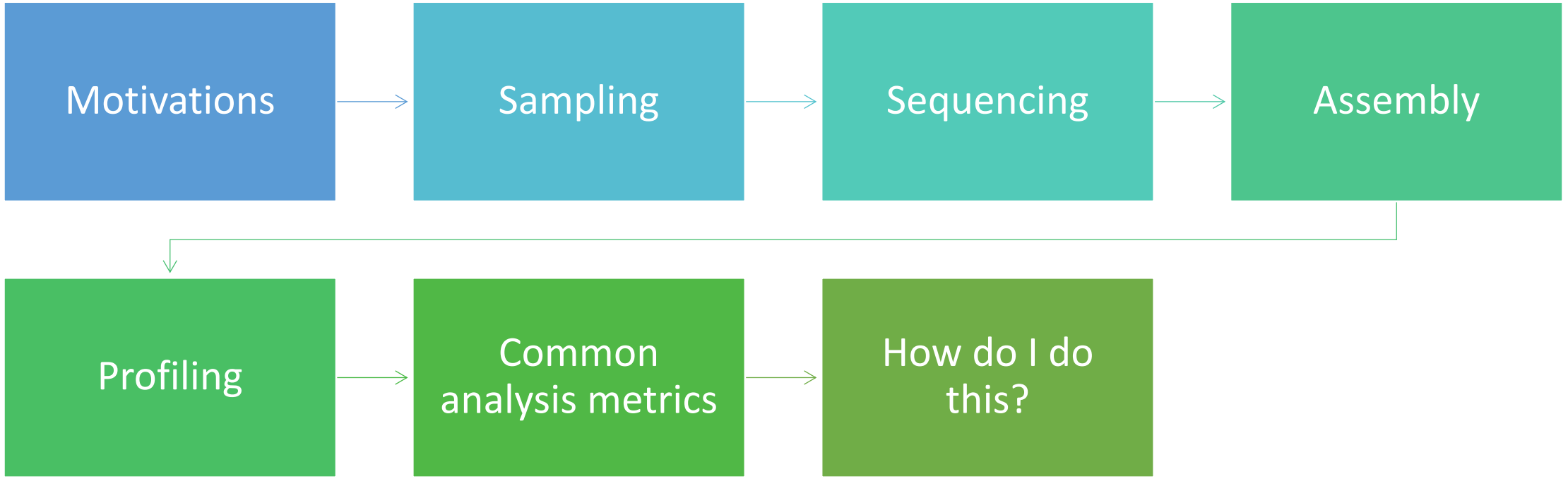
A Primer on Metagenomics

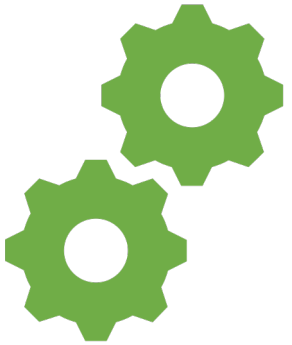
Journal Club 11/10

Outline


Motivations
Sampling
Sequencing
Assembly
Profiling
Common analysis metrics
How do I do this?

Motivations

Higher level resolution: species and strain.



Functional analysis possible.

Getting Raw Reads

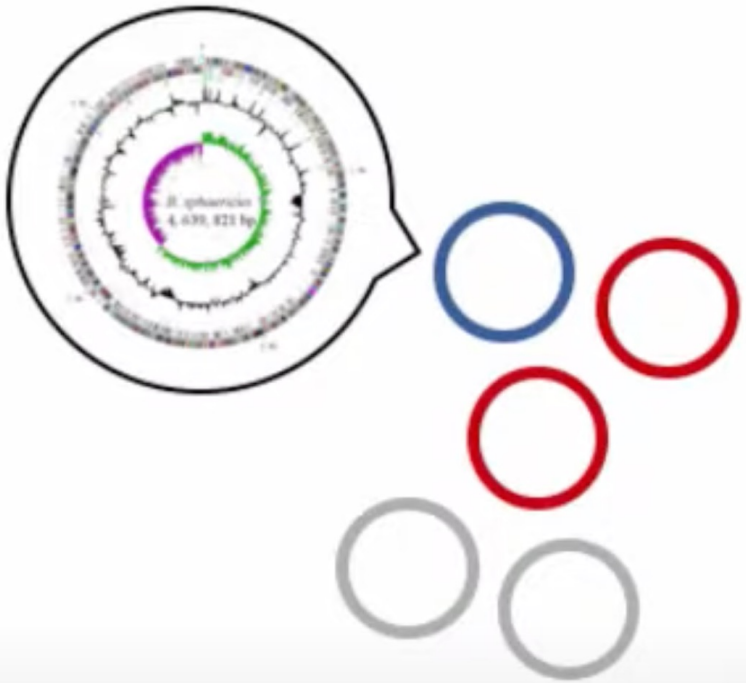
Sampling

- Representative?
- Filtering
  - Physical (viroids, protists, phages)
  - Computational (annotate and remove)

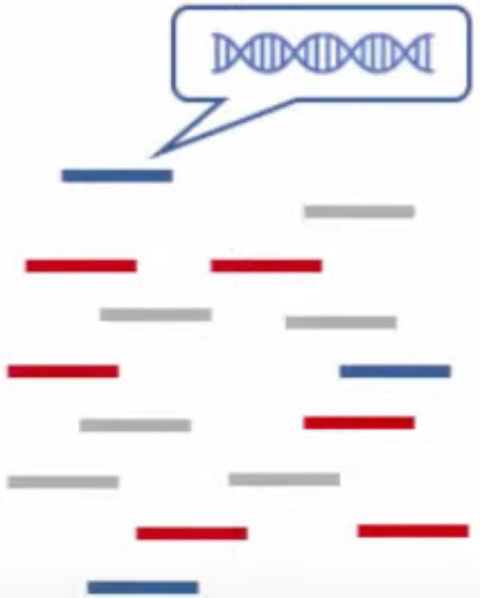
Sequencing

- Sanger sequencing wholly replaced by NGS methods.
  - Third generation in development.
- Provide high volumes of short reads (Illumina very common).
- Various biochemical methods.

Whole metagenome shotgun (WMS) sequencing



Microbial genomes
in our sample



Shatter into fragments
(~300 nucleotides each,
random, no amplification)

Sequence the fragments
(millions per sample)

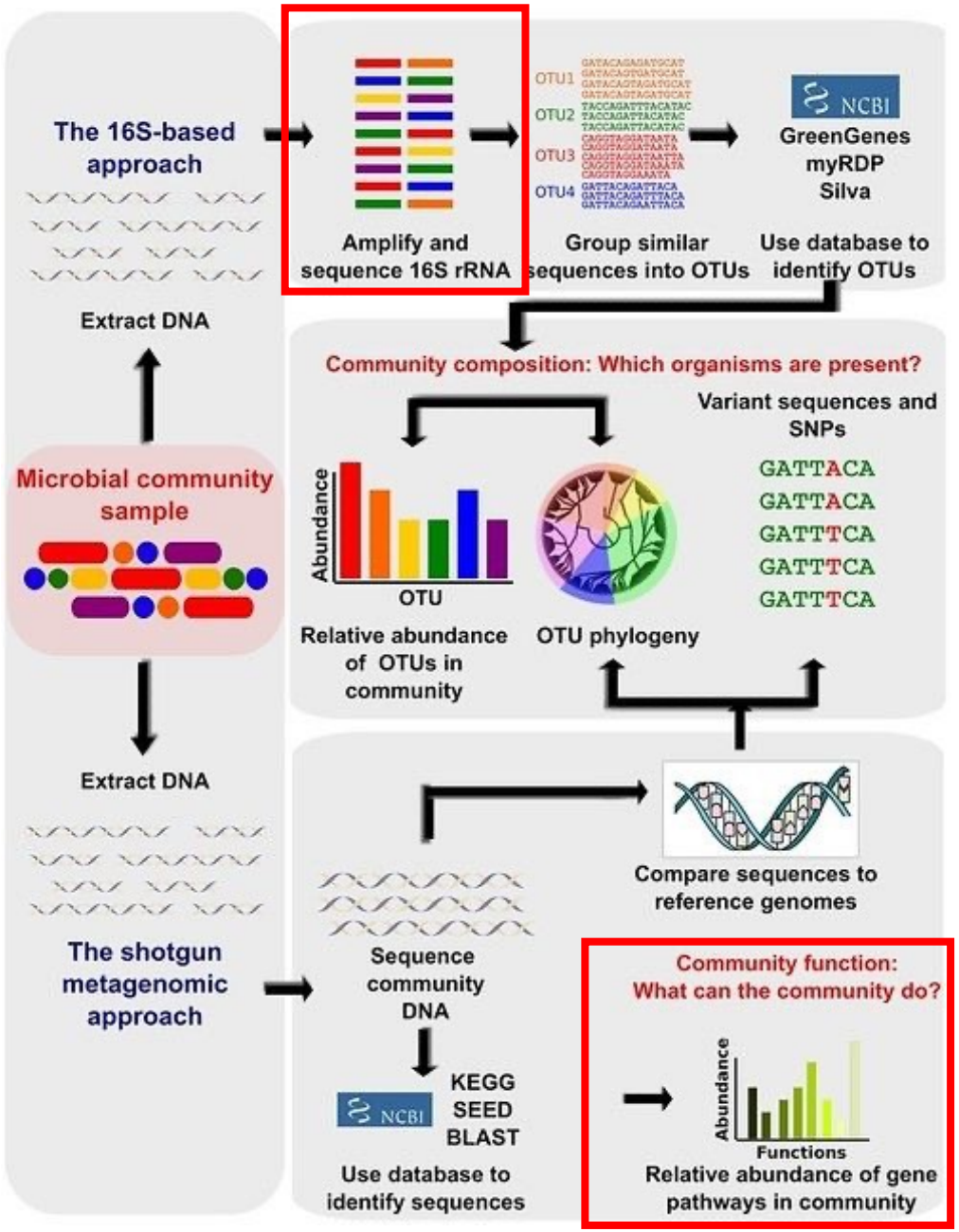
This is a metagenome

**Used to have to clone using Sanger sequencing.*

Updated Table of Sequencing Methods

NGS platforms					
Platform	Template preparation	Chemistry	Max read length (bases)	Run times (days)	Max Gb per Run
Roche 454	Clonal-emPCR	Pyrosequencing	400‡	0.42	0.40-0.60
GS FLX Titanium	Clonal-emPCR	Pyrosequencing	400‡	0.42	0.035
Illumina MiSeq	**Clonal Bridge Amplification**	**Reversible Dye Terminator**	**2x300**	**0.17-2.7**	**15**
Illumina HiSeq	**Clonal Bridge Amplification**	**Reversible Dye Terminator**	**2x150**	**0.3-11**[11]	**1000**[12]
Illumina Genome Analyzer IIX	Clonal Bridge Amplification	Reversible Dye Terminator[13][14]	2x150	2-14	95
Life Technologies SOLiD4	Clonal-emPCR	Oligonucleotide 8-mer Chained Ligation[15]	20-45	4-7	35-50
Life Technologies Ion Proton[16]	Clonal-emPCR	Native dNTPs, proton detection	200	0.5	100
Complete Genomics	Gridded DNA-nanoballs	Oligonucleotide 9-mer Unchained Ligation[17][18][19]	7x10	11	3000
Helicos Biosciences Heliscope	Single Molecule	Reversible Dye Terminator	35‡	8	25
Pacific Biosciences SMRT	Single Molecule	Phospholinked Fluorescent Nucleotides	10,000 (N50); 30,000+ (max)[20]	0.08	0.5[21]

How does this differ from 16S?



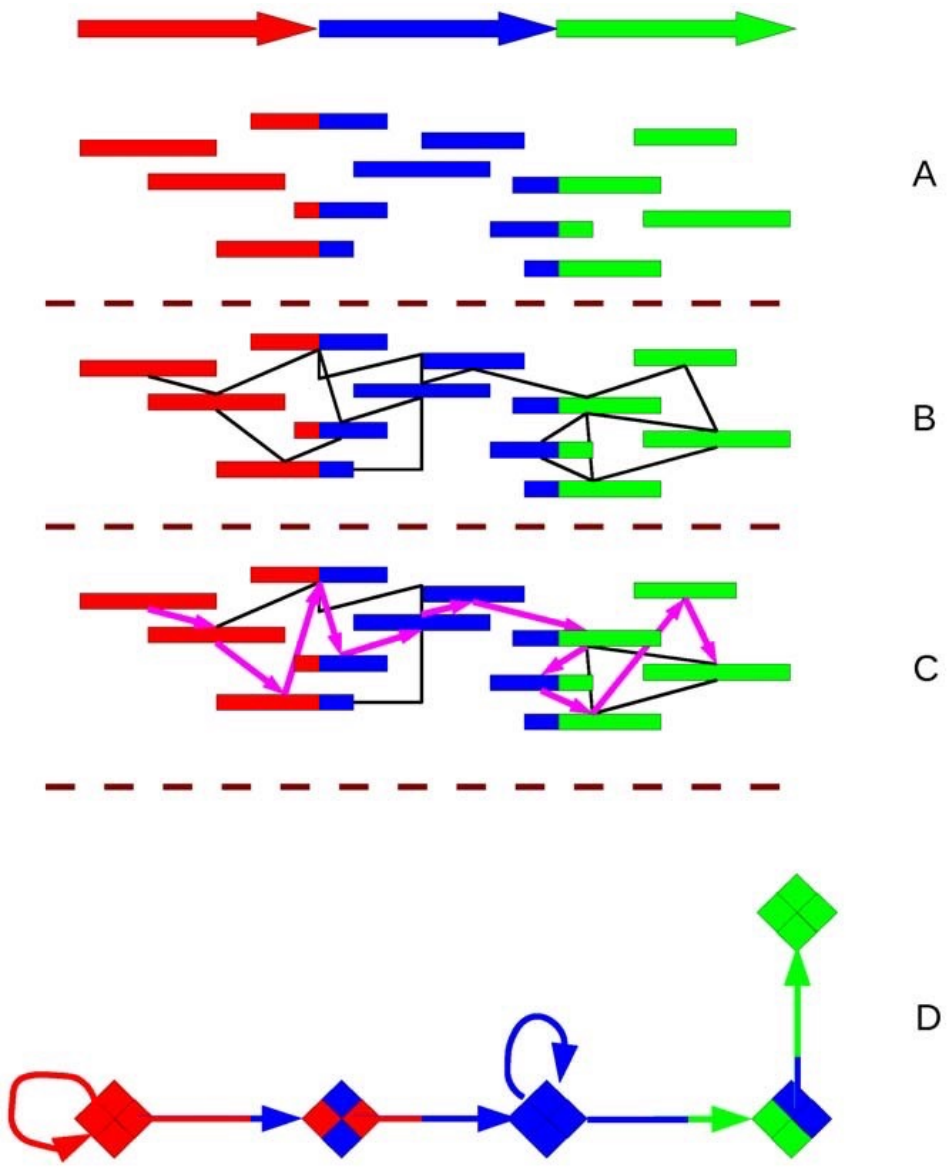
Assembly

- Computationally-intensive process to combine the millions of reads into longer reads (contigs).
- Useful for annotation of larger functional regions.
- Not strictly necessary for pure taxonomic classification.

Sequence Length (bp) Genome Element
25–75 SNPs, short frameshift mutations
100–400 Short functional signatures
500–1,000 Whole domains, single domain genes
1,000–5,000 Short operons, multidomain genes
5,000–10,000 Longer operons, some cis-control elements
>100,000 Prophages, pathogenicity islands, various mobile insertion elements
>1,000,000 Whole prokaryotic chromosome organization

The Process of Assembly

- **Hamiltonian Path (A-C)**
  - Represent each read as a vertex on a graph.
  - Correct assembly is visiting each overlapping read once.
  - NP-complete: exponential time complexity.
- **Eulerian Path (D)**
  - Use k-mer words, which is not affected by repeat reads.
  - Reads as edges rather than vertices: Eulerian path.
  - Time complexity: O(n).



Annotation / Profiling / Classification

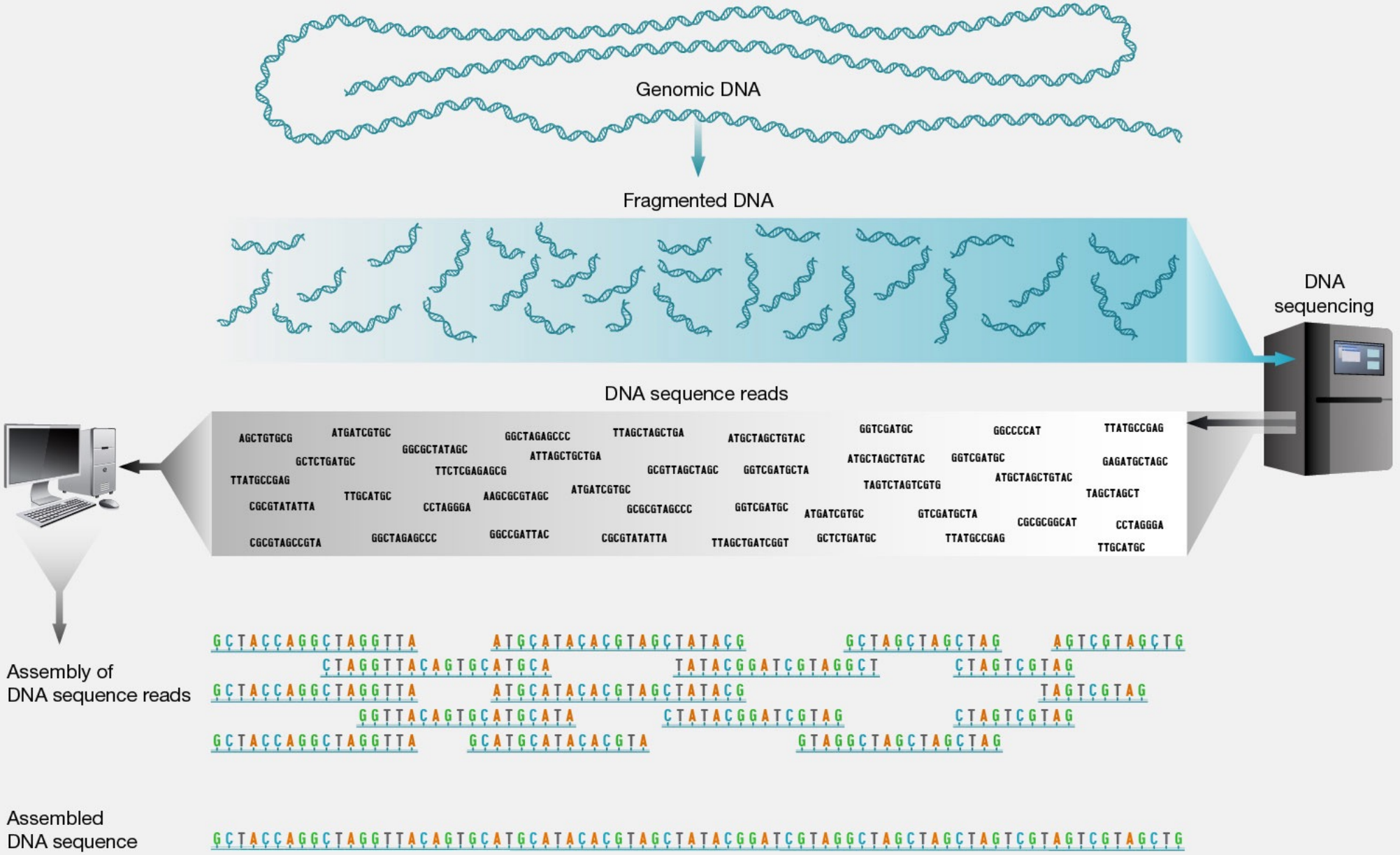
Taxonomic

- Tools:
  - *Kraken2* – k-mer approach
  - *MetaPhlAn* – species-marker approach
  - *Woltka* – OGU approach
- Yield taxonomic information about the organisms and their abundances in the samples.

Functional Analysis

- Tools:
  - HUMAnN2 – Biobakery
  - HMMER – Hidden Markov Models
- Yield information about proteins/products/enzymes.
- What is it *capable* of doing, as opposed to **metatranscriptomics / metaproteomics.**

Review


Genomic DNA
Fragmented DNA
DNA sequencing
DNA sequence reads
AGCTGTGCG
ATGATCGTGC
GCTCTGATGC
GGCGCTATAGC
TTATGCCGAG
TTCTCGAGAGCG
TTGCATGC
CGCGTATATTA
CCTAGGGA
AAGCGCGTAGC
CGCGTAGCCGTA
GGCTAGAGCCC
GGCCGATTAC
GGCTAGAGCCC
ATTAGCTGCTGA
TTAGCTAGCTGA
GCGTTAGCTAGC
ATGATCGTGC
GCGCGTAGCCC
CGCGTATATTA
ATGCTAGCTGTAC
GGTCGATGCTA
GGTCGATGC
TTAGCTGATCGGT
GGTCGATGC
ATGCTAGCTGTAC
TAGTCTAGTCGTG
ATGATCGTGC
GCTCTGATGC
GGCCCCAT
GGTCGATGC
ATGCTAGCTGTAC
GTCGATGCTA
TTATGCCGAG
TTATGCCGAG
GAGATGCTAGC
TAGCTAGCT
CGCGCGGCAT
CCTAGGGA
TTGCATGC
Assembly of
DNA sequence reads
Assembled
DNA sequence

Common Analysis Metrics

Alpha Diversity

- Measures the number of different features of a given community (i.e., sample).
- Shannon Diversity Index
- Inverse Simpson Index

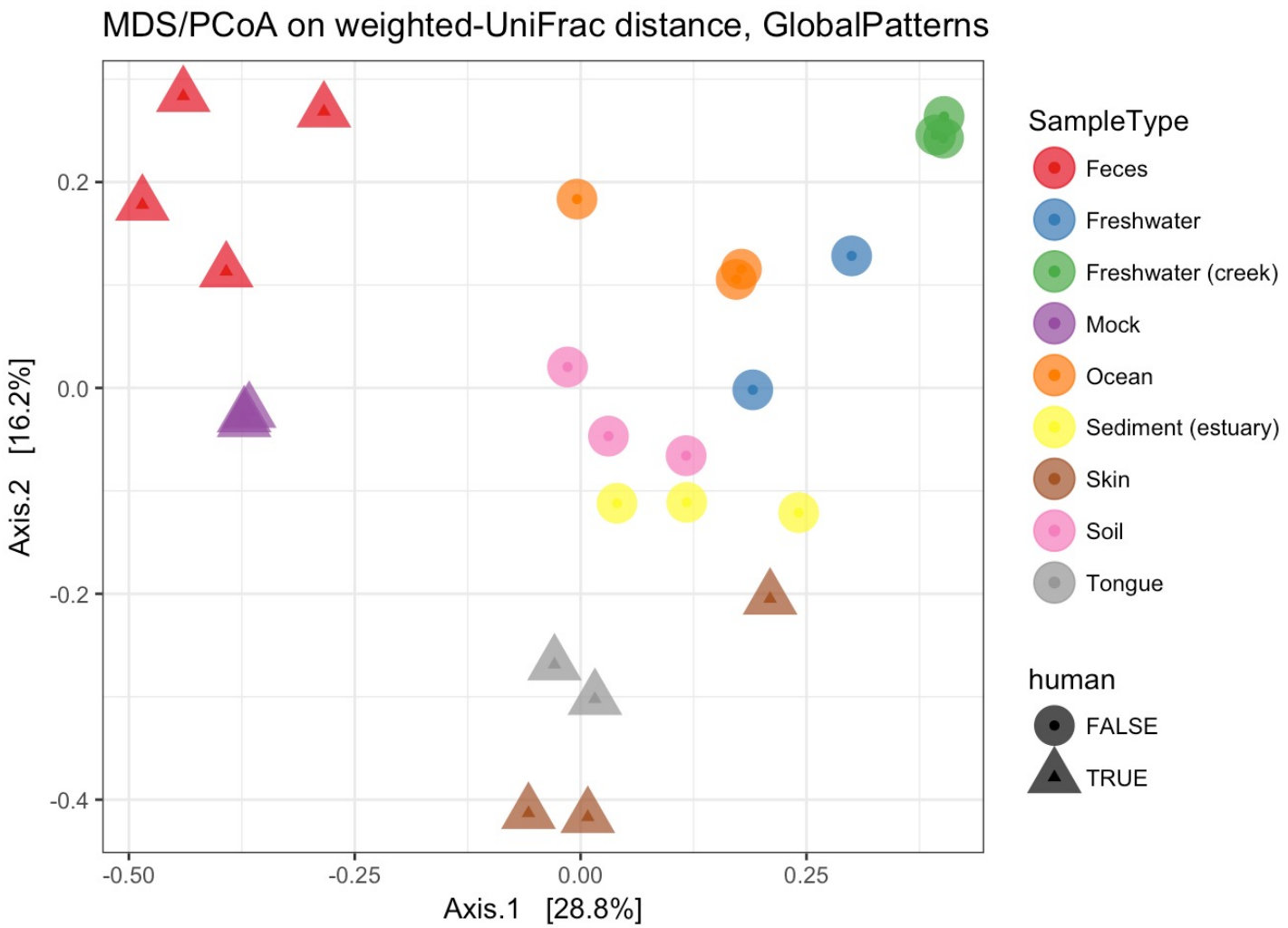
Beta Diversity

- Measures the similarity (distance) between two communities.

- Bray-Curtis Dissimilarity
- Unifrac Distance
- Jaccard Distance
- **Aitchison Distance (CLR-transformed)**

Ordination Plots

- Take the beta diversity metrics from the previous plots.
- Plot the samples on fewer dimensions to observe clustering.
- Can reveal relationships between variables if clustering is observed.



How do I do this?

- Several useful, publically-available packages.
  - JAMS
  - Biobakery
  - Nephele (WGSA2)
  - Woltka
  - Many others