Microbiome (culture-independent) analysis

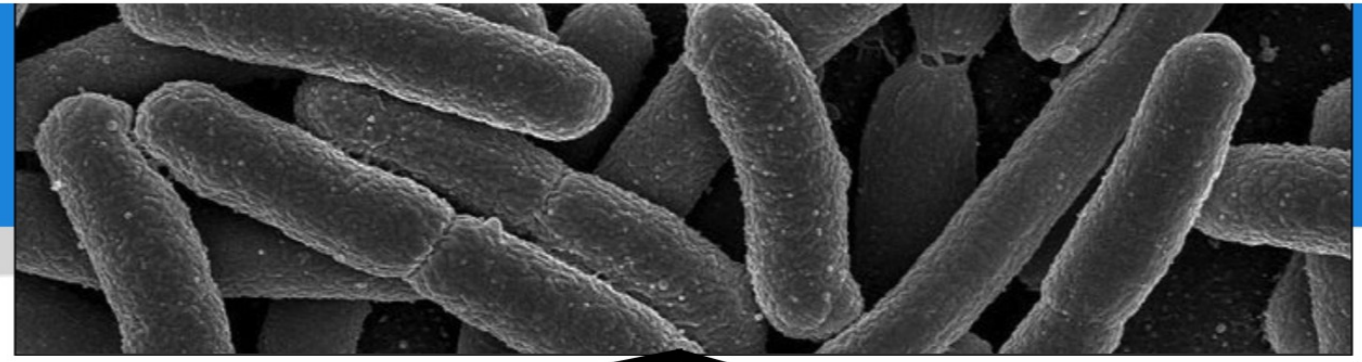
How to identify microbial community?

- **We are capable of sequencing and analyzing the genomes of culturable species**

- **Culture dependent analysis**:
  - Culture and obtain pure colonies
  - Complete genome sequencing of DNA
  - Organism has to be cultured in the laboratory

- **Culture-independent analysis**
  - 16S ribosomal RNA (rRNA) sequencing

- **Metagenomics:** sequence based analysis of complete microbial communities without need for culturing



16S rRNA sequencing

Taxonomic Content

Metagenomic sequencing

Functional predictions based on gene content

Long Reads + Better Assembly = group genes by genome

Metatranscriptomic sequencing

Functional predictions based on gene expression

Other 'Omics
Proteome
Glycome
etc

Changes in Sequencing Technologies

ABI 3730*xl* ***1-2 Mb/day***

Illumina GA IIx
50 Gb/12day run

ABI SOLiD
100Gb/12 day run

454 GS FLX +
0.6Gb/23hr run

Illumina HiSeq 2000 (2500†)
600 Gb/11day run

Ion Torrent
1Gb/2hr run

Ion Proton
100Gb/4 hr Run

†***HiSeq 2500 upgrade***: *up to 120Gb/27 hour run (available now for $50K)*

Higher capacity for
lower cost in
less time
than HiSeq 2000

Illumina NextSeq 500

Main platform for 16S rRNA

Illumina MiSeq

Shotgun Metagenomics

“The study of genetic material recovered directly from environmental samples”

Why Metagenomics?

- What is there?
- How many are there?
- What are they doing?
- Experimental manipulations
- Diagnostics


icanhasGIF.com

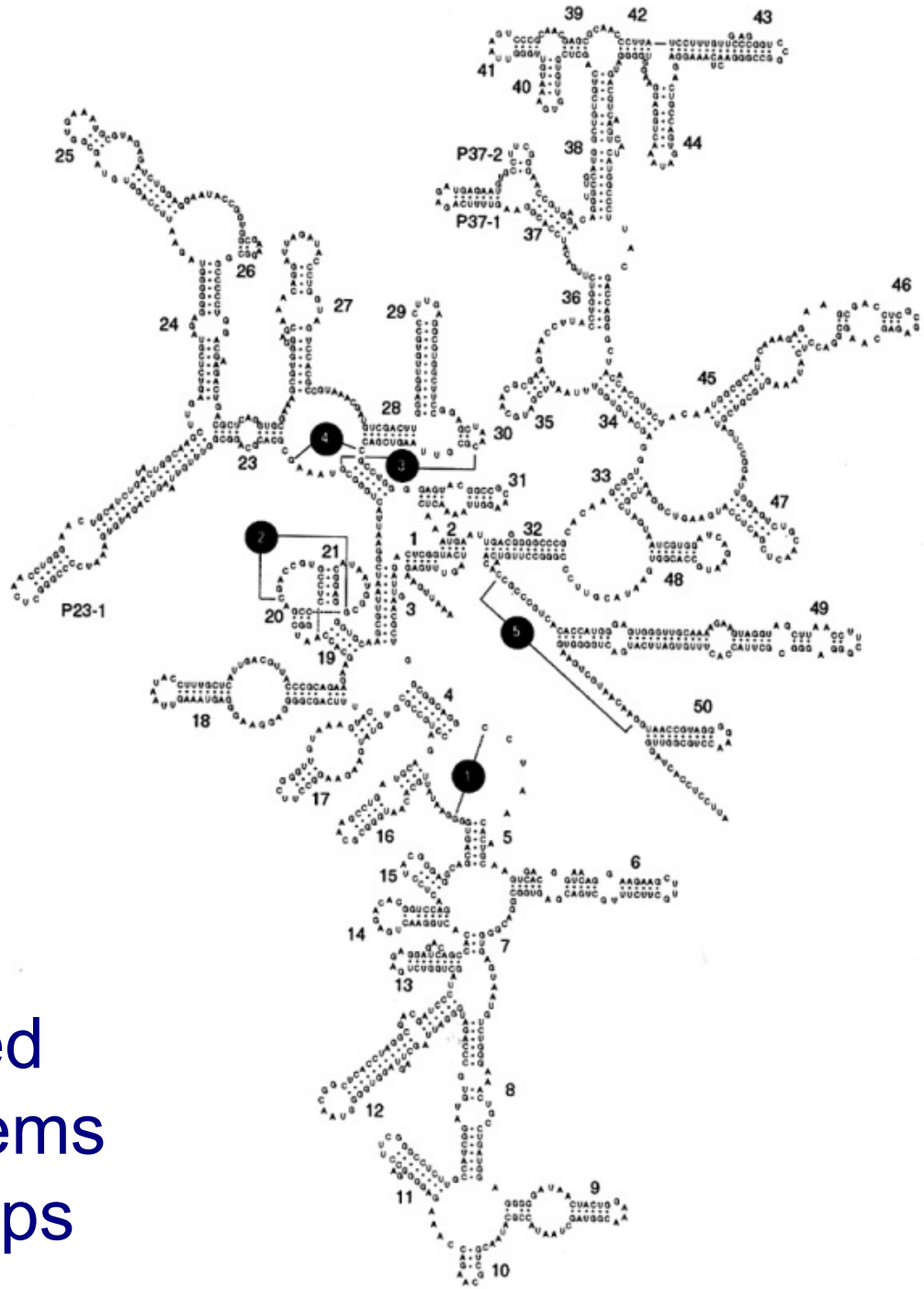
Method	Read-based	Assembly-based	Detection-based
Description	Read-based metagenomics analyzes unassembled reads. It was one of the first methods to be used. It is still valuable for quantitative analysis, especially if relevant references are available.	Assembly-based workflows attempt to assemble the reads from one or more samples, “bin” contigs from these samples into genomes then analyze the genes and contigs.	Detection-based workflows attempt to identify with very high-precision but lower sensitivity (recall) the presence of organisms of interest, often pathogens.

Method	Read-based	Assembly-based	Detection-based
Typical questions	• What is the bulk taxonomic/functional composition of these samples? • What new kinds of a particular functional gene family can I find? • How do my sites or treatments differ in taxonomic/functional composition?	• What are the functional and metabolic capabilities of specific microbes in my sample? • What is the phylogeny of gene families in my samples? • Do the organisms that inhabit my samples differ? • Are there variants within taxa in my population?	• Are known organisms of interest present in my sample? • Are known functional genes e.g. beta-lactamases, present in my sample?
Examples	MG-Rast	IMG from JGI	Taxonomer, Surpi, One Codex, CosmosID

16S sequencing

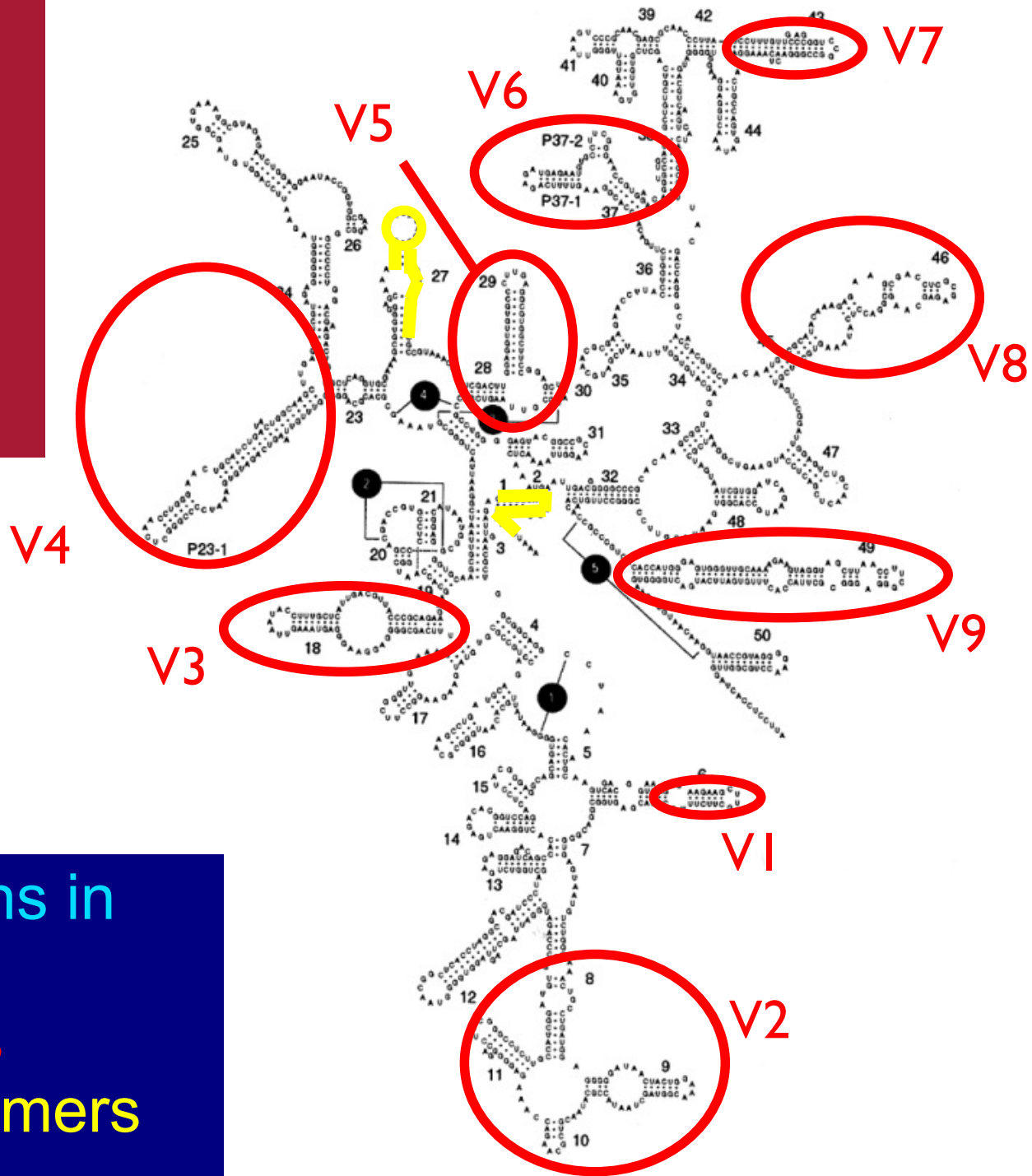
- Catalogs the bacteria that are present
- PCR amplify the 16S gene with standard primers
- Sequence the primers
- Compare to known databases

E. coli 16S rRNA secondary structure



- Highly conserved
- Base pairs = stems
- No pairing = loops

E. coli 16S rRNA secondary structure



Variable regions in the 16S rRNA.

Vn – 9 regions

forward/rev primers

16S Primers

- 27F – 1492R full length

1,465 base pairs

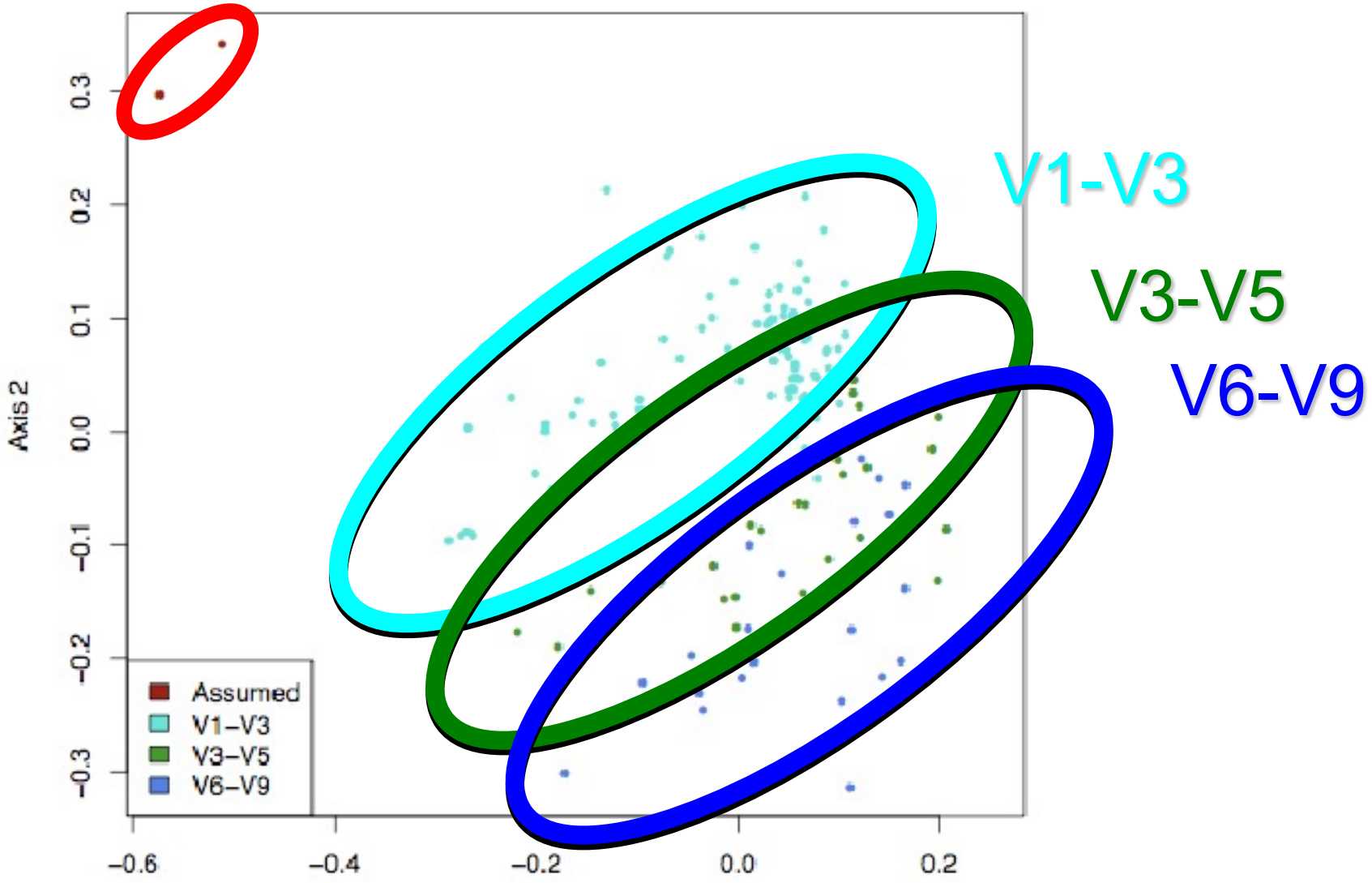
- 967F – 1046R V6 region

79 base pairs

- 1380F – 1510R V9 region

130 base pairs

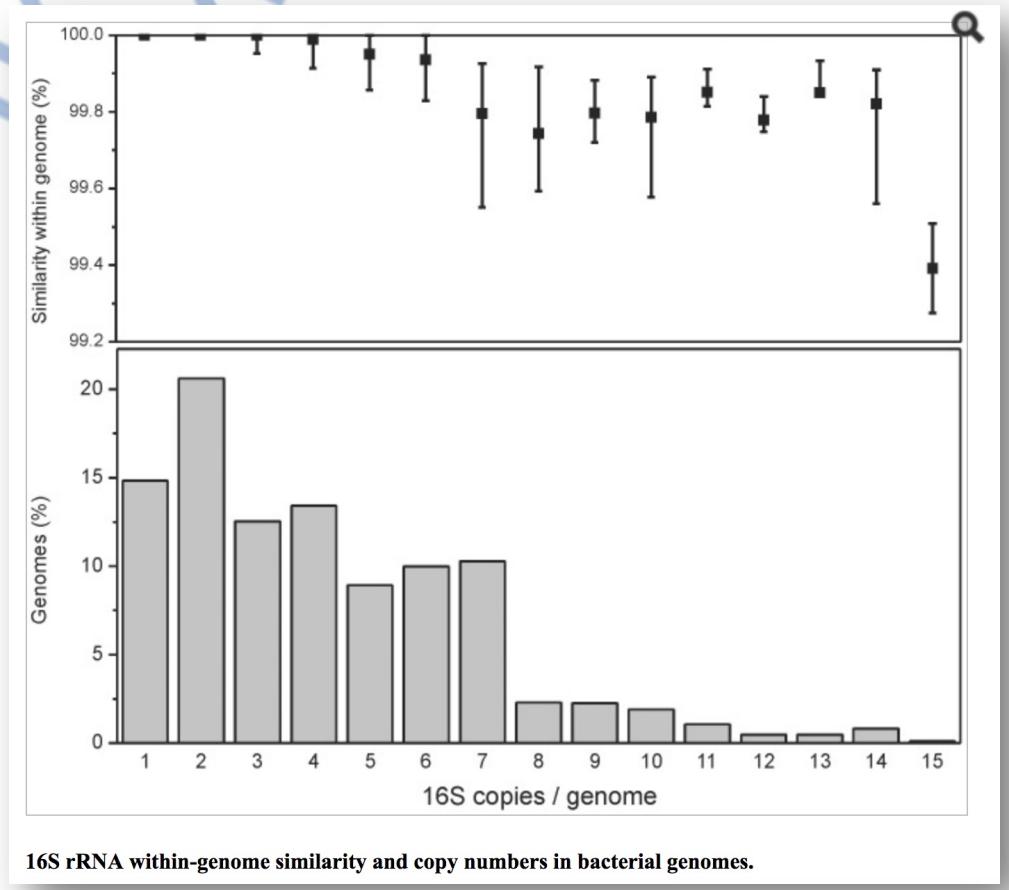
Variable regions = Variable results!



16S databases

- Greengenes
  - http://greengenes.lbl.gov/
  - Gary Andersen, Lawrence Berkeley National Laboratory
- SILVA – ARB
  - http://www.arb-silva.de/
  - Frank Oliver Glöckner, MPI, Bremen, Germany
- VAMPS
  - http://vamps.mbl.edu/
  - Mitch Sogin, Woods Hole, USA
- Ribosomal Database Project (RDP)
  - http://rdp.cme.msu.edu/
  - James Cole, Michigan State University, USA

16S sequencing



16S rRNA within-genome similarity and copy numbers in bacterial genomes.

PCR bias

Variable regions give variable answers

Only tells you which organisms are present & abundance

Does not explain much of the variance of the data

What does 16S sequencing actually tell you?

What would you do?

- **Sequencing**
  - 16S rRNA amplicon sequencing
  - $20 - $40 per sample
  - Simple, fast (1 month turn-around)

  - Shotgun metagenomics
  - $300 - $500 per sample
  - Complex (6 months turn-around)

- What would you do ?
  - Hypothesis
  - Number of samples
  - Sequencing technique
  - Rationale behind choice of sequencing technique

Microbiome 16S Analysis: A Quick-Start Guide

When to Use Marker Gene

- 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

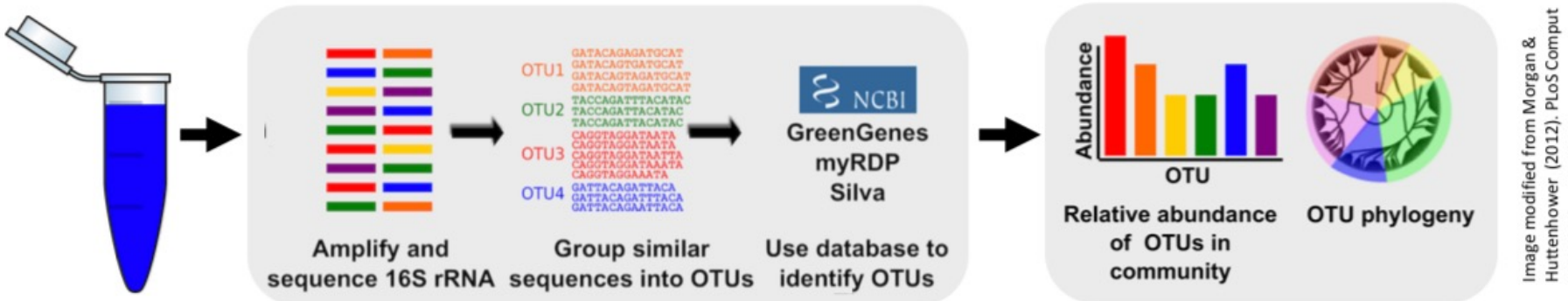


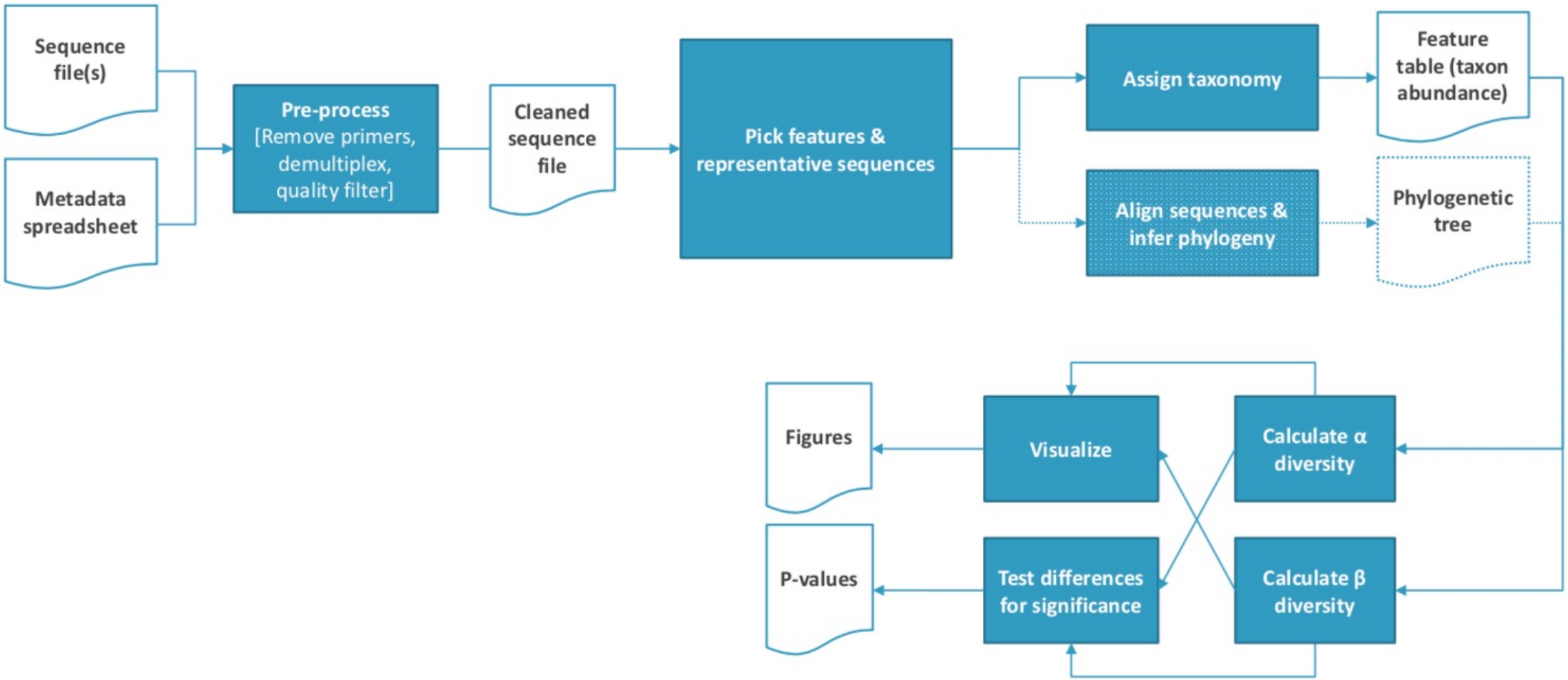
Image modified from Morgan & Huttenhower (2012). PLoS Comput Biol 8(12): e1002808.

When to Use Marker Gene

- 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

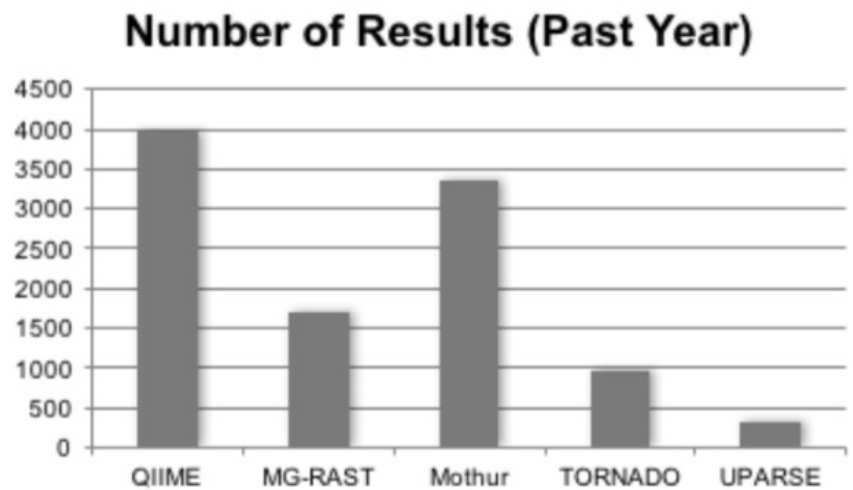
- The good news:
  - Target gene studies are slightly cheaper to prep and sequence than shotgun ones
  - Analysis software is mature, and many studies can be analyzed on a laptop
  - Known taxa can be detected with very low (100s of reads) sequence depth
- The bad news
  - No target gene distinguishes all microbes well
    - 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



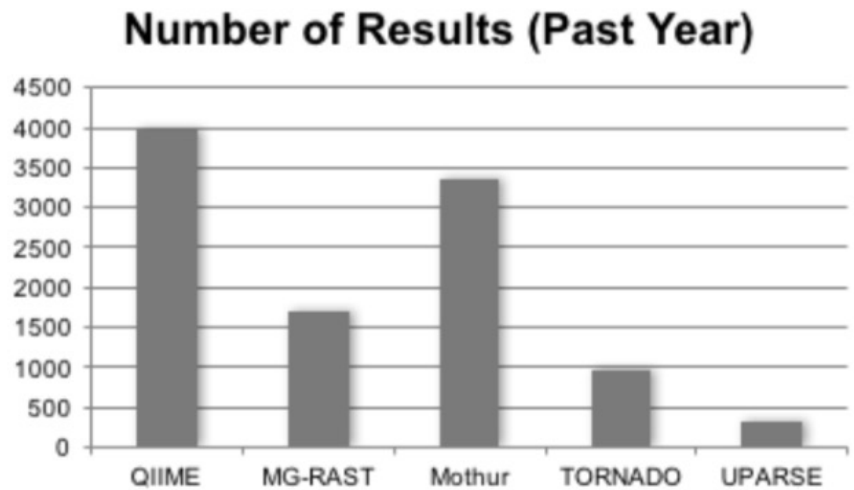
Software Selection

- Google “16S analysis <program name>”; main contenders are
- Mothur
  - Name: not an acronym (play on DOTUR, SONS)
  - Philosophy: single piece of re-implemented software
  - Top pro: easy to install
  - Top con: re-implementations could be buggy
  - Language: C++
  - Model: open-source
  - License: GPL
  - Published: 2009
  - Developed: at Umichigan



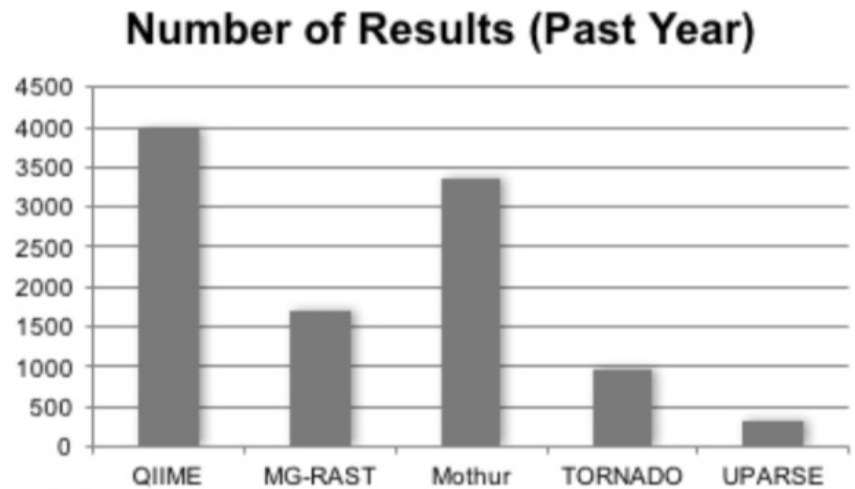
Software Selection

- Google “16S analysis <program name>”; main contenders are
- QIIME
  - Name: Quantitative Insights Into Microbial Ecology
  - Philosophy: wrapper of best-in-class software
  - Top pro: extremely flexible
  - Top con: QIIME 2 not yet feature-complete
  - Language: python (wrapper)
  - Model: open-source
  - License: mixed
  - Published: 2010
  - Developed: At UCSD, NAU



Software Selection

- Google “16S analysis <program name>”
  - Main contenders are Mothur and QIIME
  - Both widely used
  - Both pride themselves on quality of support
- Will discuss only QIIME in this tutorial
- QIIME 1 vs QIIME 2
  - QIIME 1 won’t be supported after end of 2017
  - QIIME 2 not yet feature-complete
    - But already much easier to use!
  - This tutorial uses QIIME 2 **only**
- **I’m not a QIIME 2 developer**
  - I’m not taking credit for this tool, just demonstrating it!



Getting the Software & Data

- Not covered in this tutorial, for sake of time
- QIIME 2 is very easy to install with the Conda environment- and package-manager
  - Conda is also very easy to install—either Miniconda or Anaconda versions

```
wget https://data.qiime2.org/distro/core/qiime2-2021.2-py36-linux-conda.yml

conda env create -n qiime2-2021.2 --file qiime2-2021.2-py36-linux-conda.yml

# OPTIONAL CLEANUP
rm qiime2-2021.2-py36-linux-conda.yml

conda activate qiime2-2021.2
```