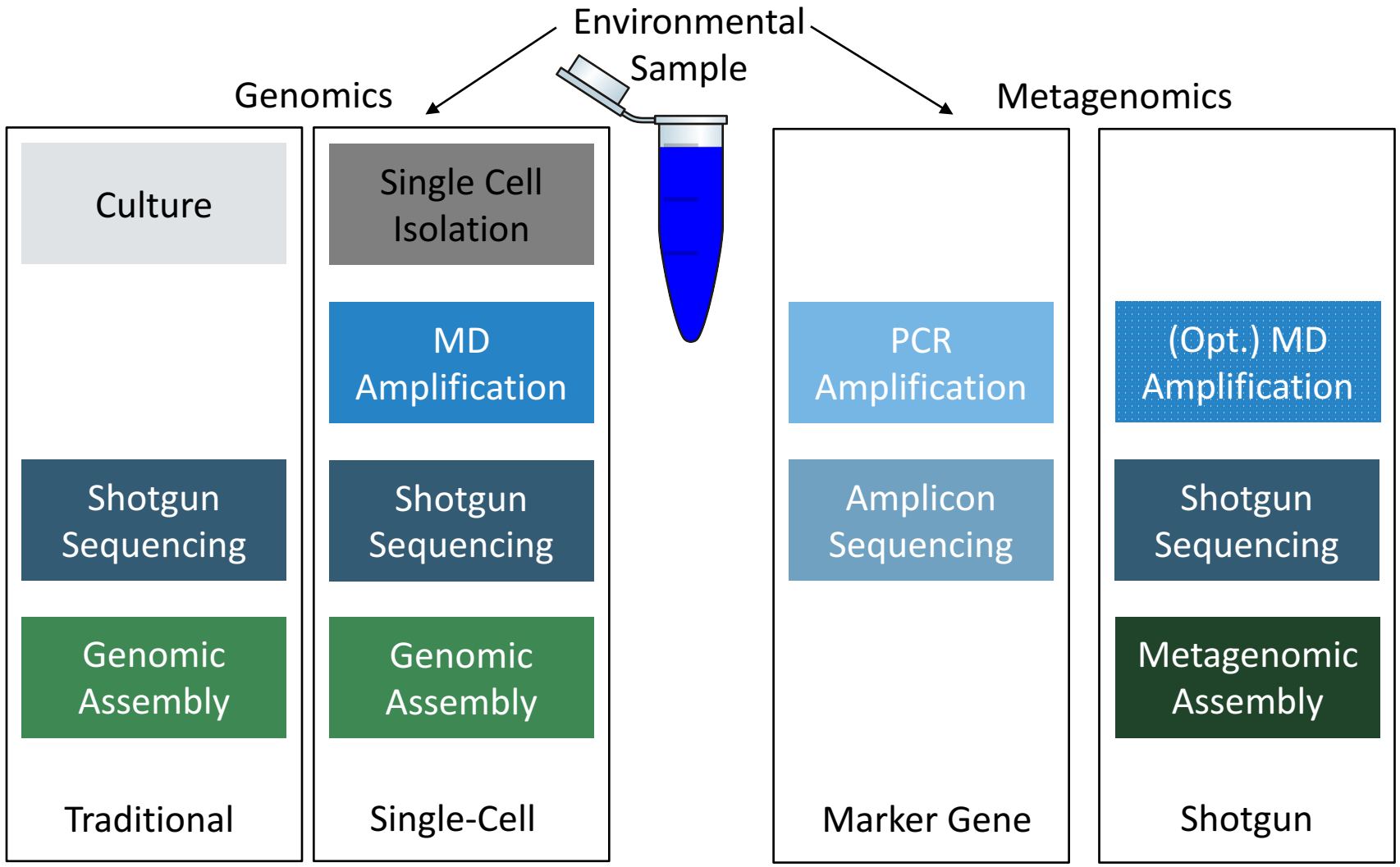
Choosing the Right Strategies for Conducting a Microbiome Study

Amanda Birmingham

Center for Computational Biology & Bioinformatics
University of California at San Diego

Microbial Genomics & Metagenomics

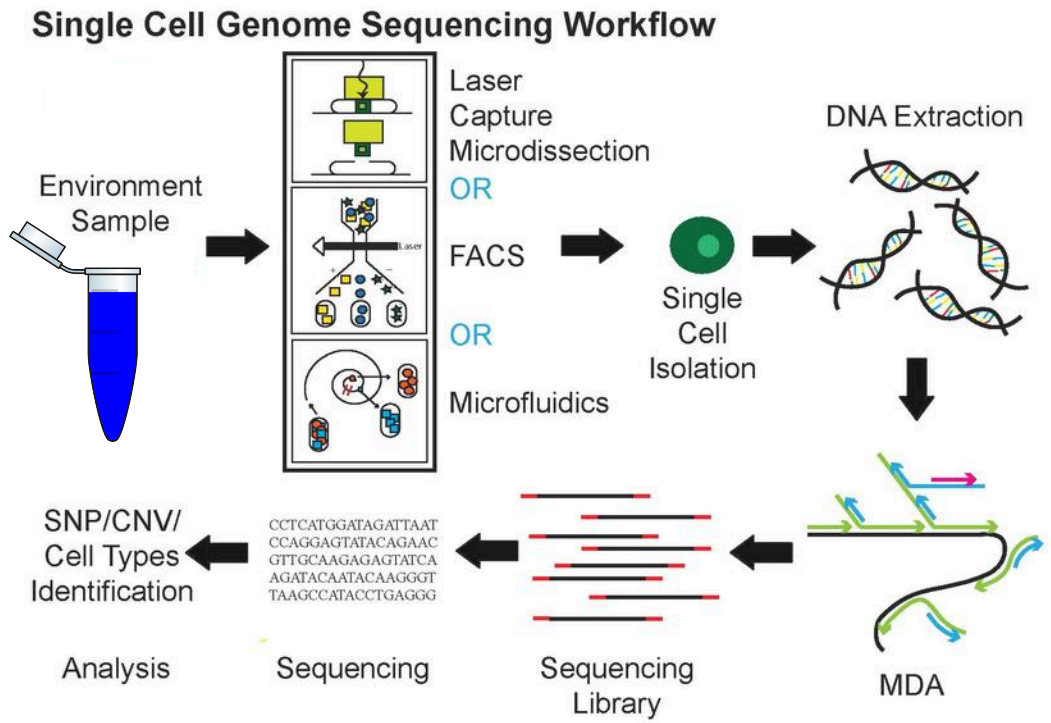


When to Use Traditional Genomics

- When you have:
  - Only one or a couple of microbes of interest
  - And they are culturable
  - And you care about their genome sequences, not their abundance in the sample(s)
- The good news:
  - This approach can identify plasmids associated with the bacterial chromosome
  - Software for *de novo* (from scratch) assembly of short reads into genomes is getting better
    - With high coverage, existing tools can sometimes reach ~98% completeness
- The bad news:
  - Repeat regions (like those from transposons) are really hard to assemble
    - 100% complete reference genomes still require specialized skills and protracted effort
  - Most microbes aren't easily culturable in the lab

When to Use Single-Cell Genomics

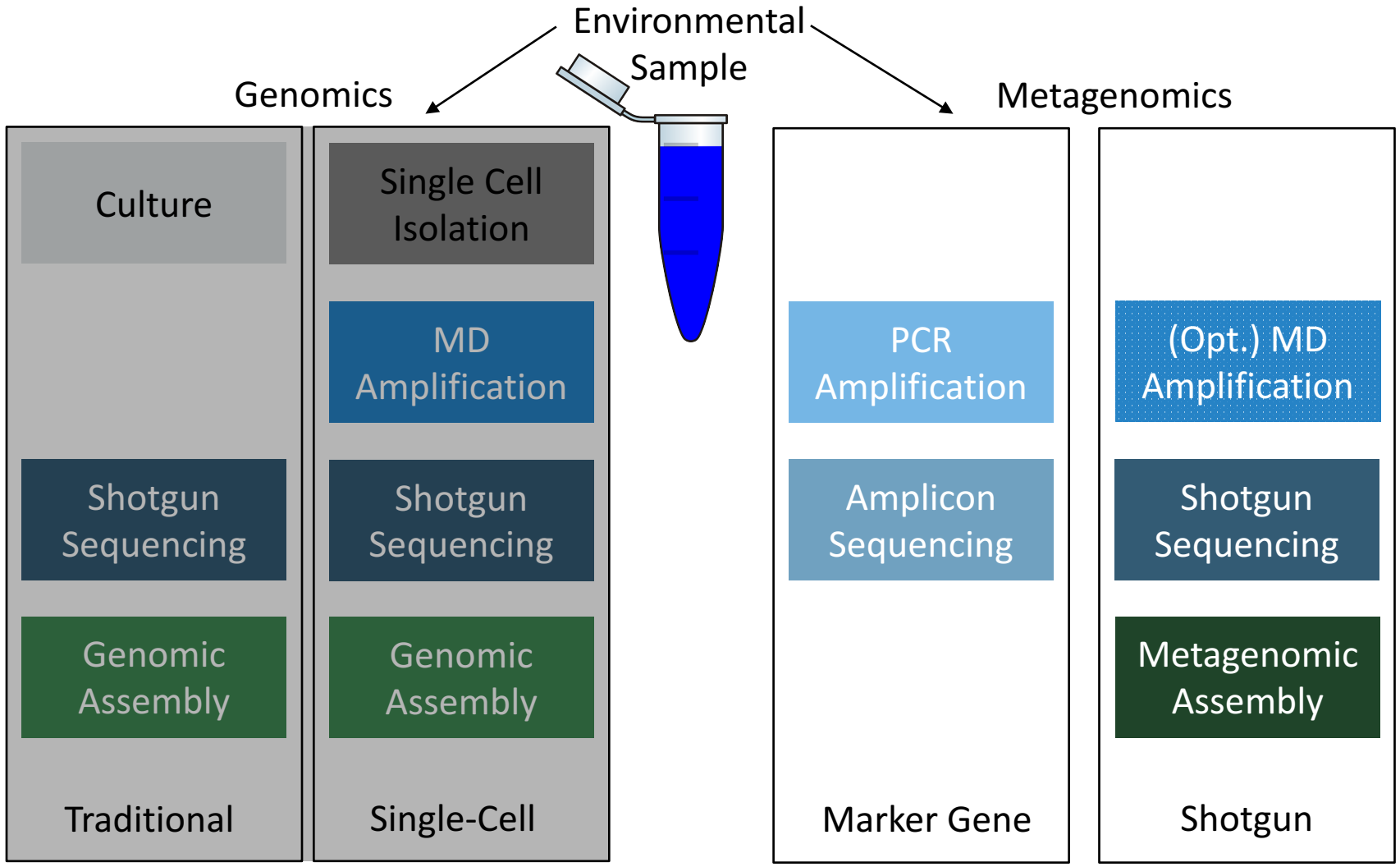
- When you need reference genome(s) and can't culture, e.g. functional analysis of soil microbes
  - Community is EXTREMELY heterogeneous (most common organism ~1% of total); shotgun won't assemble
  - Community members are too hard to culture (~1% grow in standard medium)
  - So: consider single-cell genomics to generate a reference database, then shotgun for abundance information



https://commons.wikimedia.org/wiki/File:Single_Cell_Genome_Sequencing_Workflow.pdf

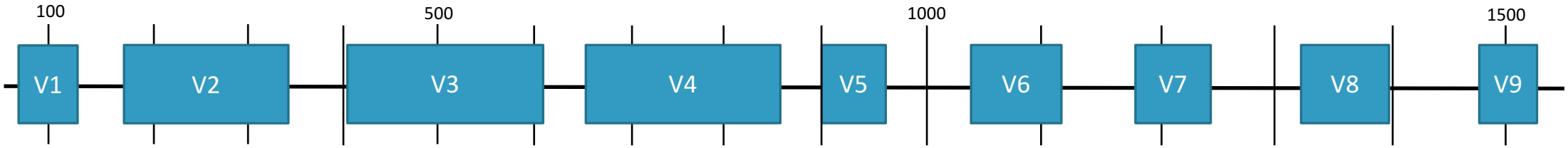
- When you have the time, money, and equipment
  - Having single-cell-level understanding of microbial communities is wonderful if we can get it!
- The good news:
  - See traditional genomics, plus … this is really possible!
- The bad news:
  - Getting single cells is expensive and/or time-consuming
  - This gets around culture issues but not assembly ones
  - Lots of amplification is required, with potential for bias

Microbial Genomics & Metagenomics



Marker Gene Metagenomics Basics

- Approach: PCR amplicons of a conserved constitutive gene (a "marker gene") to determine identity and abundance of microbes present
  - Usually the “conserved constitutive gene” of choice is 16S rRNA
    - The small sub-unit (SSU) of bacteria’s ribosome
    - Excludes eukaryotic DNA as eukaryotes’ SSU is 18S



- 16S rRNA is widely conserved across bacteria and archaea, providing shared primer sites
  - Yet also has 9 hypervariable regions: can be used to id different “species” and build phylogenetic trees
    - Not really species but Operational Taxonomic Units (OTUs)—groups of organisms defined *only* by sequence similarity
- Can’t study fungi with 16S (they don’t have it) nor 18S (evolves too slowly)
  - Internal transcribed spacer (ITS) is standard fungi marker gene

When to Use Marker Gene Metagenomics

- When your sample is MOSTLY made up of host DNA, e.g. tumor samples
  - Shotgun reads will also be mostly host DNA, with few left over for the microbes
  - Use 16S rRNA instead, as the primers exclude eukaryotic DNA from amplification
- The good news:
  - Target gene studies are *slightly* cheaper to prep and sequence than shotgun ones
  - Analysis software is mature, and many studies can be analyzed on a laptop
  - Known taxa can be detected with very low (100s of reads) sequence depth
- The bad news
  - No target gene distinguishes all microbes well
  - No other genome information (outside target gene) is captured

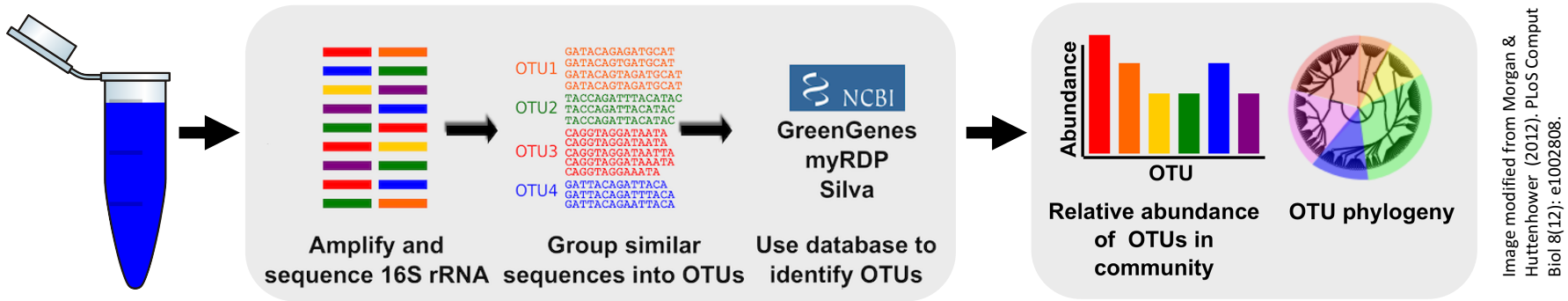
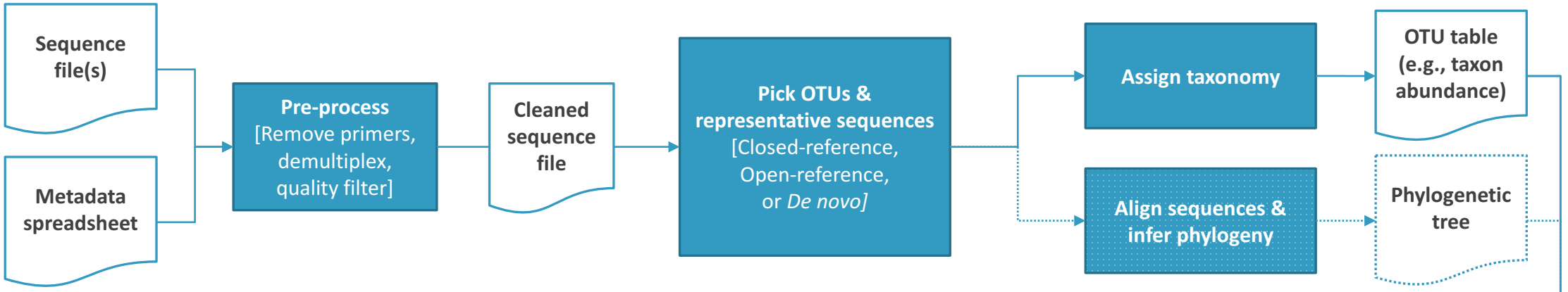
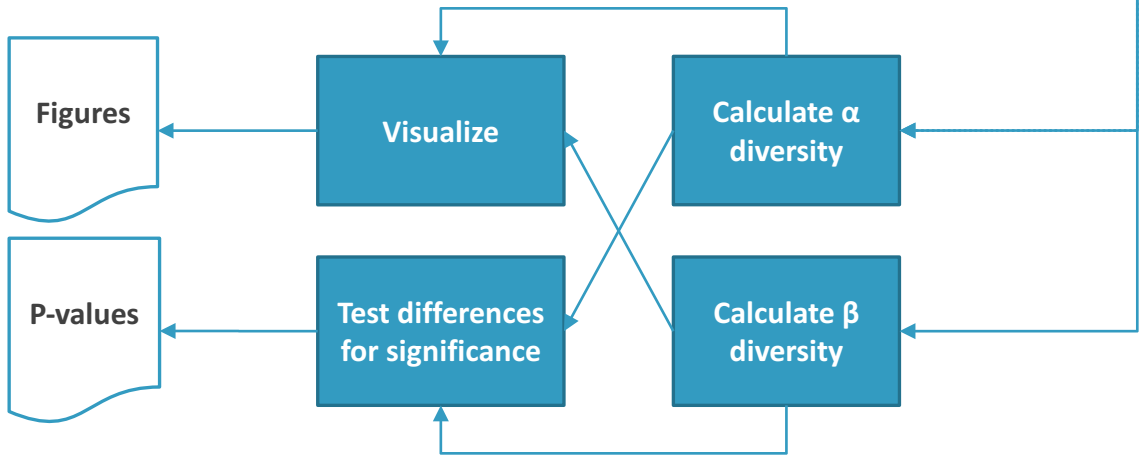


Image modified from Morgan & Huttenhower (2012). PLoS Comput Biol 8(12): e1002808.

Marker Gene Analysis Workflow



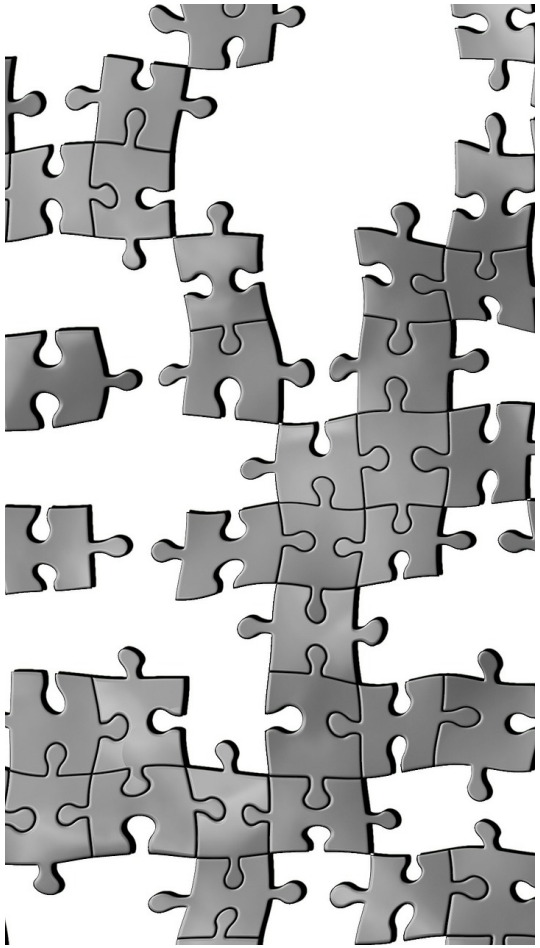
- Most critical analysis choices:
  - What kind of OTU picking to perform
    - Closed reference is fast
    - Open reference is usually a good compromise
    - *De novo* is necessary if no reference db available
  - What α and β metrics to pick
    - Some are phylogenetically aware, some aren't



Common Issues in Marker Gene Studies

- Neglecting metadata
  - Analysis can not test for effects of, or discard bias from, features you didn't record!
- Picking novel 16S primers—not all created equal
  - Earth Microbiome Project recommends 515f-806r primers, error-correcting barcodes
- Not taking precautions to support amplicon sequencing
  - Some Illumina machines require high PhiX, low cluster density
- Selecting an inappropriate reference database
  - E.g., Greengenes (16S) reference database when sequencing ITS
- Expecting species-level taxonomy calls
  - Most OTUs only specified to family or genus level
- Using inappropriate statistical tests
  - Taxa abundance requires a compositionality-aware test like ANCOM
  - Differences in β diversity distances across groups requires test like PERMANOVA, not ANOVA

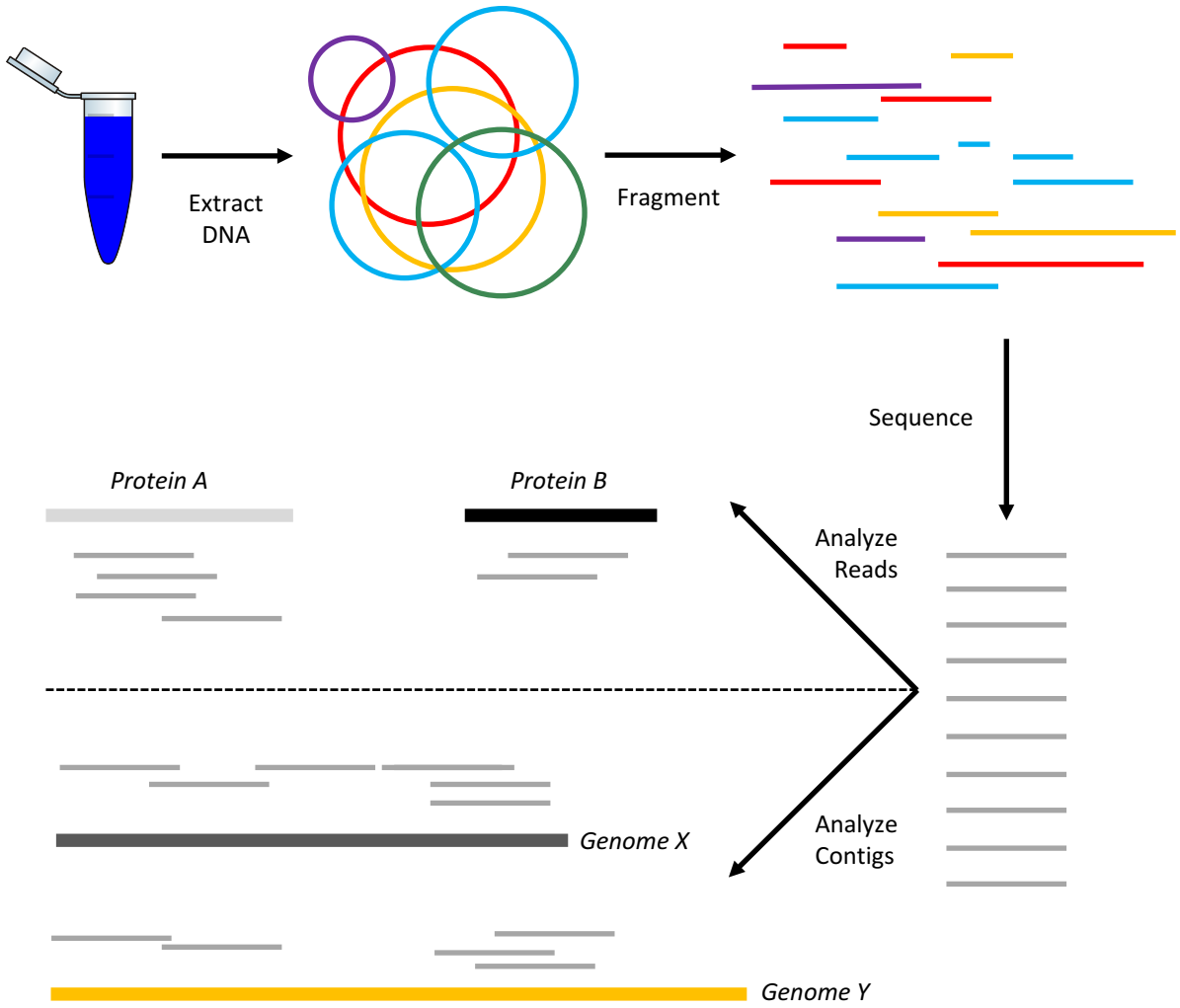
Shotgun Metagenomics Basics



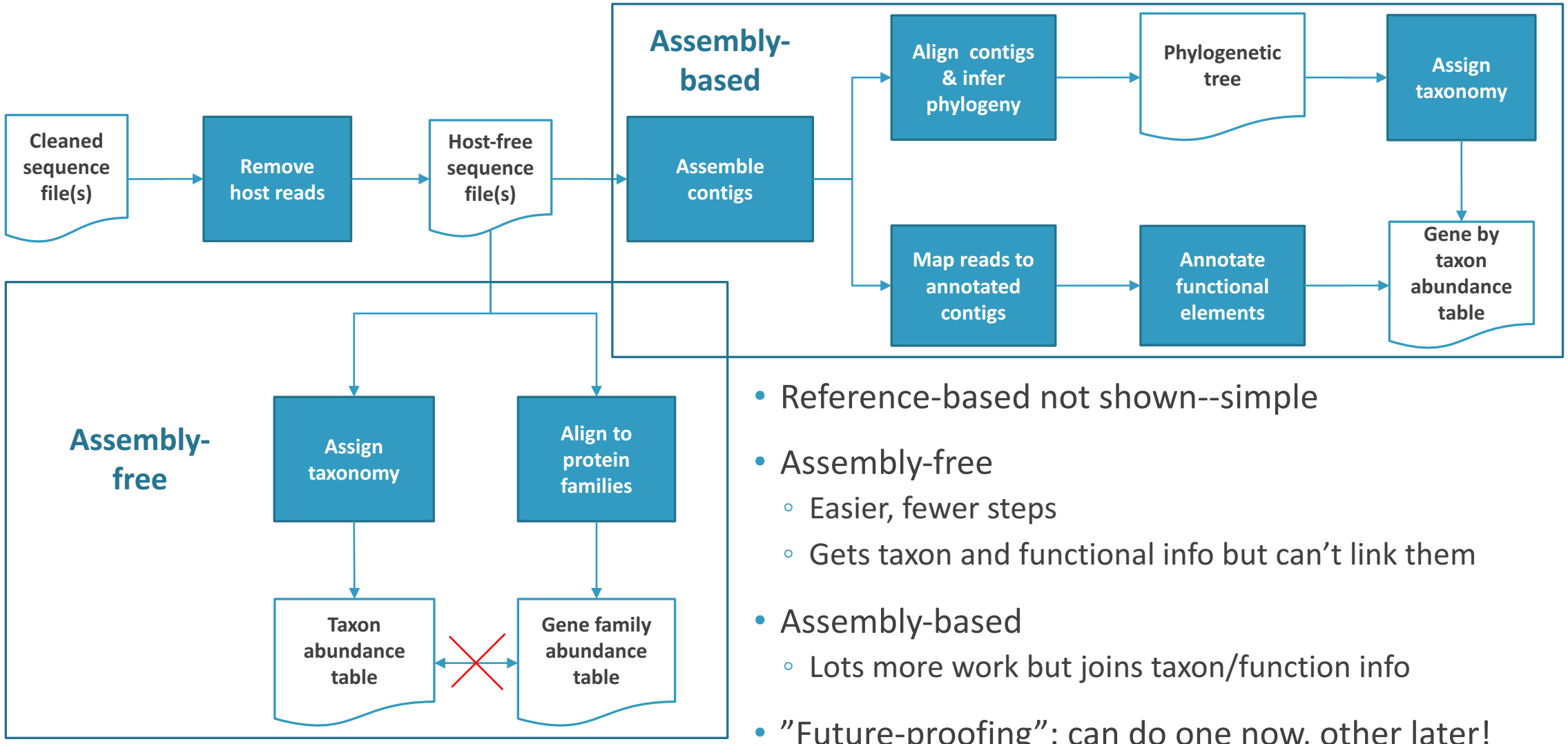
- Just fragment and sequence, try to figure out what it means in analysis!
- Reference-based
  - Map shotgun reads to database of known, complete reference genomes
  - Find identity and abundance
    - Analogous to approach for single-organism RNASeq—but more complex
  - Usually not feasible: too few references known
    - Exceptions: human gut, mouth, vagina
- Assembly-free
  - Map reads to database of known marker genes to guess taxonomic identity
  - Translate reads and map to protein family database to find functionality
- Assembly-based
  - Assemble reads into (multiple) genomes—or at least contigs
  - Place contigs in phylogeny to find taxonomic identity
  - Detect genes, lncRNAs, operons: find functionality linked to identity

When to Use Shotgun Metagenomics

- When target genes can't tell your microbes apart
  - E.g., *Sporosarcina psychrophila* & *Bacillus anthracis*
- When you want microscopic eukaryotes too
  - Protists, fungi, algae
- When you want to see functional detail
- The good news:
  - Sequencing has gotten cheaper, so we can do more
  - Cloud computing, better aligners, and better assemblers make analysis possible for biologists
- The bad news:
  - Can't associate plasmids with hosts
  - Read analysis is limited, contig analysis is hard
  - Data is large and analysis tools are still maturing



Shotgun Analysis Workflows



- Reference-based not shown--simple
- Assembly-free
  - Easier, fewer steps
  - Gets taxon and functional info but can't link them
- Assembly-based
  - Lots more work but joins taxon/function info
- "Future-proofing": can do one now, other later!

Common Errors in Shotgun Studies

- Not having analysis and storage plan
  - Shotgun sequencing data can easily be 10-50 Gb *compressed*
    - When uncompressed files are over 100 Gb, and analysis creates intermediate versions, doesn't take long to fill your hard drive
  - Both assembly-free and assembly-based approaches require lots of alignment
    - This is time-consuming on 10s to 100s of millions of reads, even with fast aligners
  - Assembly-based approaches are real memory hogs
- Failing to extract host reads
  - Unlike 16S, shotgun amplifies host DNA too
  - Must be aligned to host genome and removed
    - This is a big problem if you don't **have** a host genome
- Not filtering amplified duplicates
  - Amplifying low-abundance inputs creates uninformative duplicates
  - These can swamp real reads
- Throwing away raw reads
  - "Future-proofing" only works if you have the original data to reanalyze later!

Conclusions

- Microbiome research reinforces that life is inherently interconnected and interdependent
- Metagenomic studies allow insight into a whole interdependent community at once
- 16S metagenomics is a tried-and-true workhorse
  - But it is about ready to be put out to pasture for most experiments
- Shotgun metagenomics is no longer “bleeding edge”
  - But analysis stage can still cause some pain!
  - Assembly-free analysis is easier
  - Assembly-based gets us closer to what we really want to know
- Shotgun metagenomics offers best chance of “future-proofing” data collection