

Phylogenetics

Nichola Hill, Postdoc – Runstadler Lab

Email: nhill@mit.edu

I am a wildlife disease ecologist

My favorite organisms to study
are wild birds & viruses

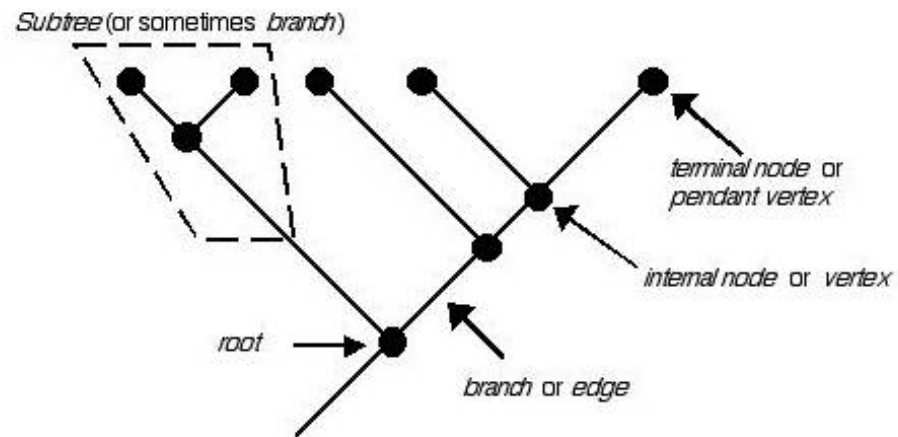


Tree

thinking

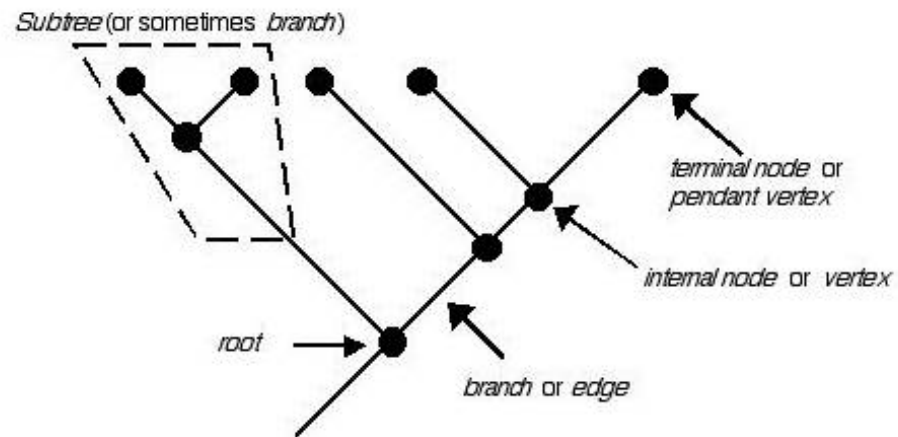
Phylogenetic trees...

- are graphs with **nodes & edges**
- organisms are connected by the passage of genes along branches of the tree
- models evolution as a bifurcating process



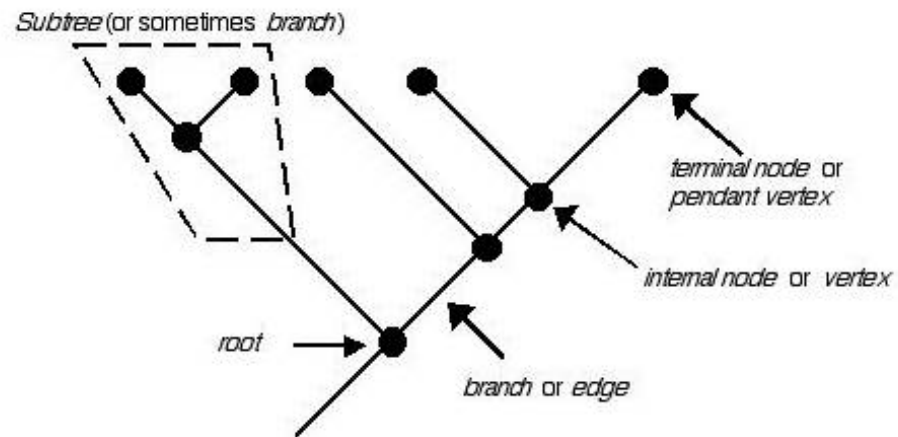
Phylogenetic trees...

- are graphs with **nodes** & **edges**
- organisms are connected by the passage of **genes** along branches of the tree
- models evolution as a bifurcating process



Phylogenetic trees...


- are graphs with **nodes** & **edges**
- organisms are connected by the passage of **genes** along branches of the tree
- models evolution as a **bifurcating** process




Phylogeny

Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.

Time

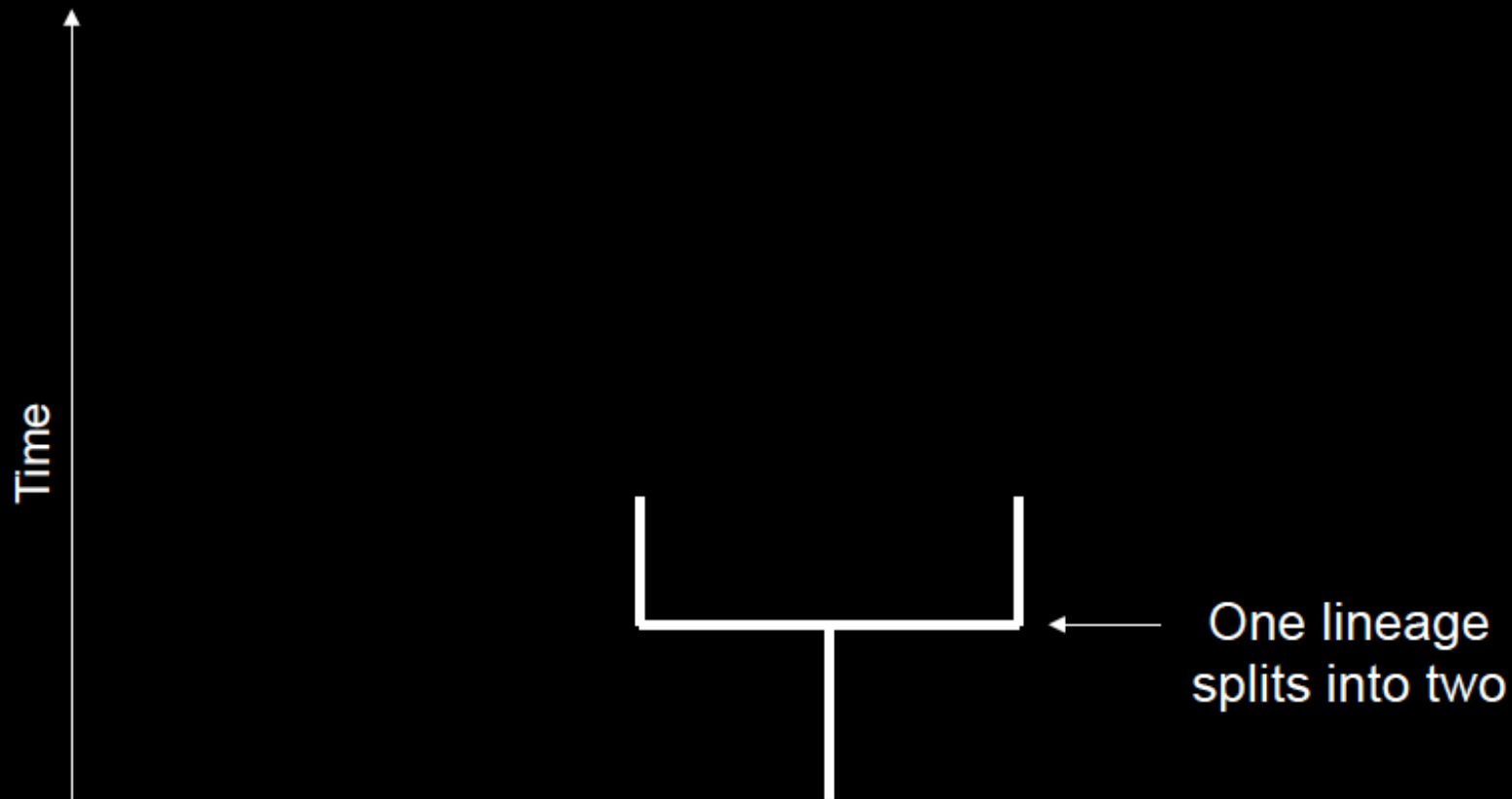


Consider an ancestral lineage
(e.g., descendants from one HIV virus)



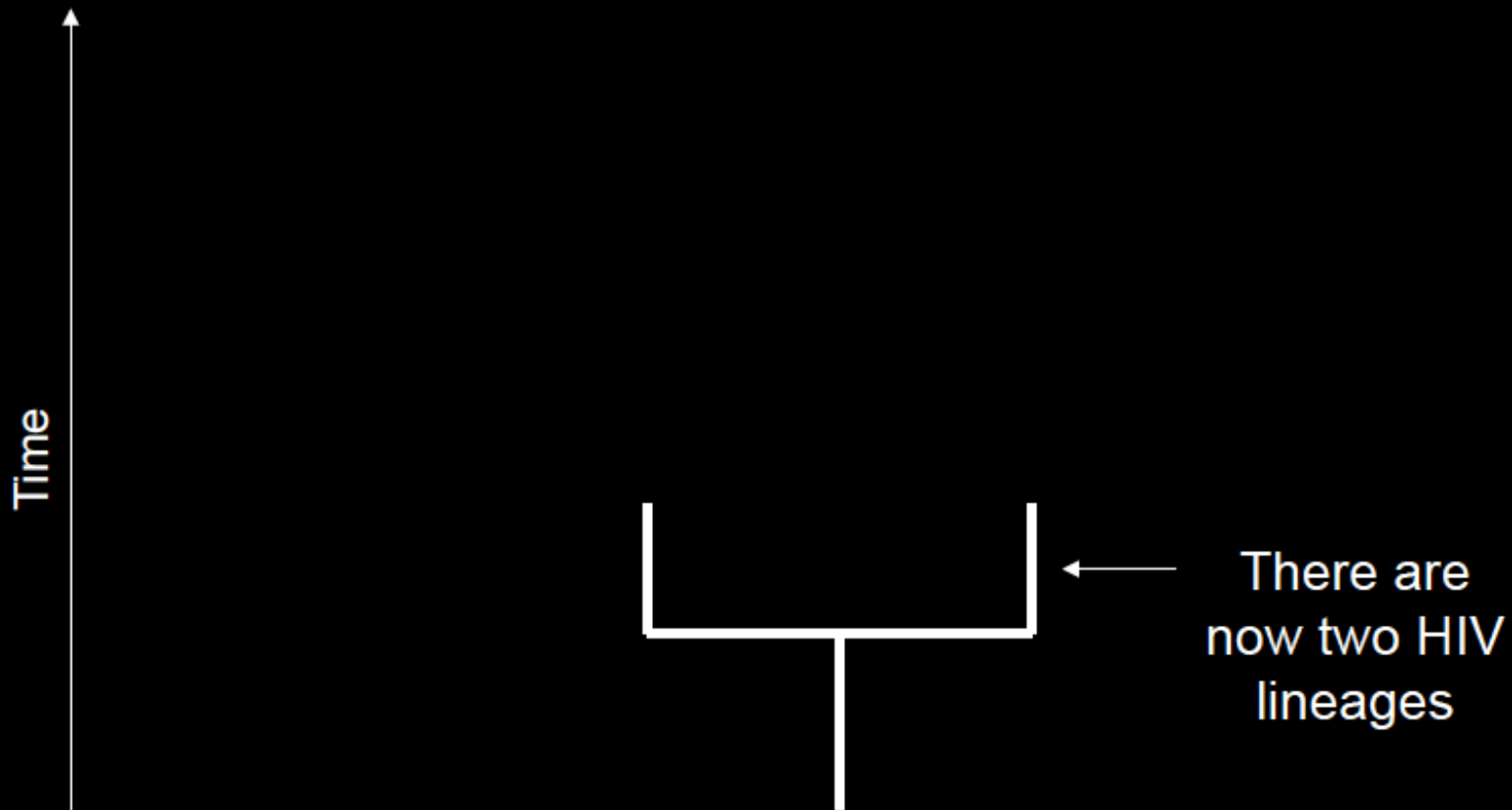
Phylogeny

Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.



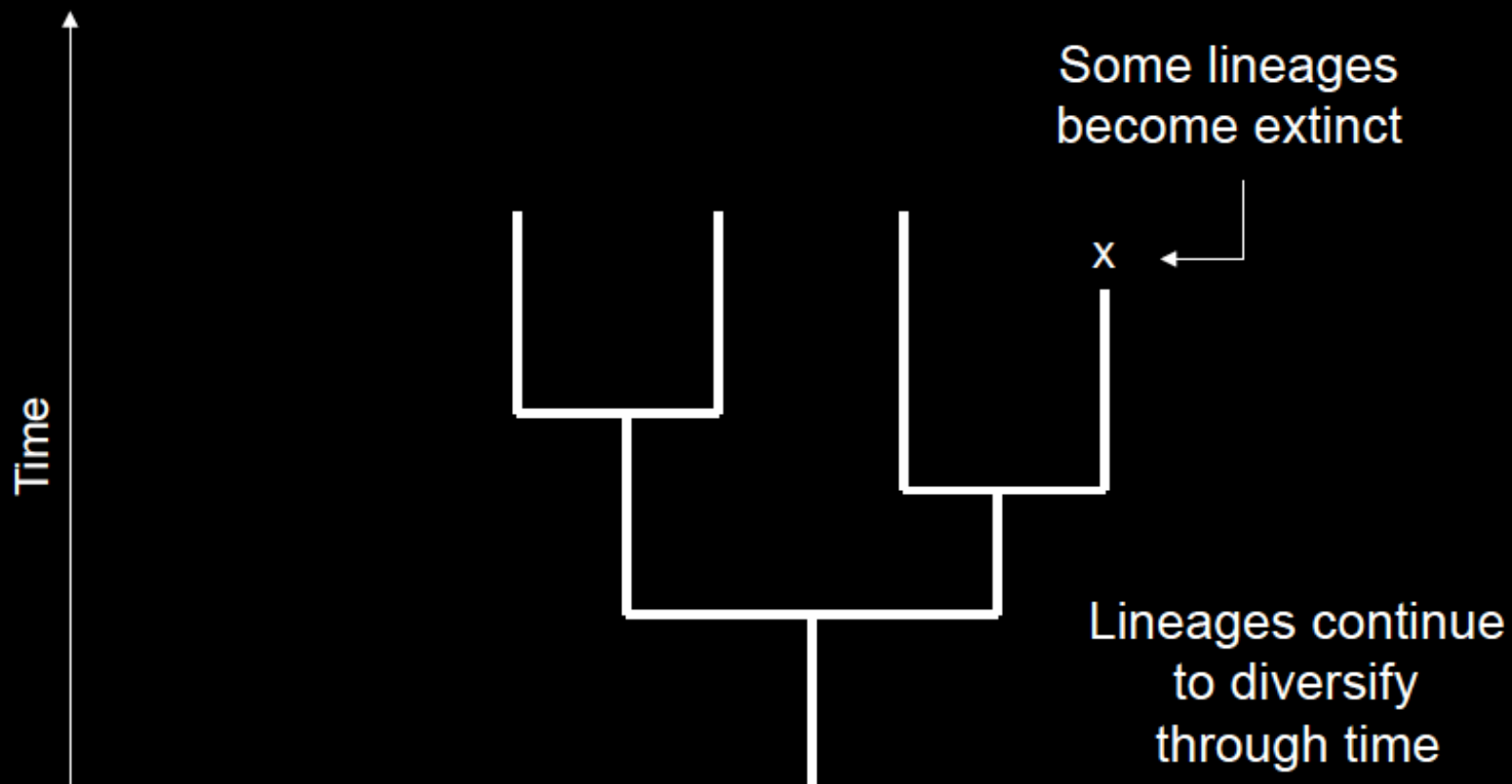
Phylogeny

Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.



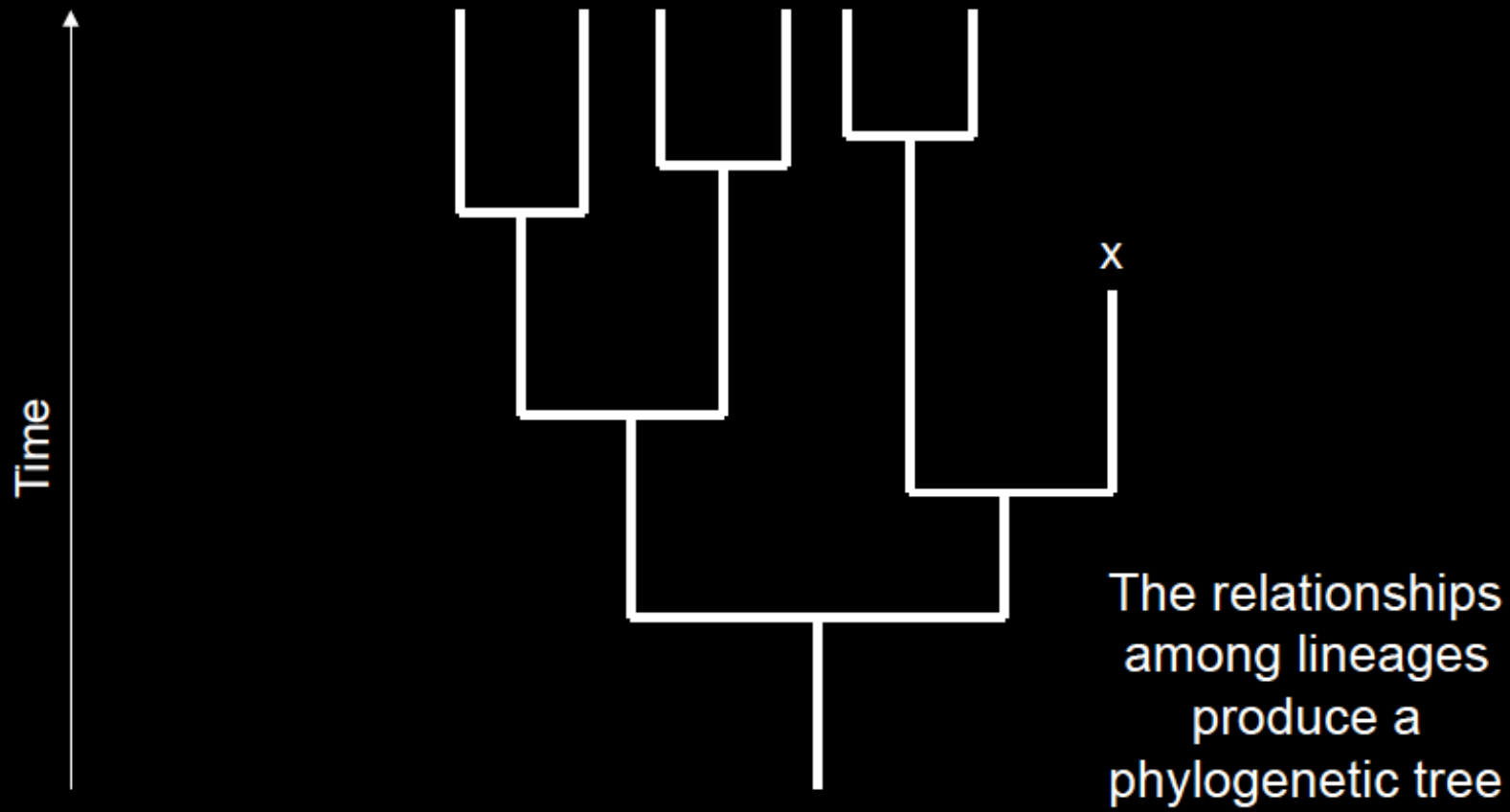
Phylogeny

Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.



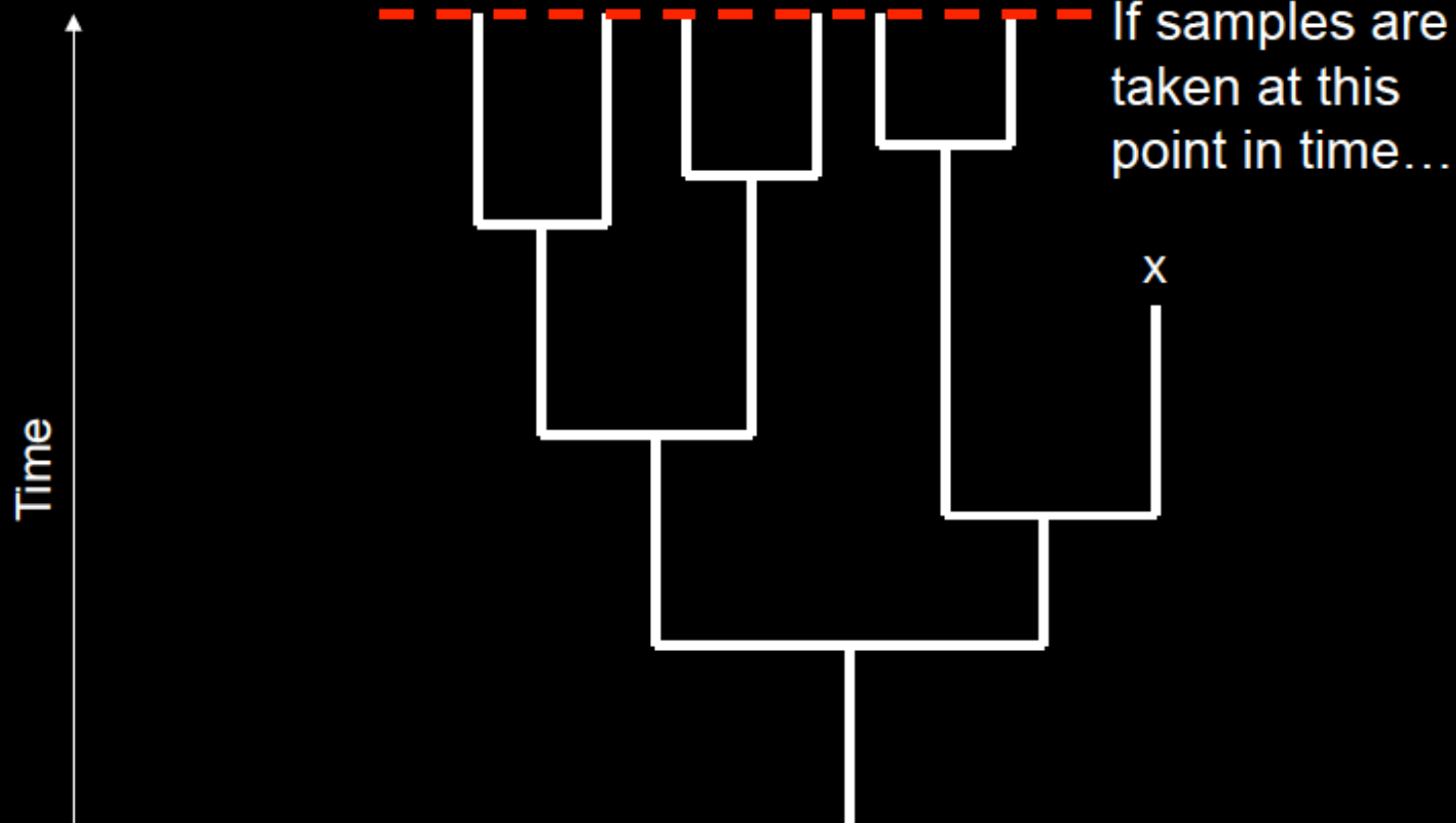
Phylogeny

Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.



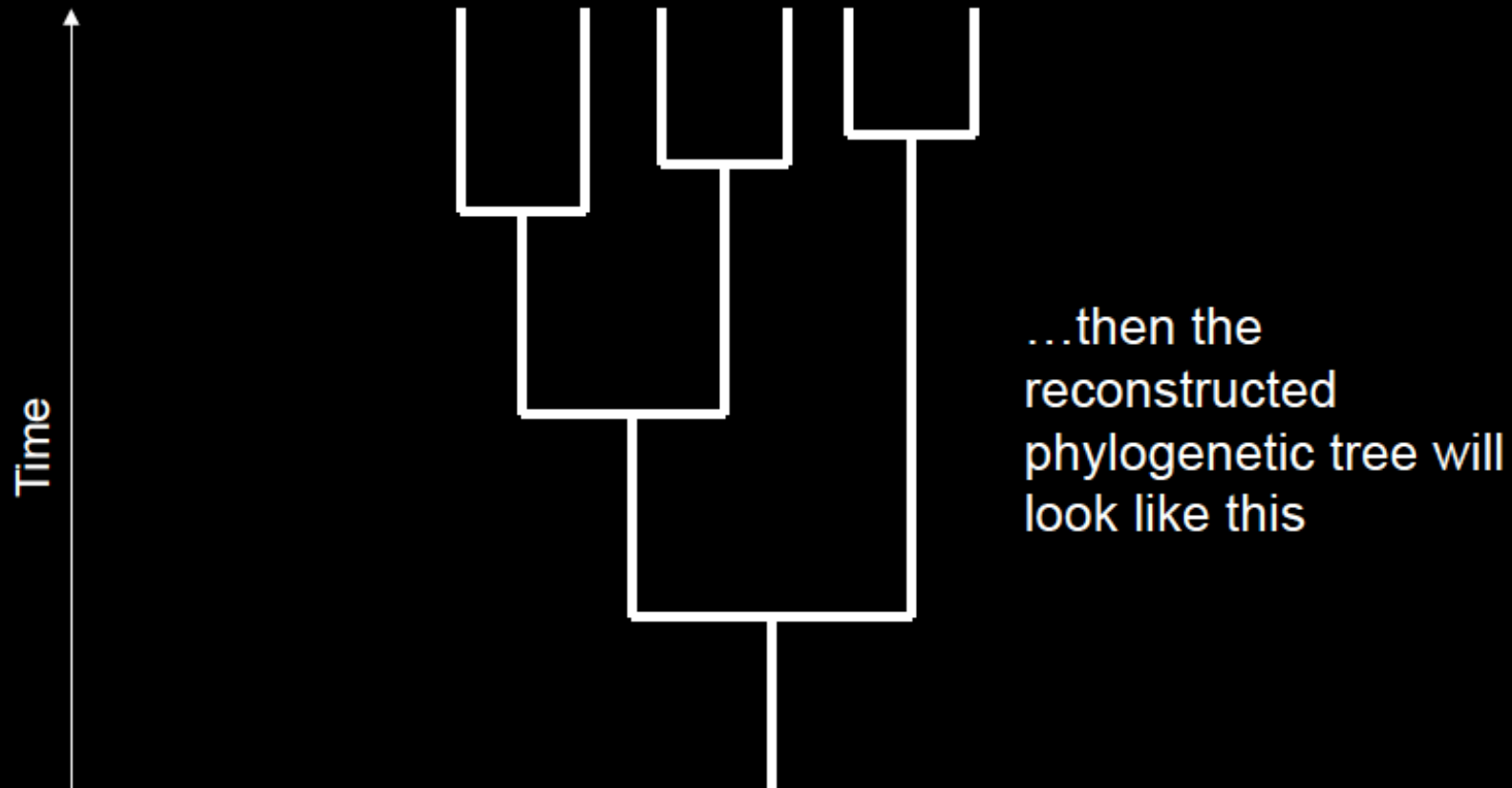
Phylogeny

Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.



Phylogeny

Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.



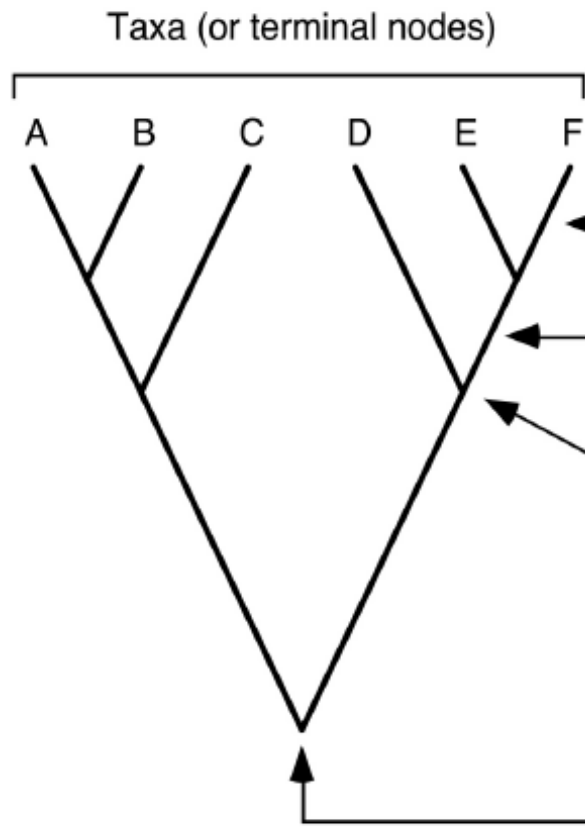
Phylogenetic trees are useful for **inferring** evolutionary relationships...

...but usefulness is influenced by **sampling** (i.e. how well the samples represent the population)

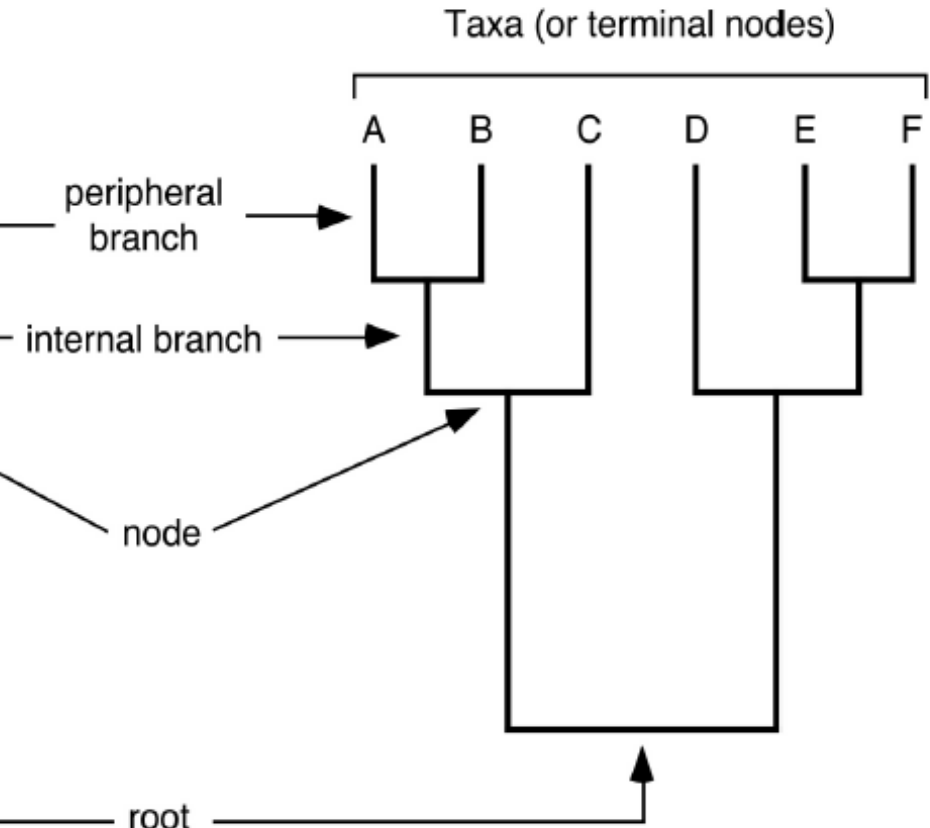
Tree terminology

Terminology for Trees

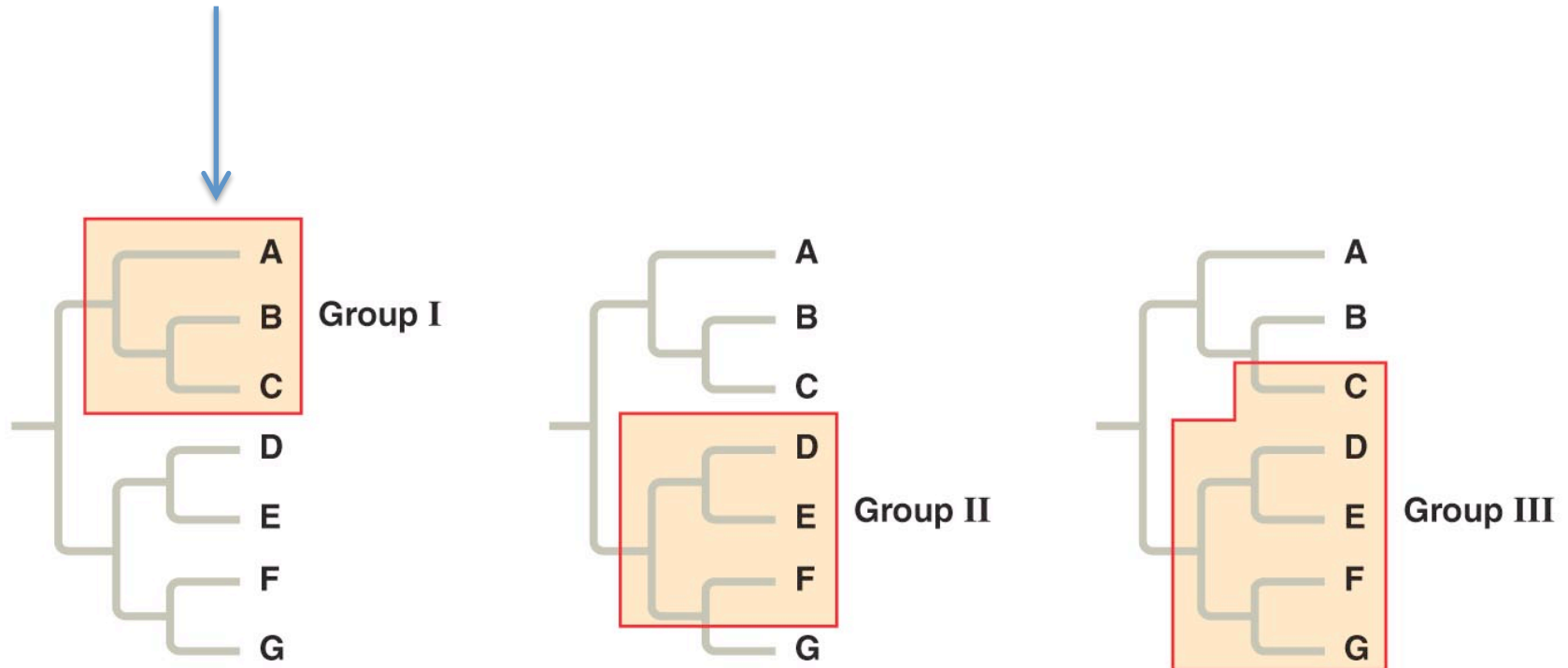
(a)



(b)



Monophyletic

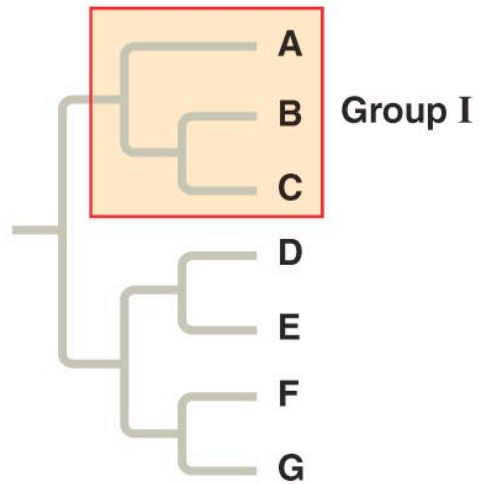


(a) Monophyletic group (clade)

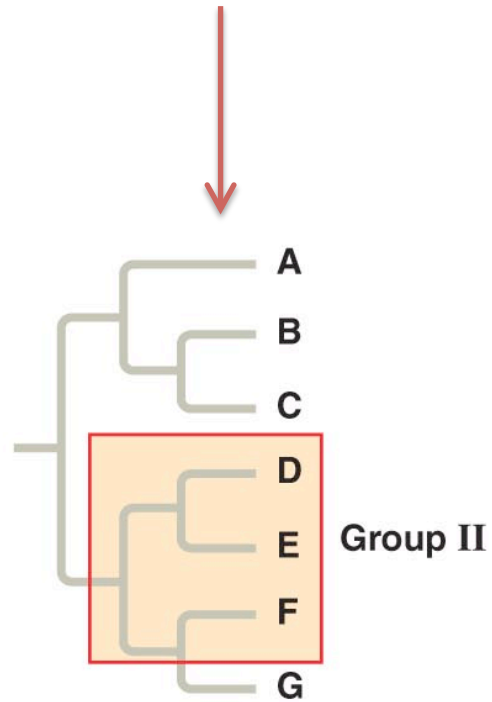
(b) Paraphyletic group

(c) Polyphyletic group

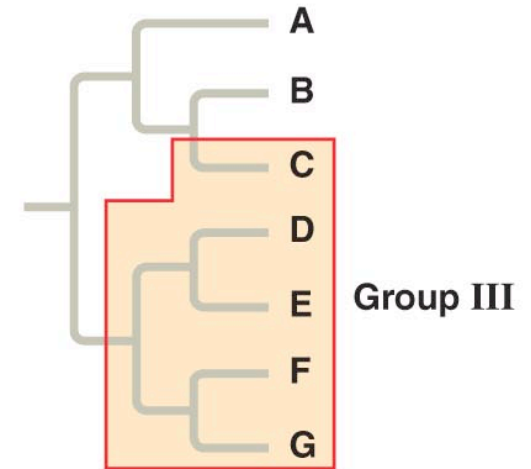
Paraphyletic



(a) Monophyletic group (clade)

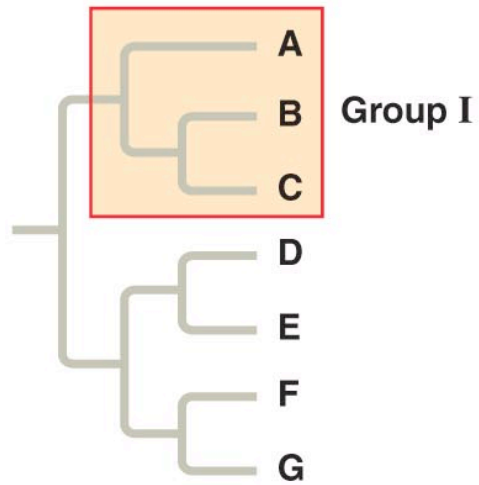


(b) Paraphyletic group

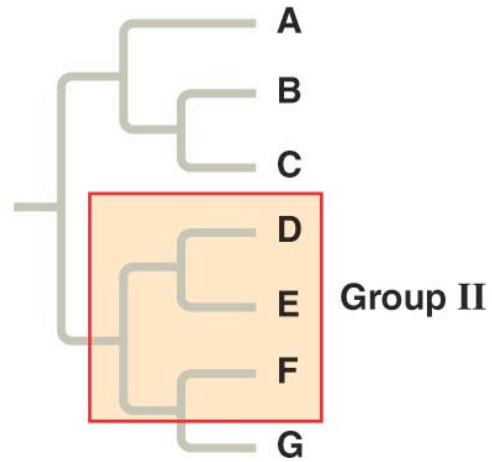


(c) Polyphyletic group

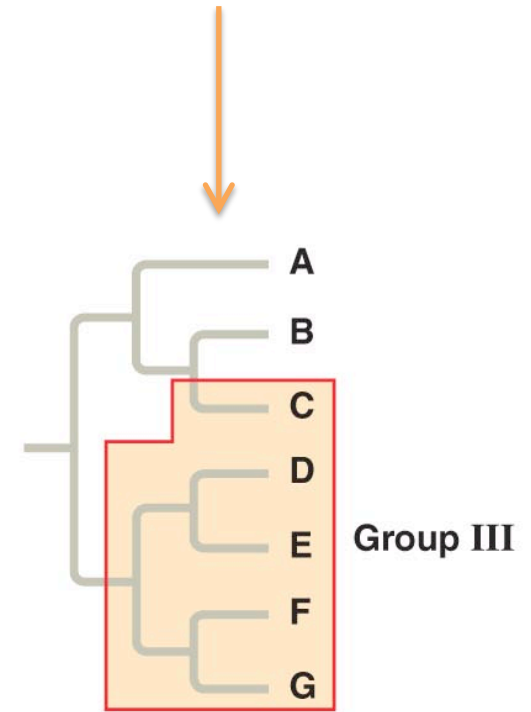
Polyphyletic



(a) Monophyletic group (clade)

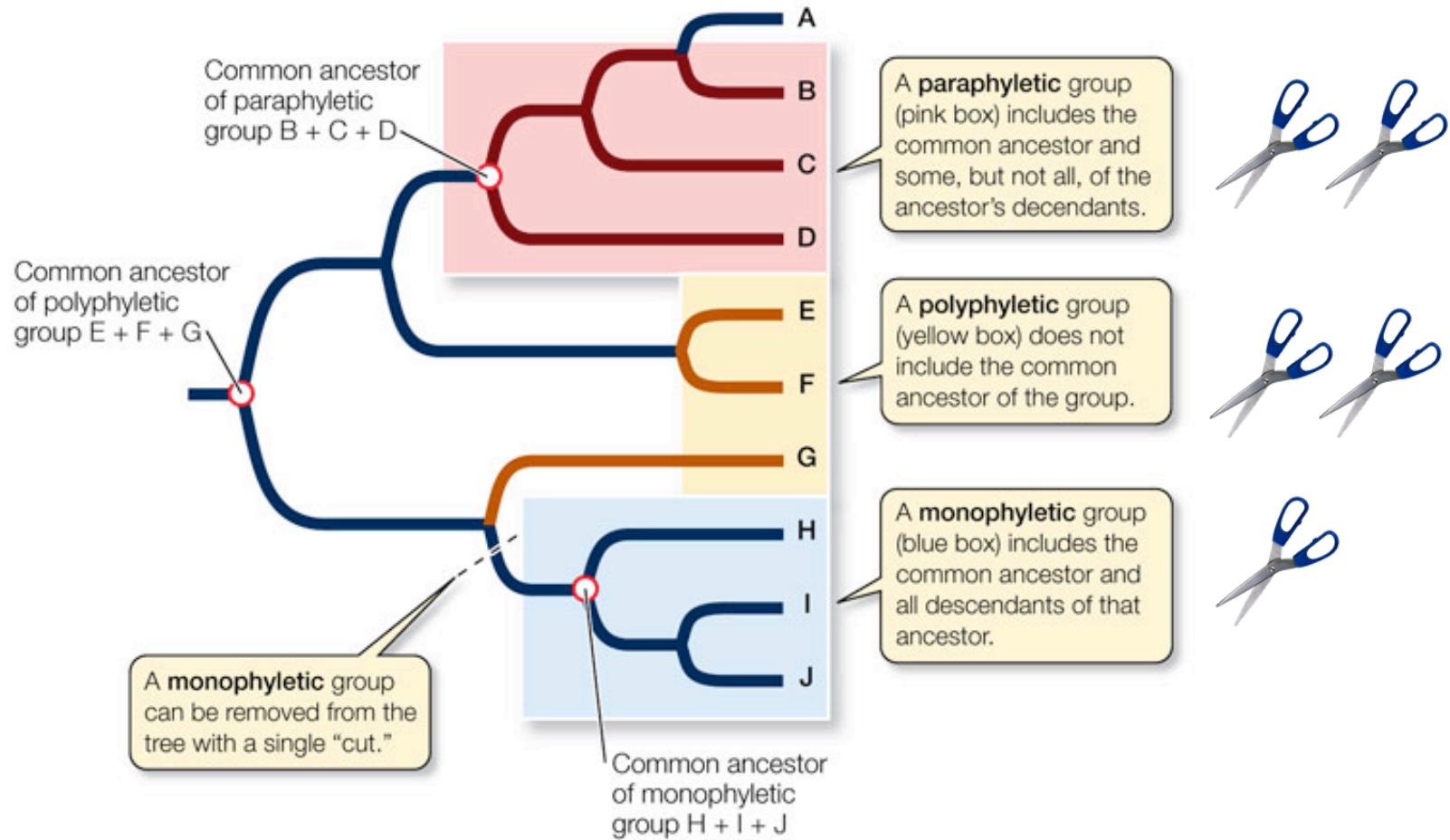


(b) Paraphyletic group



(c) Polyphyletic group

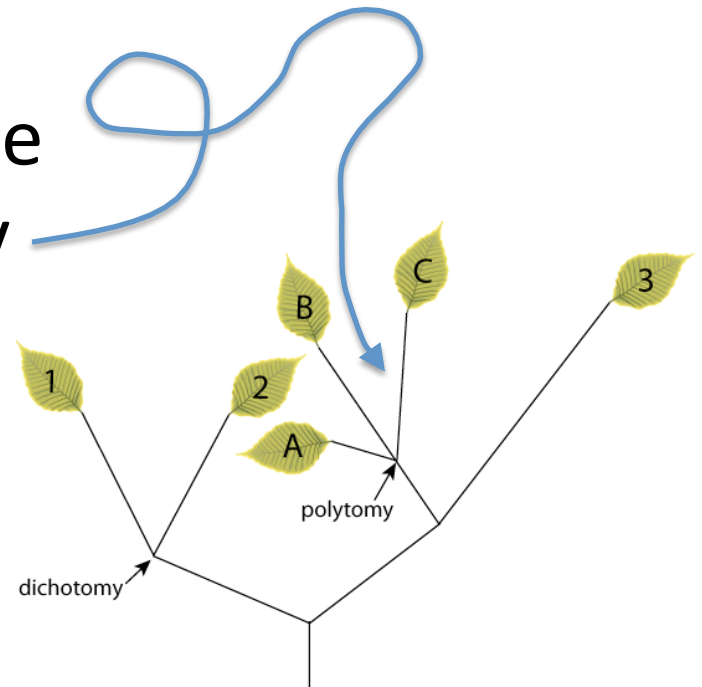
Cutting with scissors



Branching

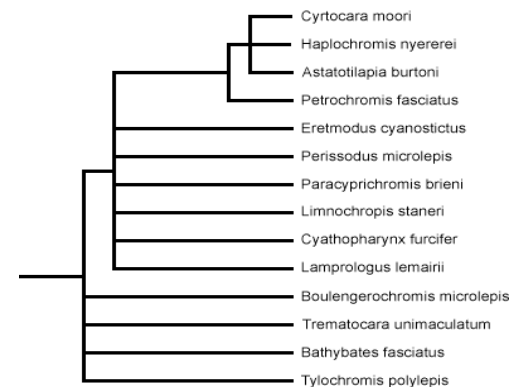
Binary branching

- A fully binary tree is called 'fully resolved'
- What about cases where one lineage splits simultaneously into multiple descendants?



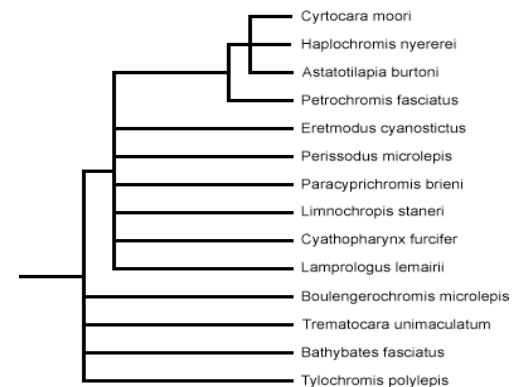
Deviations from binary branching

- **Polytomy** a node with >2 descendent lineages
 - Hard polytomy = real speciation event involving > 2 lineages diverging from a common ancestor
 - Soft polytomy = insufficient phylogenetic information, uncertain tree topology (uh oh)

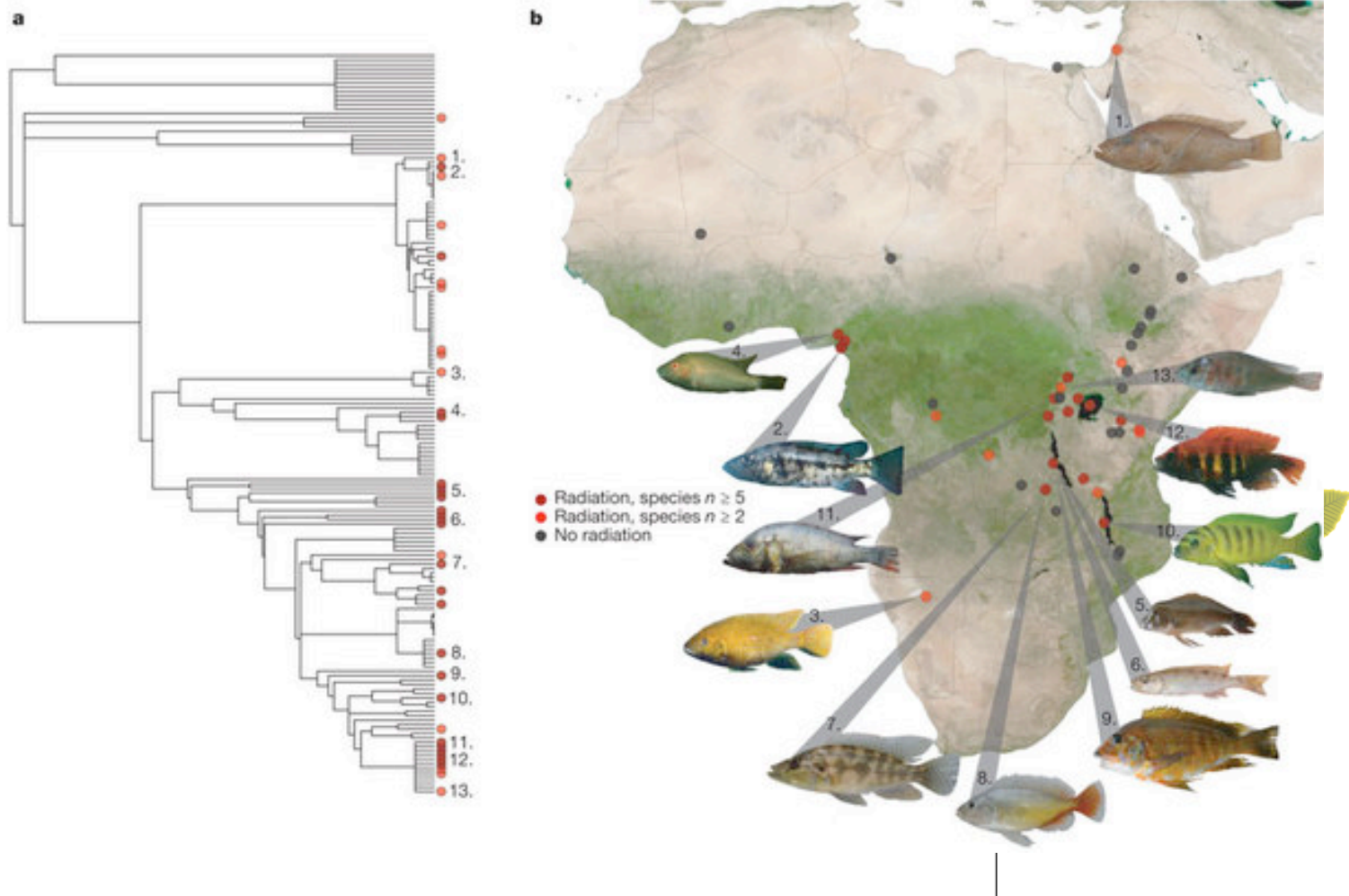


Deviations from binary branching

- **Polytomy** a node with >2 descendent lineages
 - **Hard polytomy** = real speciation event involving > 2 lineages diverging from a common ancestor
 - Soft polytomy = insufficient phylogenetic information, uncertain tree topology (uh oh)



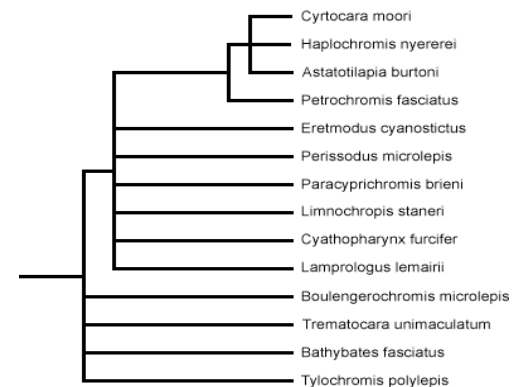
Deviations from binary branching



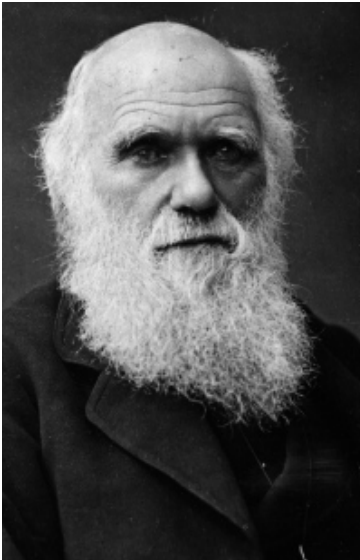
Wagner et al. (2012) Ecological opportunity and sexual selection together predict adaptive radiation. Nature: doi:10.1038/nature11144

Deviations from binary branching

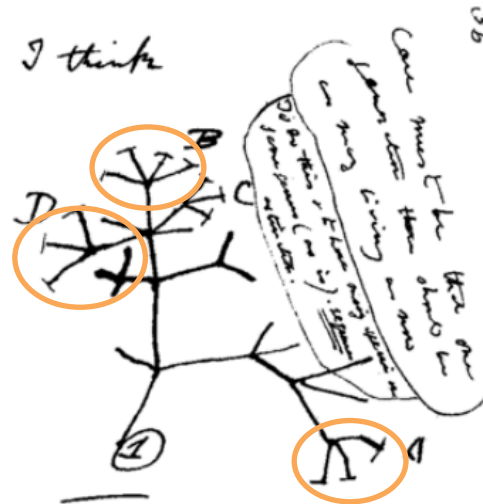
- **Polytomy** a node with >2 descendent lineages
 - Hard polytomy = real speciation event involving > 2 lineages diverging from a common ancestor
 - **Soft polytomy** = insufficient phylogenetic information, uncertain tree topology (uh oh)



The Origin of Species

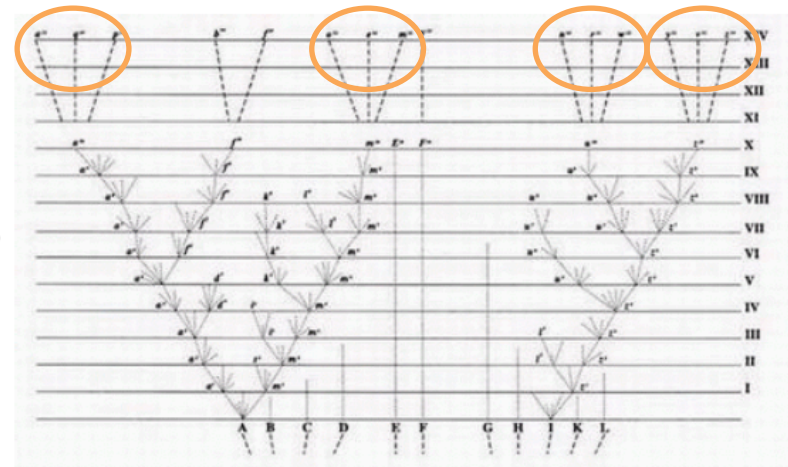


Father of evolution:
Chuck D.



There between A & B. various
 sort of relation. C + B. The
 first gradation, B & D
 rather greater distinction
 than genera would be
 formed. - binary relation

Darwin's notebooks contain the sketch (left, from 1837) that was the basis for the only figure in *The Origin of Species* (below, 1859), a conceptual drawing of a phylogenetic tree

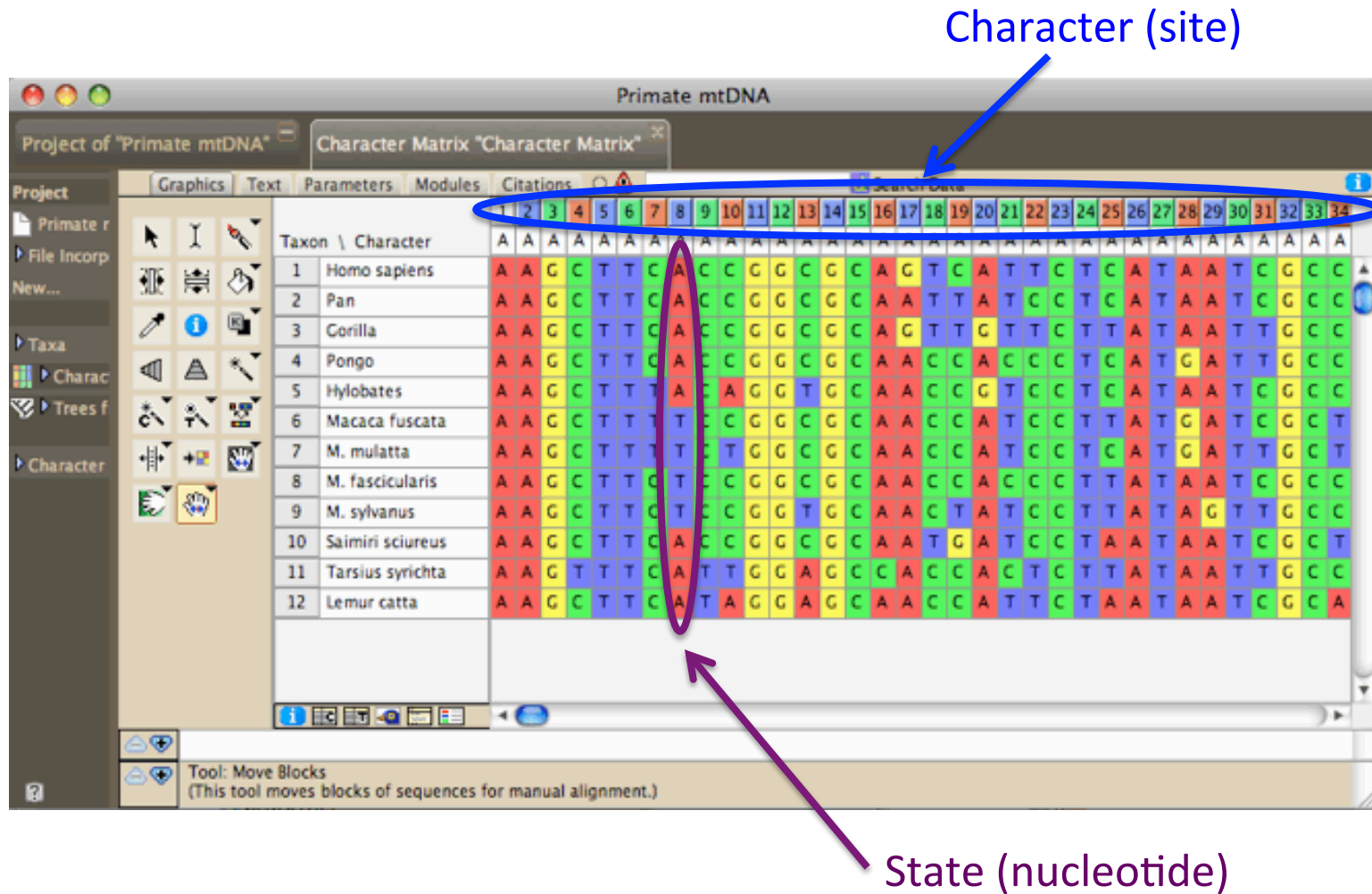


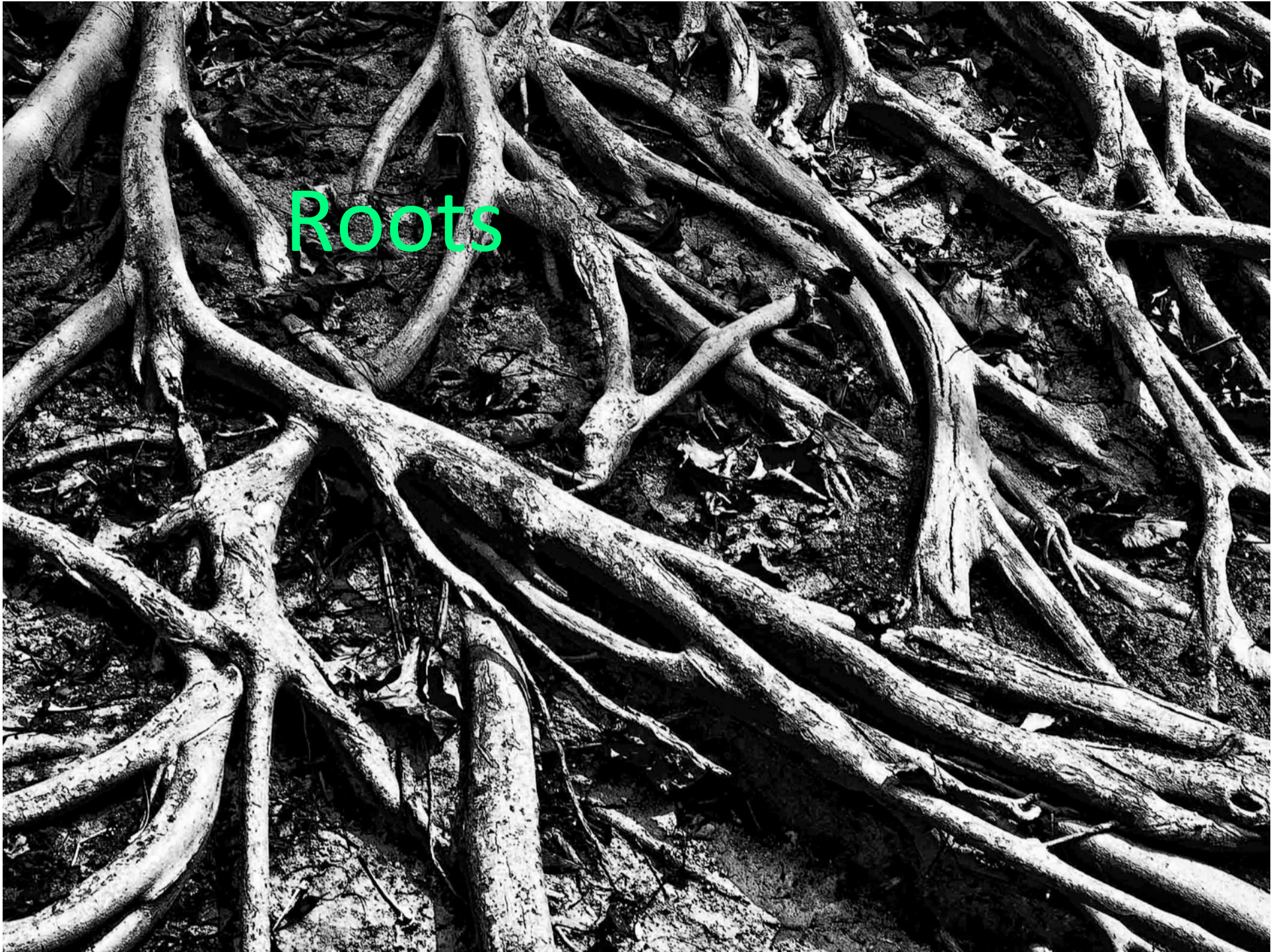
Characters vs States

Characters vs States

- **Character** is an attribute that can potentially vary at the tips (ie. hair color)
- **State** are alternative versions of the same character (ie. black, brown, blonde)

Example: DNA sequence

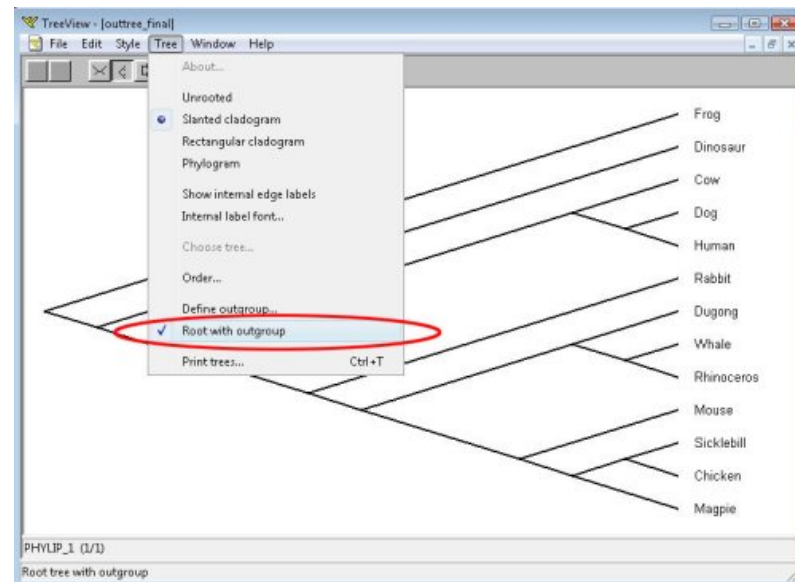




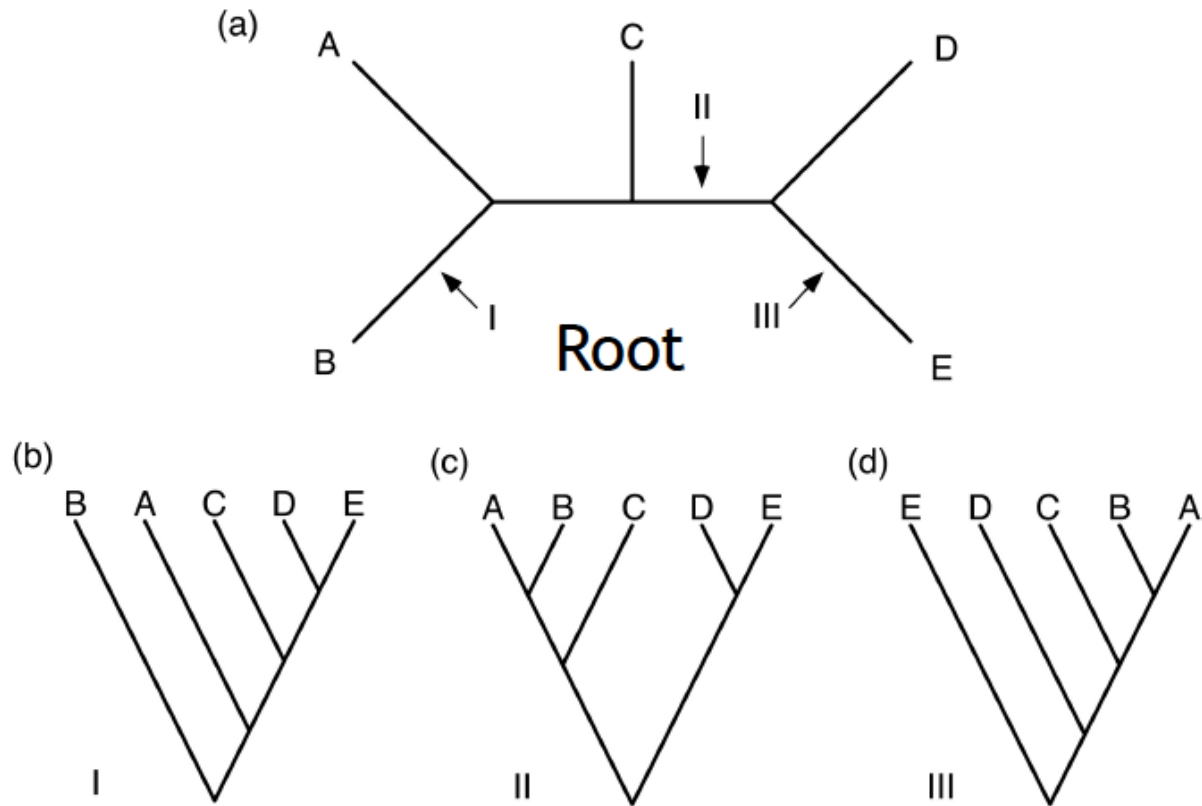
Roots

Rooting trees

- Trees can be rooted or unrooted
- Rooted trees indicate flow of time i.e. **time-calibrated tree**
- An **outgroup** is often used to root (ie. taxa known to be distantly related to ingroup)
- *One node* between outgroup and ingroup is identified as the root

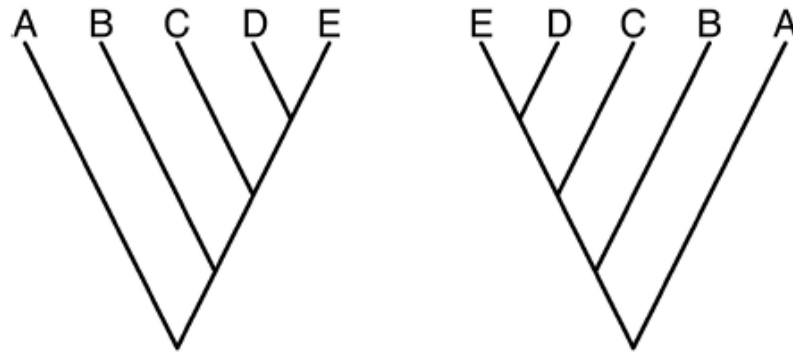


Rooting trees

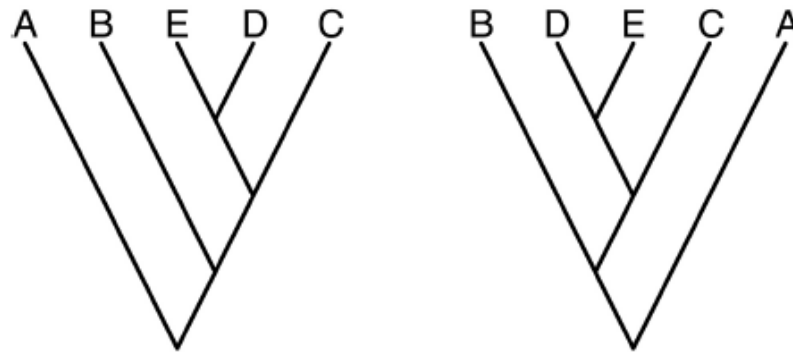


Flipping branches

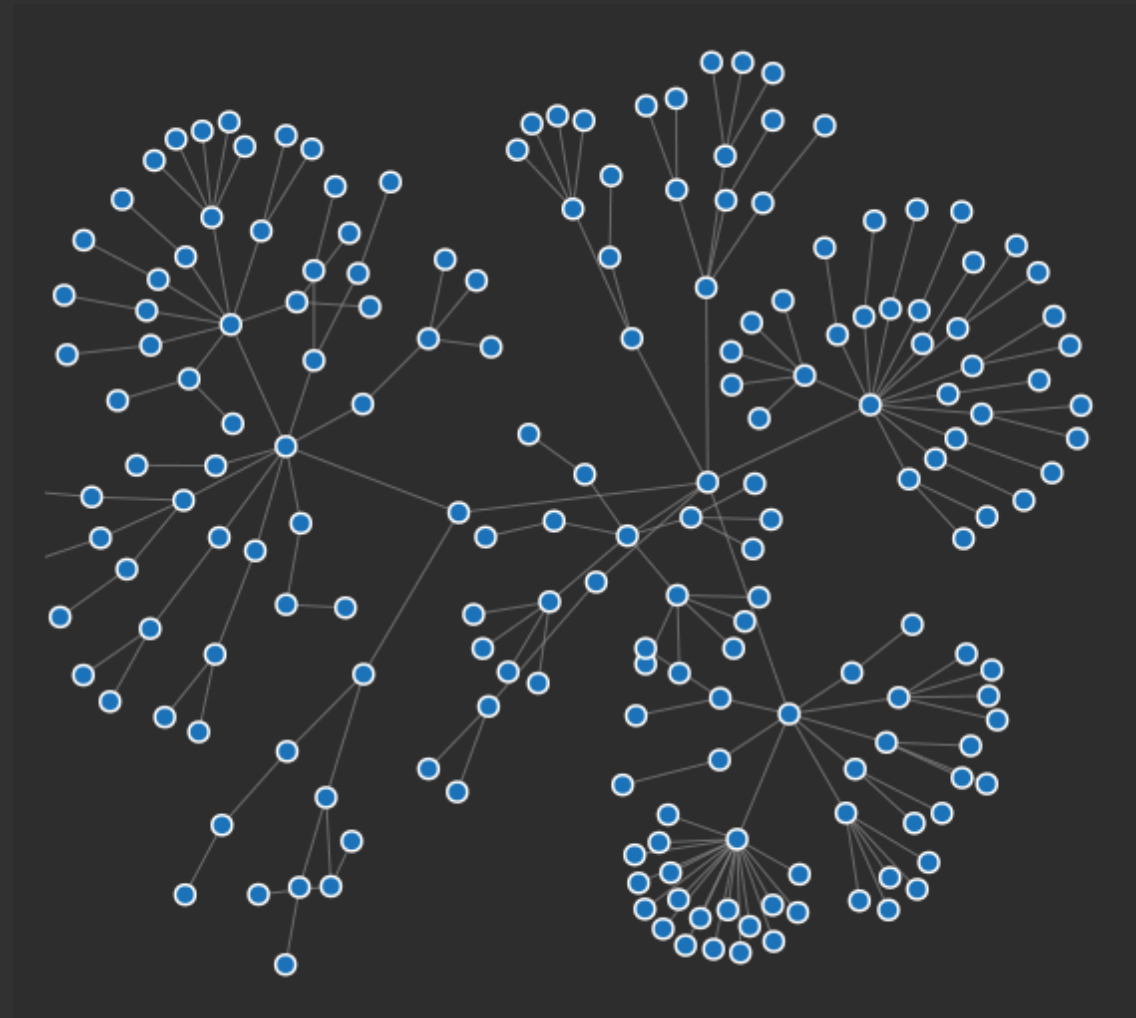
Which are Different?



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



Dynamic trees



Tree
building

How to build a phylogenetic tree

1

- Collect data i.e. DNA

2

- Retrieve homologous sequences

3

- Multiple sequence alignment

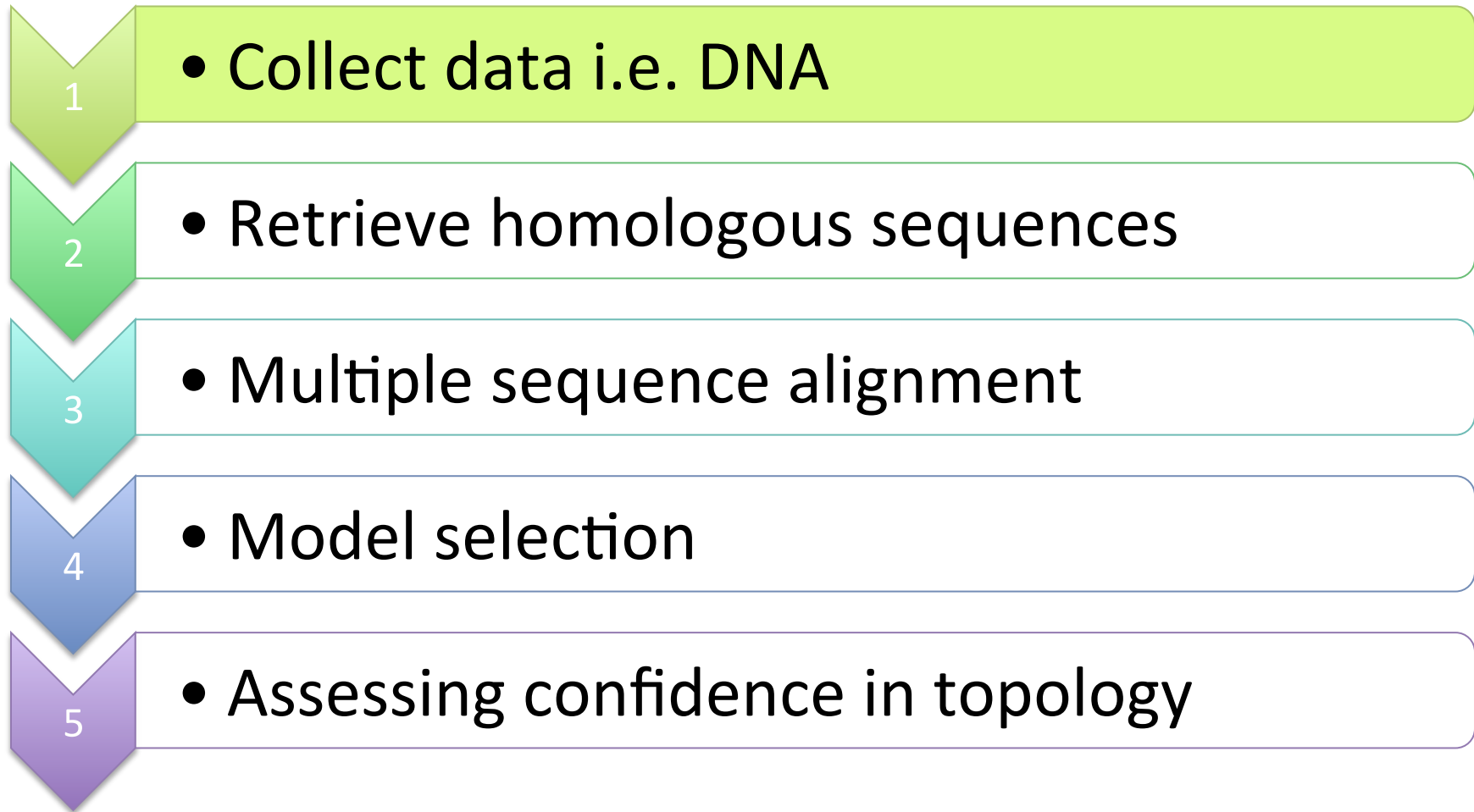
4

- Model selection

5

- Assessing confidence in topology

How to build a phylogenetic tree



How to build a phylogenetic tree

1

- Collect data i.e. DNA

2

- Retrieve homologous sequences

3

- Multiple sequence alignment

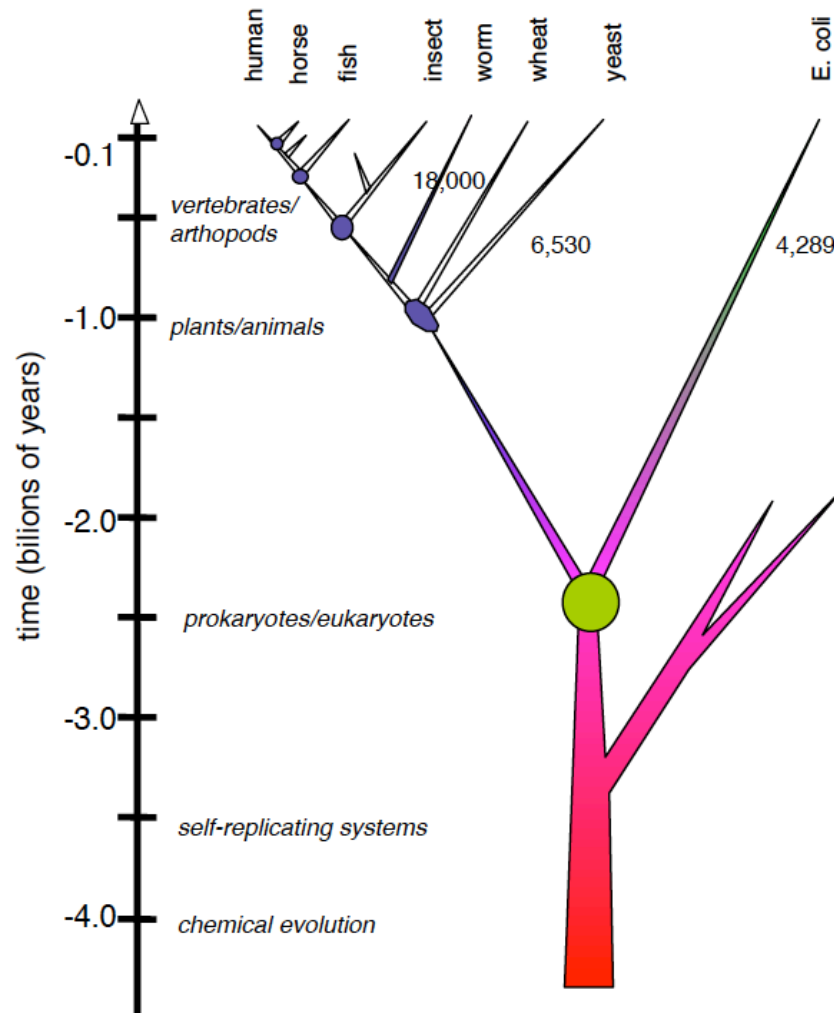
4

- Model selection

5

- Assessing confidence in topology

Homologues share a common ancestor



Retrieve homologous sequences

- Common tool: **BLAST** (Basic Local Alignment Search Tool) used to find homologs
- BLAST finds homologs by locating **short matches** between sequences (aa=3, nt=11)
- Pros: **quick** and easy, relatively accurate
- Question: what **bacterial species** share common ancestry with my isolate of interest?

How to score homologues

- Use **E-values**, not percent identify to infer homology
- E-value = number of hits one can **expect** to see by chance
- The lower the E-value the more **significant** the match (ie. the better!)
- E- value **<0.001** is significant for most searches

E-values

Descriptions

Legend for links to other resources: [U](#) UniGene [E](#) GEO [G](#) Gene [S](#) Structure [M](#) Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NM_003689.2	Homo sapiens aldo-keto reductase family 7, member A2 (aflatoxin ;	2484	2484	100%	0.0	100%	U E G
BK000395.1	TPA: TPA exp: Homo sapiens aflatoxin B1-aldehyde reductase (AKF	2457	2457	98%	0.0	100%	G
BC012171.1	Homo sapiens aldo-keto reductase family 7, member A2 (aflatoxin ;	2439	2439	98%	0.0	100%	U E G
BC007352.2	Homo sapiens aldo-keto reductase family 7, member A2 (aflatoxin ;	2439	2439	98%	0.0	100%	U E G
BC010852.1	Homo sapiens aldo-keto reductase family 7, member A2 (aflatoxin ;	2414	2414	97%	0.0	99%	U E G
AF026947.1	Homo sapiens aflatoxin aldehyde reductase AFAR mRNA, complete	2396	2396	96%	0.0	99%	U E G
Y16675.1	Homo sapiens mRNA for aflatoxin B1-aldehyde reductase	2379	2379	95%	0.0	100%	U G
BC013996.2	Homo sapiens aldo-keto reductase family 7, member A2 (aflatoxin ;	2356	2356	94%	0.0	100%	U E G
CR617181.1	full-length cDNA clone CS0DB008YK02 of Neuroblastoma Cot 10-no	2352	2352	94%	0.0	100%	U G
BC011586.1	Homo sapiens aldo-keto reductase family 7, member A2 (aflatoxin ;	2349	2349	94%	0.0	100%	U E G
BC004111.2	Homo sapiens aldo-keto reductase family 7, member A2 (aflatoxin ;	2349	2349	94%	0.0	100%	U E G
CR597954.1	full-length cDNA clone CS0DD009Y007 of Neuroblastoma Cot 50-nc	2349	2349	94%	0.0	100%	U G
CR606608.1	full-length cDNA clone CS0DE002YD02 of Placenta of Homo sapiens	2324	2324	93%	0.0	100%	U G
CR614593.1	full-length cDNA clone CS0DK008YI20 of HeLa cells Cot 25-normaliz	2302	2302	92%	0.0	100%	U G
CR606766.1	full-length cDNA clone CS0DI068YG11 of Placenta Cot 25-normalize	2286	2286	92%	0.0	100%	U G
CR625016.1	full-length cDNA clone CS0DK008YF01 of HeLa cells Cot 25-normali:	2275	2275	91%	0.0	100%	U G
CR603343.1	full-length cDNA clone CS0DJ011YO15 of T cells (Jurkat cell line) Cc	2266	2266	91%	0.0	100%	U G
CR610843.1	full-length cDNA clone CS0DI041YB06 of Placenta Cot 25-normalize	2248	2248	90%	0.0	100%	U G
XM_001092177.1	PREDICTED: Macaca mulatta aldo-keto reductase family 7, member	2192	2192	98%	0.0	95%	U G

E-values

Two important parameter can influence the E-value...

1) Number of sequences in the database

- Problem: greater chance of finding 'matches' in larger databases (too much noise)
- Fix: use smaller databases, likely to contain your isolate to reduce overestimating the E-value

2) Length of the search query

- Problem: greater chance of finding a match if your sequence is short
- Fix: use shallow scoring matrices to reduce overestimating the E-value for short reads

E-values: quick fix

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/ BLAST/ blastn suite Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

Enter Query Sequence BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Enter accession number(s), gi(s), or FASTA sequence(s) Clear Query subrange From To

Or, upload file No file selected.

Job Title Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database Human genomic + transcript Mouse genomic + transcript Others (nr etc.):
Nucleotide collection (nr/nt)

Organism Optional Enter organism name or id—completions will be suggested Exclude Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude Optional Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query Optional Enter an Entrez query to limit search

Program Selection

Optimize for Highly similar sequences (megablast) More dissimilar sequences (discontiguous megablast) Somewhat similar sequences (blastn) Choose a BLAST algorithm



BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/ BLAST/ blastn suite Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

Enter Query Sequence BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Enter accession number(s), gi(s), or FASTA sequence(s) Clear Query subrange From To

Or, upload file No file selected.

Job Title Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database Human genomic + transcript Mouse genomic + transcript Others (nr etc.):
Nucleotide collection (nr/nt)

Organism Optional Enter organism name or id—completions will be suggested Exclude Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude Optional Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query Optional Enter an Entrez query to limit search

Program Selection

Optimize for Highly similar sequences (megablast) More dissimilar sequences (discontiguous megablast) Somewhat similar sequences (blastn) Choose a BLAST algorithm

Genomic plus Transcript

- Human genomic plus transcript (Human G+T)
- Mouse genomic plus transcript (Mouse G+T)

Other Databases

- Nucleotide collection (nr/nt)
- Reference RNA sequences (refseq_rna)
- Reference genomic sequences (refseq_genomic)
- NCBI Genomes (chromosome)
- Expressed sequence tags (est)
- Genomic survey sequences (gss)
- High throughput genomic sequences (HTGS)
- Patent sequences (pat)
- Protein Data Bank (pdb)
- Human ALU repeat elements (alu_repeats)
- Sequence tagged sites (dbsts)
- Whole-genome shotgun contigs (wgs)
- Transcriptome Shotgun Assembly (TSA)
- 16S ribosomal RNA sequences (Bacteria and Archaea)

Narrow down the size of your search by nominating a specific database

How to build a phylogenetic tree

1

- Collect data i.e. DNA

2

- Retrieve homologous sequences

3

- Multiple sequence alignment

4

- Model selection

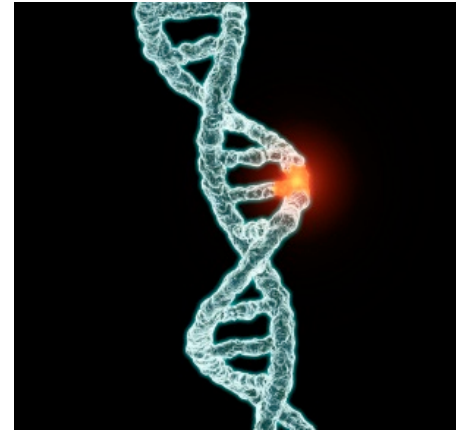
5

- Assessing confidence in topology

Evolution of DNA sequences

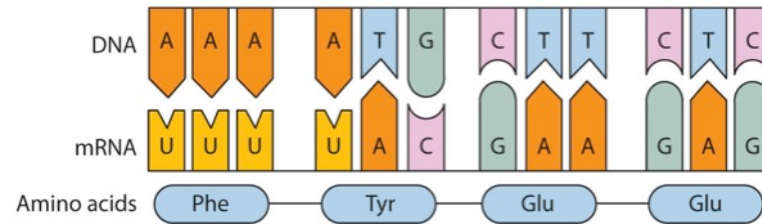
refresher

- Evolution of visible (phenotypic) characters is the result of changes at the **molecular** level
- 3 types of **mutations**:
 - insertions (frameshift)
 - deletions (frameshift)
 - substitutions

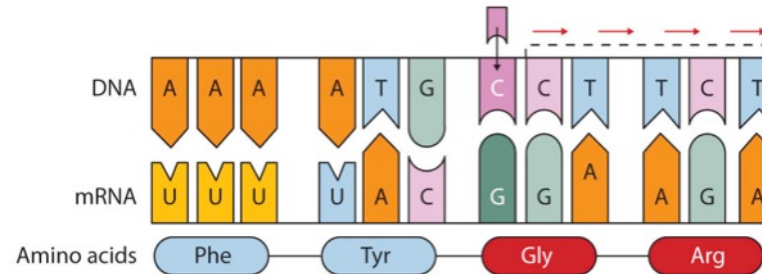


Types of mutations: Frameshift

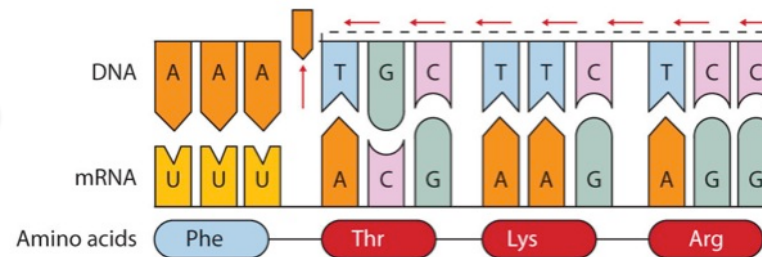
Normal



Insertion: addition of cytosine



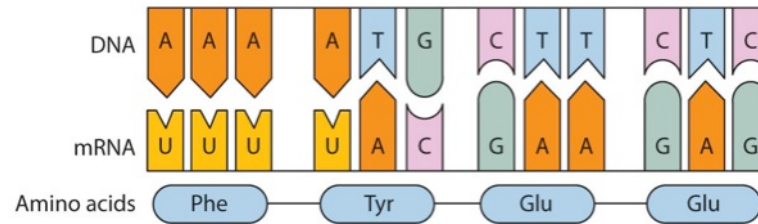
Deletion: removal of adenine



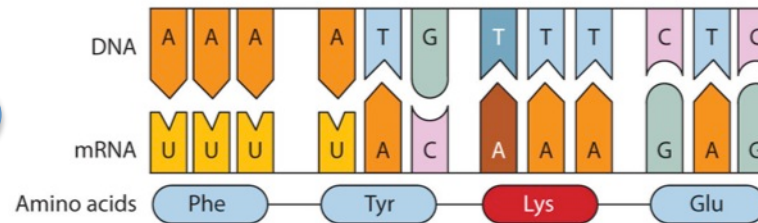
Result: reading frame downstream of mutation is changed → codons are misaligned → translation of non-functional protein → deleterious effect on organism

Types of mutations: Substitution

Normal



Substitution:
cytosine - thymine



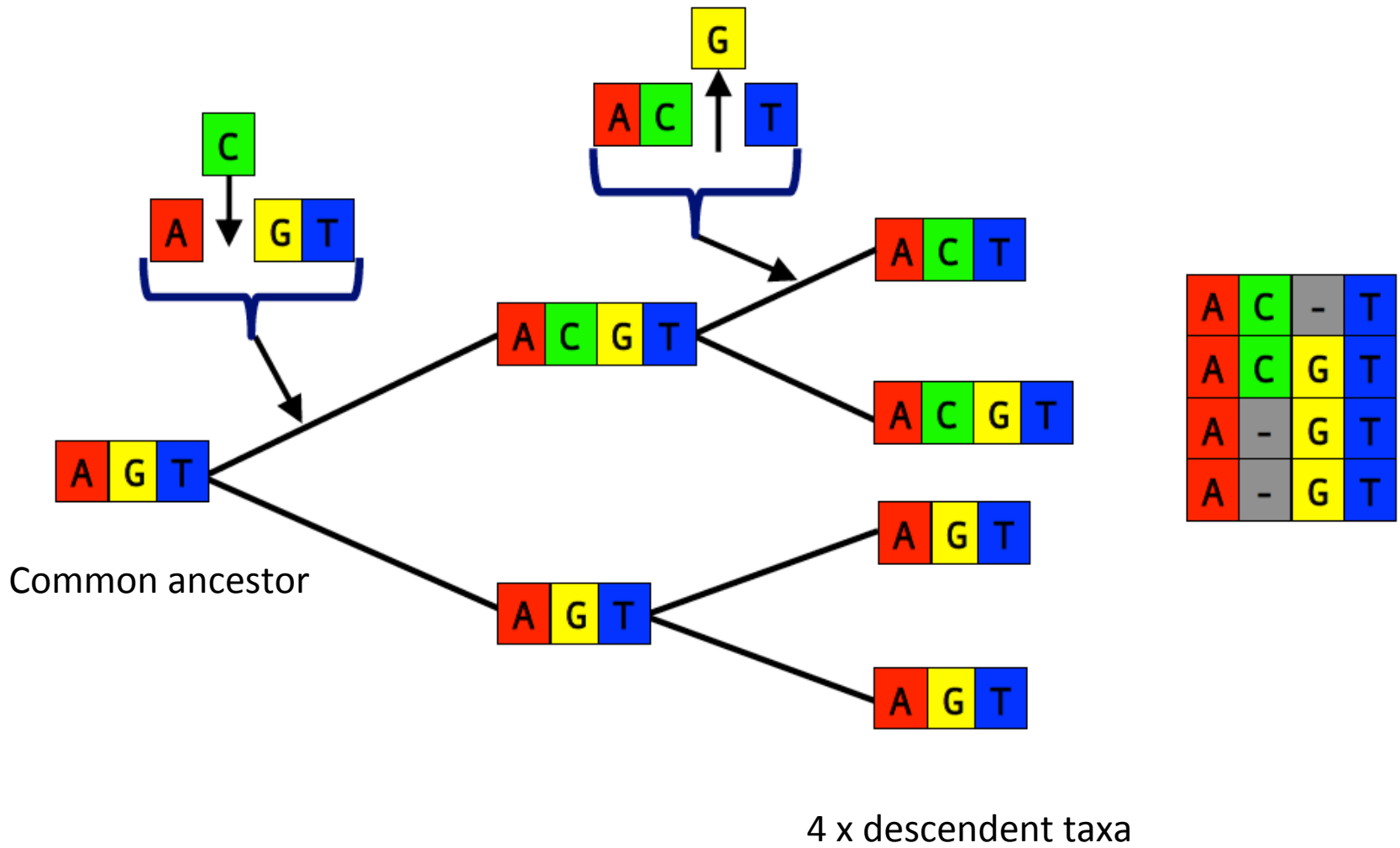
Result: reading frame preserved → different protein is translated
→ if neutral or beneficial impact on phenotype then selection occurs

Multiple sequence alignment

- Insertions & deletions ('**indels**') obscure sites that are homologous (= traits descended from common ancestor)
- Goal of MSA is to introduce **gaps** so that nucleotides in same column are homologous

Scarites	C	T	T	A	G	A	T	C	G	T	A	C	C	A	A	-	-	-	A	A	T	A	T	T	A	C
Carenum	C	T	T	A	G	A	T	C	G	T	A	C	C	A	C	A	-	T	A	C	-	T	T	T	A	C
Pasimachus	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	T	A	T	A	A	G	T	T	T	A	C
Pheropsophus	C	T	T	A	G	A	T	C	G	T	T	C	C	A	C	-	-	-	A	C	A	T	A	T	A	C
Brachinus armiger	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	T	A	T	A	T	T	C
Brachinus hirsutus	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	T	A	T	A	T	A	C
Aptinus	C	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	C	A	A	T	T	A	C
Pseudomorpha	C	T	T	A	G	A	T	C	G	T	A	C	C	-	-	-	-	-	A	C	A	A	A	T	A	C

MSA in action



MSA in action: great apes

- Finding homology by **scoring** the matrix:
 - Reward matches (+ scores)
 - Penalize substitutions (- scores)
 - Penalize gaps (-- scores)
- Goal is to find an alignment that **maximizes** the total score



<i>Pongo</i>	V	D	E	V	G	G	E	L	G	R	L	-	F	V	V	P	T	Q
<i>Gorilla</i>	V	-	E	V	A	G	D	L	G	R	L	L	I	V	Y	P	S	R
Score	4	-8	5	5	0	6	2	4	6	5	4	-8	0	4	-1	7	4	1

How to build a phylogenetic tree

1

- Collect data i.e. DNA

2

- Retrieve homologous sequences

3

- Multiple sequence alignment

4

- Model selection

5

- Assessing confidence in topology

Model selection

Model selection

- Scoring a matrix can't tell you which traits are derived and which are ancestral
- Need trees to infer evolutionary relationships
- Choose the **simplest** or **most likely** tree corresponding to the matrix



Simplest vs most likely

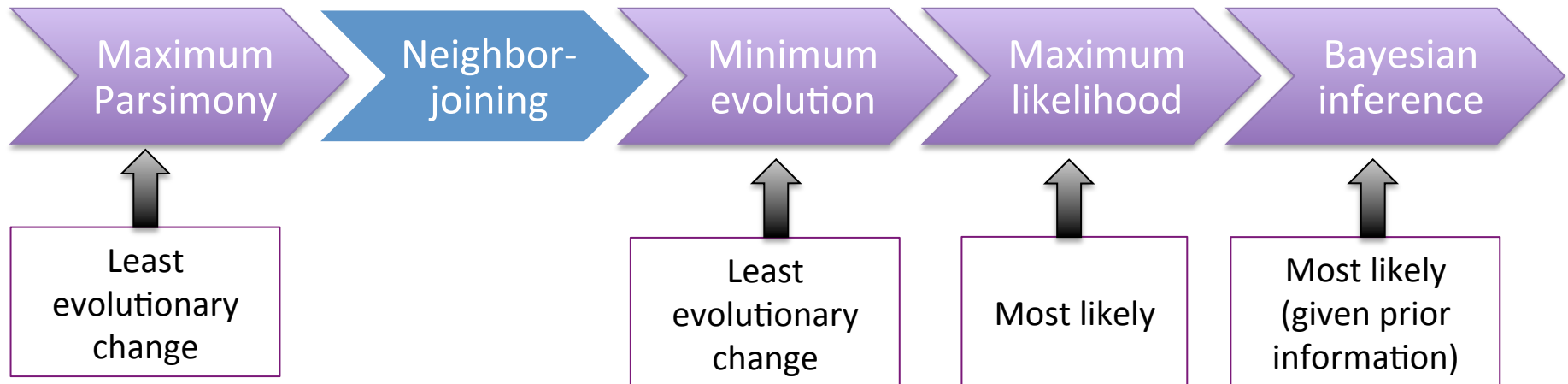
- We need a **metric** to decide which trees are better and which trees are worse
- **Optimality criterion** = a metric of quality (i.e. tree length, parsimony or likelihood) used to assess the **optimal tree**

Optimality criteria

Which methods use an optimality criteria to decide on best tree?



Optimality criteria



Neighbor-joining

Assessing confidence

Assessing confidence

- Trees obtained by phylogenetics are subject to **error** like all other scientific hypotheses
- A tree will be generated regardless of whether there is a phylogenetic signal
- Need to **quantify** how strongly data supports each of the relationships in the tree
- What is the extent to which characters within a matrix **contradict** each other?

Bootstrapping

- Typically tackled with a statistical test called **bootstrapping**
- Assesses chances of recovering a particular clade again if we randomly re-sample our data
- Data matrix is sampled with replacement to produce **pseudo-replicate** datasets
- Measures which parts of the tree are weakly supported with a low bootstrap %

Bootstrap cut-offs

- Exact interpretation of bootstrap % is elusive
- Higher is better but what is a reasonable cut-off? 70%?
- Warning: bootstrapping predicts whether the same result would occur if more data were collected not whether the result is correct

