

20.109
Laboratory Fundamentals in
Biological Engineering

Module 1
Nucleic Acid Engineering
Lecture 6

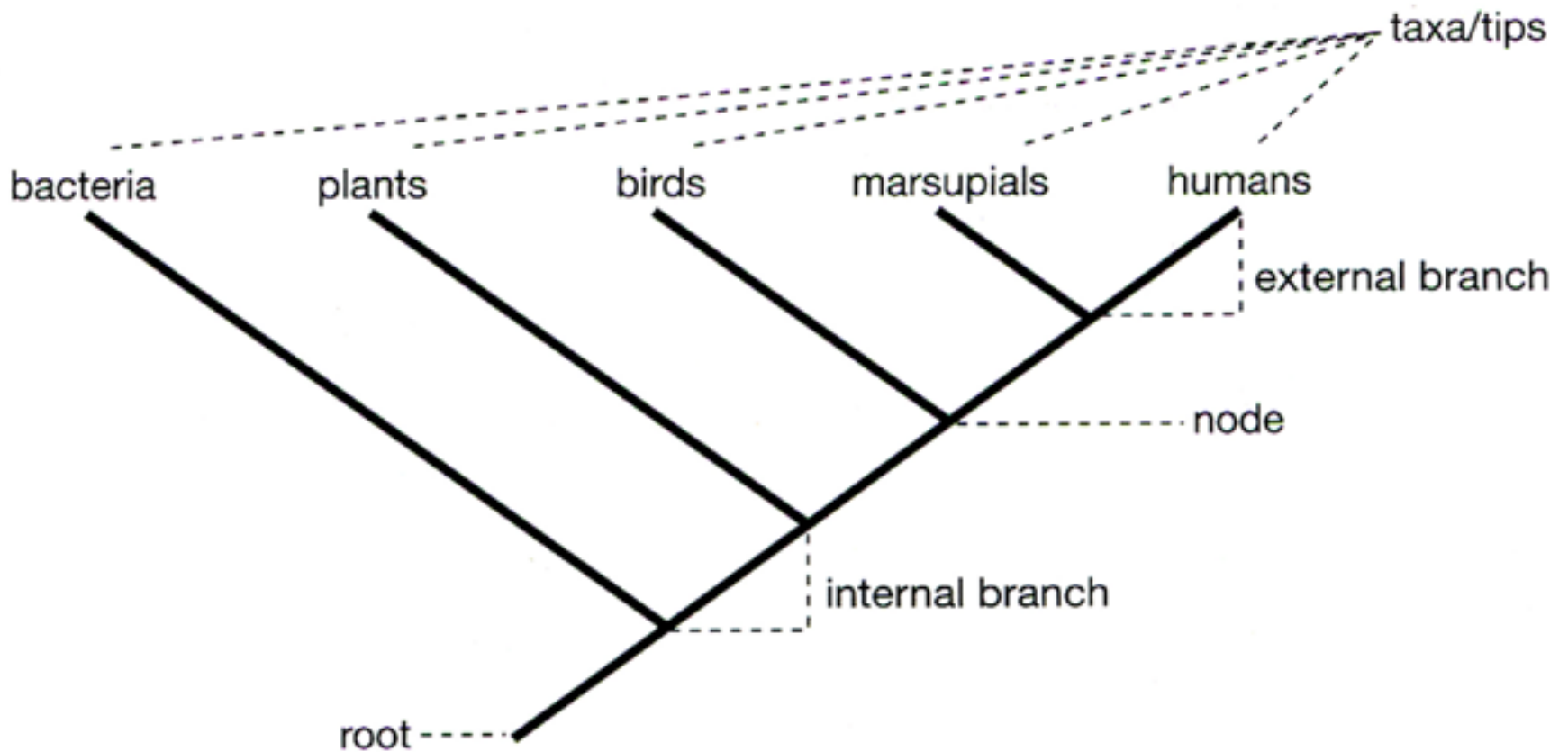
Today

Finish Phylogenetics 1.0

The other side of sequencing

Comparing microbial communities

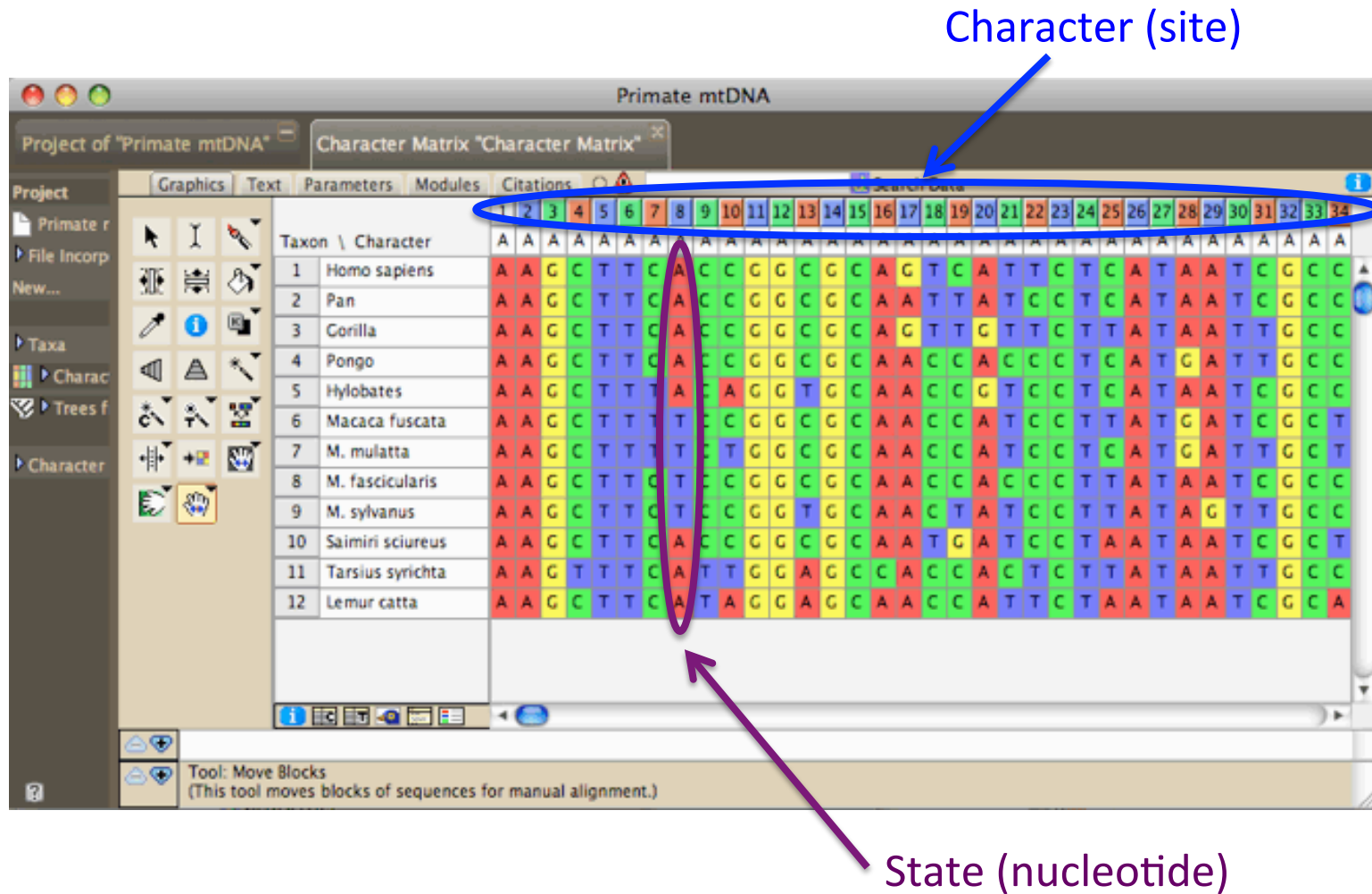
Some important terms used to describe phylogenetic trees

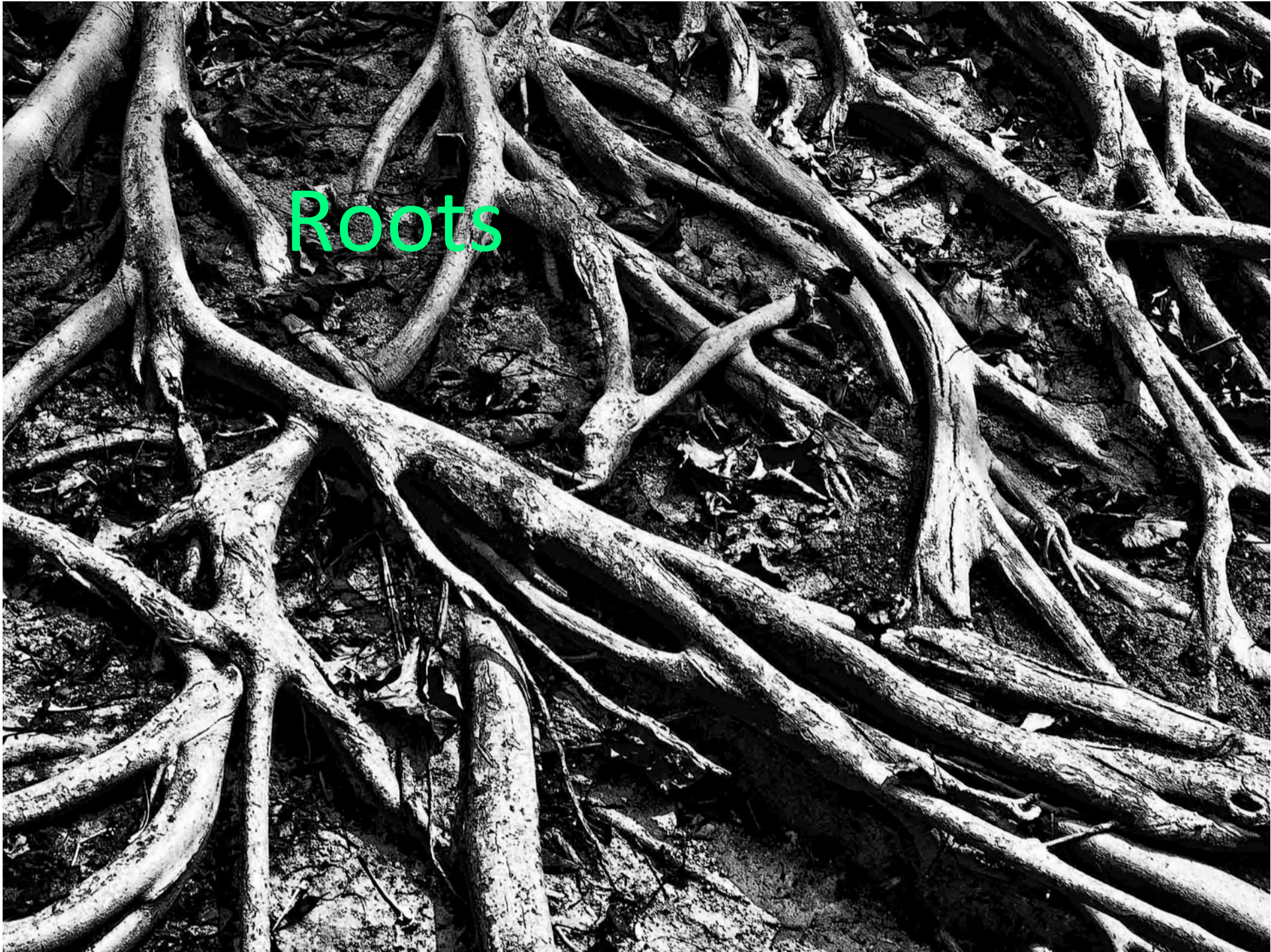


Characters vs States

- **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

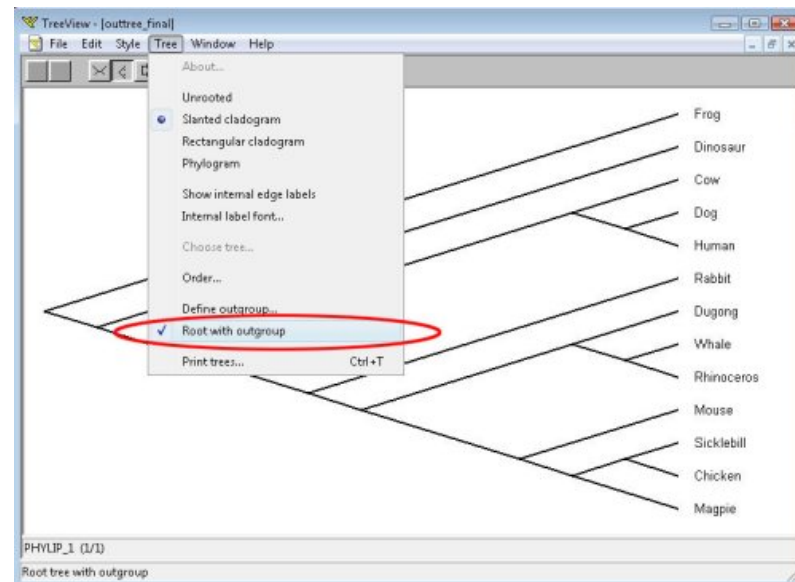




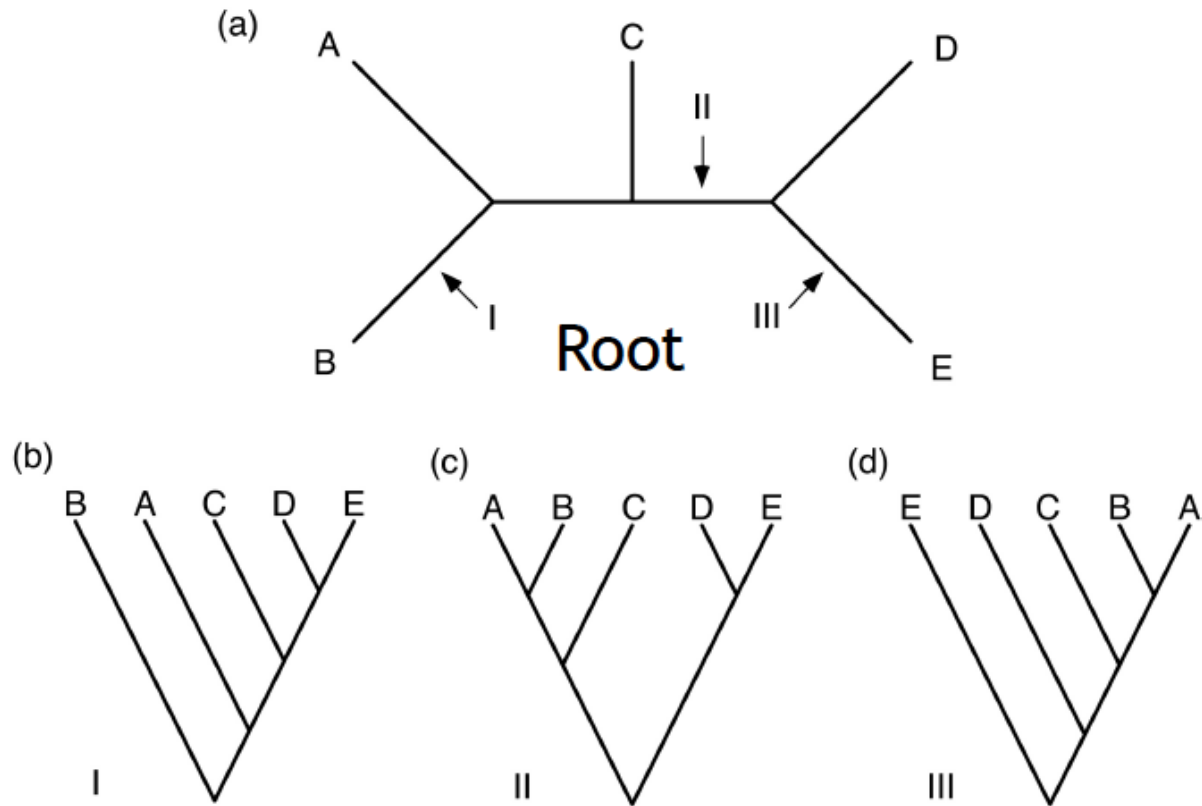
Roots

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

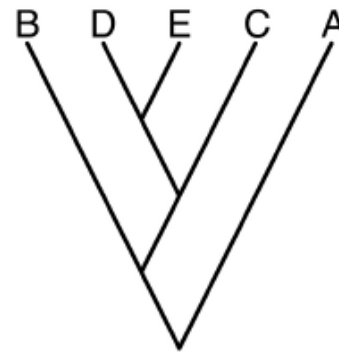
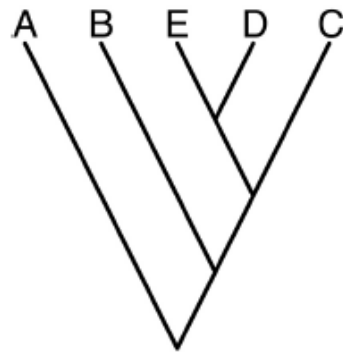
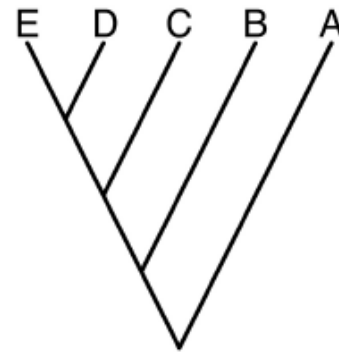
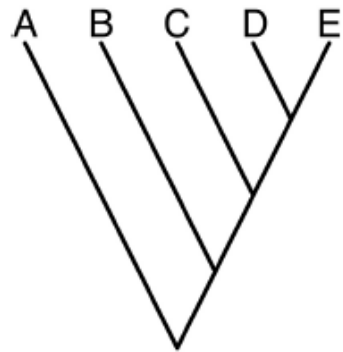


Rooting trees



Flipping branches

Which are Different?



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

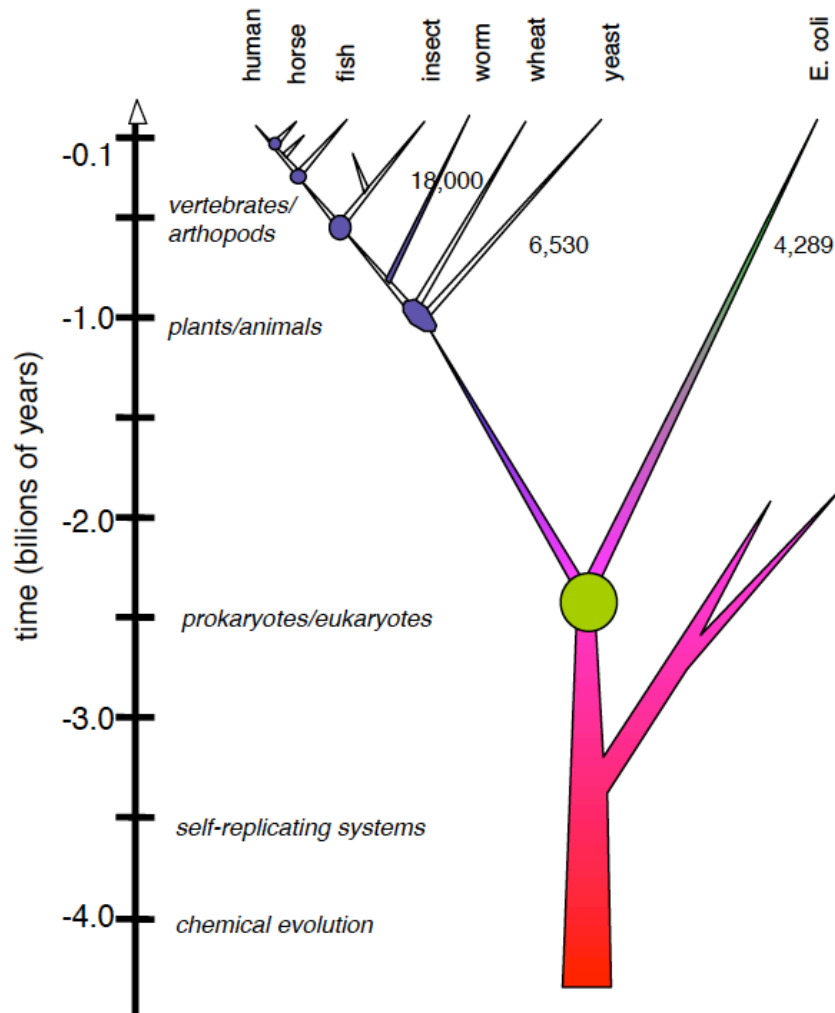
4

- 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 reflects 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-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) 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

How to build a phylogenetic tree

1

- Collect data i.e. DNA

2

- Retrieve homologous sequences

3

- Multiple sequence alignment

4

- Model selection

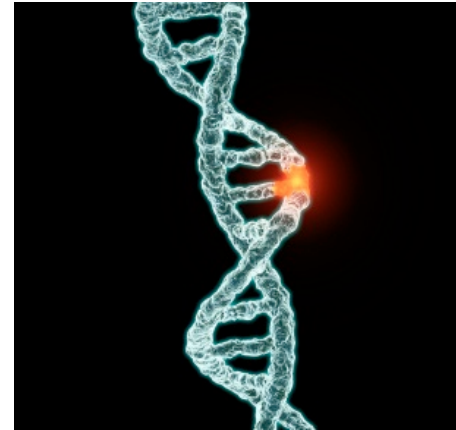
5

- 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

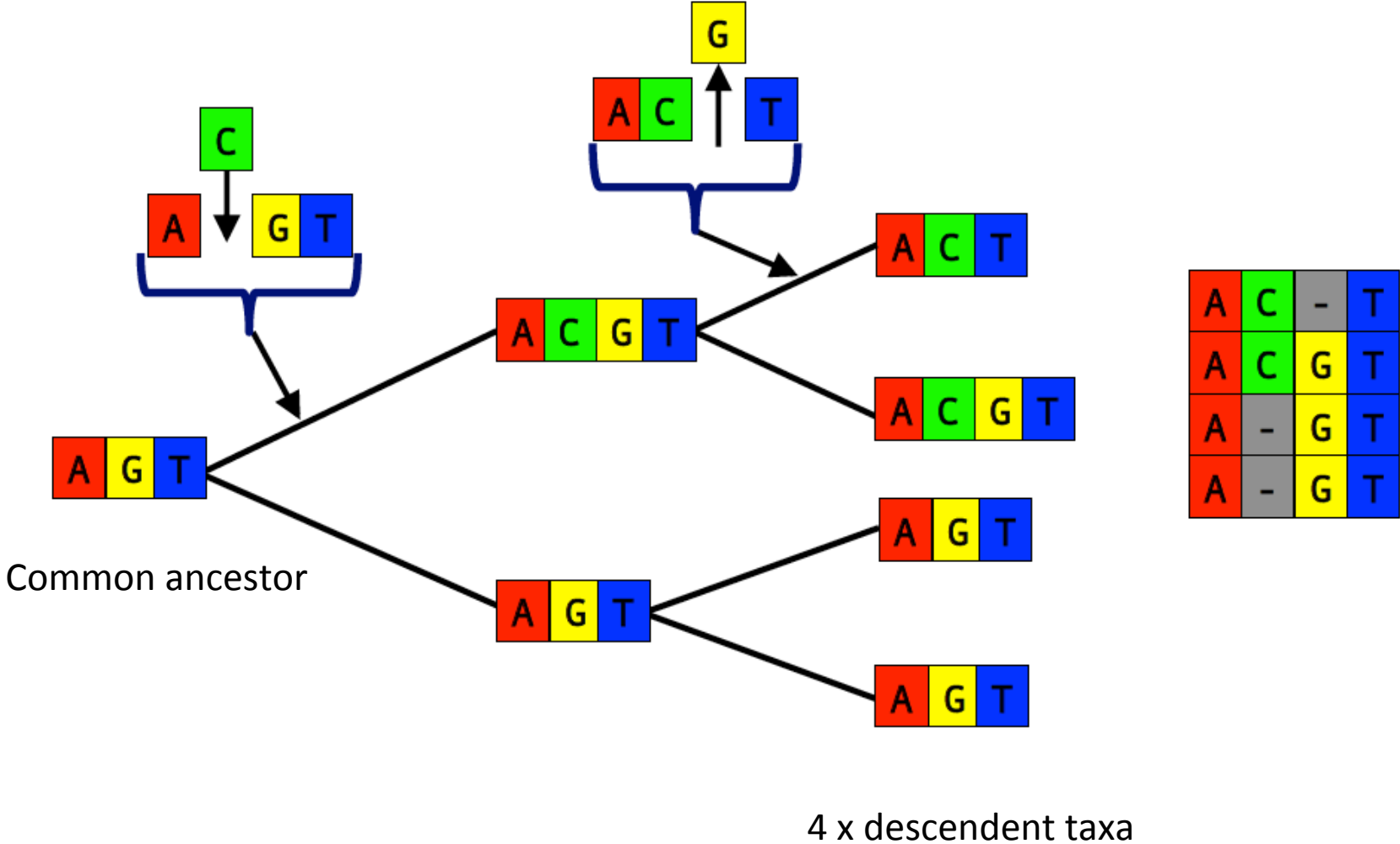


Multiple sequence alignment

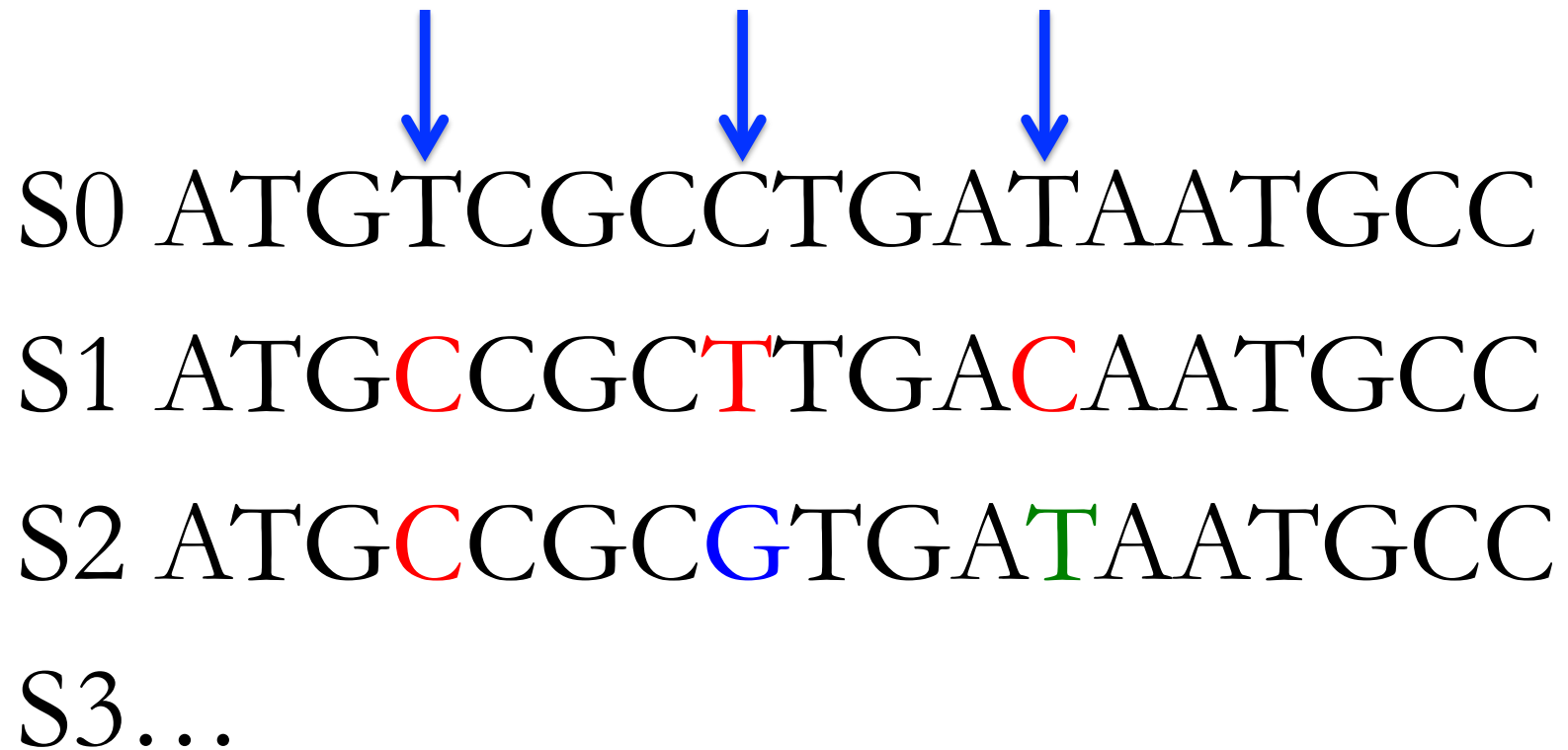
- Insertions & deletions ('indels') obscure sites that are homologous (= traits descended from common ancestor)
- Goal of MSA is to introduce **gaps** so that nucleotides in same column are homologous

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



S0 ATGTCGCCTGATAATGCC
S1 ATGCCGCTTGACAATGCC
S2 ATGCCGC GTGATAATGCC
S3...



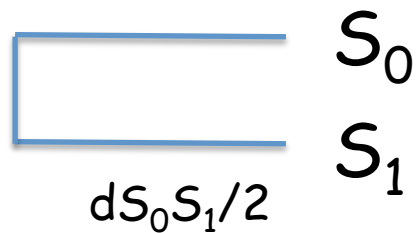
The diagram illustrates a sequence alignment between a reference sequence (S0) and three variant sequences (S1, S2, S3). The reference sequence S0 is ATGTCGCCTGATAATGCC. Three blue arrows point to the 4th, 7th, and 10th positions of S0. In S1, the 4th, 7th, and 10th positions are mutated to C, T, and C, respectively. In S2, the 4th, 7th, and 10th positions are mutated to C, G, and A, respectively. S3 is shown as an ellipsis, indicating further sequences.

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

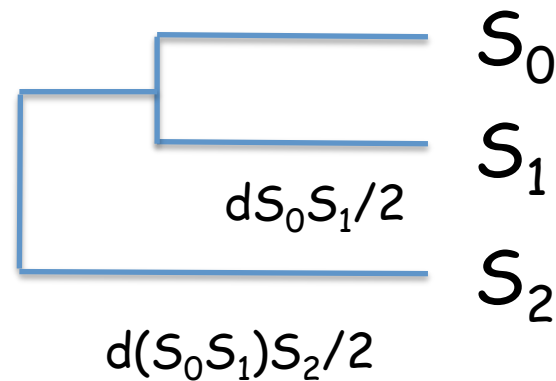
Distance - UPGMA

OTU	S0	S1	S2
S1	$d_{S_0S_1}$		
S2	$d_{S_0S_2}$	$d_{S_1S_2}$	
S3	$d_{S_0S_3}$	$d_{S_1S_3}$	$d_{S_2S_3}$

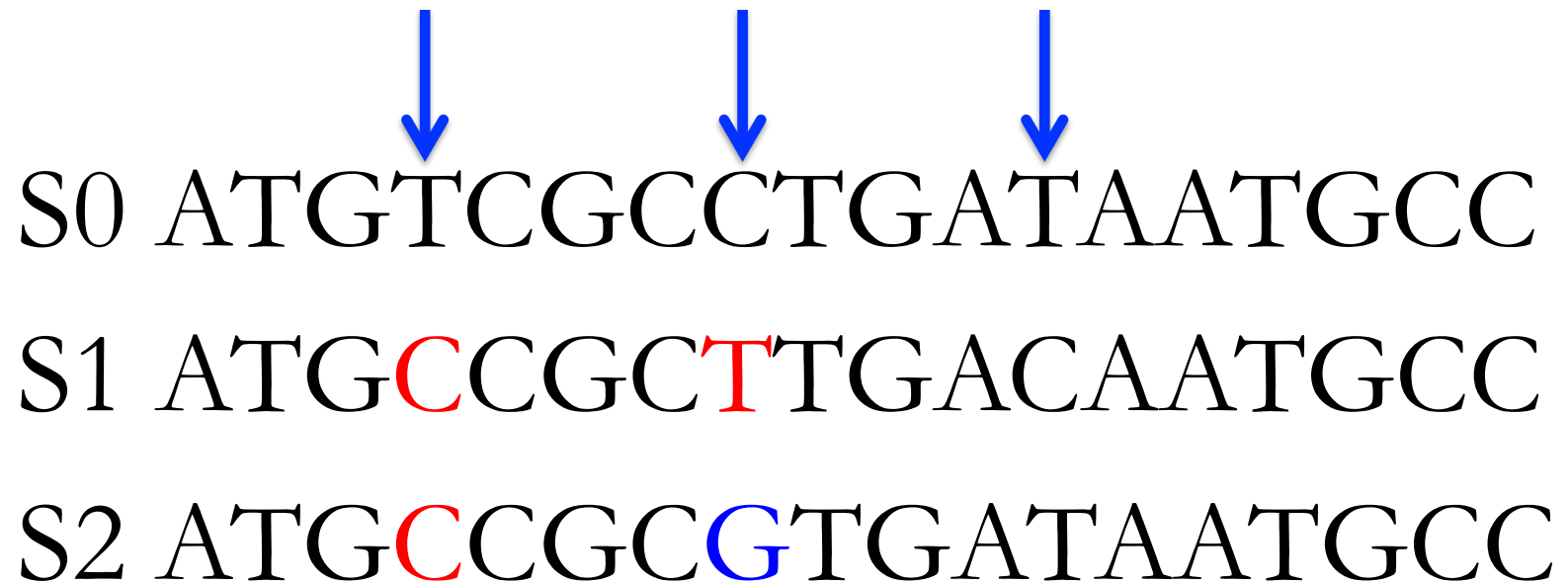


Distance - UPGMA

OTU	(S ₀ S ₁)	S ₂
S ₂	$D(S_0S_1)S_2$	
S ₃	$D(S_0S_1)S_3$	dS_2S_3



S0 ATGTCGCCTGATAATGCC
S1 ATGCCGCTTGACAATGCC
S2 ATGCCGC GTGATAATGCC



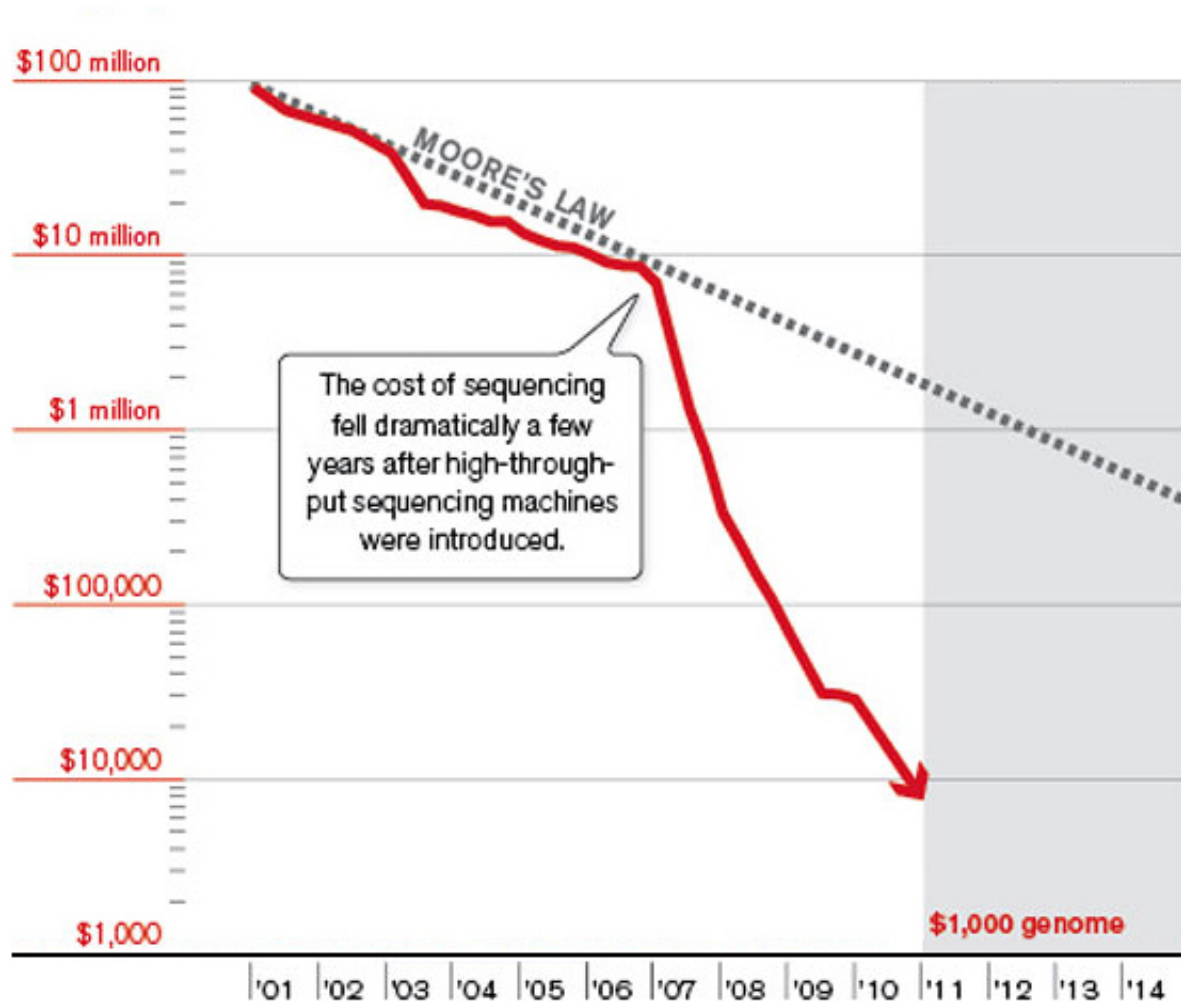
Bases to Bytes (Technology Review April 2012)

Cheap sequencing technology is flooding the world with genomic data.

Can we handle the deluge?

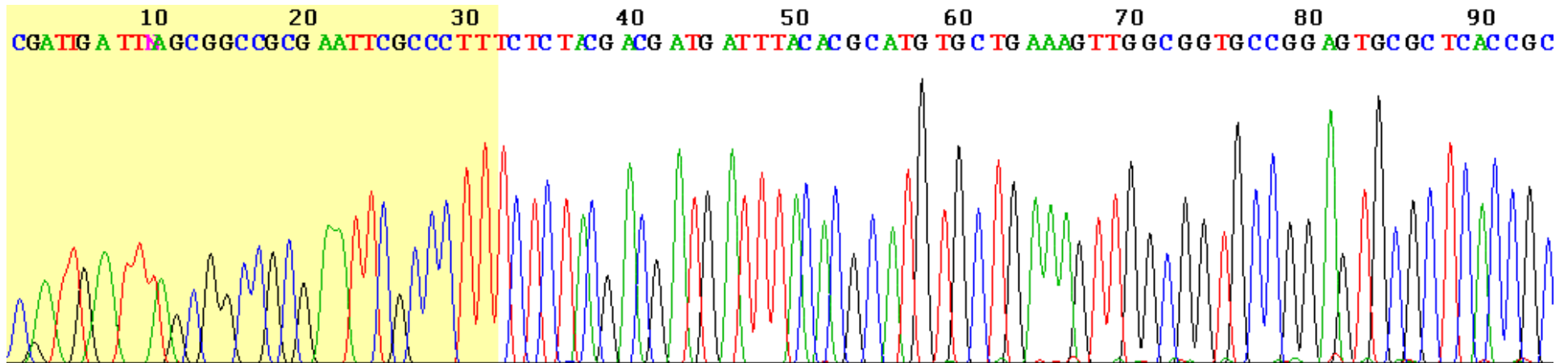
Sequencing Costs Plummeting

Cost per genome

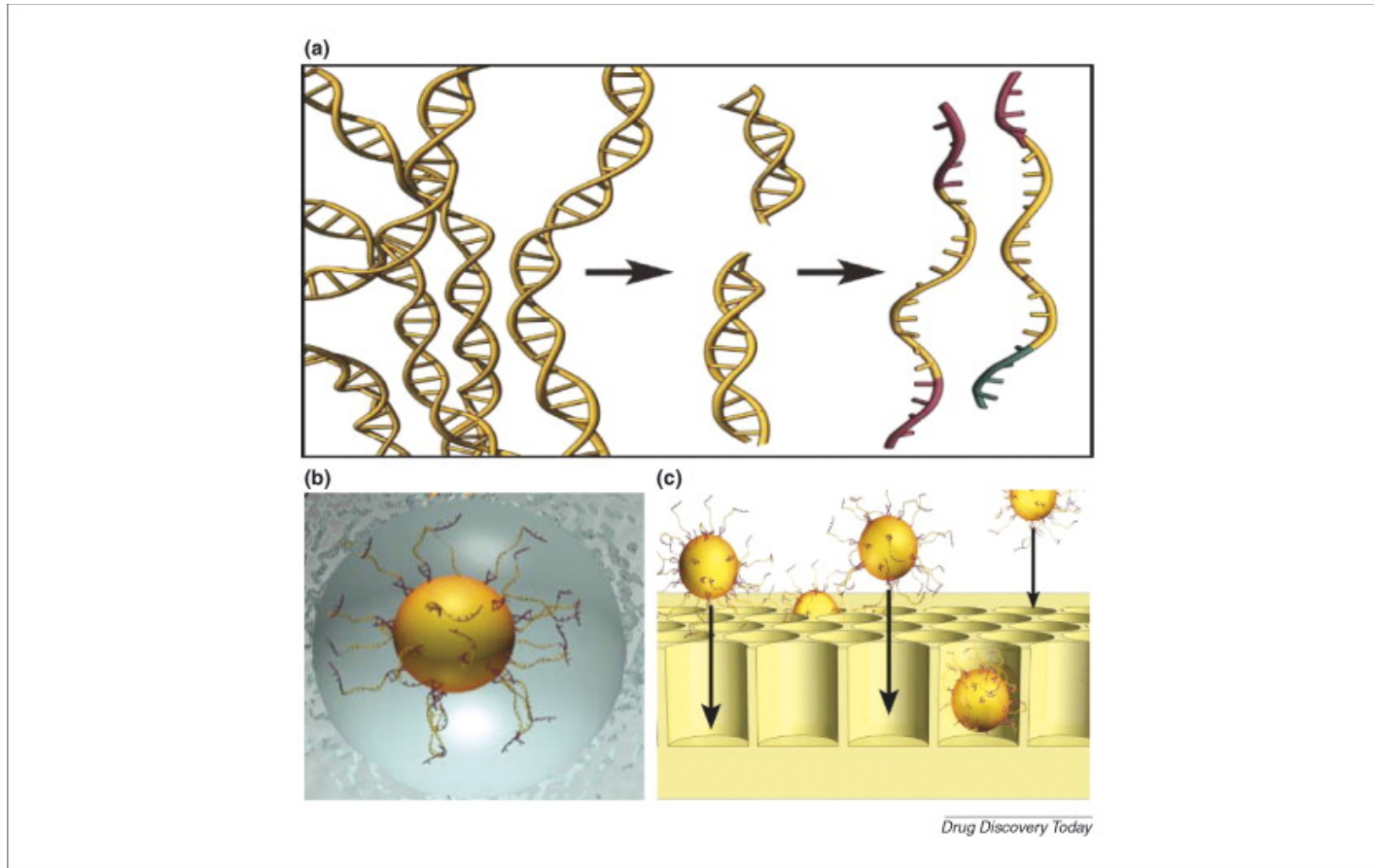


Sanger sequencing

<http://www.wellcome.ac.uk/Education-resources/Education-and-learning/Resources/Animation/WTDV026689.htm>

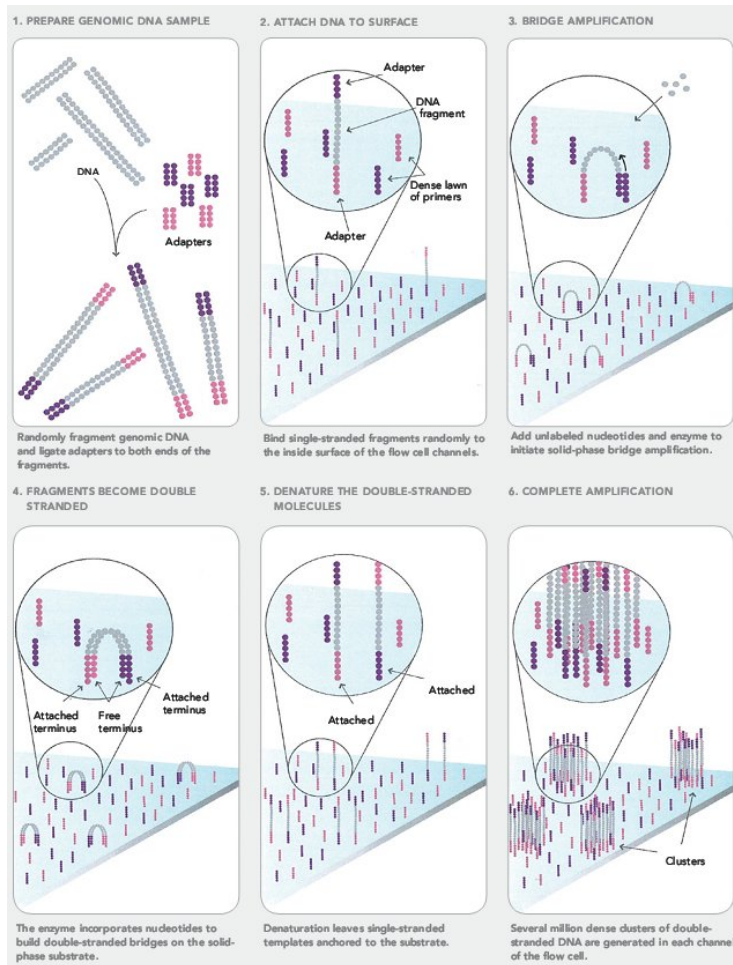


Pyrosequencing



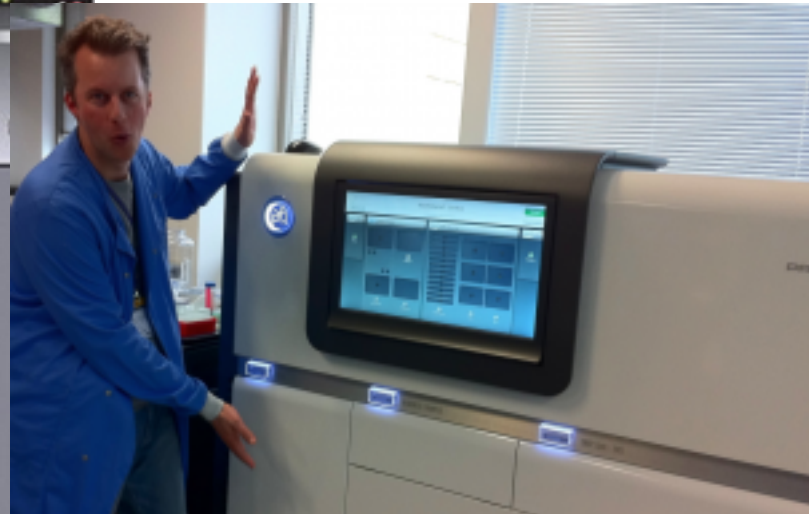
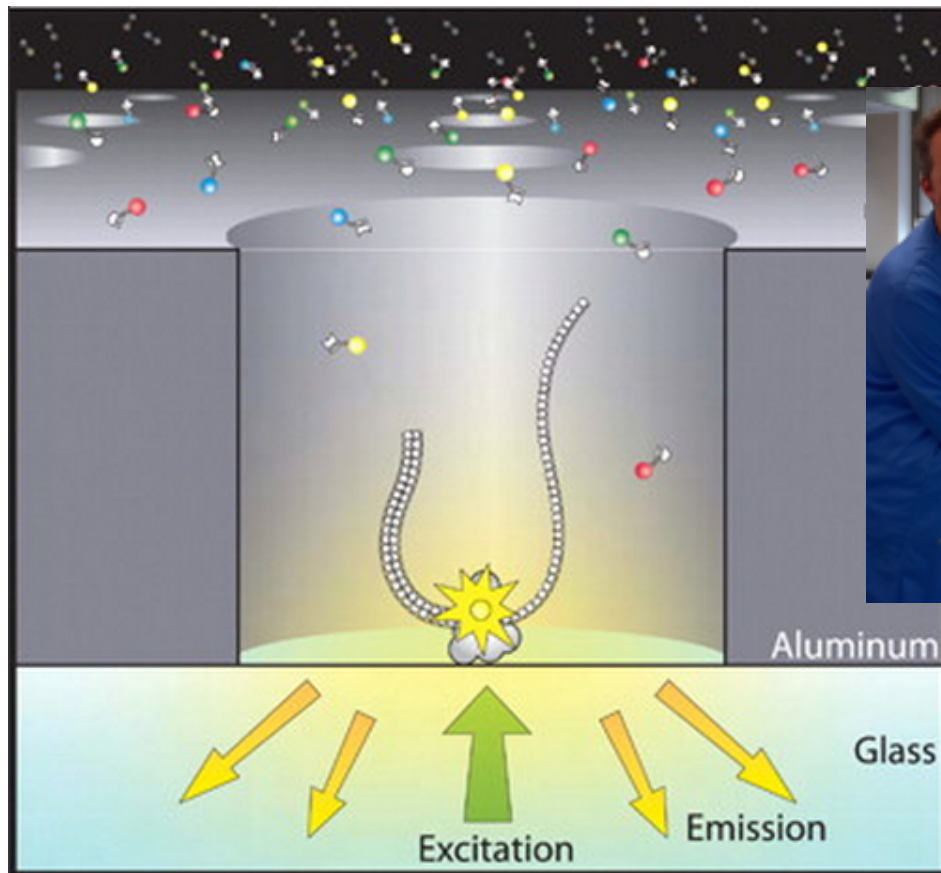
<https://www.youtube.com/watch?v=bFNjxKHP8Jc>

Illumina sequencing



<http://www.dnatube.com/video/27892/What-is-Illumina-Solexa-sequencing->

PacBio sequencing



<http://www.youtube.com/watch?v=NHCJ8PtYCFc>

Nanopore Technology



<https://www.nanoporetech.com/news/movies#movie-28-minion>

<https://www.nanoporetech.com/news/movies#movie-24-nanopore-dna-sequencing>

Sanger sequencing vs NextGen (3rd gen?)

- Read length
- Cost
- Error
- Sampling depth
- Bioinformatics

Comparing sequencing platforms in microbiome analysis

<u>Platform</u>	<u>Method</u>	<u>Reads</u>	<u>16S</u>	<u>Metagenome</u>	<u>Notes</u>
Sanger	Dideoxy terminator	750 bp	2-3 reads to cover	Good for database comparisons	Accurate, costly, slow
Pyroseq.	Light emission	400 bp			Good for 16S but not meta
Illumina	Flourescent step-by-step	100-150		More coverage makes up for short reads	High coverage, low cost
3 rd generation	Electronic signal	10-100 kb		Great for assembly	Unknown error, usability

How do you compare the composition of two microbial communities?

Diversity

- α diversity: taxa within a sample
- β diversity: between sample comparisons

UniFrac

