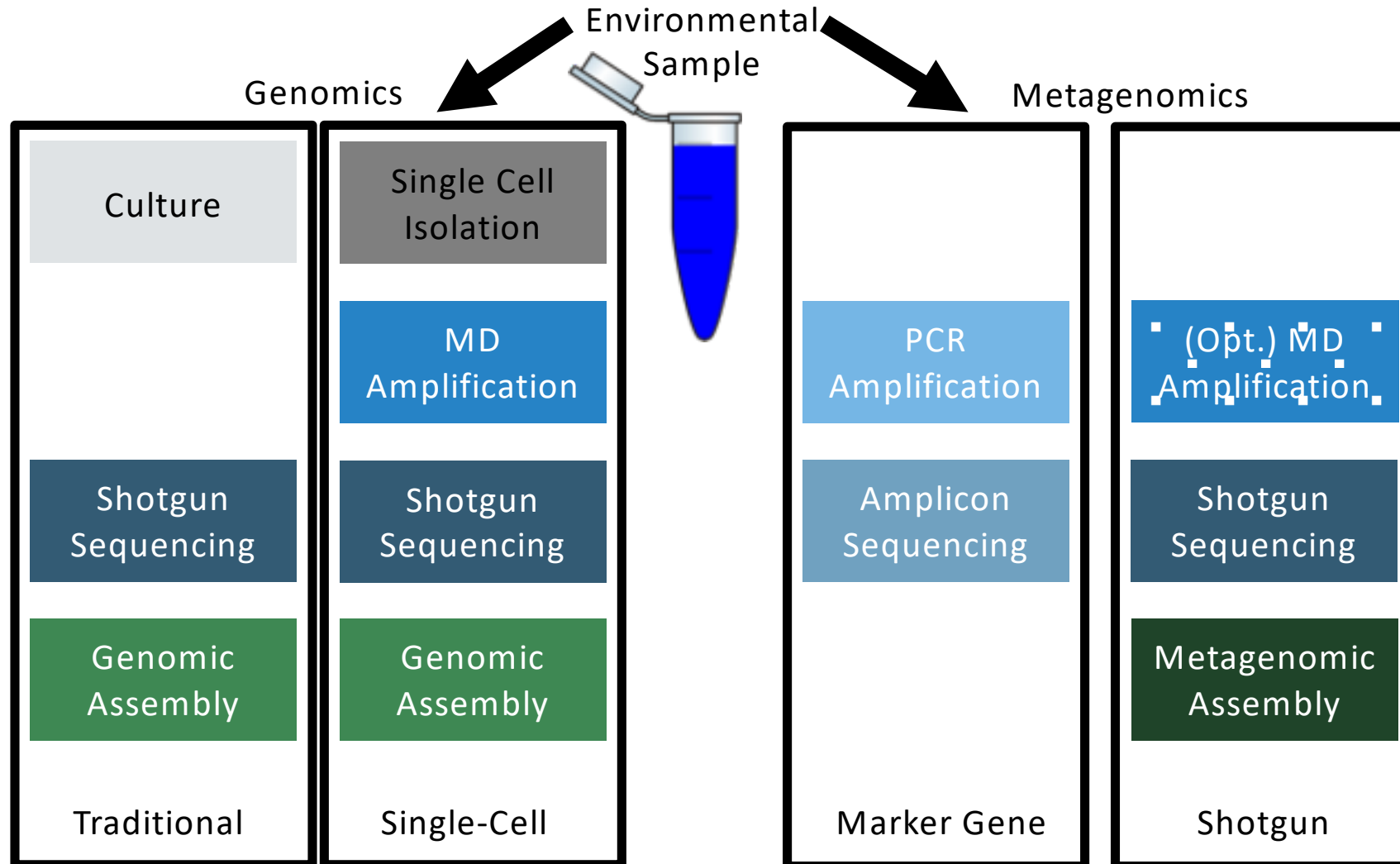


Choosing the Right Strategies for Conducting a Microbiome Study

Amanda Birmingham

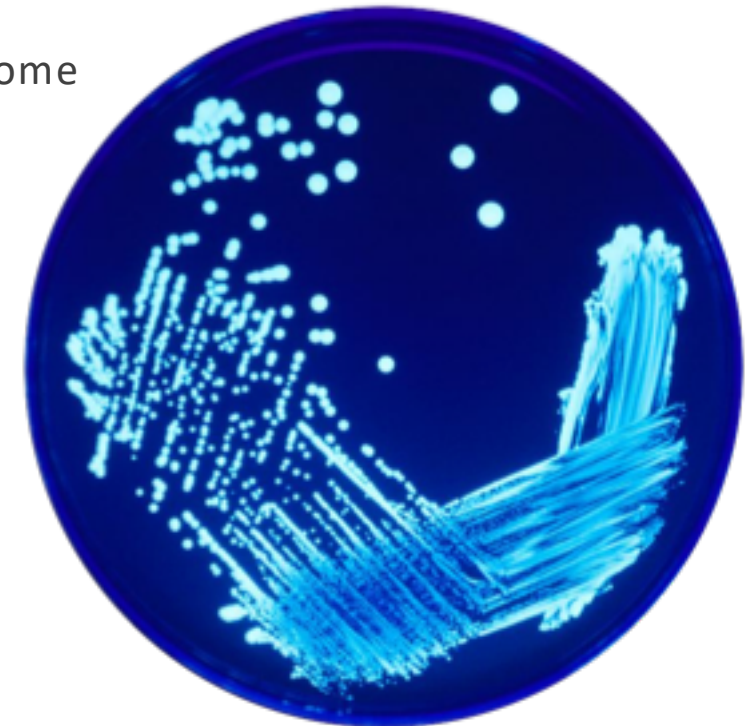
Center for Computational Biology & Bioinformatics
University of California at San Diego

Microbial Genomics & Metagenomics

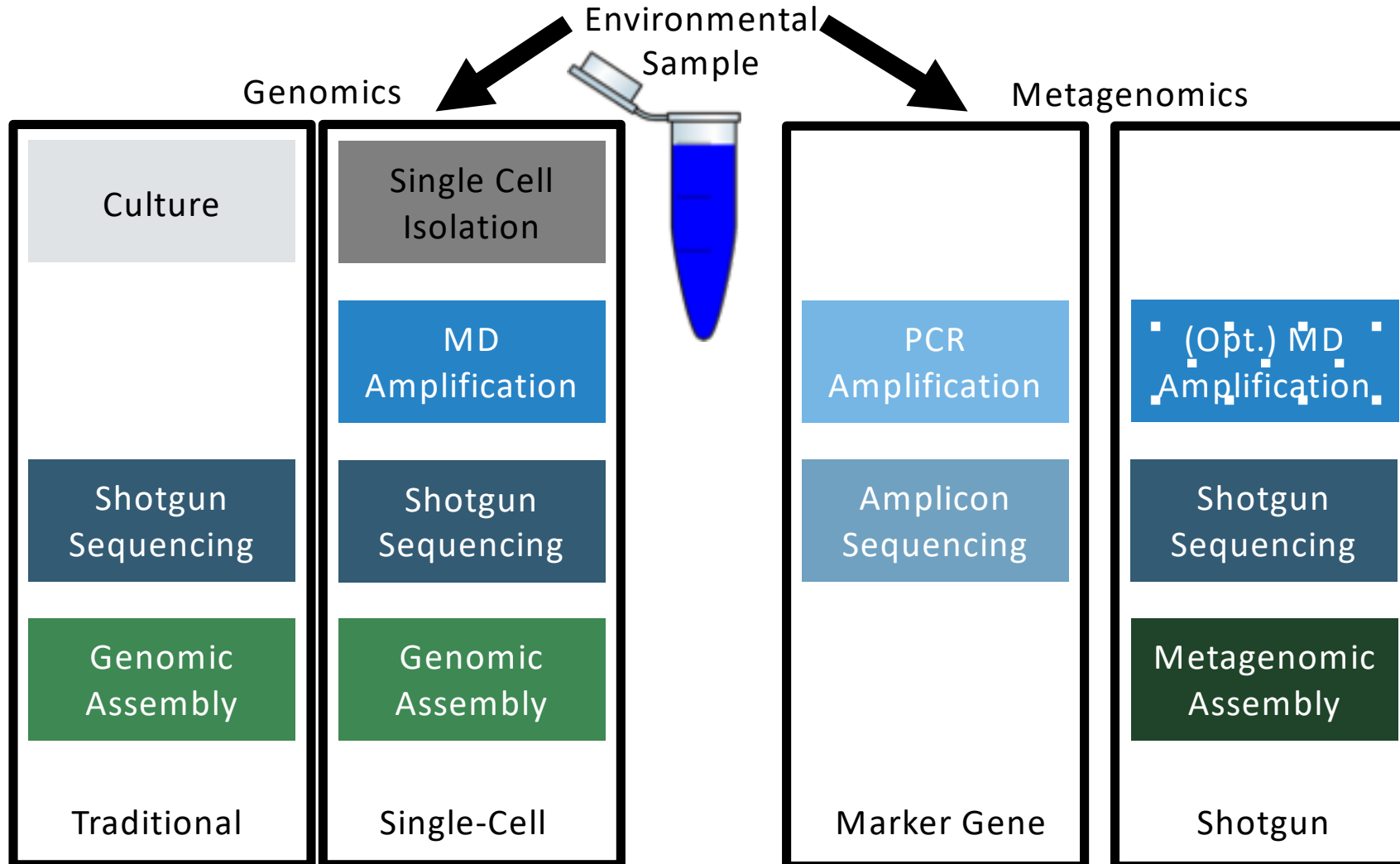


When to Use Traditional Genomics

- 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



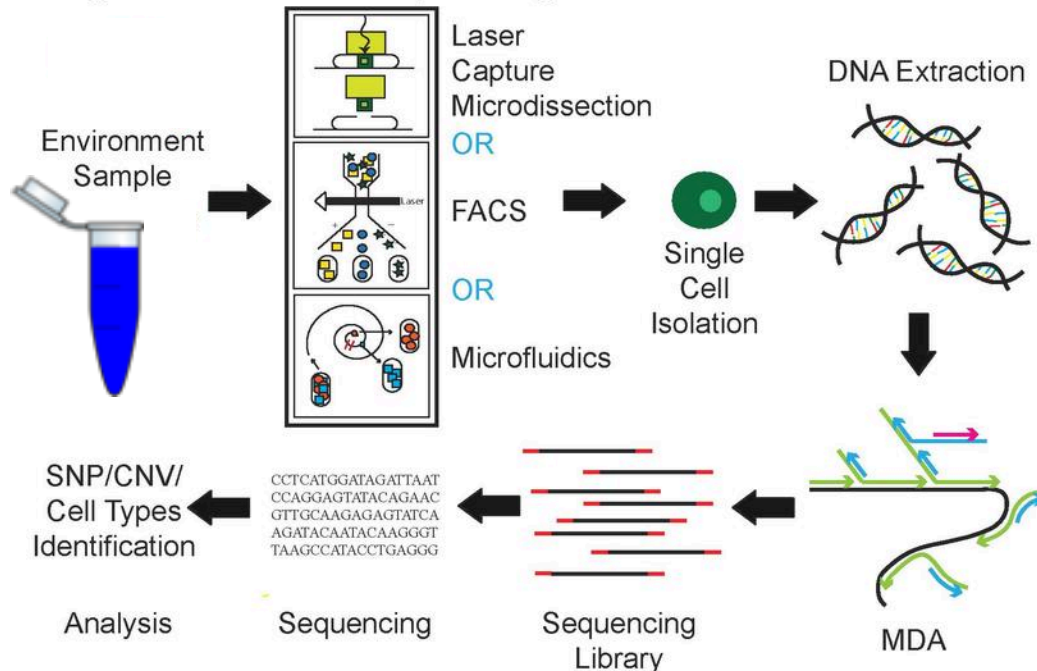
Microbial Genomics & Metagenomics



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

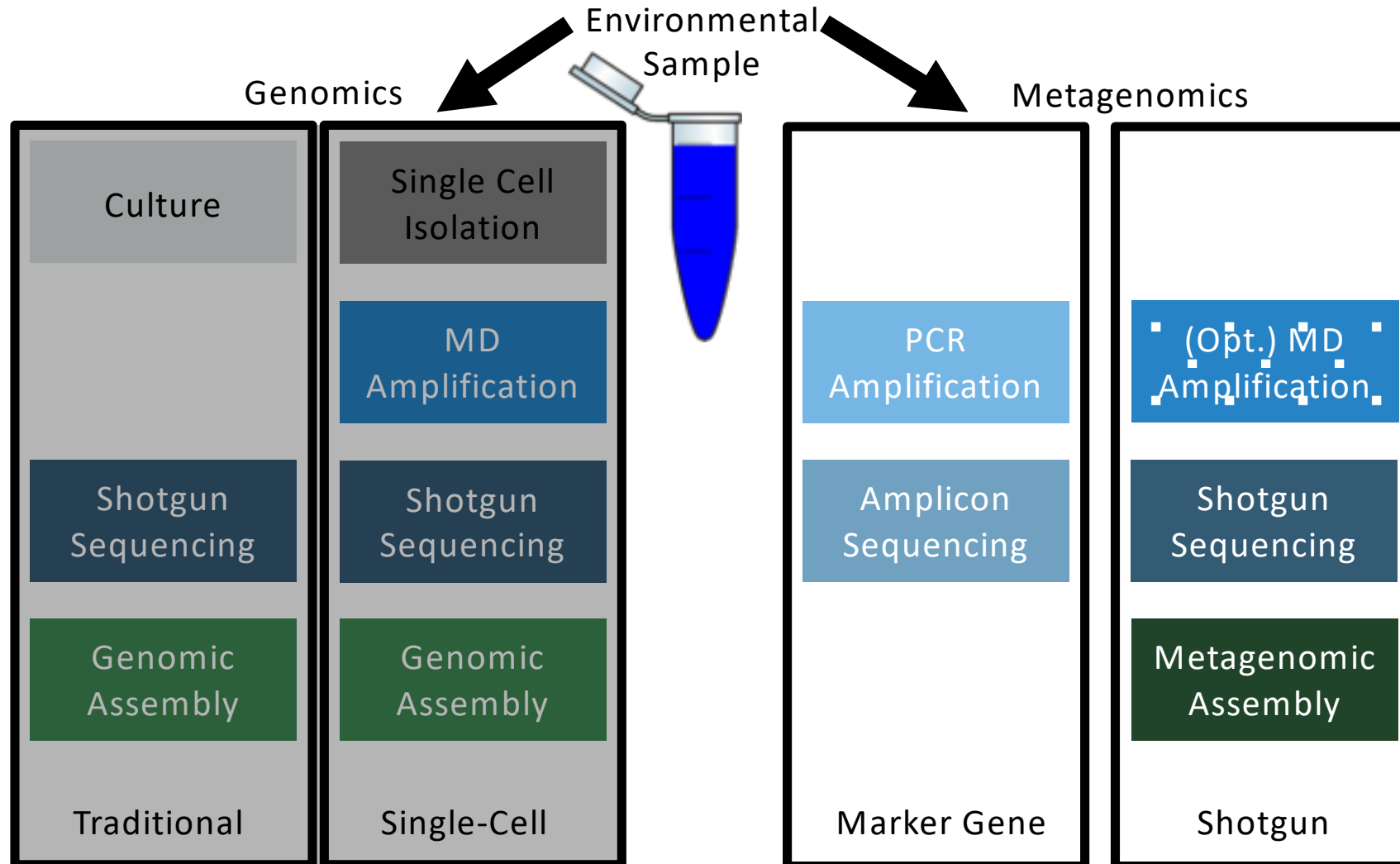
Single Cell Genome Sequencing Workflow



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

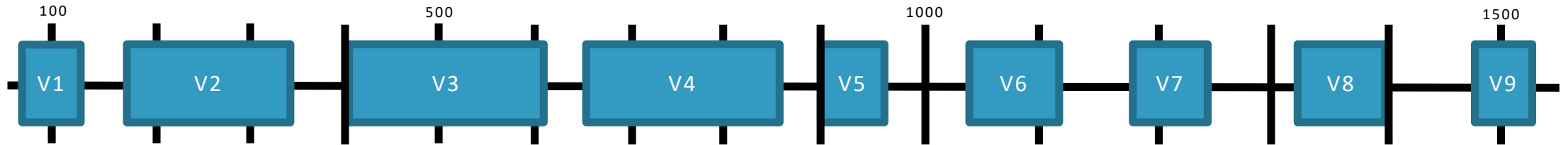
- When you have the time, money, and equipment
 - Having single-cell-level understanding of microbial communities is wonderful if we can get it!
- The good news:
 - See traditional genomics, plus ... this is really possible!
- The bad news:
 - Getting single cells is expensive and/or time-consuming
 - 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

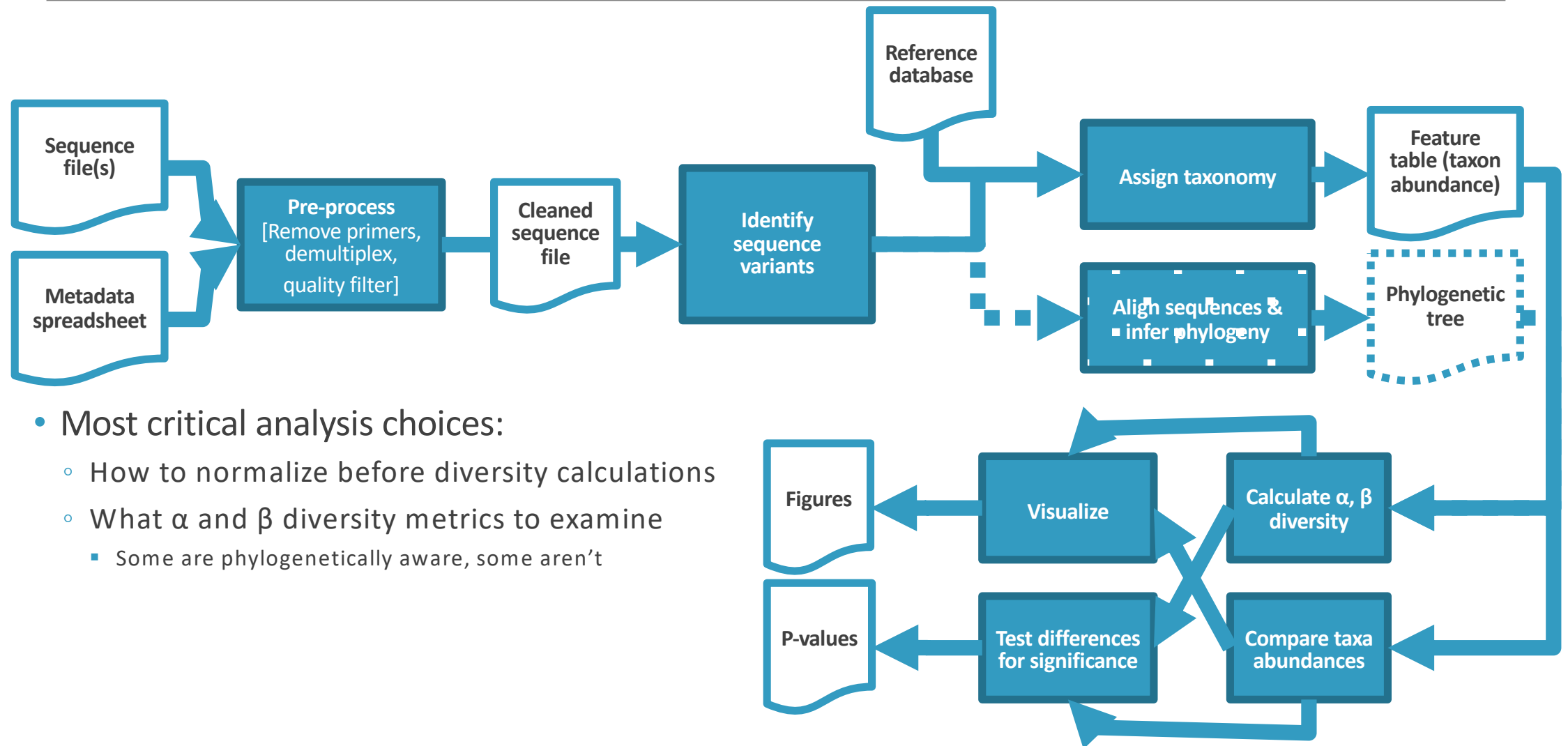


- 16S rRNA is widely conserved across bacteria and archaea, providing shared primer sites
 - Yet also has 9 hypervariable regions: can be used to differentiate organisms and build phylogenetic trees
- Can't study fungi with 16S (they don't have it) nor 18S (evolves too slowly)
 - Internal transcribed 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
- When you're cheap 😊
- 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
 - And, for a given gene, no primer pair distinguishes all microbes well
 - No other genome information (outside target gene) is captured

Marker Gene Analysis Workflow



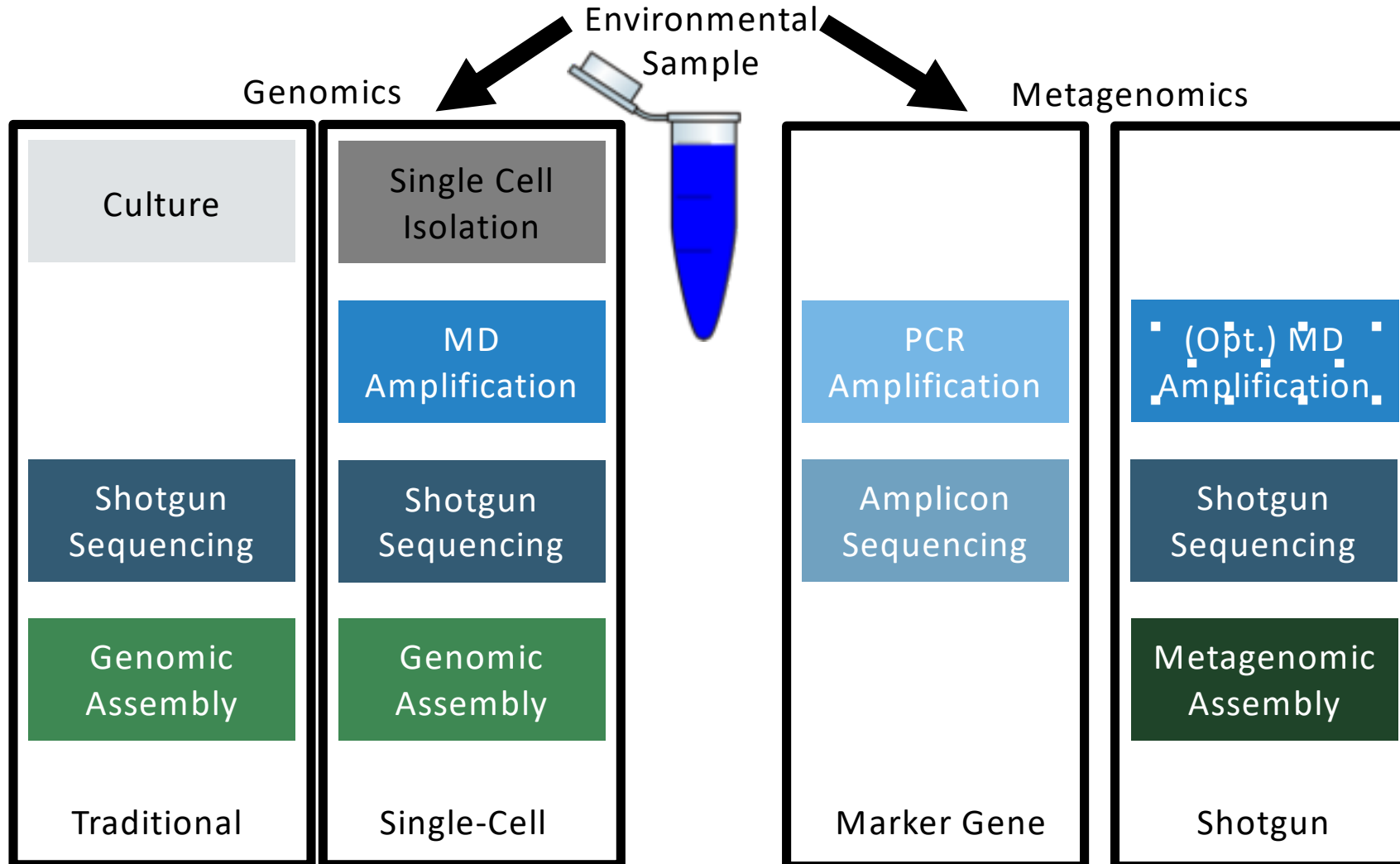
- Most critical analysis choices:
 - How to normalize before diversity calculations
 - What α and β diversity metrics to examine
 - Some are phylogenetically aware, some aren't

Common Issues in Marker Gene Studies

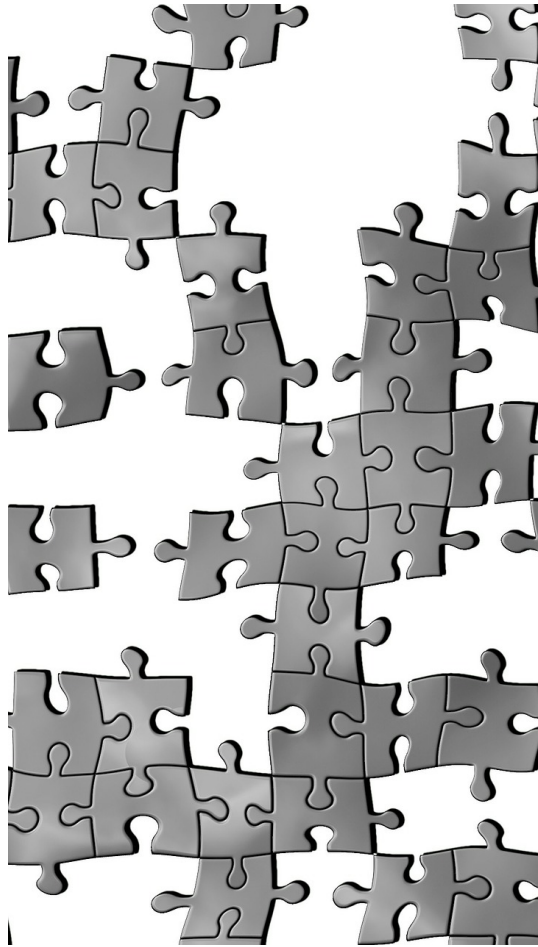
- Neglecting metadata
 - Analysis can not test for effects of, or discard bias from, categories you didn't record!
- Picking novel 16S primers—not all created equal
 - Earth Microbiome Project recommends 515f-806r primers, error-correcting barcodes
- Not taking precautions to support amplicon sequencing
 - 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 sequence variants only specify 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



Microbial Genomics & Metagenomics



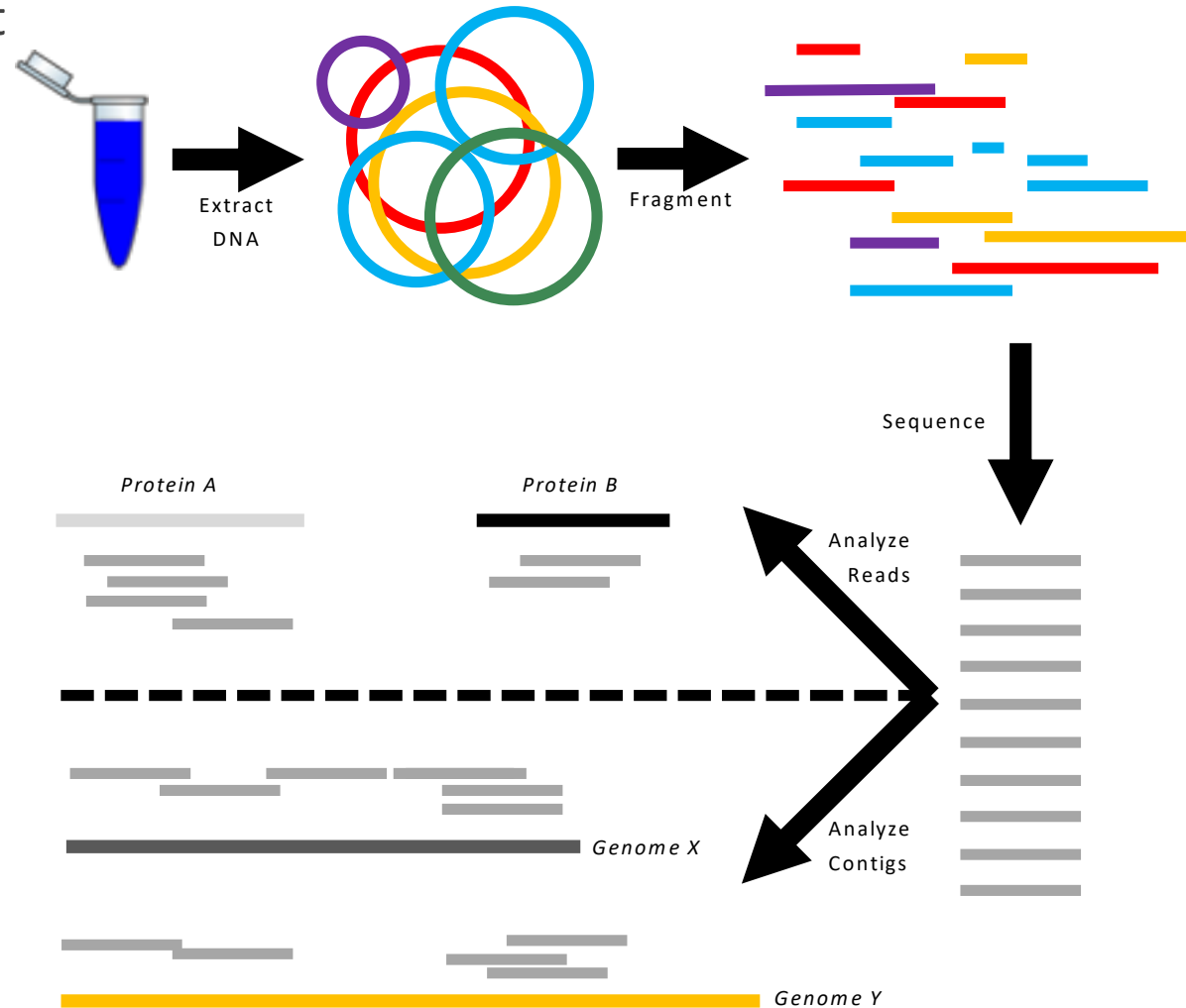
Shotgun Metagenomics Basics



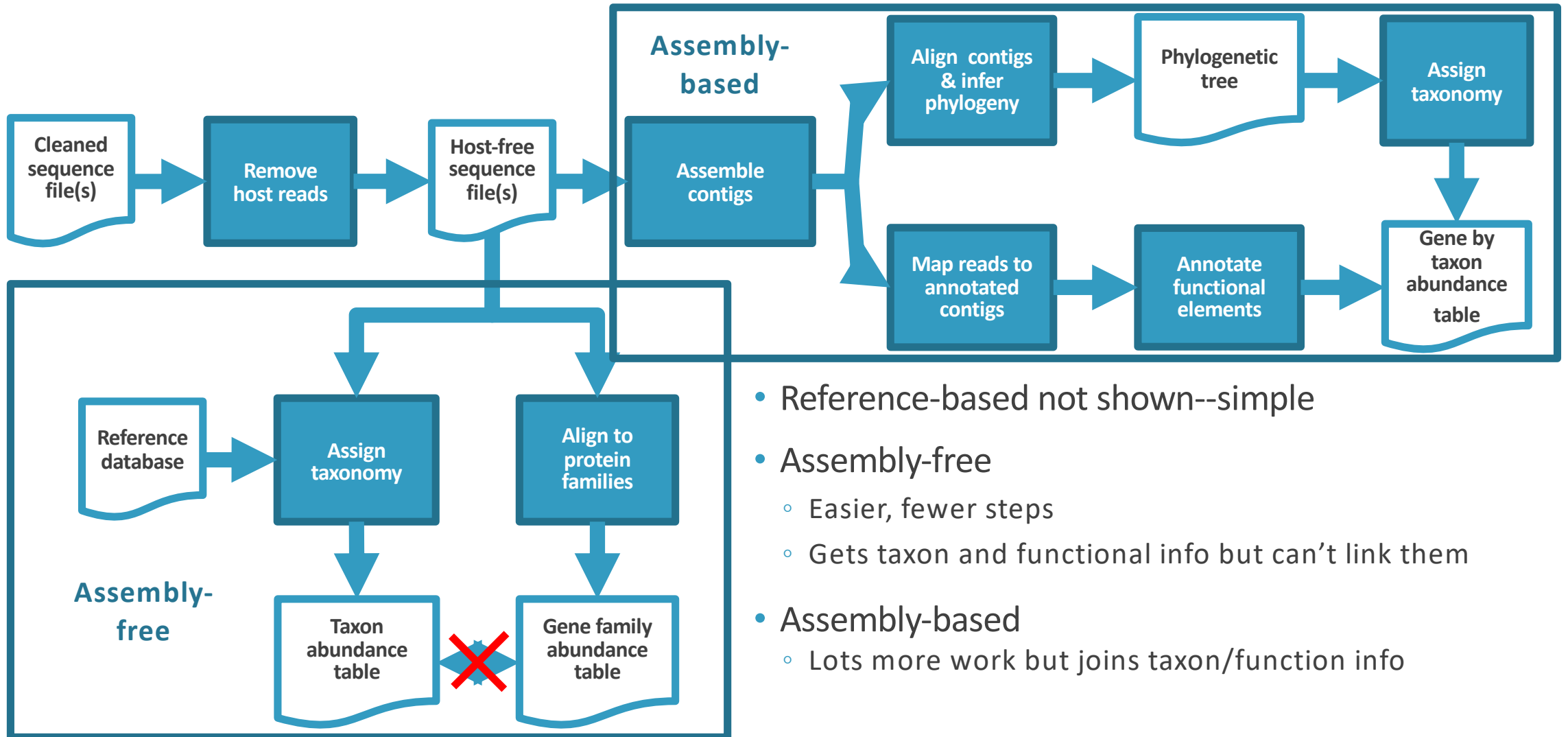
- Just fragment and sequence, try to figure out what it means in analysis!
- Reference-based
 - Map shotgun reads to database of known, complete reference genomes
 - Find identity and abundance
 - Analogous to approach for single-organism RNASeq—but more complex
 - Usually not feasible: too few references known
 - Exceptions: human gut, mouth, vagina
- Assembly-free
 - Map reads to known genomes and 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

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



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
 - However, it can still bite if you mishandle it (e.g., use non-compositionality-aware statistics)
- Shotgun metagenomics is no longer “bleeding edge”
 - But analysis stage still causes pain for non-computationalists!
 - Assembly-free analysis is easier
 - Assembly-based gets us closer to what we really want to know
- The only constant is change
 - New techniques are always being developed
 - New statistical differential abundance tests
 - Easier shotgun analysis pipelines
 - Longer-read sequencers
 - If you can, *save your samples*: may be easier to re-sequence later than to reanalyze

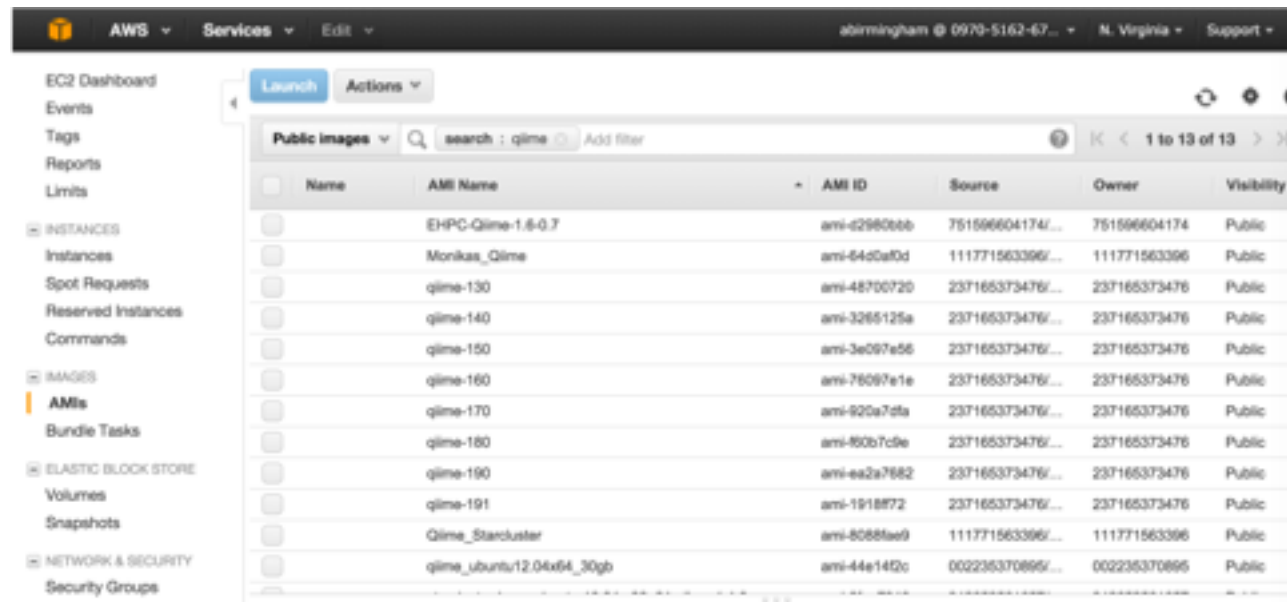
Acknowledgements

- Center for Computational Biology & Bioinformatics, University of California at San Diego
 - Executive Director Kathleen Fisch
- Greg Caporaso, Northern Arizona University
- Knight lab, UCSD
 - Jon Sanders
 - Antonio Gonzalez
 - Daniel McDonald

Supplementary Slides

Analysis in the Cloud

- For many applications, no longer necessary to buy, administer, and upgrade dedicated clusters
- Microsoft, Google, and Amazon all sell computing capacity on the open market
- Amazon Web services offers good combination of ease of use with customization
 - http://qiime.org/tutorials/working_with_aws.html
 -

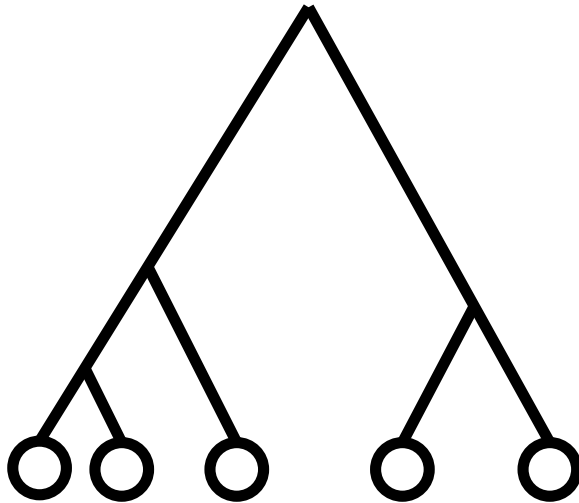


The screenshot shows the AWS Management Console interface for the 'Public images' section. The search filter is set to 'qime', and 13 results are displayed in a table. The table columns are Name, AMI Name, AMI ID, Source, Owner, and Visibility.

Name	AMI Name	AMI ID	Source	Owner	Visibility
	EHPC-Qime-1.6-0.7	ami-c2980bbb	751596604174/...	751596604174	Public
	Monikas_Qime	ami-64d0af0d	111771563396/...	111771563396	Public
	qime-130	ami-48700720	237165373476/...	237165373476	Public
	qime-140	ami-3265125a	237165373476/...	237165373476	Public
	qime-150	ami-3e097e56	237165373476/...	237165373476	Public
	qime-160	ami-76097e1e	237165373476/...	237165373476	Public
	qime-170	ami-920a70fa	237165373476/...	237165373476	Public
	qime-180	ami-60b7c9e	237165373476/...	237165373476	Public
	qime-190	ami-ea2a7682	237165373476/...	237165373476	Public
	qime-191	ami-1918872	237165373476/...	237165373476	Public
	Qime_Starcluster	ami-8088fae0	111771563396/...	111771563396	Public
	qime_ubuntu12.04x64_30gb	ami-44e140c	002235370895/...	002235370895	Public

Functional Prediction

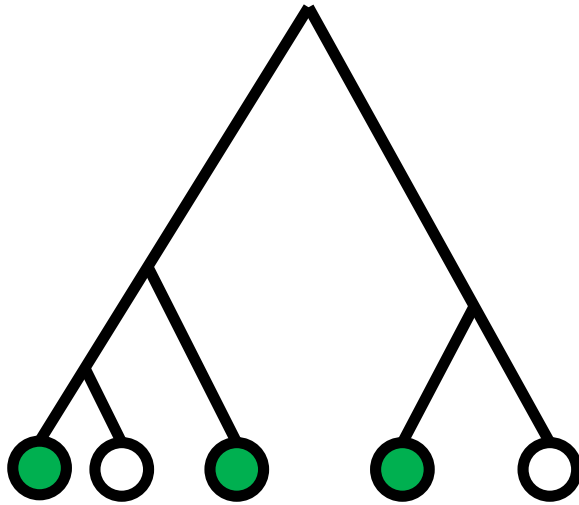
- Technique: predict contents of novel genomes from known ones
- Goal: Use 16S data to infer functional profiles of metagenomes
- Purpose: identify functional make-up of microbial communities




Phylogenetic tree of observed OTUs

Functional Prediction

- Technique: predict contents of novel genomes from known ones
- Goal: Use 16S data to infer functional profiles of metagenomes
- Purpose: identify functional make-up of microbial communities

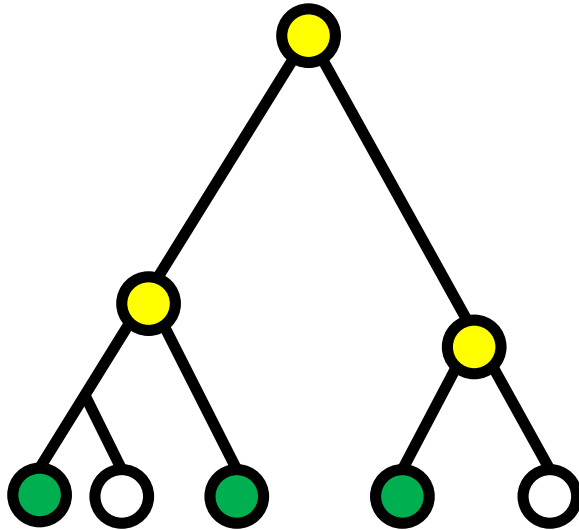


Phylogenetic tree of observed OTUs

 Known genomes

Functional Prediction

- Technique: predict contents of novel genomes from known ones
- Goal: Use 16S data to infer functional profiles of metagenomes
- Purpose: identify functional make-up of microbial communities

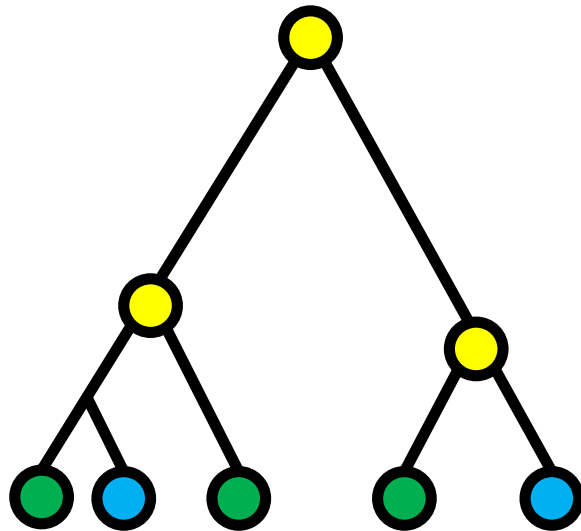


Phylogenetic tree of observed OTUs

- Known genomes
- Reconstructed ancestral genomes

Functional Prediction

- Technique: predict contents of novel genomes from known ones
- Goal: Use 16S data to infer functional profiles of metagenomes
- Purpose: identify functional make-up of microbial communities



Phylogenetic tree of observed OTUs



Known genomes



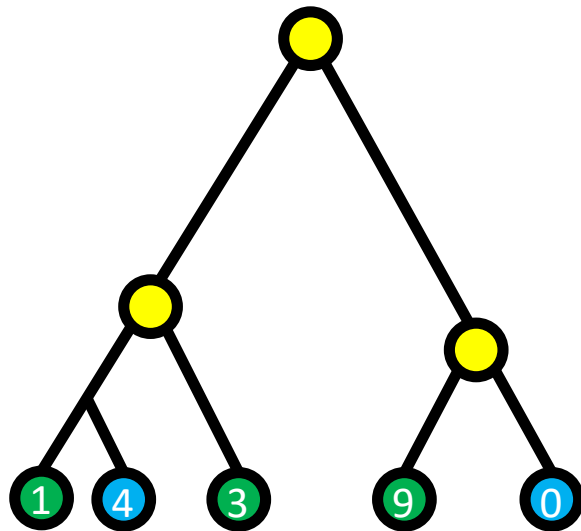
Predicted genomes



Reconstructed ancestral genomes

Functional Prediction

- Technique: predict contents of novel genomes from known ones
- Goal: Use 16S data to infer functional profiles of metagenomes
- Purpose: identify functional make-up of microbial communities



		X		
3	0	14	1	0
3	0	42	9	0

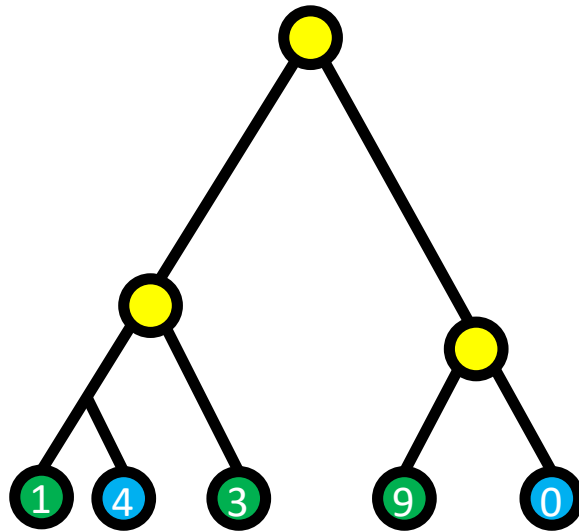
of genes in genome with given functionality

of genome instances (OTUs) observed in a sample

of functional instances in predicted metagenome for sample

Functional Prediction

- Technique: predict contents of novel genomes from known ones
- Goal: Use 16S data to infer functional profiles of metagenomes
- Purpose: identify functional make-up of microbial communities



- Check NSTI (Nearest Sequenced Taxon Index)
 - Low values give better accuracy
- Choose a functional source relevant to your study
 - E.g., KEGG Orthology, COG, etc
- To infer metagenomes: PICRUSt
- To assess findings: QIIME

		X		
3	0	14	1	0
3	0	42	9	0

of genes in genome with given functionality

of genome instances (OTUs) observed

of functional instances in predicted metagenome for sample

Assembly-Free Pipeline

- Several software tools exist, e.g., MG-RAST, MEGAN, Biobakery
- Biobakery suite (from Huttenhower lab, Harvard)
 - Kneaddata: data QC and prep
 - FastQC (optional)
 - Trimmomatic
 - Bowtie against host
 - FastQC (optional)
 - Metaphlan: taxonomic profiling
 - HUMAnN2: functional profiling
 - Also estimates pathway abundance and coverage

Assembly-Based Pipeline

- The pieces exist but no pipeline wrapper—yet
 - Data QC and prep: FastQC, Trimmomatic, Bowtie, etc
 - Could use kneaddata
 - Assembly Options
 - MegaHIT—blazing fast
 - MetaSPAdes—developed at UCSD (Pavel Pevsner)
 - Taxonomic Profiling Options
 - PhyloPhlAn (Biobakery)
 - TIPP—developed at UCSD (Siavash Mirarab)
 - Genomic Annotation Options
 - Prokka
 - Micronota—developed at UCSD (Rob Knight)
 - Binning useful when combining samples (Concoct)