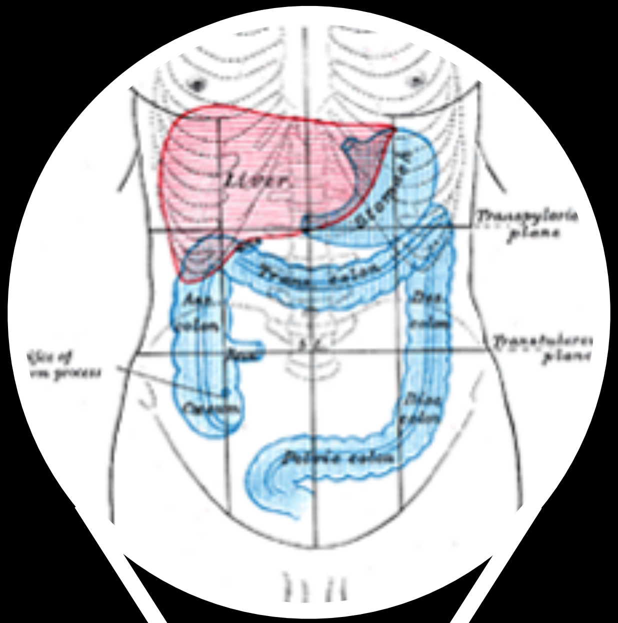
Metagenomics

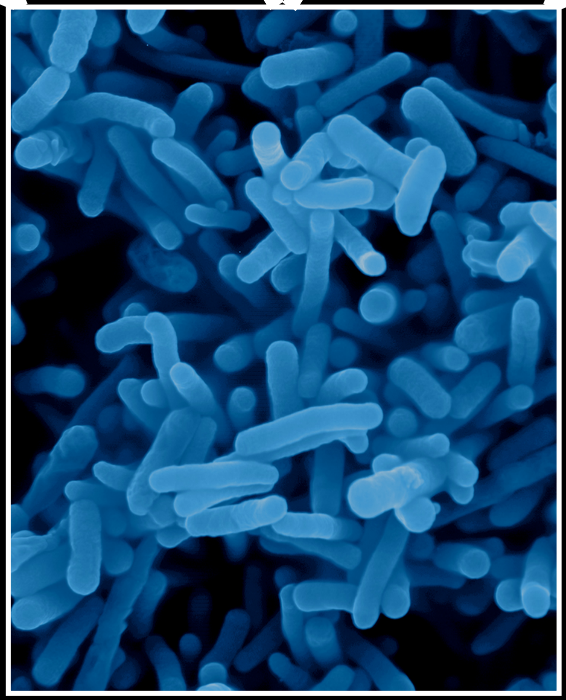
Katie Pollard

BMI 206
docpollard.org/bmi206
October 3, 2016

Microbes are Everywhere



But only ~1% have been cultured!



Who is there? What are they doing?

The Human Microbiome

Microbes in our bodies

- Equal numbers to human cells
- Contribute 100x more genes
- Make up ~5 lbs. of body weight (most of which is gut microbes)
- Directly contact human cells in our organs and body fluids
- Communicate and exchange molecules with human cells
- Interact with human genetics to make us who we are

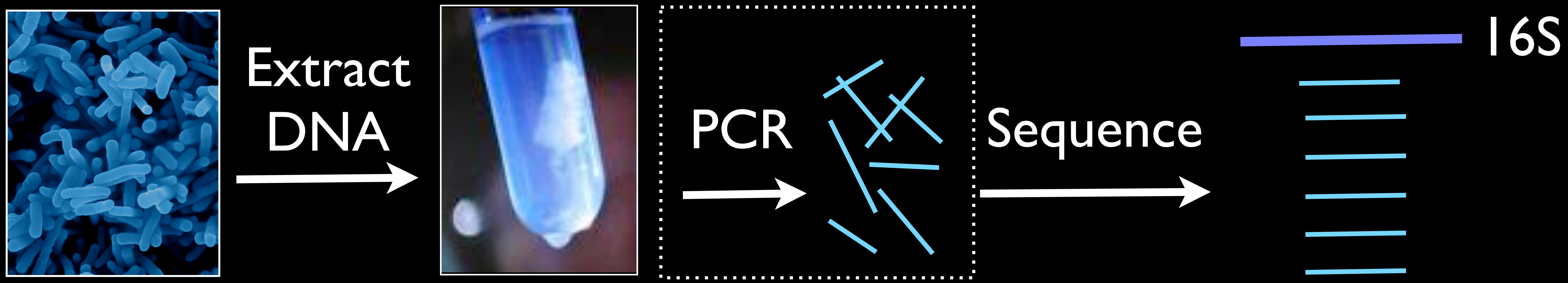
Microbiome Changes with Disease

- Microbial community composition is associated with many diseases:
  - Obesity and malnutrition
  - Colitis after antibiotic treatment
  - Inflammatory bowel diseases
  - HIV progression
  - Tooth and gum diseases
  - Ear infections

Why does the microbiome change? Is it causing disease?

Idea: Microbiome manipulation could lead to novel cures

Studying Microbes *In Situ*



Extract DNA

PCR

Sequence

16S

PCR-Based Sequencing (16S rRNA gene)

Two general approaches to 16S analysis

Reference based:

1. Compare reads to reference database of 16S sequences using BLAST like algorithms
2. Count reads homologous to each taxon
3. Normalize to quantify taxon (relative) abundance

De novo operational taxonomic units (OTUs):

1. Cluster reads based on percent sequence identity
2. Normalize cluster sizes to quantify relative abundance
3. Optionally label clusters based on similarity to reference database sequences

Quantifying Community Alpha Diversity

RICHNESS

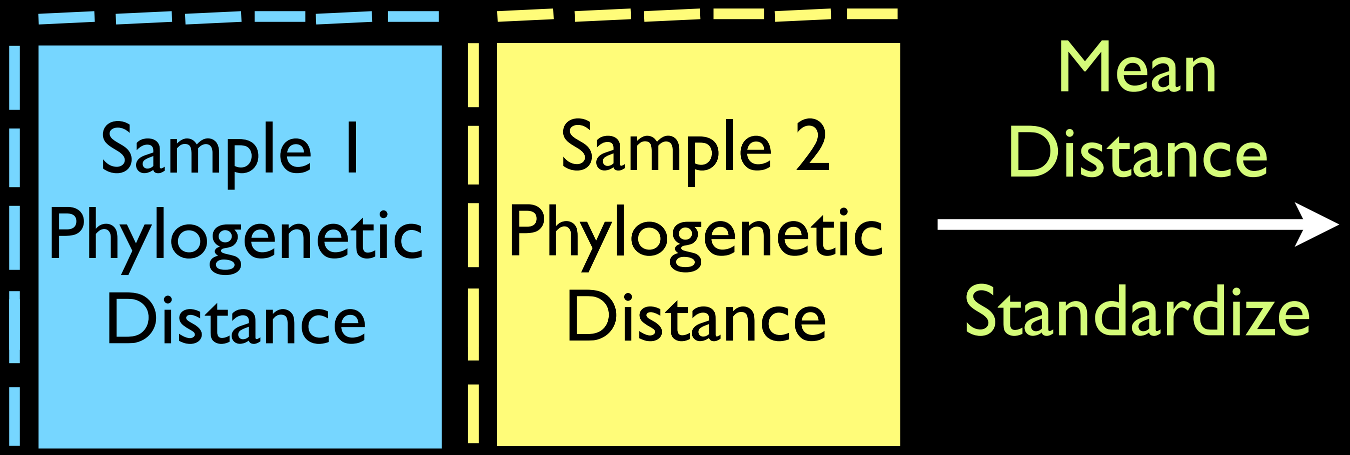
- Number of OTUs or protein families

SHANNON DIVERSITY

- Evenness of OTUs or protein families

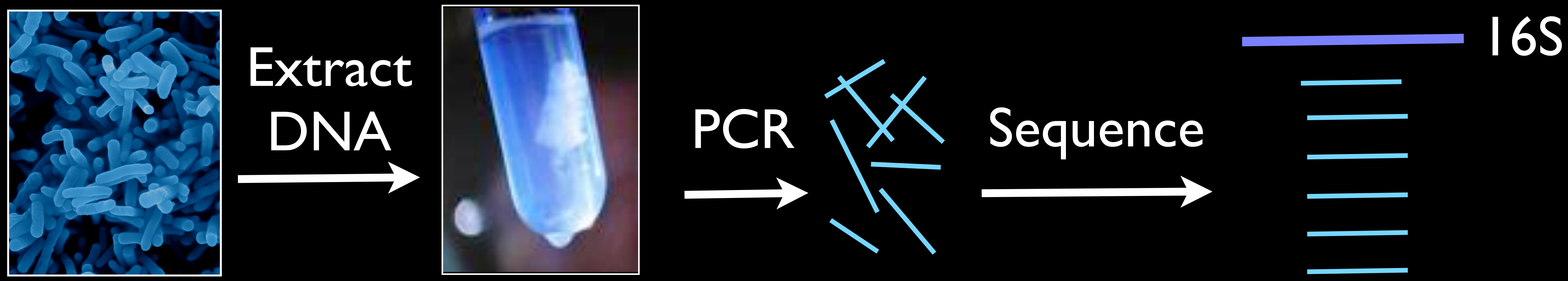
PHYLOGENETIC DIVERSITY

- Z-score of pairwise branch lengths


Sample 1 Phylogenetic Distance
Sample 2 Phylogenetic Distance
Mean Distance
Standardize

	Phylogenetic Diversity	Depth	Biomass
Sample 1	-2.4	10m	7.21
Sample 2	0.9	200m	1.66

Studying Microbes *In Situ*



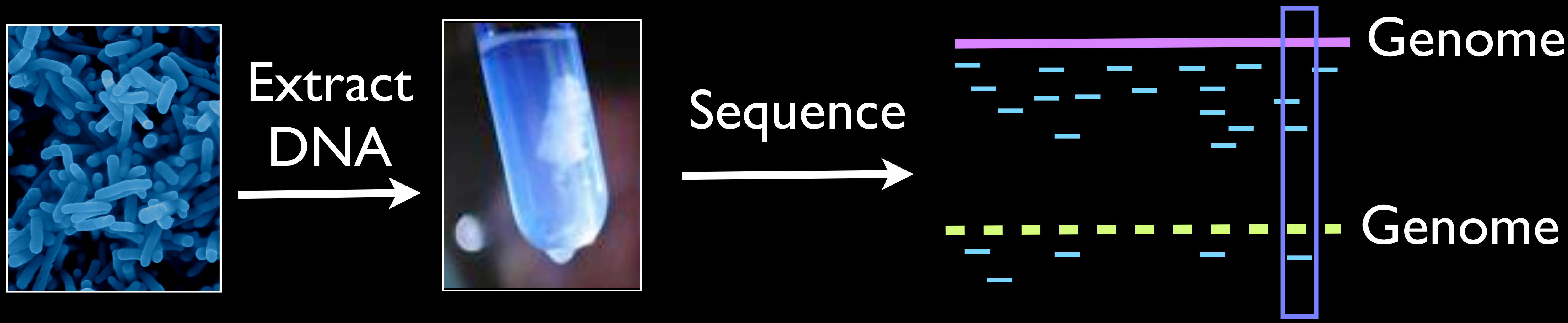
Extract DNA

PCR

Sequence

16S

PCR-Based Sequencing (16S rRNA gene)



Extract DNA

Sequence

Genome

Genome

Metagenomic Shotgun Sequencing

16S

Metagenomics: Promises & Challenges

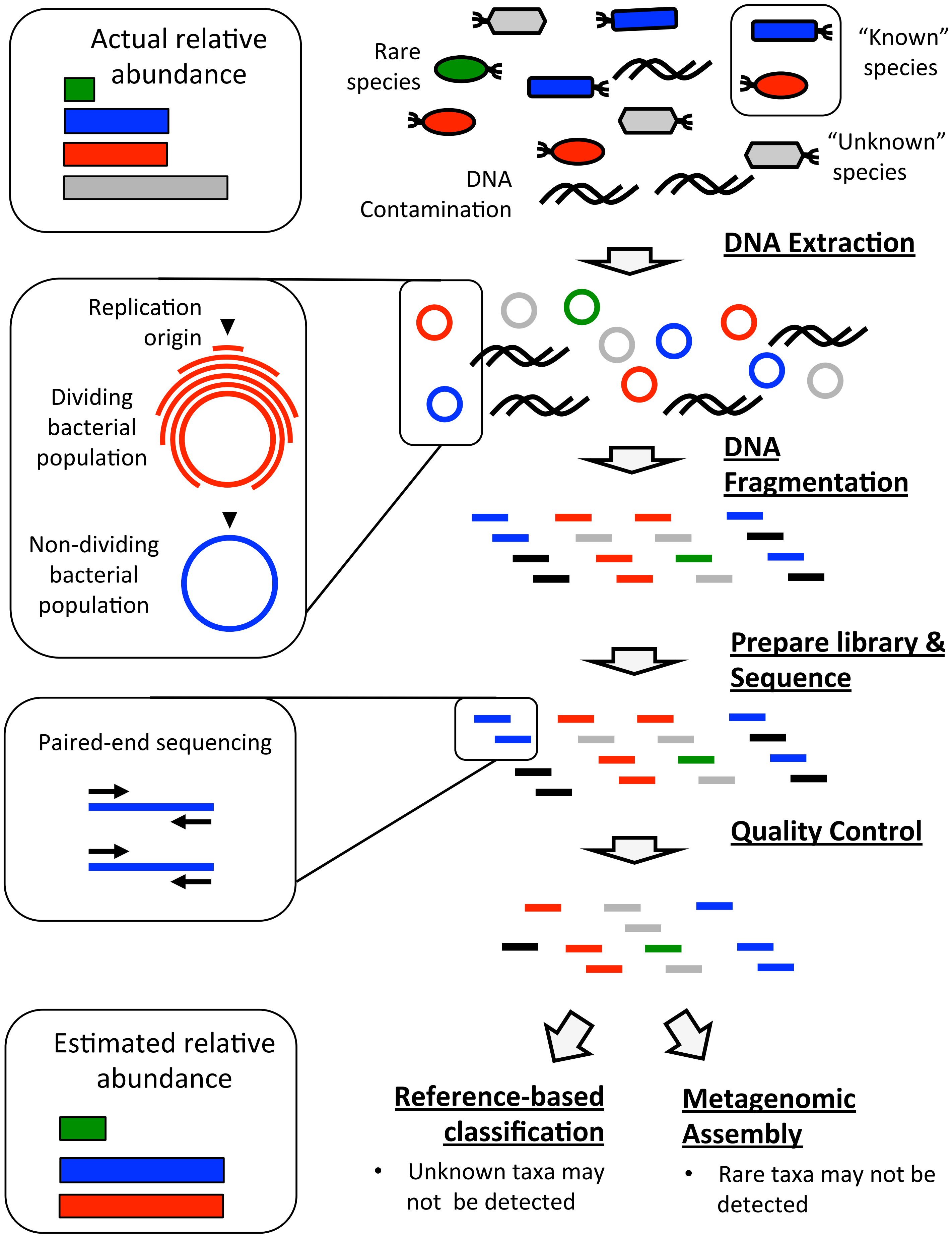
Shotgun sequencing enables:

1. Identification of new microbes & genes
2. Better quantification of microbial diversity
3. Associate microbiome taxa & functions with traits
4. Strain-level analysis of genes within species

But new methods are required to:

1. Minimize effects of experimental error
2. Reduce informatics biases
3. Estimate meaningful abundance parameters

Sample from microbial community


Actual relative abundance
Rare species
"Known" species
"Unknown" species
DNA Contamination
Replication origin
Dividing bacterial population
Non-dividing bacterial population
Paired-end sequencing
Estimated relative abundance

A

- Unknown species can dominate microbial communities (Nayfach and Pollard 2016) and are not detected by reference-based methods
- DNA from the host (Ames et al. 2015) or laboratory environment (Salter et al. 2014) can contaminate a biological sample

DNA Extraction

B

- Extraction efficiency varies between taxa (Kennedy et al. 2014)
- Dividing bacterial genomes have higher and less even genomic coverage (Korem et al. 2015)

DNA Fragmentation

C

- Extracted DNA is fragmented at breakpoints which preferentially occur at certain di-nucleotides (Poptsova et al. 2014)

Prepare library & Sequence

D

- Library preparation protocol affects estimated community composition (Jones et al. 2015)
- Sequencing technologies have different read lengths and error rates (Quail et al. 2012)

Quality Control

E

- Duplicate reads eliminated
- Read-tails trimmed
- Low quality reads filtered
- DNA contamination removed

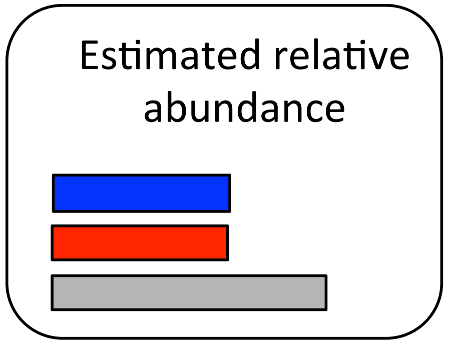
Reference-based classification

- Unknown taxa may not be detected

Metagenomic Assembly

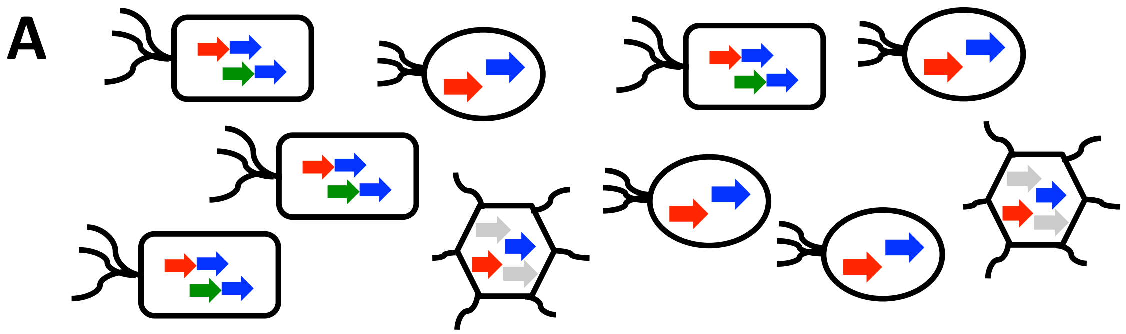
- Rare taxa may not be detected

F

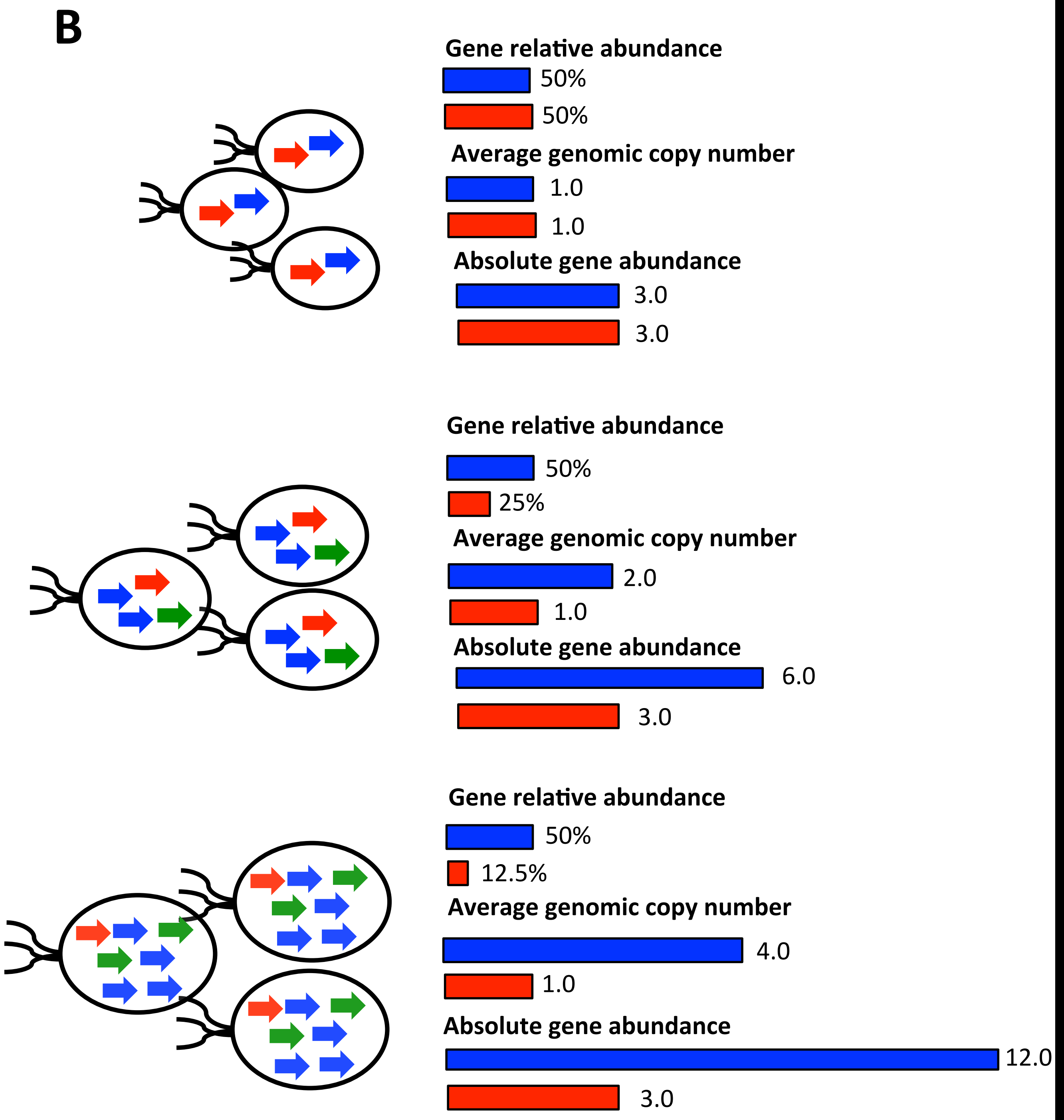


Nayfach & Pollard (2016) Cell

Taxon & gene abundance parameters



	CELLS	GENES
Relative Abundance	▭ = 4/10	➡ = 4/32
Absolute Abundance	▭ = 4	➡ = 4
Average Copy Number	n/a	➡ = 4/10
Per Taxon Copy Number	n/a	➡ = (1,0,0)



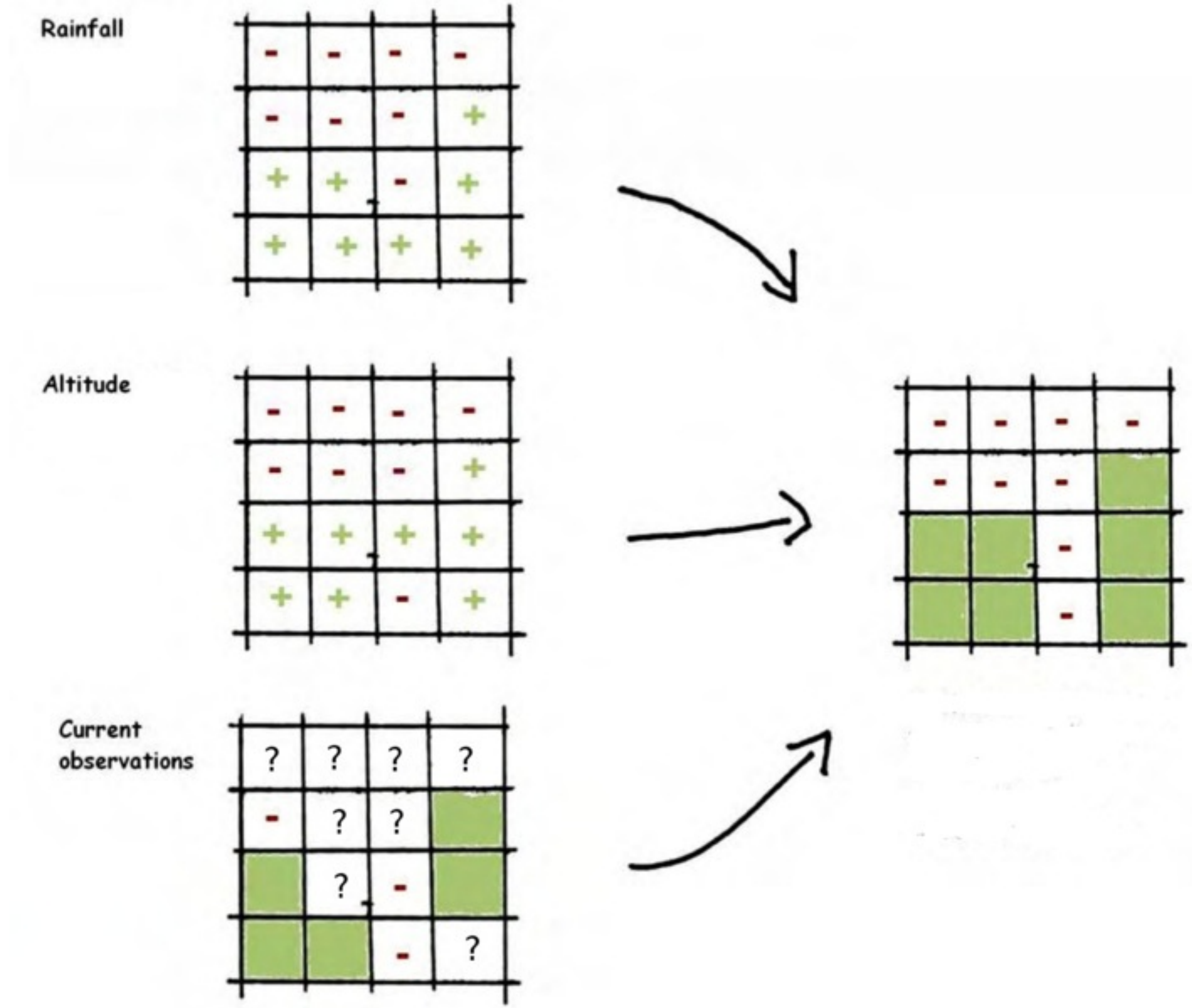
Nayfach & Pollard (2016) Cell

Other quantitative problems in metagenomics

1. Gene and genome assembly
2. Binning
3. Strain-level analysis
4. Covariation analysis
5. Metabolic modeling
6. Longitudinal analysis

Niche modeling

Niche Modeling: Predicting microbial distributions



Rainfall

-	-	-	-
-	-	-	+
+	+	-	+
+	+	+	+

Altitude

-	-	-	-
-	-	-	+
+	+	+	+
+	+	-	+

Current observations

?	?	?	?
-	?	?	🟩
🟩	?	-	🟩
🟩	🟩	-	?

-	-	-	-
-	-	-	🟩
🟩	🟩	-	🟩
🟩	🟩	-	🟩

Franklin and Miller, 2009, Mapping Species Distributions

Niche Modeling: Predicting microbial distributions

Input

1. OTUs or genes at sparse sampling locations

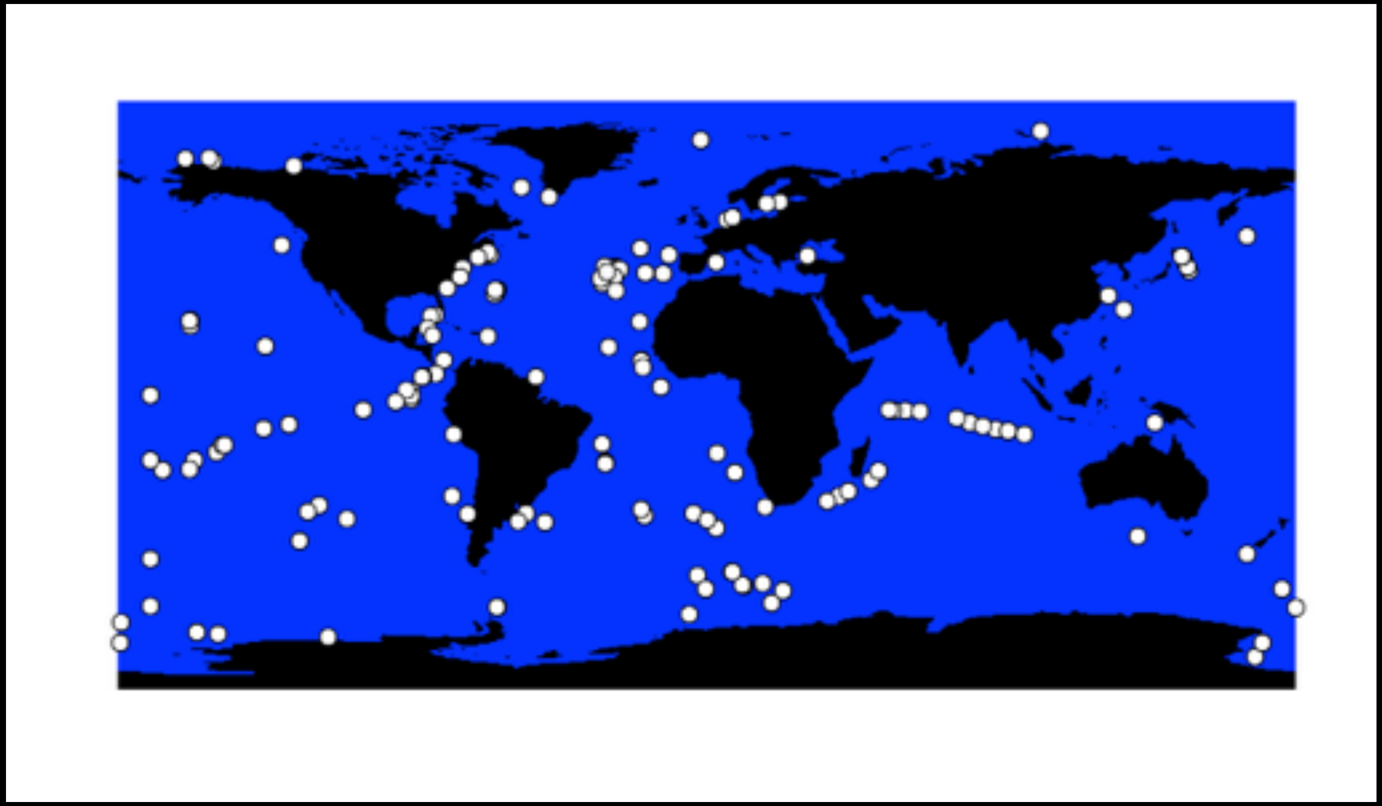
Sequence Data

377 samples, 164 unique locations

Marine surface waters (epipelagic zone)

16S sequences clustered into OTUs

2. Environmental data across globe



Model

Diversity ~ Month + Environment

Output

Predicted diversity across globe

Ladau et al. (2013) ISME Journal

Niche Modeling: Predicting microbial distributions

Input

1. OTUs or genes at sparse sampling locations

2. Environmental data across globe

Model

Diversity ~ Month + Environment

Output

Predicted diversity across globe

Environmental Data

surface temperature
depth (above thermocline)
chlorophyll concentration
salinity
day length
phosphate concentration
sea ice concentration

Ladau et al. (2013) ISME Journal

Niche Modeling: Predicting microbial distributions

Input

1. OTUs or genes at sparse sampling locations

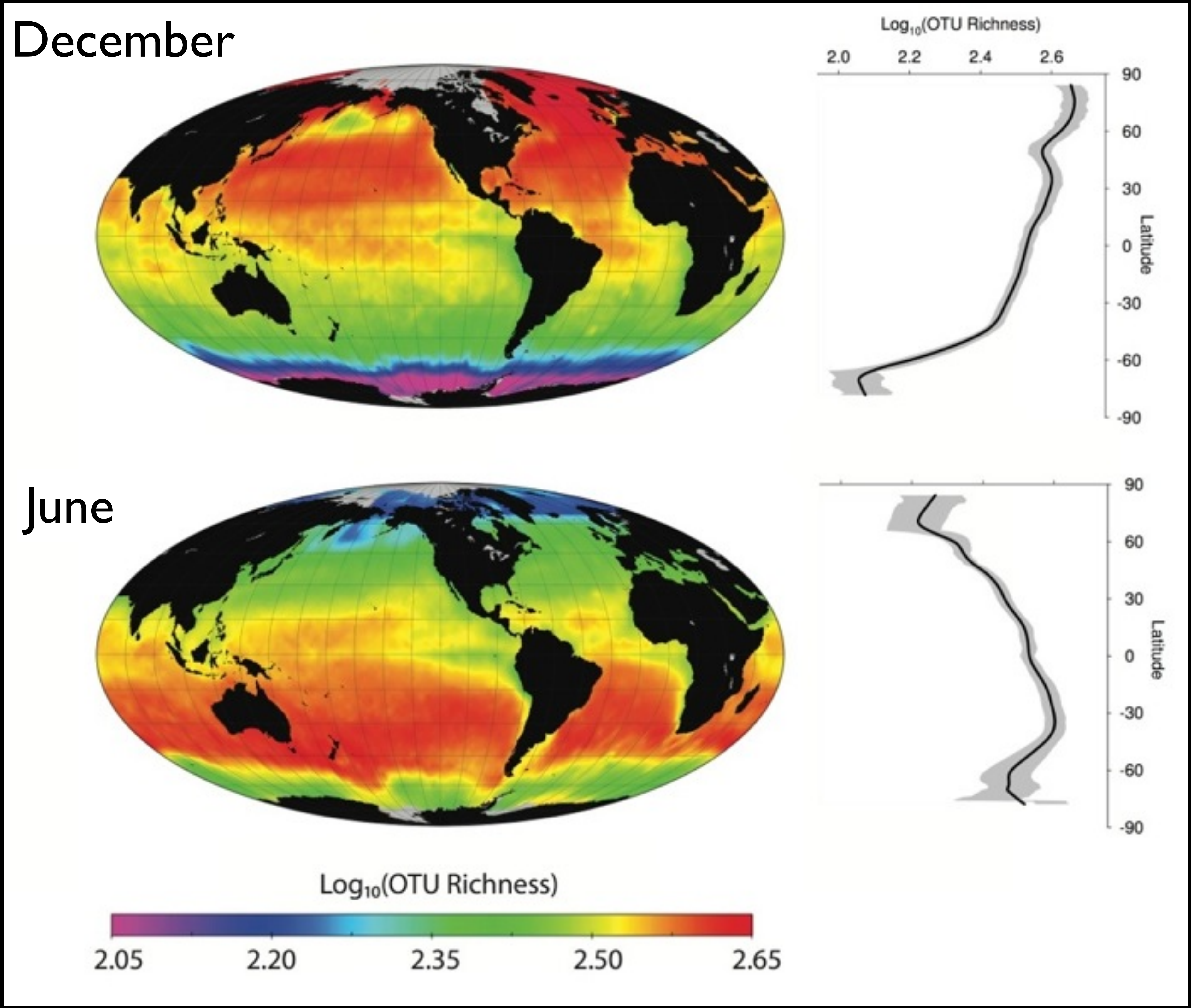
2. Environmental data across globe

Model

Diversity ~ Month + Environment

Output

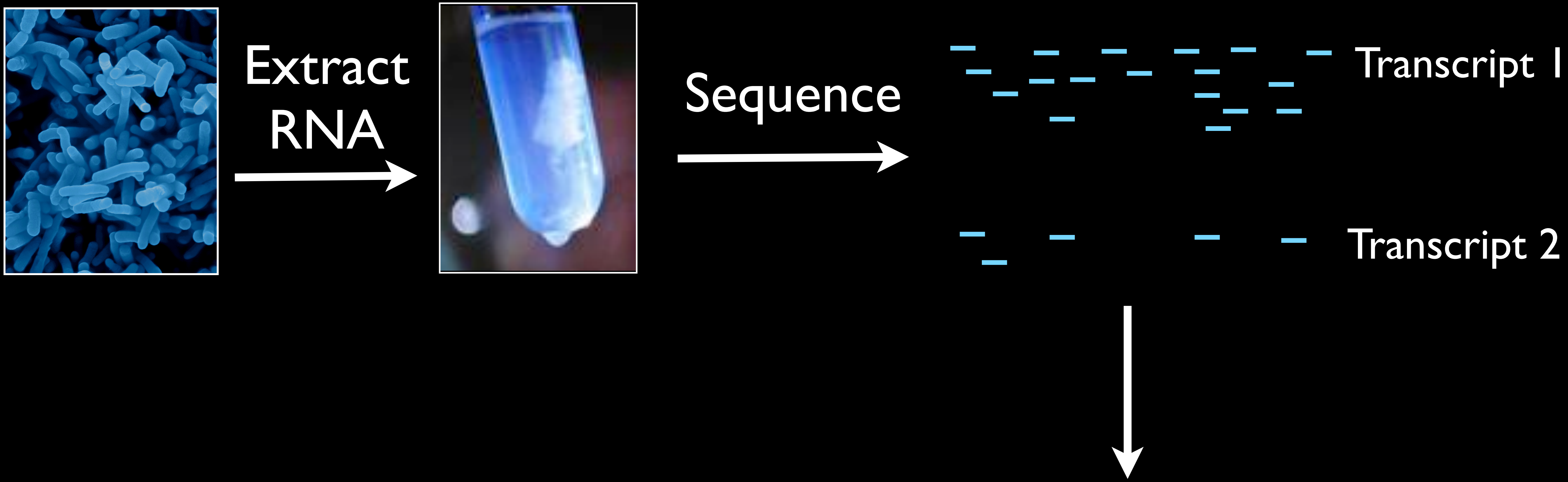
Predicted diversity across globe



Ladau et al. (2013) ISME Journal

Other Meta ‘Omics

Metatranscriptomics

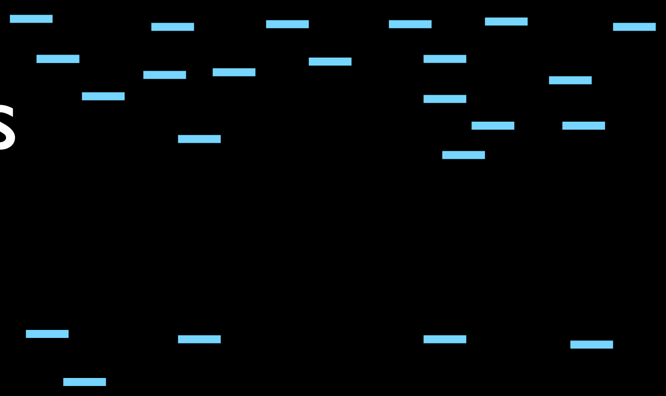


Expression levels of genes from many different genomes

Metaproteomics

Extract Protein

Fragment into peptides



Protein 1

Protein 2

Mass Spec

Abundances of detected proteins from many genomes

Metametabolomics

Extract small molecules

Separation methods

Analytes

Mass Spec or NMR

Abundances of detected metabolites from many organisms

Additional Details

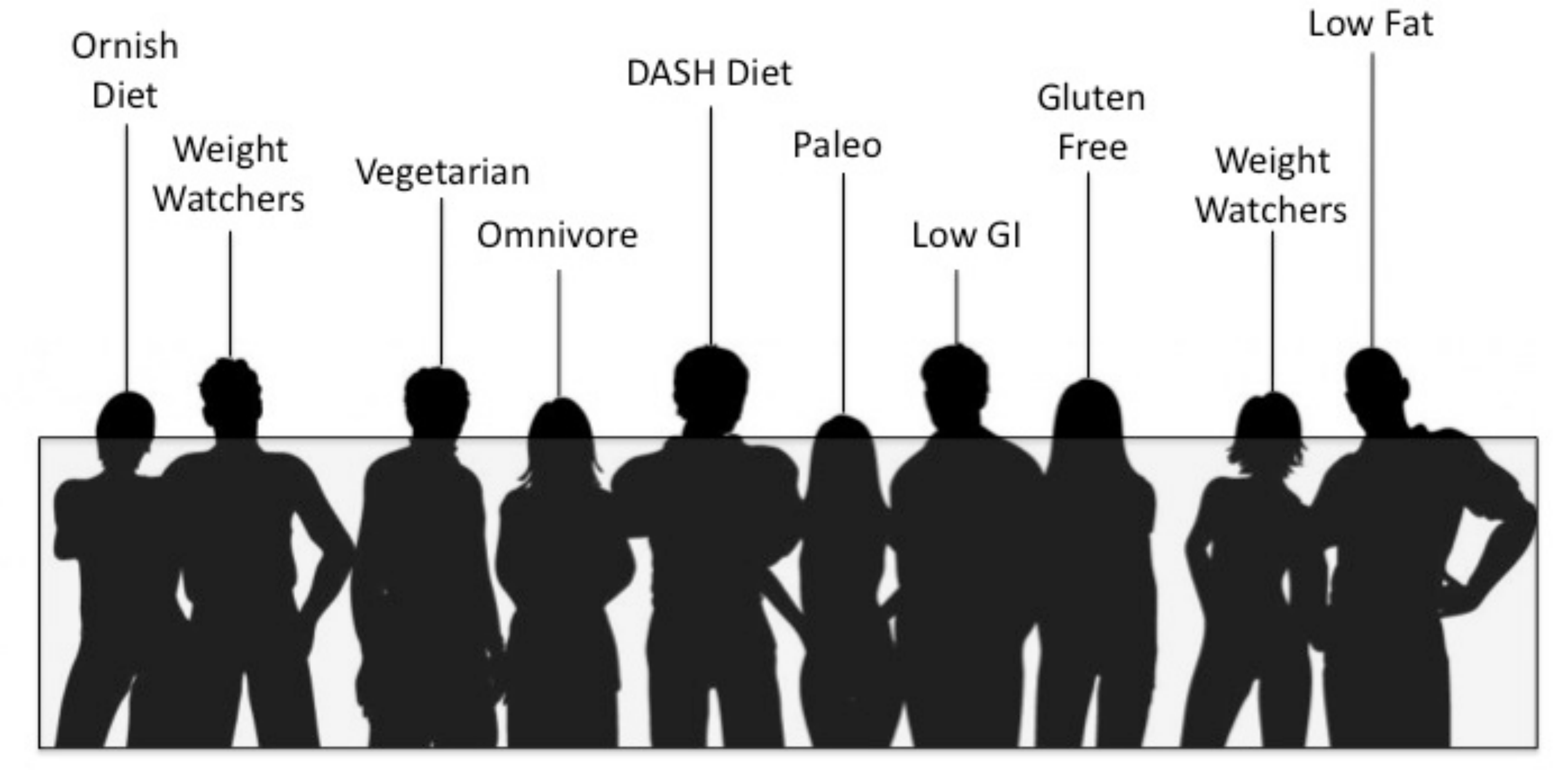
Microbiome Integral to Digestion

- Gut microbes:
  - Help us harvest energy from our food
  - Synthesize vitamins and metabolites for us
  - Produce anti-inflammatory molecules that allow us to tolerate their presence

- Gut microbes also affect other organs
  - Immunity
  - Hormones
  - Brain

Jon Berkeley

Microbiome Shaped By Diet



- Breastfeeding vs. formula in infants
- Microbiome composition changes within two days when switching diets (vegan vs. meat)
- Obesity and metabolism can be transferred via fecal transplant or coprophilia (mice)

How to estimate who is there?

1. Compare reads to sequence databases
   - Uses BLAST or related algorithms
     - Works if identical or similar to known microbes
     - Typically can't classify >50% of reads
   - Profile searches (HMMs for protein markers, SCFGs for RNA) can help with long reads, but not short

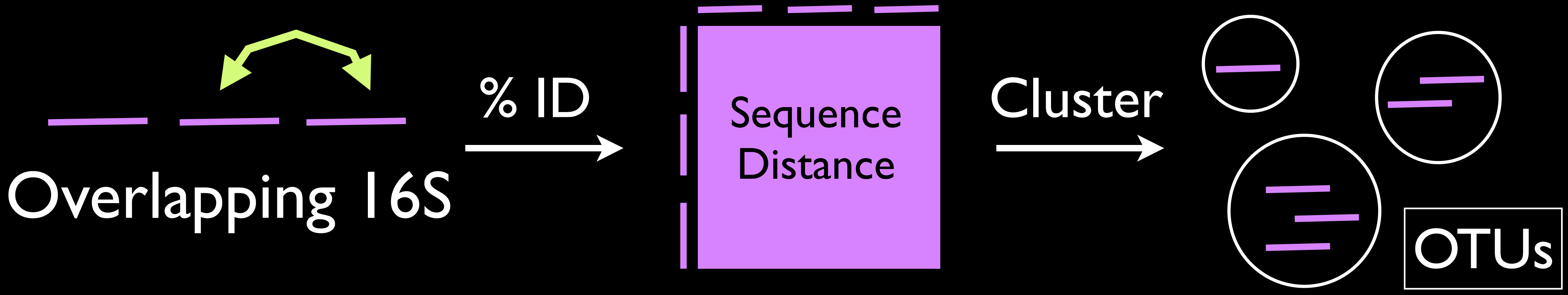
green genes
16S rRNA gene database and workbench compatible with ARB
greengenes.lbl.gov

silva
comprehensive ribosomal RNA databases
http://www.arb-silva.de

RIBOSOMAL DATABASE PROJECT
http://rdp.cme.msu.edu

How to estimate who is there?

1. Compare reads to sequence databases
2. Cluster reads from marker genes (16S, proteins) into Operational Taxonomic Units (OTUs)



MOTHUR/ESPRIT: http://plaza.ufl.edu/sunyijun/ESPRIT.htm
UCLUST/QIIME: http://qiime.org

How to estimate who is there?

1. Compare reads to sequence databases

2. Cluster reads from marker genes into OTUs
   - Typically requires overlapping reads (whole gene, pyrotags)
     - PhylOTU enabled computation of distance between non-overlapping reads using phylogeny

     PhylOTU: https://github.com/sharpton/PhylOTU
   - The challenge: Who are they?

Both approaches are being extended to detect strain-level variation in shotgun metagenomes

How to estimate what they are doing?

1. Compare reads to sequence databases
    - Pairwise searches (BLAST and fast-BLAST) work if identical or similar to known proteins

MEGAN: http://ab.inf.uni-tuebingen.de/software/megan/
MG-RAST: http://metagenomics.anl.gov
Phymm & PhymmBL: Brady & Salzberg (2009) Nature Methods

How to estimate what they are doing?

1. Compare reads to sequence databases
   - Pairwise searches (BLAST and fast-BLAST) work if identical or similar to known proteins
   - Profile searches can help for more distant homology (<30% aa identity), but perform poorly for some gene families and for short reads (BLAST generally better if <200bp)

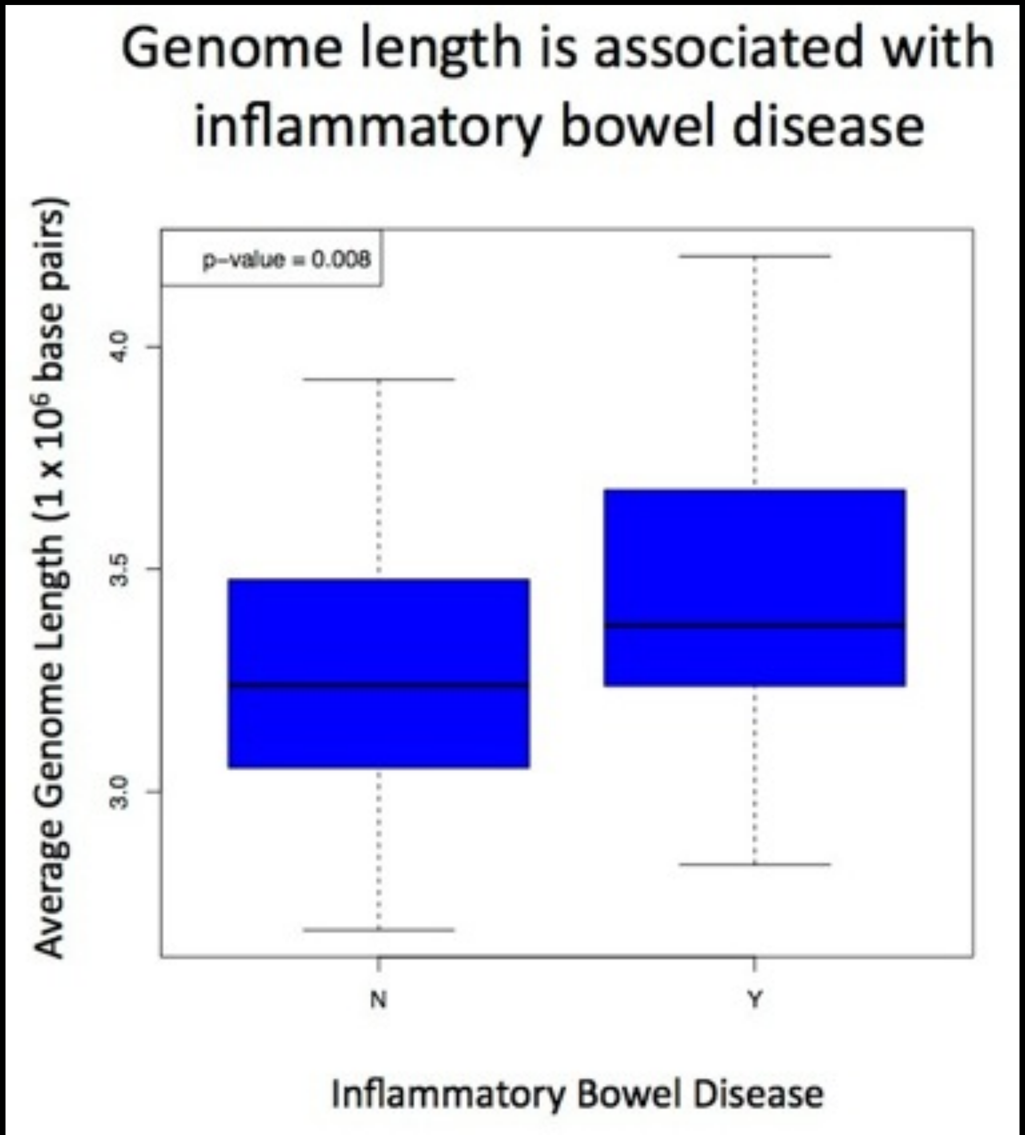
Pfam: http://pfam.sanger.ac.uk
FIGfams: http://www.theseed.org/wiki/FIGfams/
TIGRFAMS: http://www.jcvi.org/cgi-bin/tigrfams/index.cgi
SFams: Sharpton et al. BMC Bioinformatics 2012

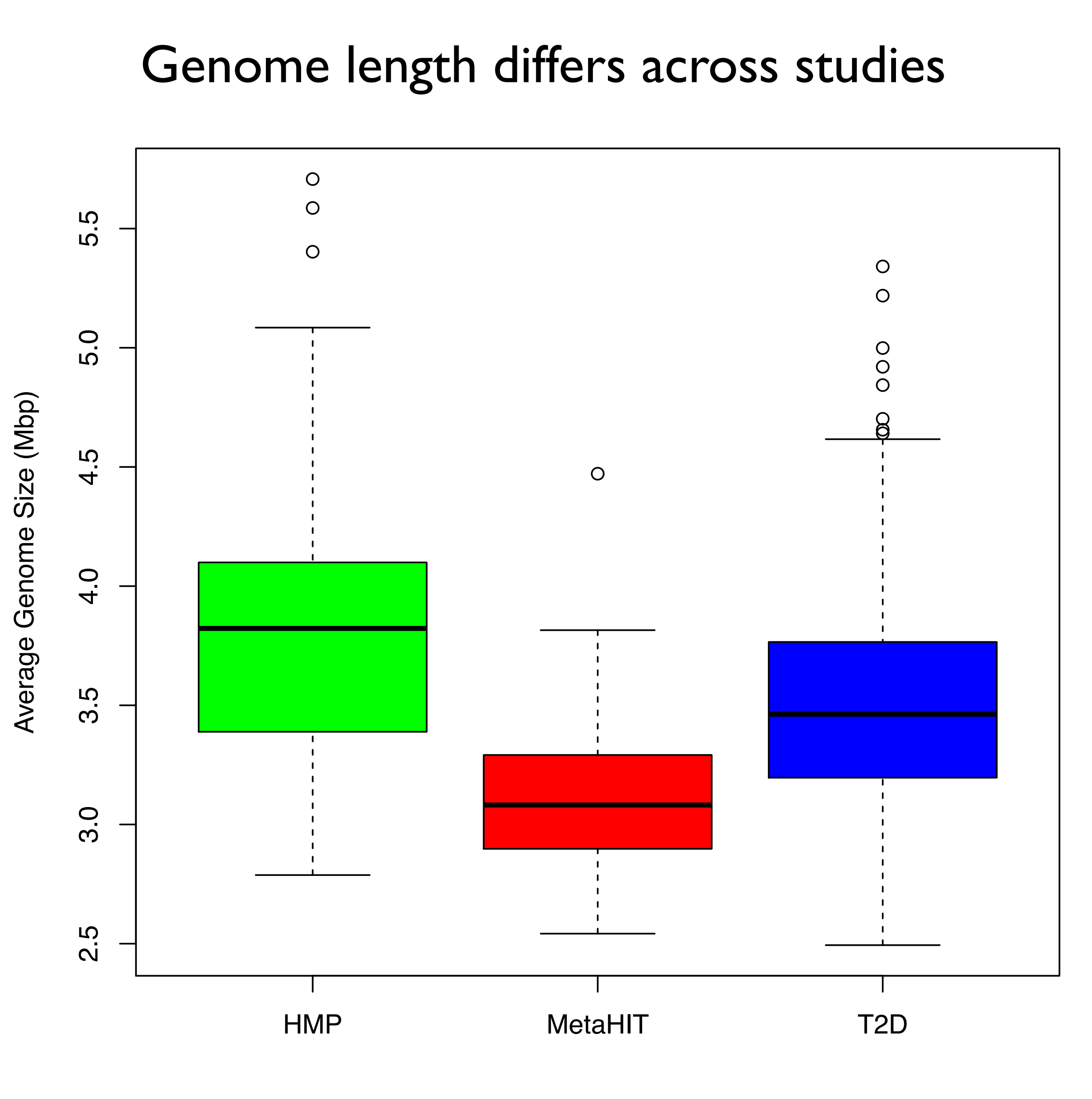
How to estimate what they are doing?

1. Compare reads to sequence databases
2. Cluster reads into Operational Protein Families
    - The challenge: What are their functions?

Schloss & Handelsman, BMC Bioinformatics 2008

Average genome size matters





Longer genomes → Fewer reads per gene → Systematic underestimate of abundance

Nayfach & Pollard *Genome Biology* (2014); Manor & Borenstein *Genome Biology* (2014)