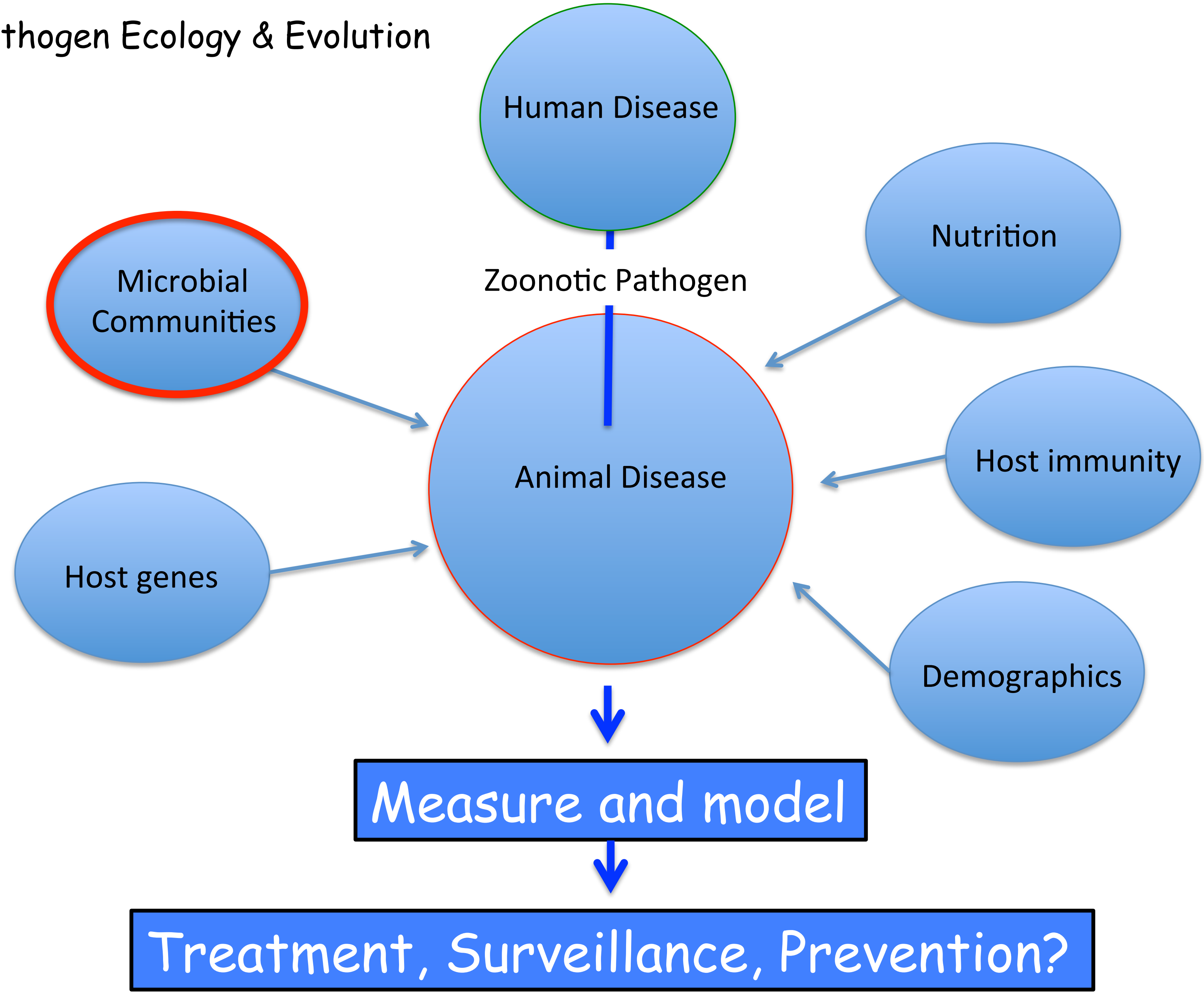
20.109
Laboratory Fundamentals in Biological Engineering

Module 1
Nucleic Acid Engineering
Lecture 5

Pathogen Ecology & Evolution



Human Disease

Zoonotic Pathogen

Microbial Communities

Nutrition

Animal Disease

Host immunity

Host genes

Demographics

Measure and model

Treatment, Surveillance, Prevention?

Today

Microbiome 1.0

Phylogenetics 1.0

Microbes in our World

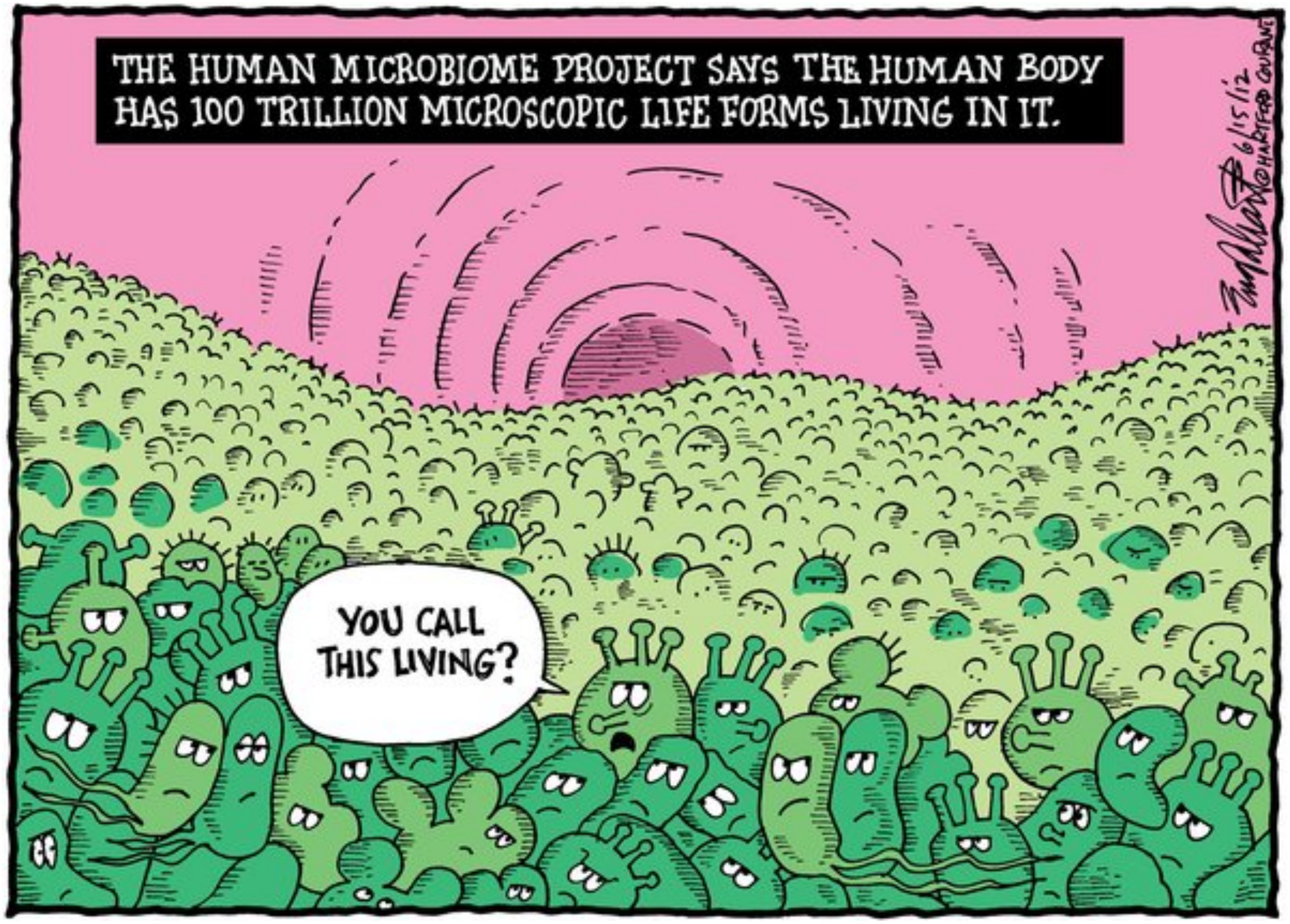
- Fermentation (e.g. – beer)
- Industrial products (e.g. – medicinals, cosmetics, etc…)
- Environmental communities (e.g. – human gut, ocean, soil, etc…)
- Nitrogen fixation, nutrient cycles in ecosystems.

The microbial environment

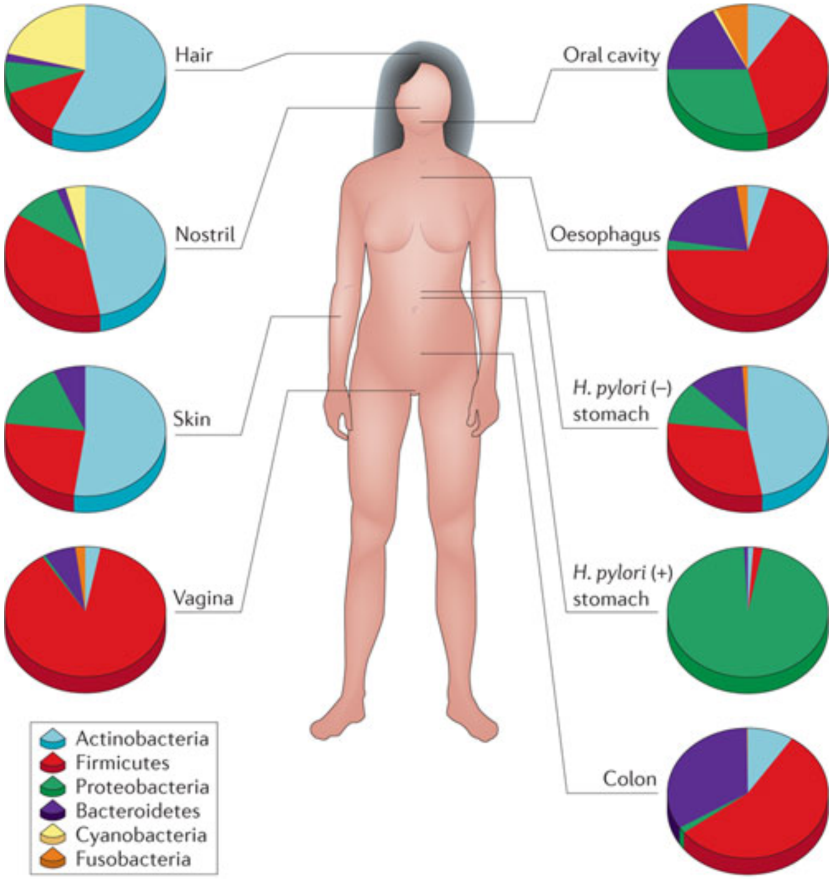
[Scientific American interactive microbiome](#)

Scitechdaily.com

The Human Microbiome

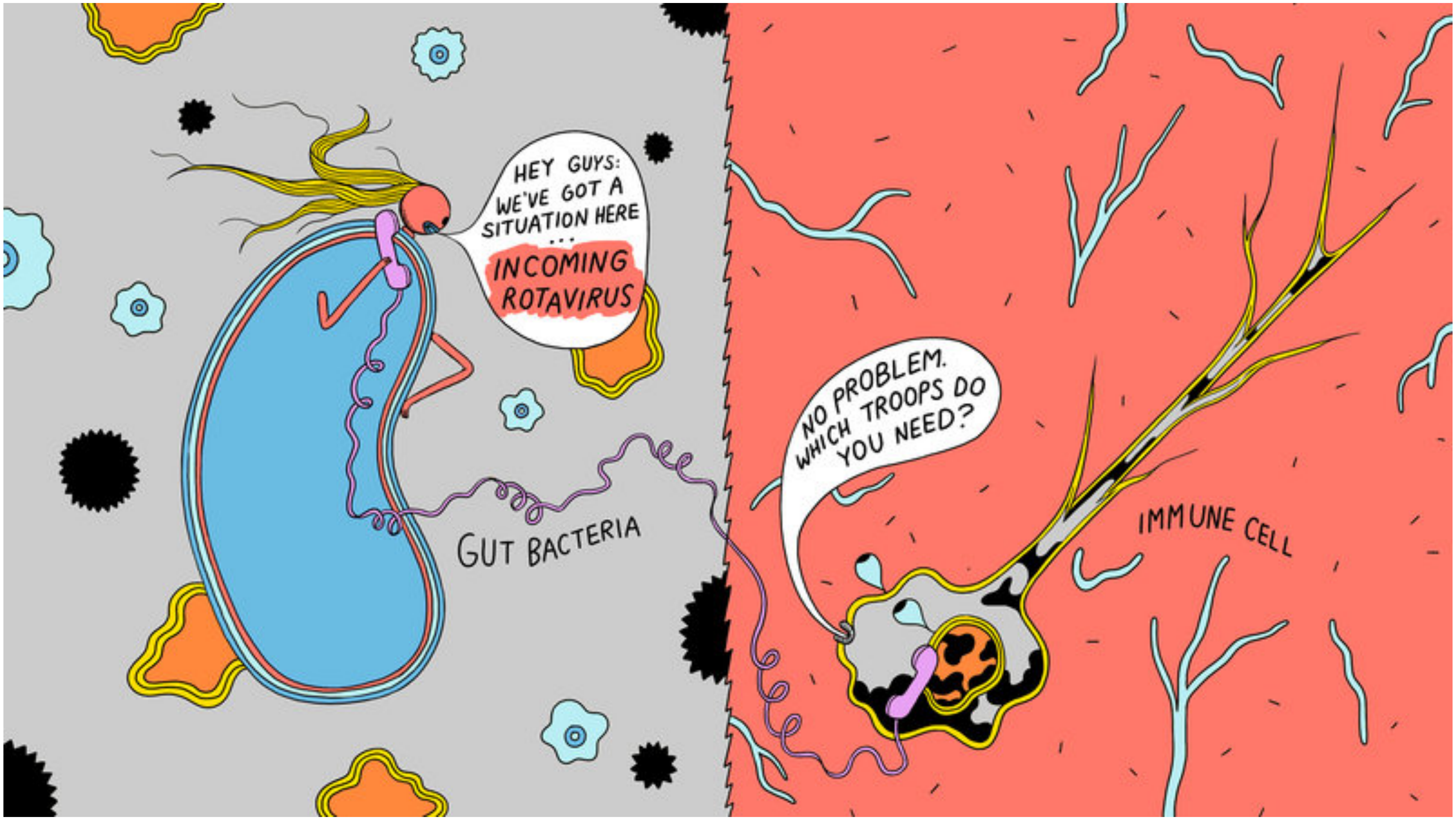


A world of environments



Nature Reviews | Genetics

Virus-fighting bacteria



Microbial community profiling

- What criteria make a good molecular target?

16S rRNA

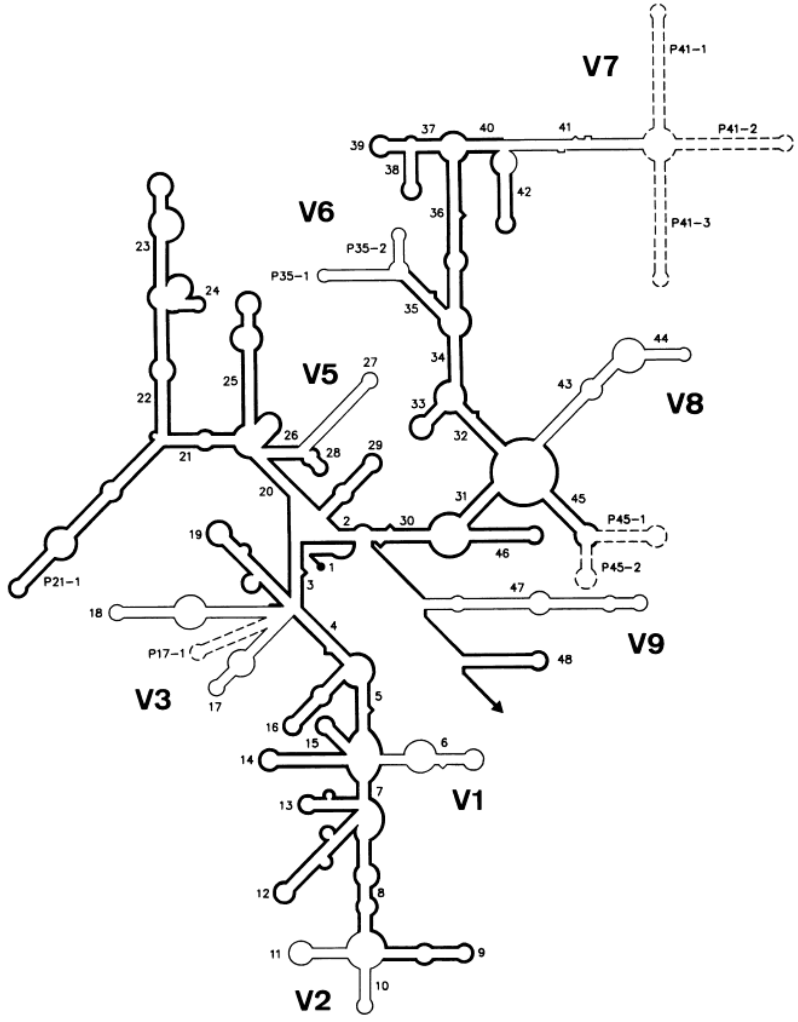
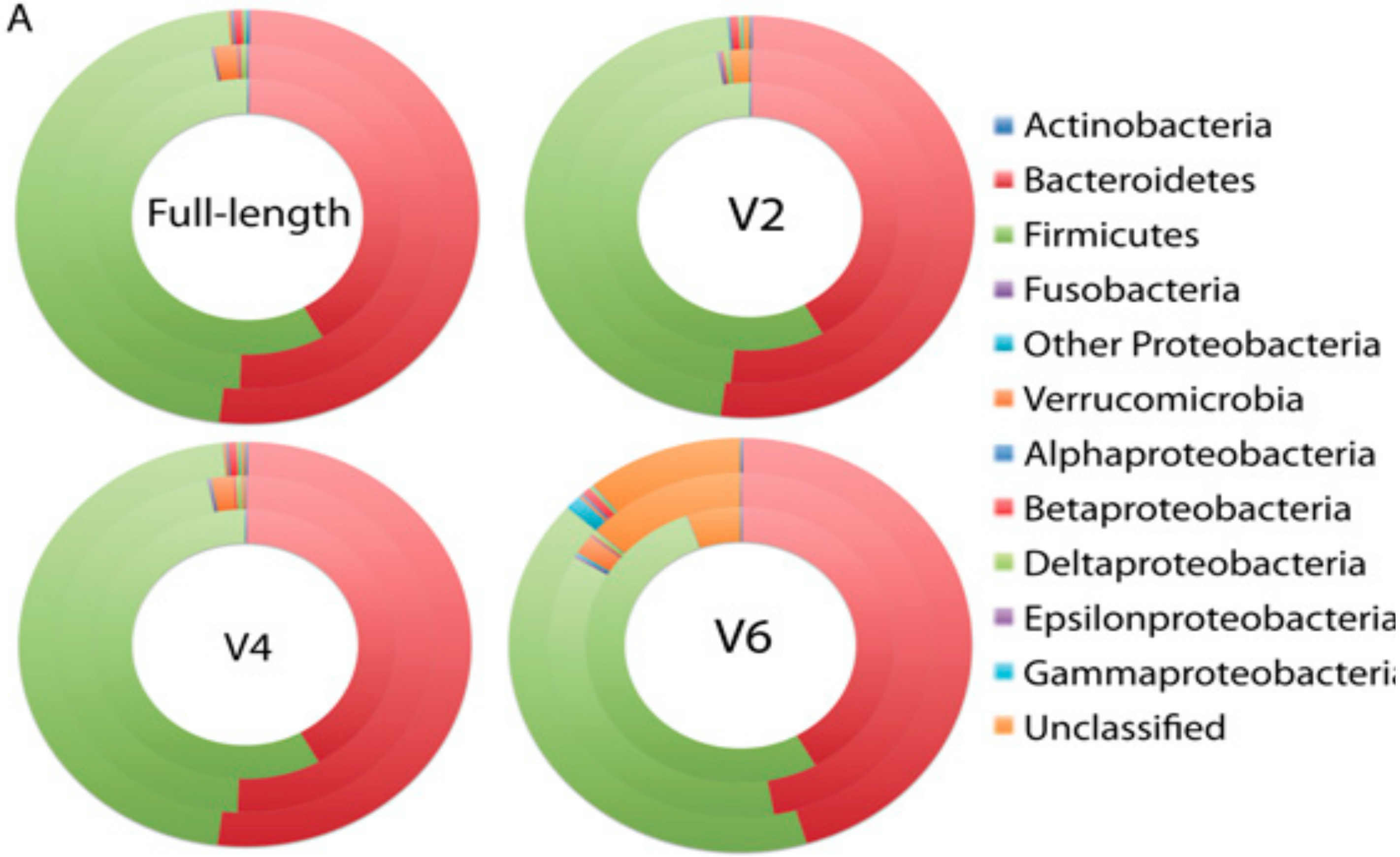
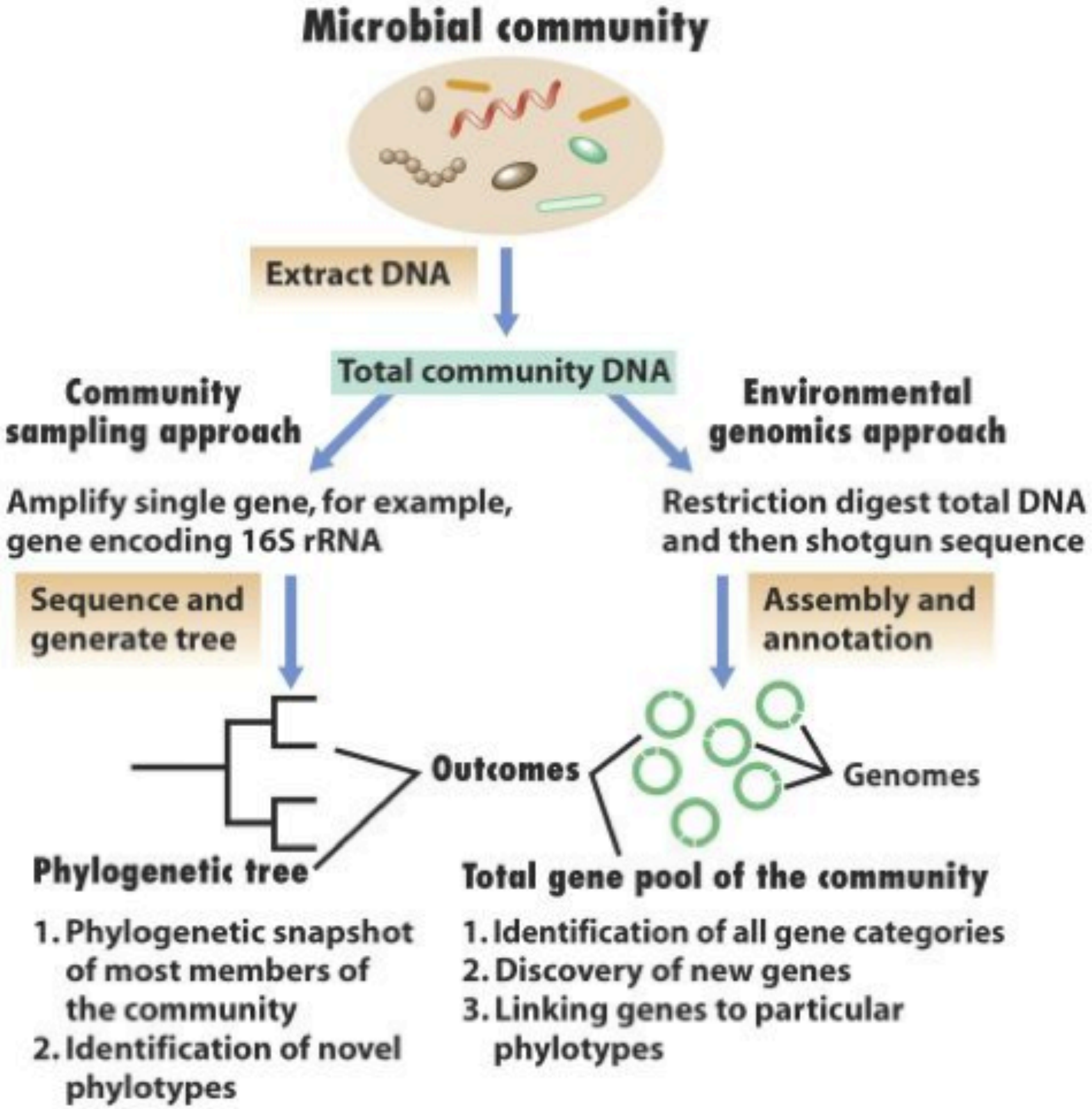


Fig. 1. Secondary structure model for prokaryotic srRNAs.
The 5′-terminus is symbolized by a filled circle and the 3′-terminus by an arrowhead. Helices are numbered in the order of occurrence from 5′- to 3′-terminus. Helices bearing a single number are common to the prokaryotic and eukaryotic (Fig. 2) models. A composite number preceded by P points to a prokaryote-specific helix. Relatively conserved areas are drawn in bold lines, areas of sequence – and length variability in thin lines. Eight variable areas, numbered V1 to V9, are distinguished, V4 being absent in prokaryotic srRNAs. Helices drawn in broken lines are present in a small number of known structures only. Archaebacterial sequences follow the prokaryotic pattern except for helix 35, which is unbranched as in eukaryotes.

Primer bias is a problem



Hamady and Knight, 2009



Microbial community

Extract DNA

Total community DNA

Community sampling approach

Amplify single gene, for example, gene encoding 16S rRNA

Sequence and generate tree

Phylogenetic tree

Outcomes

1. Phylogenetic snapshot of most members of the community
2. Identification of novel phylotypes

Environmental genomics approach

Restriction digest total DNA and then shotgun sequence

Assembly and annotation

Genomes

Total gene pool of the community

1. Identification of all gene categories
2. Discovery of new genes
3. Linking genes to particular phylotypes

Figure 18-15 Brock Biology of Microorganisms 11/e

Community profiling

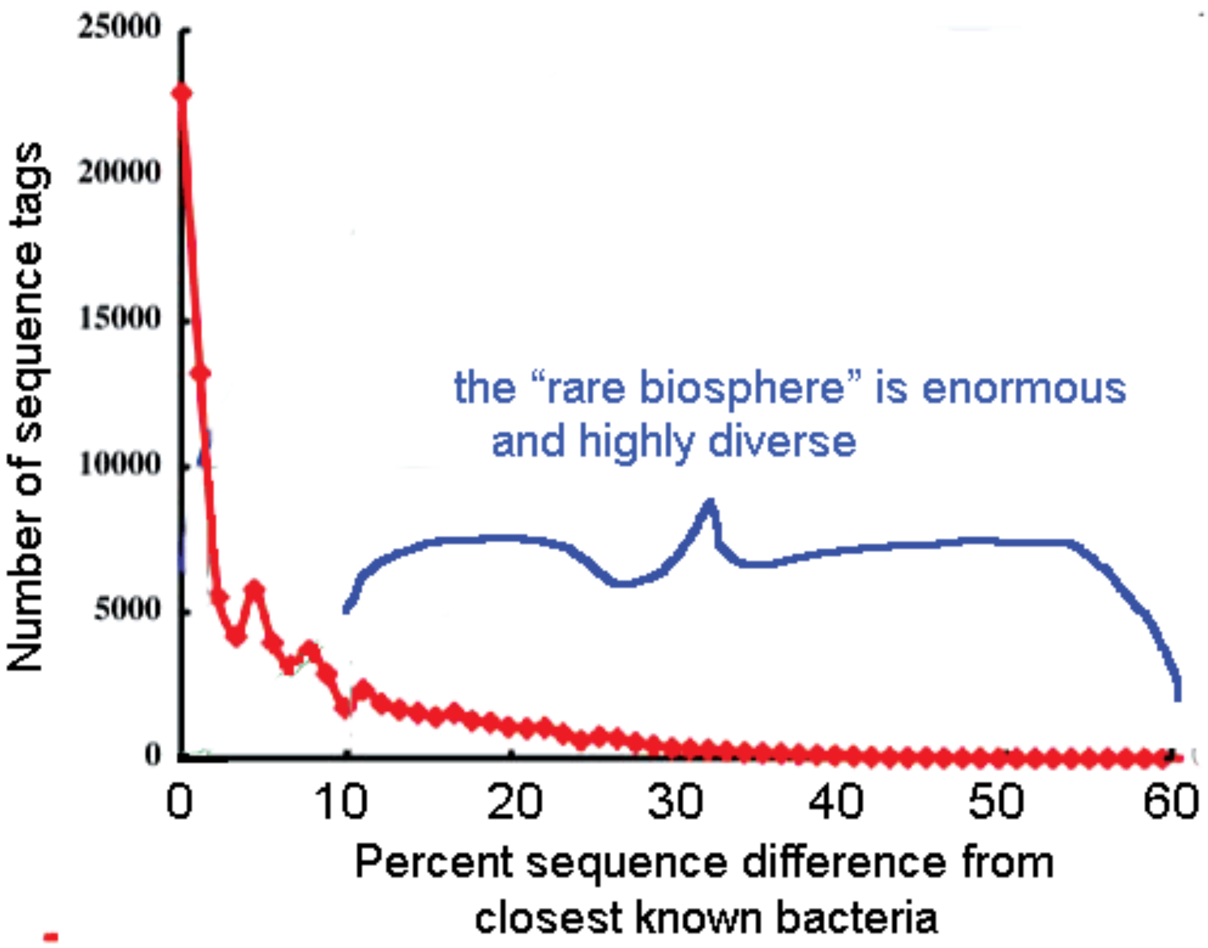
- # unique reads vs # total reads

- # sequences vs taxonomic units (OTUs)

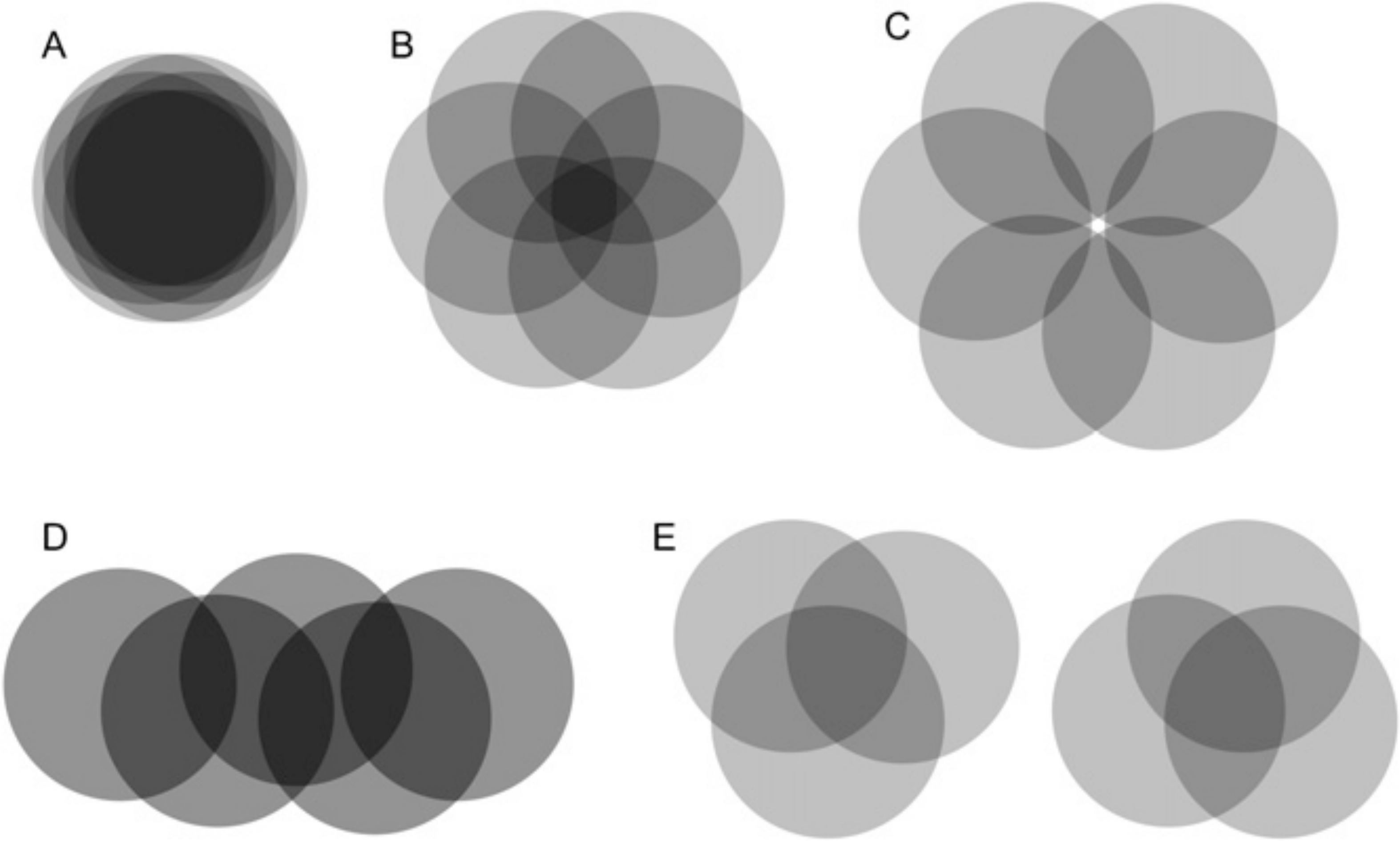
1) Does the rare biosphere matter?

2) Is there a core microbiome?

Rare biosphere

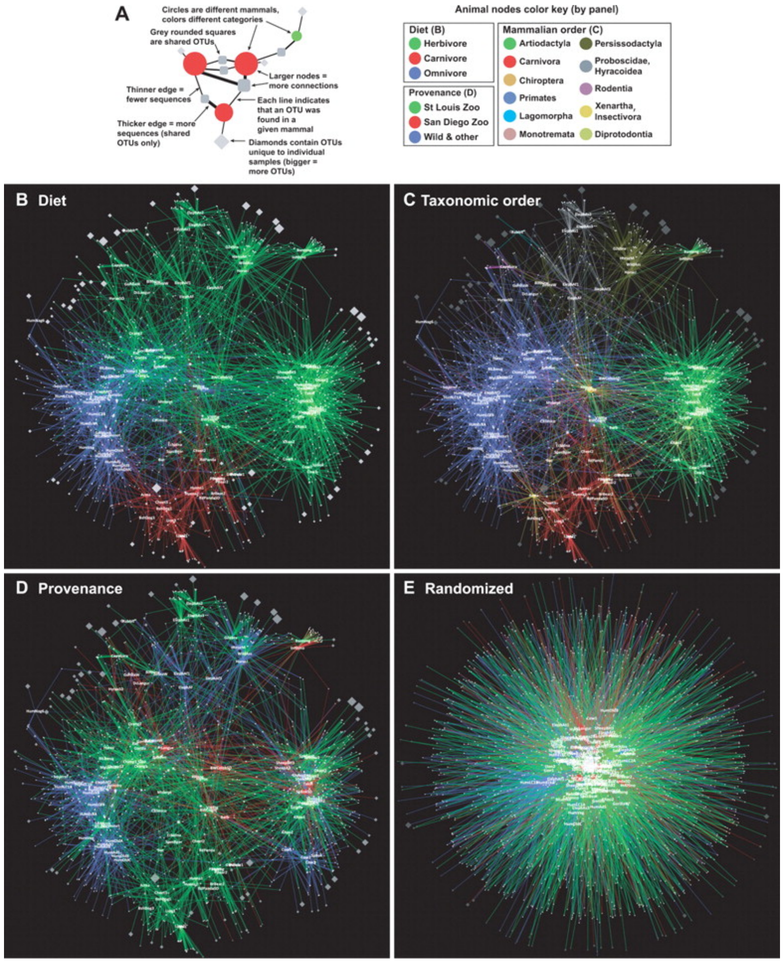


Models of a core microbiome



Hamady and Knight, 2009

Network-based analyses of fecal bacterial communities in 60 mammalian species



R E Ley et al. Science 2008;320:1647-1651

Back to the core questions

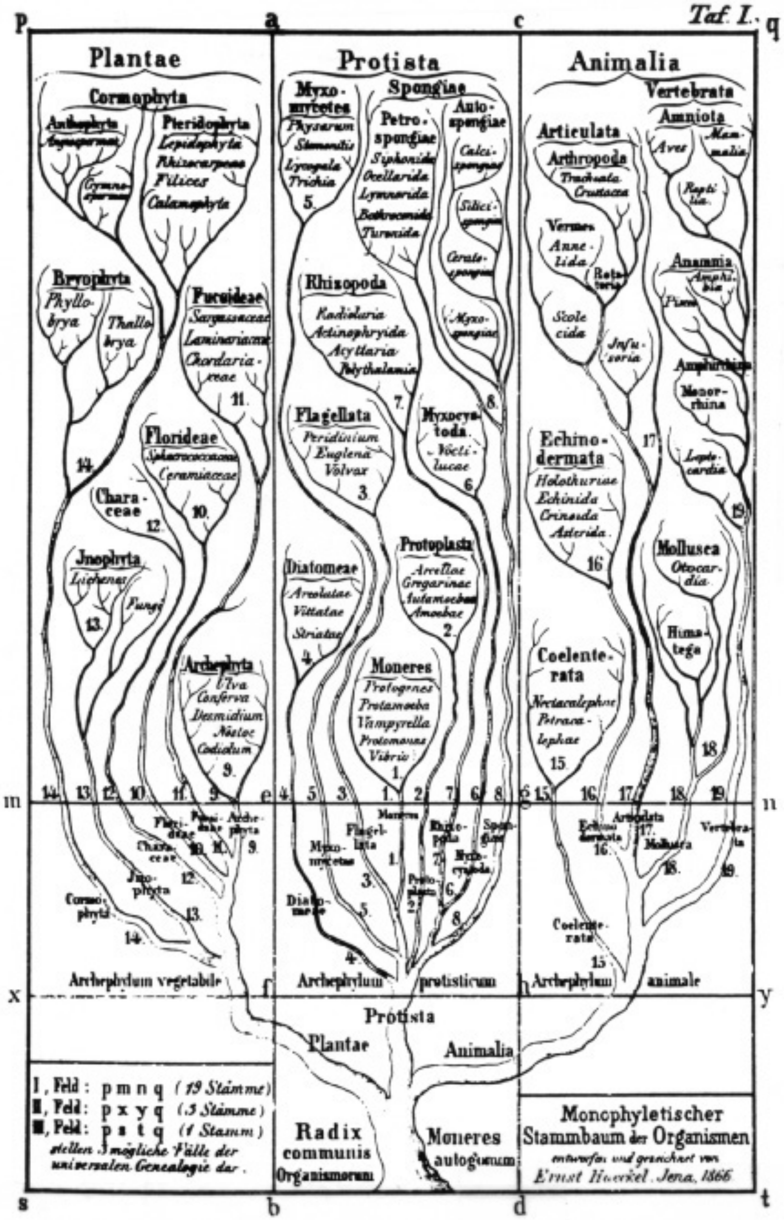
- What do we know?
- How does it go awry?
- How might we possibly manipulate it?

Phylogenetics 1

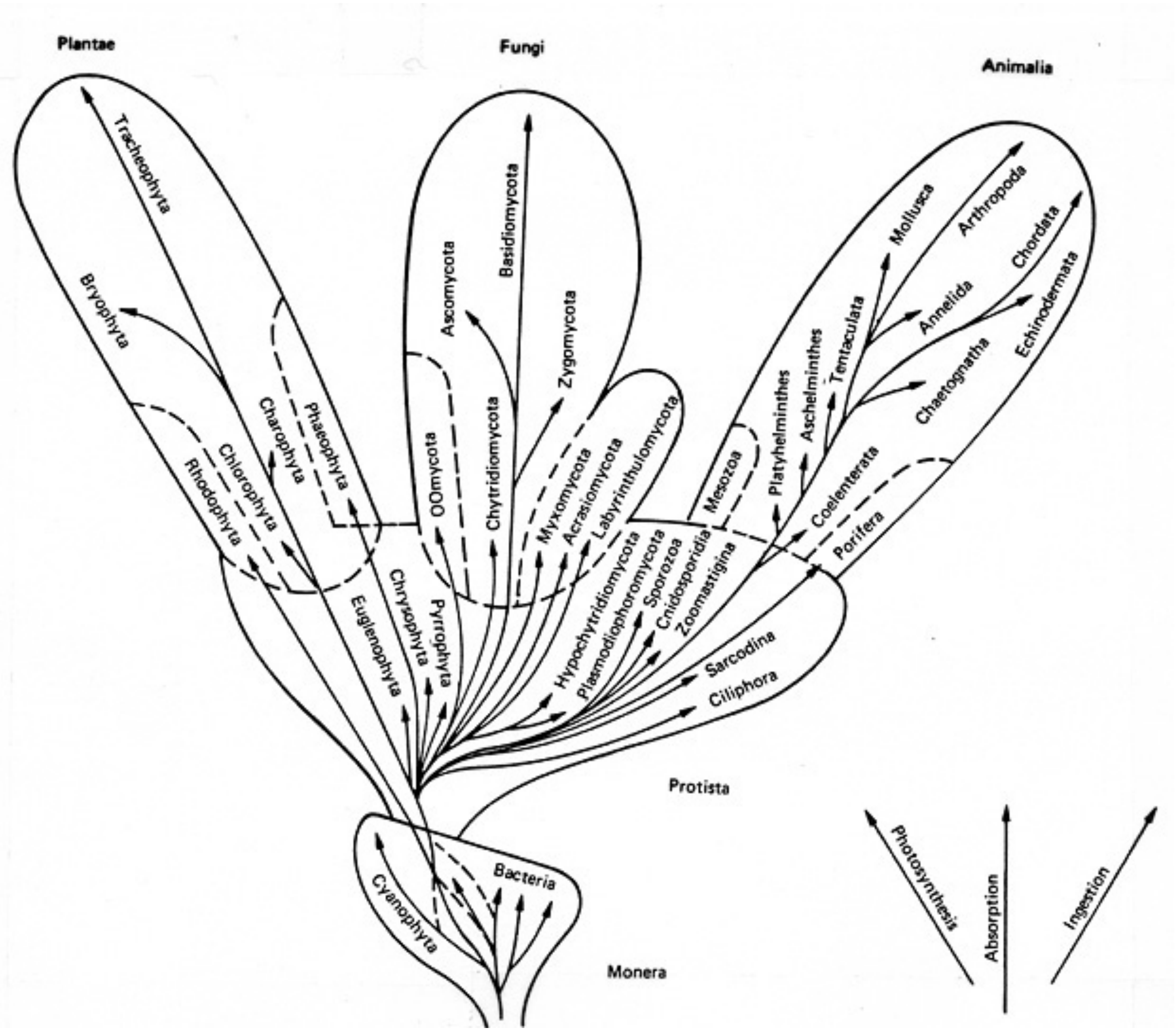
Phylogenetic reconstruction

Produce a phylogenetic tree –

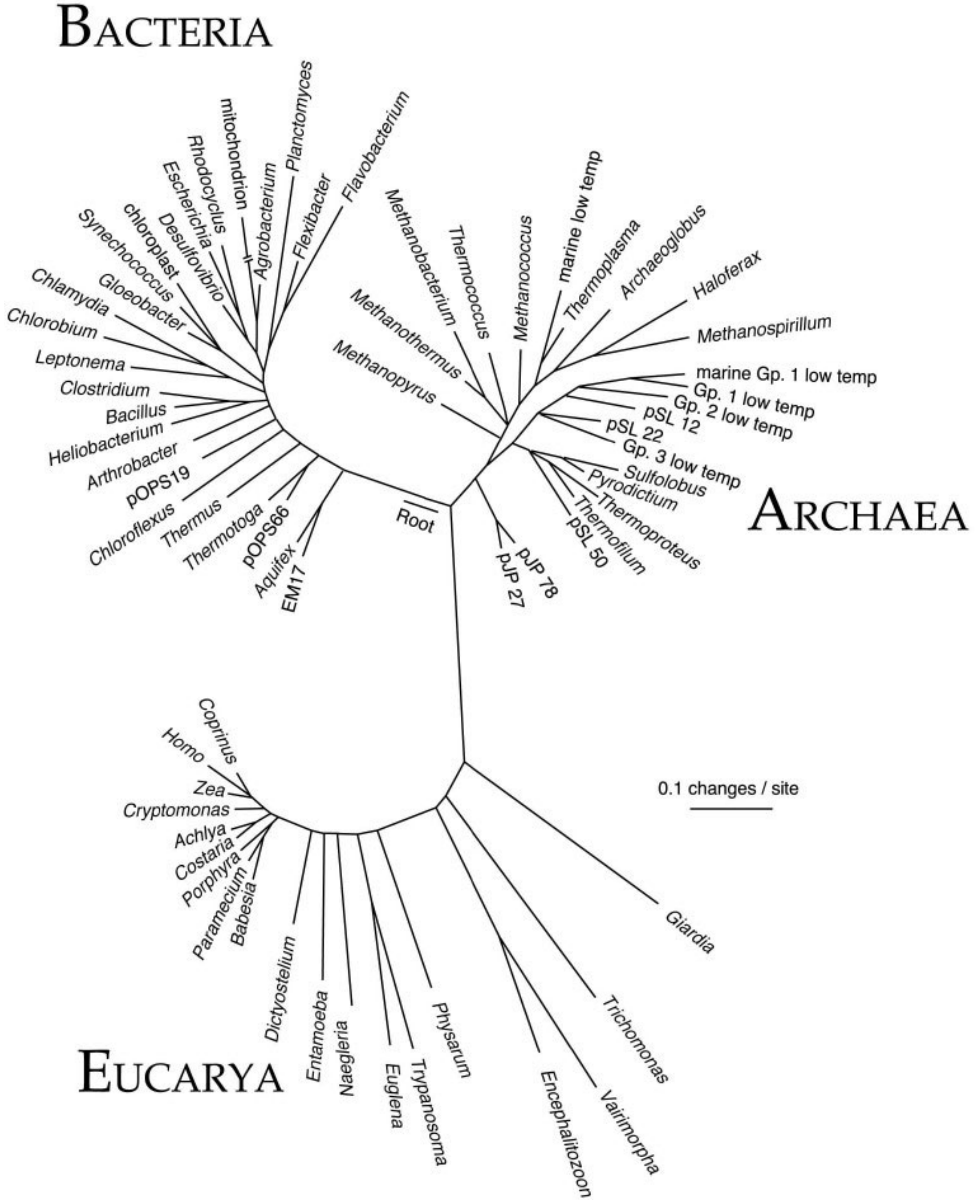
Describing likely descent from a common ancestral sequence of a set of aligned contemporary sequence.



Haeckel (1866), a Swiss naturalist, was the first to create a natural kingdom for the microbes, which had been discovered nearly two centuries before by Antony van Leeuwenhoek

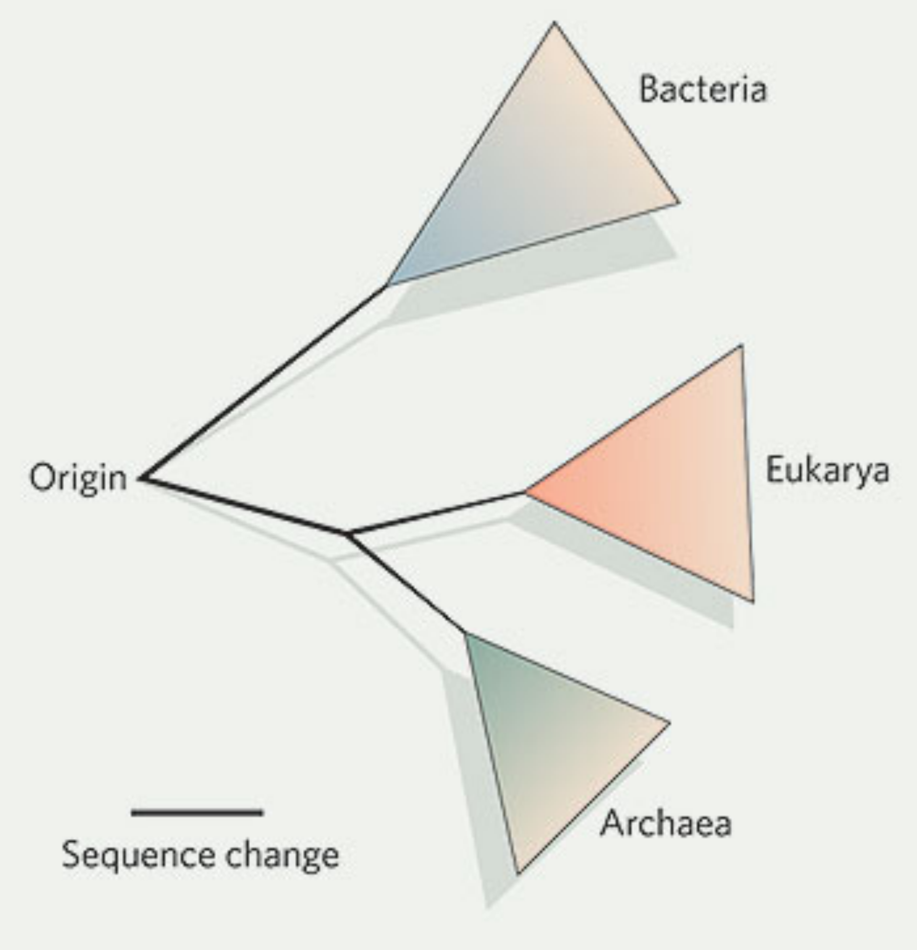


Whitaker, 1967



Pace, Science 276 1997

Revised tree of life

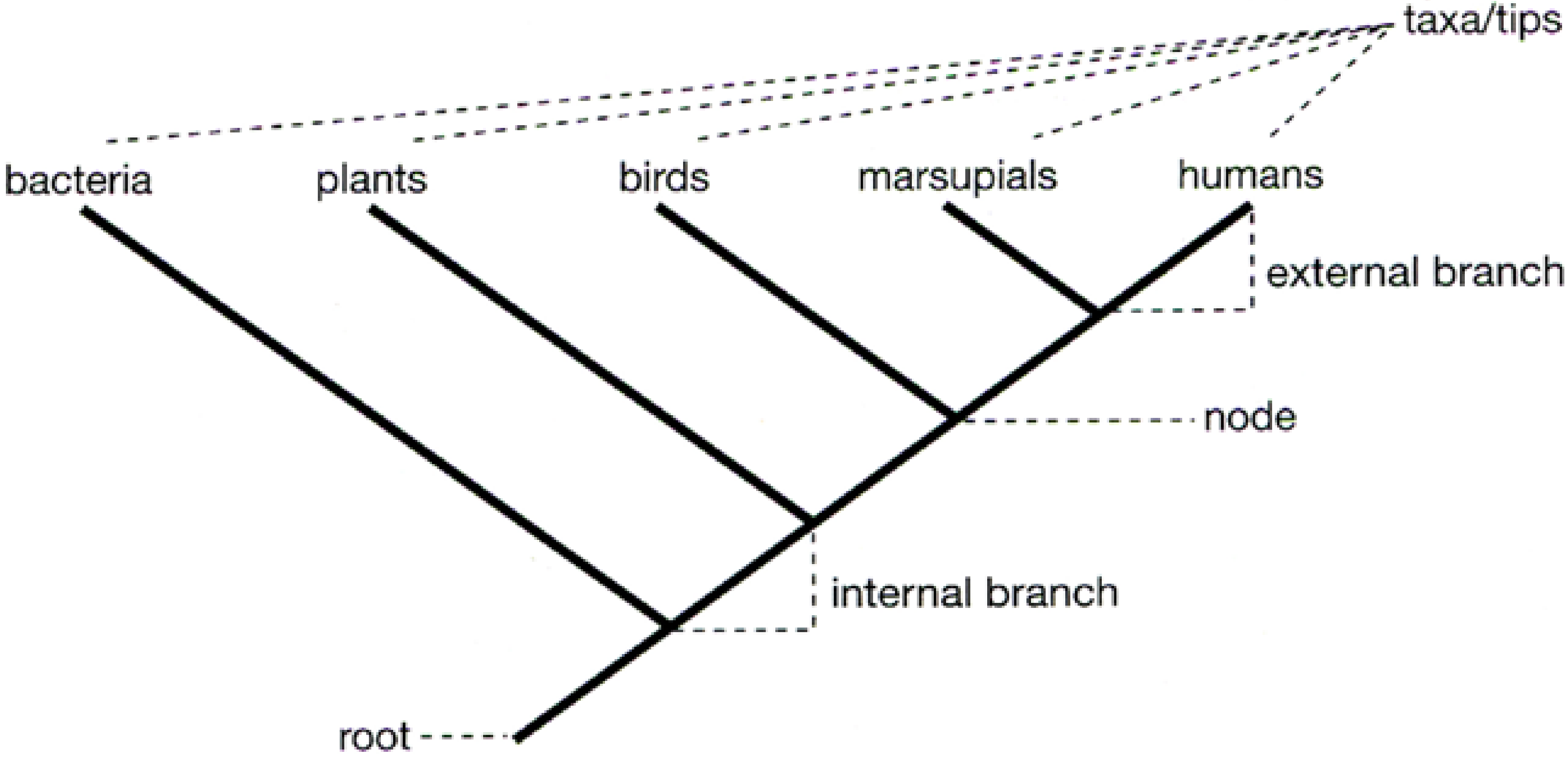


Pace Nature 2006

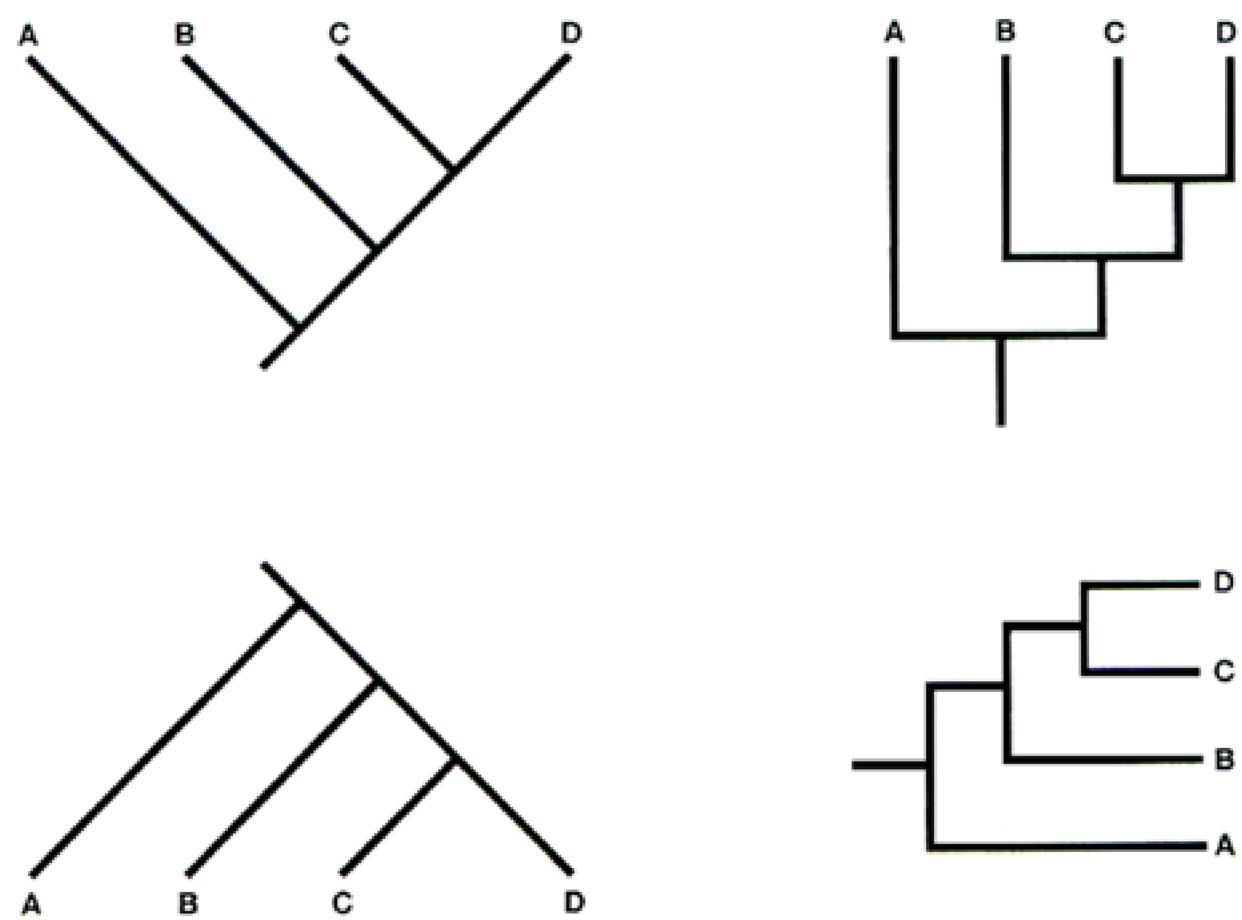
Phylogenetic trees…

- are graphs with nodes & edges
- organisms, etc… are connected by the passage of genetic information along branches of the tree
- models evolution as a bifurcating process

Some important terms used to describe phylogenetic trees

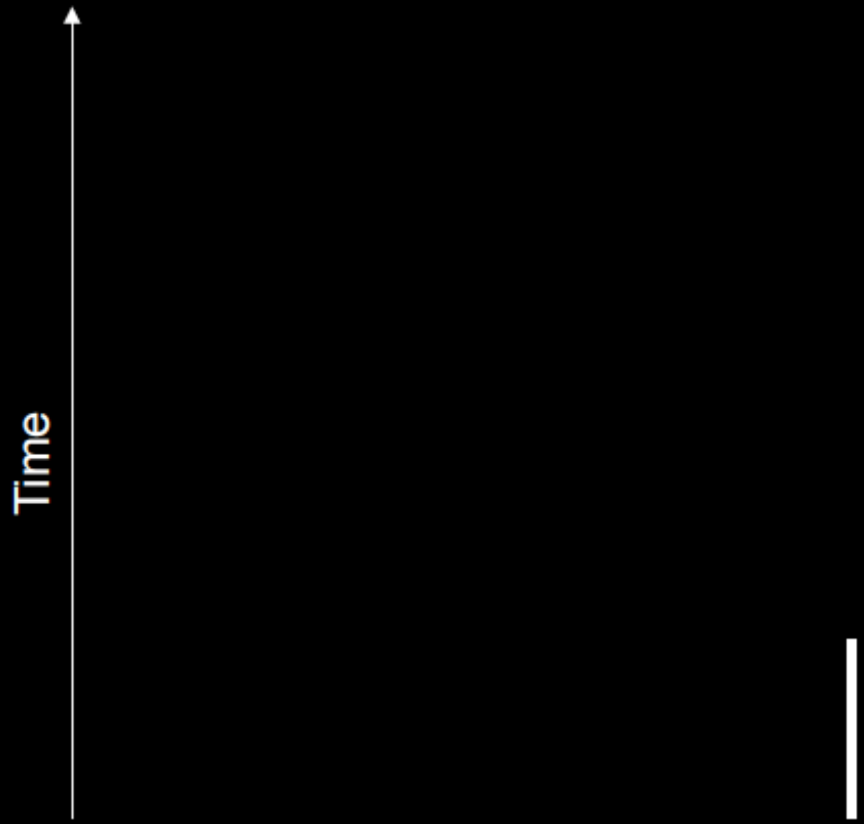


Four trees, all depict same evolutionary history



Phylogeny

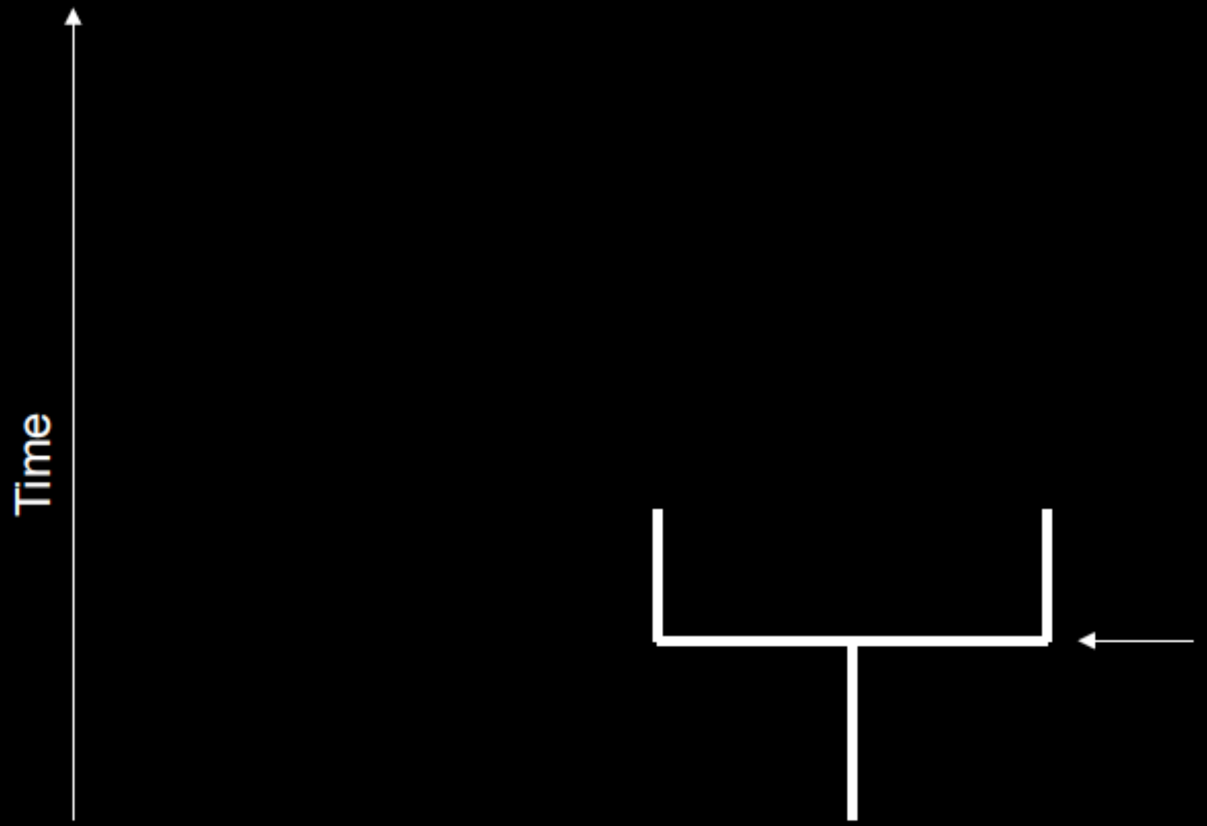
Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.



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

Phylogeny

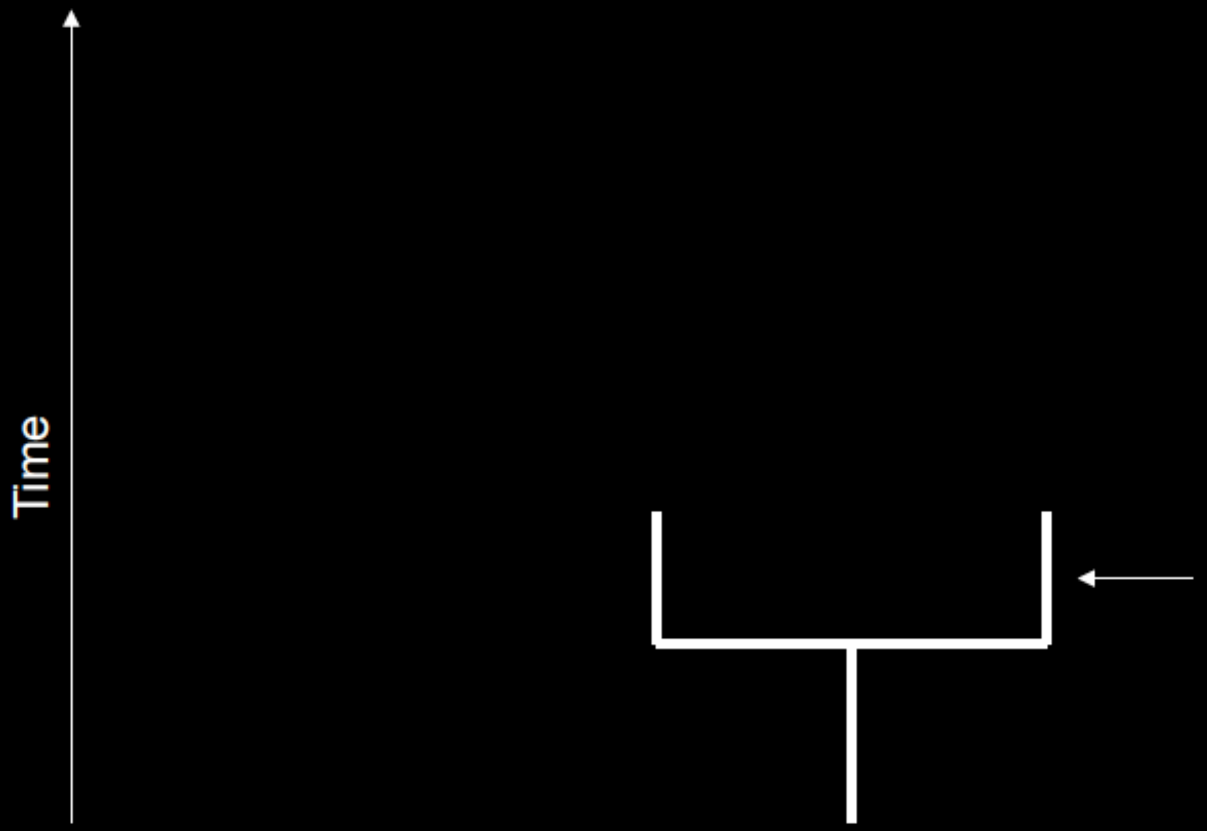
Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.



One lineage
splits into two

Phylogeny

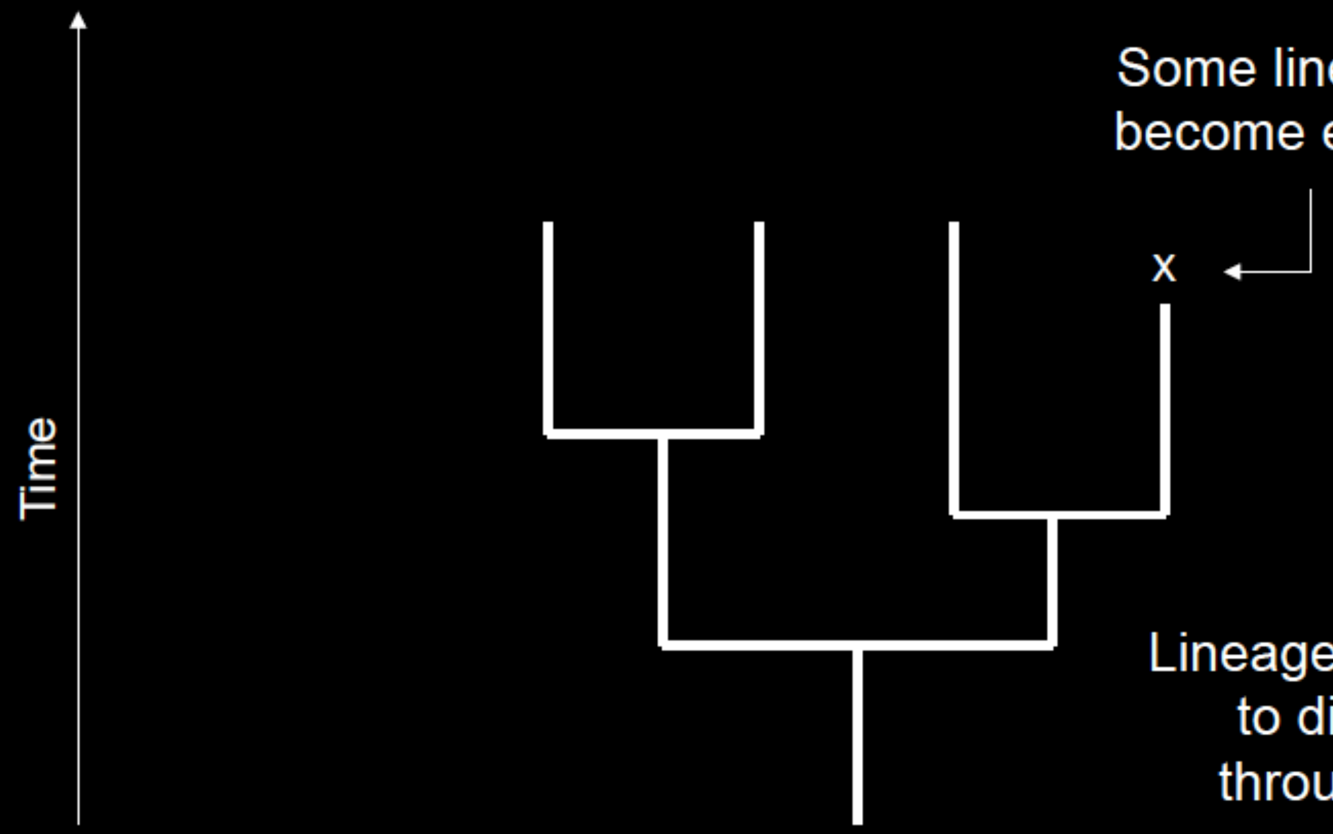
Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.



There are now two HIV lineages

Phylogeny

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

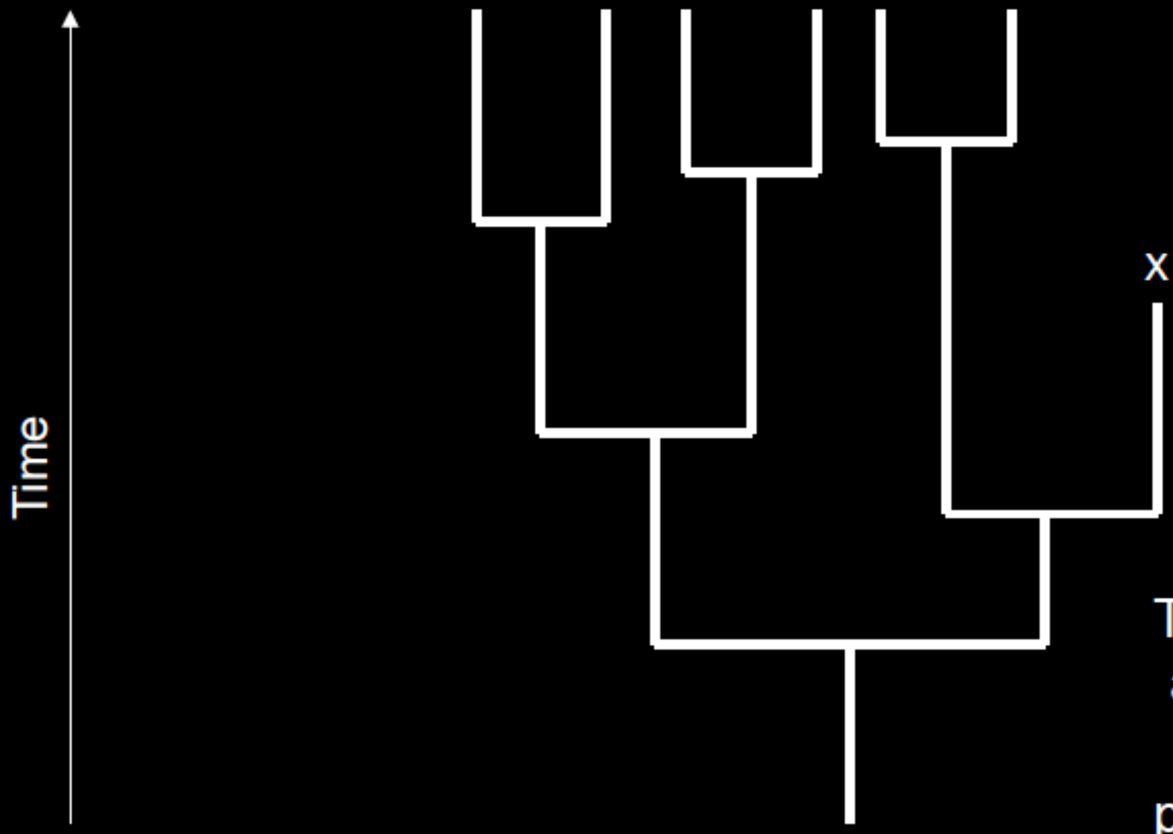


Some lineages become extinct

Lineages continue to diversify through time

Phylogeny

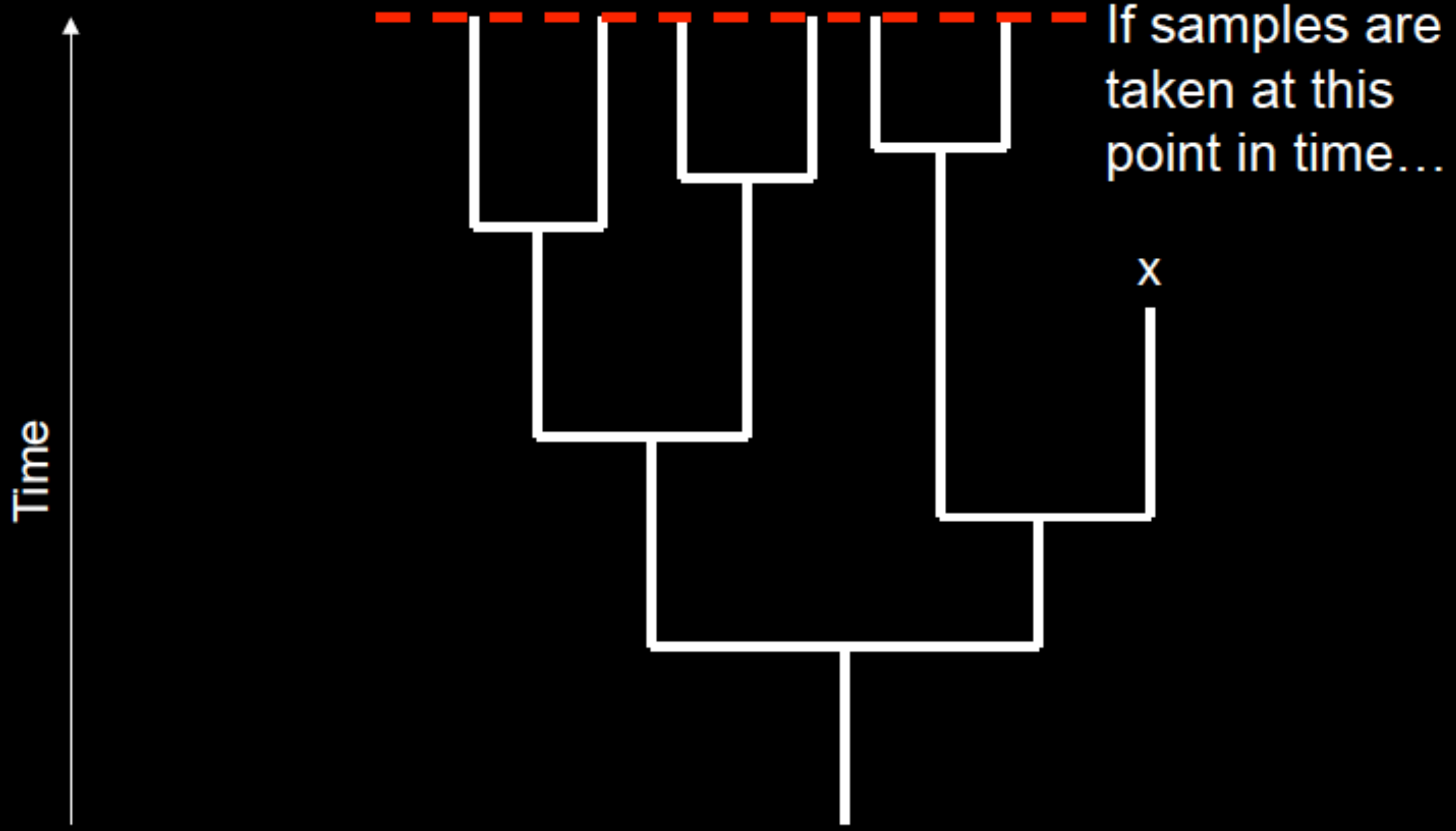
Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.



The relationships among lineages produce a phylogenetic tree

Phylogeny

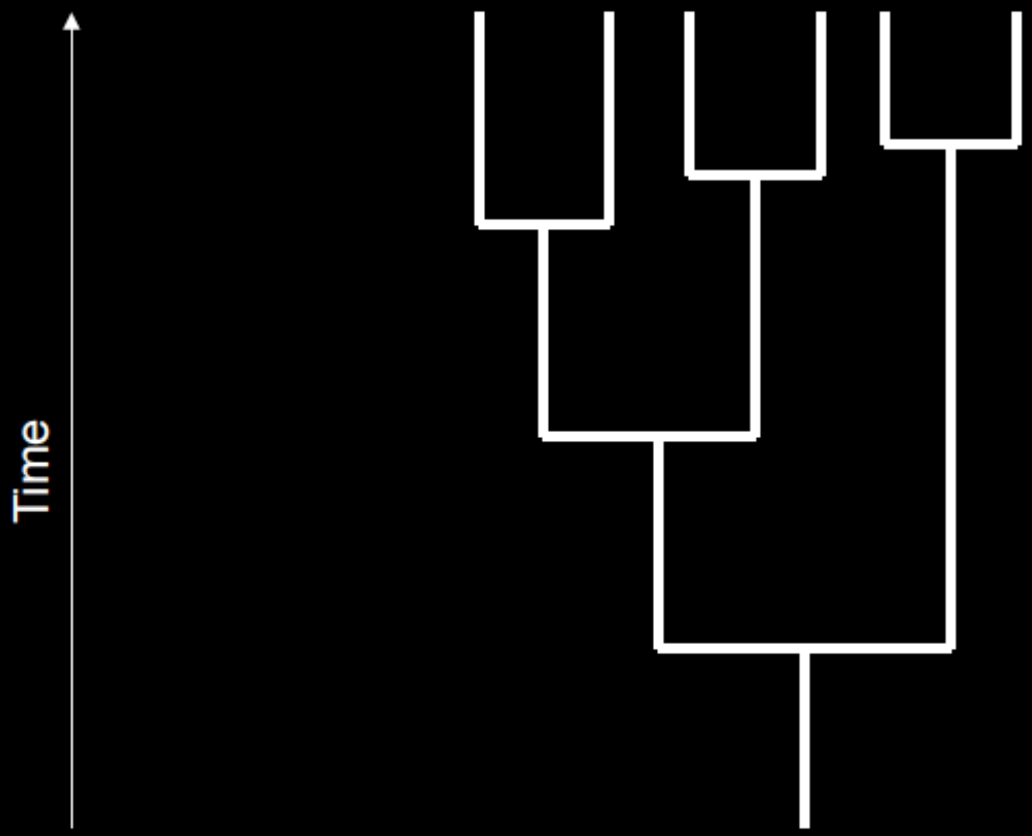
Evolutionary relationships among lineages,
such as genes, individuals, populations, species, etc.



If samples are taken at this point in time…

Phylogeny

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



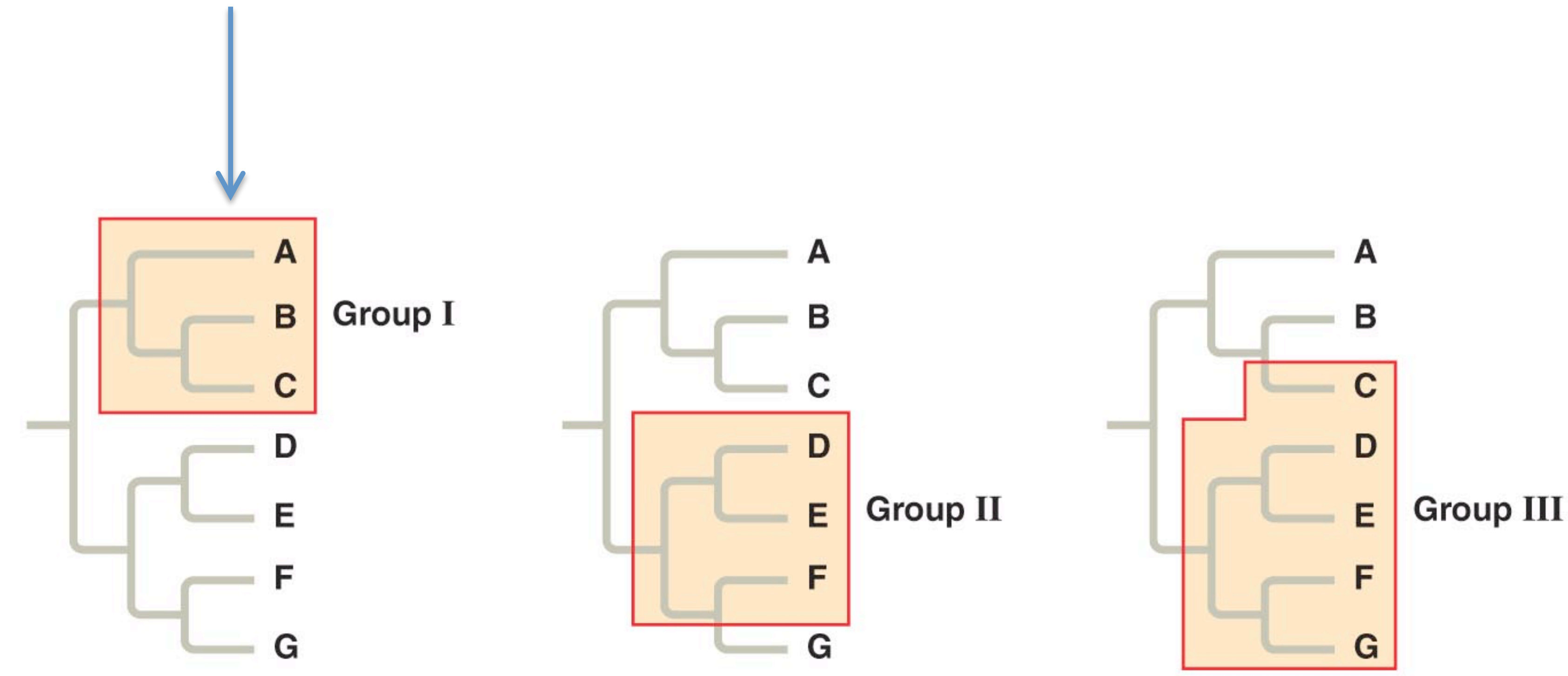
…then the reconstructed phylogenetic tree will look like this

Phylogenetic trees are useful for inferring evolutionary relationships…

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

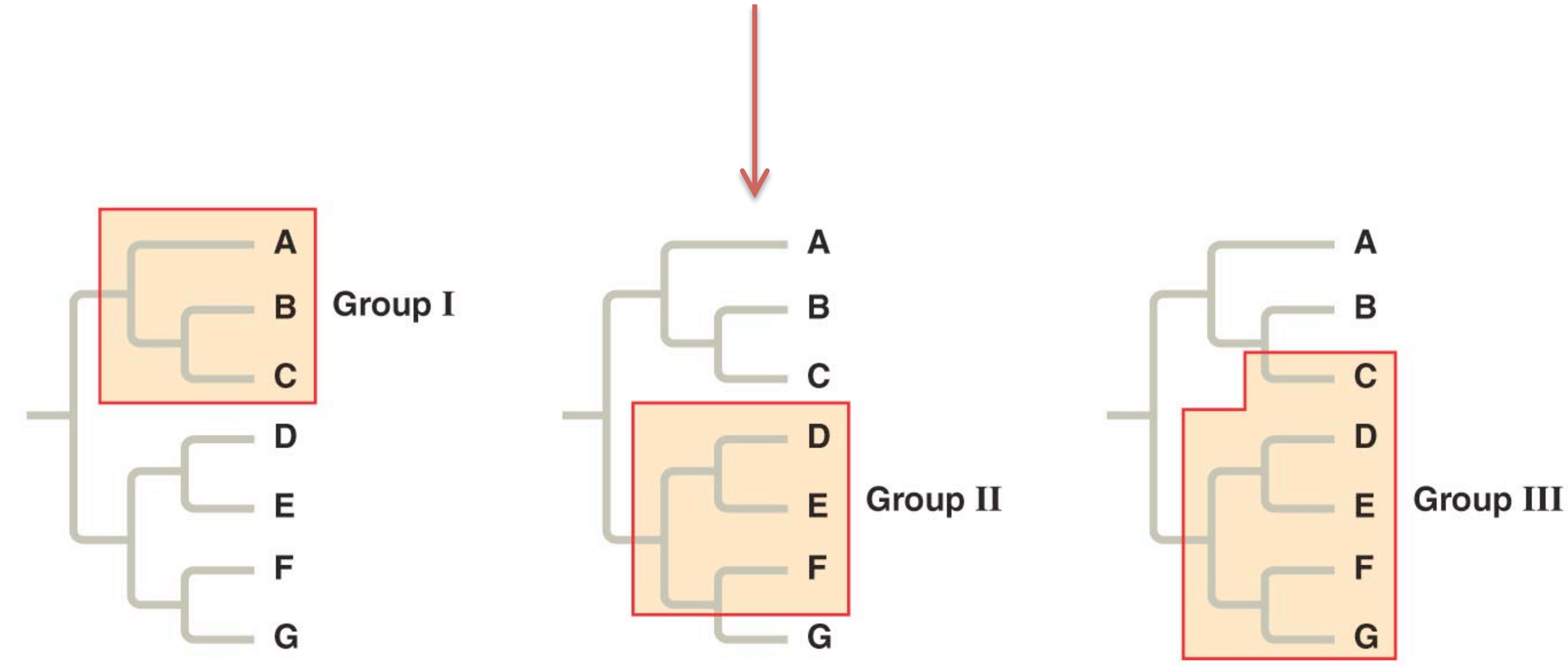
More tree terminology

Monophyletic



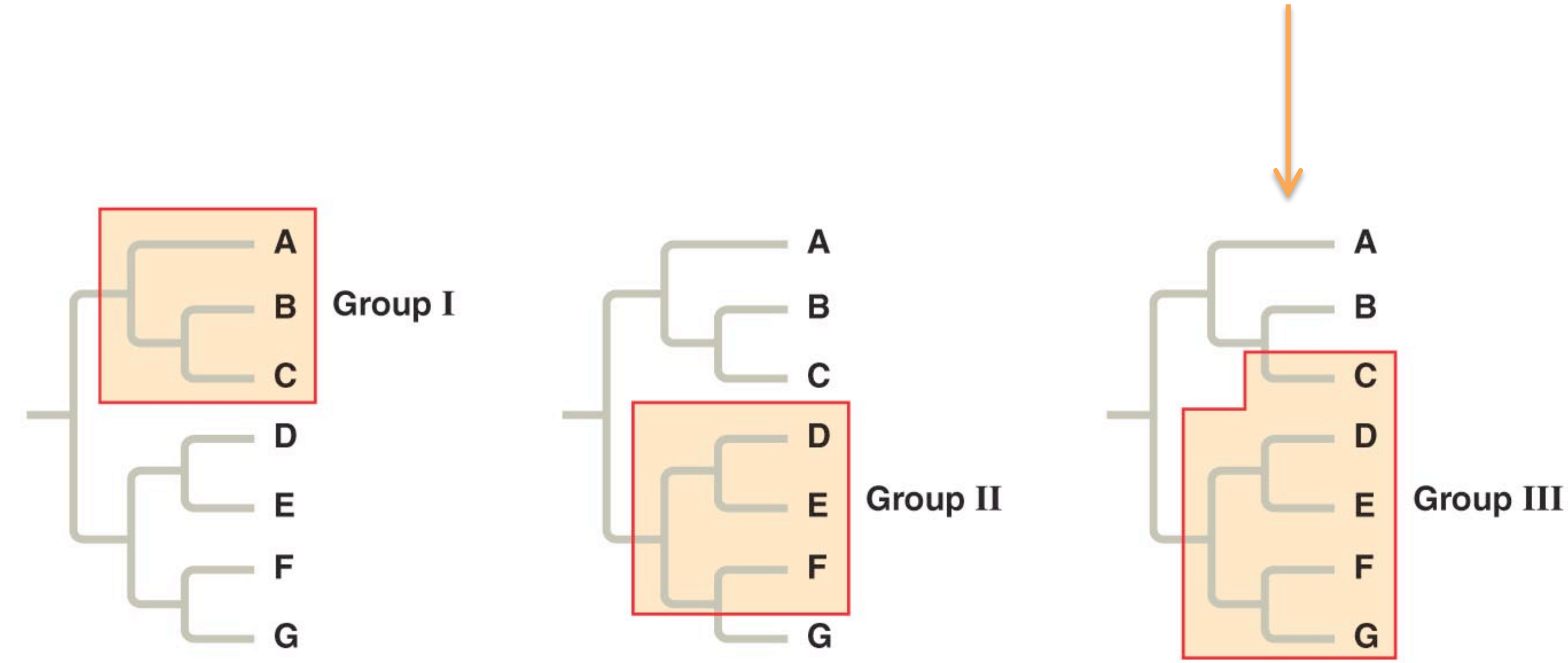
(a) Monophyletic group (clade) (b) Paraphyletic group (c) Polyphyletic group

Paraphyletic



(a) Monophyletic group (clade) (b) Paraphyletic group (c) Polyphyletic group

Polyphyletic

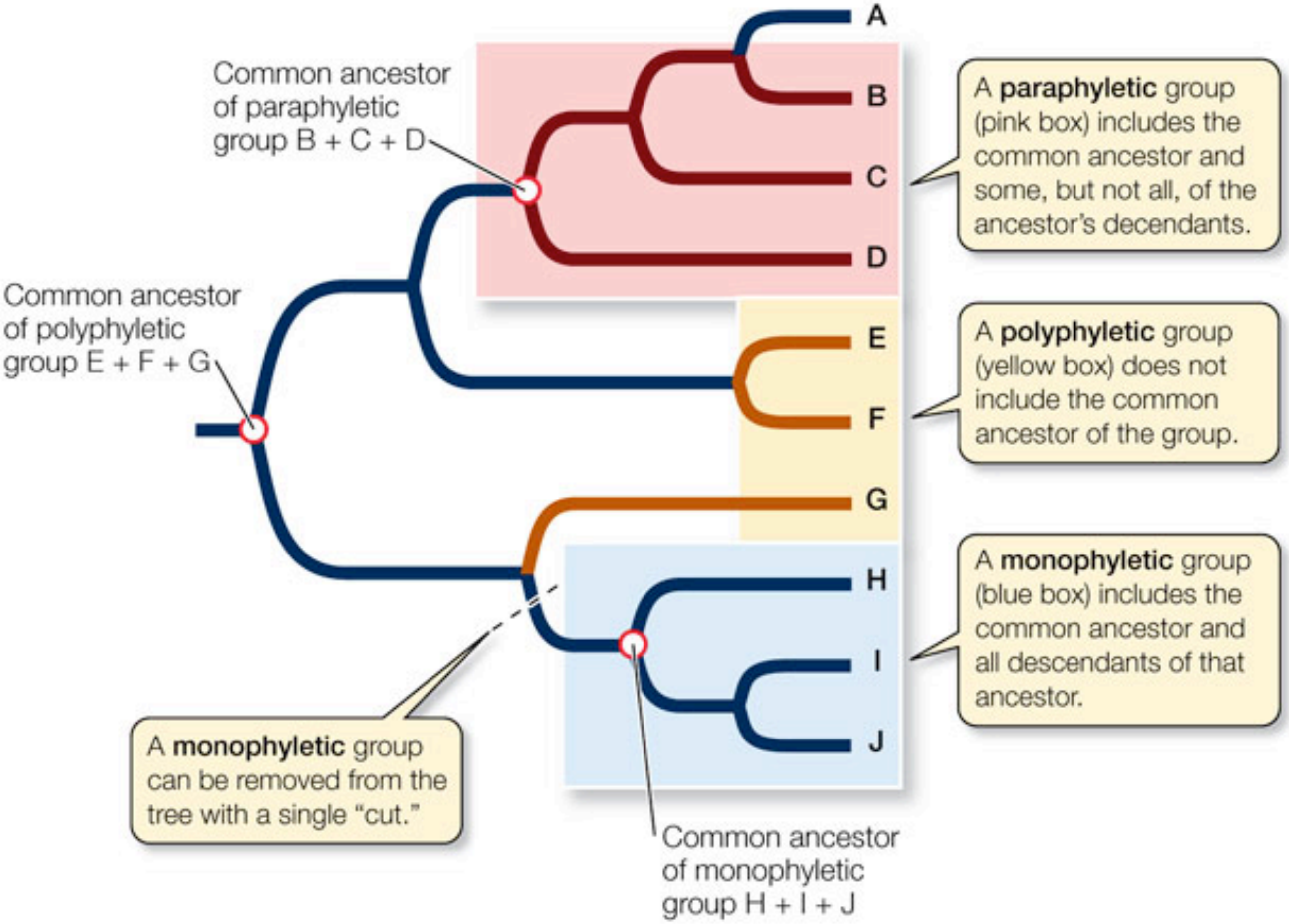


(a) Monophyletic group (clade)

(b) Paraphyletic group

(c) Polyphyletic group

Cutting with scissors



Common ancestor of paraphyletic group B + C + D

Common ancestor of polyphyletic group E + F + G

A **paraphyletic** group (pink box) includes the common ancestor and some, but not all, of the ancestor's decendants.

A **polyphyletic** group (yellow box) does not include the common ancestor of the group.

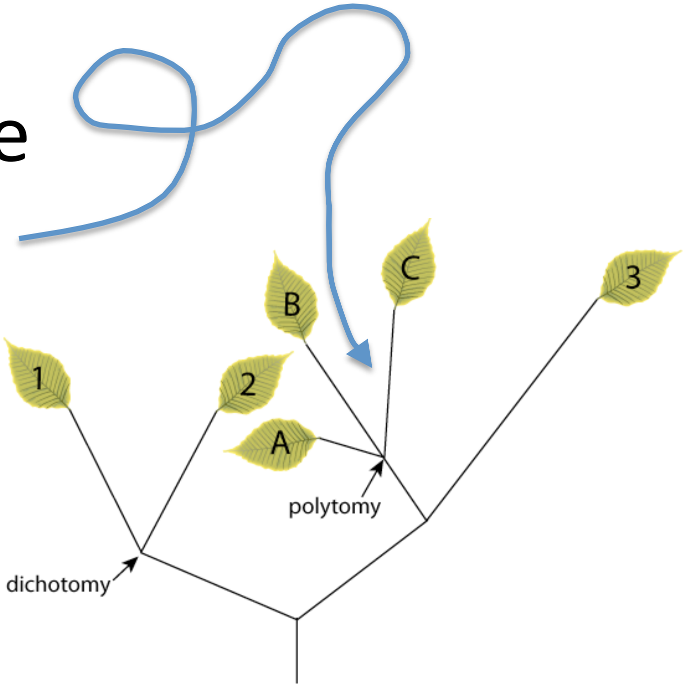
A **monophyletic** group (blue box) includes the common ancestor and all descendants of that ancestor.

A **monophyletic** group can be removed from the tree with a single "cut."

Common ancestor of monophyletic group H + I + J

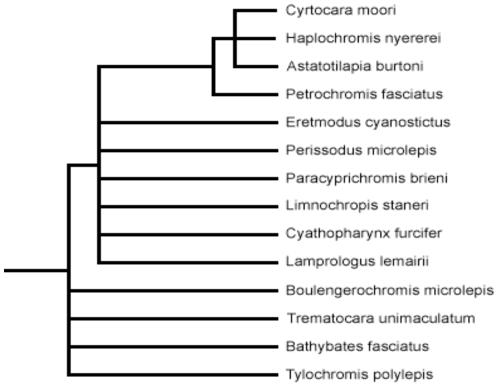
Binary branching

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

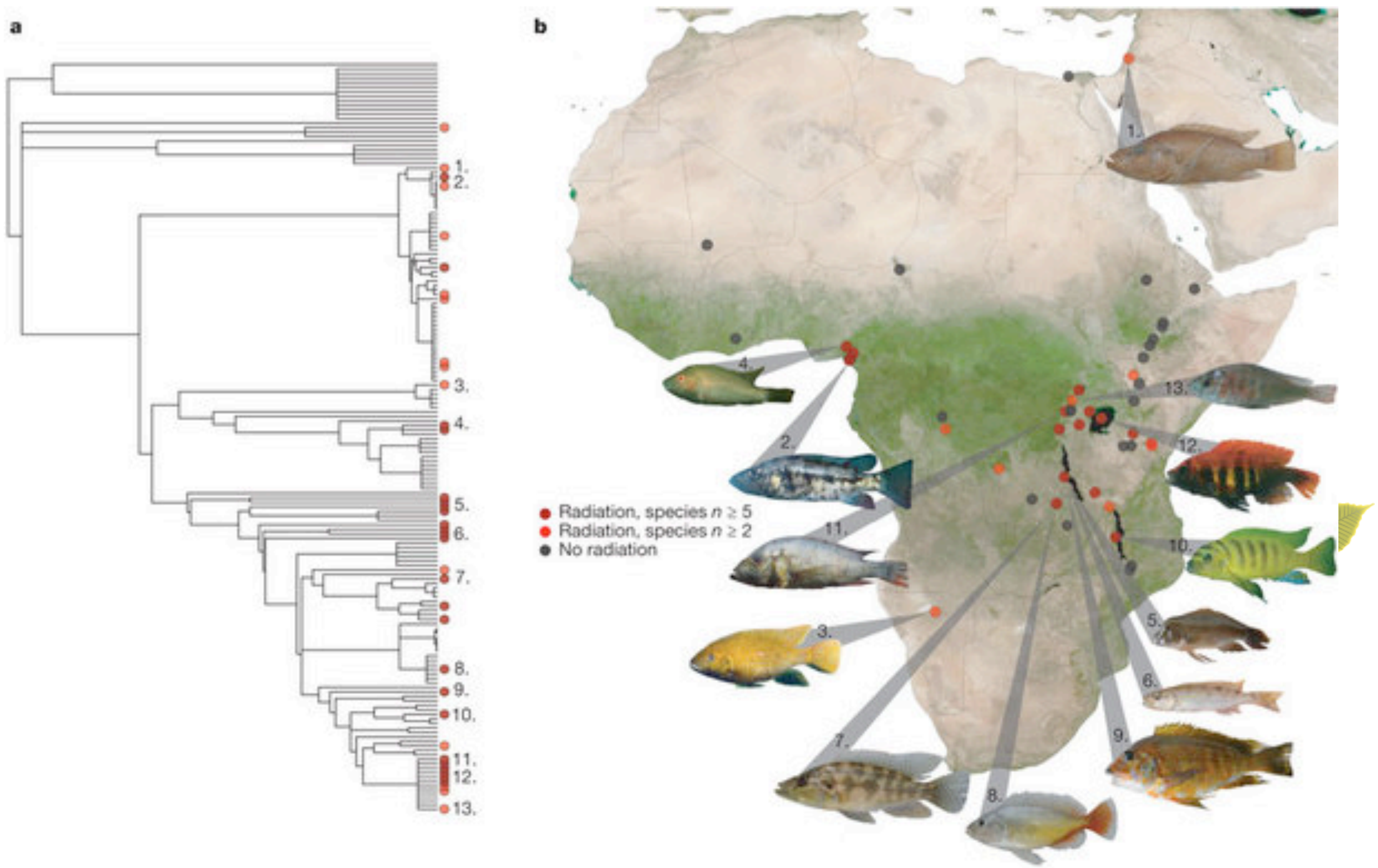


Deviations from binary branching

- 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)



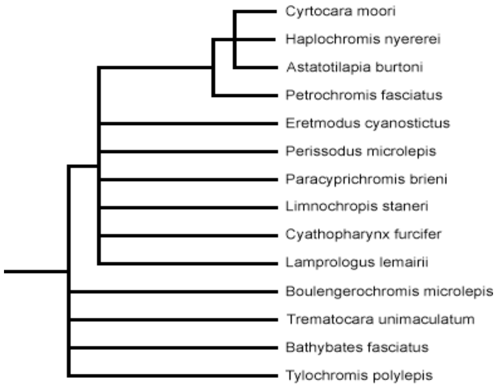
Deviations from binary branching



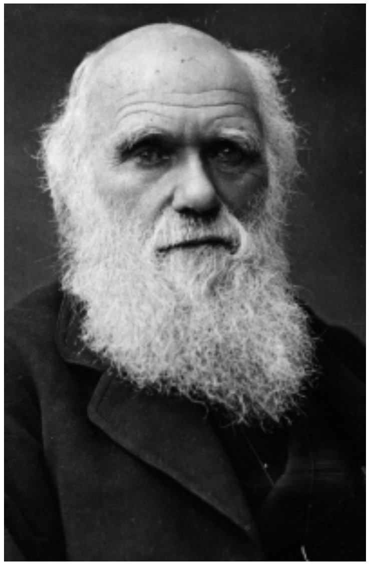
Wagner et al. (2012) Ecological opportunity and sexual selection together predict adaptive radiation. Nature: doi:10.1038/nature11144

Deviations from binary branching

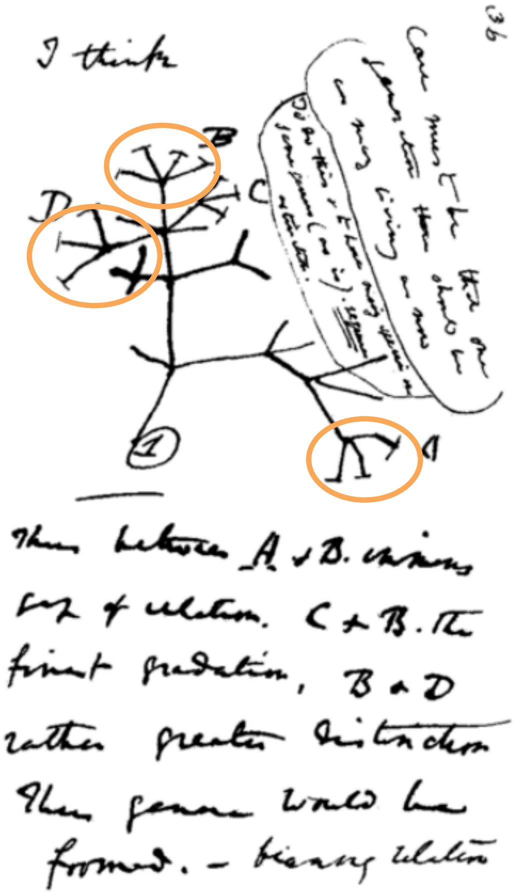
- Polytomy a node with >2 descendent lineages
  - Hard polytomy = real speciation event involving > 2 lineages diverging from a common ancestor
  - **Soft polytomy** = insufficient phylogenetic information, uncertain tree topology (uh oh)



The Origin of Species



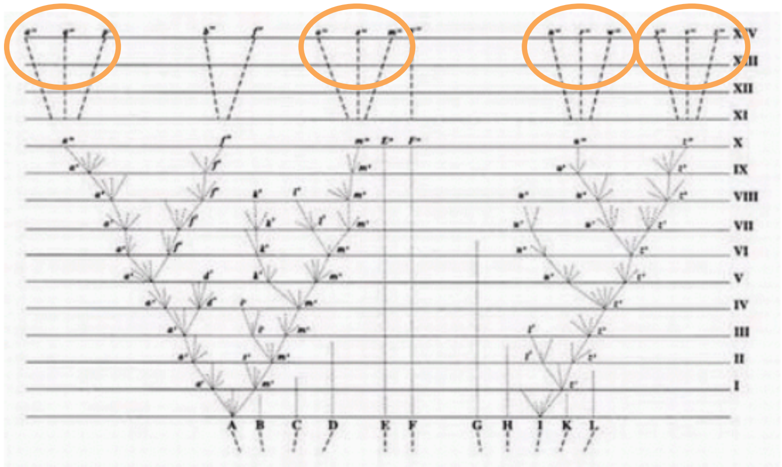
Father of evolution:
Chuck D.



Case must be that one generation then should be as many living as now. To do this & to have many species in same genus (as is) requires extinction.

Then between A & B. immens gap of relation. C & B. the finest gradation, B & D rather greater distinction Then genera would be formed. — bearing 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

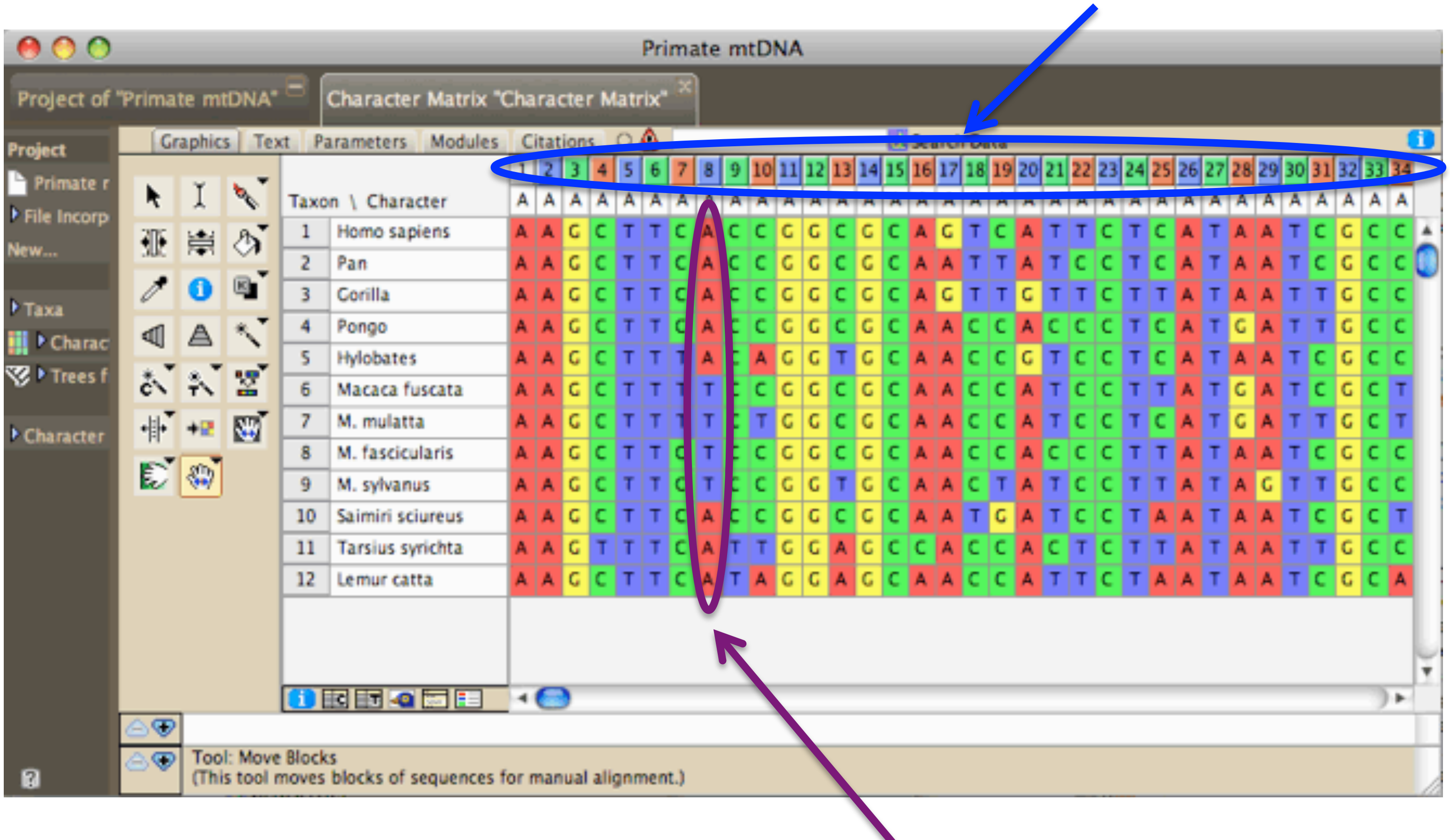


Characters vs States

- 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

Character (site)

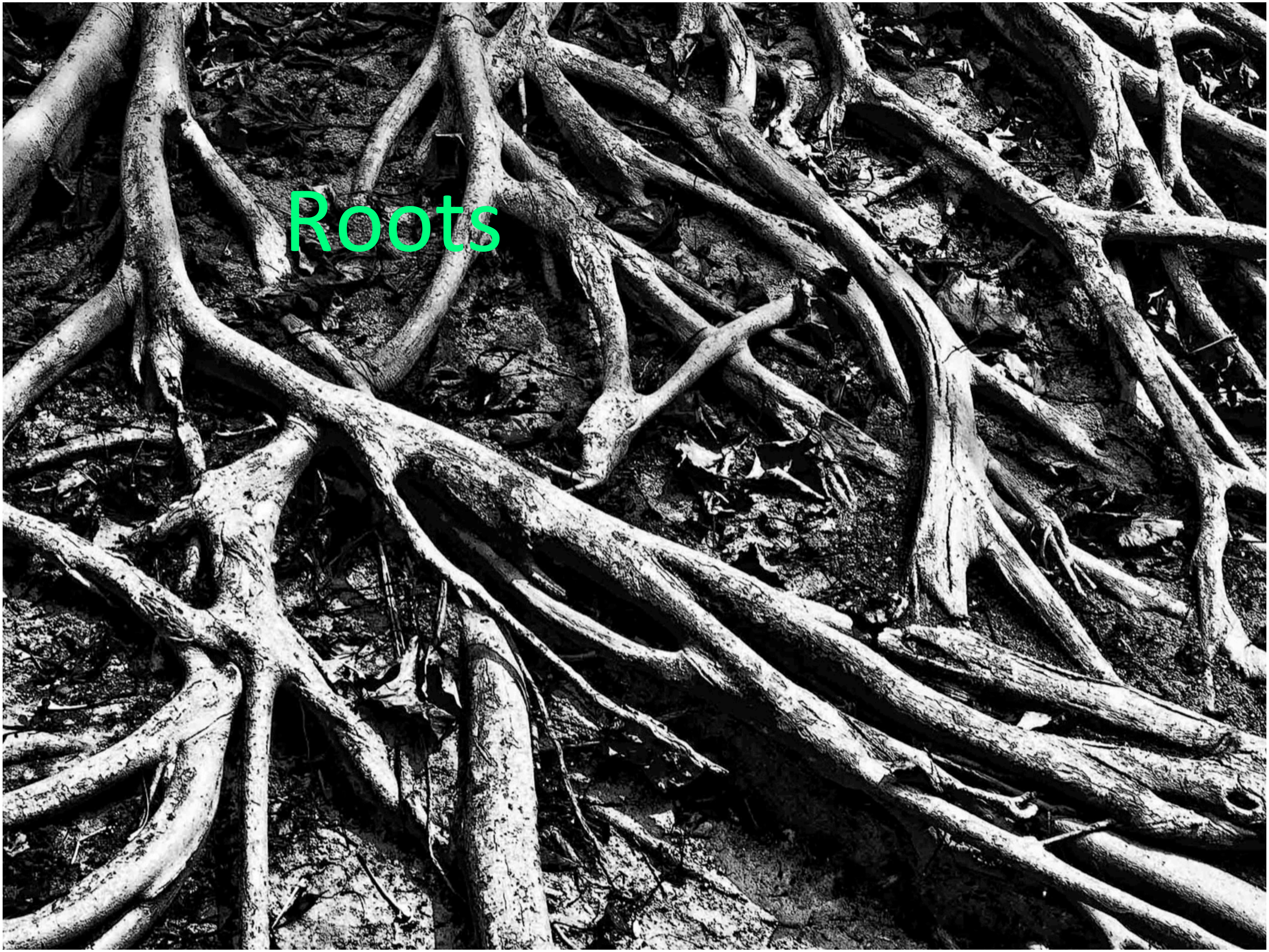


	Taxon \ Character	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
		A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
1	Homo sapiens	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	G	T	C	A	T	T	C	T	C	A	T	A	A	T	C	G	C	C
2	Pan	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	A	T	T	A	T	C	C	T	C	A	T	A	A	T	C	G	C	C
3	Gorilla	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	G	T	T	G	T	T	C	T	T	A	T	A	A	T	T	G	C	C
4	Pongo	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	A	C	C	A	C	C	C	T	C	A	T	G	A	T	T	G	C	C
5	Hylobates	A	A	G	C	T	T	T	A	C	A	G	G	T	G	C	A	A	C	C	G	T	C	C	T	C	A	T	A	A	T	C	G	C	C
6	Macaca fuscata	A	A	G	C	T	T	T	T	C	C	G	G	C	G	C	A	A	C	C	A	T	C	C	T	T	A	T	G	A	T	C	G	C	T
7	M. mulatta	A	A	G	C	T	T	T	T	C	T	G	G	C	G	C	A	A	C	C	A	T	C	C	T	C	A	T	G	A	T	T	G	C	T
8	M. fascicularis	A	A	G	C	T	T	C	T	C	C	G	G	C	G	C	A	A	C	C	A	C	C	C	T	T	A	T	A	A	T	C	G	C	C
9	M. sylvanus	A	A	G	C	T	T	C	T	C	C	G	G	T	G	C	A	A	C	T	A	T	C	C	T	T	A	T	A	G	T	T	G	C	C
10	Saimiri sciureus	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	A	T	G	A	T	C	C	T	A	A	T	A	A	T	C	G	C	T
11	Tarsius syrichta	A	A	G	T	T	T	C	A	T	T	G	G	A	G	C	C	A	C	C	A	C	T	C	T	T	A	T	A	A	T	T	G	C	C
12	Lemur catta	A	A	G	C	T	T	C	A	T	A	G	G	A	G	C	A	A	C	C	A	T	T	C	T	A	A	T	A	A	T	C	G	C	A

Tool: Move Blocks
(This tool moves blocks of sequences for manual alignment.)

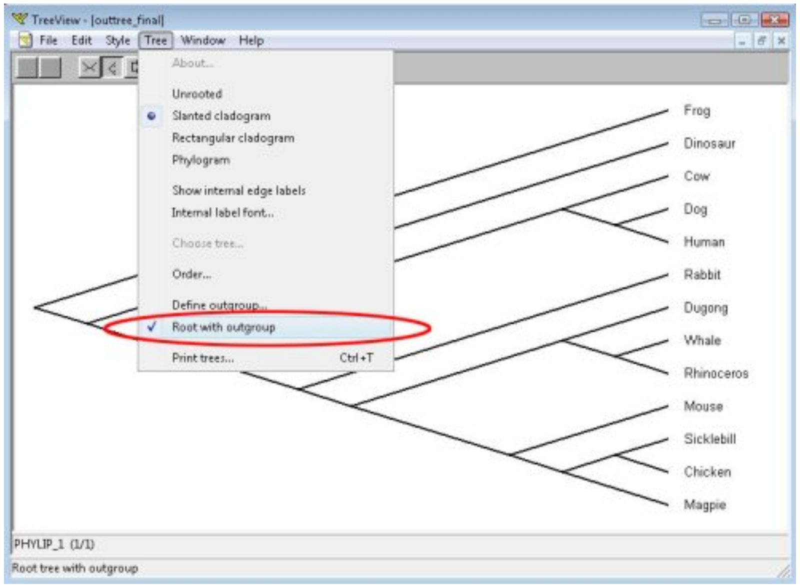
State (nucleotide)

Roots

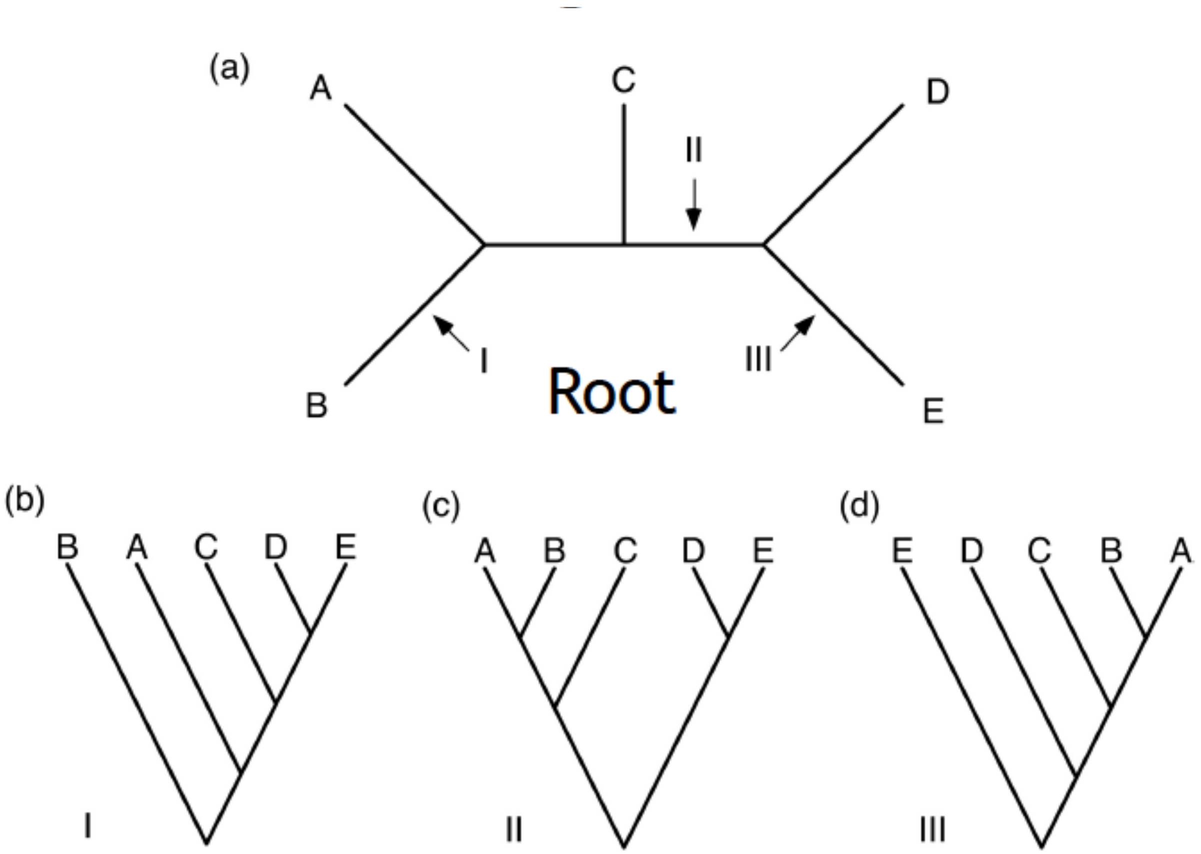


Rooting trees

- 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 <u>known</u> to be distantly related to ingroup)
- *One node* between outgroup and ingroup is identified as the root

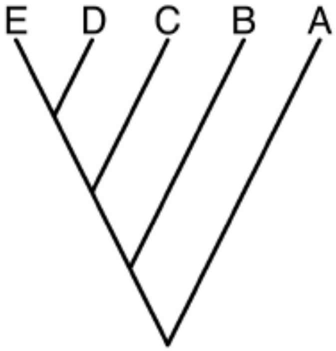


Rooting trees

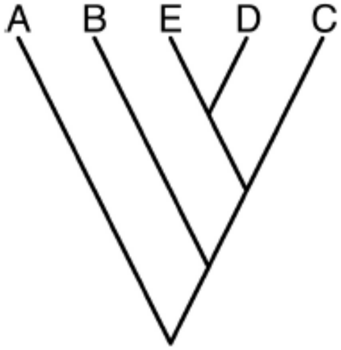


Flipping branches

Which are Different?



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



How many rooted and unrooted possibilities are there?

How many rooted and unrooted possibilities are there?

Number of OTUs	# rooted trees	# unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
6	954	105
7	10,395	954
8	135,135	10,395
9	2,027,025	135,135
10	34,459,425	2,027,025

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

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

1
- Collect data i.e. DNA

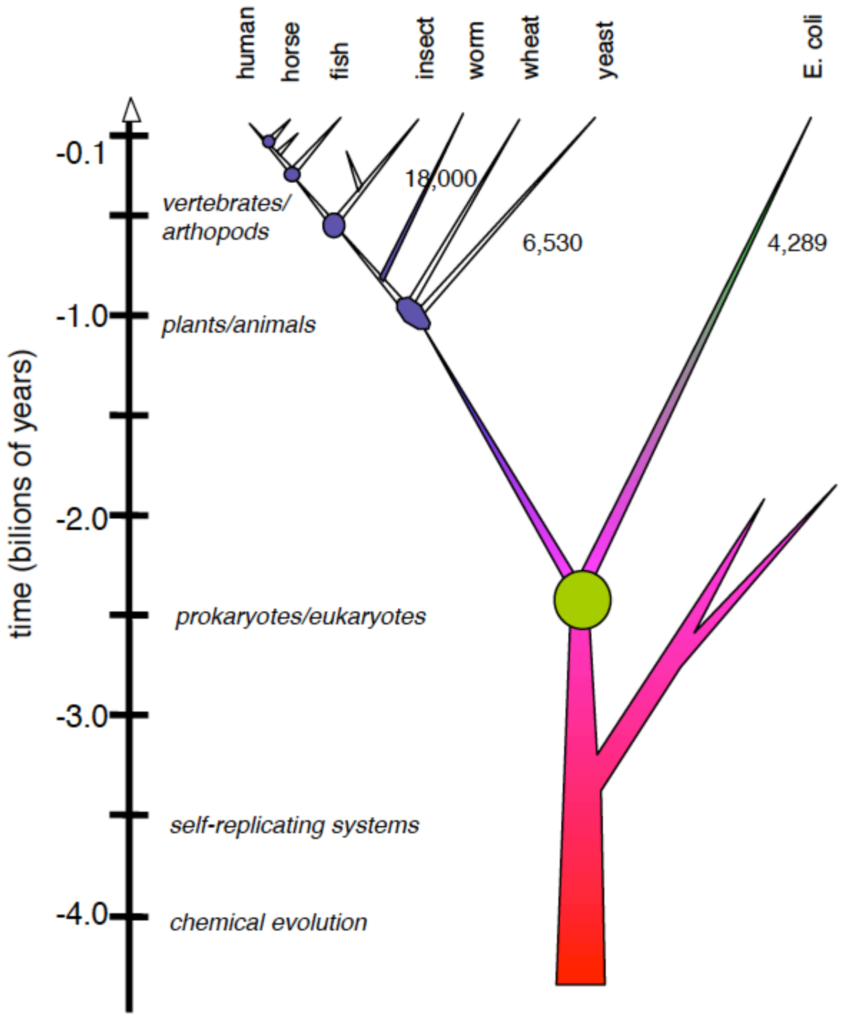
2
- Retrieve homologous sequences

3
- Multiple sequence alignment

4
- Model selection

5
- Assessing confidence in topology

Homologues share a common ancestor



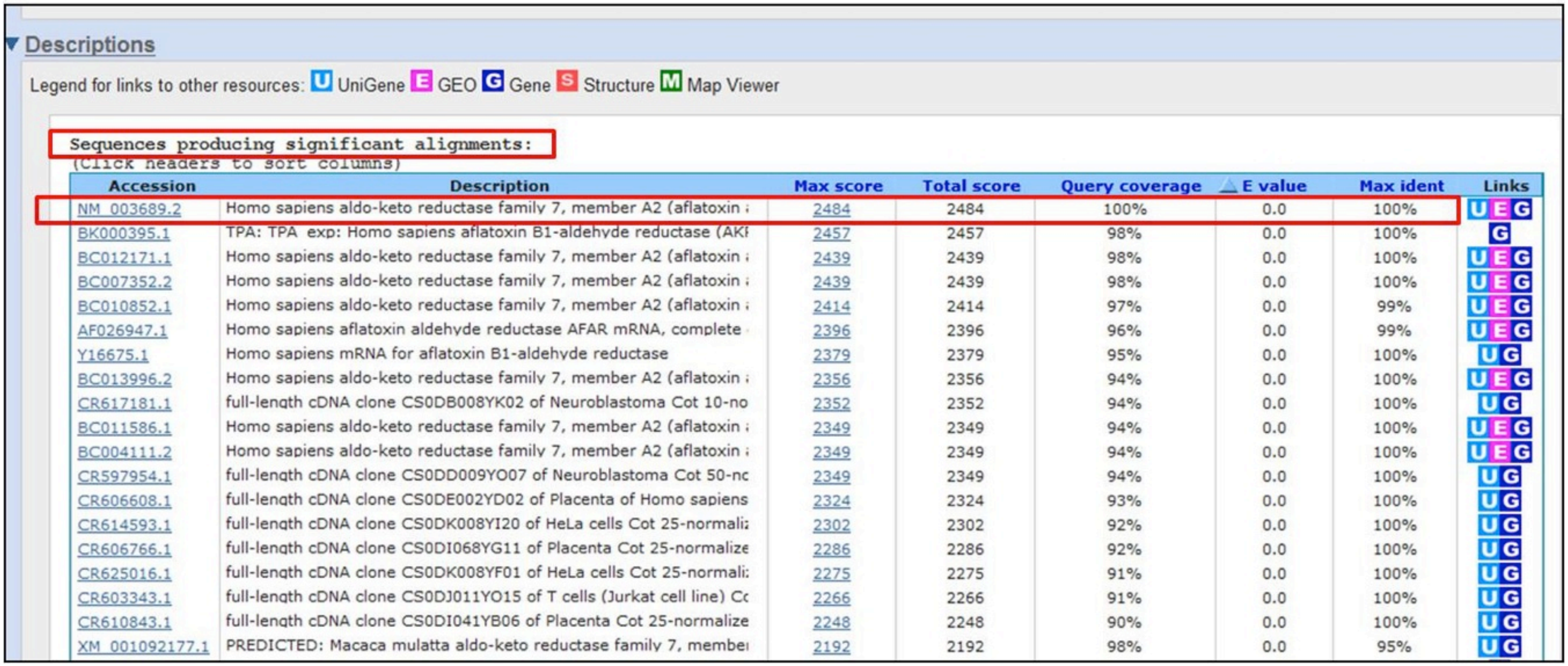
Retrieve homologous sequences

- 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 (AK	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 CS0DD009YO07 of Neuroblastoma Cot 50-n	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-normali	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) C	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, membe	2192	2192	98%	0.0	95%	U G

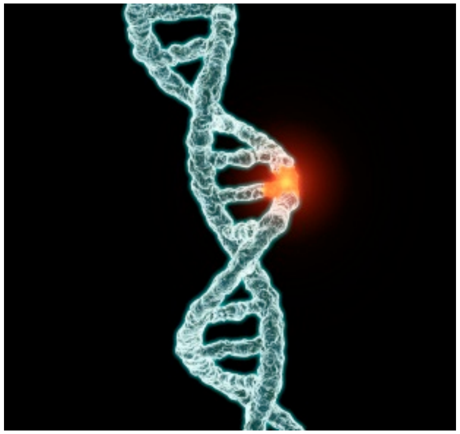
How to build a phylogenetic tree

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

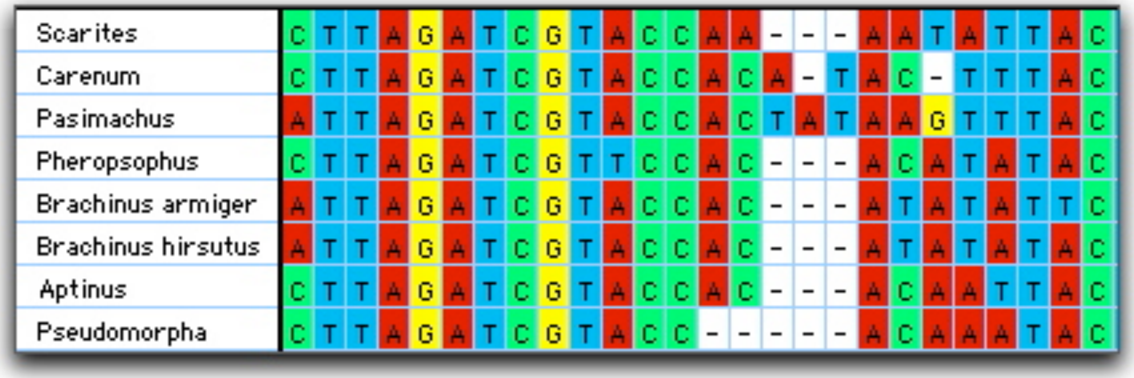
refresher

- Evolution of visible (phenotypic) characters is the result of changes at the molecular level
- types of mutations:
  - indel (frameshift)
    - Insertions/deletions
    - frameshift
  - substitutions



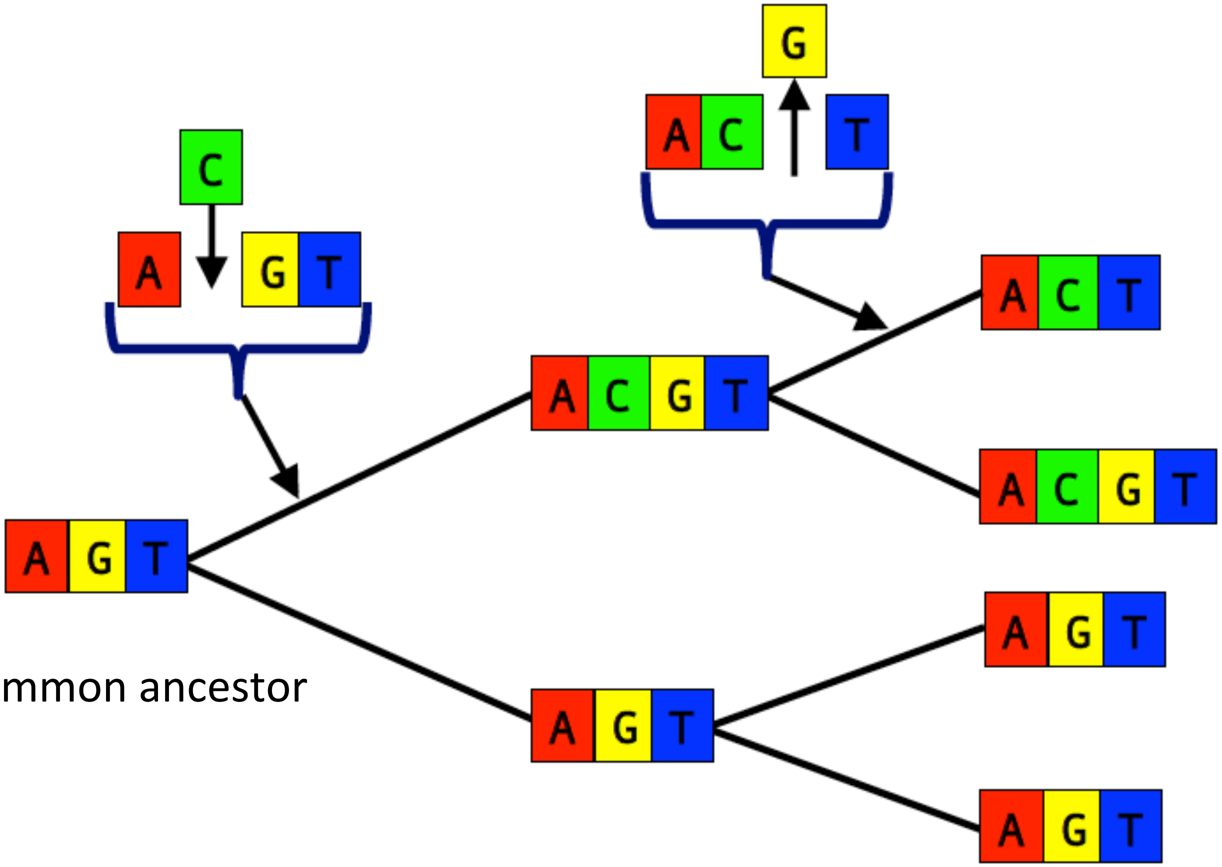
Multiple sequence alignment

- 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

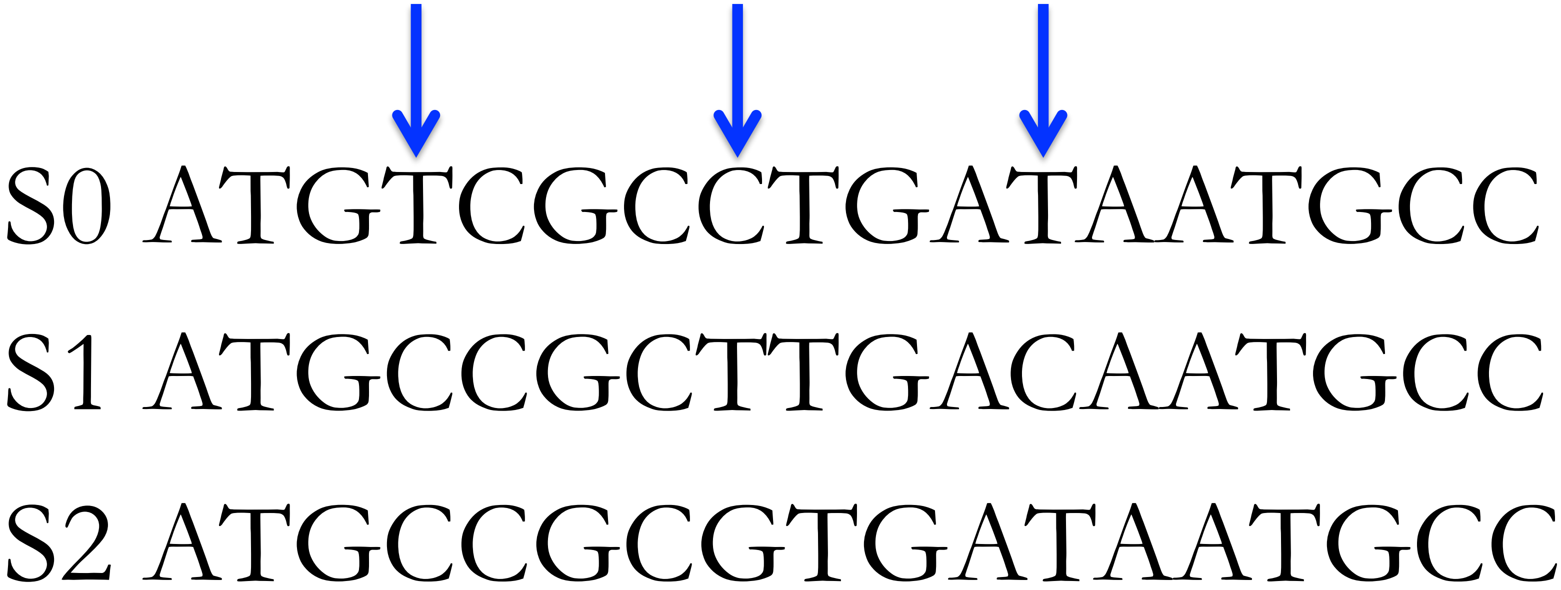
MSA in action



Common ancestor

A	C	-	T
A	C	G	T
A	-	G	T
A	-	G	T

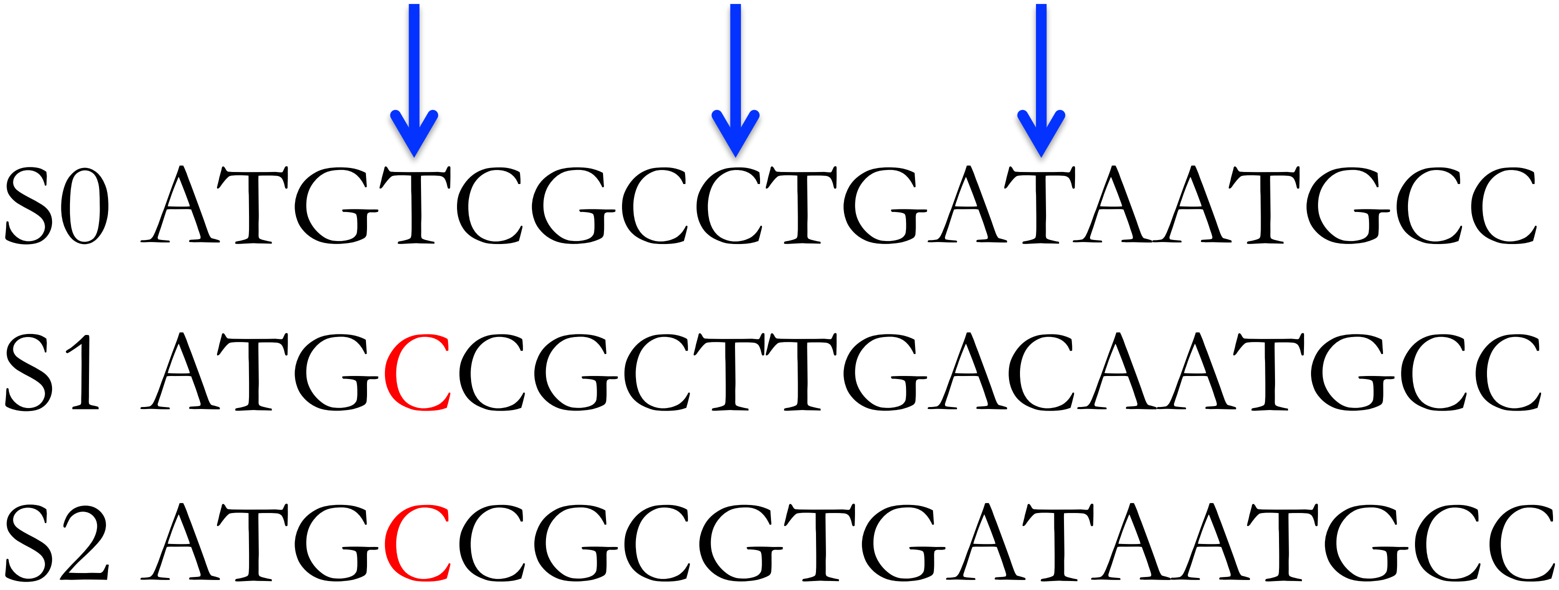
4 x descendent taxa



S0 ATGTCGCCTGATAATGCC

S1 ATGCCGCTTGACAATGCC

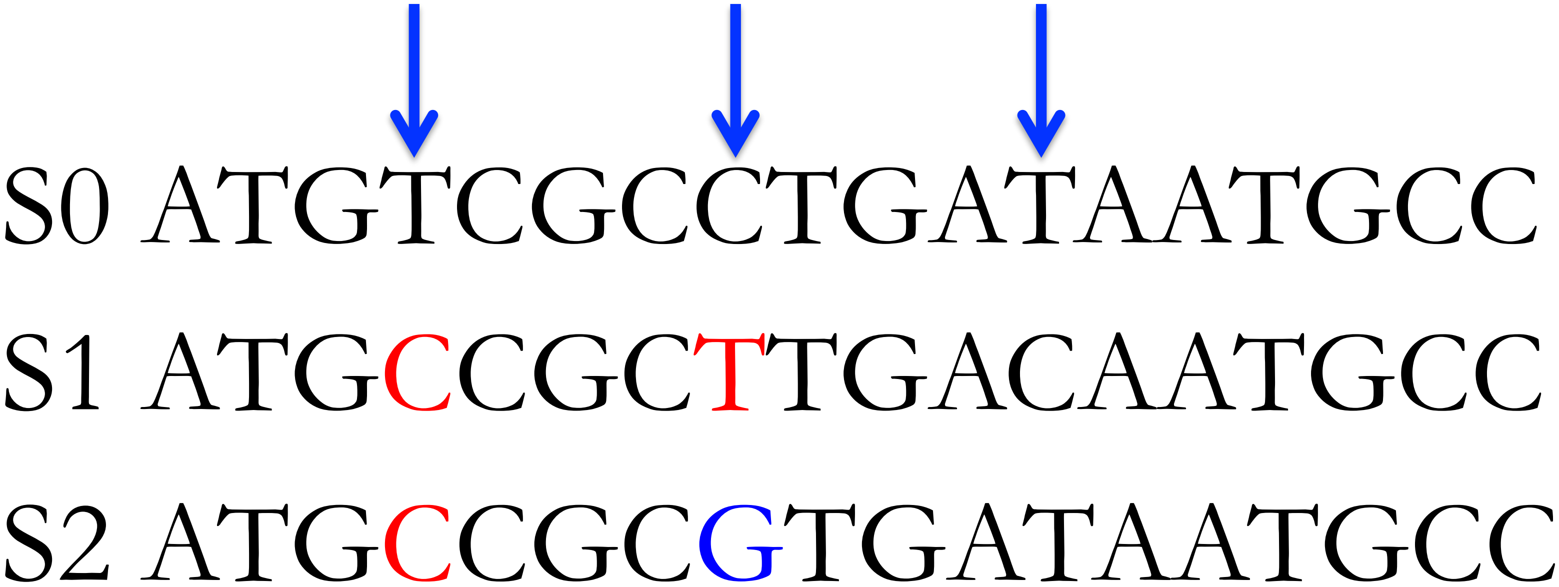
S2 ATGCCGCGTGATAATGCC



S0 ATGTCGCCTGATAATGCC

S1 ATGCCGCTTGACAATGCC

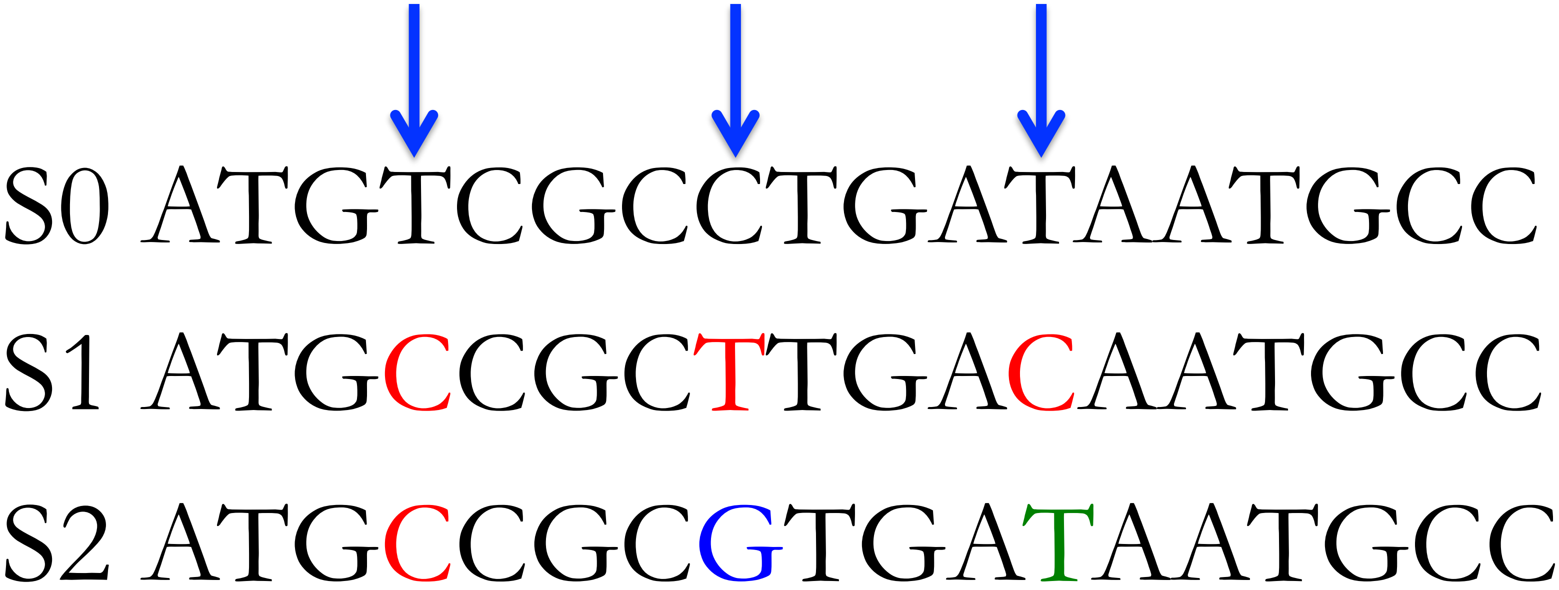
S2 ATGCCGCGTGATAATGCC



S0 ATGTCGCCTGATAATGCC

S1 ATGCCGCTTGACAATGCC

S2 ATGCCGCGTGATAATGCC



S0 ATGTCGCCTGATAATGCC

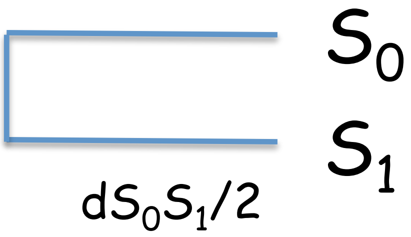
S1 ATGCCGCTTGACAATGCC

S2 ATGCCGCGTGATAATGCC

S3…

Distance - UPGMA

OTU	**S0**	**S1**	**S2**
S1	dS_0S_1		
S2	dS_0S_2	dS_1S_2	
S3	dS_0S_3	dS_1S_3	dS_2S_3



Distance - UPGMA

OTU	(S0S1)	S2
S2	$D(S_0S_1)S_2$	
S3	$D(S_0S_1)S_3$	dS_2S_3