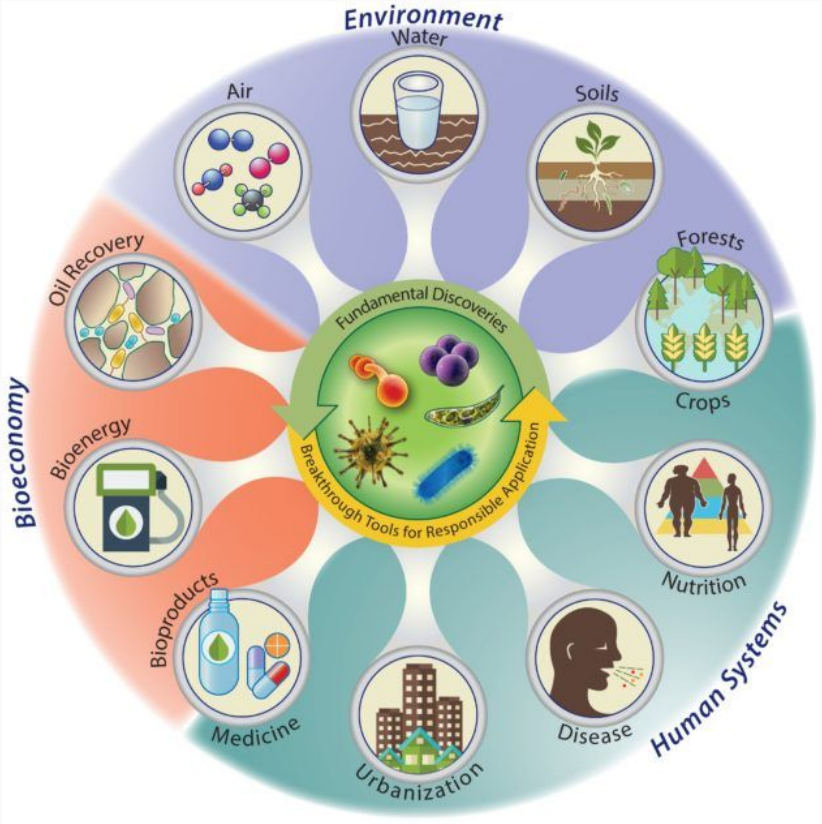
Microbiome:
16S rRNA Sequencing

Microbiome

- **Microbiome**: the microorganisms that live in an environment
- Includes:
  - Bacteria
  - Archaea
  - Fungi
  - Viruses

Every Earth Environment Has A Microbiome



https://newscenter.lbl.gov/2016/05/13/national-microbiome-initiative/

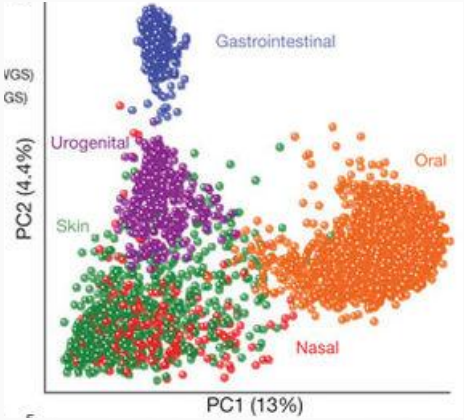
Human Microbiome Project

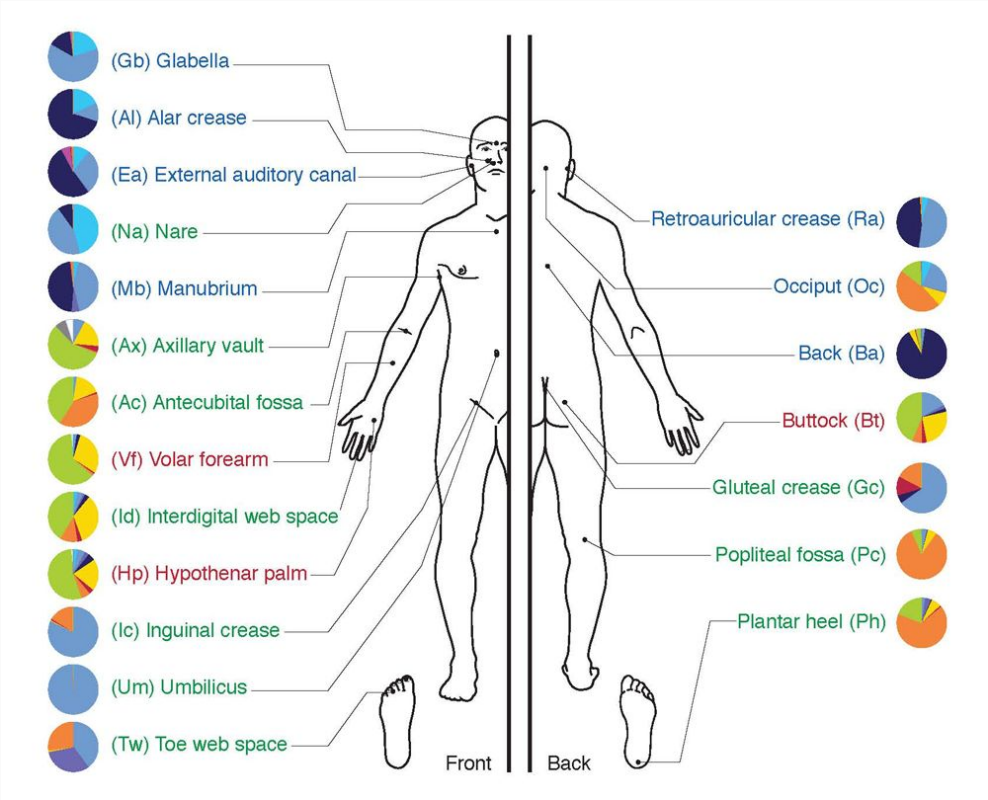
- Launched 2008, Finished (phase 1) 2012
  - Second phase is ongoing
- Championed culture-independent methods
  - 16S rRNA Sequencing
  - Whole Genome Sequencing
    - Created reference genomes!
- Not the first study to survey microbiomes
  - Just the first to survey the *human* microbiome large-scale

https://hmpdacc.org/
https://commonfund.nih.gov/hmp/

Human Body Microenvironments

Different microbes colonize different body areas

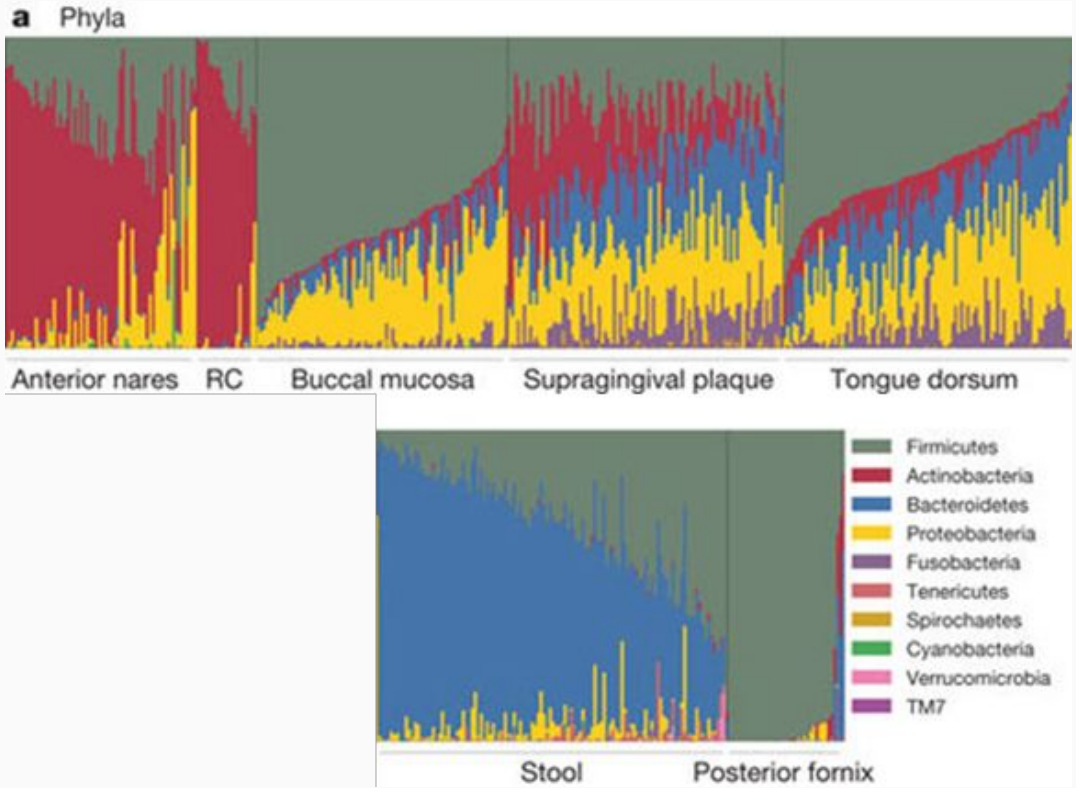




(Credit: Darryl Leja, NHGRI)

Human Microbiome Composition

Lots of variation between individuals



The Human Microbiome Project Consortium. (**2012**). Structure, function and diversity of the healthy human microbiome. *Nature*

Microbiome

Not all microbes are bad

- Still trying to understand microbe-microbe and microbe-host interactions

Earlier approaches rely on **isolating** microbes

- We are unable to culture many microbial species

→ Need to directly sequence

Studying the Microbiome

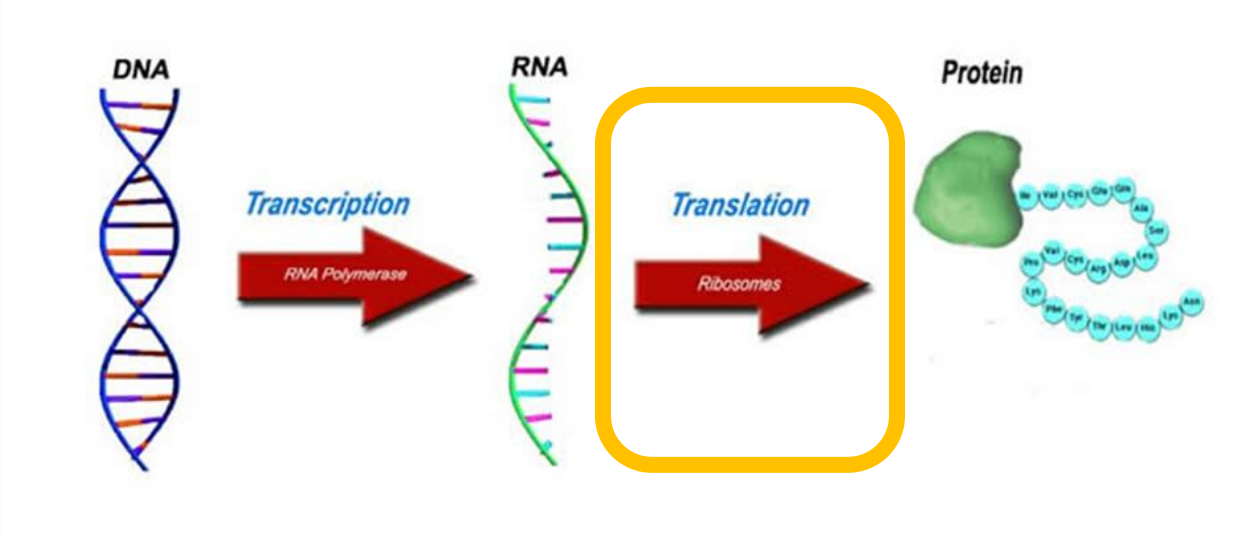
Want to understand community:

- **Structure**: species diversity
  - “Who is there?”
  - Studied with 16S ribosomal RNA (rRNA) Sequencing
- **Function**: genes + metabolic pathways
  - “What can/do they all do?”
  - Studied with metagenomics approaches

Ribosomal RNA

- Essential molecules for life
  - Every organism has them
- Form 60% of ribosomal mass
- Life is organized by which subunits present:
  - Prokaryotes: 5S, 23S, **16S**
  - Eukaryotes: 5S, 5.8S, 28S, and **18S**
- S = “Svedberg”
  - Svedberg = unit of molecular size

Why rRNA?



(Almost) all known life is driven by proteins

Translation requires **ribosomes**...

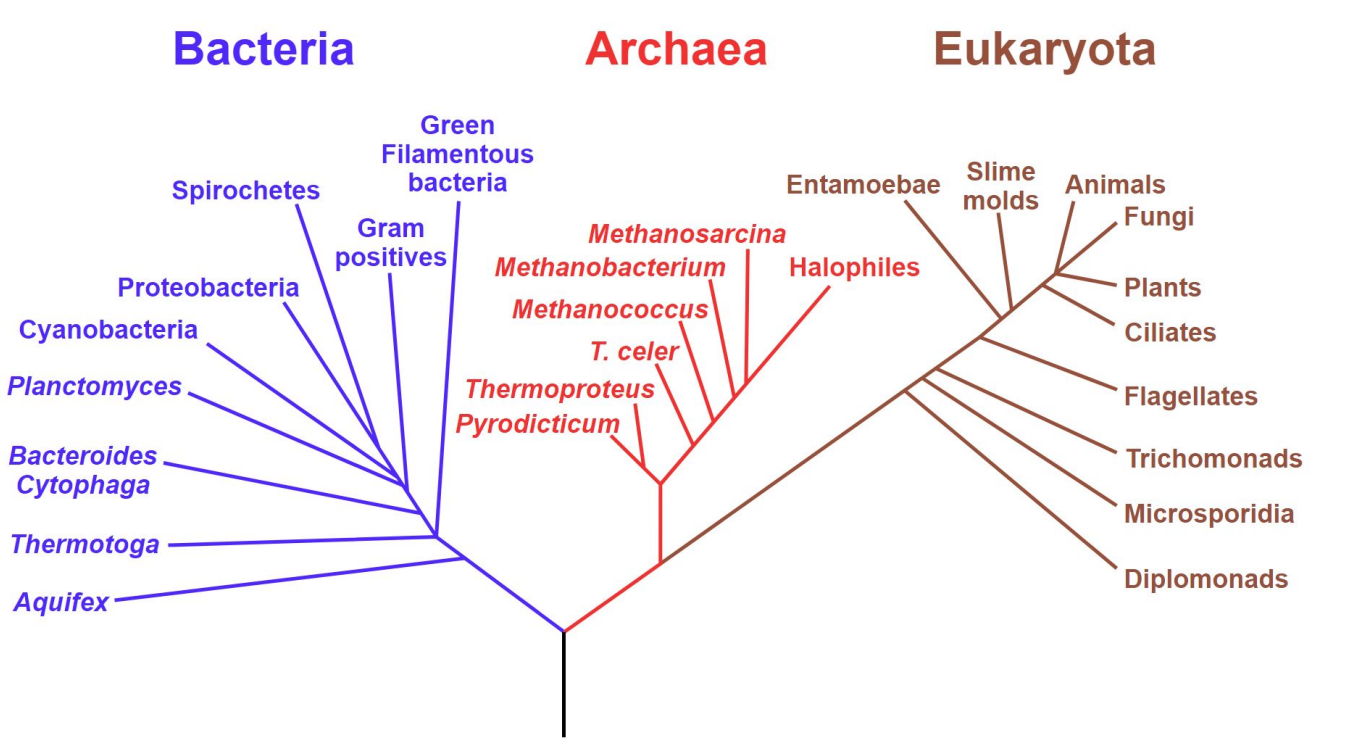
...**ribosomal RNA** (**rRNA**) encode ribosomes

Why rRNA?

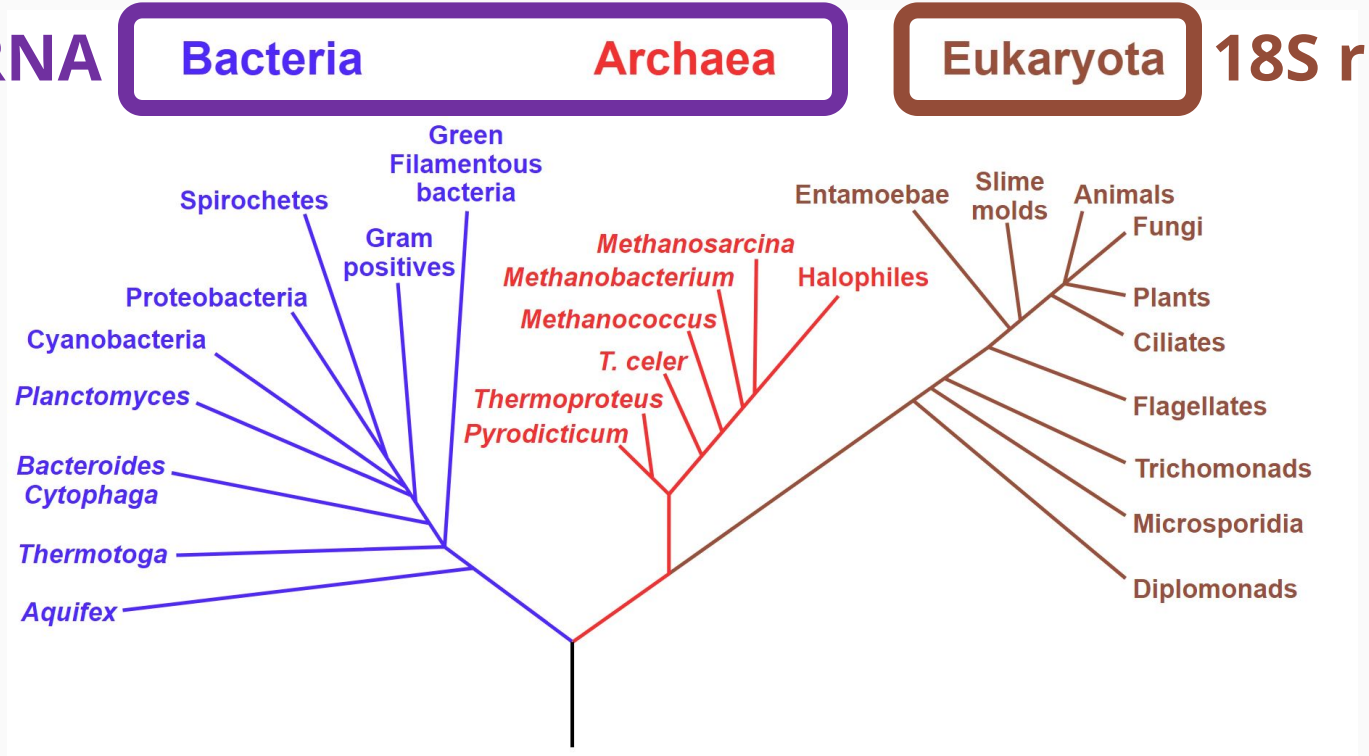
- Highly conserved
- Relatively short (~1.5 kb)
  - Short = cheap
- Generally different between species
  - smaller (variable) regions == cheap

Woese & Fox (**1977**). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *PNAS*
Lane, Pace, Olsen, Stahl, Sogin, & Pace (**1985**). Rapid determination of 16S ribosomal RNA sequence for phylogenetic analyses. *PNAS*

Phylogenetics



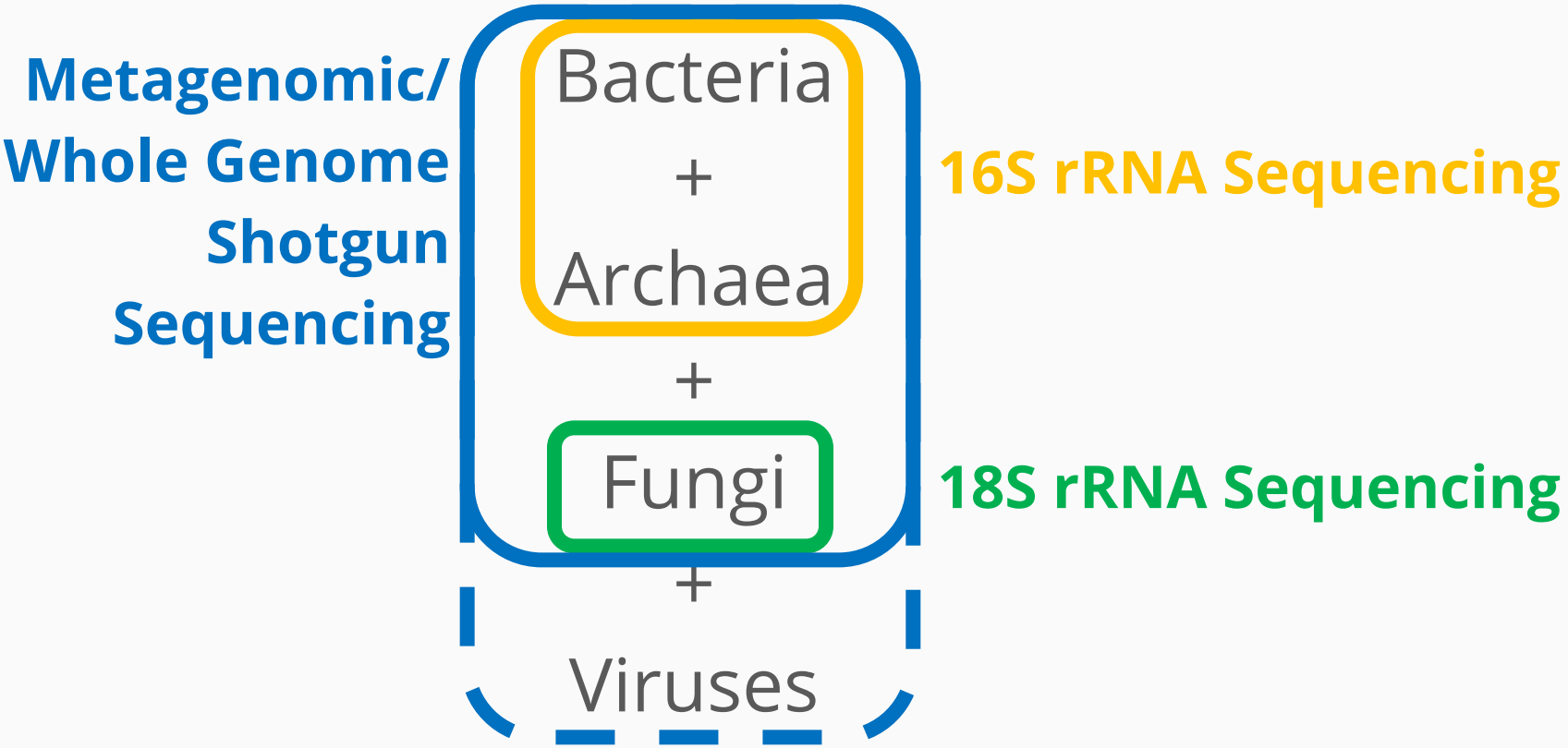
rRNA Genes as a Marker



Woese used rRNA to classify life into three domains

Woese, Kandler, & Wheelis (**1990**). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *PNAS*

Microbiome


Metagenomic/
Whole Genome
Shotgun
Sequencing
Bacteria
+
Archaea
+
Fungi
+
Viruses
16S rRNA Sequencing
18S rRNA Sequencing

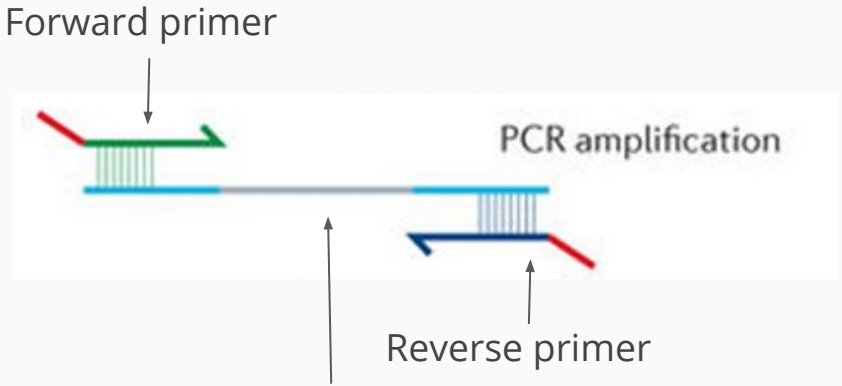
PCR

PCR: Polymerase Chain Reaction

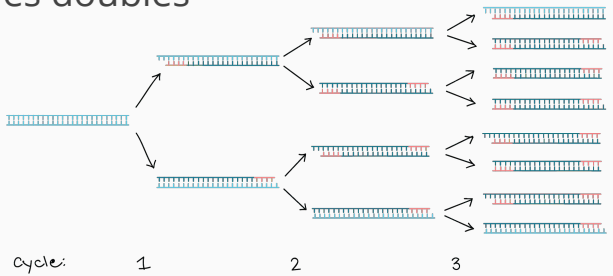
- "Molecular photocopying"
- Technique to amplify/copy small segments of DNA

How do you know you discovered something important?

- Win Nobel Prize
- Have people stop citing your work

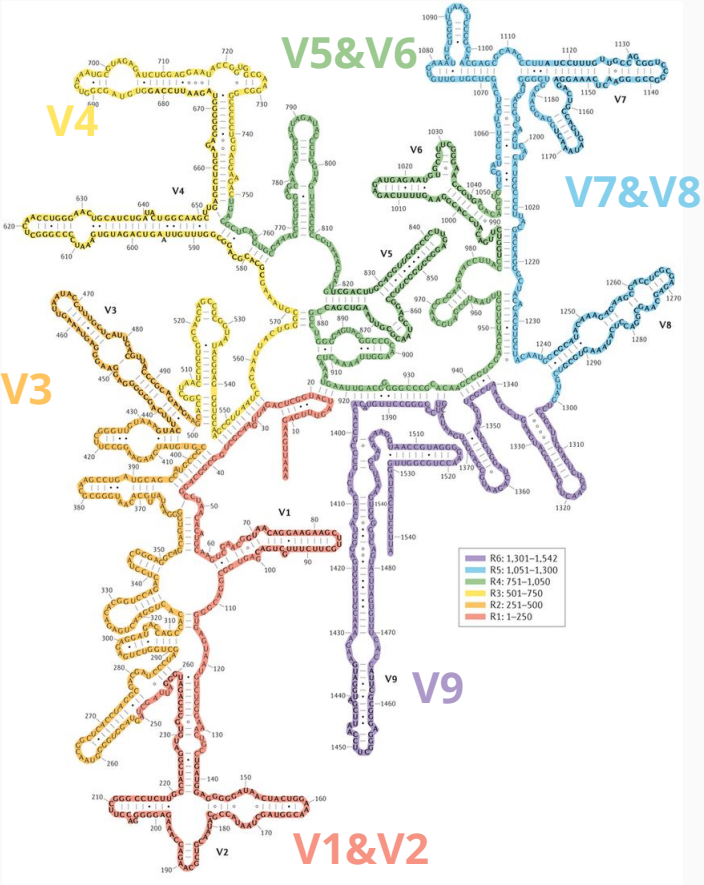


Each time process repeats, the number of copies doubles



https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/biotechnology/a/polymerase-chain-reaction-pcr

16S rRNA Gene



Ribosomal RNA **gene**

- Encodes the 30S small subunit

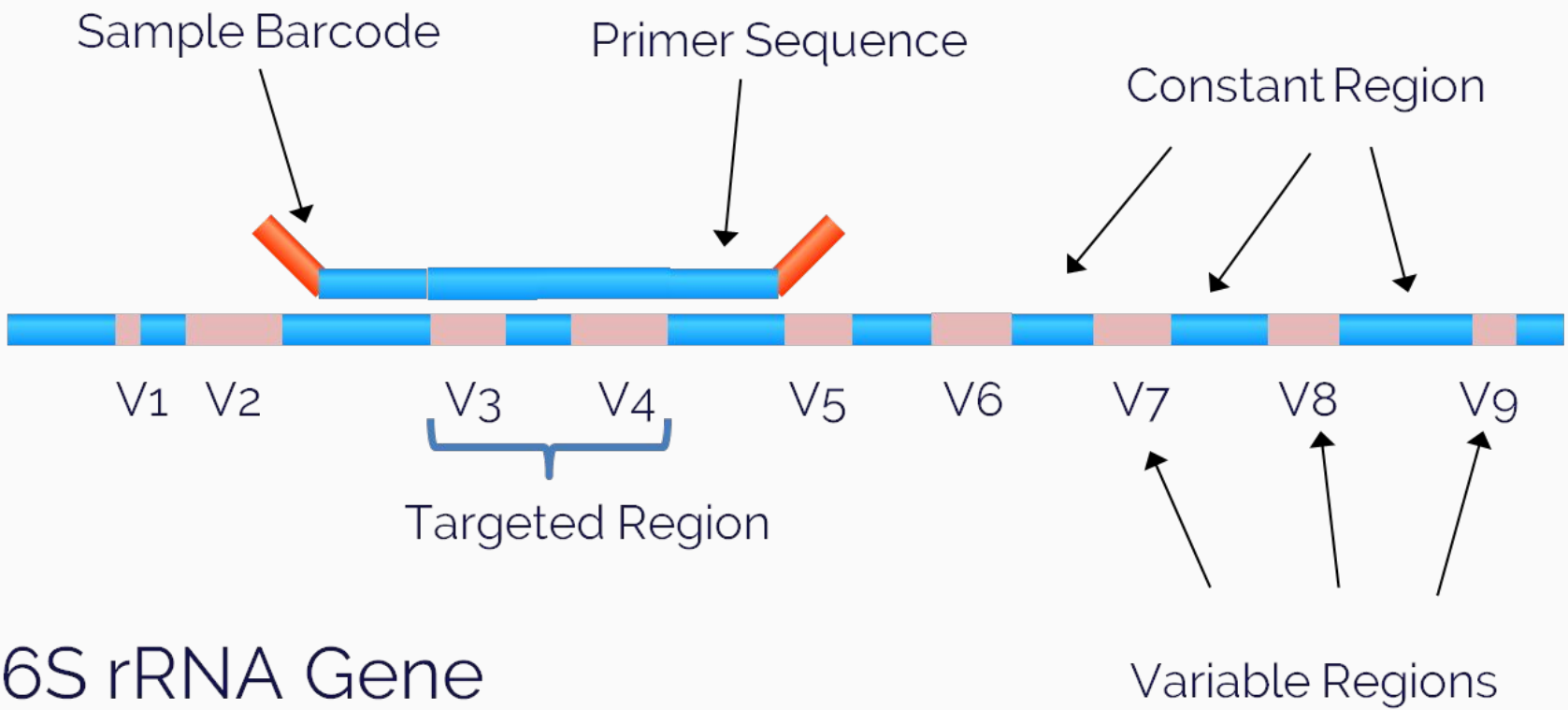
10 universal/constant regions

- Used to create primers

9 variable interstitial regions

- Used as signatures to determine phylogeny
- Choose which region(s) to sequence

Using Constant Regions



https://www.lcsciences.com/discovery/applications/genomics/16s-mobile/

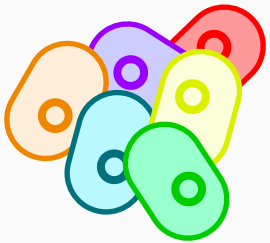
16S rRNA Sequencing

- Sanger sequencing (only for single isolate)
- High-throughput sequencing
  - Sequence PCR product
  - Sample contains many different species
  - Very few (50-100k) reads required to identify species
  - Compare with 30-80M for typical human RNA-Seq

16S rRNA Sequencing

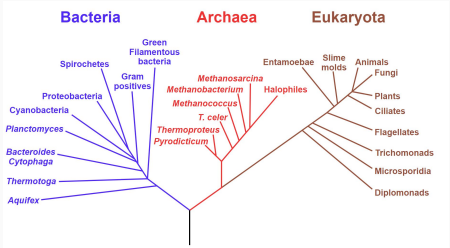
Survey:
Who's there?

Sample



16S rRNA Sequencing

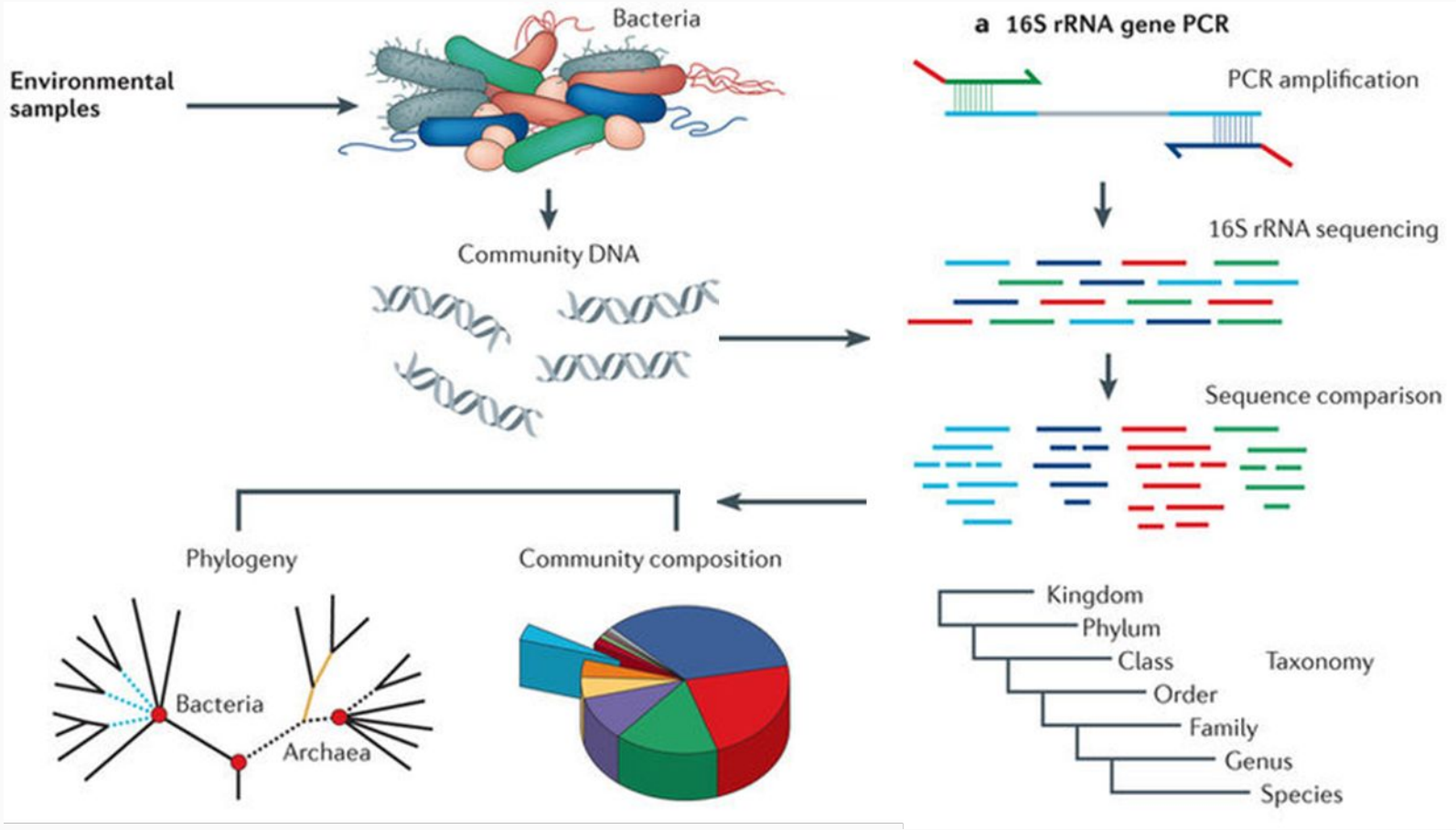
Taxonomic Classification



Population Analysis

- Alpha diversity
- Beta diversity
- Over/under-representation

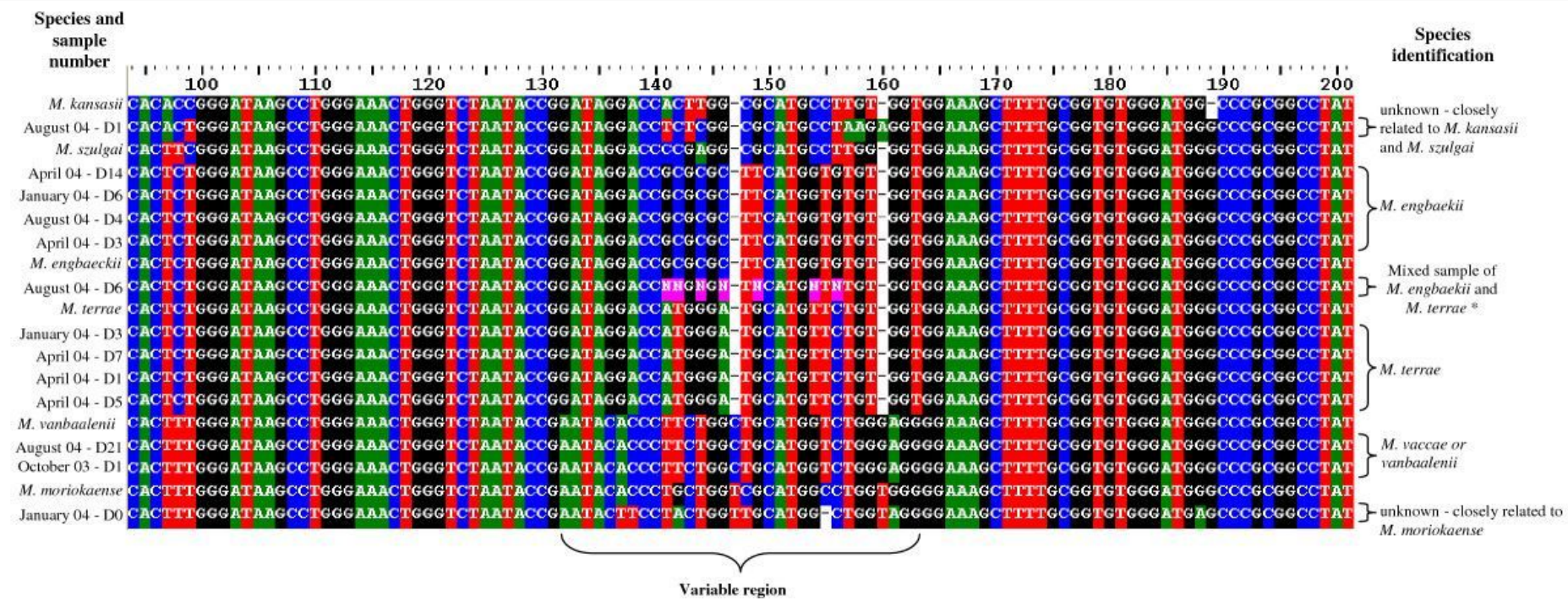
Basic Workflow



Preprocessing & Alignment

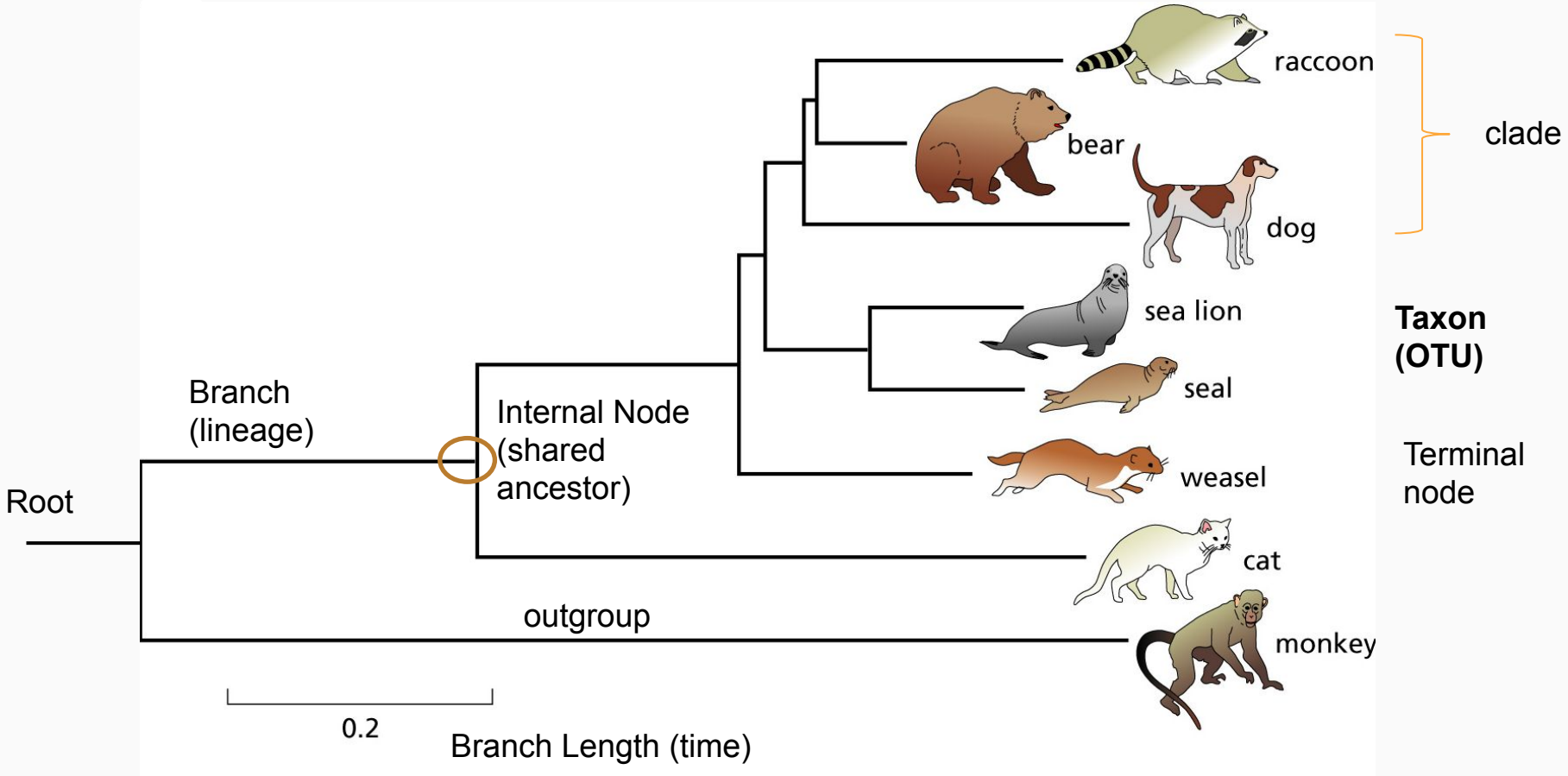
- Quality Assessment & Trimming
  - Remove adapters, PCR primers, and low-quality bases
  - Demultiplex using barcodes, discard reads without a barcode
- Each read (pair) is an rRNA sequence
- Collapse to unique reads
- Align against rRNA database (e.g. Silva)
- Cluster sequences by similarity

rRNA Sequence Encodes Relatedness



Michel, Anita L., et al. 2007. “Bovine Tuberculosis in African Buffaloes: Observations Regarding Mycobacterium Bovis Shedding into Water and Exposure to Environmental Mycobacteria.” *BMC Veterinary Research* 3 (September): 23.

Phylogenetic Tree Terminology



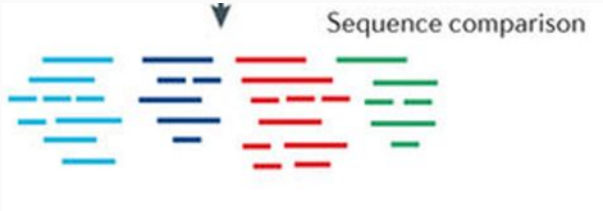
Binning

OTU: Operational Taxonomic Unit (also “phylotype”)

- A group of sequences clustered together based purely on similarity and an arbitrary threshold (e.g. 90% identical, 95%...)
- May or may not be equivalent to taxonomical entities (species, genera, *etc.*)

Can cluster based on similarity to a reference database or *de novo* (compared to each other)

- Can also cluster *de novo* and then assign taxonomy



(Linnaean) Taxonomy

Domain	Eukaryota	Bacteria
Kingdom	Animalia	Bacteria
Phylum	Chordata	Proteobacteria
Class	Mammalia	Gammaproteobacteria
Order	Primates	Enterobacteriales
Family	Hominidae	Enterobacteriaceae
Genus	Homo	Escherichia
Species	*Homo sapiens*	*Escherichia coli*

Taxonomic Resolution

Domain
Kingdom
Phylum
Class
Order
Family
Genus
Species

16S/18S rRNA does not yield species-level information
→ **genus**-level, at best, but usually higher

- Closely-related species have a high sequence similarity across the 16S gene
- Typically don’t sequence the whole gene, just 1(+) variable regions

Population Measurements

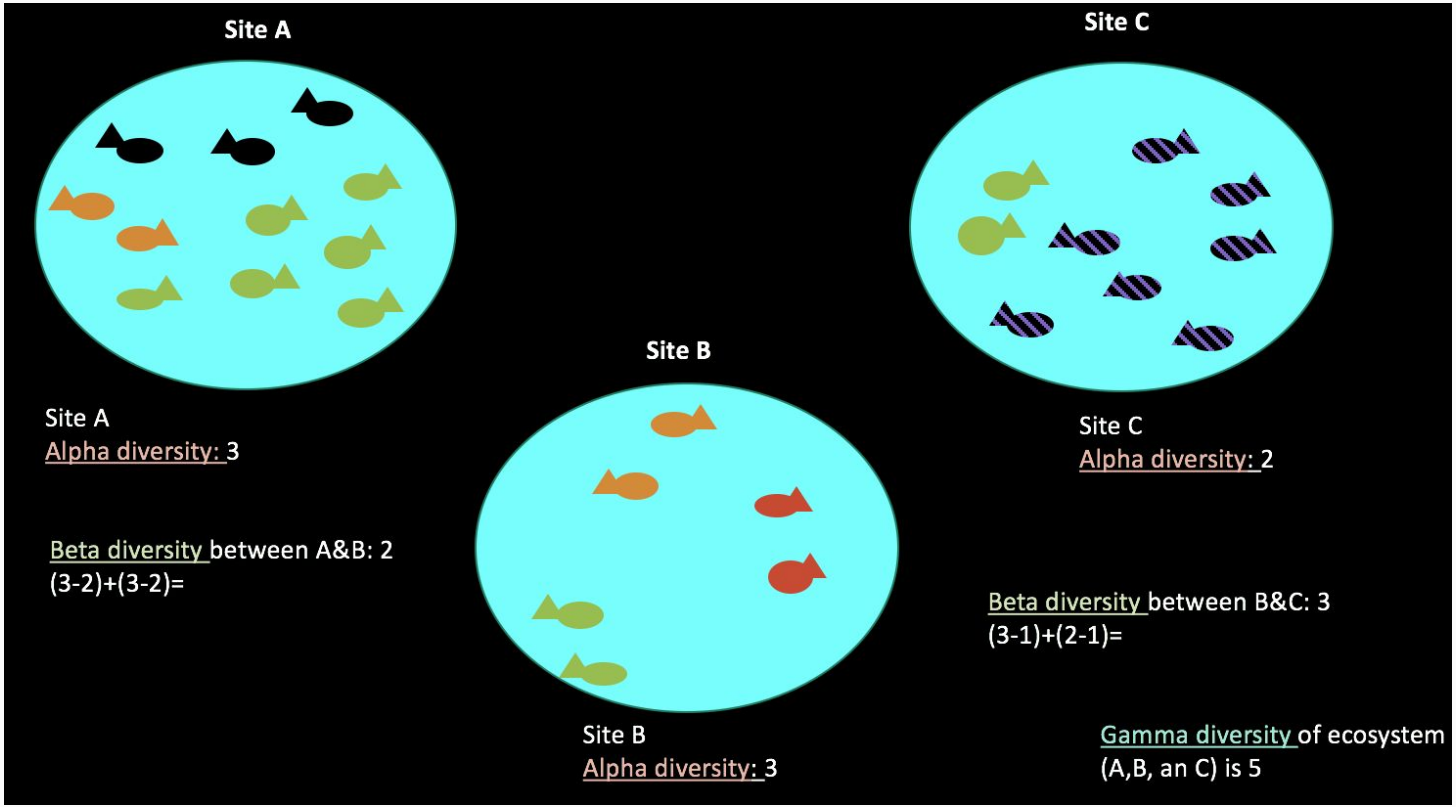
Alpha Diversity: Diversity within a sample

Beta Diversity: Diversity between samples

Evenness: Distribution of taxa

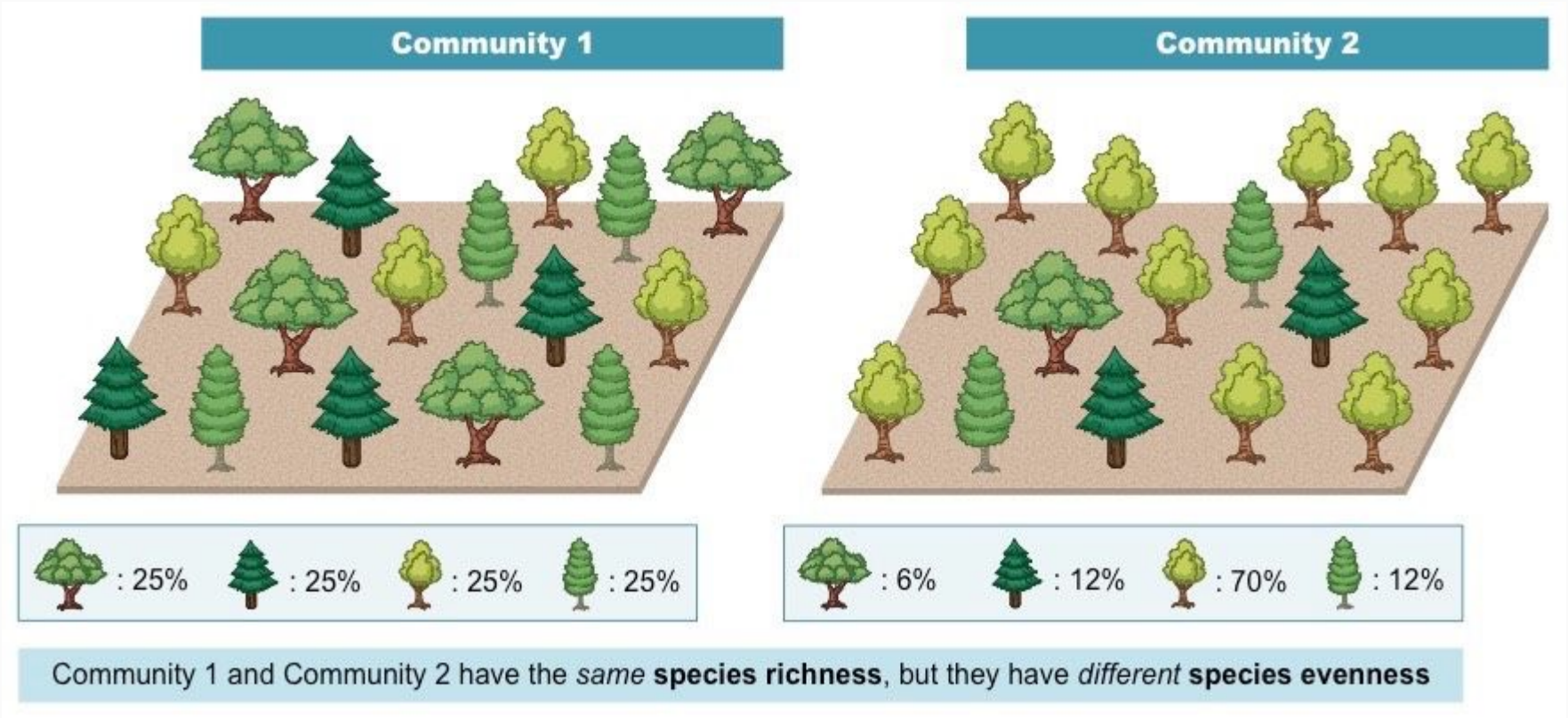
Richness: Number of taxa

Alpha and Beta Diversity



https://socratic.org/questions/what-are-alpha-beta-and-gamma-diversity

Evenness vs Richness

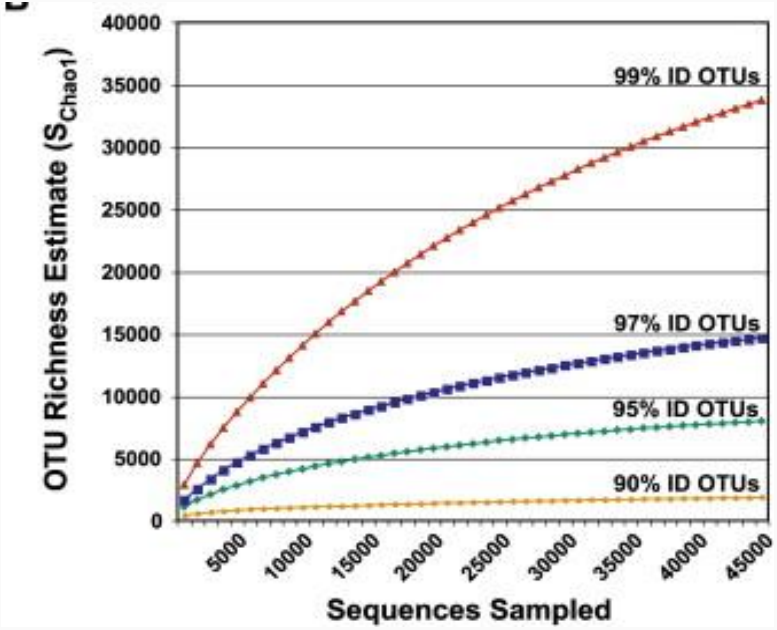


http://ib.bioninja.com.au/options/option-c-ecology-and-conser/c4-conservation-of-biodiver/biodiversity.html

Richness: Rarefaction Curves

Cutoffs: What percent sequence identity should you use?

→ Will depend on the error rate, *etc.*



Things to Keep in Mind

- Sequencing platform
  - Error rate, biases, read length, noise
- Choice of variable region(s)
- Amplification process
  - Error rate, biases, choice of primer, DNA template concentration, PCR cycle number, introduction of chimeras
- Coverage/Depth

Software Packages

mothur (your project)
QIIME (“chime”)
bioBakery* (PICRUSt)
CloVR*

*Not currently on the SCC, except for specific projects

Taxonomy Databases

Greengenes - 16S rRNA database (older)
MG-RAST - Metagenomic database
NCBI - Microbial genome & gene database
RDP - 16S & 18S Ribosomal Database
SILVA - 16S/18S/23S/28S rRNA Database

*A different database may give you different taxonomic results

Notes

16S rRNA Sequencing analysis is qualitative

- Surveying who is there

PCR depends on the *a priori* knowledge assumption of universal primers

- May yield an altered/incomplete estimation of diversity
- Also, uneven primer annealing, uneven amplification, *etc*...

Are converting either to binary data (presence/absence) or normalizing (relative abundance)

- Make sure you are using the appropriate statistical/analysis tools for binary and normalized data!

Notes

Taxonomic classification will depend on

- The resolution which variable region of the 16S rRNA gene is used
- The primers used for PCR
- Which database is used
- Which software package is used
- What cutoff is used
- Sequence coverage/depth
- Sequencing platform
- The species composition in the community being analyzed
- When you perform the analysis
- ...

Considerations

Define the question as precisely as possible.

- What controls do you need?

What sequencing platform will you use?

- Illumina is the typical platform (right now)

What region of the 16S rRNA gene will you amplify?

- V4 usually yields genus-level

How many reads do you need per sample?

- Coverage/Depth

What are hidden technical issues?

- Example: Chimeras

What analysis tool will you use? How will you display your data? How will you compare your results with other published studies?