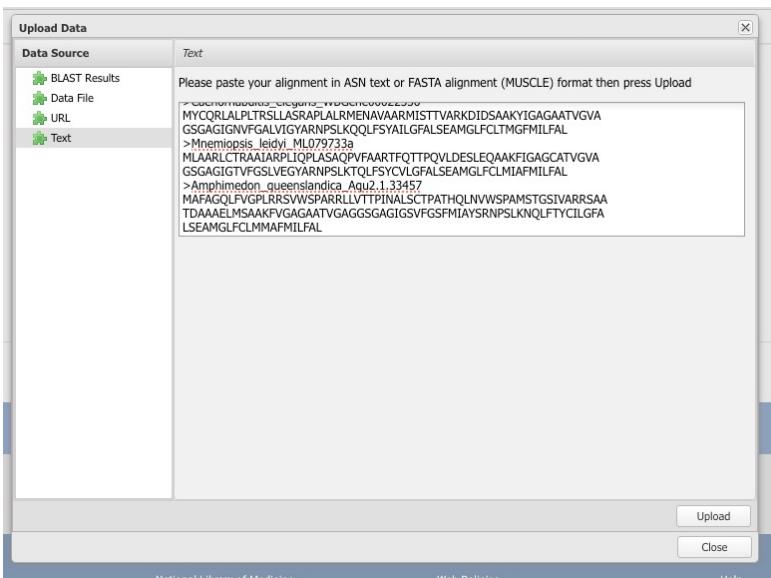
How is a bootstrap support score similar to a p-value? How is it different?

Assessing our confidence in our chosen tree using bootstrap resampling of the original data

The 'whippo' clade is actually not very well-supported

However, the clade that places whales as an artiodactyl is well-supported





National Library of Medicine
National Center for Biotechnology Information

NCBI MSA Viewer 1.25.0 Help Release Notes

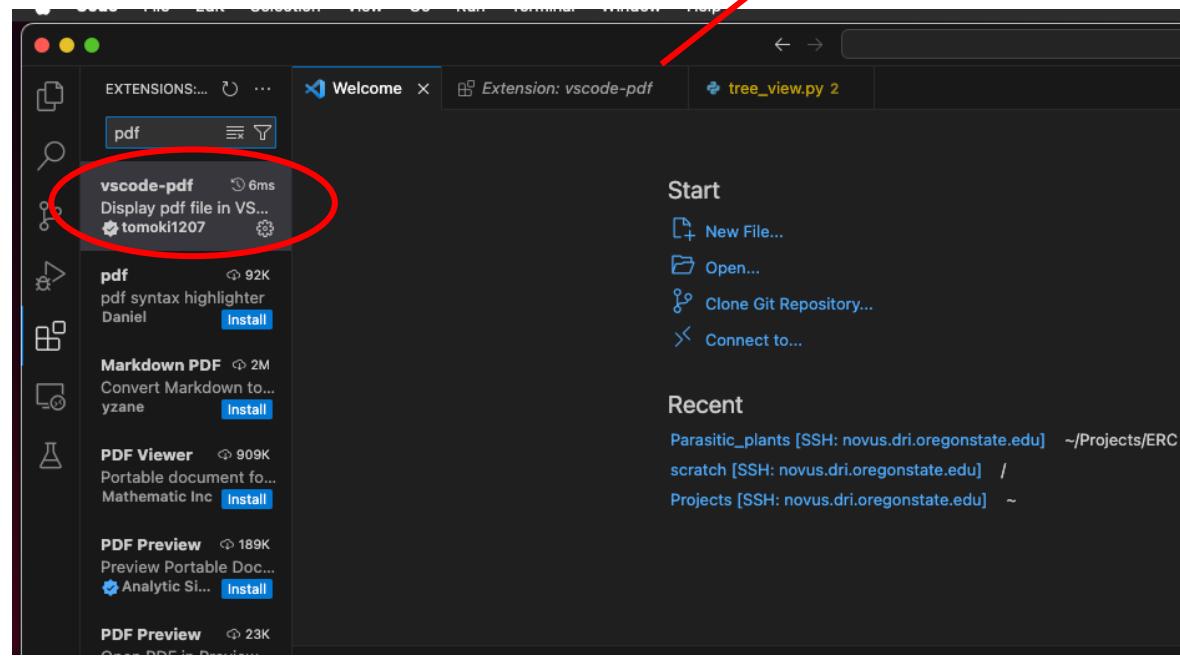
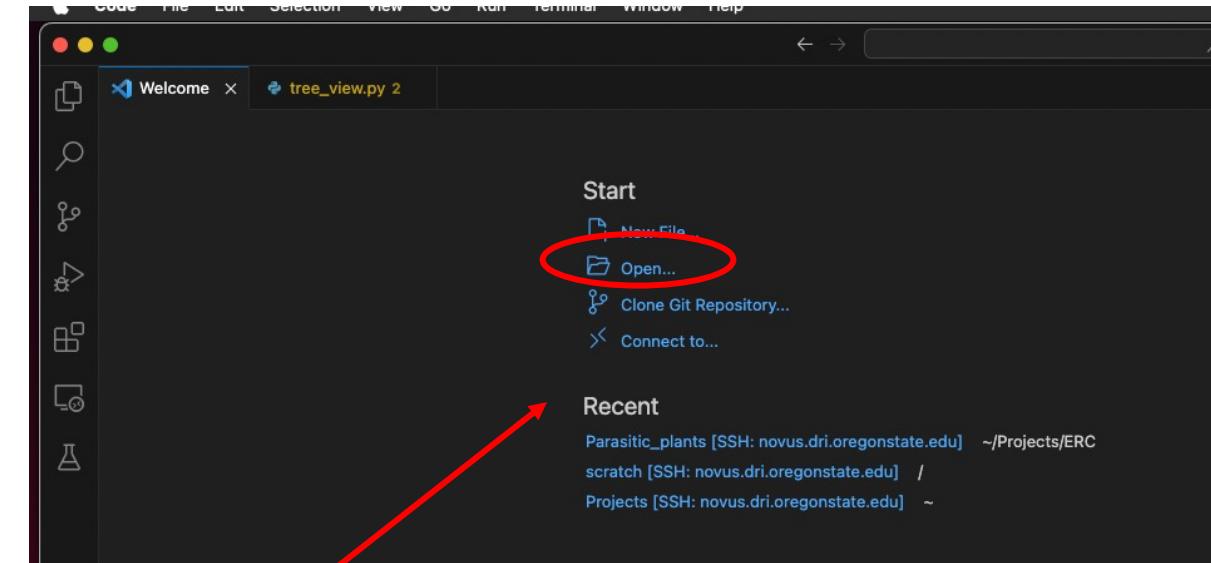
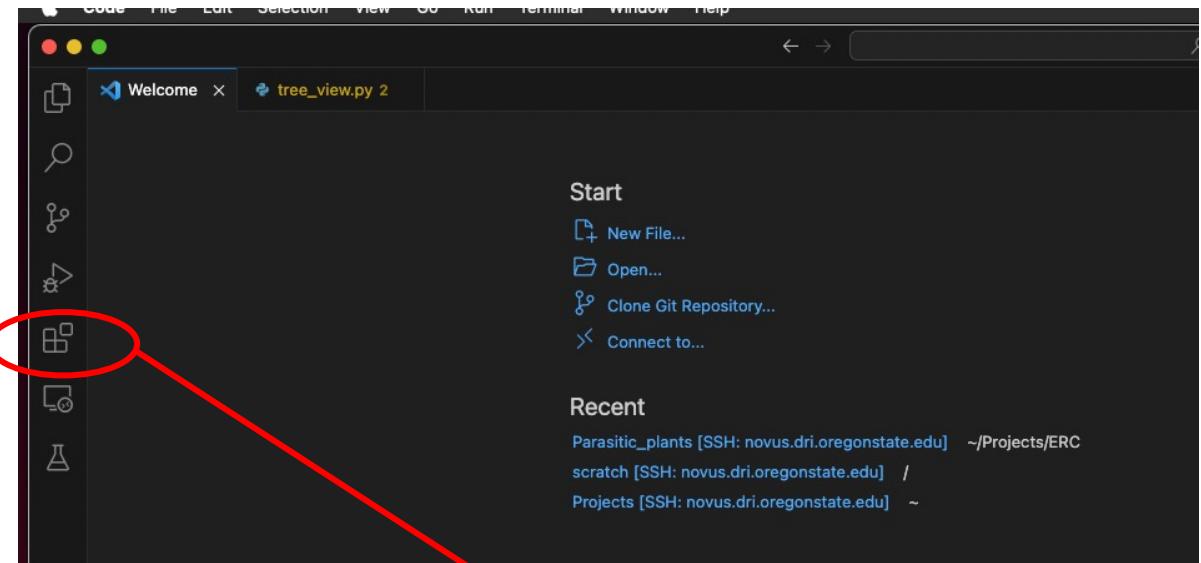
NCBI Multiple Sequence Alignment Viewer

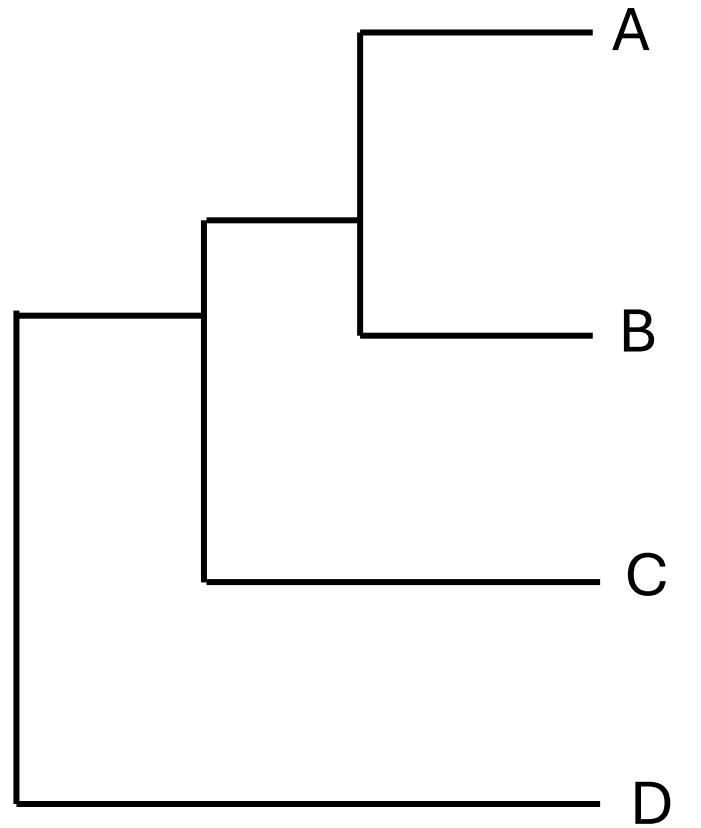
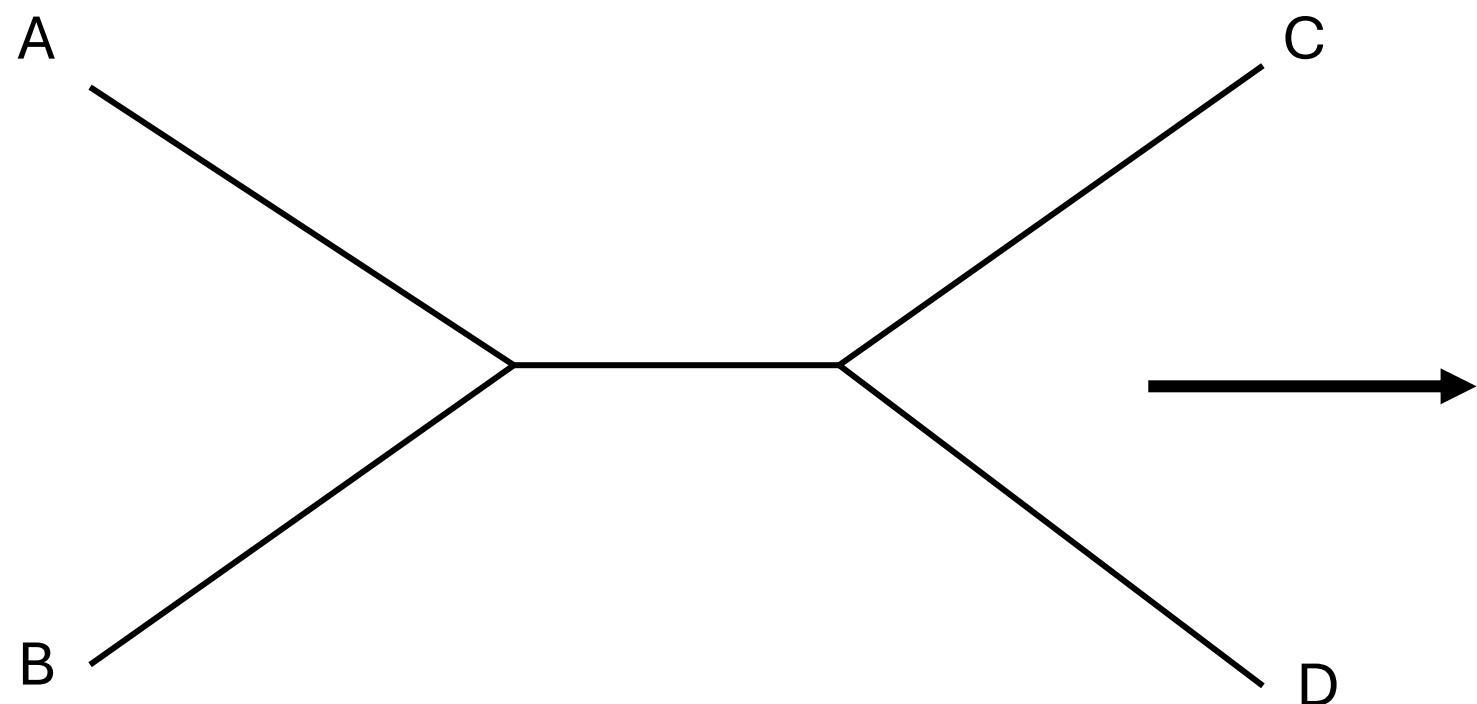
The NCBI Multiple Sequence Alignment Viewer (MSA) is a graphical display for nucleotide and protein sequence alignments.

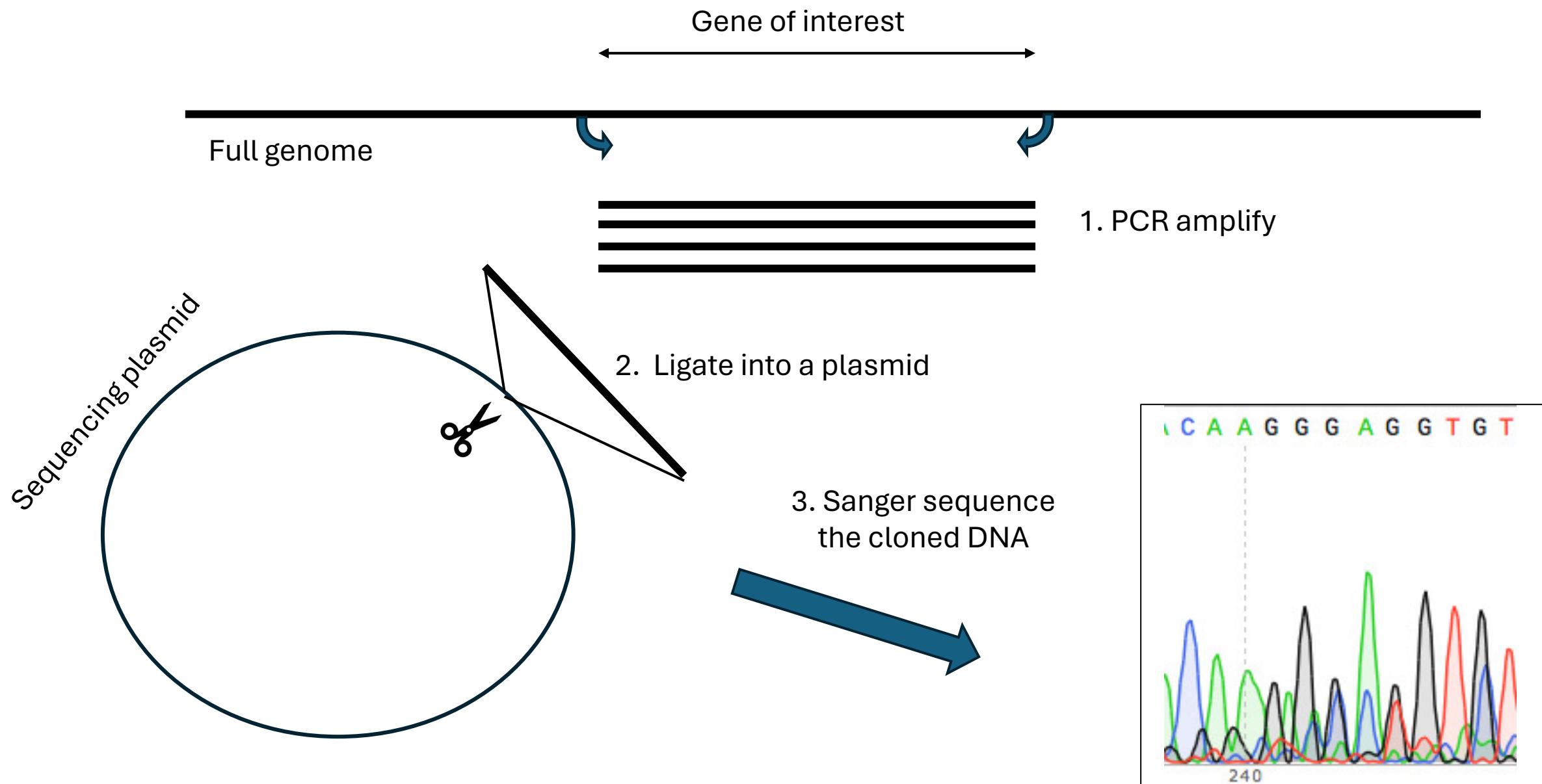
[Review documentation or watch a video tutorial.](#)

To see your own alignment, [Upload](#) your data

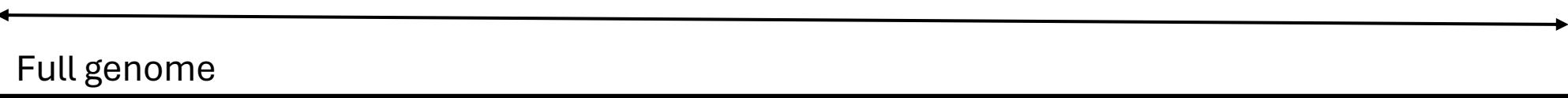








Genome of interest



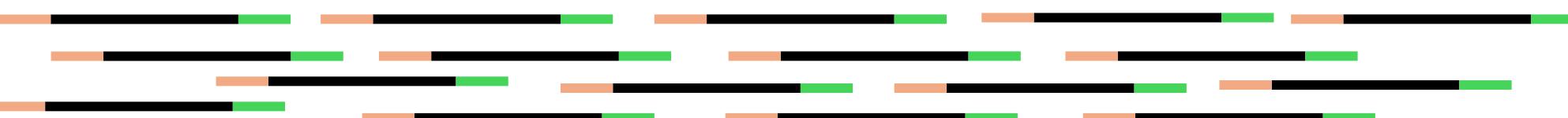
1. Obtain many copies of the genome



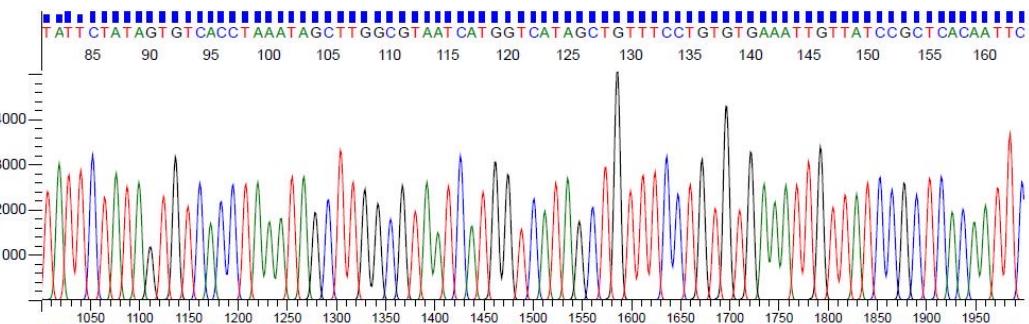
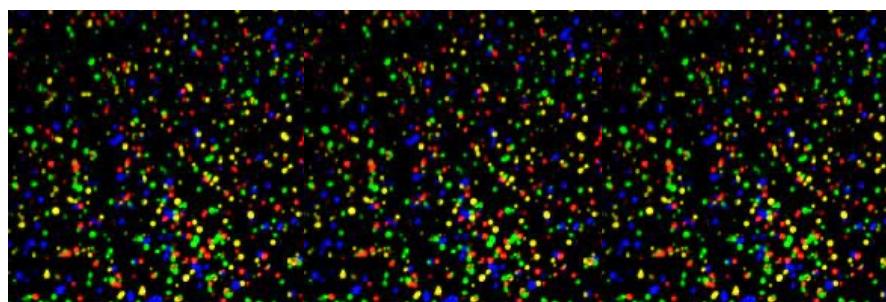
2. Fractionate (break into small pieces)

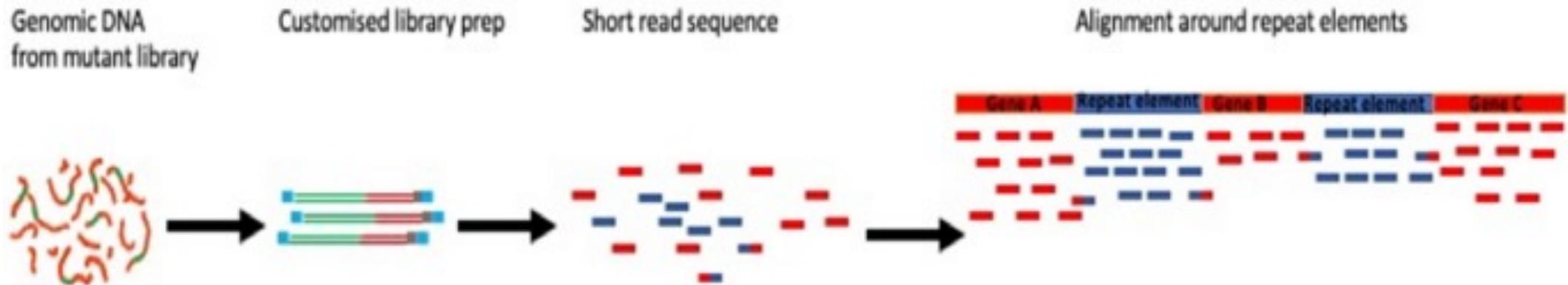
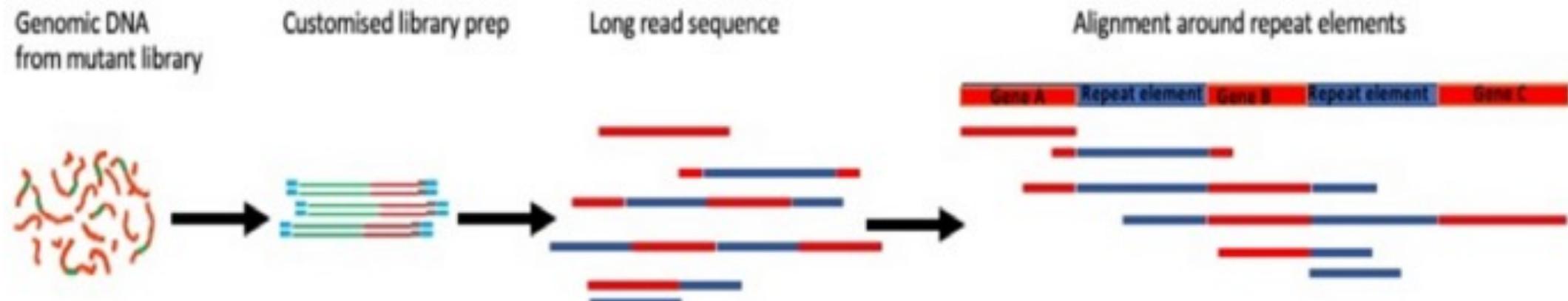


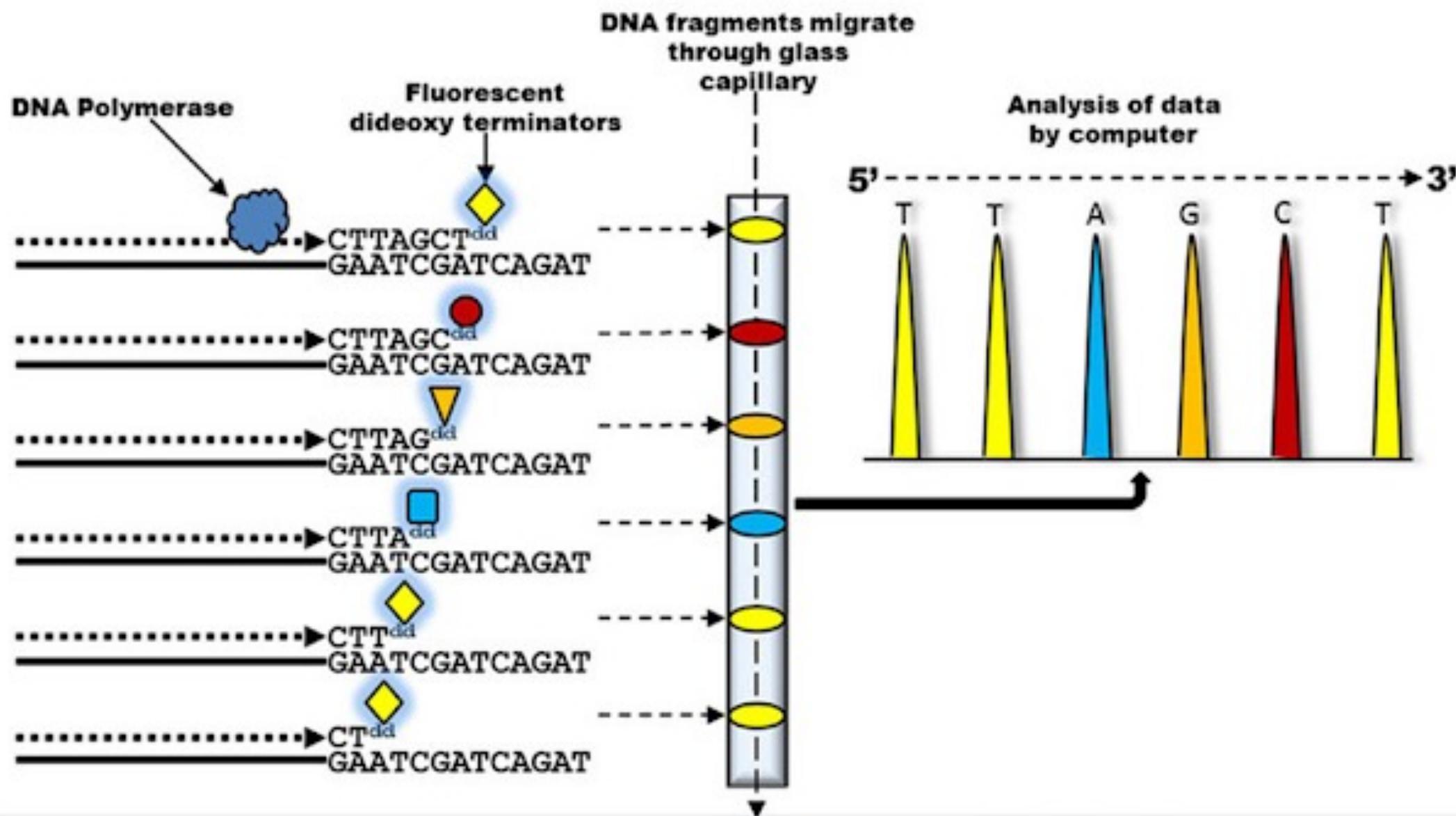
3. Ligate adapters

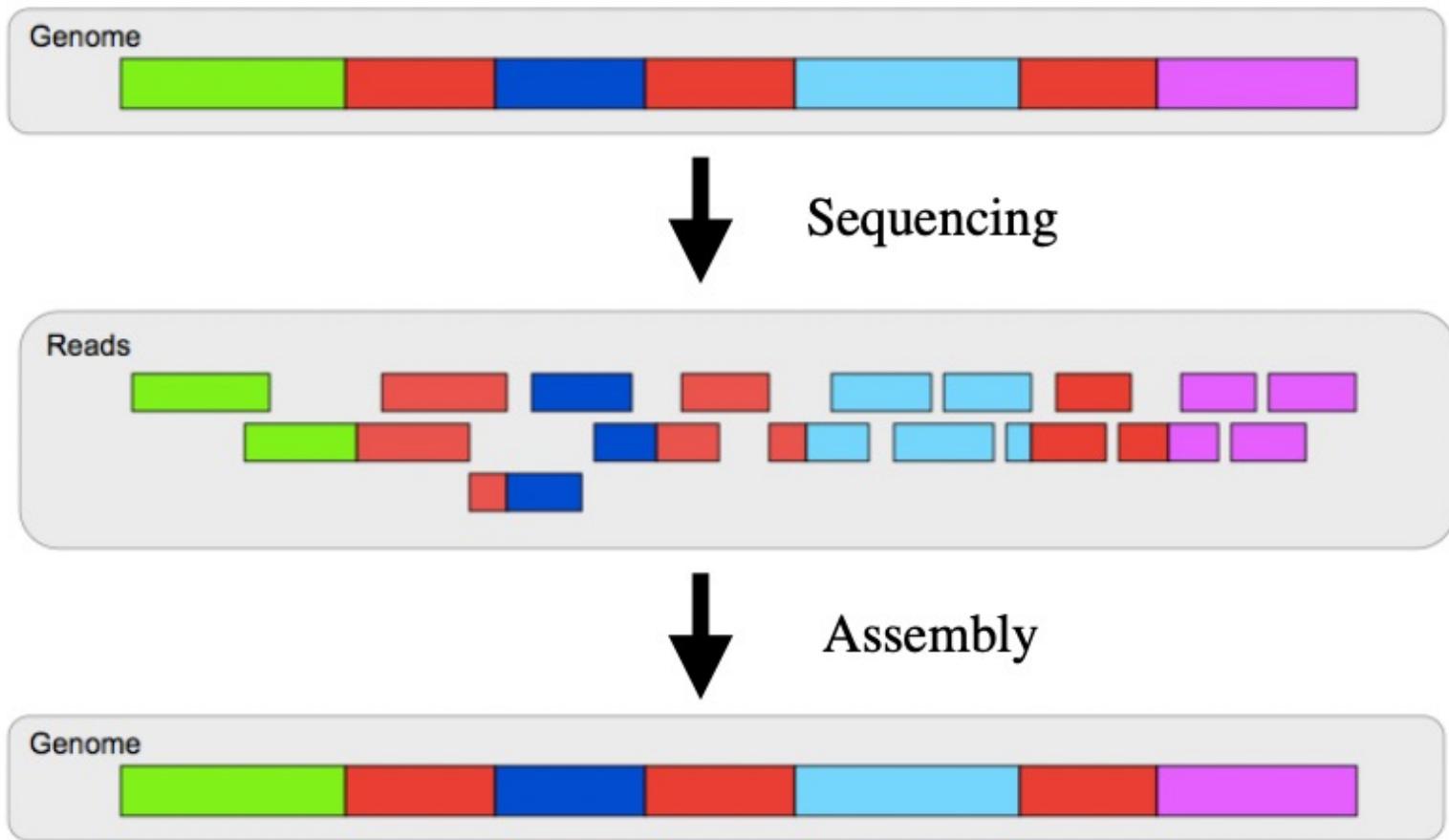


4. Send to next-gen sequencing



A**B**





Whole-genome “shotgun” sequencing starts by copying and fragmenting the DNA

(“Shotgun” refers to the random fragmentation of the whole genome; like it was fired from a shotgun)

Input: **GGCGTCTATATCTGGCTCTAGGCCCTCATTTTT**

Copy: **GGCGTCTATATCTGGCTCTAGGCCCTCATTTTT**
GGCGTCTATATCTGGCTCTAGGCCCTCATTTTT
GGCGTCTATATCTGGCTCTAGGCCCTCATTTTT
GGCGTCTATATCTGGCTCTAGGCCCTCATTTTT

Fragment: **GGCGTCTA TATCTGG CTCTAGGCCCTC ATTTTT**
GGC GTCTATAT CTCGGCTCTAGGCCCTCA TTTTTT
GGCGTC TATATCT CGGCTCTAGGCCCT CATTTTTT
GGCGTCTAT ATCTCGGCTCTAG GCCCTCA TTTTTT

Assume sequencing produces such a large # fragments that almost all genome positions are covered by many fragments...

Reconstruct
this

CTAGGCCCTCAATTTT
CTCTAGGCCCTCAATTTT
GGCTCTAGGCCCTCATTTTT
CTCGGCTCTAGCCCCTCATTTT
TATCTCGACTCTAGGCCCTCA
TATCTCGACTCTAGGCC
TCTATATCTCGGCTCTAGG
GGCGTCTATATCTCG
GGCGTCGATATCT
GGCGTCTATATCT

From these

→ GGC GTCT ATAT CTC GG CTCT AGG CCC CTCA TTTTTT



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

...but we don't know what came from where

Reconstruct
this

CTAGGCCCTCAATTTT
GGCGTCTATATCT
CTCTAGGCCCTCAATTTT
TCTATATCTCGGCTCTAGG
GGCTCTAGGCCCTCATTTTT
CTCGGCTCTAGCCCCTCATTTT
TATCTCGACTCTAGGCCCTCA
GGCGTCGATATCT
TATCTCGACTCTAGGCC
GGCGTCTATATCTCG

→ GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTT

From these



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Key term: coverage. Usually it's short for average coverage: the average number of reads covering a position in the genome.

CTAGGCCCTCAATTTT	
CTCTAGGCCCTCAATTTT	
GGCTCTAGGCCCTCATTTTT	
CTCGGCTCTAGCCCCTCATTTT	
TATCTCGACTCTAGGCCCTCA	177 nucleotides
TATCTCGACTCTAGGCC	
TCTATATCTCGGCTCTAGG	
GGCGTCTATATCTCG	
GGCGTCGATATCT	
GGCGTCTATATCT	
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTT	35 nucleotides
Average coverage = 177 / 35 ≈ 7x	

Coverage could also refer to the number of reads covering a particular position in the genome:

CTAGGCCCTCAATTTT
CTCTAGGCCCTCAATTTT
GGCTCTAGGCCCTCATTTTT
CTCGGCTCTAGCCCCTCATTTT
TATCTCGACTCTAGGCCCTCA
TATCTCGACTCTAGGCC
TCTATATCTCGGCTCTAGG
GGCGTCTATATCTCG
GGCGTCGATATCT
GGCGTCTATATCT

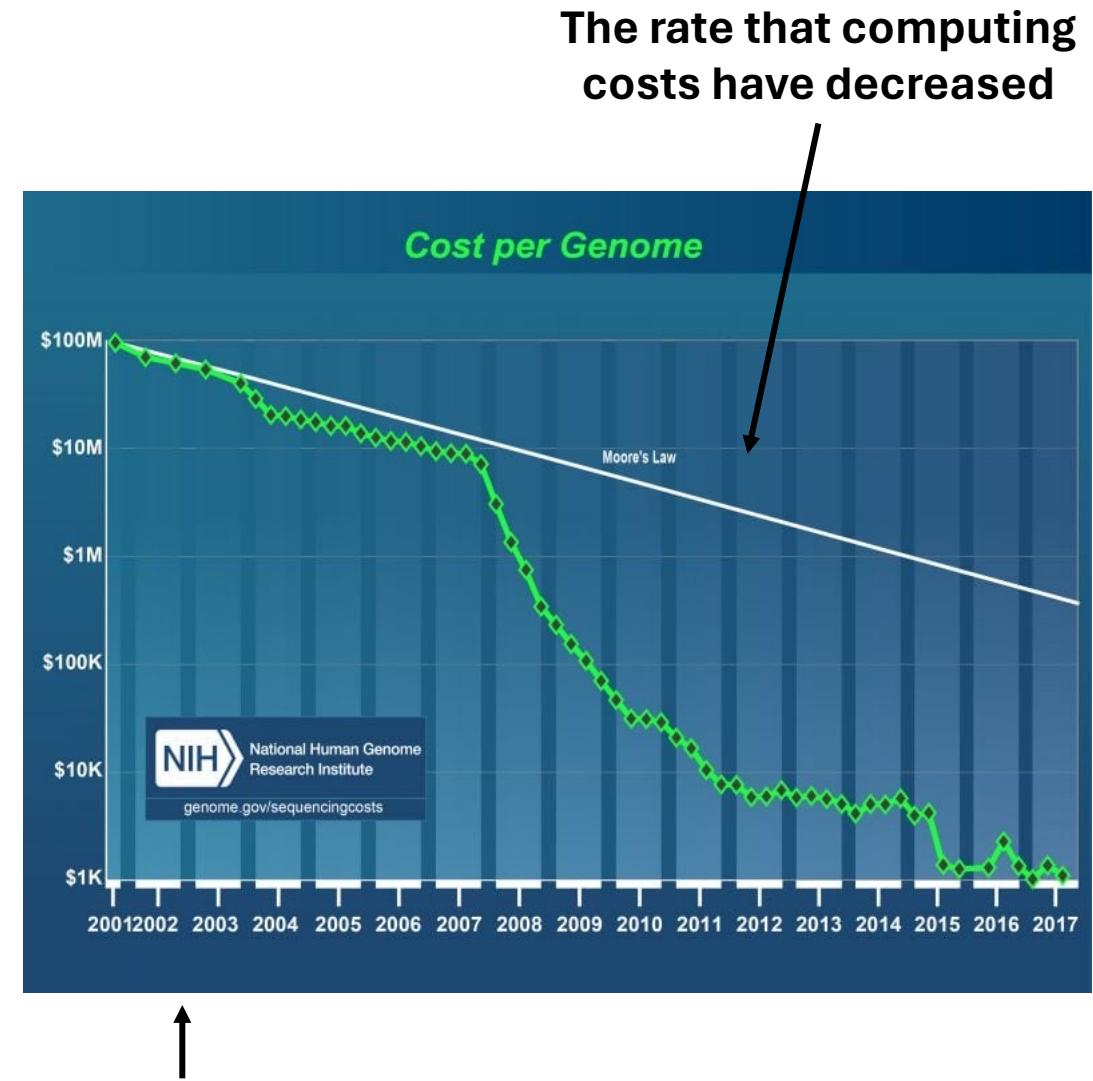
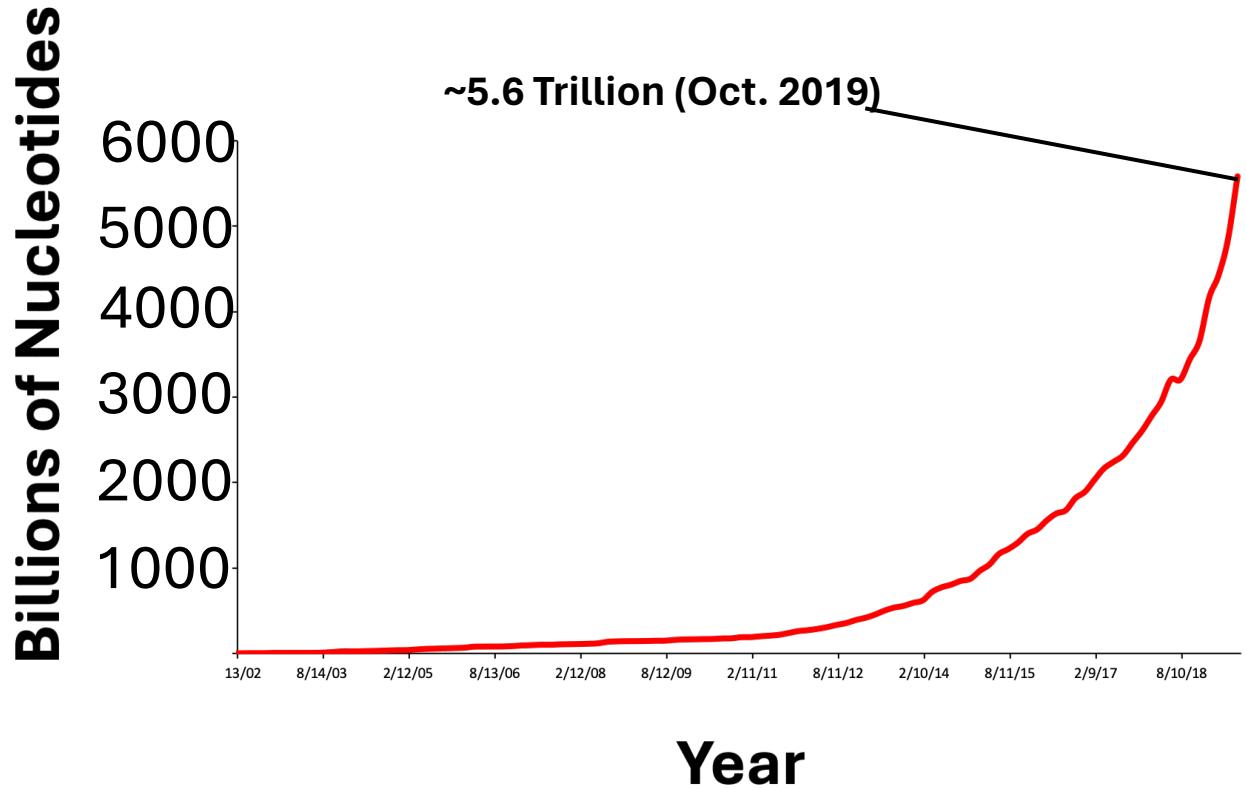
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTT

Coverage at this position = 6

Assembly for Short and Long Reads

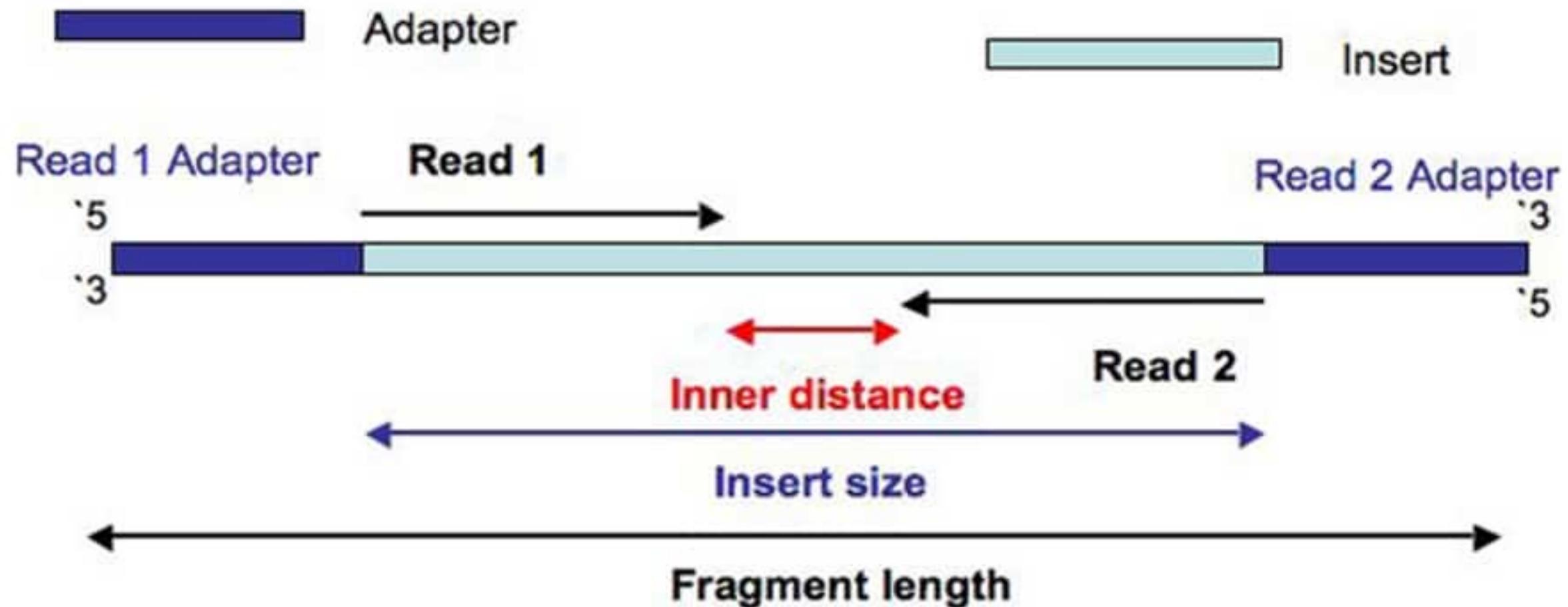


- Long reads (Pacbio/Nanopore)
 - >10,000bp reads are common
 - Higher error rate (5-15%)
 - **Key computational challenge:** overcome high error rate
- Short reads (Illumina):
 - High accuracy, very high throughput
 - Short read length limits ability to resolve repeats
 - **Key computational challenge:** efficiently assemble large numbers of short reads

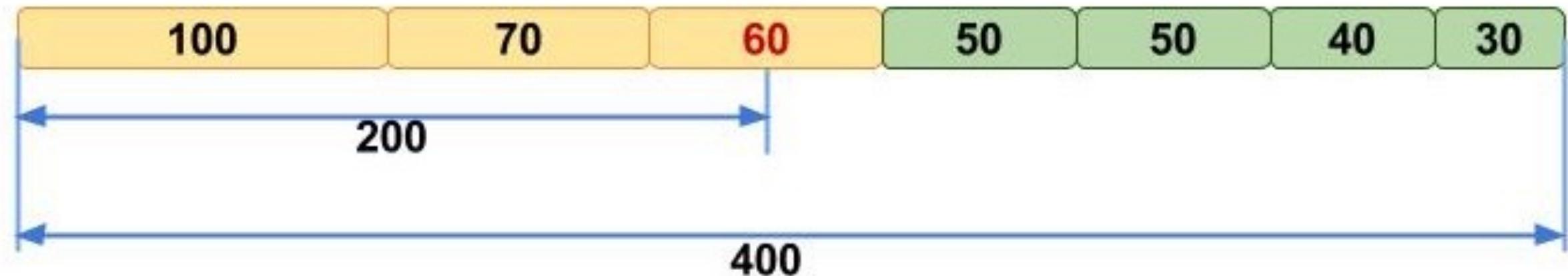


Human genome sequenced:

- Cost more than \$2 billion
- 13-year effort by hundreds of scientists



N50 (high = good)

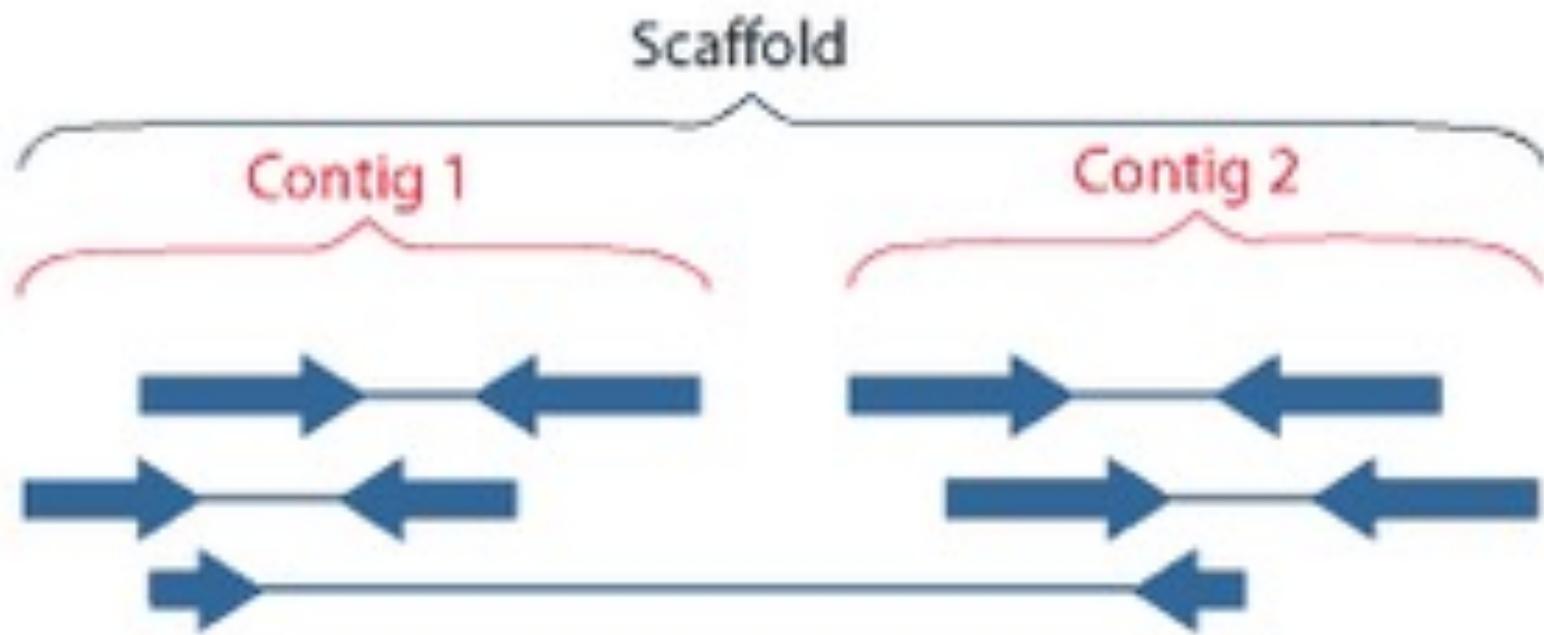


Genome fraction of the reference genome (high = good)



Number of contigs (high = bad)

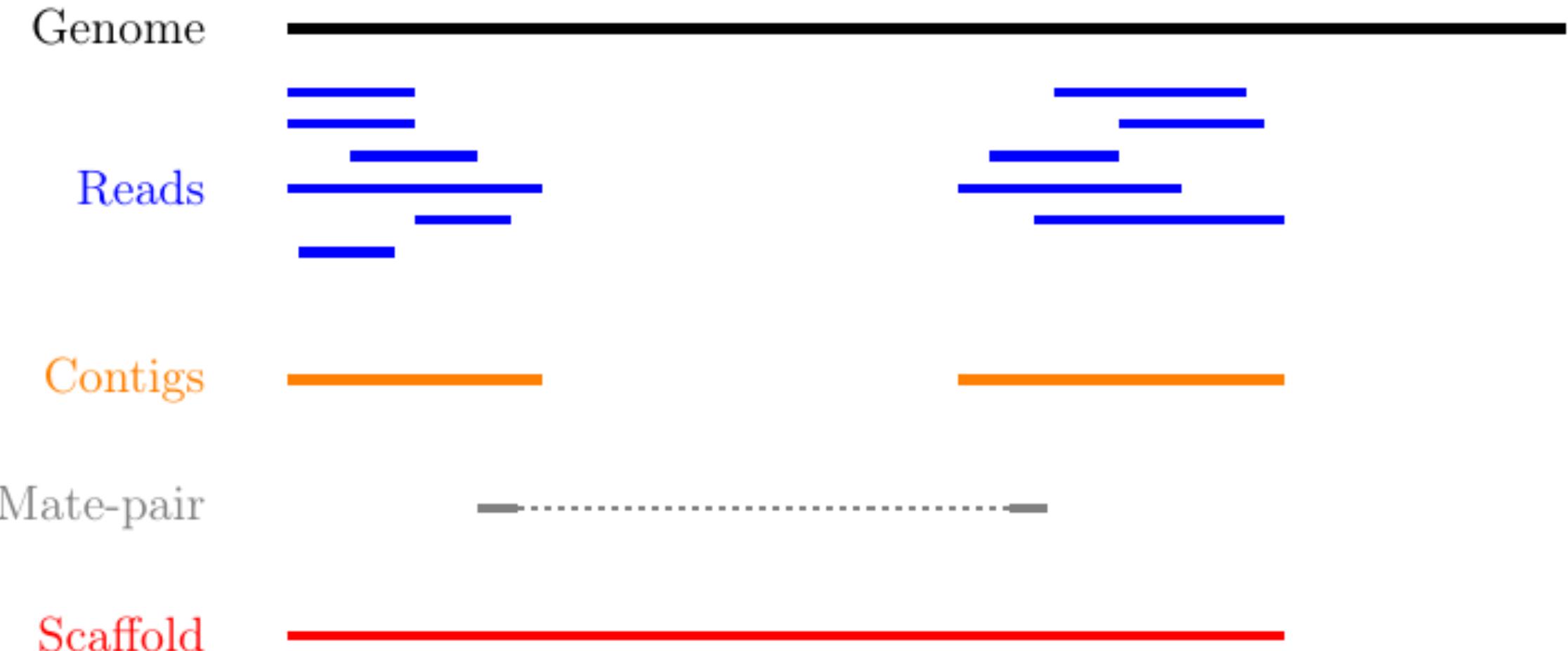


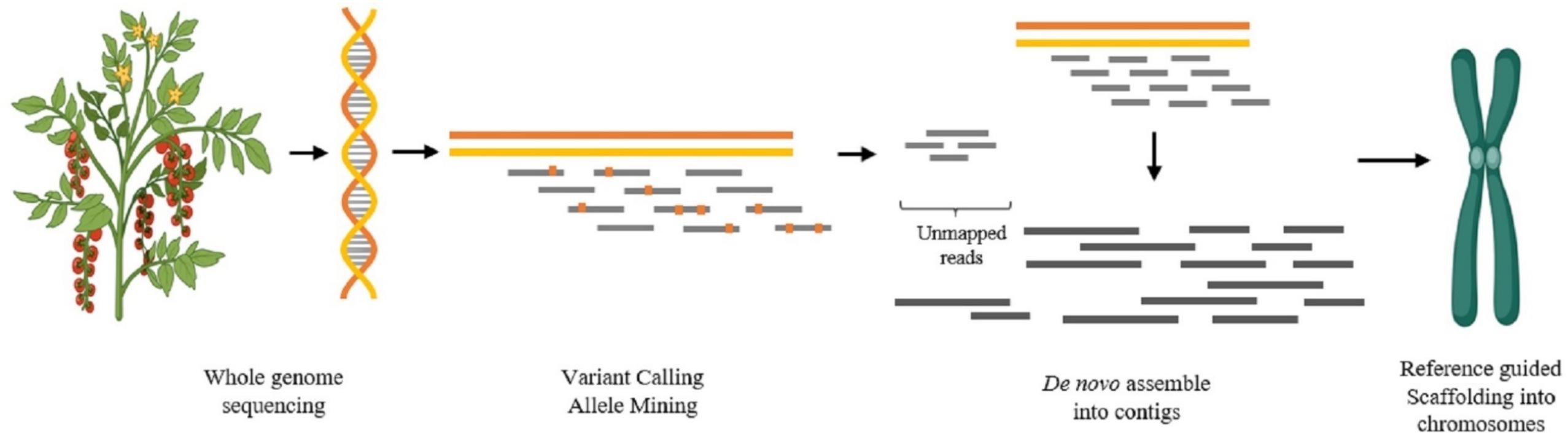


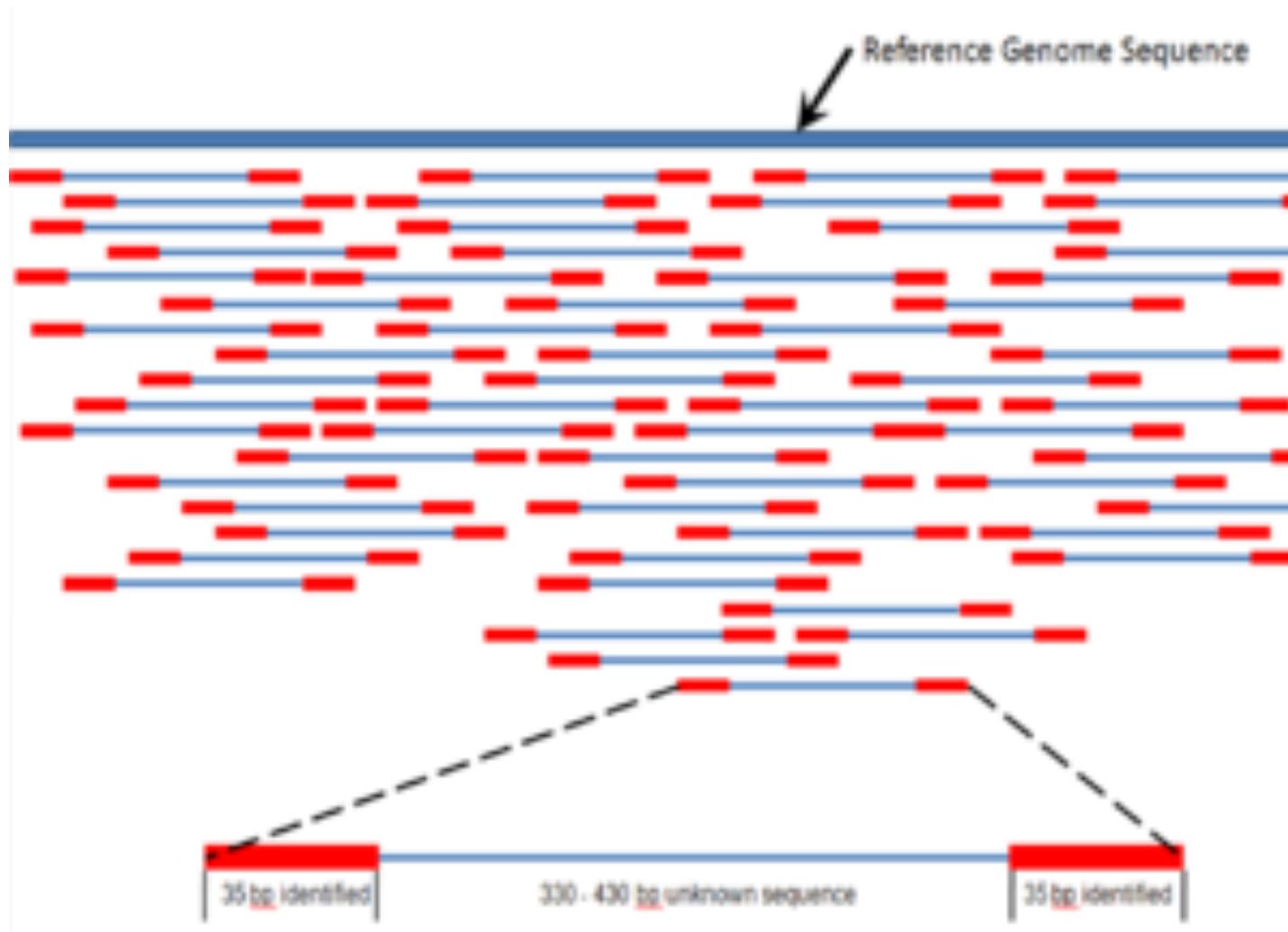
↔↔ Fragment

←→ Read (known sequence)

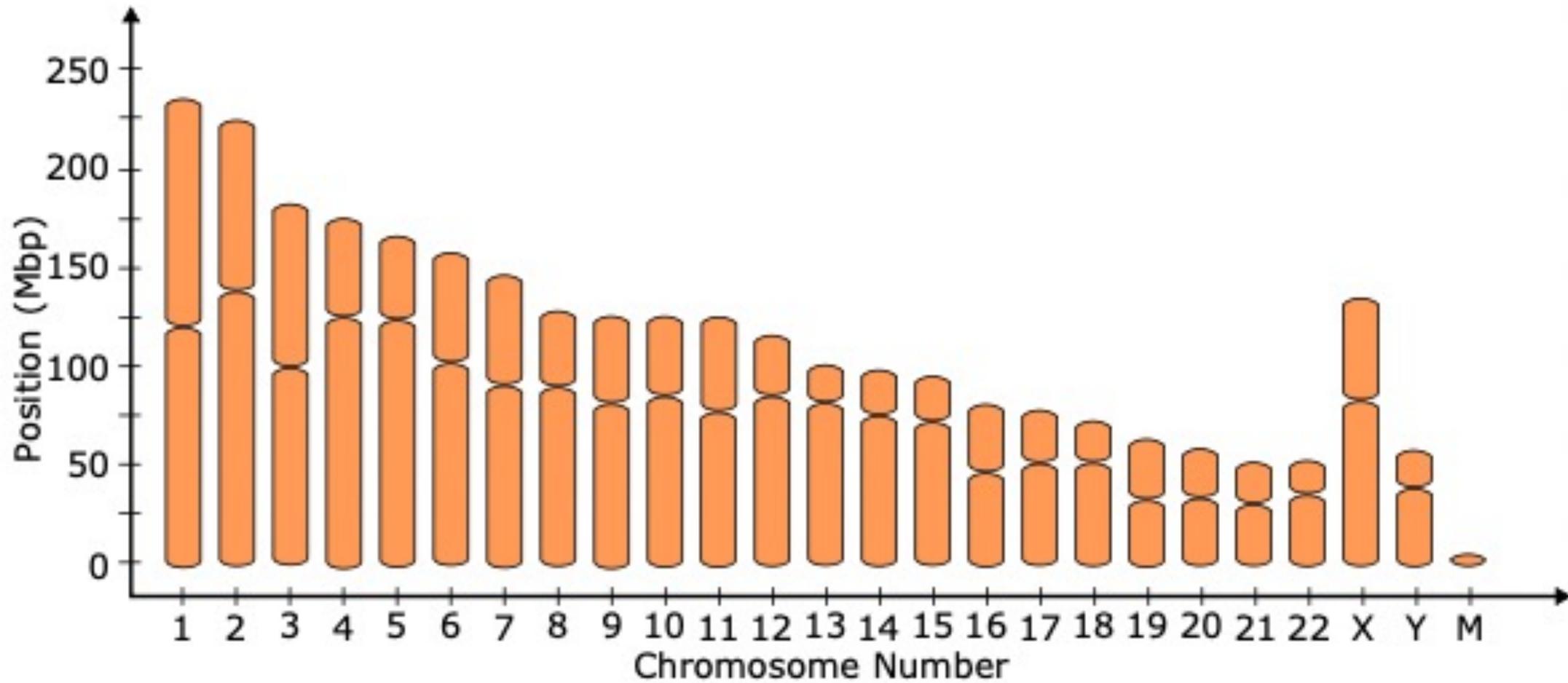
— Roughly known length but not known sequence

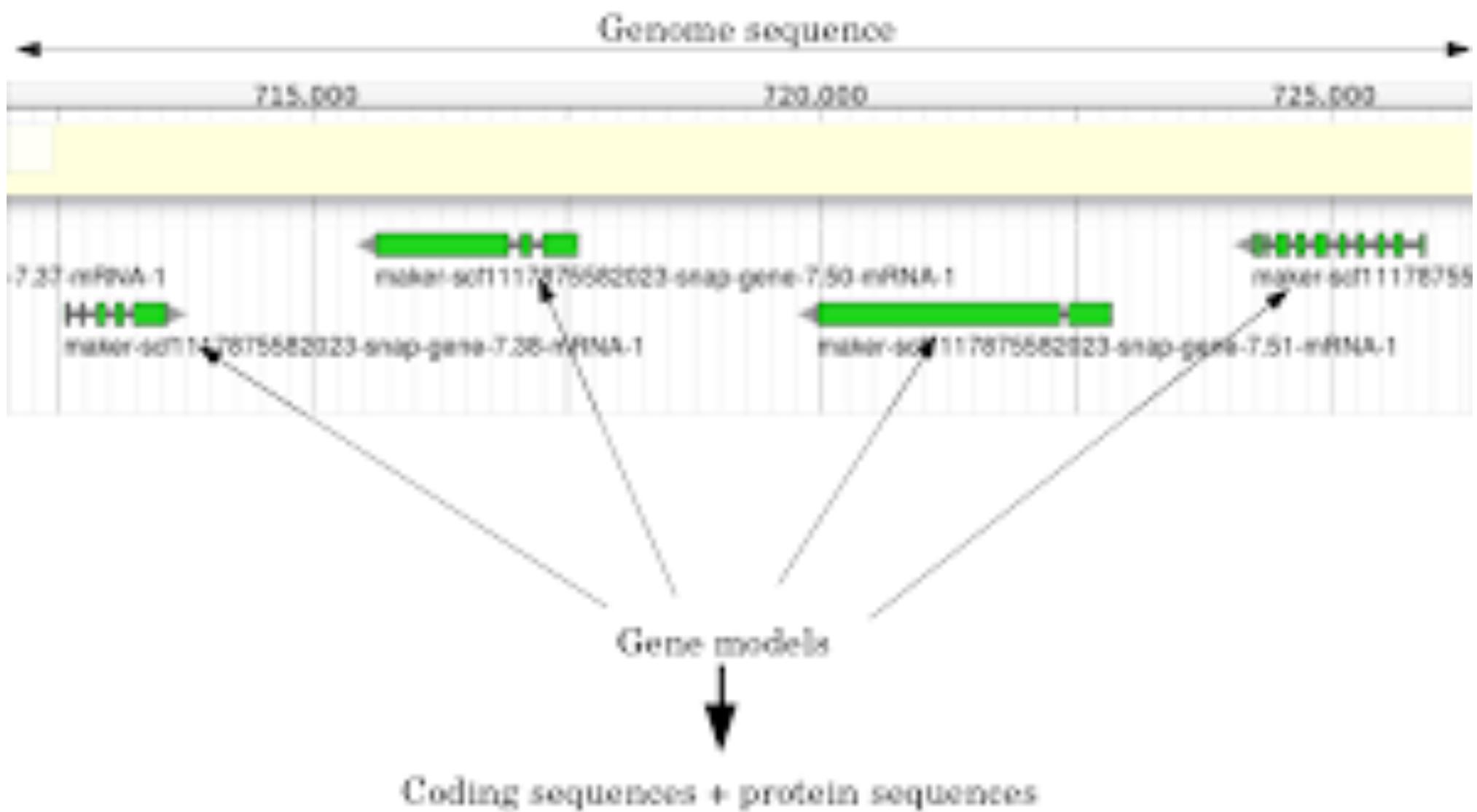


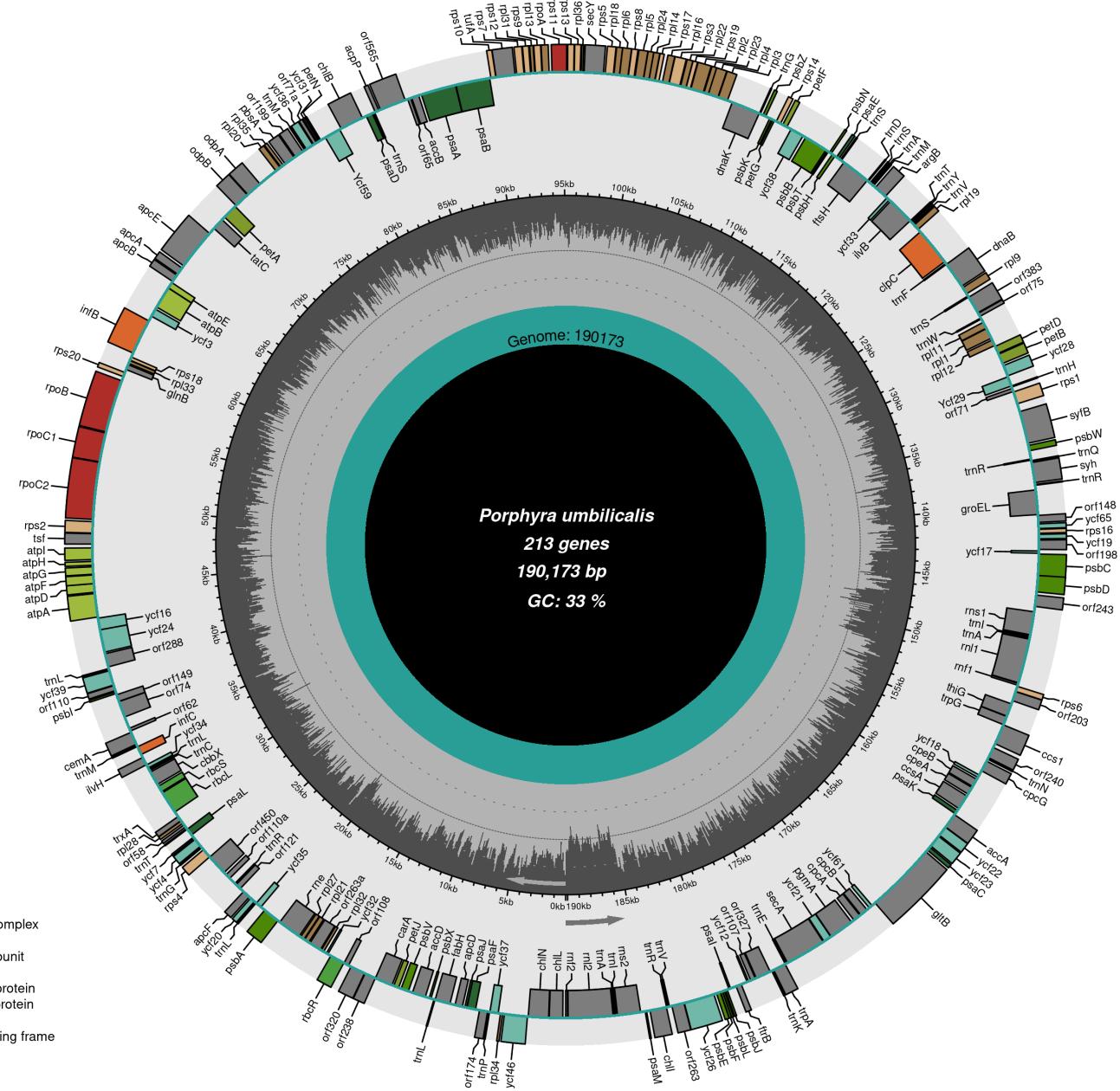




Week 5







- photosystem I
 - photosystem II
 - cytochrome b/f complex
 - ATP synthesis
 - RubisCO large subunit
 - RNA polymerase
 - small ribosomal protein
 - large ribosomal protein
 - clpP, matK, infA
 - hypothetical reading frame
 - transfer RNA
 - other

Panda DataFrame objects for working with data tables in python

My_data_tab in table format

Columns	
Animal	Color
bear	brown
swan	white
tiger	orange
frog	green
crow	black

Pandas module
for DataFrames in
python

My_data_df in DataFrame format

Keys	
Animal	Color
bear	brown
swan	white
tiger	orange
frog	green
crow	black

#Create a dictionary object

```
animal_dict = {"Animal": ["bear", "swan", "tiger", "frog",  
"crow"], "Color": ["brown", "white", "orange", "green",  
"black"] }
```

#Import the pandas module and store as pd for short
import pandas as pd

#Convert from dictionary to a dataframe object

#using the pandas function .DataFrame()

```
animal_df= pd.DataFrame(animal_dict)
```

```
print(animal_df)
```

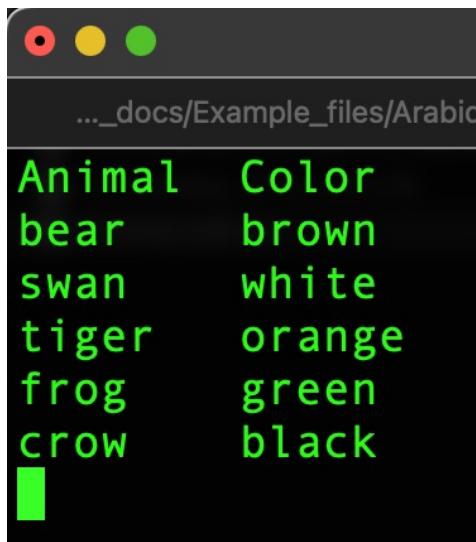
	Animal	Color
0	bear	brown
1	swan	white
2	tiger	orange
3	frog	green
4	crow	black

Tab-separated values (.tsv) format stores tables in plain-text

Looking at a table in Excel

	A	B
1	Animal	Color
2	bear	brown
3	swan	white
4	tiger	orange
5	frog	green
6	crow	black
7		
8		

Looking at a table in Vim



Animal	Color
bear	brown
swan	white
tiger	orange
frog	green
crow	black

Looking at the plain
text for a table

Animal	Color
bear	brown
swan	white
tiger	orange
frog	green
crow	black

There are ‘tab symbols’
between the items on each line

- Tab symbols are encoded by “\t” in python (and many other languages)
- Tab symbols tell programs like Excel and Pandas which values belong in separate fields
 - **Tab Separated Values (.tsv) format** is an extremely common format for storing table data

```
#Specify the Full path to Animal_data.tsv
animal_file_path="/home/hub_data_share/Examples/Mod06/Animal_tab.tsv"
```

```
# Read in the Animal_data.tsv file
my_df=pd.read_csv(animal_file_path,
delimiter= "\t")
```

```
## Explore the dataframe a little bit
```

```
#Get a list of all the column headers
print(list(my_df.keys()))
```

```
#Get the number of rows in the df
print(len(my_df.index))
```

GFF3 format files store gene annotation data in tsv format

```
>Chr5
TATACCATGTACCCCTAACCTAAACCTAAACCTATACTATAAATCTTAAACCTATACTCTAAACCATAAGGGTT
TGTGAGTTGCATAAAGTGTACGTATAAGTGTCTAACATGTGAGTTGCATAAGAGTCGACTATGTGTTGTC
AAAAGTGACGTAAGTGTAGACTAGAGCGGCCGTGAGCACAAGCGGCCAAGGCCATGCTGCGGAGATTACCTA
TAATTATGTTGCGGCTTACAATTGAAATTGTTGTTAGTGGTCAGTCAGGGATGAGTTGTTCTAAA
CATCTCAAATTCTAATCTTCAAGTGTAGGTCACTCTACCTTATTTTTTTTTTTATTAAATTCTAT
TAGAAAGAACTTGAACCTTATTCAGTGAACATTTAAAGAATTGTAACCCACCTATTCTATTATTGAATATG
TGTAAAGTAAATTTCACAATAAGCATGTATAAGTAAACATATTGTTCTTCTATAGCCTTGAAATCATAG
GGACGGATCTAGTTAGATAAAAGTGTGCATAAGTCTCGACTAACATGTGAGTCCATTGAGTTGCATAA
CTTTGCGCCTTCGACAATTGAATTGATAAGTGTGACTAATTATGTTGCTAAGTGATTGACAT
TTCGACTAAGTGTGATTGTCATTGCGTCTGCGATCCGGAAAGCGTGTGTTGTTCTGTCTTACT
TGCCGGGTTATGAAGTCCTTGTGAGAGATCTGCTCATCATGATAGAAGATGGTAGATGTTATCTGGATTGC
CGCGGGTTGAATTGCGGGCATCTTAAATCTCATCACATGTTGAGGAAGAGTTTTGTTTTA
TTATTATTGTTCCCTAACAGATTATACTTATTCTCTATTGAGGACTTTATTTGCTGTTGGCTTAG
```

A genome sequence in fasta format:

- How do we know where the genes are??

GFF3 (.tsv) format

seqid	source	type	start	end	strand	attributes
Chr5	phyt ozonev12	gene	995	5156	-	ID=AT5G01010.Araport11.447;Name=AT5G01010
Chr5	phyt ozonev12	mRNA	995	4994	-	ID=AT5G01010.2.Araport11.447;Name=AT5G01010.2;pacid=37423170;longest=1;Parent=AT5G01010.Araport11.447
Chr5	phyt ozonev12	CDS	4765	4924	-	ID=AT5G01010.2.Araport11.447.CDS.1;Parent=AT5G01010.2.Araport11.447;pacid=37423170
Chr5	phyt ozonev12	five_prime_UTR	4925	4994	-	ID=AT5G01010.2.Araport11.447.five_prime_UTR.1;Parent=AT5G01010.2.Araport11.447;pacid=37423170
Chr5	phyt ozonev12	CDS	4552	4679	-	ID=AT5G01010.2.Araport11.447.CDS.2;Parent=AT5G01010.2.Araport11.447;pacid=37423170
Chr5	phyt ozonev12	CDS	4335	4467	-	ID=AT5G01010.2.Araport11.447.CDS.3;Parent=AT5G01010.2.Araport11.447;pacid=37423170
Chr5	phyt ozonev12	CDS	4102	4258	-	ID=AT5G01010.2.Araport11.447.CDS.4;Parent=AT5G01010.2.Araport11.447;pacid=37423170
Chr5	phyt ozonev12	CDS	3927	4005	-	ID=AT5G01010.2.Araport11.447.CDS.5;Parent=AT5G01010.2.Araport11.447;pacid=37423170
Chr5	phyt ozonev12	CDS	3762	3802	-	ID=AT5G01010.2.Araport11.447.CDS.6;Parent=AT5G01010.2.Araport11.447;pacid=37423170
Chr5	phyt ozonev12	CDS	3543	3659	-	ID=AT5G01010.2.Araport11.447.CDS.7;Parent=AT5G01010.2.Araport11.447;pacid=37423170