Metagenomics and Metabarcoding

Slides thank to Jenn DeBruyn

Associate Professor

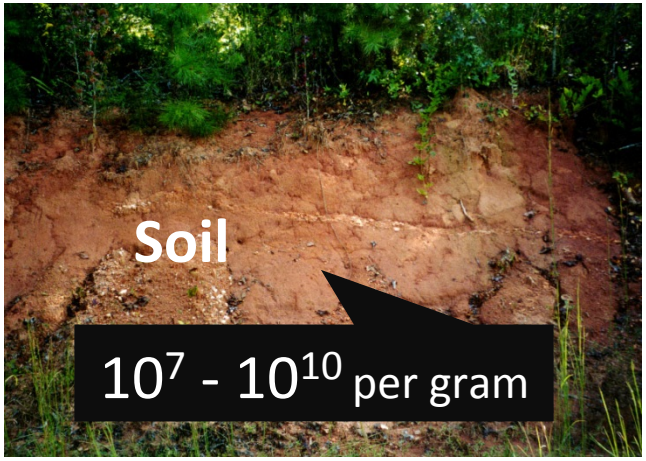
BIOSYSTEMS ENGINEERING & SOIL SCIENCE

UTIA INSTITUTE OF AGRICULTURE

THE UNIVERSITY OF TENNESSEE

Complex Microbial Communities

10^5 - 10^6 per ml



10^7 - 10^{10} per gram

10^{11} - 10^{12} per ml

10^9 - 10^{11} per ml

(Bio)diversity = variation in an ecosystem

Questions we may wish to answer by studying biodiversity:

1. Functional diversity: Microbes are the "gatekeepers" of ecosystem processes
   - Environment shaping the community (e.g. biogeography, host lifestyle, clinical interventions)
   - Community shaping the environment (e.g. biogeochemistry, pathogenesis)

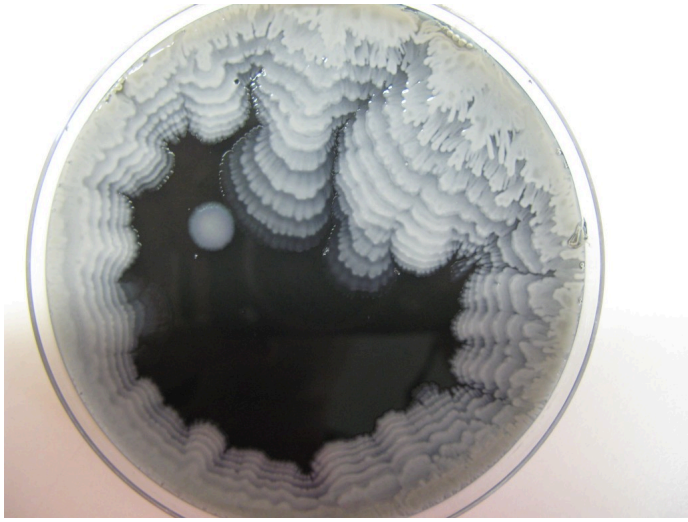
2. Genetic diversity: Possibility for novel genes/ enzymes/ compounds of commercial value

   Examples?

3. Organismal diversity: Understanding the nature of "life"

Why use sequencing to assess microbial biodiversity?

- For many years, labs cultured bacterial and other microbes to assess what microbes lived in an environment
- We now know that the majority of microbial life cannot be cultured in a lab (<20%*)
- We were missing huge amounts of information!
  - Who is there?
  - What are they doing?
- mid1980s - established sequencing as the primary culture-free method of profiling microbial communities



*Ward et al 1990

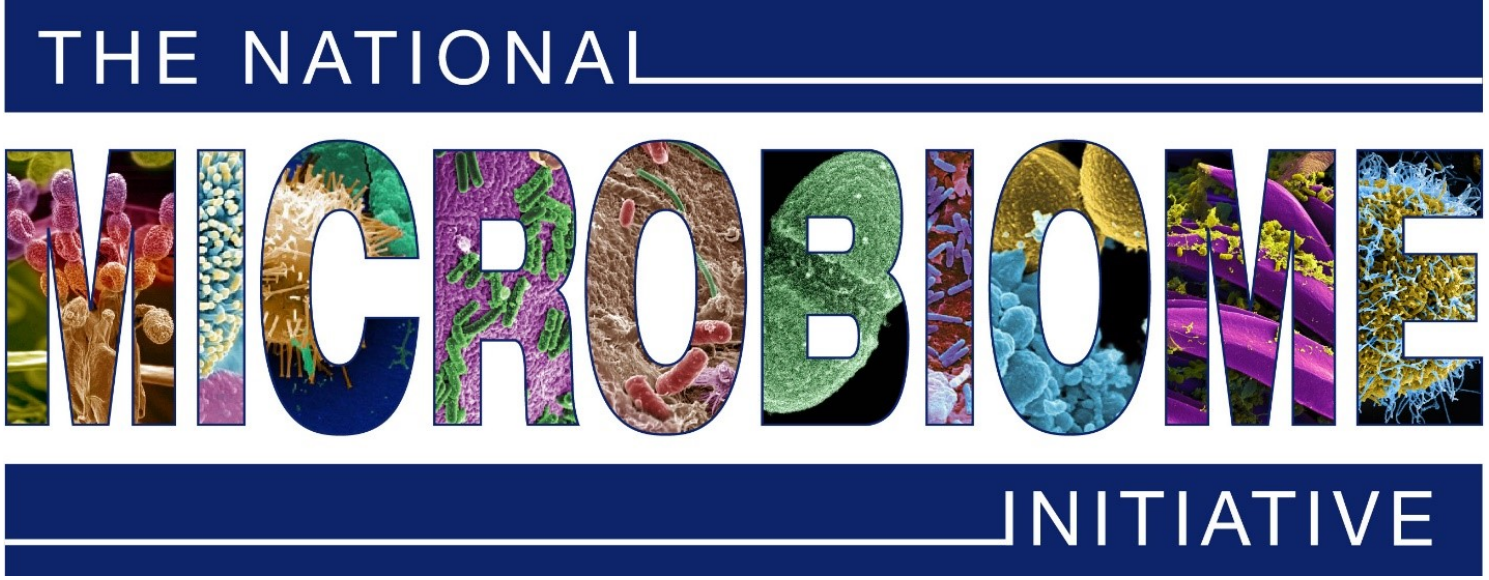
NIH HUMAN MICROBIOME PROJECT

"The NIH Common Fund Human Microbiome Project (HMP) was established in 2008, with the mission of generating resources that would enable the comprehensive characterization of the human microbiome and analysis of its role in human health and disease."

- HMP has characterized the microbial communities found at several different sites on the human body: nasal passages, oral cavity, skin, gastrointestinal tract, and urogenital tract

Goals:

- Development of a reference set of 3,000 isolate microbial genome sequences
- Initial 16S & mWGS metagenomic studies to generate an estimate of the complexity of the microbial community at each body site, providing initial answers to the questions of whether there is a "core" microbiome at each site
- Demonstration projects to determine the relationship between disease and changes in the human microbiome
- Development of new tools and technologies for computational analysis, establishment of a data analysis and coordinating center (DACC), and resource repositories
- Examination of the ethical, legal and social implications (ELSI) to be considered in the study and application of the metagenomic analysis of the human microbiota



- "The NMI aims to advance understanding of microbiome behavior and enable protection and restoration of healthy microbiome function."
- Goals:
  - **Supporting interdisciplinary research** to answer fundamental questions about microbiomes in diverse ecosystems.
  - **Developing platform technologies** that will generate insights and help share knowledge of microbiomes in diverse ecosystems and enhance access to microbiome data.
  - **Expanding the microbiome workforce** through citizen science, public engagement, and educational opportunities.
- Federal agency investment of $121M for FY2016 and 2017

Approaches

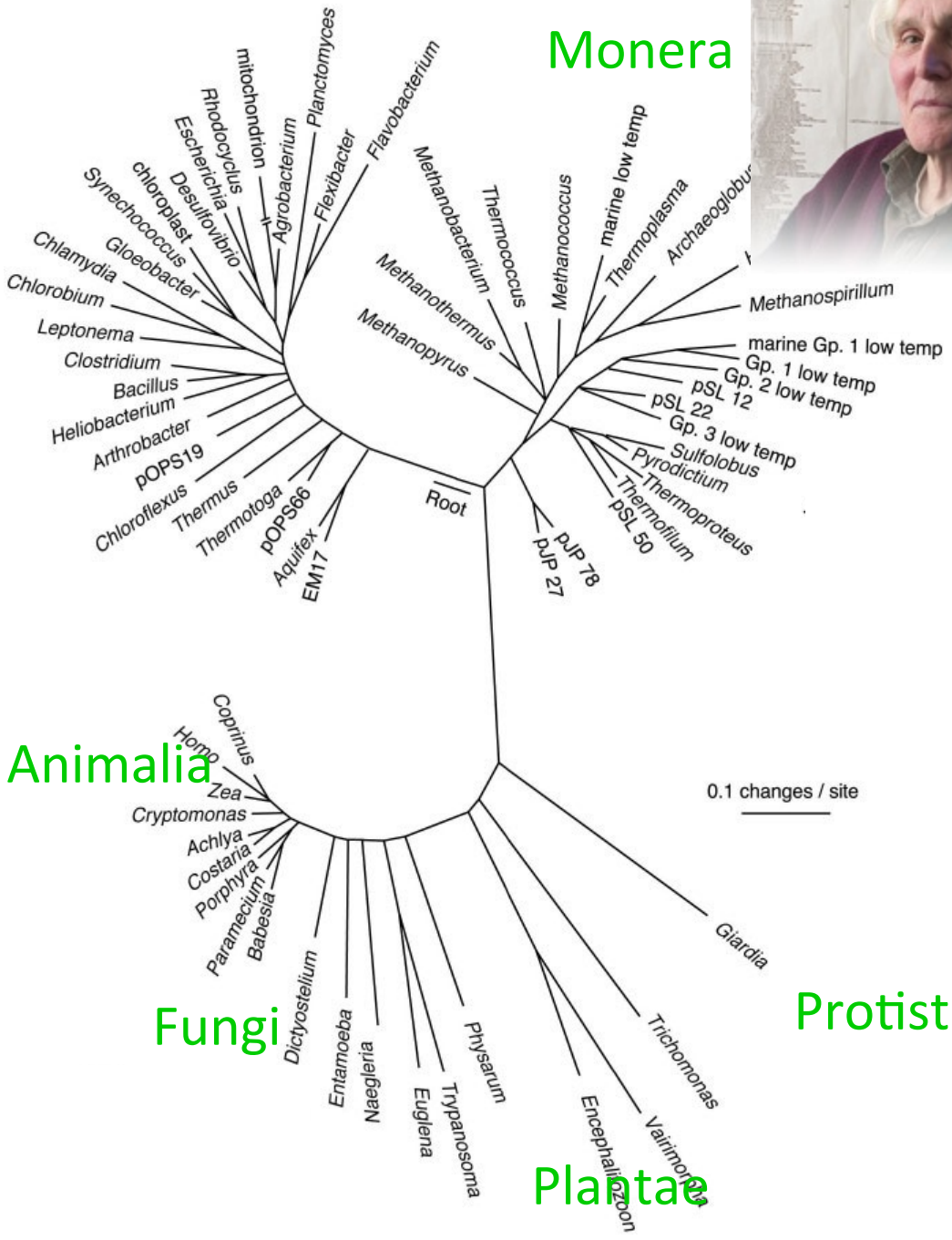
- Targeted metagenomics
  - Synonyms: DNA Profiling, DNA Barcoding, Metabarcoding, Amplicon Sequencing
  - Focus on a single gene or set of genes for sequencing
  - rRNA (16S, 18S), ITS
- Metagenomics
  - Sequence all DNA
- Metatranscriptomics
  - Sequence all RNA

TARGETED METAGENOMES: AMPLICON LIBRARIES

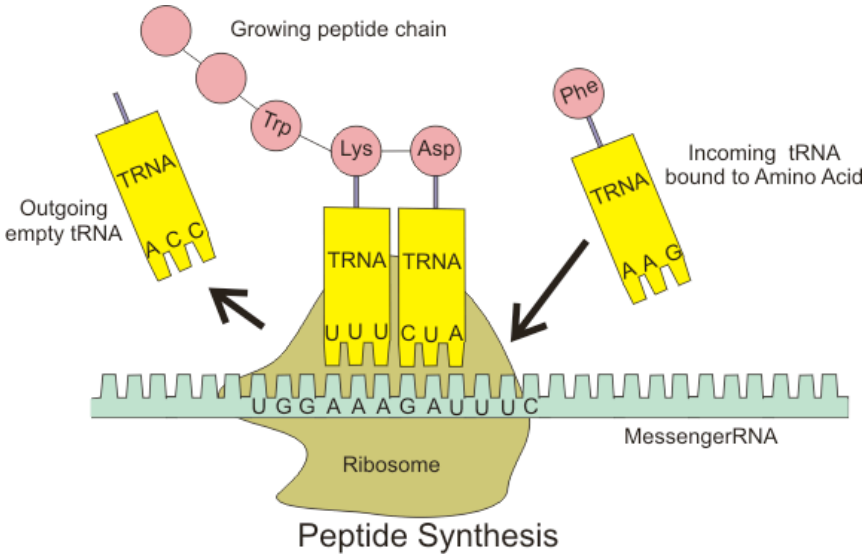
Molecular phylogeny

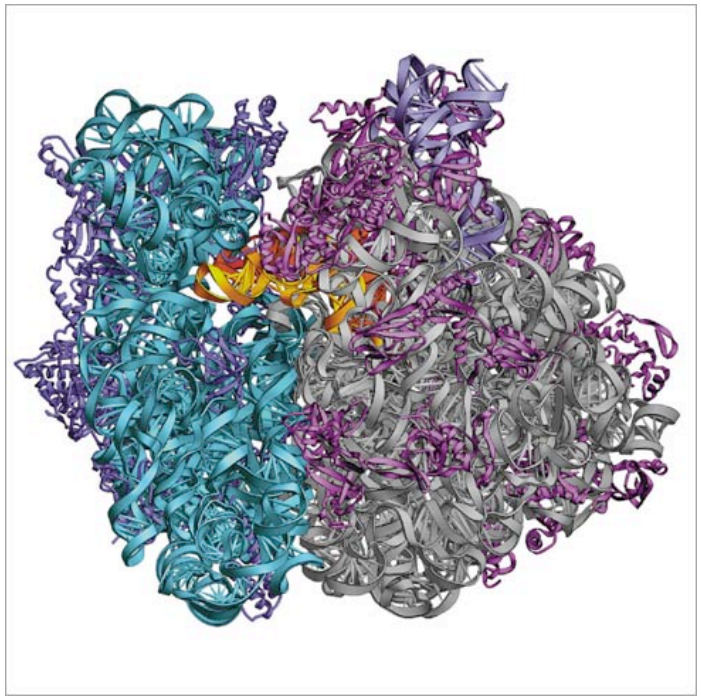
- Molecular phylogeny uses the **sequence of a common gene** to determine how related two organisms are
  - Genes used for comparison = **phylogenetic markers**
  - Organisms that are more closely related have more similar gene sequences

- 1977: Carl Woese revolutionized the field of phylogenetics by introducing the first phylogeny based on ribosomal RNA (small subunit – 16S or 18S)
  - Revealed great diversity in the prokaryotes
  - Identification of Archaea
  - Move from a 5 kingdom to 3 domain system of biological classification



Phylogenetic Marker: Ribosomal RNA

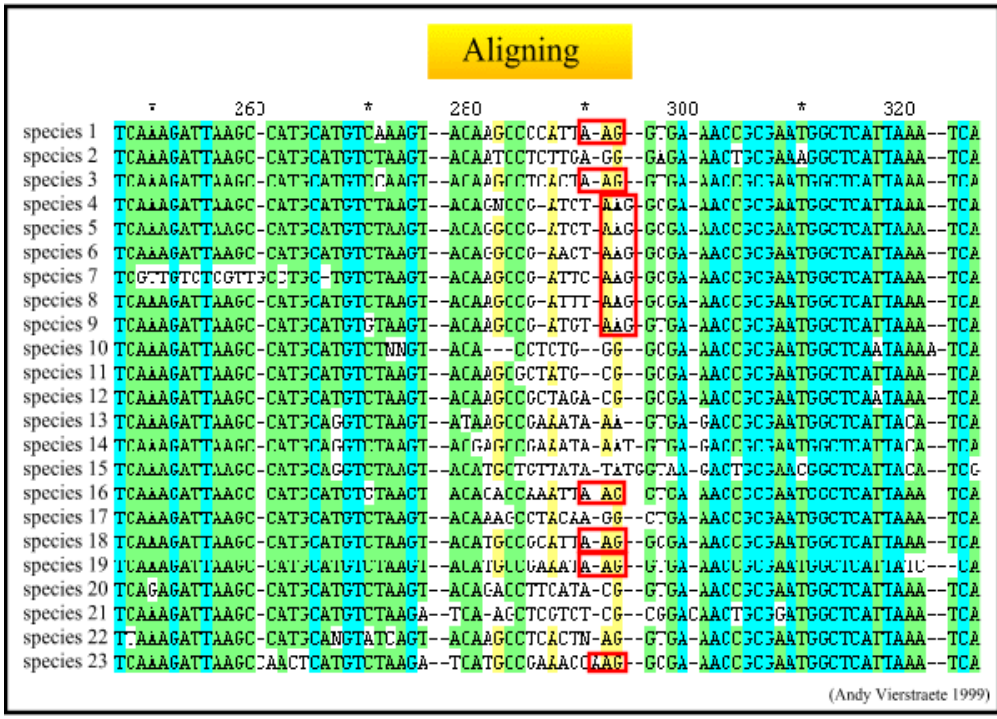


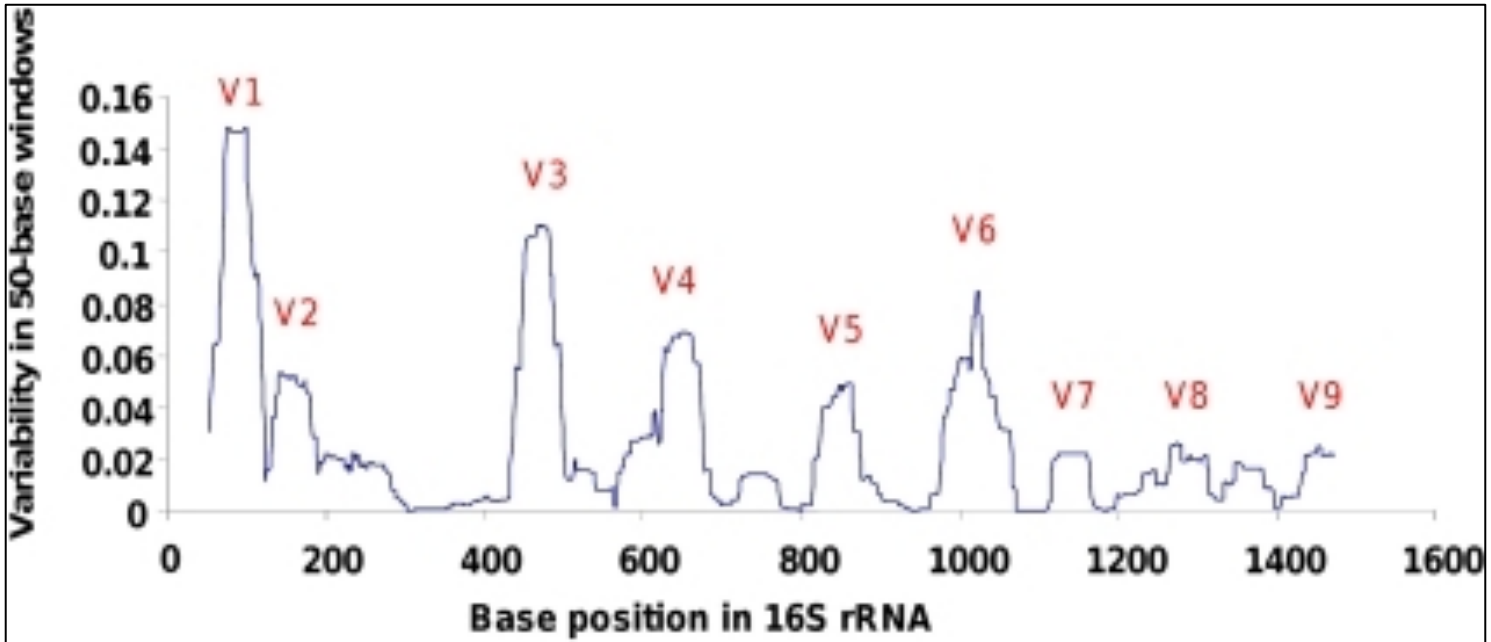


- Ribosomes are an essential part of translating DNA code to proteins (cell building blocks) and consist of protein and RNA components
  - ALL organisms have ribosomes
  - Highly conserved
- Most common phylogenetic marker: Small subunit ribosomal RNA gene
  - Prokaryotes (bacteria and archaea): 16S
  - Eukaryotes: 18S

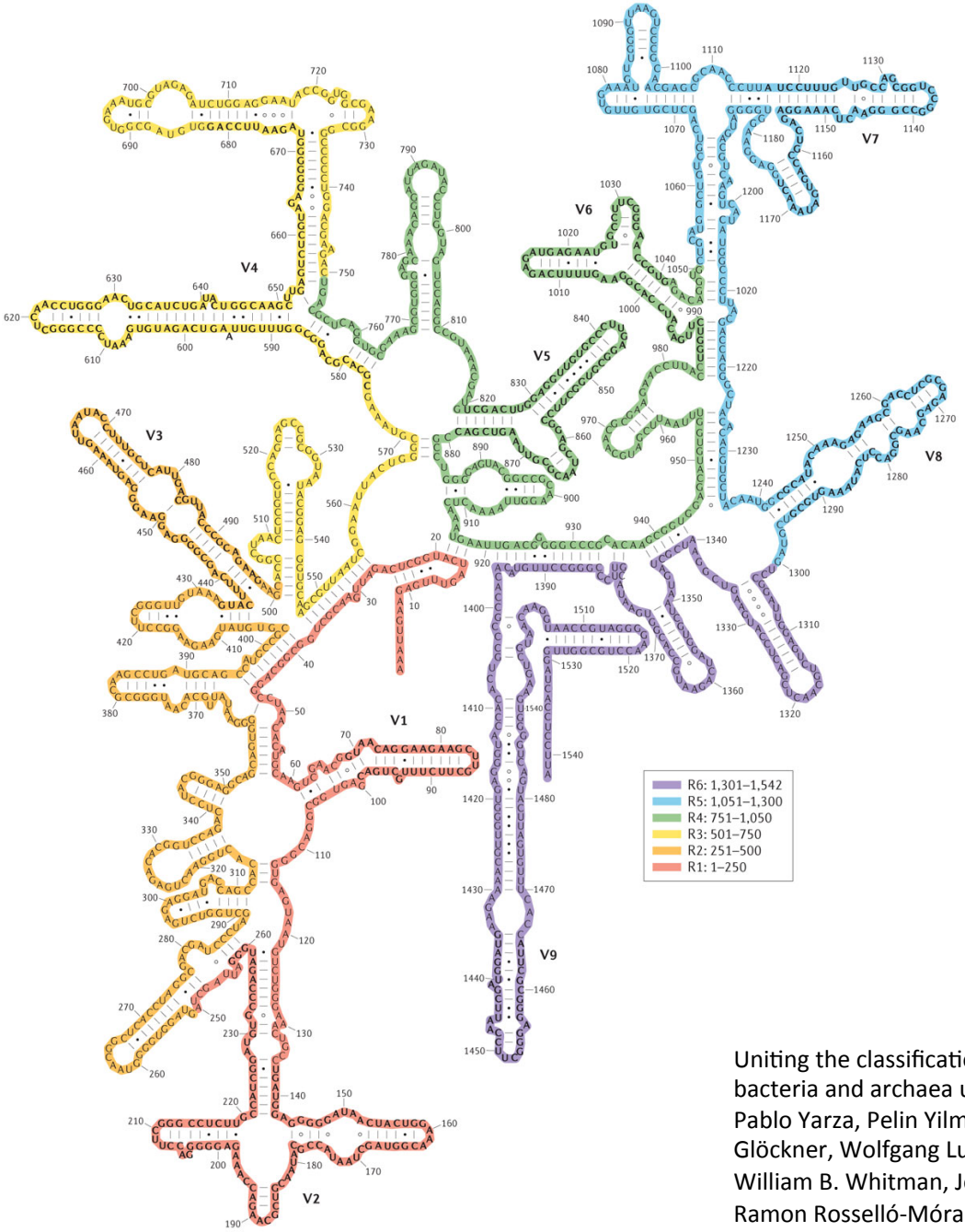
16S ribosomal RNA gene

- Conserved enough to design "universal" PCR primers that are used to amplify the 16S gene from any organism (i.e. allows us to "find" the gene in every organism)
- Variable regions allow us to distinguish evolutionary relationships
  - Assumption based on **neutral theory of evolution**





Bodilis et al 2012



Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Pablo Yarza, Pelin Yilmaz, Elmar Pruesse, Frank Oliver Glöckner, Wolfgang Ludwig, Karl-Heinz Schleifer, William B. Whitman, Jean Euzéby, Rudolf Amann & Ramon Rosselló-Móra. **Nature Reviews Microbiology 12, 635–645 (2014)** doi:10.1038/nrmicro3330

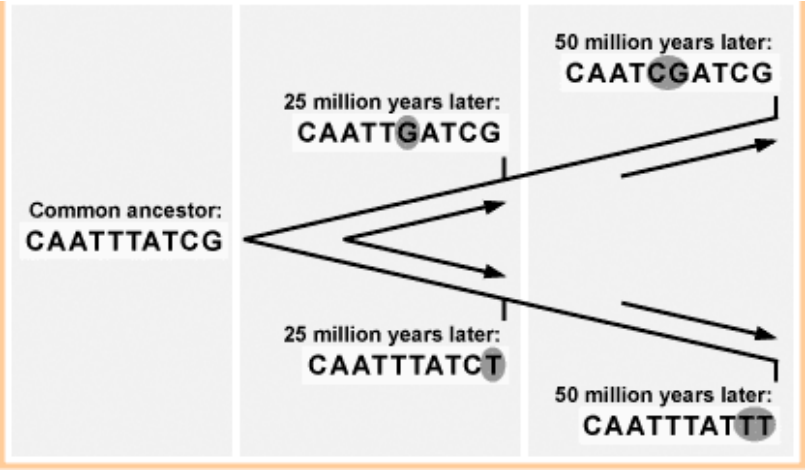
Neutral theory of evolution

Neutral theory of evolution

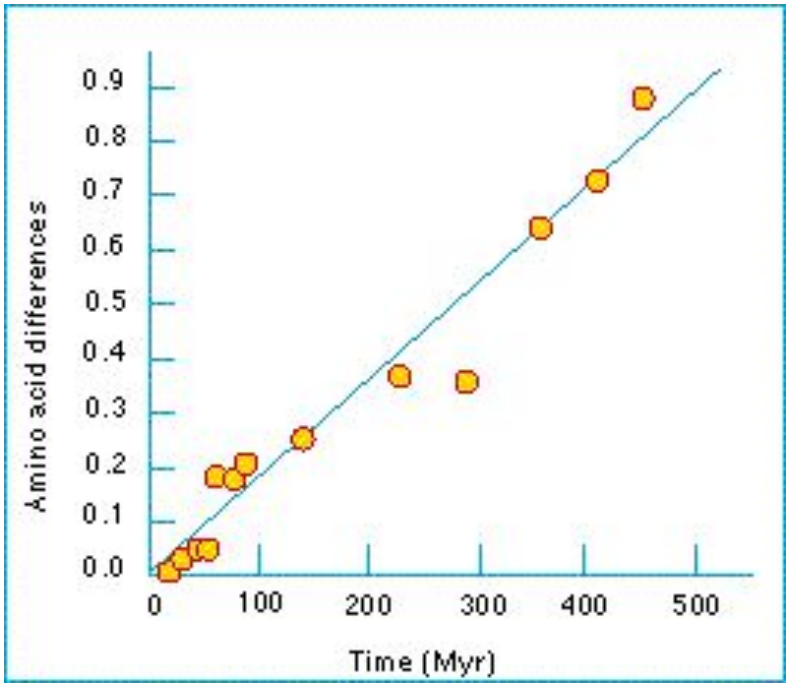
-Organisms randomly pick up nucleotide mutations (changes) due to copying errors

-These mutations don't help or hurt the organism, so there is no selection on them (neutral)

-Because these mutations are not "weeded out" by natural selection, they accumulate over time at a fairly regular rate.



Molecular Clock

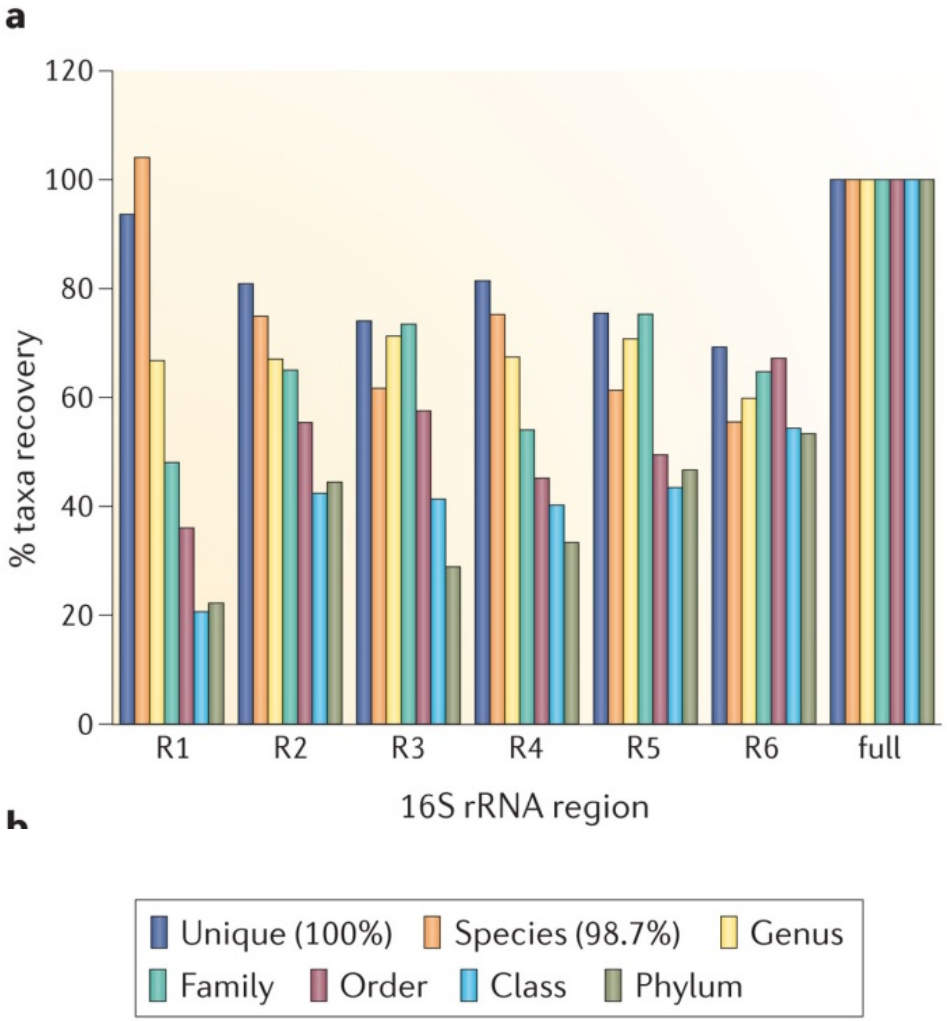


- the regular acquisition of mutations means that the number of nucleotide differences approximates evolutionary time
- Organisms that have are more closely related (have a recent common ancestor) will have fewer nucleotide differences compared to organisms that diverged earlier.

"Molecular Clock"

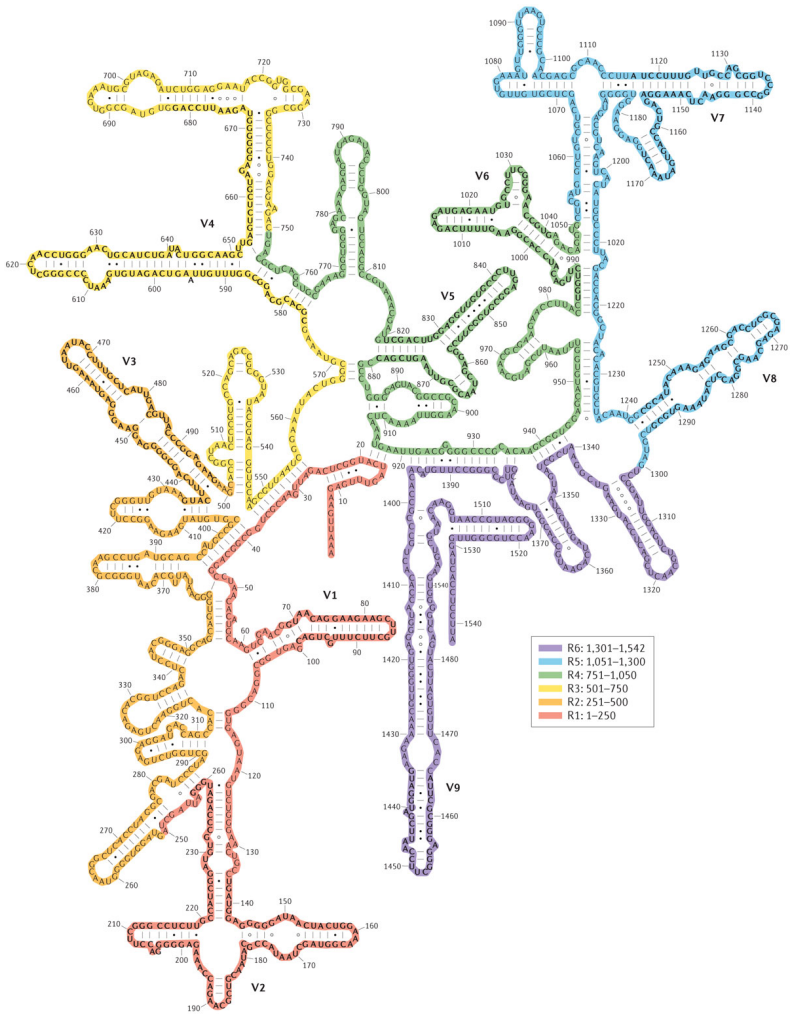
Taxonomic identification with 16S profiling

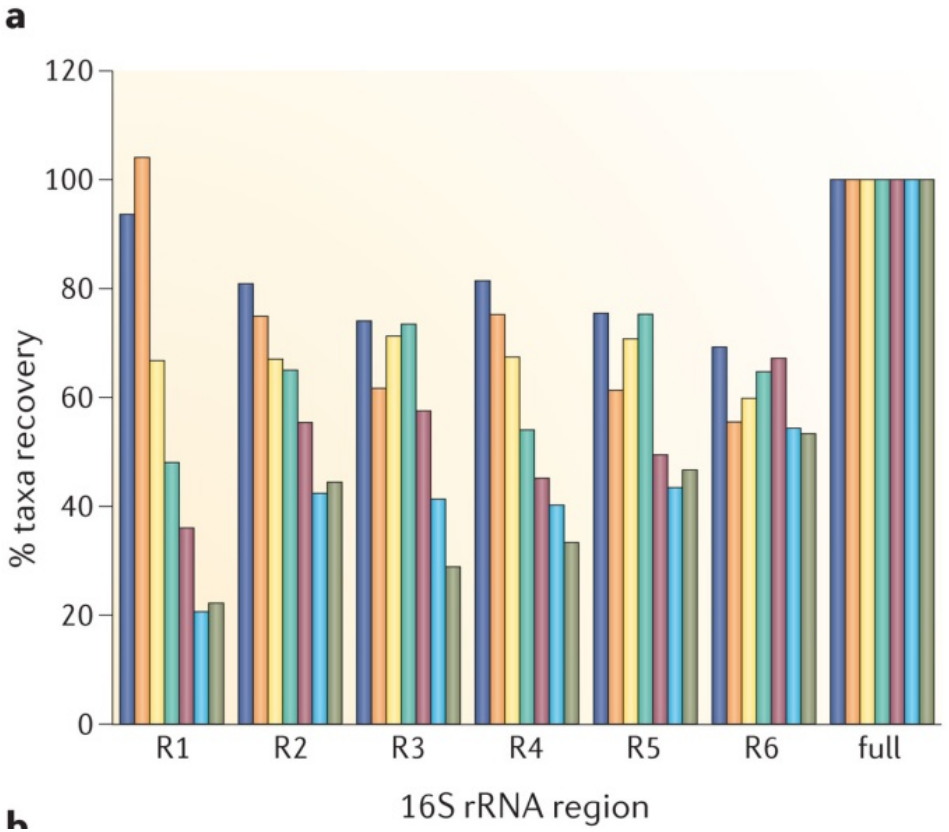
- Short reads don't capture enough variation to assign reads to specific taxa
- Some labs are experimenting with PacBio and MinIon technologies – longer reads will help



Nature Reviews | Microbiology

Taxonomic identification with 16S profiling



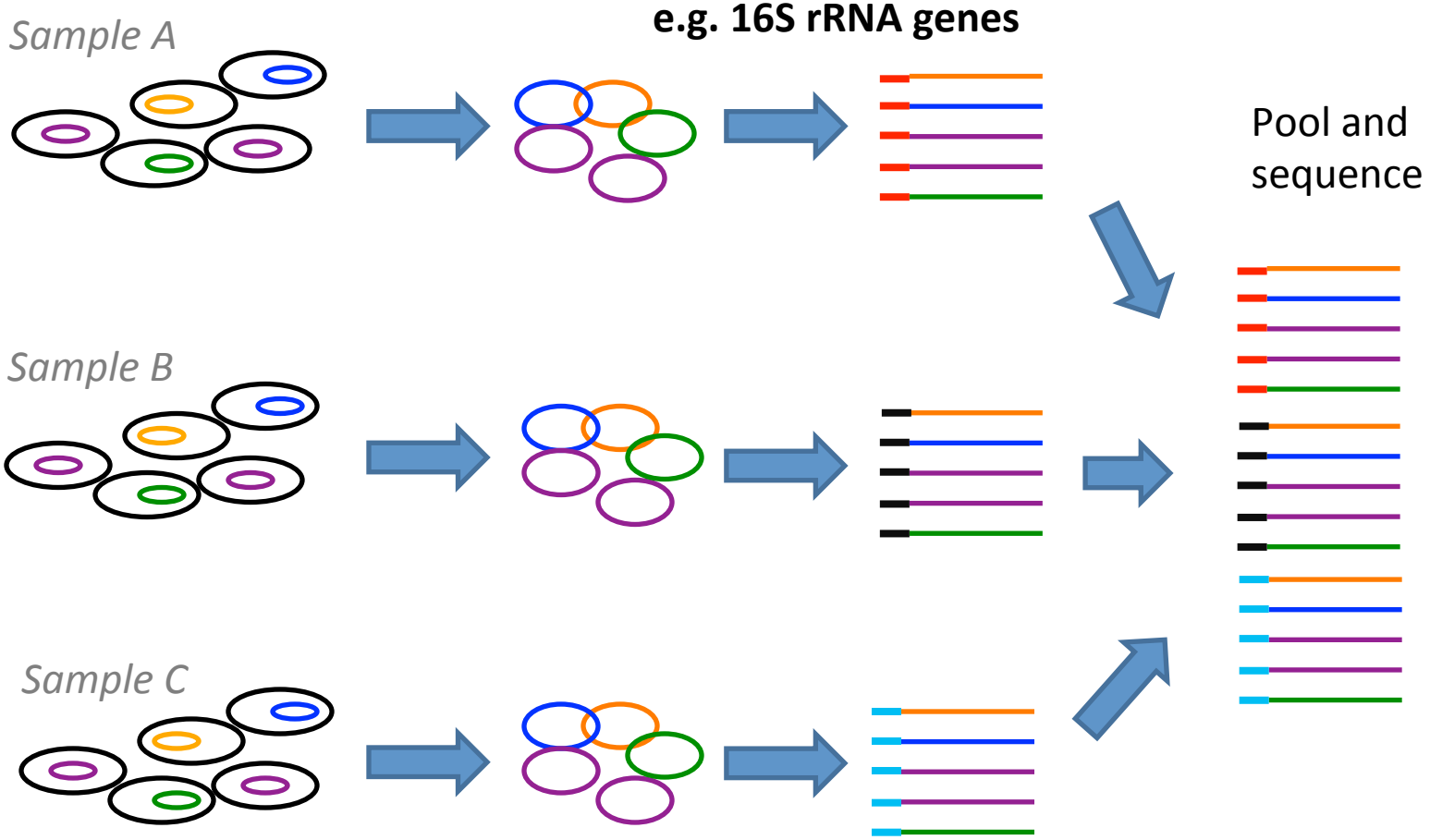


Nature Reviews | Microbiology

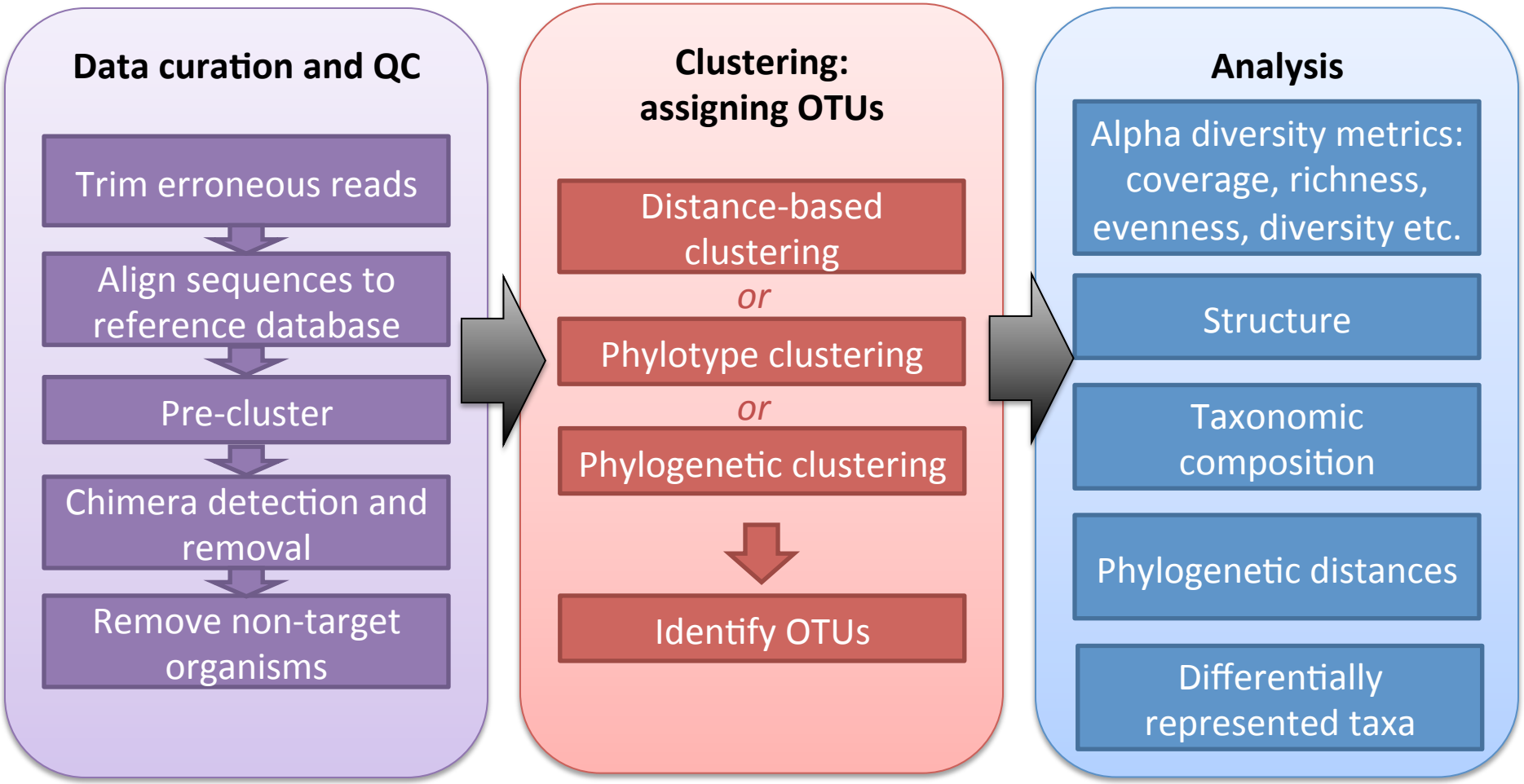
Amplicon Libraries: Library preparation

Extract all DNA from samples

Use universal (conserved) primers with barcodes to PCR amplify target gene: **e.g. 16S rRNA genes**

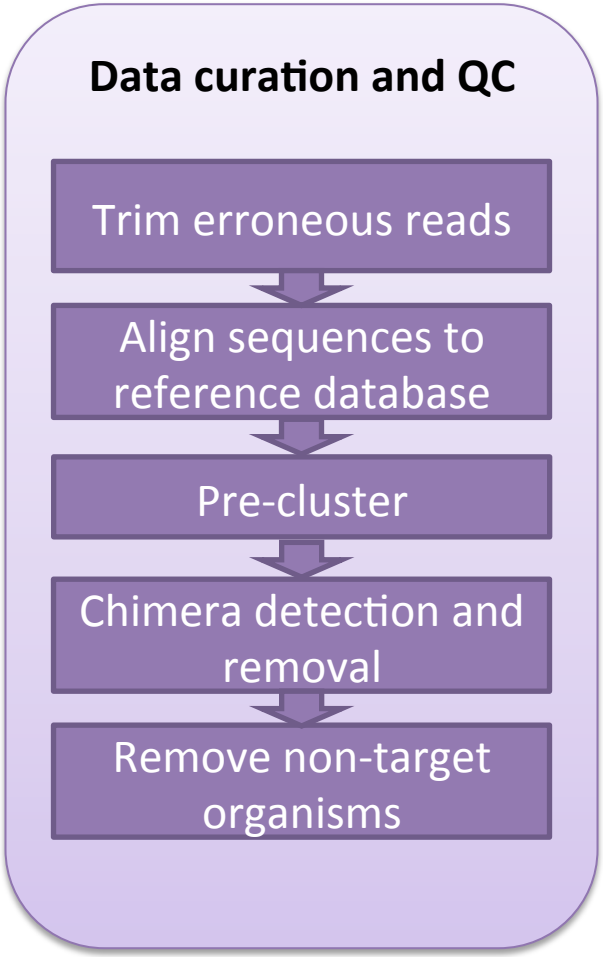


Amplicon libraries: data processing and analysis



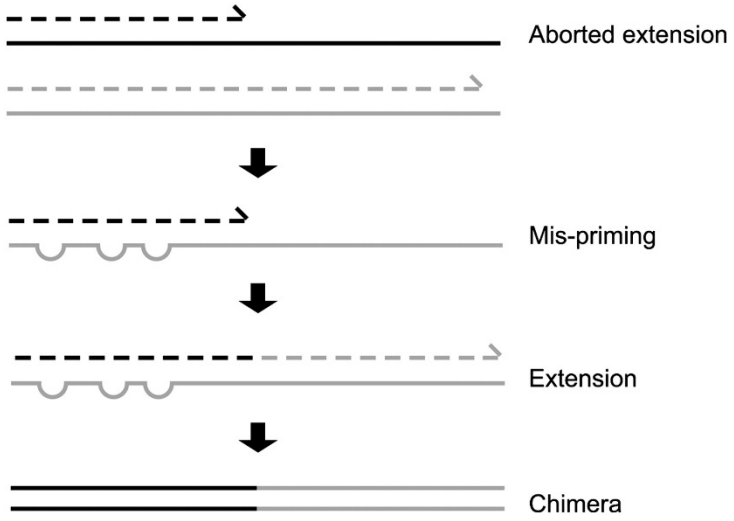
Amplicon libraries: data processing and analysis

1. Data curation and QC



Data curation and QC

- Primary source of error: PCR
  - Polymerase mistakes
    - Taq polymerase: 10^{-5}
    - High fidelity Taq: 10^{-7}
    - For 400 bp, 30 cycles, this means:
      - Regular Taq: 27% of products have errors
      - HiFi Taq: 0.5% of products have errors
  - Bias and draft:
    - GC% - GC rich runs can be compressed; fast ramp times don't allow enough time to completely denature
  - Chimera formation

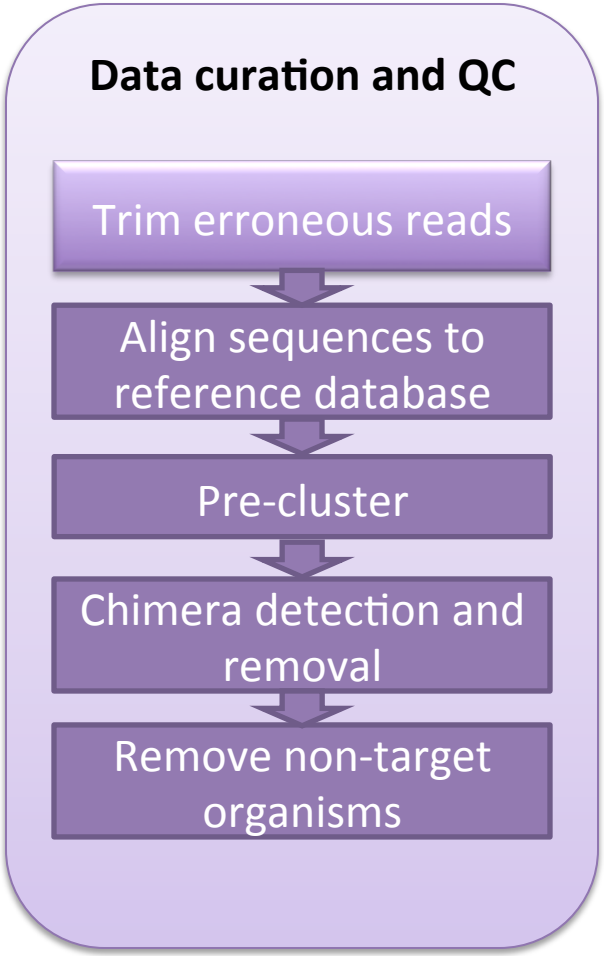


Two curation pipeline options:

MOTHUR (Schloss et al. 2009)

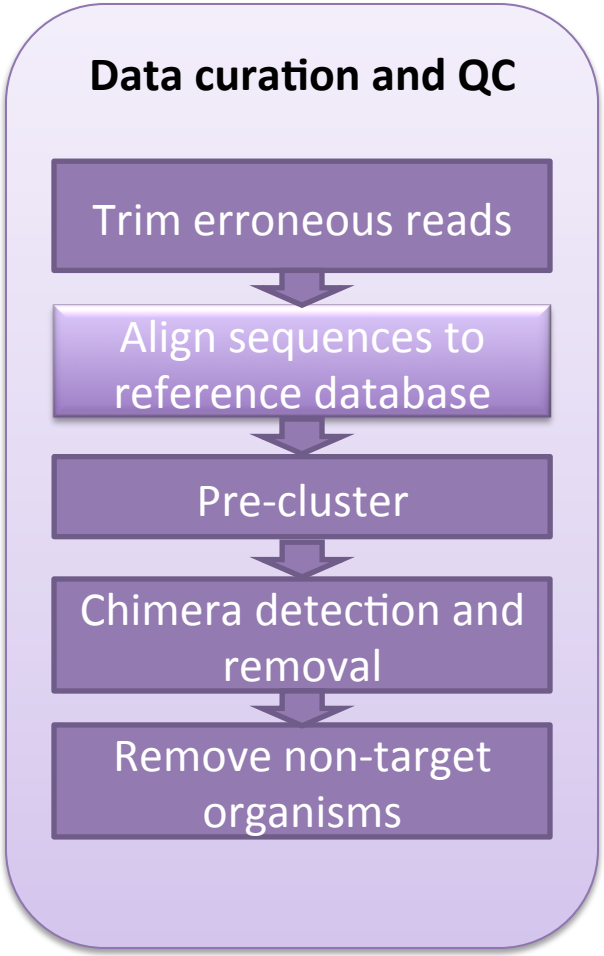
QIIME (Caporaso et al. 2010)

Trim out obviously erroneous reads



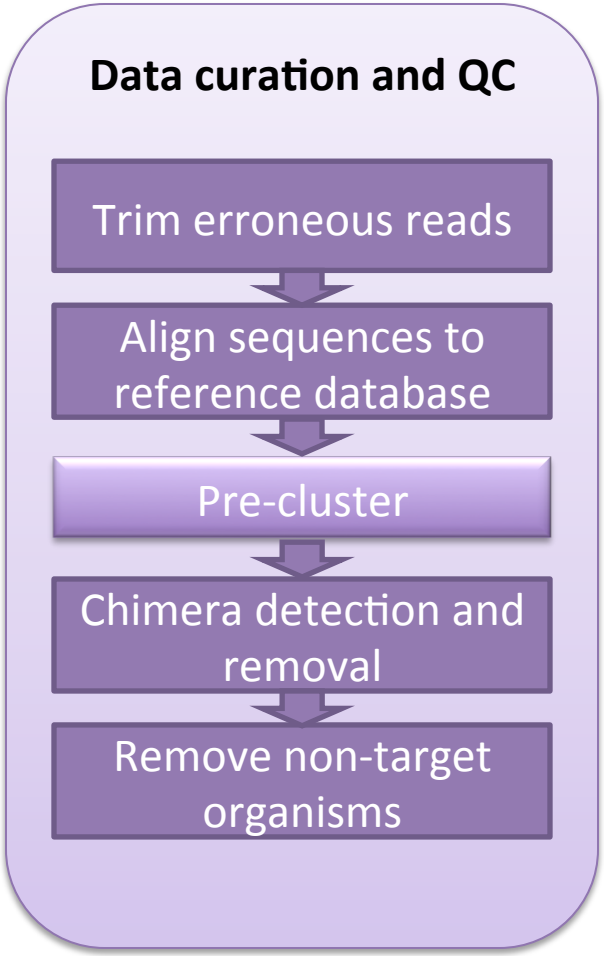
- Remove sequences that have
  - Unambiguous bases (N)
  - >8 homopolymers
  - Too long or too short
  - Any mismatches to primer or barcode
  - Drops errors from 0.5% to 0.2%

Alignment

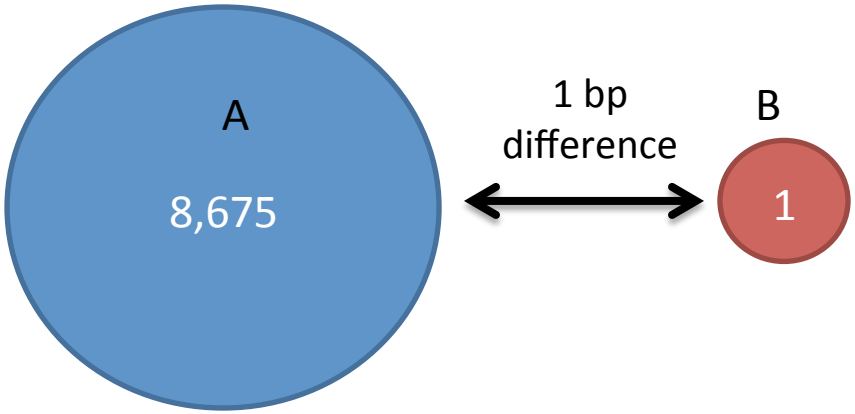


- Align your sequences to reference alignment
- Database options:
  - Ribosomal Database Project (RDP)
    - Populated from GenBank
    - Has variety of tools for classifying and aligning
    - Contains both short and long reads
    - Follows Bergy's Taxonomy
  - Greengenes
  - Silva/ARB
- Classification options:
  - BLAST
  - Kmers
- Following alignment, remove any sequences that didn't align properly (they are likely erroneous)

Pre-cluster

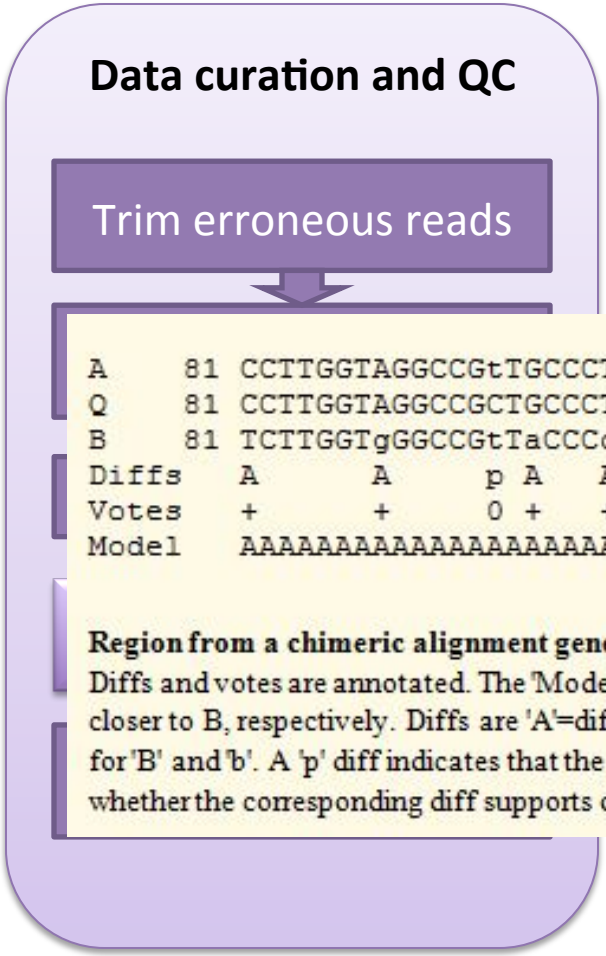


- "Ironing out the wrinkles" (Huse 2007)
  - More abundant sequences are more likely to have sequence errors
  - Ranks sequences by abundance, then finds sequences that are 1 bp different and lumps them together (this is more likely a sequence error than a new taxon)
  - Drops error rate from 0.2% to 0.04%



Chimera detection and removal

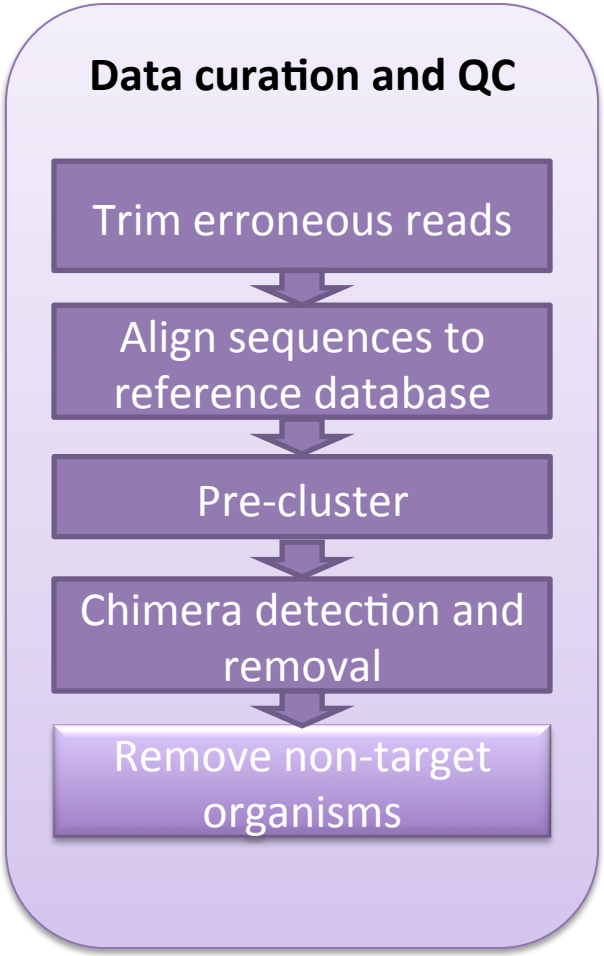
- Chimera detection algorithm: UCHIME (Edgar et al. 2011)
  - Database dependent or independent




```
A     81 CCTTGGTAGGCCGtTGCCCTGCCAACTAGCTAATCAGACGCgggtCCATCtcaCACCaccggAgtTTTtcTCaCTgTacc 160
Q     81 CCTTGGTAGGCCGCTGCCCTGCCAACTAGCTAATCAGACGCATCCCCATCCATCACCGATAAATCTTTAATCTCTTTCAG 160
B     81 TCTTGGTgGGCCGtTaCCCcGCCAACaAGCTAATCAGACGCATCCCCATCCATCACCGATAAATCTTTAAaCTCTTTCAG 160
Diffs    A      A     p A   A      A              BBBB     BBB    BBBBB BB   BBa B   B BBB
Votes    +      +     0 +   +      +              ++++     +++    +++++ ++   ++! +   + +++
Model    AAAAAAAAAAAAAAAAAAAAAAAAAAxxxxxxxxxxxxxxxBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
```

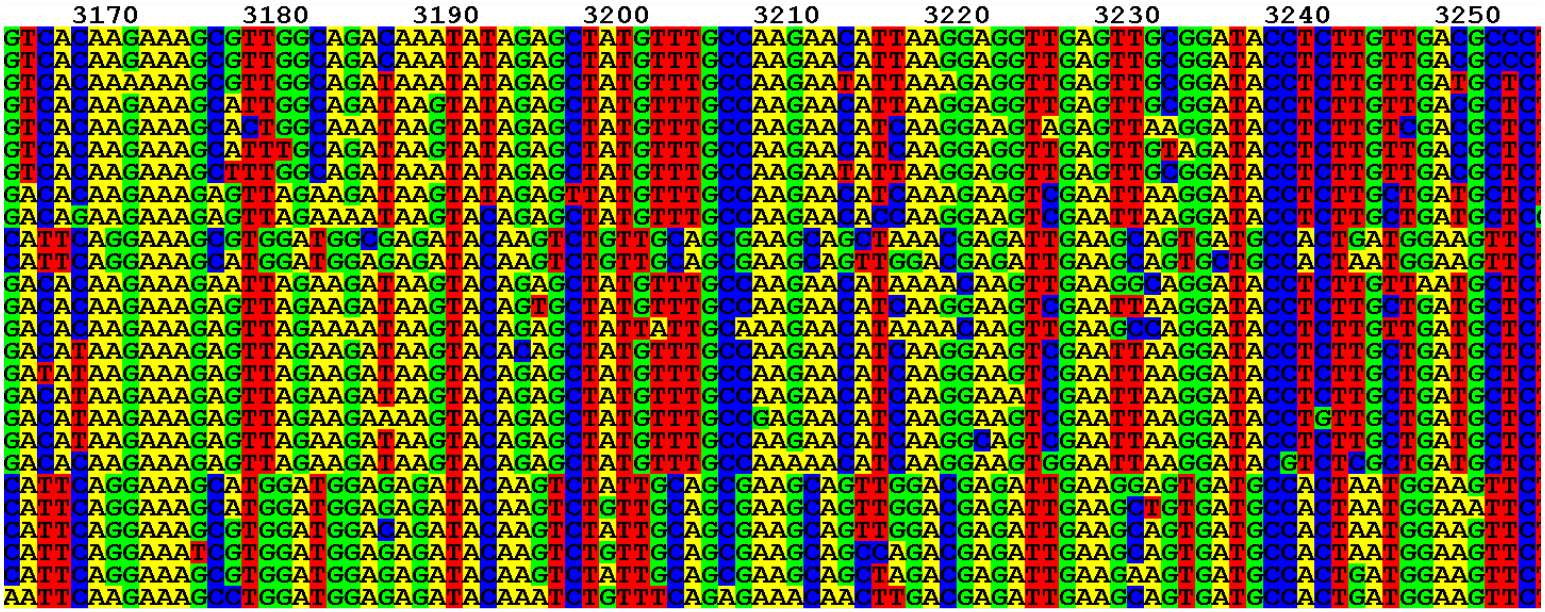
Region from a chimeric alignment generated by UCHIME.
Diffs and votes are annotated. The 'Model' row indicates the three segments of the alignment which are closer to A, the crossover (X), and closer to B, respectively. Diffs are 'A'=diff with Q closer to A in the A segment, 'a'=diff with Q closer to A in the B segment, and similarly for 'B' and 'b'. A 'p' diff indicates that the parents agree but are different from Q. Votes are '+' (yes), '!' (no) and '0' (abstain), indicating whether the corresponding diff supports or contradicts the model.

Remove non-targets



- Taxonomically classify sequences, and remove anything that's not what you meant to get
- E.g. if doing a 16S library, remove anything classifying as:
  - Eukaryota
  - Chloroplast
  - Mitochondria
  - Unknown
  - Archaea (maybe)

Result: Set of clean, aligned, trimmed sequences

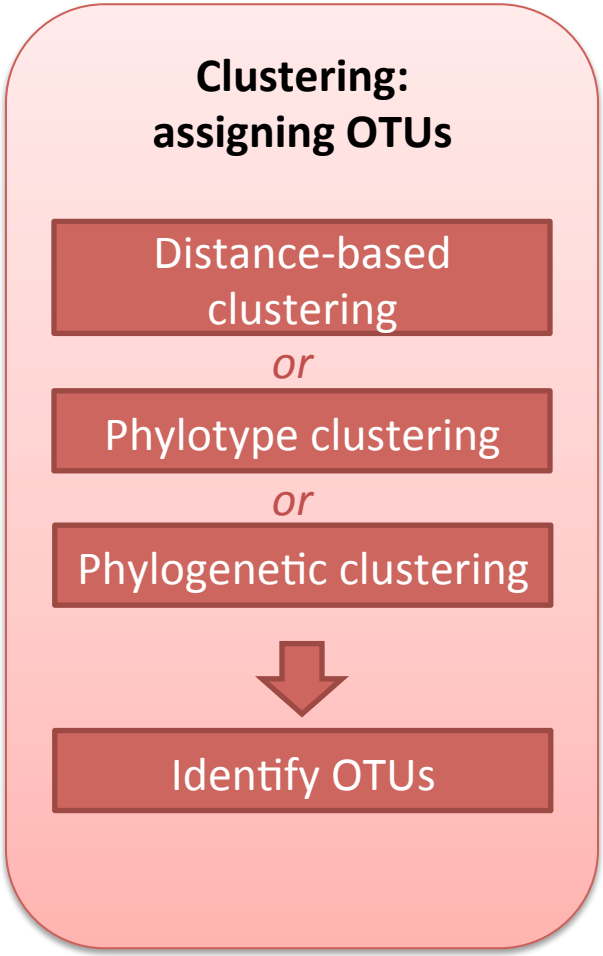


Next step:

Turn sequences into something biologically meaningful

Amplicon libraries: data processing and analysis

2. Clustering: Assigning OTUs



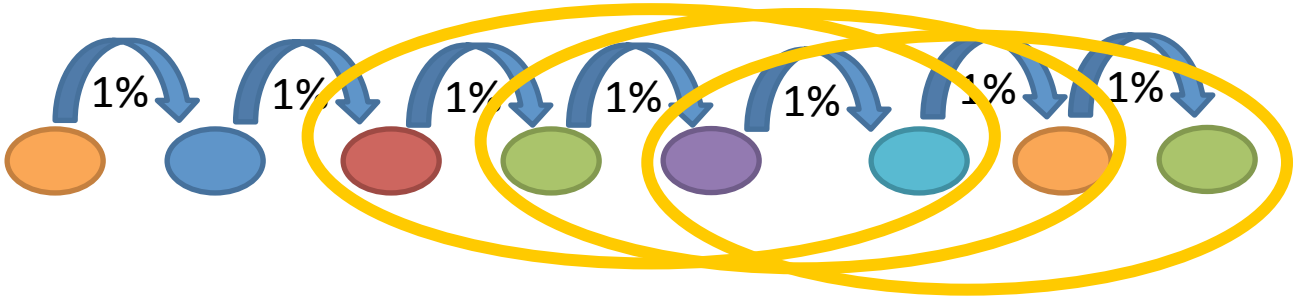
OTUs: Operational Taxonomic Units

- Biological species concept doesn't work well for prokaryotes
- OTU definitions rely on nucleotide distances – usually 97% or 98%
- BUT: because of fast generation times, and rapid mutation rates and divergence, asexual reproduction, prokaryotes represent the whole spectrum of similarity

While we refer to them as a taxonomic unit, may or may not actually behave biologically in a similar manner

OTU Decisions:

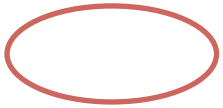
- Method for clustering
- To use a database of existing clusters or not?
- Cut-off of sequence similarity



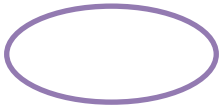
Distance-Based Clustering Methods

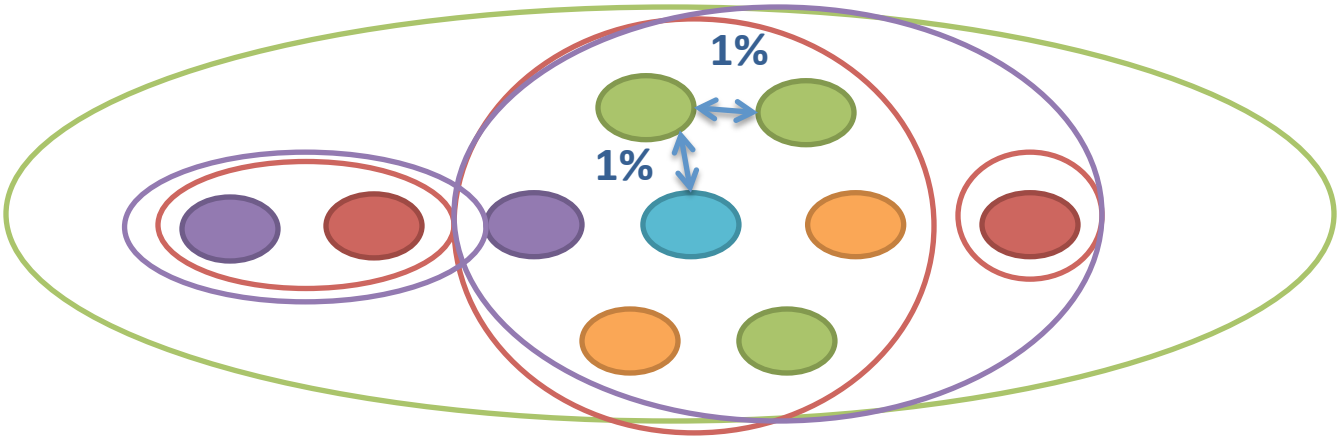
- Nearest neighbor/single linkage – each sequence is within 3% of at least one other sequence in the cluster
- Furthest neighbor/complete linkage – no two sequences are further than 3% apart
- Average neighbor – average distance from a sequence to every other sequence is less than 3%.

NN can underestimate # of OTUs



FN can overestimate # of OTUs





Distance methods

Sequence 1	A	T	T	C	A	G	C
Sequence 2	A	T	C	C	A	G	C
Sequence 3	A	T	C	C	G	C	
Sequence 4	A	T	C	C	G	T	C

Create a consensus alignment representing minimal change

Consensus alignment

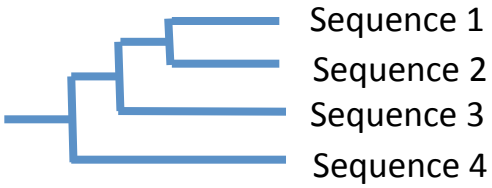
Sequence 1	ATTCAG–C
Sequence 2	ATCCAG–C
Sequence 3	ATCC–G–C
Sequence 4	ATCC–GTC

Convert differences to distance matrix

Nucleotide distance matrix

	Seq1	Seq2	Seq3	Seq4
Sequence 1	**0**			
Sequence 2	**1**	**0**		
Sequence 3	**2**	**2**	**0**	
Sequence 4	**3**	**2**	**1**	**0**

Use clustering algorithm to represent distances as a **dendogram**



Final results?

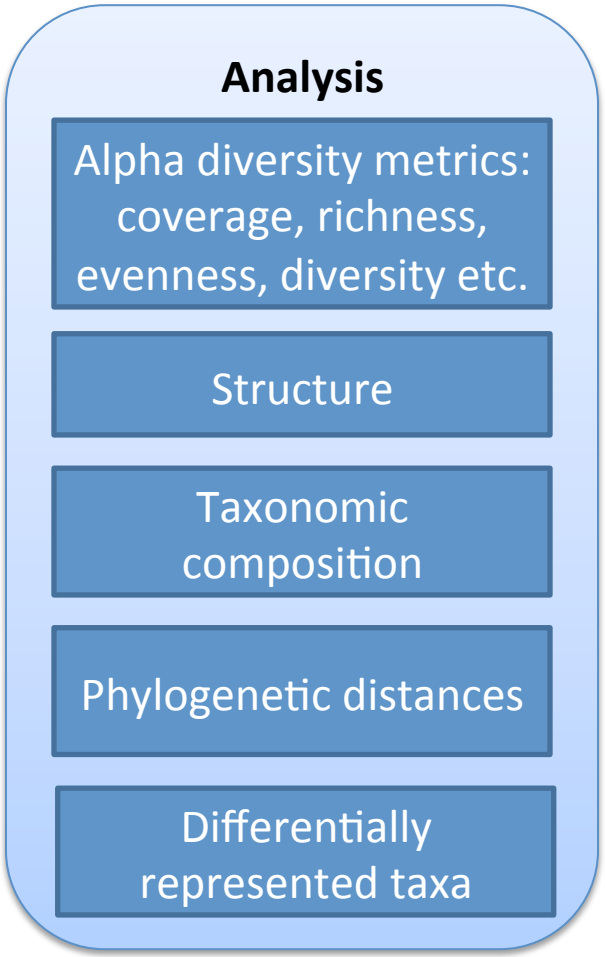
OTU table (aka species count table)…

Group	numOtus	Otu001	Otu002	Otu003	Otu004	Otu005	Otu006	Otu007	Otu008	Otu009	Otu010	Otu011	Otu012	Otu013	Otu014	Otu015	Otu016	Otu017	Otu018	Otu019	Otu020
12AA0031	839	29	0	77	0	96	2989	18333	125	12122	319	1154	879	597	356	91	565	94	194	20	62
12AA0073	839	10	0	344	0	9	497	398	38	2449	709	28	489	242	79	176	105	26	239	0	138
12AA0075	839	2	1	126	18	89	329	891	95	267	720	354	1818	296	898	73	9	3	26	0	119
12AA0106	839	7	0	228	2	79	223	1604	356	299	35	164	1800	374	3003	65	177	16	120	14	141
12AA0171	839	2012	0	151	0	39	679	516	14	805	101	101	125	107	12	204	38	152	3	0	329
12AA0215	839	0	0	566	1	166	1662	4120	0	1767	20	1283	3005	2260	2276	322	4	29	6	0	728
12AA0242	839	50292	1	423	9	7	1268	701	0	16	48	19	0	18	80	465	125	14	2	34	628
12AA0345	839	0	1	533	13	87	921	1176	151	1647	186	1993	3339	1331	3490	311	62	17	37	19	493
12AA0362	839	2	0	93	0	50	280	1367	3	604	37	293	875	1112	201	197	312	11	5232	0	221
12AA0416	839	23	0	587	1	97	1028	5596	43	2323	562	512	235	265	840	45	747	19	147	20	155
12AA0418	839	20746	2	256	5	49	276	958	36	1422	409	104	417	419	167	100	106	4	815	8	97
12AA0423	839	3292	5	49	1	7	260	8	5	866	126	0	0	52	25	84	0	4	0	5	78
12AA0430	839	1	6	26	0	9	254	6	13	839	345	0	0	20	27	77	0	3	0	9	101
12AA0571	839	16322	1	665	0	47	1978	225	328	33	1722	5	10	93	39	528	0	25	0	856	857
12AA0840	839	31	2	83	0	43	1867	1194	129	314	1145	760	1037	398	730	410	386	68	3548	183	396
12AA0849	839	609	0	127	7	309	946	4952	30	646	139	917	800	730	2688	275	286	76	578	58	315
12AA0851	839	13	0	604	3	62	3225	124	27	72	2737	123	141	62	169	114	9	1342	19	11	107
12AA0852	839	3928	0	17	0	177	278	1883	26	227	297	284	481	196	1456	114	25	7	124	0	131
12AA0863	839	9	1	47	0	98	187	1083	66	416	2710	430	444	212	841	62	36	1	236	5	80
12AA0866	839	1	0	23	0	12	253	564	23	1494	99	93	466	174	106	106	14	6	161	17	101
12AA0882	839	83	0	138	0	38	2044	3472	42	1212	244	637	503	187	342	205	218	84	736	54	232
12AA0922	839	16160	1	54	0	248	302	1143	285	1577	8581	43	32	53	30	72	20	6	24	720	225
12AA0924	839	2695	1	35	0	139	376	3151	340	1050	15	210	663	210	494	147	39	6	22	1161	434
12AA0926	839	76	0	654	124	34	2222	2908	1007	207	249	696	880	527	1405	426	106	33	476	2564	1195
12AM0001	839	17	7799	905	182	111	14023	957	795	574	98	11	182	458	492	5825	26	213	111	2913	5289
12AM0002	839	1	11286	140	0	305	2649	12246	603	13497	1961	510	484	2155	34	1390	11	32	67	439	2072
12AM0003	839	46839	477	1	0	31	193	832	243	506	674	41	2	69	85	102	3	3	2	267	94
12AM0004	839	0	5824	11	0	148	911	11160	253	4017	199	1052	2865	1628	289	391	4	28	74	846	279
12AM0005	839	0	29084	30	0	2436	180	21	268	136	1074	30	0	15	11	54	5	62	4	487	120

…and a taxonomy table

OTU	Size	Taxonomy					
Otu001	2139018	Bacteria(100)	Proteobacteria(100)	Alphaproteobacteria(100)	Rickettsiales(100)	Rickettsiaceae(100)	Rickettsia(100)
Otu002	1528159	Bacteria(100)	Proteobacteria(100)	Gammaproteobacteria(100)	Thiotrichales(100)	Francisellaceae(100)	Francisella(100)
Otu003	336263	Bacteria(100)	Proteobacteria(100)	Gammaproteobacteria(100)	Gammaproteobacteria_unclassified(100)	Gammaproteobacteria_unclassified(100)	Gammaproteobacteria_unclassified(100)
Otu004	506329	Bacteria(100)	Chlamydiae(100)	Chlamydiae(100)	Chlamydiales(100)	Chlamydiales_unclassified(100)	Chlamydiales_unclassified(100)
Otu005	49051	Bacteria(100)	Proteobacteria(100)	Proteobacteria_unclassified(100)	Proteobacteria_unclassified(100)	Proteobacteria_unclassified(100)	Proteobacteria_unclassified(100)
Otu006	386031	Bacteria(100)	Proteobacteria(100)	Gammaproteobacteria(100)	Pseudomonadales(100)	Pseudomonadaceae(100)	Pseudomonas(100)
Otu007	605590	Bacteria(100)	Proteobacteria(100)	Alphaproteobacteria(100)	Sphingomonadales(100)	Sphingomonadaceae(100)	Sphingomonas(100)
Otu008	269133	Bacteria(100)	Proteobacteria(100)	Gammaproteobacteria(100)	Pseudomonadales(100)	Moraxellaceae(100)	Acinetobacter(100)
Otu009	409180	Bacteria(100)	Proteobacteria(100)	Alphaproteobacteria(100)	Rhizobiales(100)	Methylobacteriaceae(100)	Methylobacterium(100)
Otu010	173363	Bacteria(100)	Proteobacteria(100)	Gammaproteobacteria(100)	Xanthomonadales(100)	Xanthomonadaceae(100)	Xanthomonadaceae_unclassified(100)
Otu011	100995	Bacteria(100)	Proteobacteria(100)	Alphaproteobacteria(100)	Sphingomonadales(100)	Sphingomonadaceae(100)	Sphingomonadaceae_unclassified(100)
Otu012	144756	Bacteria(100)	Planctomycetes(100)	Planctomycetacia(100)	Planctomycetales(100)	Planctomycetaceae(100)	Singulisphaera(100)
Otu013	154956	Bacteria(100)	Proteobacteria(100)	Alphaproteobacteria(100)	Rhizobiales(100)	Rhizobiales_unclassified(100)	Rhizobiales_unclassified(100)
Otu014	78968	Bacteria(100)	Bacteria_unclassified(100)	Bacteria_unclassified(100)	Bacteria_unclassified(100)	Bacteria_unclassified(100)	Bacteria_unclassified(100)
Otu015	119103	Bacteria(100)	Proteobacteria(100)	Betaproteobacteria(100)	Burkholderiales(100)	Comamonadaceae(100)	Delftia(100)
Otu016	68815	Bacteria(100)	Proteobacteria(100)	Betaproteobacteria(100)	Burkholderiales(100)	Oxalobacteraceae(100)	Oxalobacteraceae_unclassified(100)
Otu017	11264	Bacteria(100)	Proteobacteria(100)	Gammaproteobacteria(100)	Pseudomonadales(100)	Pseudomonadaceae(100)	Pseudomonadaceae_unclassified(100)
Otu018	117154	Bacteria(100)	Proteobacteria(100)	Betaproteobacteria(100)	Burkholderiales(100)	Oxalobacteraceae(100)	Massilia(100)
Otu019	132649	Bacteria(100)	Bacteroidetes(100)	Flavobacteria(100)	Flavobacteriales(100)	Flavobacteriaceae(100)	Cloacibacterium(100)
Otu020	136666	Bacteria(100)	Bacteroidetes(100)	Flavobacteria(100)	Flavobacteriales(100)	Flavobacteriaceae(100)	Elizabethkingia(100)

3. Analysis (the FUN part!)



- OTU table is used as the basis for many different analyses
- Analysis approach depends on the question you are asking
- But there are some common themes…

Alpha vs. Beta Diversity

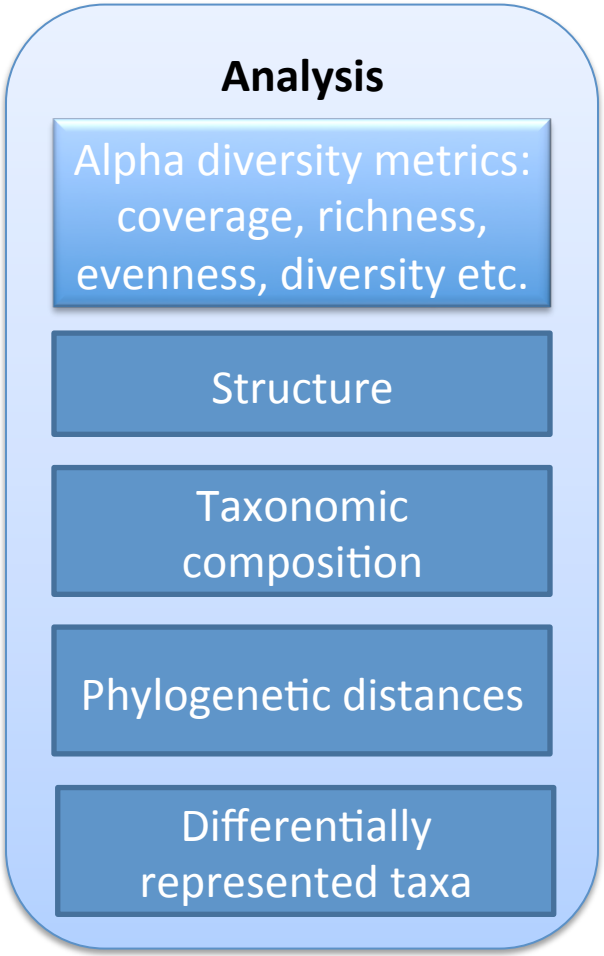
Alpha Diversity

- What is the phylogenetic relationship?
- How many types? (Richness)
- How are the types distributed? (Evenness)
- What is the diversity? (Richness + Evenness)
- What is the structure of the community? i.e. What types are there? (Assign taxonomy)

Beta Diversity

- How does one community compare to another?
- Are there geographical/temporal patterns to communities?
  - Biogeography
  - Successional patterns
- Are particular taxa of interest changing in abundance or diversity?

Alpha Diversity



Biodiversity takes into account:

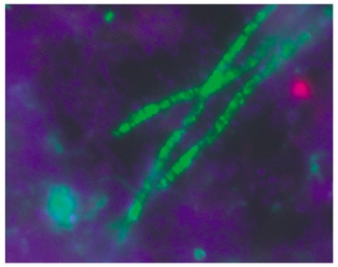
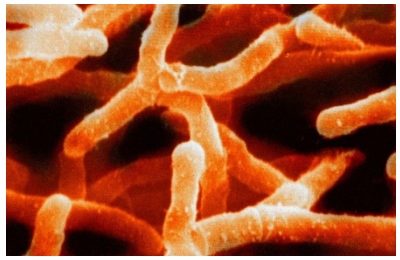
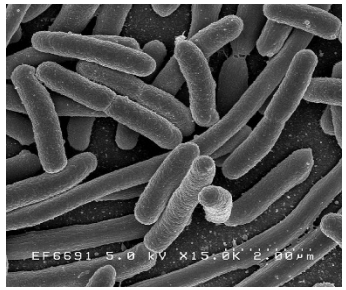
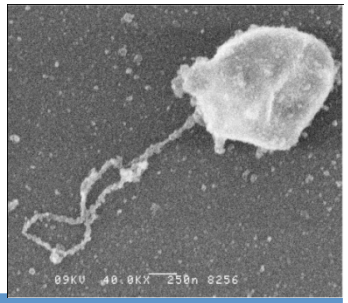
- Species richness: the number of species in a region or specified area
- Species evenness: the degree of equitability in the distribution of individuals among a group of species. Maximum evenness is the same number of individuals among all species.

Species	Number of Individuals
E. Redbud	1
Black Oak	9
Post Oak	2
White Pine	4
Honey Locust	5

Example

A soil ecologist goes out into the field and collects soil from two separate wooded plots with one big difference: Plot 1 has been treated with elevated CO_2 and Plot 2 is a control (ambient CO_2). The ecologist is interested in the types of soil microorganisms that are found in the plots and whether there is a difference between the two plots.

What will we find out?

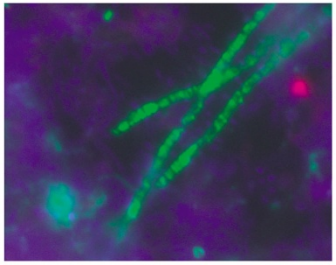





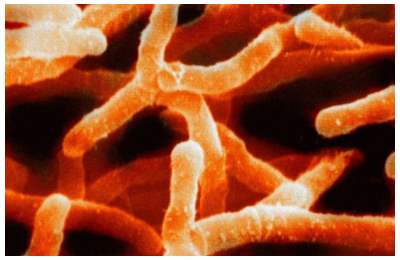
Phylum	Plot 1 Elevated CO_2	Plot 2 Ambient CO_2
Acidobacteria	50	10
Actinobacteria	36	50
Proteobacteria	35	0
Verrucomicrobia	55	39

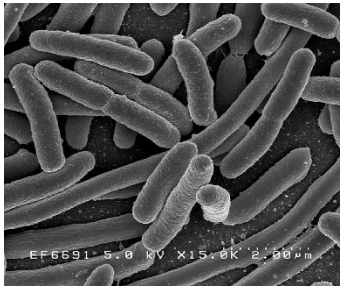
Which plot as more richness?
Which plot has more evenness?
Which plot has more biodiversity?

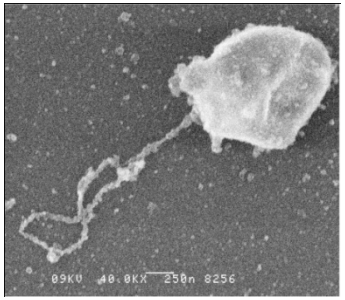
What if your data looked like this?

Phylum	Plot 1 Elevated CO_2	Plot 2 Ambient CO_2
Acidobacteria	50	1
Actinobacteria	36	1
Proteobacteria	35	30
Verrucomicrobia	55	39
Planctomycetes	0	40









Diversity Metrics

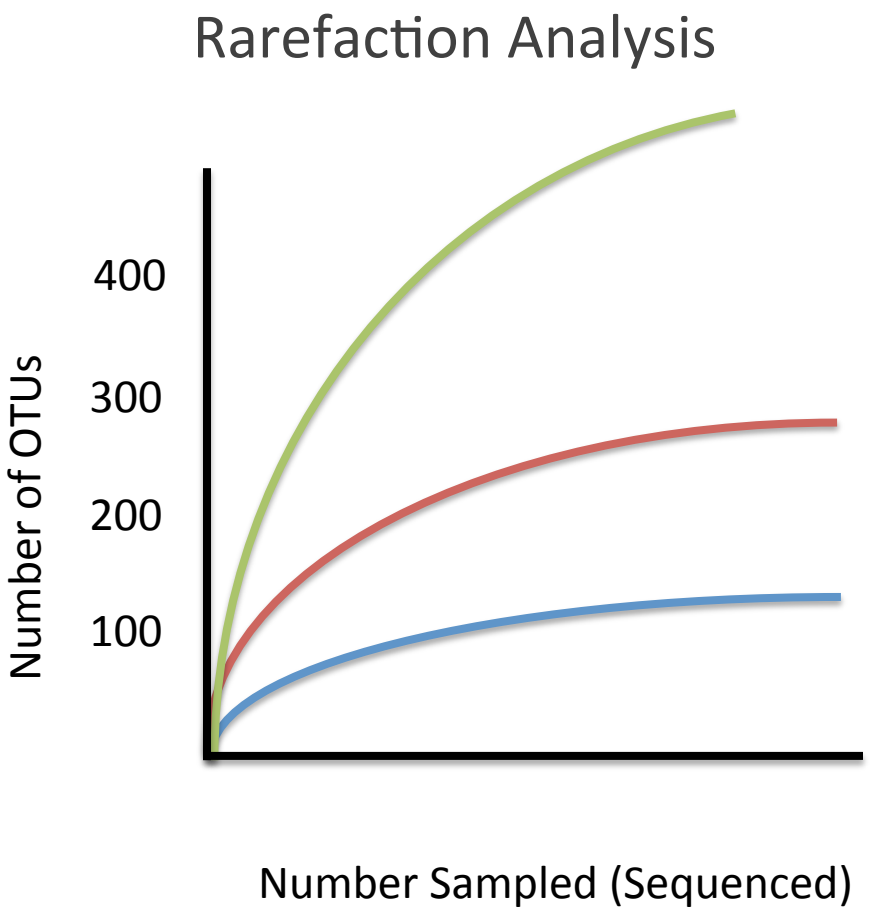
- Simpson's Index

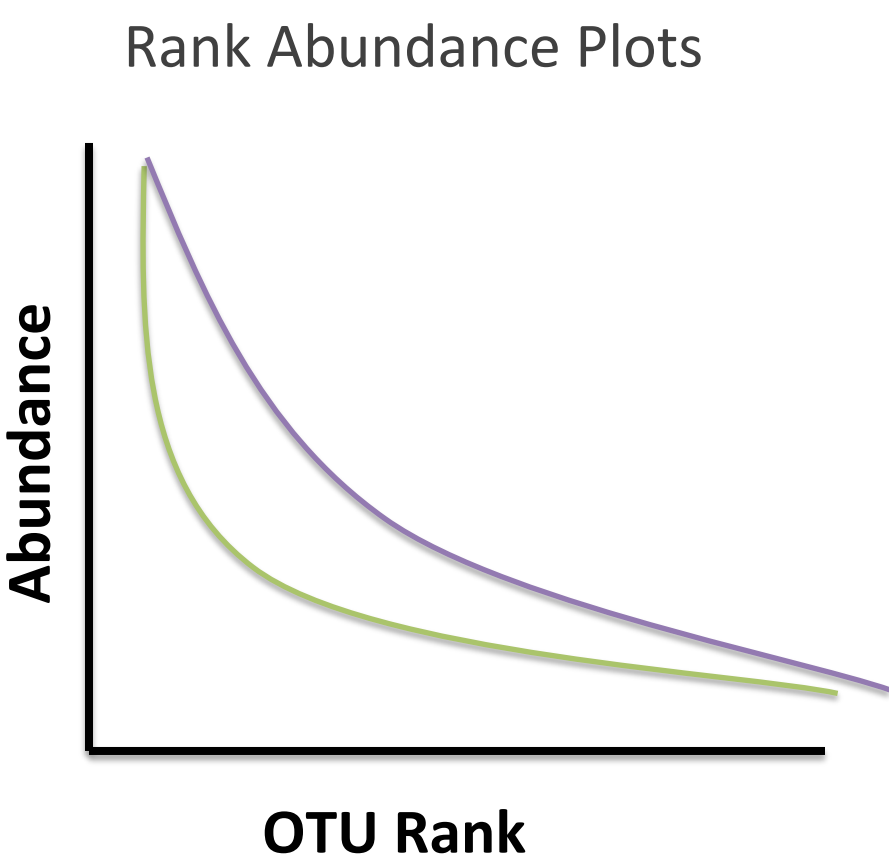
$$D = 1 - \frac{\sum_{i=1}^{S} n_i(n_i - 1)}{N(N-1)}$$

- Shannon Index

$$H' = -\sum_{i=1}^{S} p_i \ln p_i$$

Visualizing Richness and Evenness

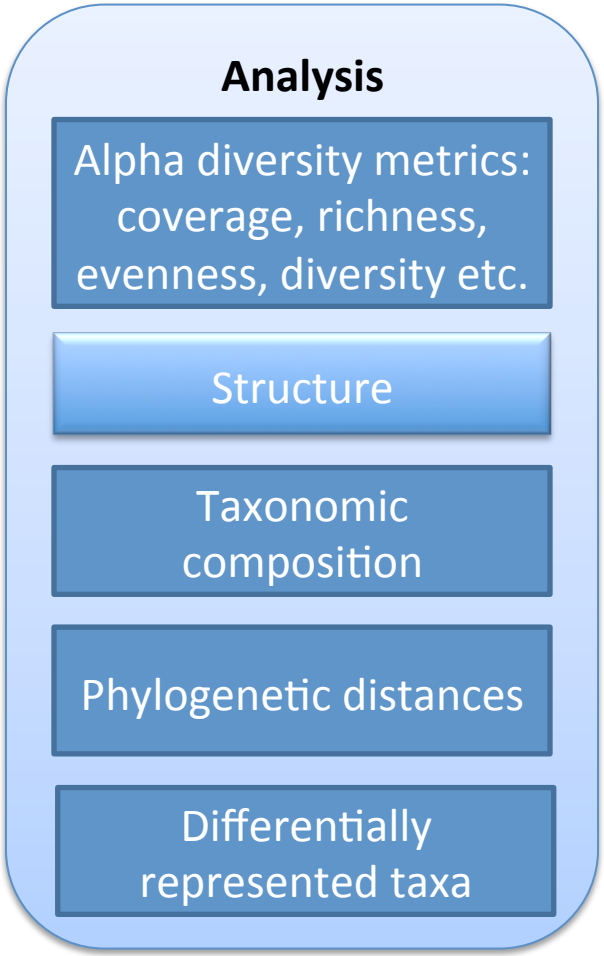




Beta-diversity: Comparing communities

- Reasons that researchers want to compare communities:
  - Determine if *variable X* has an impact on microbial communities
  - Determine if community structure changes in time or space
  - Determine if particular groups may be changing abundance
  - Identify taxa driving the differences

The rest...



- To be continued on Wed

Limitations of targeted metagenomics

- PCR error, chimeras and other artificial sequences
- PCR bias may result in missing some of the diversity
- Only provides taxonomic information (function prediction programs are limited by what's in the database)
- Novel or highly diverged microbes are difficult to study (show up as "unknown" or "unclassified" in the libraries)

Mothur

- Wrapper of many previous tools and techniques
- One interface to go from raw sequence data to diversity indices and visualization
- Efficiency - Collapses identical reads but keeps track of how many times they were seen
- Online wiki and manual to guide analsis
  - Mothur.org

Interactive mode

- Terminal directly for mothur
- Prompt will become

```
mothur >
```

- Important commands:

```
quit()    or  quit
help()
system(ls)
set.dir(input=../inputFiles)
set.dir(output=outputFiles)
```

- With the exception of quit, all commands require you to provide an open and close parentheses.
- If you supply any options there cannot be a space between the option, the '=', and the option setting. To separate options, use a comma.

```
mothur > cluster(method=furthest, cutoff=0.10)
```

Modes

- Three modes
- Interactive
- Command line

```
mothur
"#read.dist(phylip=98_sq_phylip_amazon
.dist, cutoff=0.1); cluster();“
```

  - Note quotes, # sign, semicolons

- Batch

```
mothur commands.batch
```

  - Convenient for saving commands and rerunning them exactly later on