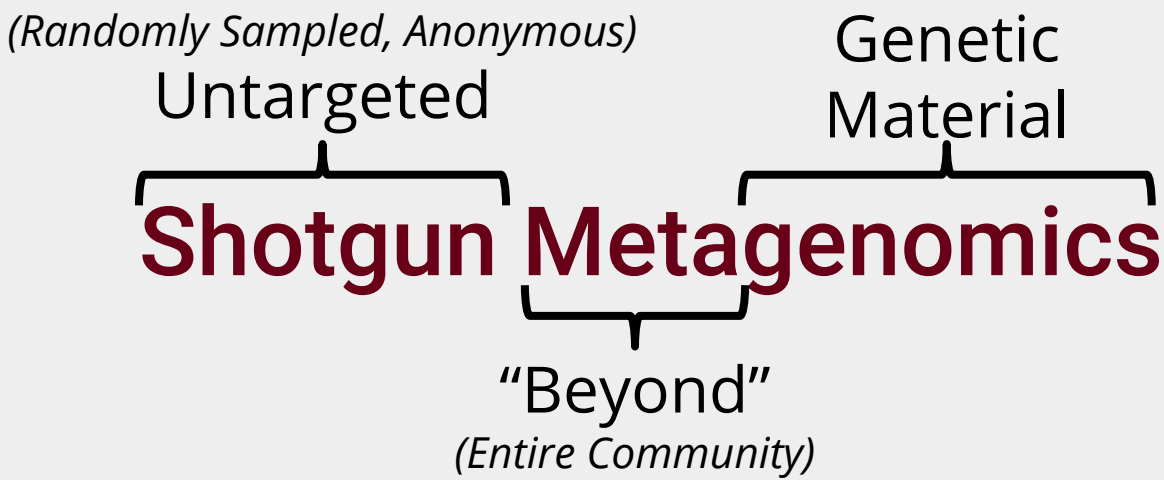
Microbiome: Metagenomics


(Randomly Sampled, Anonymous)
Untargeted
Genetic Material
Shotgun Metagenomics
“Beyond”
(Entire Community)

Why Shotgun Metagenomics?

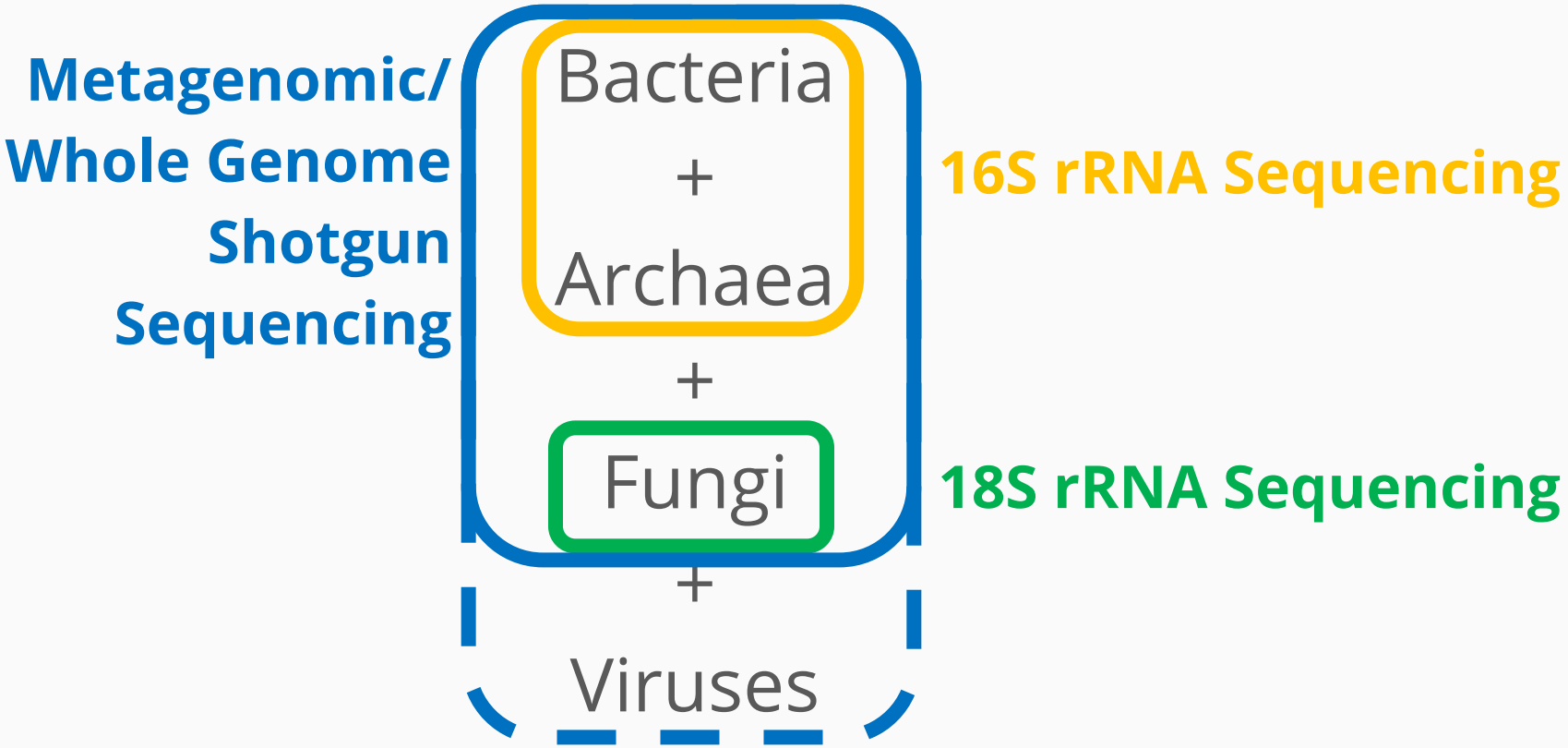
Recover **whole genome** sequences of all microbial community members, not just selected organisms or single marker genes

- “All” == above a certain threshold (where the threshold depends on coverage/depth and the microbe’s abundance in the sample)

Can perform **large-scale** investigations of **complex** microbial communities

- **Structure:** Taxonomic composition (who’s there?)
- **Function:** Metabolic potential (what can they do?)

Microbiome


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

Taxonomic Resolution

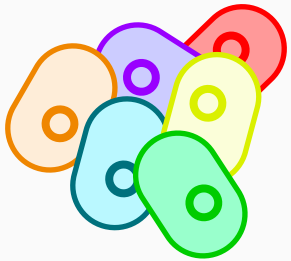
Domain
Kingdom
Phylum
Class
Order
Family
Genus
Species

Shotgun metagenomics yield **species**-level information

- Closely-related species have a high sequence similarity
  - Similar issue to 16S sequencing

Metagenomic Sequencing

Sample



Shotgun Metagenomic Sequencing

Who's there?

Taxonomic Classification

Population Analysis

Functional Analysis

What can they do?

Microbiome Analysis

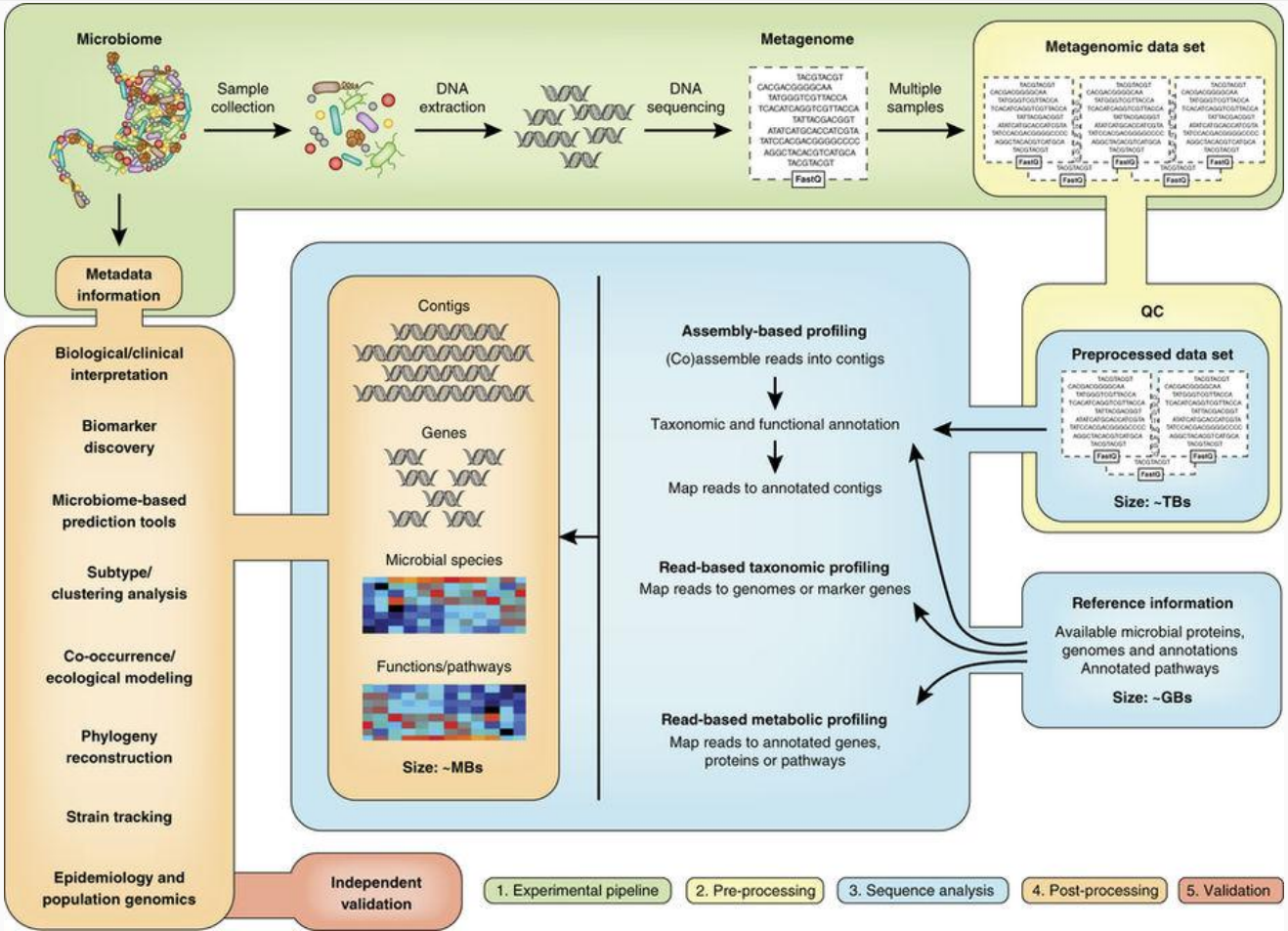
16S rRNA sequencing helps us understand get a better understanding of community **structure**...

...but metagenomics gives us a better understanding of the metabolic **potential** of a community ("function")...

...and next we need to focus on what the metabolic **activity** of a community (function)!

Shotgun Metagenomics Sequencing Bioinformatics

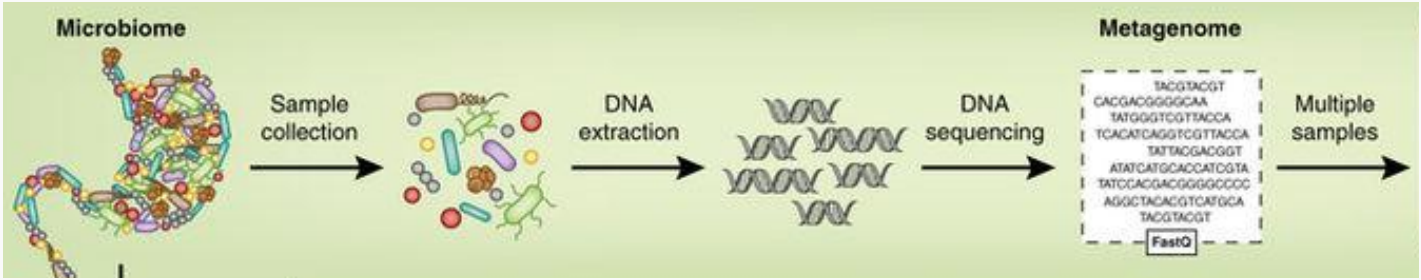
Metagenomics Workflow



Sample Collection

Sample collection and preservation should be standardized → Avoid systematic biases

- Can be difficult if samples are collected by different research groups
- Time in frozen storage may vary in longitudinal studies



1. Collection, processing, and sequencing of the samples

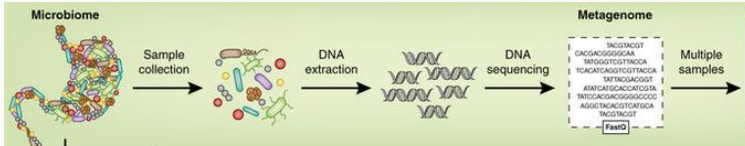
DNA Extraction

Very small amounts of DNA are (usually) sufficient for sequencing

- Any contamination can overwhelm the “real” signal → need to be as sterile as possible
- Should include a “blank” as a sequencing control

Need to ensure that there is sufficient microbial biomass for sequencing

- There are many enrichment methods available (if needed) → often introduce their own biases



1. Collection, processing, and sequencing of the samples

Whole-Genome Amplification

Advantages

Generates sufficient DNA for sequencing

- Even from tiny amounts of starting material

Can be applied directly to extracted environmental DNA

Can amplify DNA from the whole range of species present within a given sample

Limitations

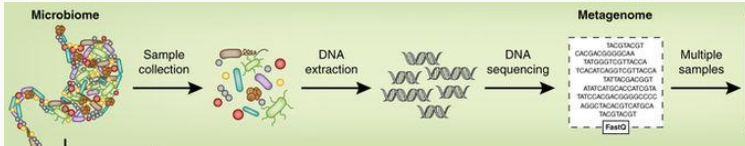
Can introduce significant biases

- Can skew resulting metagenomics profiles

Chimeric molecules can form

- Can confound assembly

Is unlikely to improve proportional abundance of DNA from a species of interest



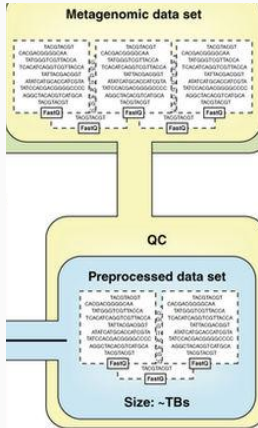
1. Collection, processing, and sequencing of the samples

Preprocessing

Quality Assessment & Trimming

- Remove adapters, low-quality bases, PCR primers (if used)
- Demultiplex using barcodes, discard reads without a barcode
- Filter out “foreign” DNA (*e.g.* human)

Tools: FastQC, Trimmomatic, Picard

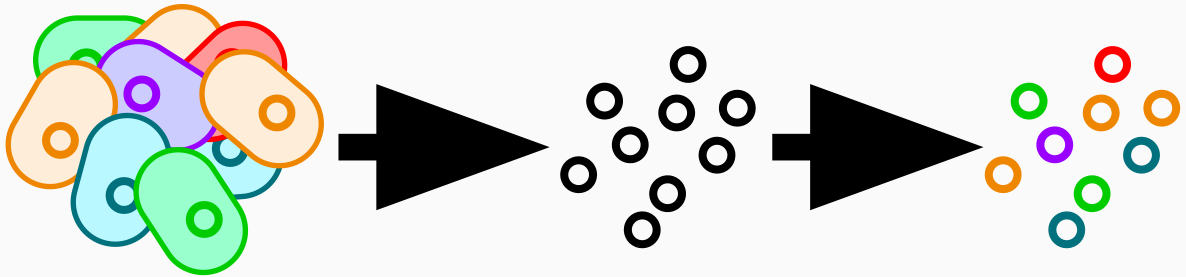


Sequence Analysis

Assembly: Putting short sequences together to reconstruct a longer, source sequence

Mapping: Locating where one short sequence is found in a longer sequence

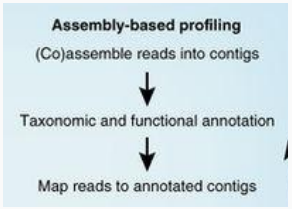
Binning: Identifying the “owner” of each anonymous DNA fragment → Classification of DNA sequences/reads/contigs

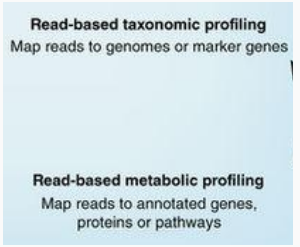


Sequence Analysis

Next-Gen Sequencing yields a large volume of data (FastQ files) in the form of short reads

Can either **assemble** the reads into **contigs**
OR
Can **map** the reads to reference databases



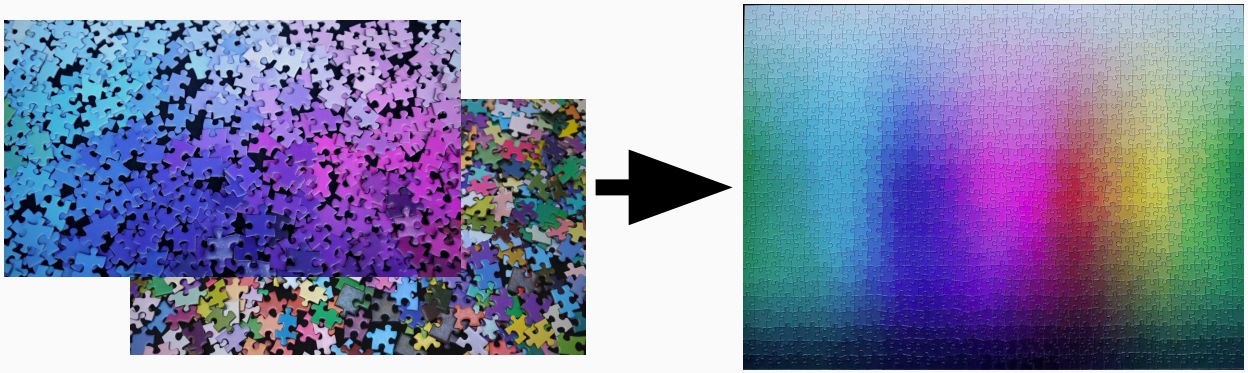


3. Sequence analysis to profile taxonomic, functional, and genomic features of the microbiome

Assembly

We can put together a picture of the community profile like a puzzle (or a set of puzzles with the pieces all mixed together)

- We can use a reference genome or we can perform assembly *de novo*



Similar to whole-genome assembly

Assembly

De Novo

Does not require reference genomes

Uses graph theory algorithms to assemble sequencing reads into contigs

Computationally demanding

Guided (Reference)

Requires reference genomes

Maps sequencing reads onto reference genomes to construct contigs joined *in silico* from individual reads

Limited by the quality and availability of reference genomes

Assembly: Terminology

Contigs: Contiguous DNA sequences assembles from shorter, overlapping sequencing reads

- **N50:** Weighted median contig size (metric) → Higher N50 may mean more mis-assemblies (metric was designed for single-genome assembly)
  - size to *x* Mbp, number to *x* Mbp

Scaffold: Merged contigs

MAG: Metagenomic Assembled Genome

Binning

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

- Can also cluster *de novo* and then use a database

Supervised: Use databases to label contigs into taxonomic classes

- Requires reference genomes...but many species are not sequenced

Unsupervised: Cluster contigs based on similarity

→ Both methods use a similarity metric

Binning: MAGs

Once contigs are binned, you can assemble them again into scaffolds/MAGs

- Assess completeness by examining if single-copy core genes (*e.g.* tRNA synthetases, ribosomal proteins) are present

Note: Binning used to occur on raw reads

Mapping

Taxonomic Profiling: Map reads to reference genomes

Metabolic Profiling: Map reads to annotated genes/proteins/pathways

→ Get “mixed bag”/“enzyme soup”

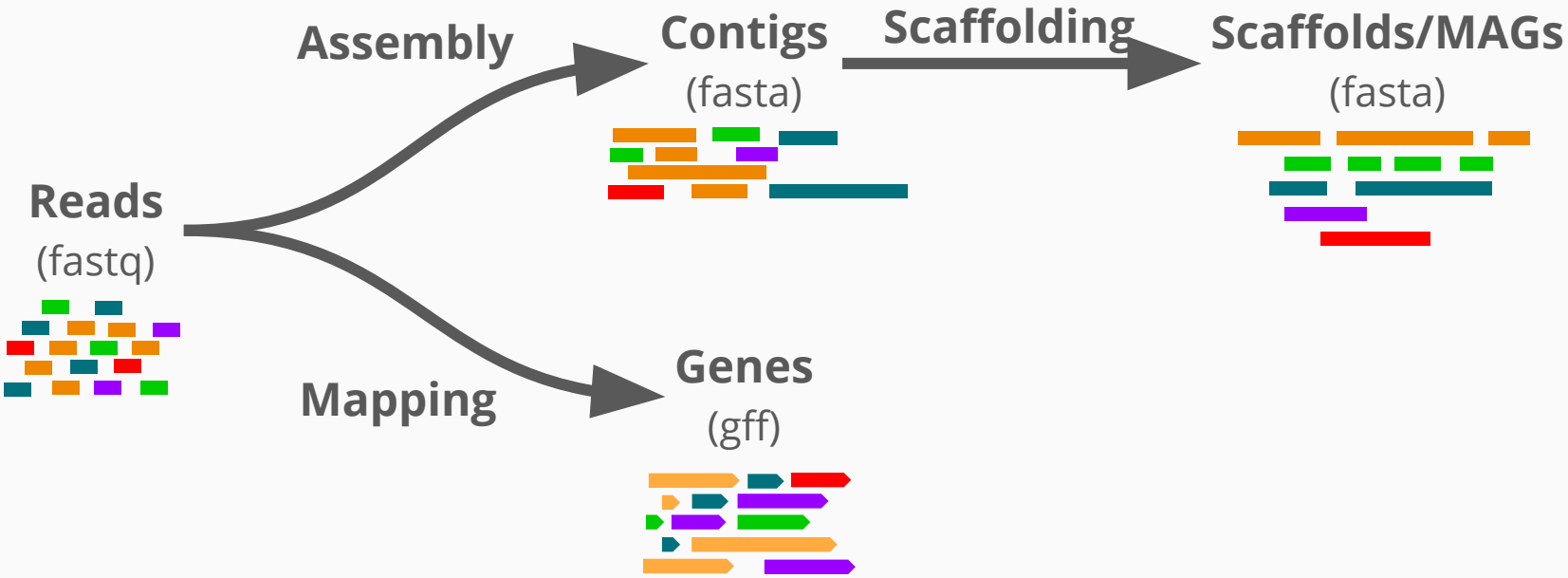
Mapping after Binning

Taxonomic Profiling: Map ~~reads~~ contigs/scaffolds to reference genomes

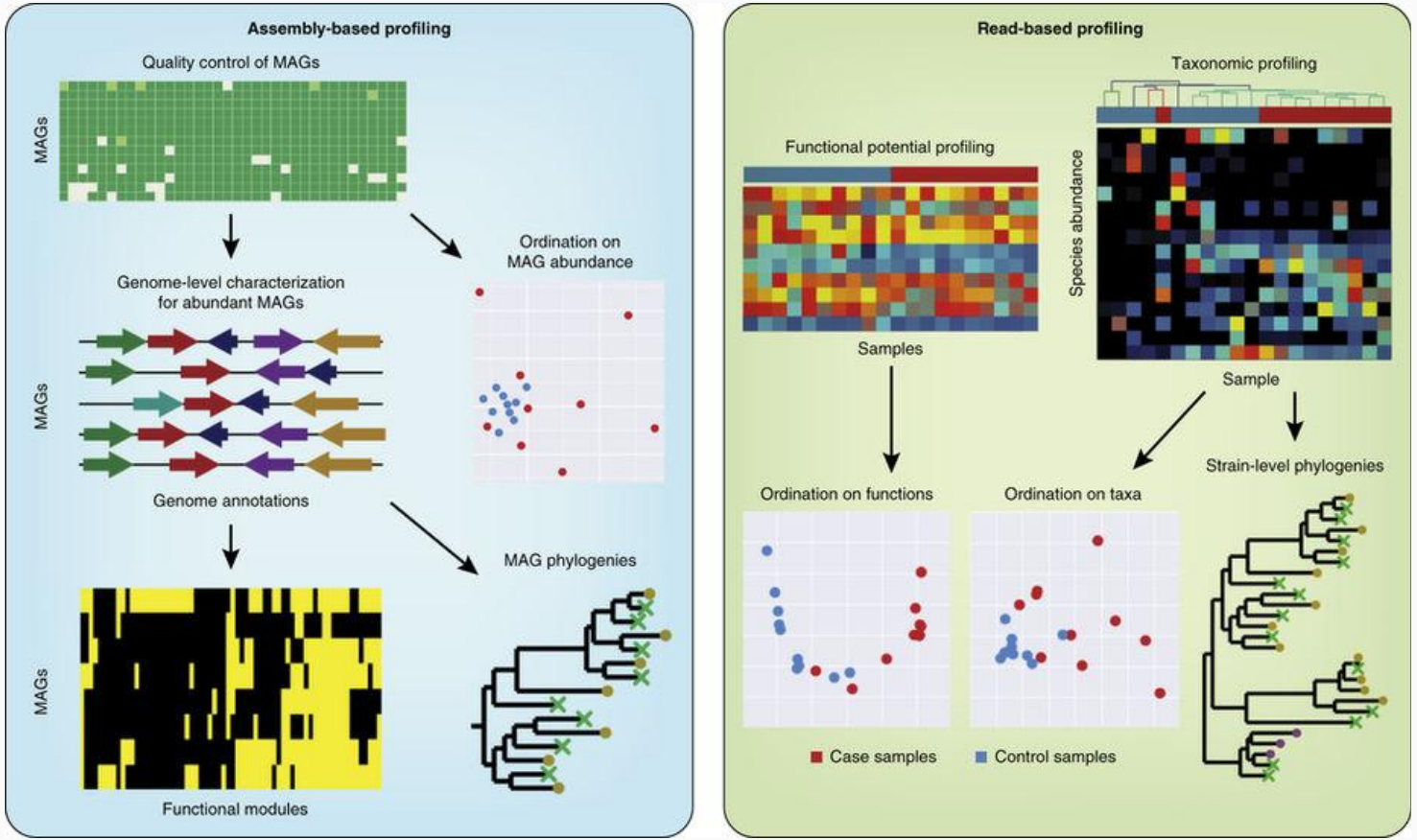
Metabolic Profiling: Map ~~reads~~ contigs/scaffolds to annotated genes/proteins/pathways

→ Know which contig/scaffold/species contains which genes/proteins/pathways

Sequence Analysis



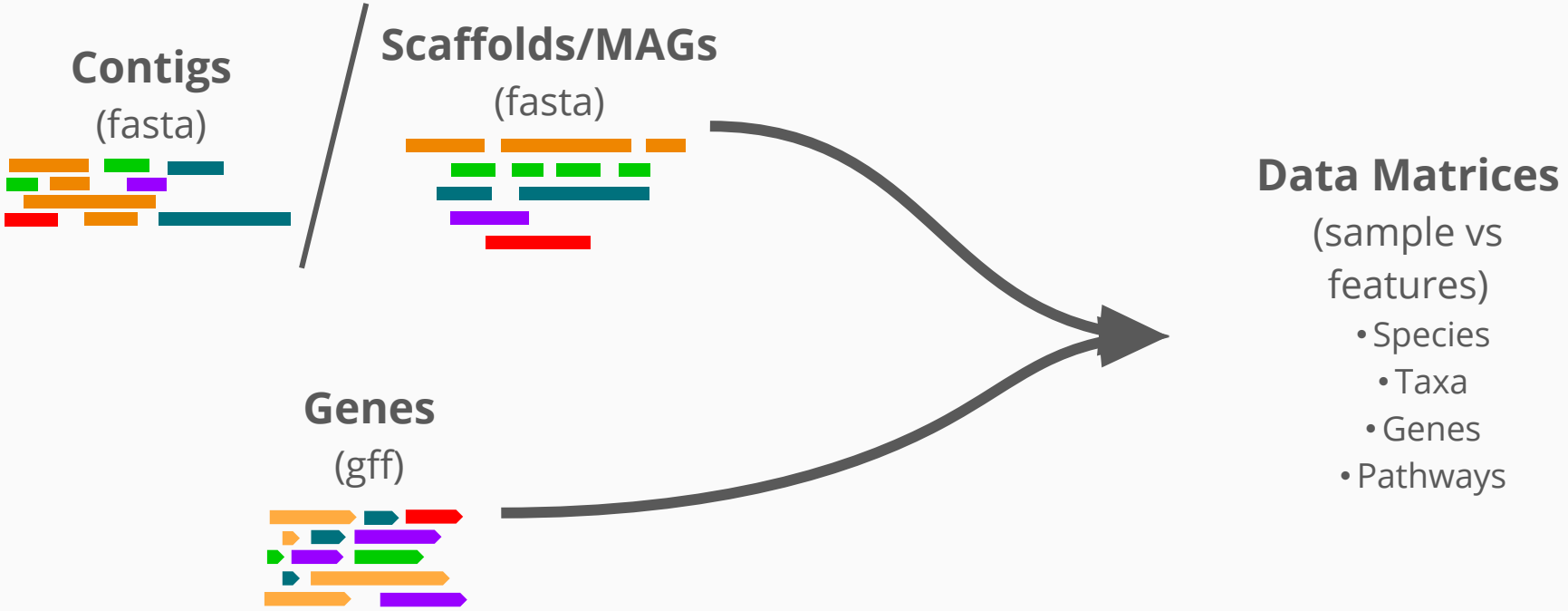
Sequence Analysis



3. Sequence analysis to profile taxonomic, functional, and genomic features of the microbiome

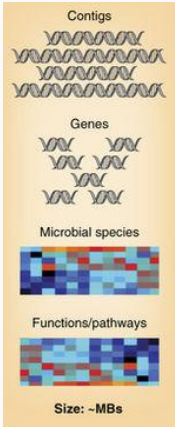
Quince, Walker, Simpson, Loman, Segata (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*

Post-Processing Analysis



Post-Processing Analysis

- Heat maps
  - Species/taxa
  - Functions/pathways
- Clustering
- Correlations
- Co-occurrence
- Phylogeny



Assemblers

Tools: Meta-IDBA, MetaVelvet, IDBA-UD, MetaSPAdes, MEGAHIT, Ray-Meta, SOAPdenovo

- Available memory is usually limiting

What's best? Depends on biological factors (*e.g.* underlying community structure) and technical factors (*e.g.* sequence platform, coverage)

- Try a few and see what is "best"

Assembly result? ~~Genomes~~! **Contigs**!

More Tools

Binners/Classifiers: MetaPhlAn, Kraken, Ray-Meta, MetaBAT, CONCOCT, Canopy
Bin Quality Assessment: CheckM
Gene/Functional Analysis: MetaProdigal
Gene/Functional Databases: PFAM, SEED, KEGG, CAZy
Statistical Analysis: HUMAnN2
Pipelines: MG-RAST, MEGAN, IMG/MER, mothur, QIIME

**Non-inclusive – there are MANY more tools out there.

Limitations and Opportunities

Metagenomics only looks at the **gene sequences** that encode proteins or functional RNAs → Tells you what the microbes a functionally **capable** of (genomic/metabolic **potential**)

- Need to examine RNA transcripts (metatranscriptomics) and/or translated proteins (metaproteomics) to see what microbes are actually **doing**

Limitations and Opportunities

Many genes are not annotated → we don’t know what protein they encode

- Our understanding of microbial communities is partial, based on what we can infer from existing knowledge (aka what is well-characterized and in databases)
- Lots of stuff to learn! But we still need to do the (expensive and low-throughput) gene-specific functional studies

Limitations and Opportunities

Available microbial genomes are biased towards model organisms, pathogens, and easily-cultivable bacteria

- All metagenomics computational tools rely on available genomes (databases), and are therefore affected by the biases in the reference sequence resources

Some reads may be unused even after assembly

- Still don't know who these microbes are (or if they are just noise)

Are we sequencing the live microbes...or are we sequencing dead or damaged cells?

Backup

Reminder…

…metagenomics is an extension of many things you have already learned!

Genomics used to be computationally difficult, and now that's metagenomics!

- Still developing tools/algorithms for data analysis (especially assembly and mapping!)

Why Sequencing?

Microbes are often difficult to culture

- (Ideally) Can study **all** microbes, not just those you can culture

Want to understand the **structure** and **function** of microbial communities

It has become (relatively) cheap

Things to Keep in Mind

Sequencing platform

- Error rate, biases, read length, noise

(Amplification process)

- Error rate, biases, choice of primer, DNA template concentration, PCR cycle number, introduction of chimeras

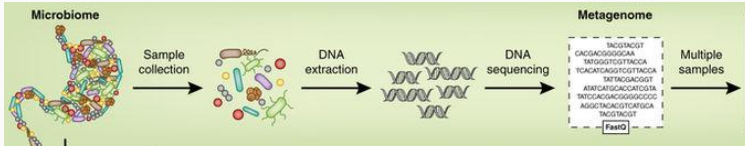
Coverage/Depth

DNA Extraction

Need to ensure that the DNA extraction methodology is stringent enough to extract DNA for all cell types in the sample → Avoid bias

- Should be effective for diverse microbial taxa → otherwise biased for easy-to-lyse microbes
- Vigorous extraction techniques can result in shortened DNA fragments → can contribute to DNA loss

Note: All extracted DNA is randomly sheared into desired fragment sizes



1. Collection, processing, and sequencing of the samples

Sequence Analysis

	Assembly-Based	Read-Based (Mapping)
Comprehensiveness	Can construct multiple whole genomes, but only for organisms with enough coverage to be assembled and binned.	Can provide an aggregate picture of community function or structure, but is based only on the fraction of reads that map effectively to reference databases.
Community Complexity	In complex communities, only a fraction of the genomes can be resolved by assembly.	Can deal with communities of arbitrary complexity given sufficient sequencing depth and satisfactory reference database coverage
Novelty	Can resolve genomes of entirely novel organisms with no sequenced relatives.	Cannot resolve organisms for which genomes of close relatives are unknown.
Computational Burden	Requires computationally costly assembly, mapping, and binning.	Can be performed efficiently, enabling large meta-analyses.

3. Sequence analysis to profile taxonomic, functional, and genomic features of the microbiome

Quince, Walker, Simpson, Loman, Segata (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*

Sequence Analysis

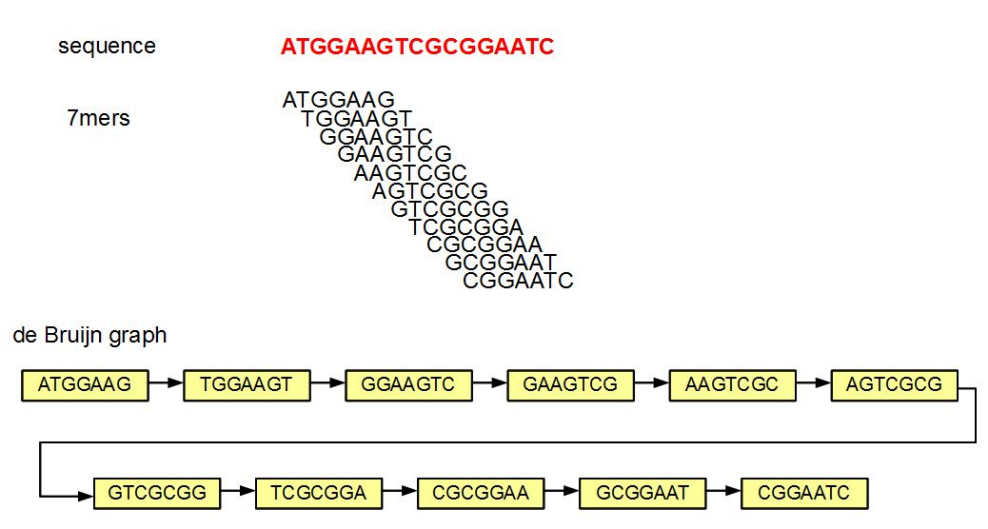
	Assembly-Based	Read-Based (Mapping)
Genome-Resolved Metabolism	Can link metabolism to phylogeny through completely assembled genomes, even for novel diversity.	Can typically resolve only the aggregate metabolism of the community, and links with phylogeny are only possible in the context of known reference genomes.
Expert Manual Supervision	Manual curation required for accurate binning and scaffolding and for misassembly detection.	Usually does not require manual curation, but selection of reference genomes to use could involve human supervision.
Integration with Microbial Genomics	Assemblies can be fed into microbial genomic pipelines designed for analysis of genomes from pure cultured isolates.	Obtained profiles cannot be directly put into the context of genomes derived from pure cultured isolates.

3. Sequence analysis to profile taxonomic, functional, and genomic features of the microbiome

Quince, Walker, Simpson, Loman, Segata (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*

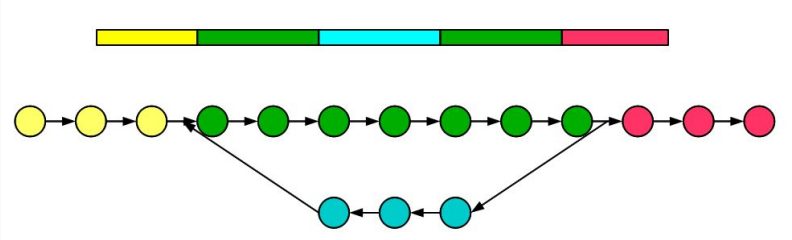
Assembly: *de novo*

Break each sequencing read into overlapping subsequences of a fixed length k ($k = 7$)



Set of overlapping "*k*-mers" define the vertices and edges of the graph

Assembler finds the path through the graph that reconstructs the genome(s)



Repetitive regions make it difficult to resolve the original sequence

Assembly: Guided (Reference)

Align reads to a (database) of reference genomes

- Similarity scores
- Sometimes only a minority of reads can be aligned

Assembly: Unique Challenges

When assembling a single genome, we typically assume that sequence coverage is ***approximately* uniform** across the genome

- Find repeat copies
- Distinguish sequencing errors from the “real” sequence
- Identify allelic variation

...but the coverage of each community member depends on the **abundance** of its genome in the community

Assembly: Unique Challenges

Communities are more biologically complex than individual microbes:

- Reads come from multiple species → Are near-identical/identical reads/configs from 1 or more species?

The presence of different strains of the same species can result in fragmented reconstructions.

Some DNA segments are repeated within the same organism, or shared between distinct organisms.

Assembly: Unique Challenges

So if the coverage of each community member depends on its **abundance**...

...then low-abundance genomes may end up fragmented/incomplete

Coverage/Depth will determine how well you can characterize low-abundance community members

(Need “enough” overlapping reads)

Trade-off between:

- Recovering low-abundance genomes
- Obtaining long, accurate contigs for high-abundance genomes

Binning

We don’t have prior knowledge about what species (or even how many) are in a sample…

…and we also don’t have prior knowledge about which contig derives from which genome…

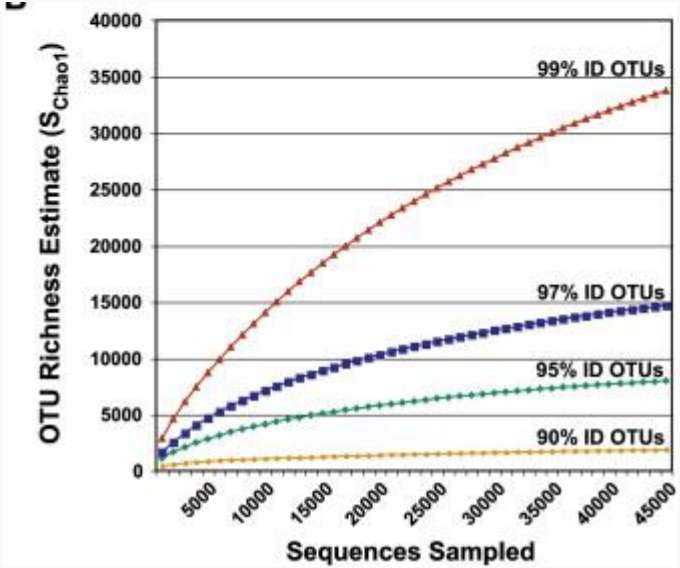
…so we need to group contigs (**bin**)

Binning

16S rRNA Sequencing:

Cutoffs: What percent sequence identify should you use?

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



Binning

Composition-Based (Unsupervised)

Does not require reference genomes

Cluster by contig sequence composition

- k-mer (usually tetramer, $k = 4$) frequencies
- GC content

Performs poorly on short reads

Homology-Based (Supervised)

Requires reference genomes

Cluster by gene homology

- Similarity to known (marker) genes

Struggles to be discriminative when there are closely-related species

Sequence Analysis: Challenges

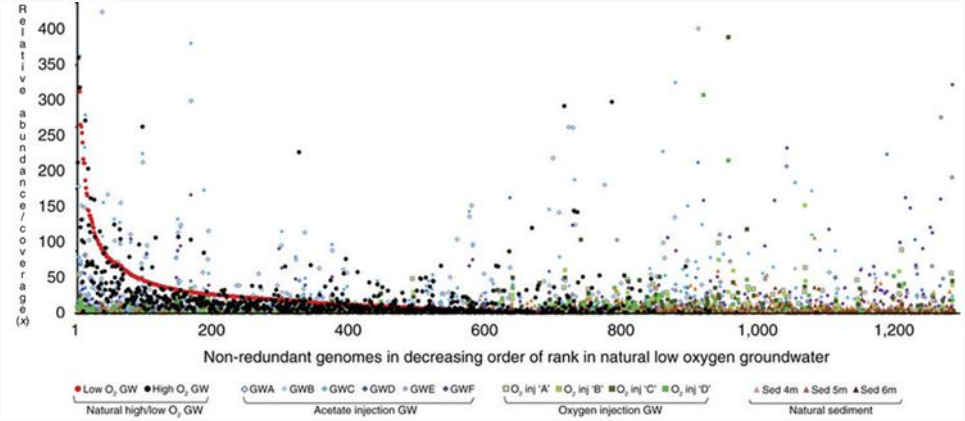
The main limiting factor in profiling the metabolic potential is the lack of annotations

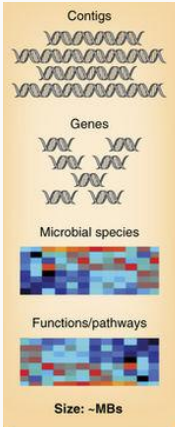
- Biased towards highly conserved pathways and housekeeping functions → may explain why function is consistent across different samples and environments event when taxonomy varies

Post-Processing Analysis: Challenges

Taxonomic and functional profiles are proportional/compositional

Abundances are (log-normal) long-tailed distributions





Things to Keep in Mind

You are taking a **sample** from a **population**

- There will be **variation** between samples from the same population
- Need to ensure that the study has enough statistical **power** to detect differences

Controls can be difficult to obtain

- Collect as much metadata as possible
  - Clinical Samples: Gender, age, antibiotic/medication use, location, diet, *etc.*
  - Environmental Samples: Location, season, pH, temperature, *etc.*
- Collect longitudinal data when possible

Things That Should Happen (But Don't)

Technical replicates to assess variability
Blank controls to assess library preparation and sequencing biases/error/contamination
Corrections for confounding factors (*e.g.* batch effects)

...although uncommon now, will hopefully become common as costs decrease

Limitations and Opportunities

Quantitative features are normalized (relative abundance)

- Make sure you are using the appropriate statistical/analysis tools for normalized data!
- Can observe false correlations otherwise...

	Sample 1	Sample 2
Organism 1	10	20
Organism 2	2	2
Organism 3	3	3

	Sample 1	Sample 2
Organism 1	0.67	0.80
Organism 2	0.13	0.08
Organism 3	0.20	0.12

Microbiome Analysis

16S rRNA sequencing: **structure**

Metagenomics: **metabolic *potential***

Metatranscriptomics/Metabolomics:
metabolic *activity*