Metagenomics: Characterization of Microbial Communities using NGS

Brandi Cantarel, PhD
BICF
08/02/2016

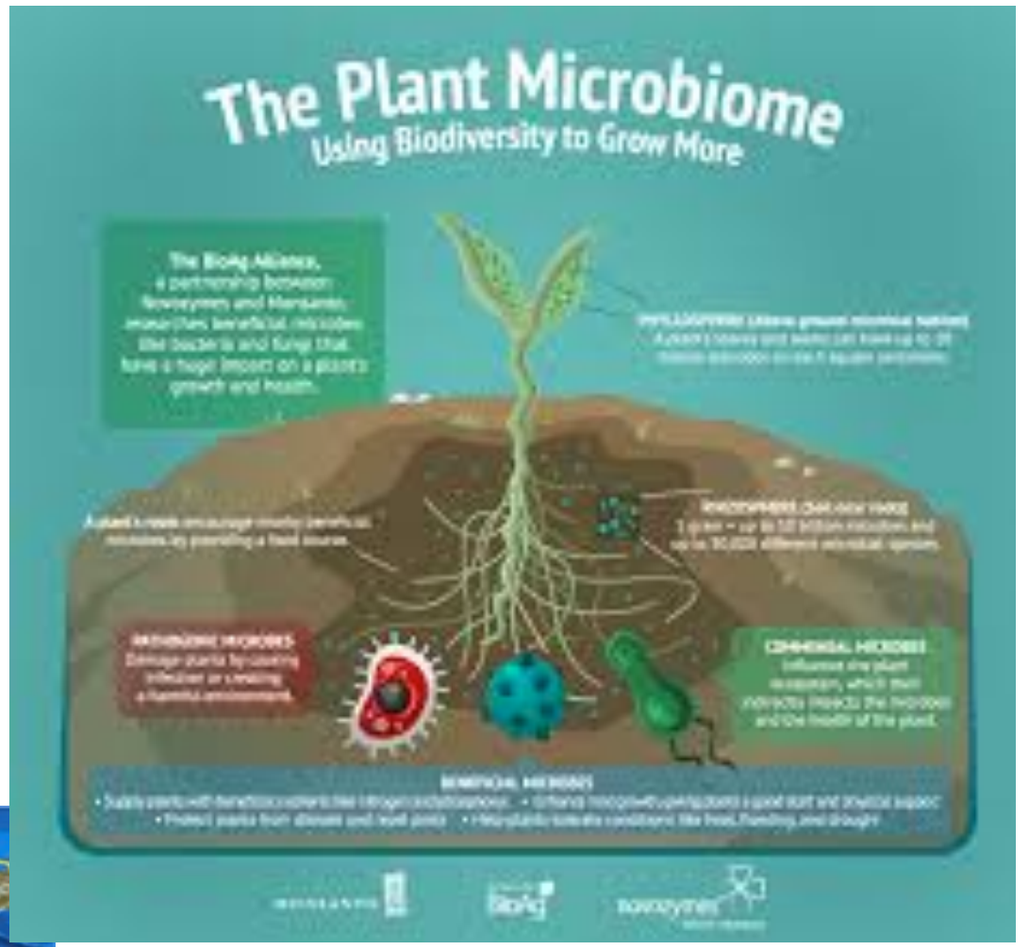
- *Introduction to Metagenomics*
  - What is a microbiome, metagenome,the relationship between the micro biome and it's environment?
  - Whole Community Sequencing Methods vs Traditional Culture Methods
  - Genetic Diversity in a Microbiome
  - Early Microbiome Projects
- *Sampling, DNA Extraction and Sequencing*
  - Comparison of Sequencing Technologies
  - Data quality: Error rates of sequencing, chimeras
  - Differences in profiles depending on sampling, DNA extraction and sequencing
- *Measuring Taxonomic Composition and Diversity using Marker Genes (16S)*
  - What is the 16S gene? Why is it so special?
  - Alternatives to 16S: ITS, CPN60, RecA
  - Data Processing workflow
  - 16S gene databases: RDP, SILVA, GreenGenes
  - Taxonomy assignment and OTUs
  - Analysis platforms: Qiime and Mother
  - Phylogenetic Trees, Rarefraction and Diversity and PCoA

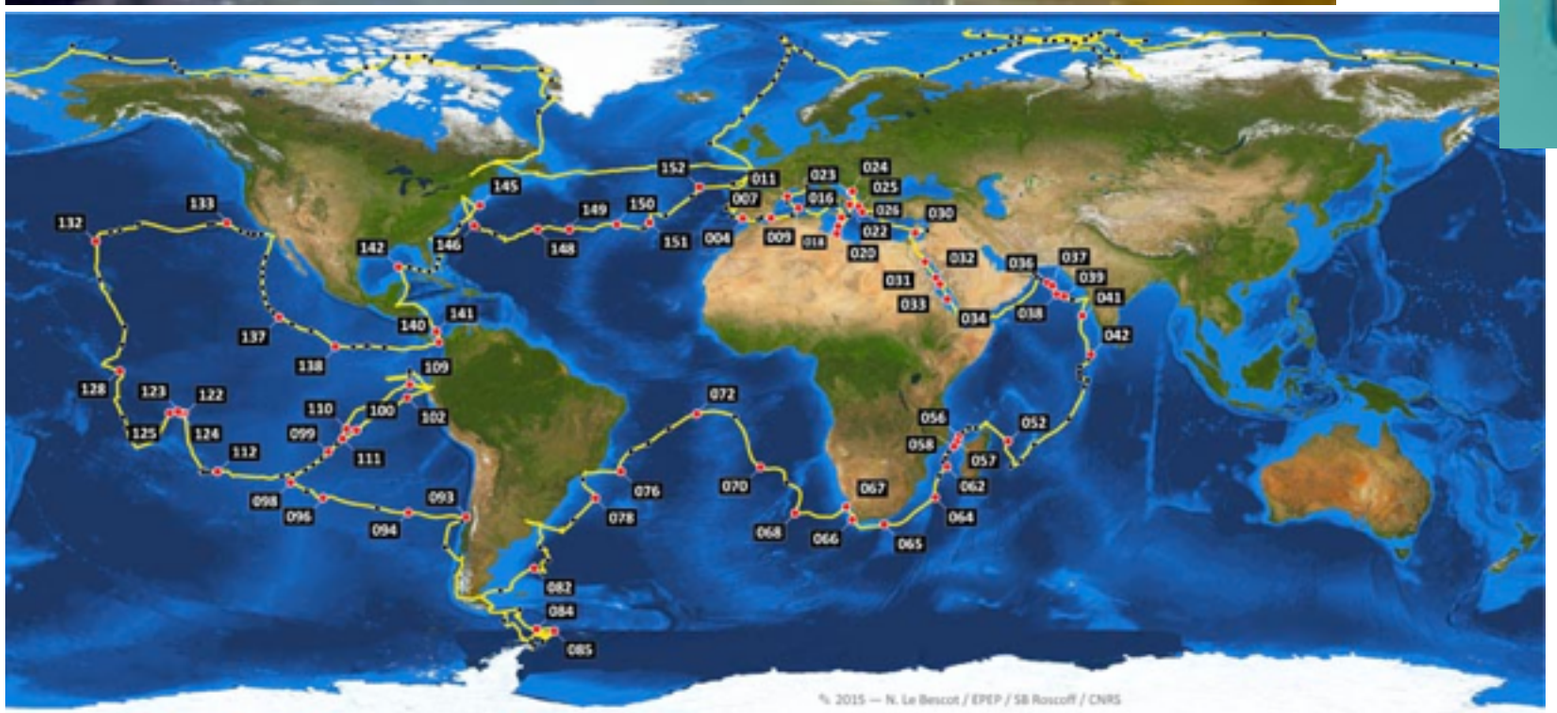
- *Introduction to Metagenomics*
  - What is a microbiome, metagenome,the relationship between the micro biome and it's environment?
  - Whole Community Sequencing Methods vs Traditional Culture Methods
  - Genetic Diversity in a Microbiome
  - Early Microbiome Projects
- *Sampling, DNA Extraction and Sequencing*
  - Comparison of Sequencing Technologies
  - Data quality: Error rates of sequencing, chimeras
  - Differences in profiles depending on sampling, DNA extraction and sequencing
- *Measuring Taxonomic Composition and Diversity using Marker Genes (16S)*
  - What is the 16S gene? Why is it so special?
  - Alternatives to 16S: ITS, CPN60, RecA
  - Data Processing workflow
  - 16S gene databases: RDP, SILVA, GreenGenes
  - Taxonomy assignment and OTUs
  - Analysis platforms: Qiime and Mother
  - Phylogenetic Trees, Rarefraction and Diversity and PCoA

What is a Microbiome?

- A term coined by Joshua Lederberg
- The ecological community of commensal, symbiotic and pathogenic microorganisms
- All plants and animals, from protists to humans, live in close association with microbial organisms.
- The hologenome theory proposes that the object of natural selection is not the individual organism, but the organism together with its associated microbial communities.

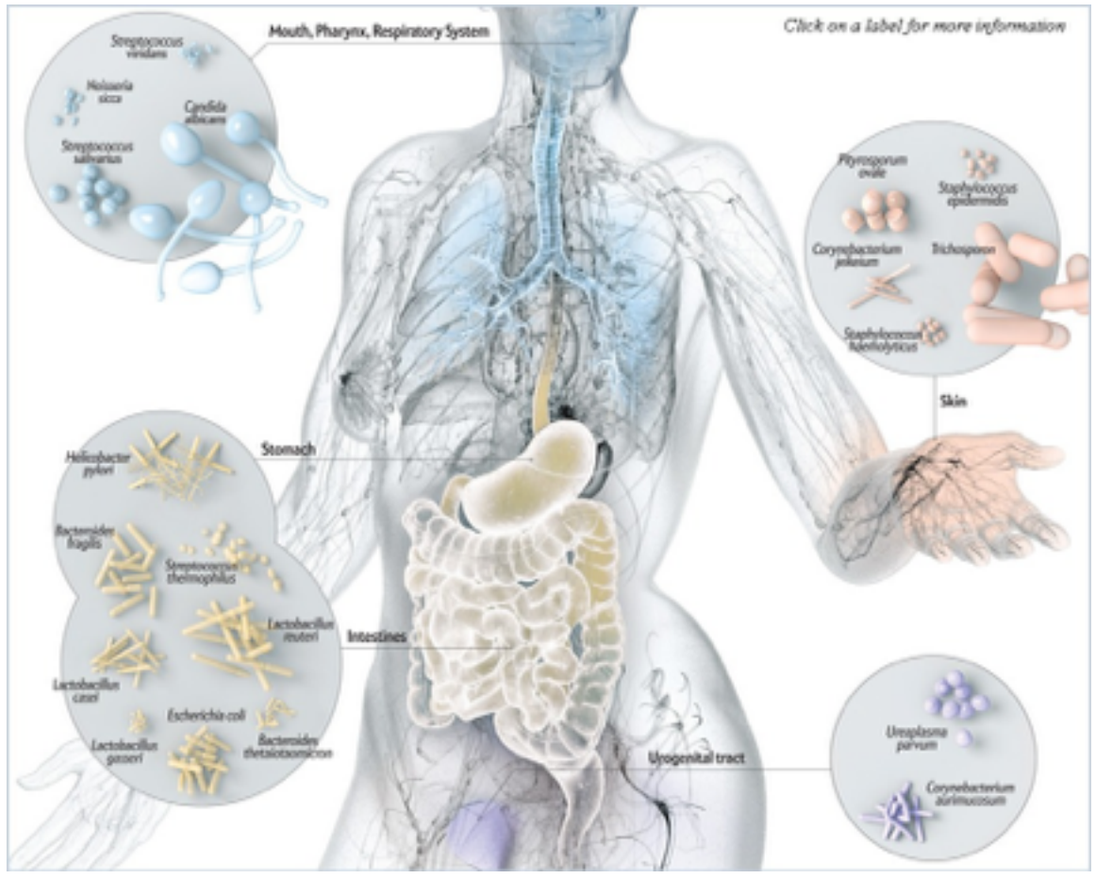
Microbiomes Are Everywhere



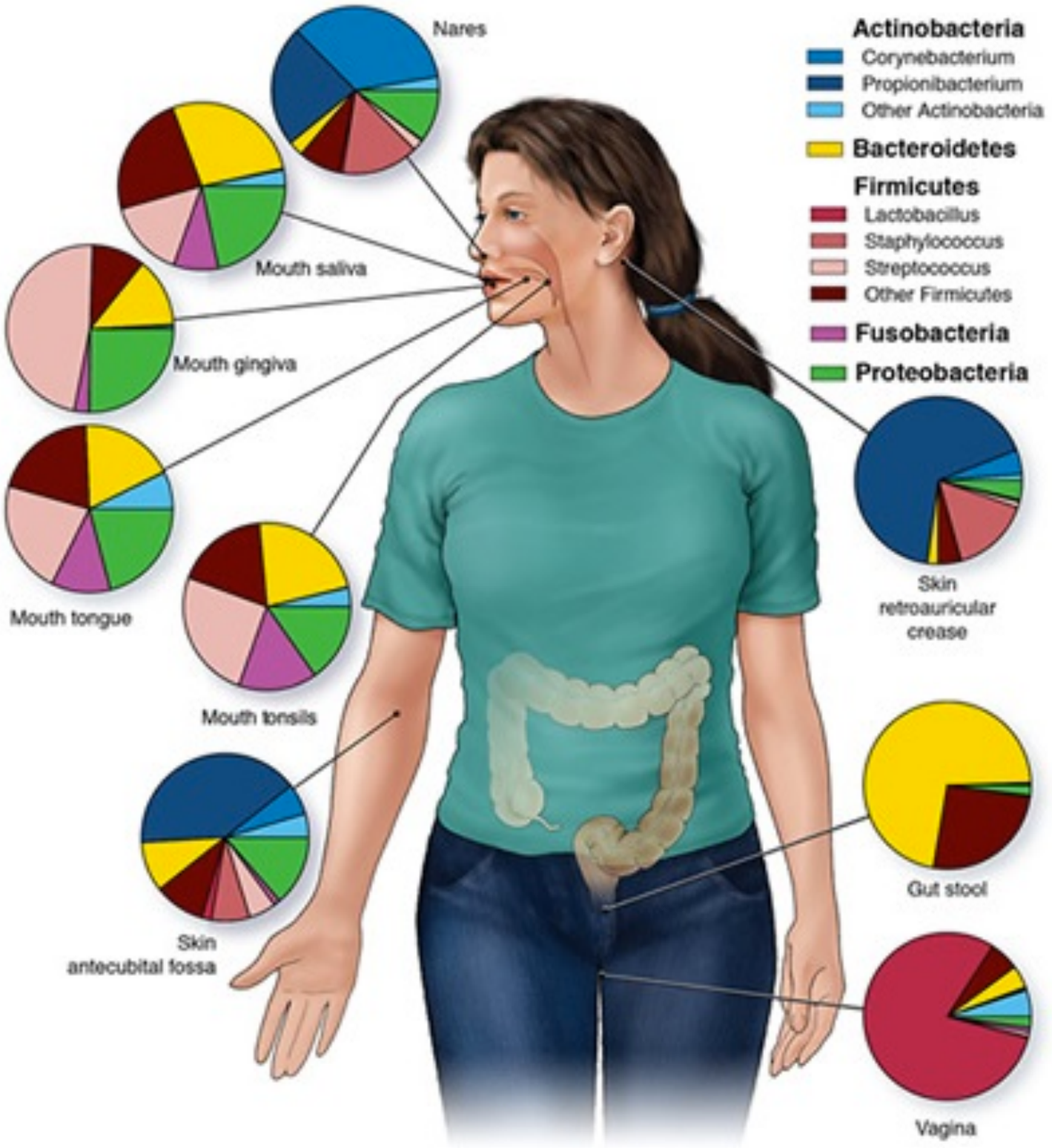


Metaorganisms (Superorganisms)

- Animal bodies (including humans) are superorganisms.
- Composed of microbial and animal cells
- Microbes are important for digestion, immune development and other functions essential for survival

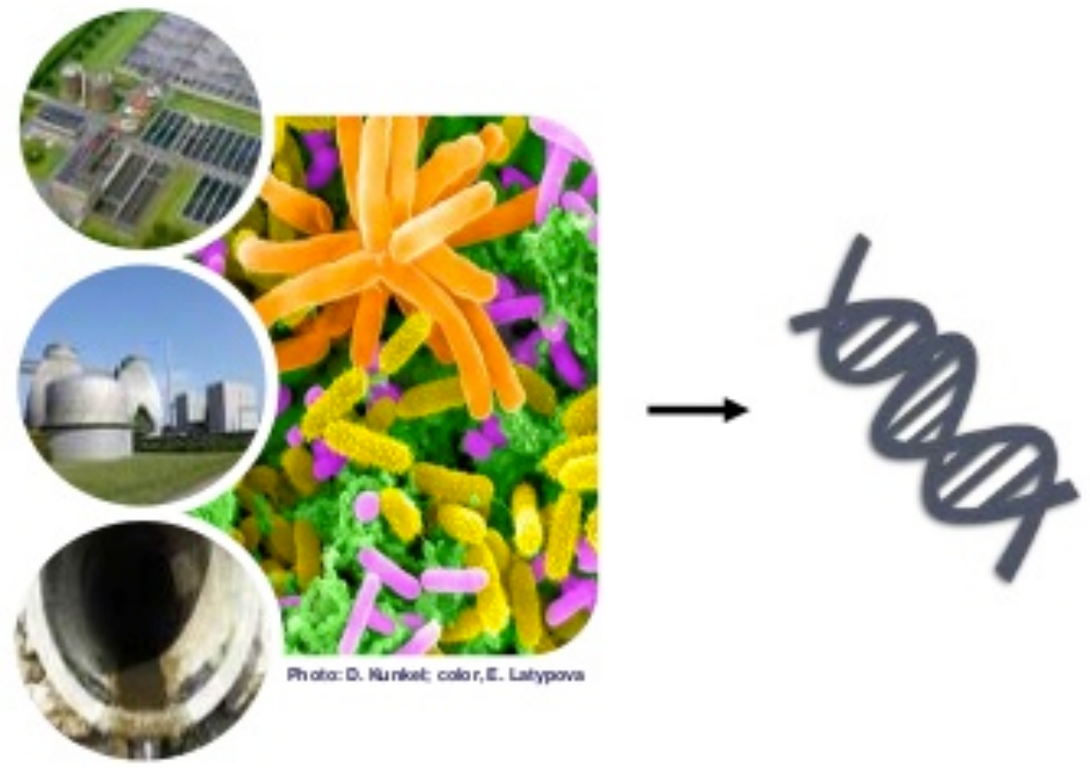


The Human Body is Composed of Many Small Ecosystems (Microbiomes)



What is a Metagenome?

- The term "metagenomics" was first used by Jo Handelsman, Jon Clardy, Robert M. Goodman, Sean F. Brady, and others, and first appeared in publication in 1998.



- A metagenome is the collection of genes in a microbial community.

- Metagenomics is the study of genetic material from an environmental sample

- Offers a culture independent methods

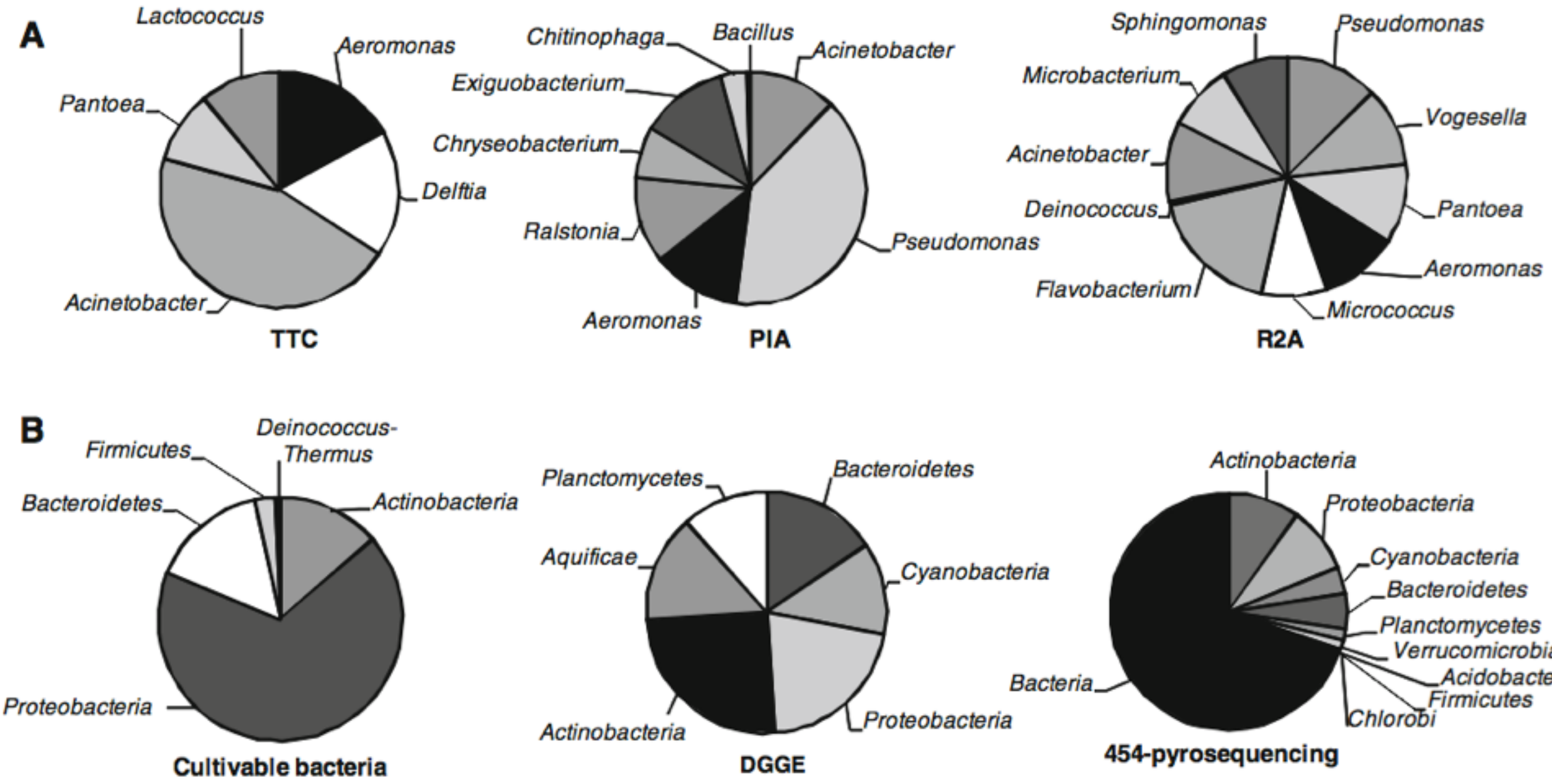
Emerging Metagenomics

- Late 17th Century, Anton van Leeuwenhoek
  - First metagenomicist who directly studies organisms from pond water and his own teeth
- 1920s
  - Cell culture evolved, 16S rRNA sequencing of cultural microbes
  - Is an organism could not be cultured, it could not be classified
- Discrepancies observed:
  - Number of organisms under microscope in conflict with amount on plates
  - Cellular activities in situ conflicted with activities in culture
  - Cells are viable but unculturable

Traditional Culture Dependent Profiling

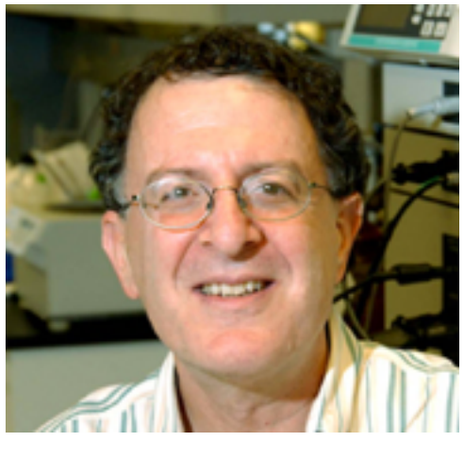
- Culture-dependent methods have traditionally been used to identify microbes living in an environment
- It's estimated that only about <1% of microorganisms can be grown in culture
  - Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev. 1995 Mar;59(1):143-69. Review. PubMed PMID: 7535888; PubMed Central PMCID: PMC239358.
- Even if all microbes could be grown in culture, it would be a daunting task to determine growth conditions for ALL microbes

Culture Dependent Profiles Depends on Culture Methods



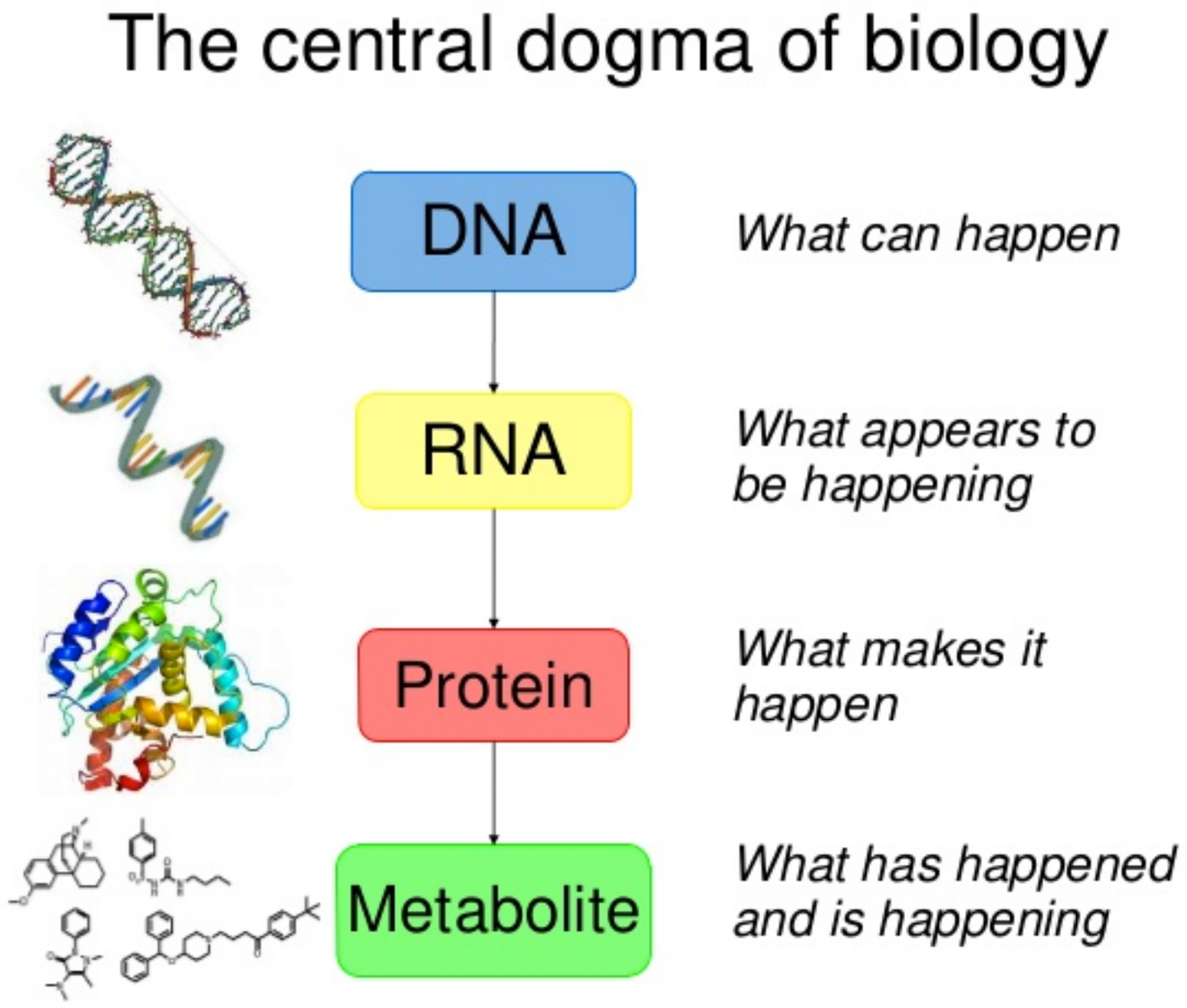
“Modern” Metagenomics

- Norman Pace proposes the idea of cloning DNA directly from environmental samples in 1985
  - First report by Pace et al in 1991
- Ribosomal Database Project (RDP) is funded in 1989
- Metagenomics is defined 1998
- Metagenomic discovery of proteorhodopsin in 2000
- Term Microbiome is coined in 2001
- Sargasso and Acid Mine Drainage Projects published 2004
- Global Ocean Sampling Expedition in 2007
- Human Microbiome and MetaHit projects are launched in 2008



Understanding Interactions between Microbial Communities and Environment

- Experimental and computational techniques are necessary to make inferences about the community:
  - Community Structure
  - Gene Content
  - Expression
  - Translation
  - Metabolites



Much of the Early Studies Focused on DNA

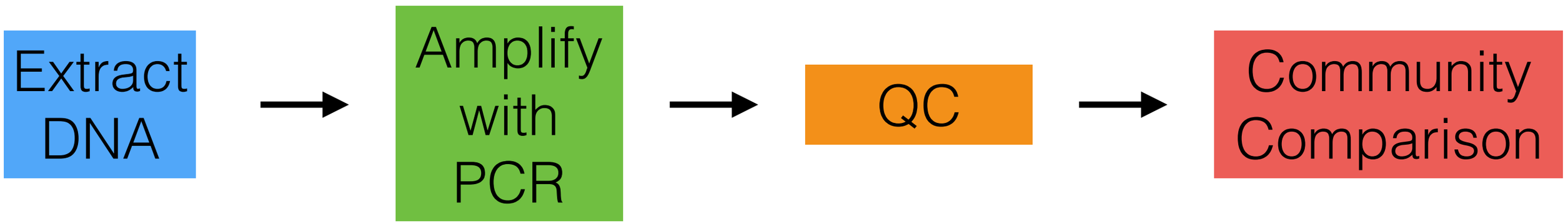
DNA

What can happen

- Community Structure
- Gene Content

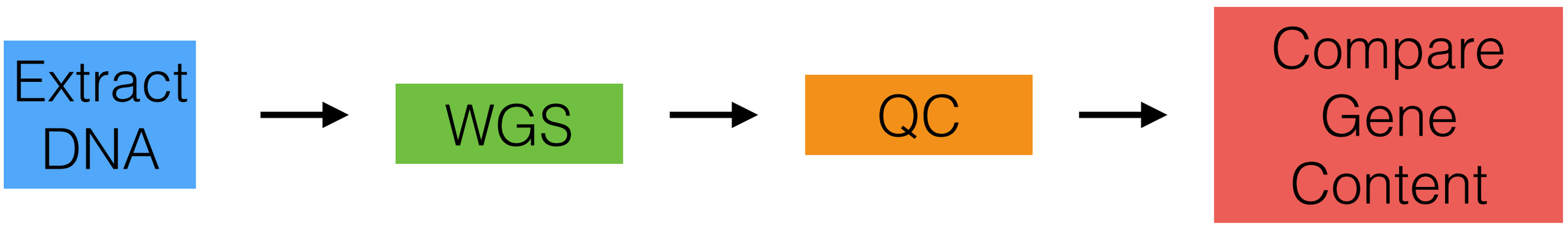
Using Marker Genes to Examine Community Structure

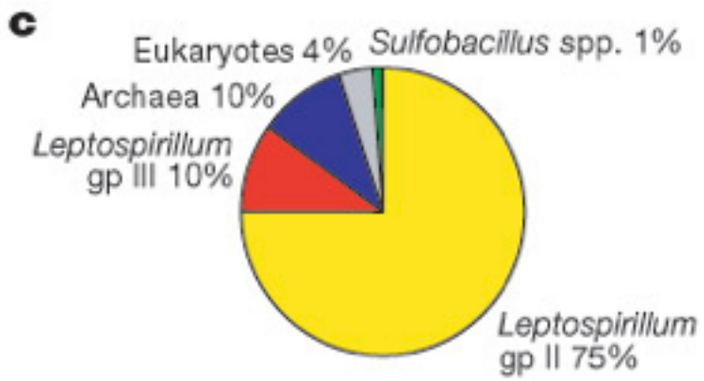
- Asking the questions: what organisms are there?
- A good marker gene needs to be:
  - Present in all organisms compared
  - Vertically and slowly evolving
  - Amplify-able with small set of “universal primers”
  - Have an established database of reference

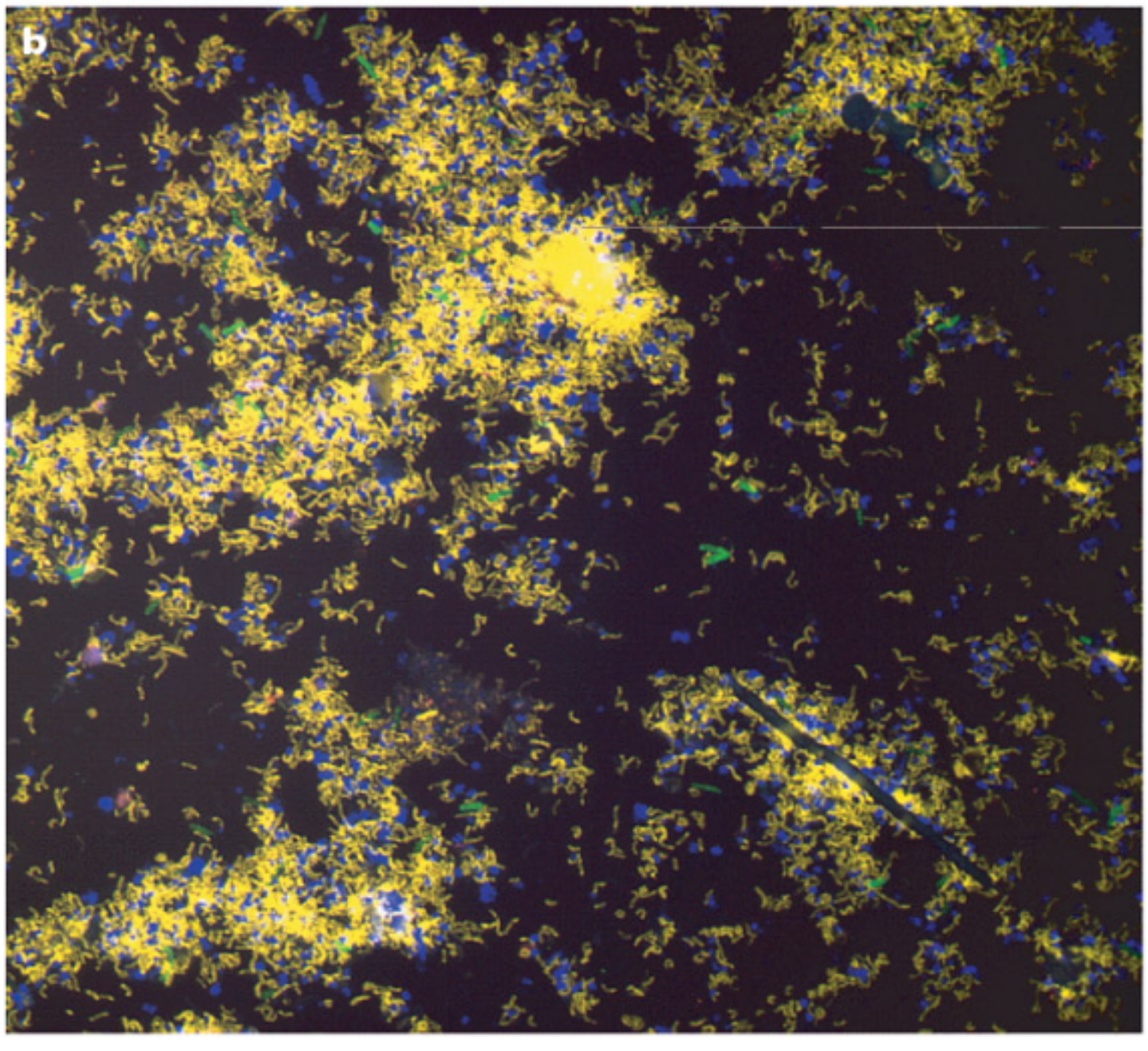


Whole Genome Shotgun

- Asking the questions: what is the functional capability?
- WGS can
  - Identify organisms present — if closely related to organisms with sequenced genomes
  - Identify gene families present — if homologs have been functionally characterized
  - Identify functional pathways present — if homologs have been annotated to gene pathways
  - Identify new species/strains — if assemblies are of sufficient depth

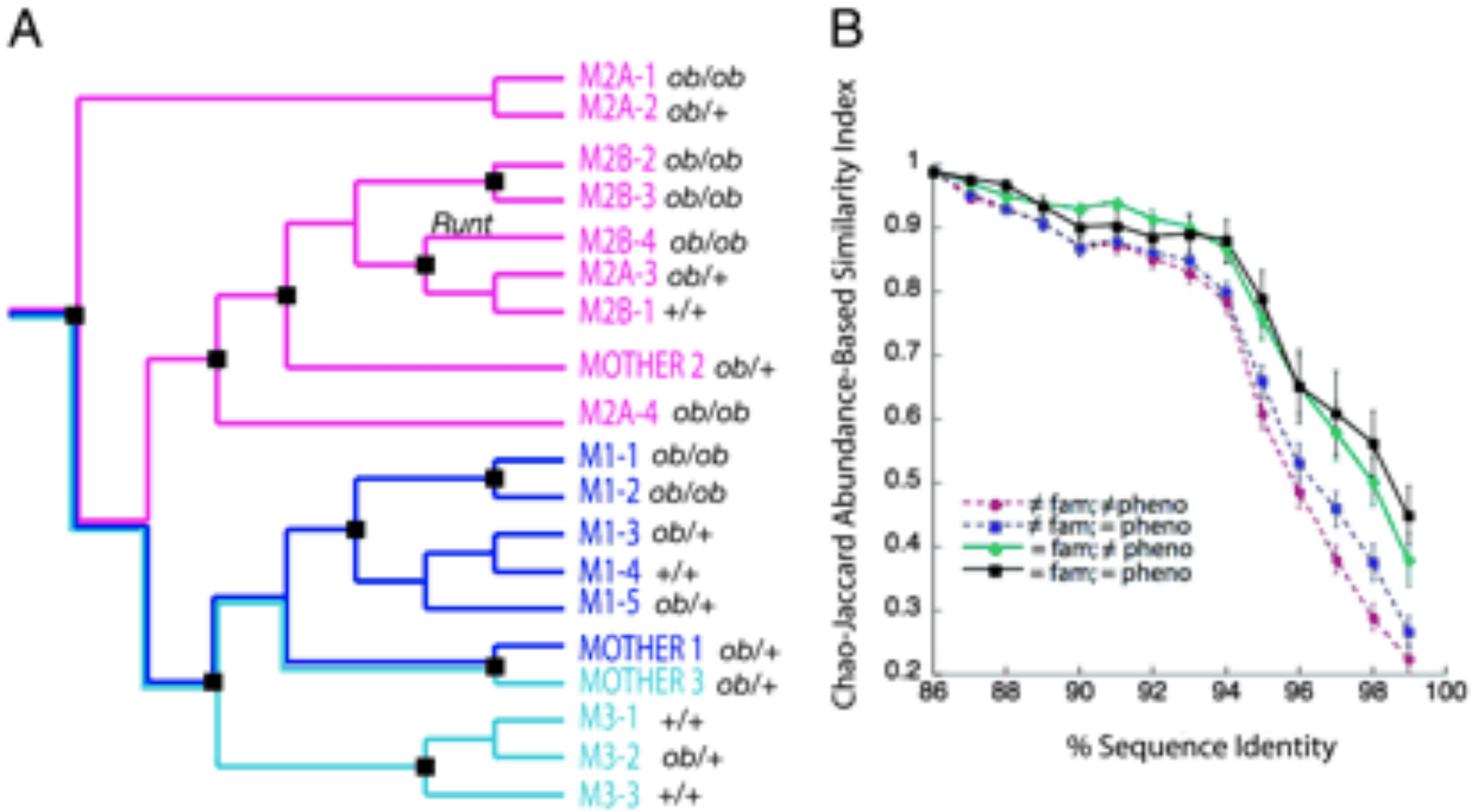


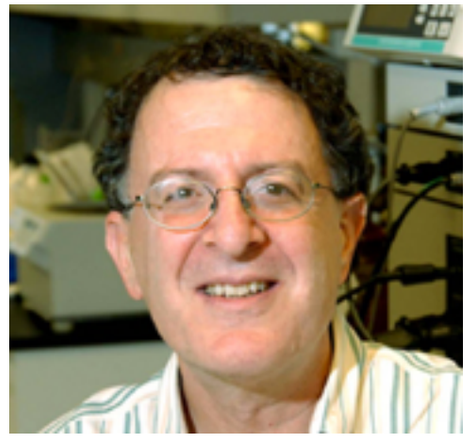


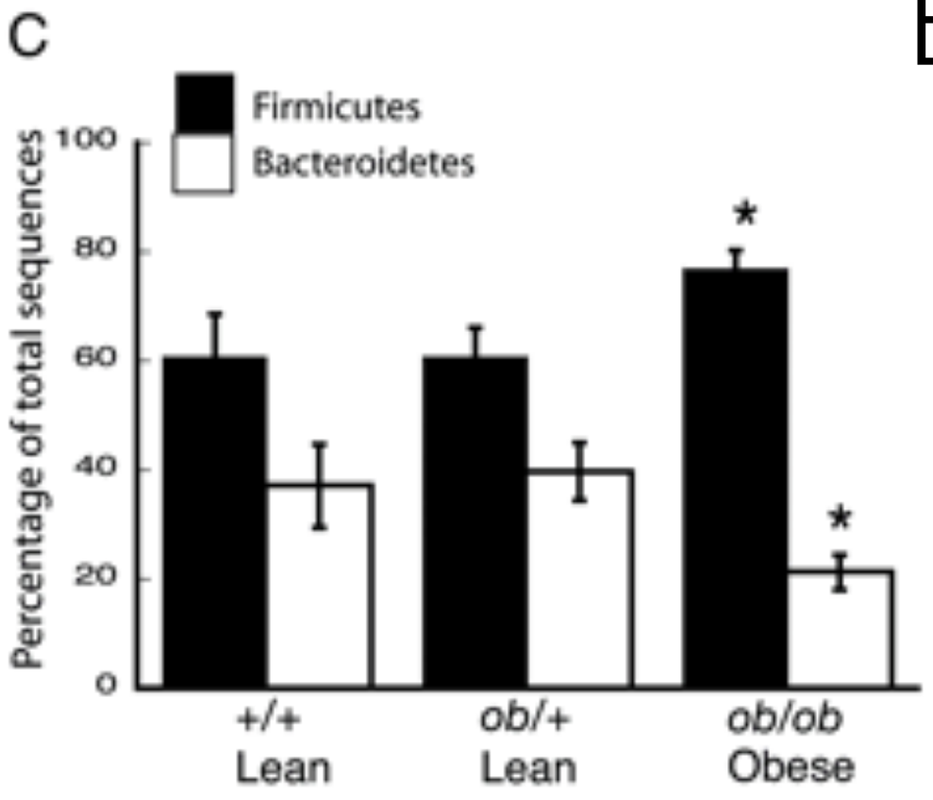


Acid Mine Drainage

Community structure and metabolism through reconstruction of microbial genomes from the environment Gene W. Tyson, Jarrod Chapman, Philip Hugenholtz, Eric E. Allen, Rachna J. Ram, Paul M. Richardson, Victor V. Solovyev, Edward M. Rubin, Daniel S. Rokhsar and Jillian F. Banfield
Nature 428, 37-43(4 March 2004)



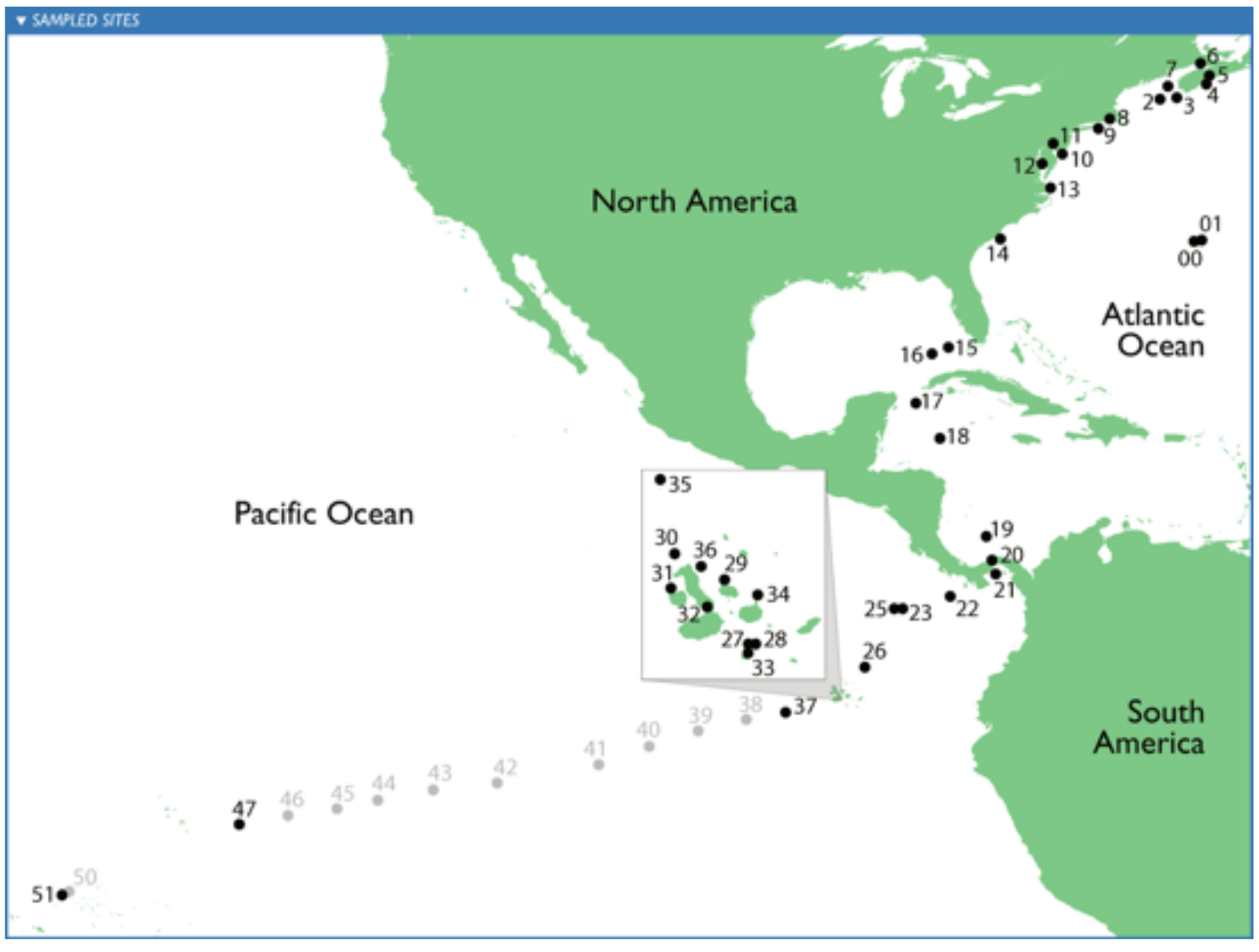


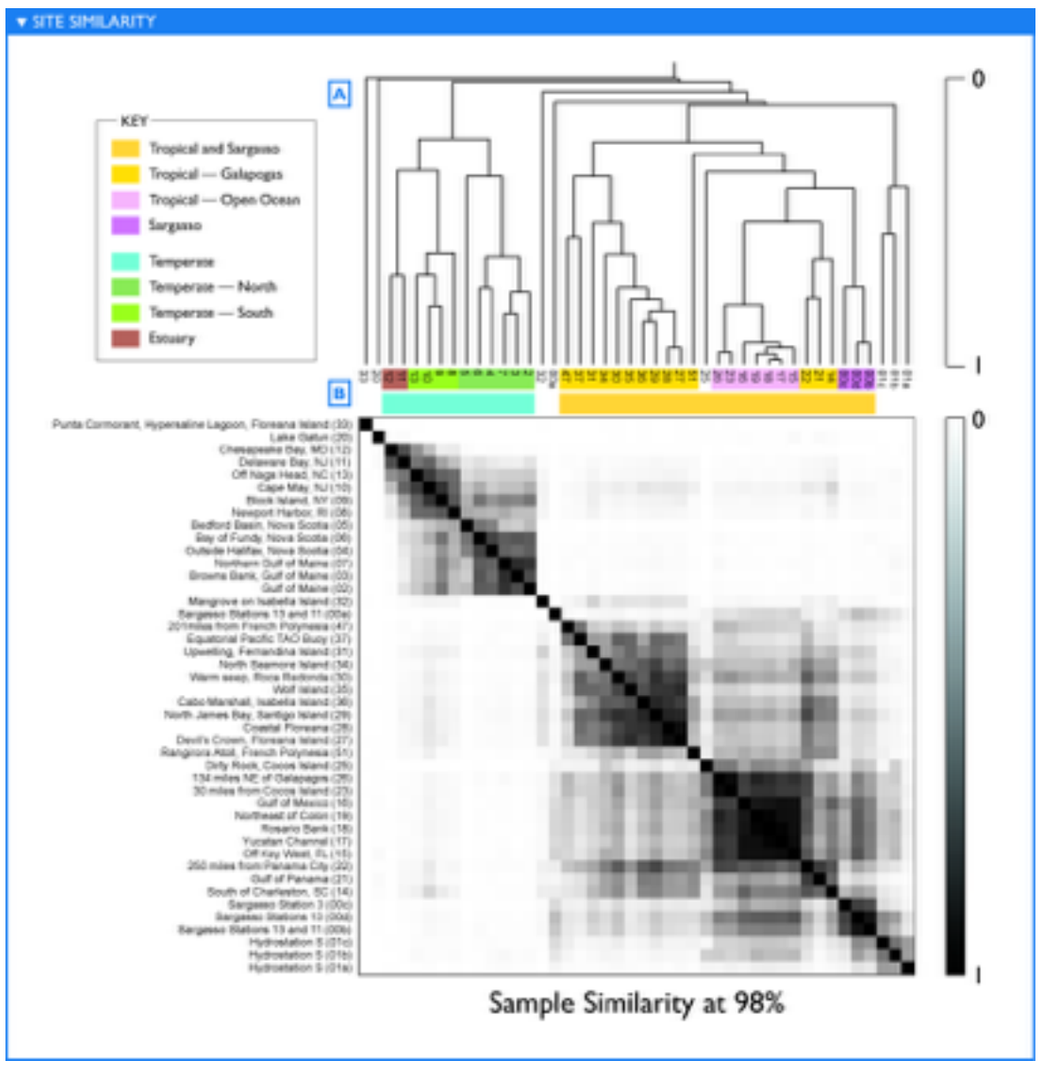


Effects of kinship and obesity on gut microbial ecology.

Ley RE, Bäckhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A. 2005 Aug 2;102(31):11070-5. Epub 2005 Jul 20. PubMed PMID 16033867; PubMed Central PMCID: PMC1176910.

Global Ocean Expedition



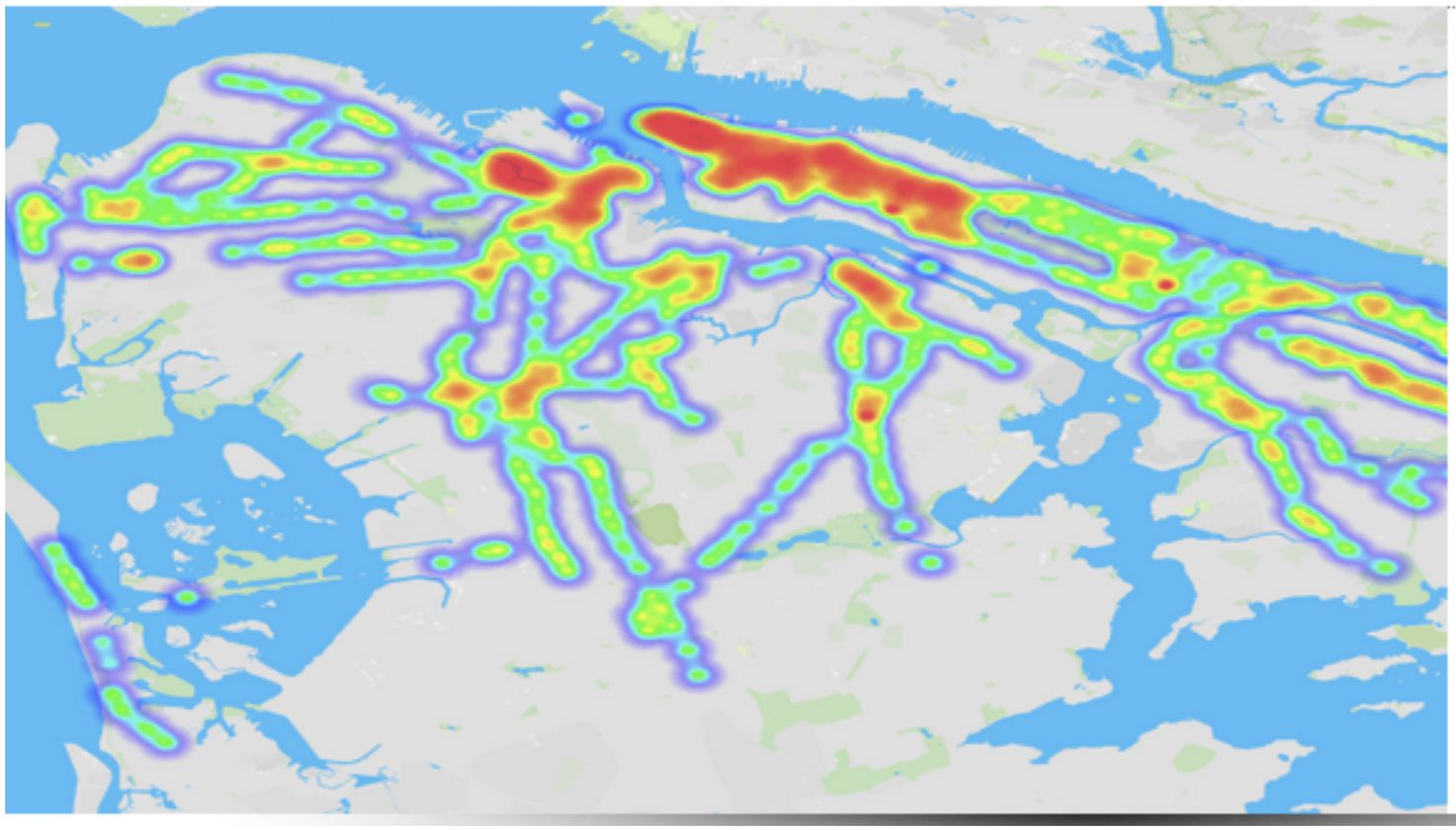


Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcón LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. 2007 Mar;5(3):e77. PubMed PMID: 17355176; PubMed Central PMCID: PMC1821060.

New York City Microbiome Project

Creating a Molecular Portrait of New York City

One swab at a time



PathoMap is a research project by **Weill Cornell Medical College** to study the microbiome and metagenome of the built environment of NYC.

Check out the full manuscript published at Cell Systems **here**.

Raw data is now online **here**.

Check out the PathoMap Ancestry Pipeline now set up on the Arvados platform **here**!

http://www.pathomap.org/

- *Introduction to Metagenomics*
  - What is a microbiome, metagenome,the relationship between the micro biome and it's environment?
  - Whole Community Sequencing Methods vs Traditional Culture Methods
  - Genetic Diversity in a Microbiome
  - Early Microbiome Projects
- *Sampling, DNA Extraction and Sequencing*
  - Comparison of Sequencing Technologies
  - Data quality: Error rates of sequencing, chimeras
  - Differences in profiles depending on sampling, DNA extraction and sequencing
- *Measuring Taxonomic Composition and Diversity using Marker Genes (16S)*
  - What is the 16S gene? Why is it so special?
  - Alternatives to 16S: ITS, CPN60, RecA
  - Data Processing workflow
  - 16S gene databases: RDP, SILVA, GreenGenes
  - Taxonomy assignment and OTUs
  - Analysis platforms: Qiime and Mother
  - Phylogenetic Trees, Rarefraction and Diversity and PCoA

Historic Sequencing Technologies

- Sanger
  - DNA Sequencing with chain-terminating inhibitors (1977)
  - Used in Early Microbiome Studies
  - Considered Gold Standard for Accuracy

- Roche 454
  - Pyrosequencing (2004)
  - Used for many 16S, and early WGS Sequencing studies
  - Inaccuracies accumulated in homopolymer regions

Commonly Used Sequencing Technologies

- Ion Torrent
  - 400bp reads
  - Inaccuracies accumulated in homopolymer regions
  - ~ $0.63/Mbp — Hardware ~$70K/machine
  - Low upfront and maintenance costs makes it attractive to independent labs
- Illumina HiSeq
  - 150/200 bp reads
  - $0.04 Mbp — Hardware ~ $1M
  - Used for WGS projects
- Illumina MiSeq
  - 250/300 bp reads
  - $0.05 Mbp — Hardware ~ $125K
  - Used for 16S projects — 384 samples/run
  - Desktop Sequencer

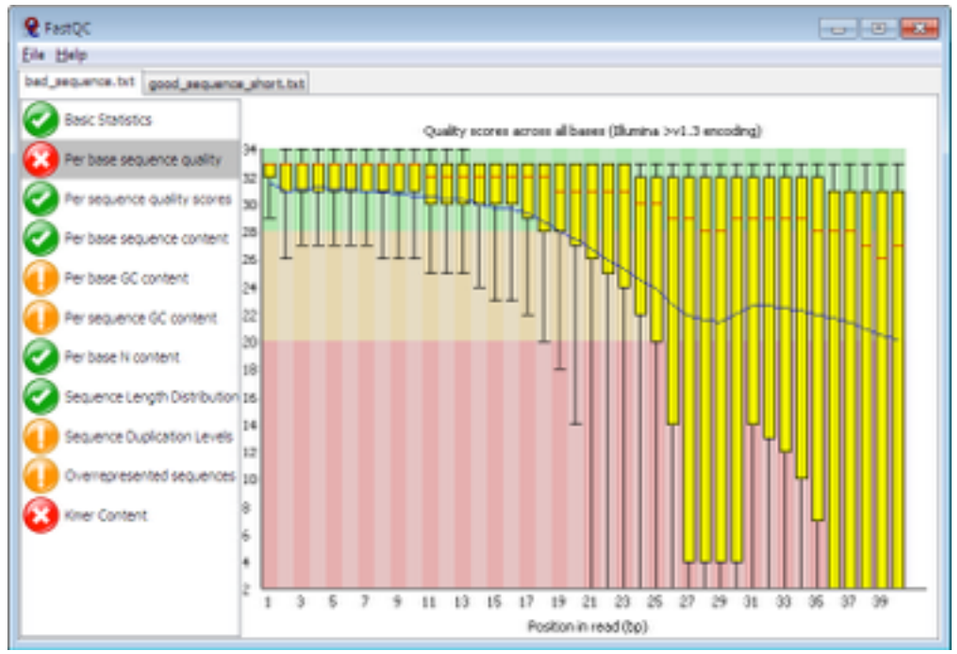
Special Sequencing Applications

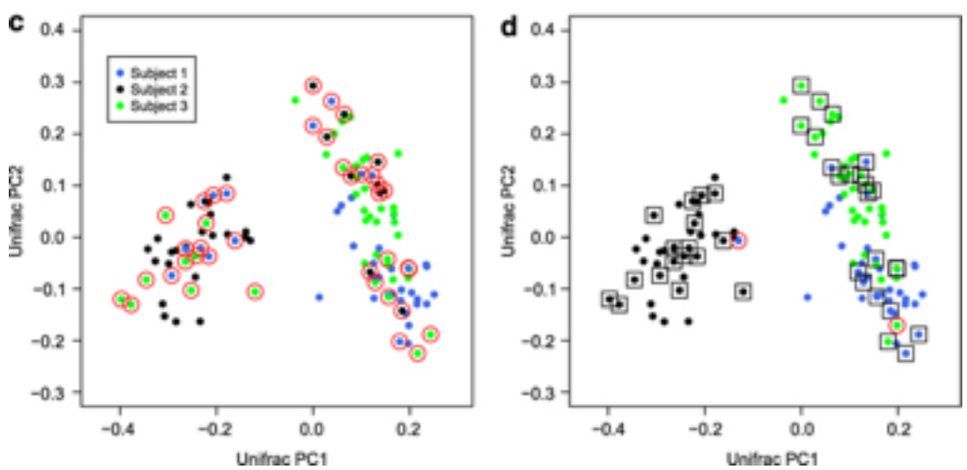
- Pacific Biosciences
  - Single Molecule Real Time (SMRT) Sequencing
  - Very High Error Rate
  - Average read length > 1kb
  - Great for Finishing Genomes by Illumina/PacBio Hybrid Assembly
  - ~$2/Mbp

- Oxford Nanopore MinIon
  - A “laptop powered” sequencing
  - Average Read Length 5.4kb
  - Light weight and low power usage makes it interesting for “in the field” applications

Quality Control

- Sequencing Errors
  - Low Quality Bases
  - Homopolymer Strings
  - Too short trimmed reads
- Contamination, Mislabeling and Sampling Error
  - Negative Controls are the best way to identify lab contamination
  - Biological and Technical Replicates



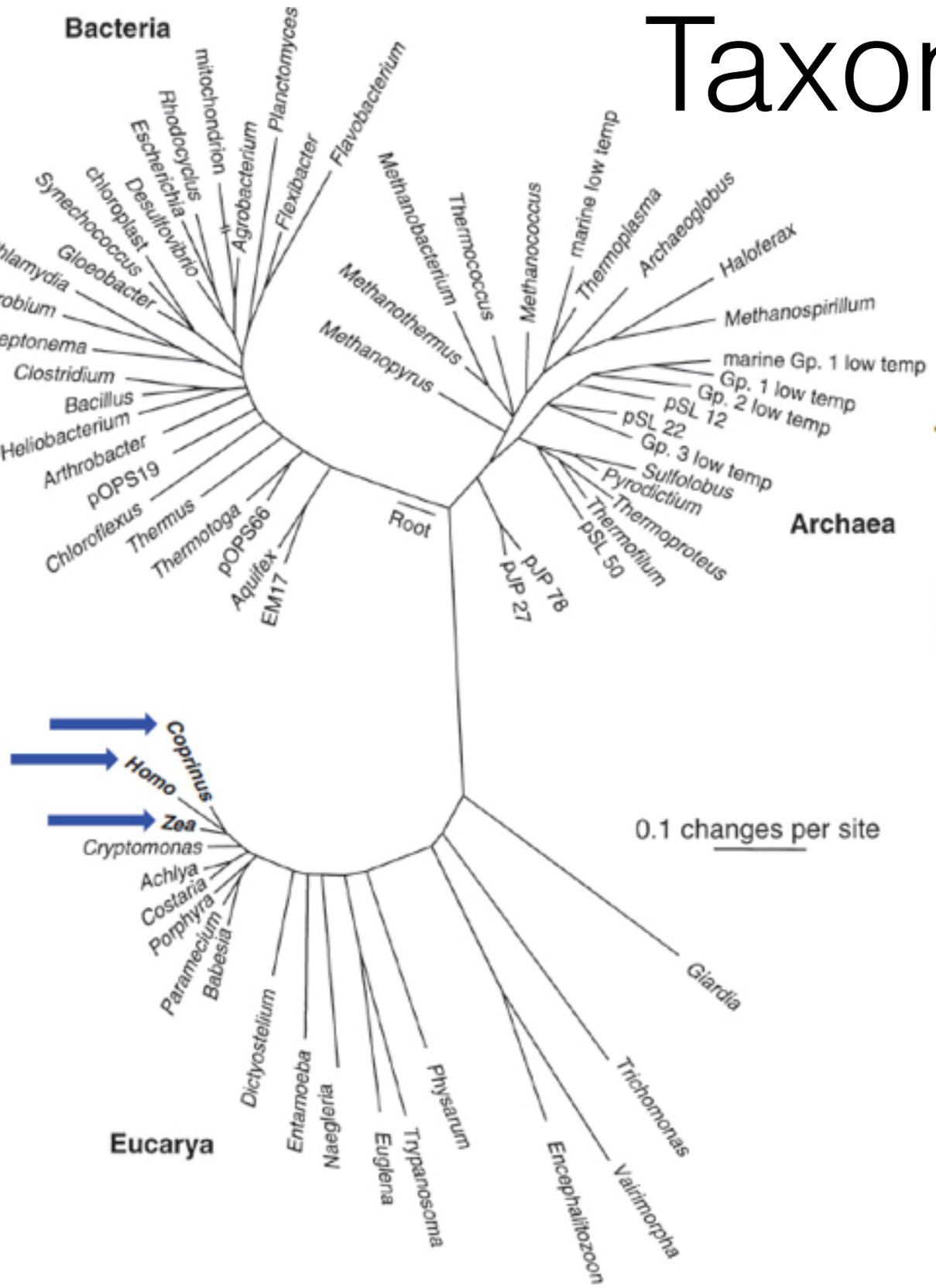


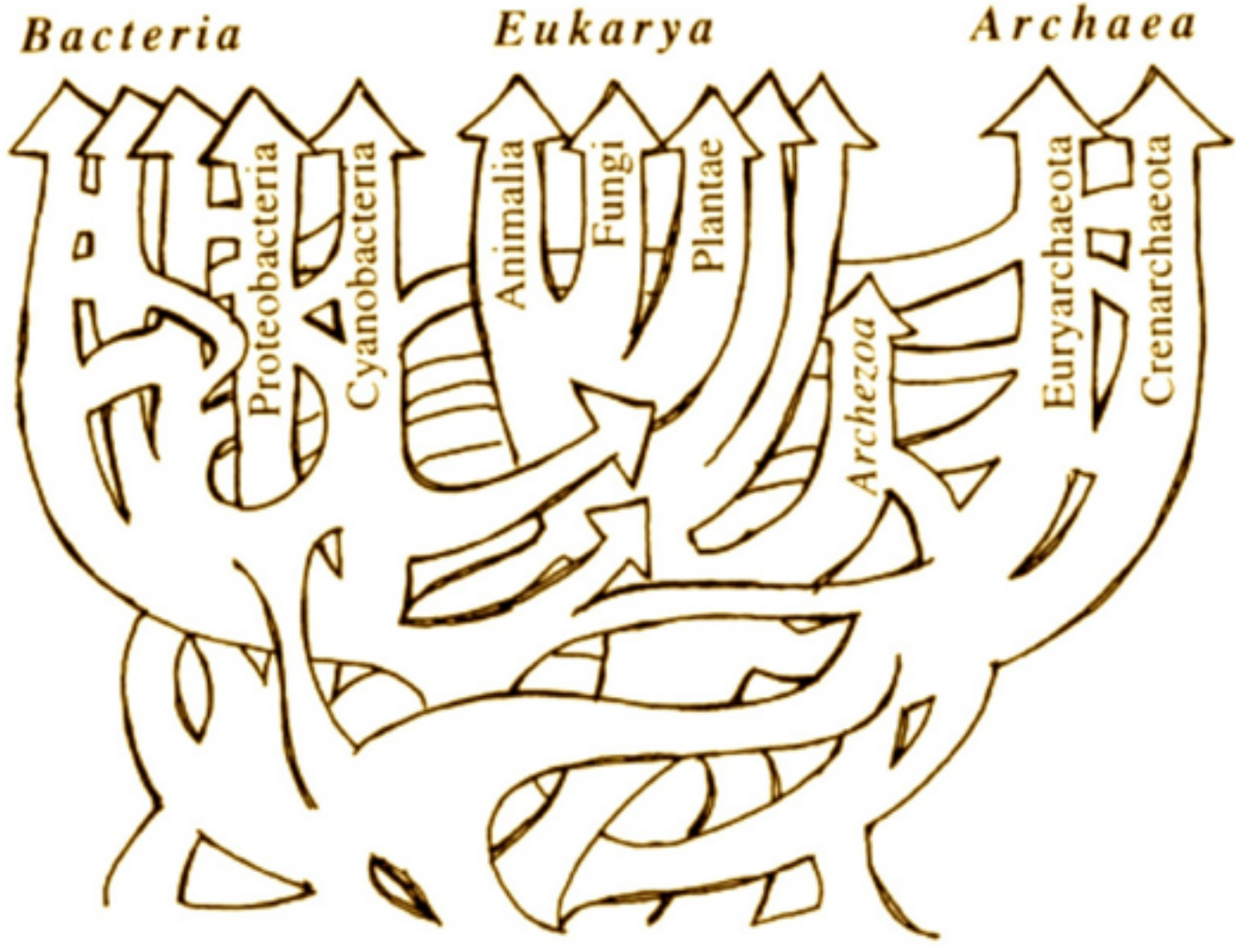
Knights D, Kuczynski J, Koren O, Ley RE, Field D, Knight R, DeSantis TZ, Kelley ST. Supervised classification of microbiota mitigates mislabeling errors. ISME J. 2011 Apr;5(4):570-3. doi: 10.1038/ismej.2010.148. Epub 2010 Oct 7. PubMed PMID: 20927137; PubMed Central PMCID: PMC3105748.

10 Minute Break

- *Introduction to Metagenomics*
  - What is a microbiome, metagenome,the relationship between the micro biome and it's environment?
  - Whole Community Sequencing Methods vs Traditional Culture Methods
  - Genetic Diversity in a Microbiome
  - Early Microbiome Projects
- *Sampling, DNA Extraction and Sequencing*
  - Comparison of Sequencing Technologies
  - Data quality: Error rates of sequencing, chimeras
  - Differences in profiles depending on sampling, DNA extraction and sequencing
- *Measuring Taxonomic Composition and Diversity using Marker Genes (16S)*
  - What is the 16S gene? Why is it so special?
  - 16S gene databases: RDP, SILVA, GreenGenes
  - Alternatives to 16S: ITS, CPN60, RecA
  - Data Processing workflow
  - Taxonomy assignment and OTUs
  - Analysis platforms: Qiime and Mother
  - Phylogenetic Trees, Rarefraction and Diversity and PCoA

Marker Genes Allow For Taxonomic Profiling





rRNAs as phylogenetic markers

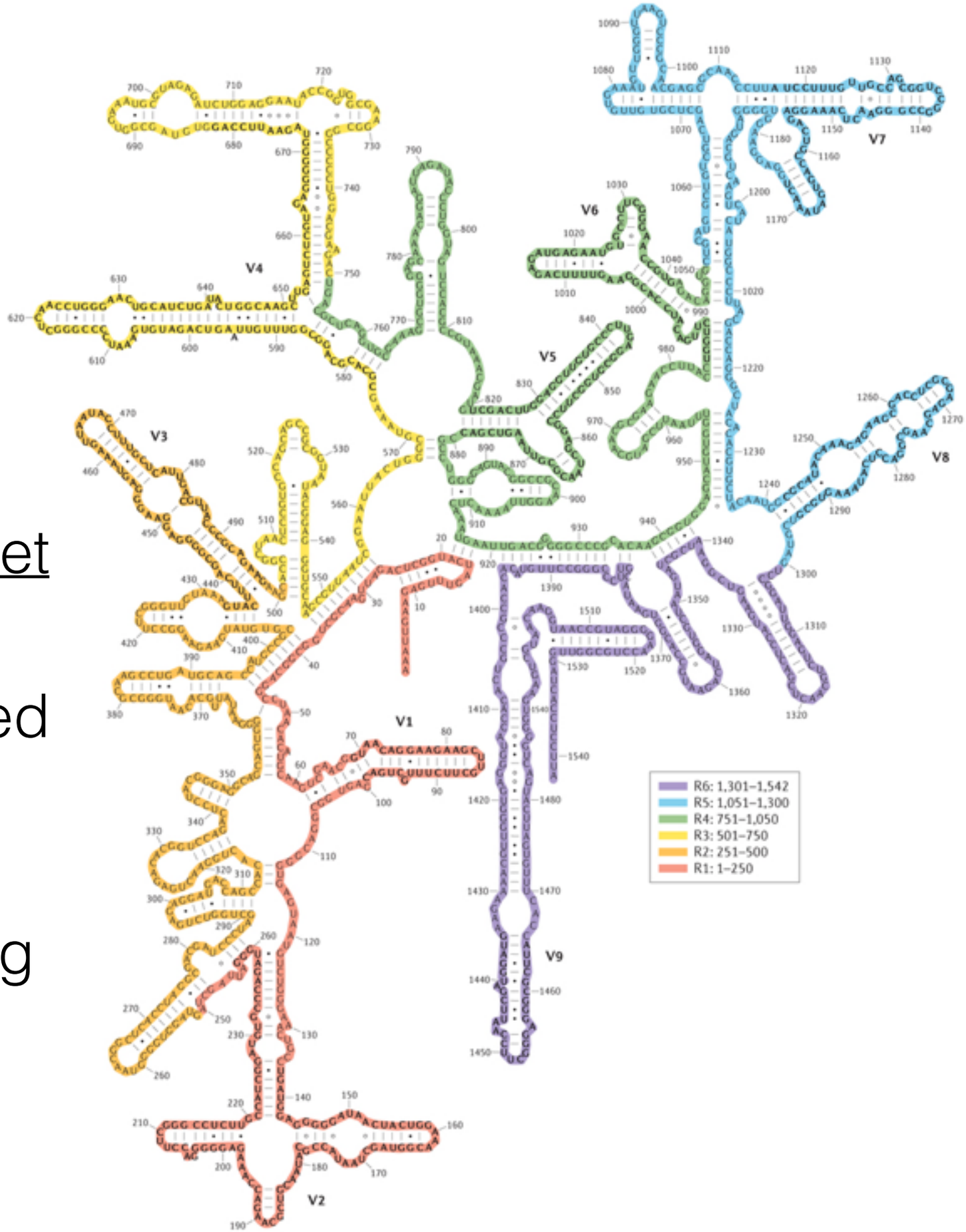
- Ribosomal RNAs are present in all living organisms
- rRNAs play a critical role in protein translation
- rRNAs are relatively conserved and rarely acquired horizontally
- Slowly evolving (molecular clock)

16S as a Marker Gene

- Present in all prokaryotic organisms compared
  - 18S is the equivalent in eukaryotes and is used for surveying fungal communities
- Vertically and slowly evolving
- Amplify-able with small set of “universal primers”
  - Better to have specialized primers for archaea
- Has an established database of reference

16S as a Marker Gene

- Amplify-able with small set of “universal primers”
- Better to have specialized primers for archaea
- Primers for 16S variable regions can give differing results



Nature Reviews | Microbiology

16S as a Marker Gene

- 16S: Different V regions generates different community profiles

- Strain-level diversity will be missed by amplicon

16S Databases

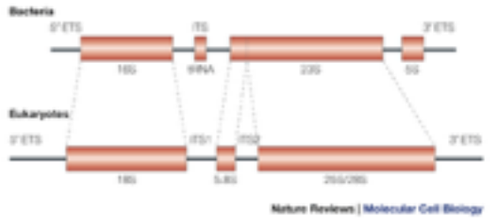
Cole, J. R., Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske, and J. M. Tiedje. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis Nucl. Acids Res. 42(Database issue):D633-D642; doi: 10.1093/nar/gkt1244 [PMID: 24288368]

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucl. Acids Res. 41 (D1): D590-D596.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. 2006. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol 72:5069-72.

Other Marker Genes

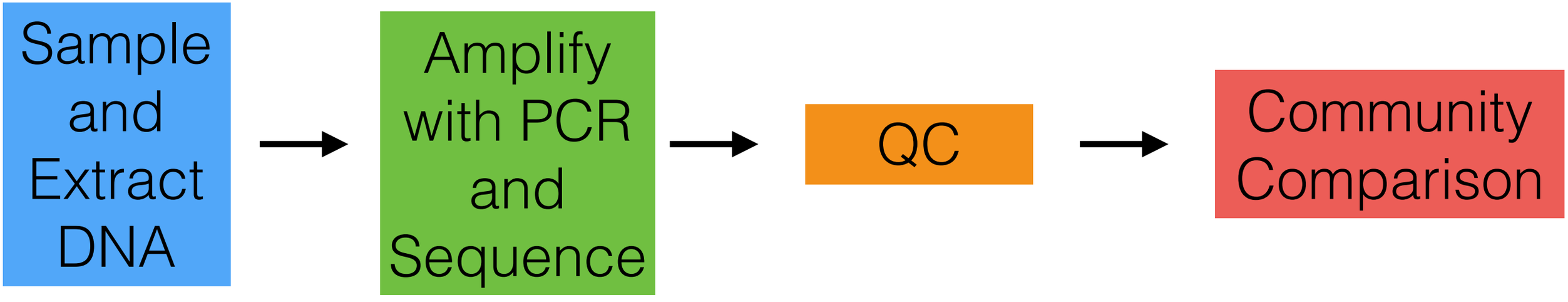
- Intergenic Transcribed Spacer (ITS)



- RecA: Response to DNA Stress in Bacteria

- Cpn60: Chaperonin Database

Workflow for Marker Gene Analysis



Sample and Extract DNA

Sampling

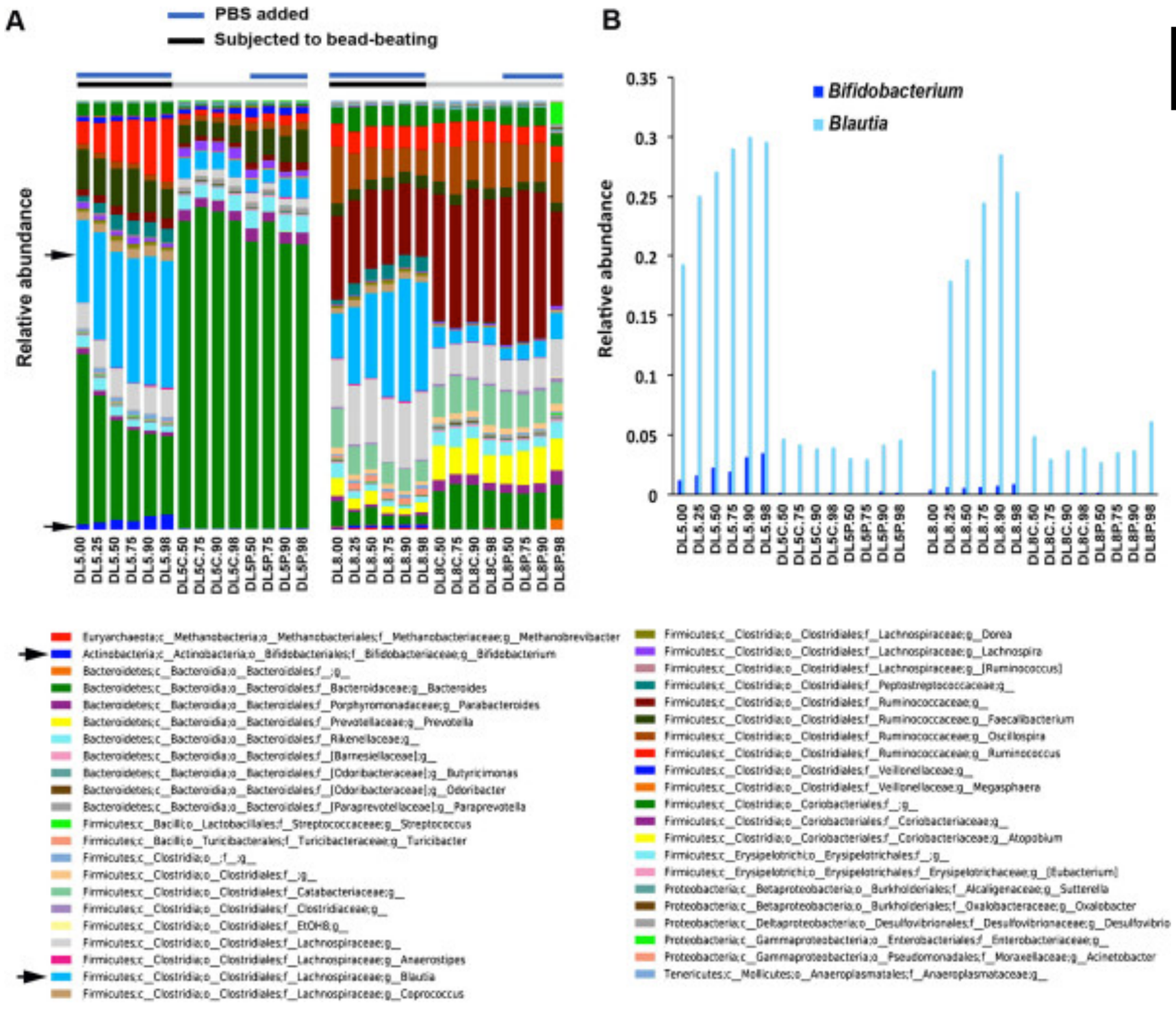
- Sampling Must be Standardized
- Samples should be
  - collected with sterile instrumentation or swabs
  - transported into a sterile tube without too much interaction with the environment
  - stabilized depending on molecule of interest
  - “frozen in time”
  - http://www.hmpdacc.org/doc/HMP_Protocol_Version_9_032210.pdf

Sample and Extract DNA

Sources of Contamination

- At Collection — use sampling protocol
  - Host DNA
  - Environmental
- In the lab
  - Use a negative control (water or stabilization buffer) sample to determine likely lab contaminiation
  - Your microbiome covers is a cloud around your body

DNA Extraction



Santiago A, Panda S, Mengels G, Martinez X, Azpiroz F, Dore J, Guarner F, Manichanh C. Processing faecal samples: a step forward for standards in microbial community analysis. BMC Microbiol. 2014 May 1;14:112. doi: 10.1186/1471-2180-14-112. PubMed PMID: 24884524; PubMed Central PMCID: PMC4021188.

Extraction Methods that Include a Bead Beating Step shows greater diversity

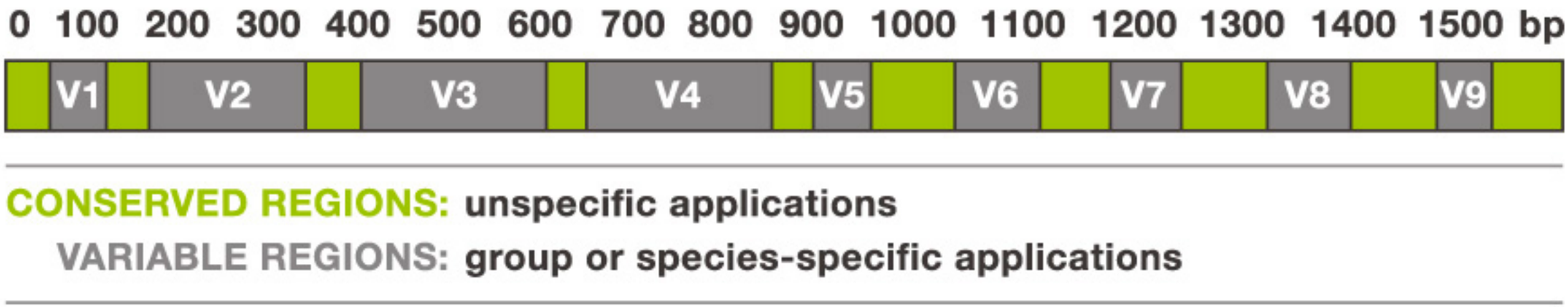
Sample and Extract DNA

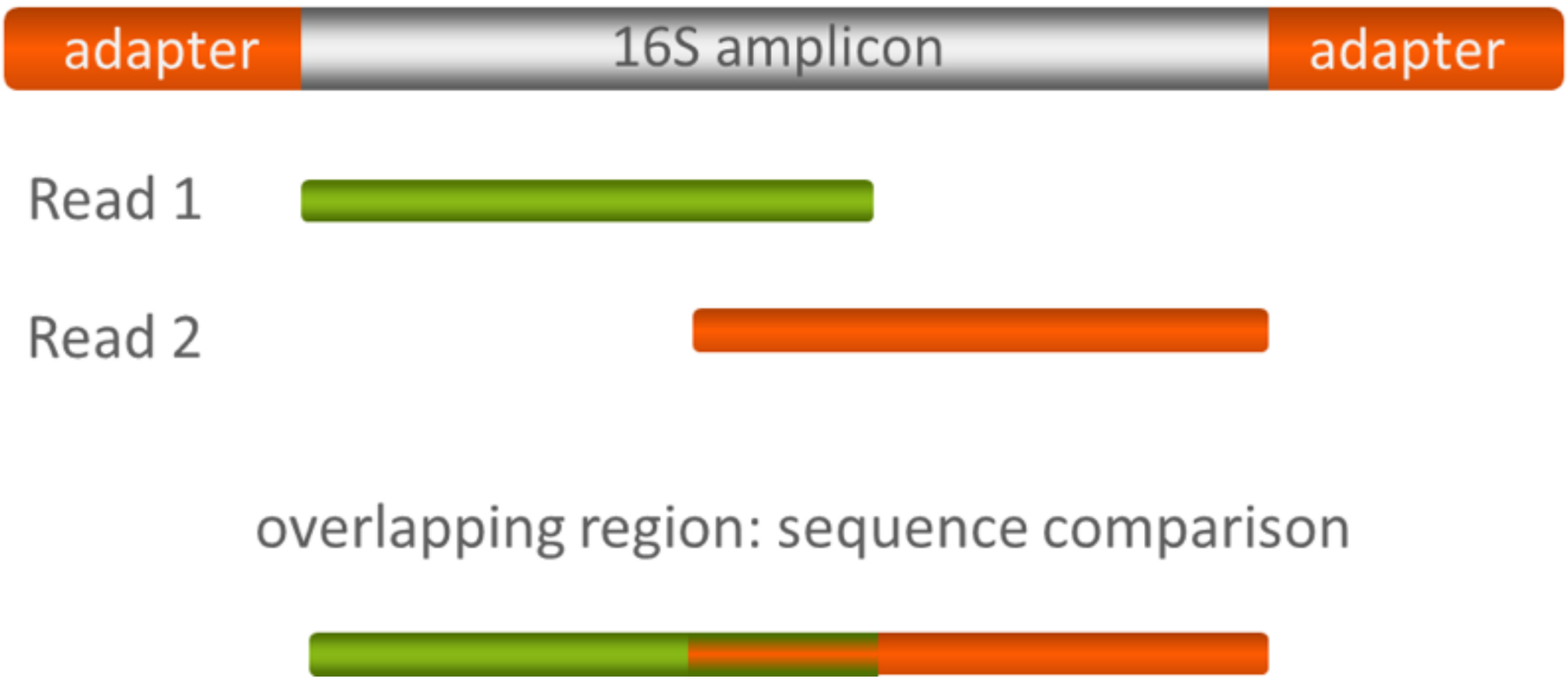
DNA Extraction

- Many protocols/kits available but kit-bias exists
  - Santiago A, Panda S, Mengels G, Martinez X, Azpiroz F, Dore J, Guarner F, Manichanh C. Processing faecal samples: a step forward for standards in microbial community analysis. BMC Microbiol. 2014 May 1;14:112. doi: 10.1186/1471-2180-14-112. PubMed PMID:24884524; PubMed Central PMCID: PMC4021188.
    - *Bead-beating is important for breaking down the cell-wall of some organisms*
- HMP and other major projects have standardized DNA extraction protocols
  - http://www.earthmicrobiome.org/emp-standard-protocols/dna-extraction-protocol/
  - http://www.hmpdacc.org/doc/HMP_MOP_Version12_0_072910.pdf

Amplify with PCR and Sequence

Target Amplification





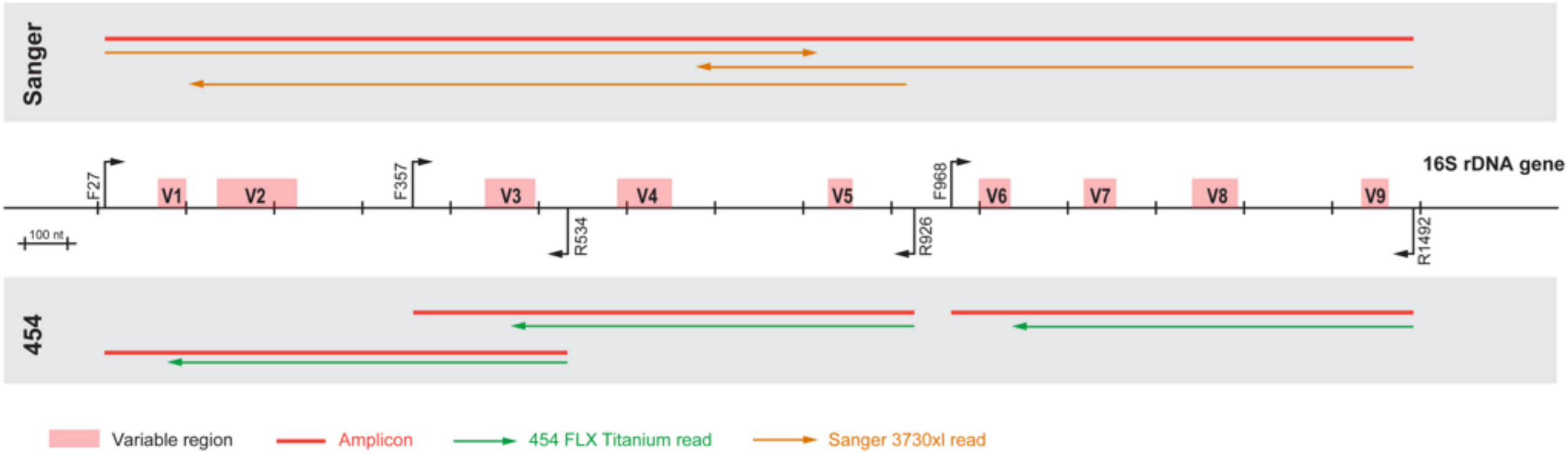
Amplify with PCR and Sequence

Target Amplification

- PCR primers are designed to amplify specific regions of the genome
  - Primer sets should be able to amplify that regions from a diverse set of taxa
  - For 16S we “prime” off of the conserved regions and use the variable regions to sample diversity
- “Dirty” samples could contain PCR inhibitors
- QC products using BioAnalyzer or gel

Amplify with PCR and Sequence

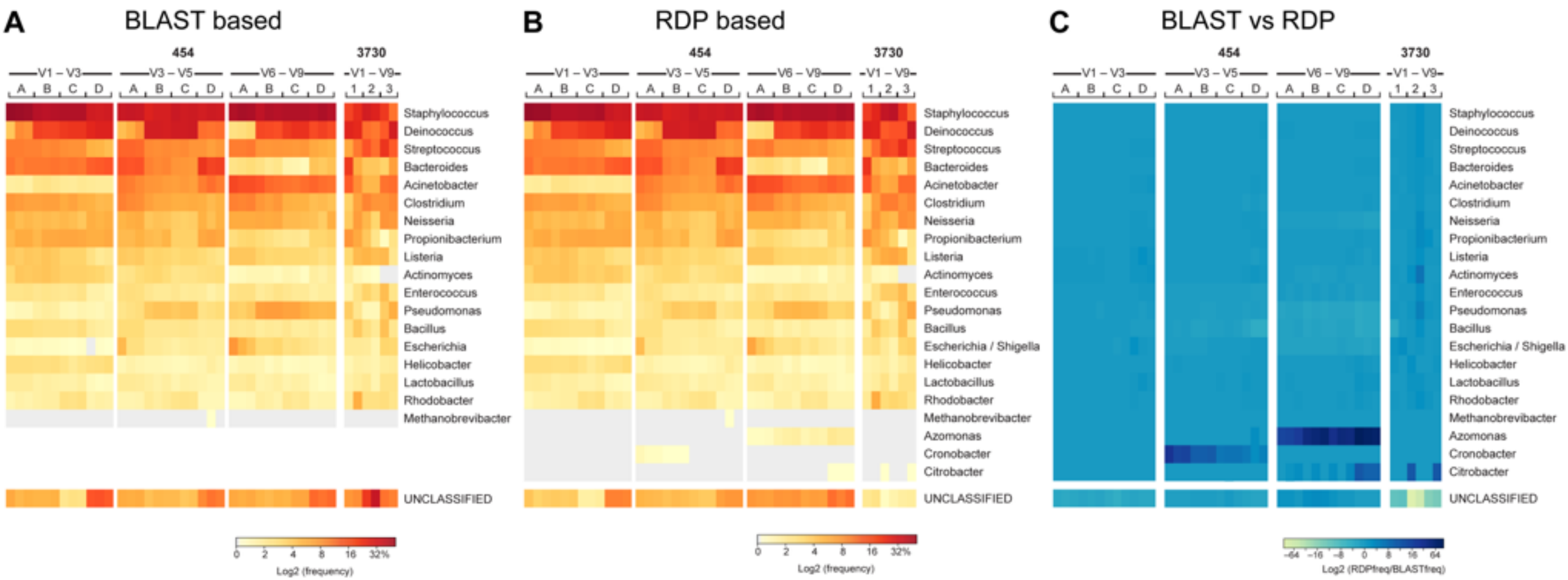
HMP Target Selection



Jumpstart Consortium Human Microbiome Project Data Generation Working Group. Evaluation of 16S rDNA-based community profiling for human microbiome research. PLoS One. 2012;7(6):e39315. doi: 10.1371/journal.pone.0039315. Epub 2012 Jun 13. PubMed PMID: 22720093; PubMed Central PMCID: PMC3374619.

Amplify with PCR and Sequence

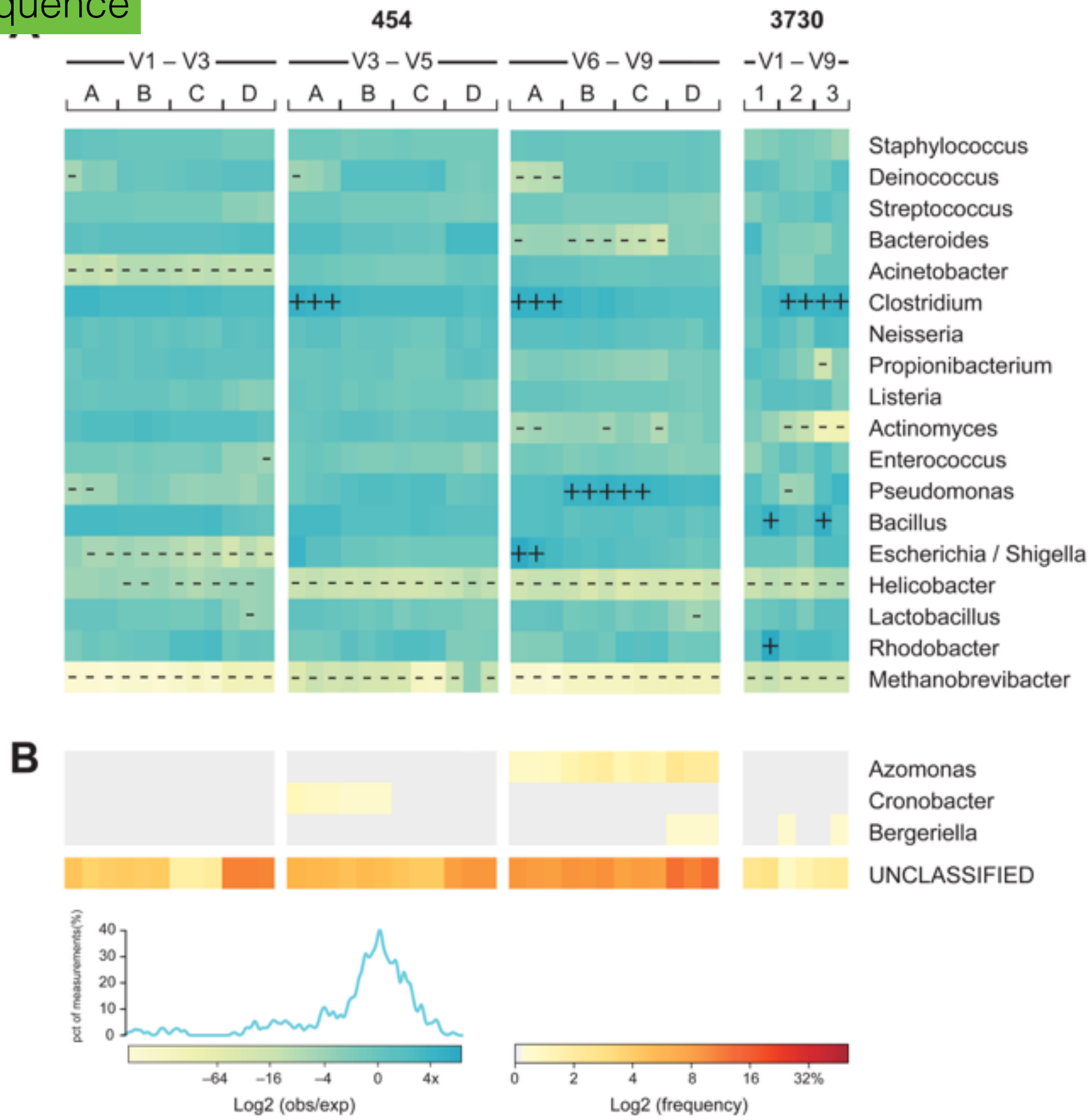
HMP Target Selection



Comparable results using BLAST and RDP but some relative abundance profile differences using different V regions compared to Sanger sequenced “whole” gene

Amplify with PCR and Sequence

HMP Target Selection



Using a Mock Community, results show differences in V regions for accurate profiling

DIY Marker Gene Analysis

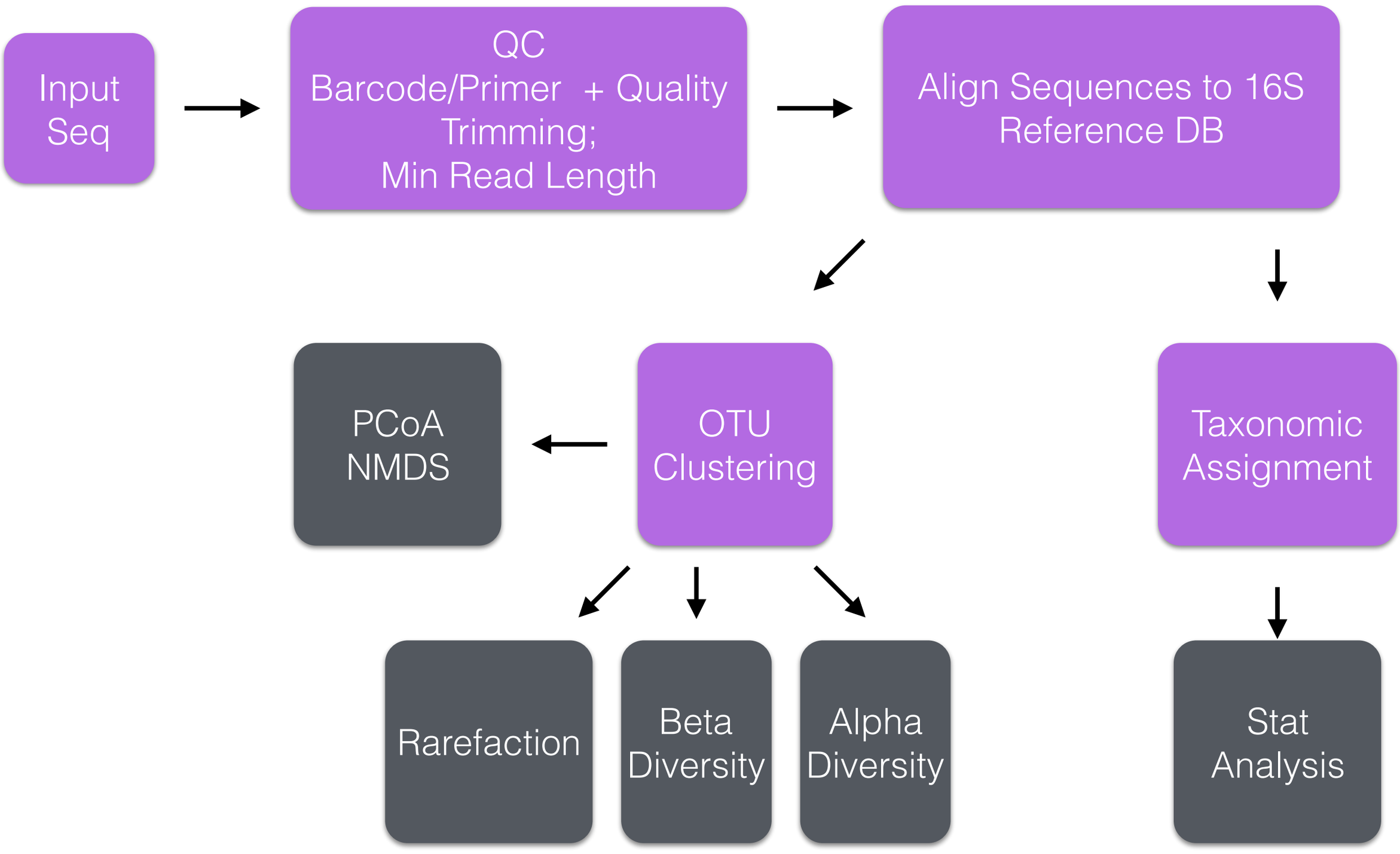
http://qiime.org/

http://www.mothur.org/

Mothur or Qiime?

- Mothur
  - Single program with all dependancies built in
  - Reimplementation/Optimization of popular algorithms
  - Easy to install, use
  - Less scalable and hard to add “your own tools”
- Qiime
  - A collection of python scripts, which are wrappers for popular algorithms
  - Large number of depedancies, VM available (for small datasets)
  - More scaleable and easy to add “your own tools”

Overall Analysis Pipeline



QC

Sequence Processing

- Primer and Barcode Trimming
- Very large (pyrosequencing) and very small sequence trimming
- Low Quality Trimming
- Remove Reads with high likelihood of errors
  - High Number of Homopolymers (pyrosequencing)
  - Ambiguous bases

QC

Sequence Analysis

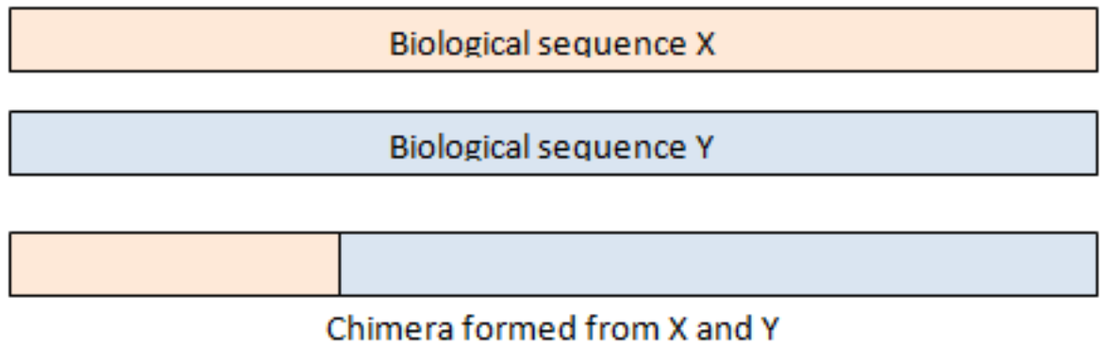
- Removal of sequences not aligned to the targeted region of the 16S Gene
- Removal of chimera
- Removal of contamination (negative control)

Community Comparison

Alignment

- Choose a reference for the alignment
  - SILVA *
  - Greengenes
  - RDP
- Choose a search algorithm (Identify the best hit)
  - blast
  - kmer *
  - suffix
- Choose an alignment algorithm (Align query to best hit)
  - Needleman (Global Aligner) *
  - Blastn
  - Gotoh
- Penalties for gaps and mismatches (Score the alignment)

Chimeras

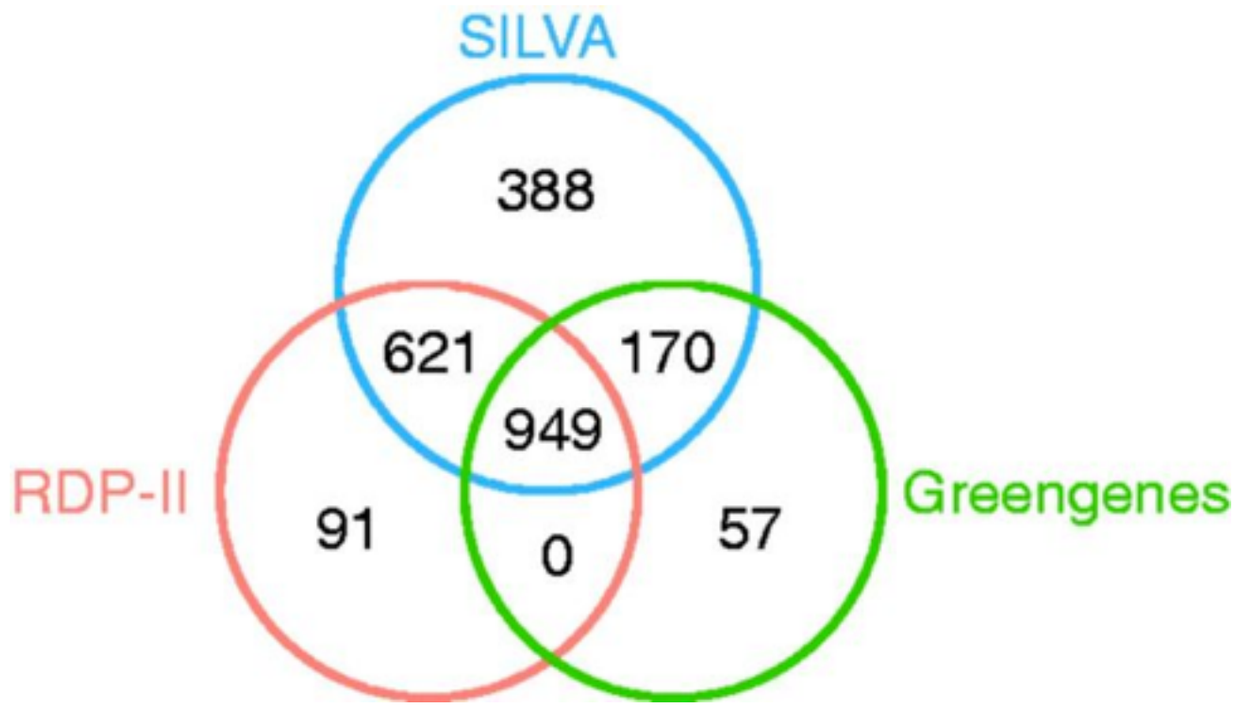


- The chimera was a beast in greek mythology made up of parts from four different creatures
- DNA sequence chimera are artifacts created by PCR
- Chimeras can be detected by comparison to a reference database
  - ChimeraSlayer
  - Uchime

Community Comparison

Taxonomic Classification

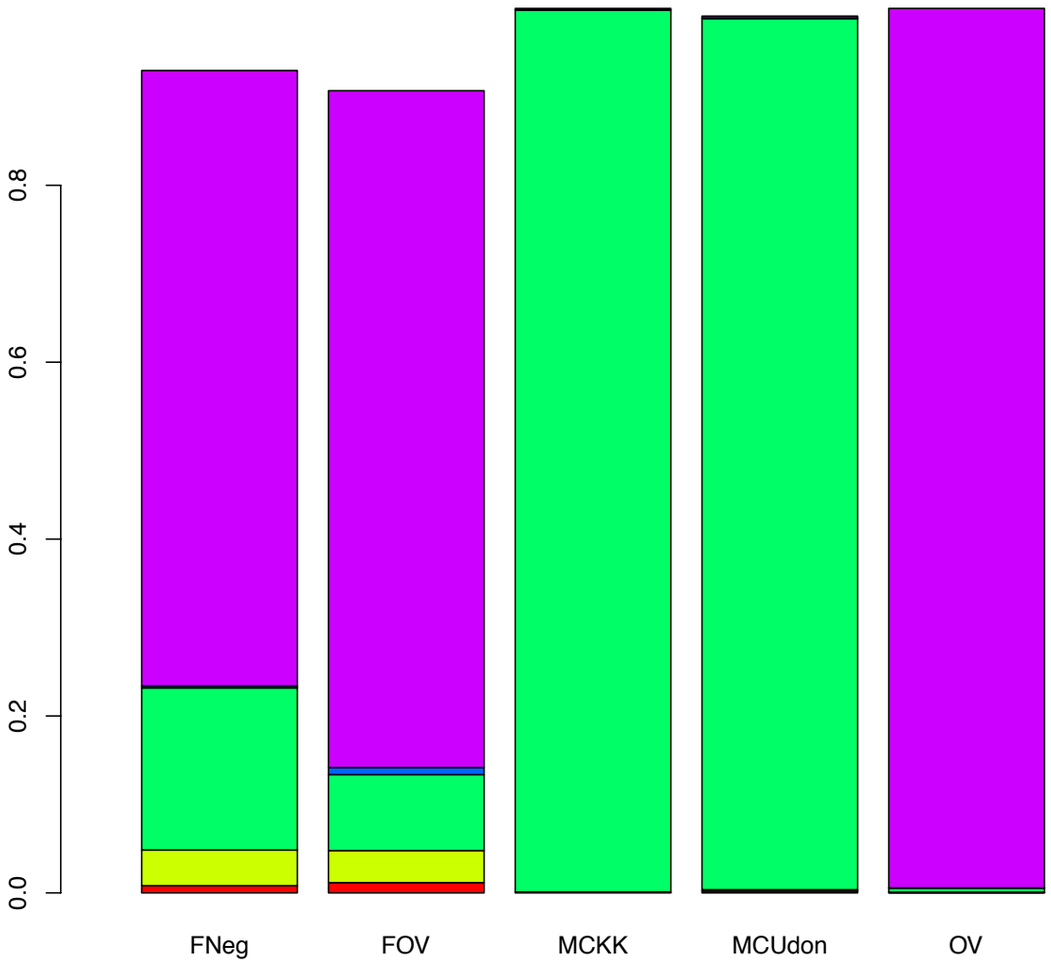
- Choose a method
  - Wang — Implemented at the RDP classifier
  - K nearest neighbor
- Choose a hierarchy
  - RDP-II
  - Greengenes
  - Silva

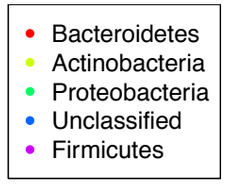


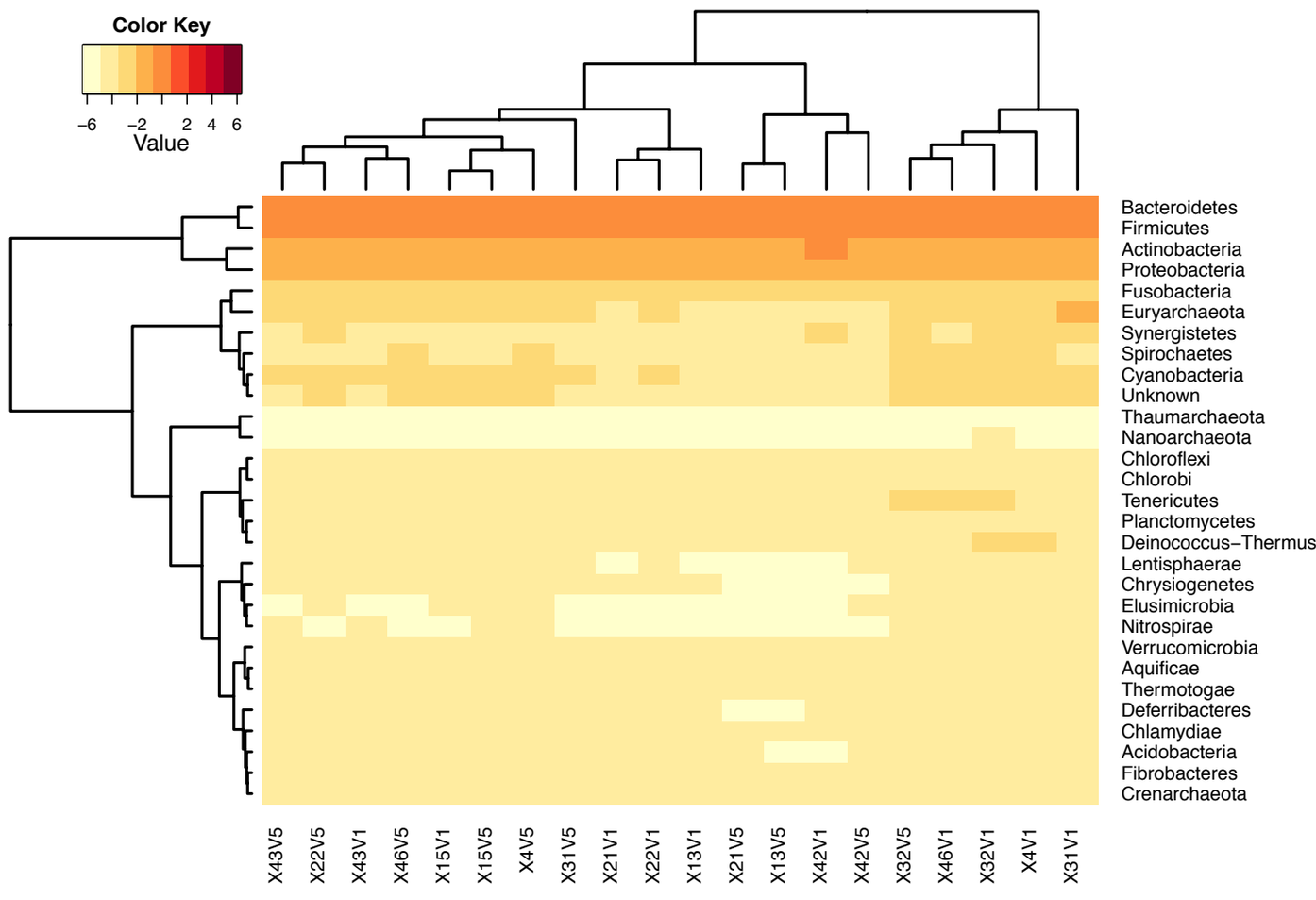
Nucleic Acids Res. 2014 Jan 1; 42(Database issue): D643–D648.

Community Comparison

Taxonomic Comparisons







Community Comparison

OTU Classification

- Calculate a distance between sequences
- Cluster Sequences by similarity
  - Nearest Neighbor
    - Each of the sequences within an OTU are at most X% distant from the most similar sequence in the OTU.
  - Average Neighbor
    - All of the sequences within an OTU are at average X% distant from all of the other sequences within the OTU.
  - Furthest Neighbor
    - All of the sequences within an OTU are at most X% distant from all of the other sequences within the OTU.

Community Comparison

Alpha Diversity

- Species richness is a survey of the number of distinct organism in a community
- Rarefaction is a method to assess species richness
- Species evenness measures how equal the community ie 2 taxa each at 50% abundance vs 9 to 1 ratio.
- Alpha diversity is a measurement composed of richness and evenness.

Community Comparison

Beta-Diversity

- Beta-diversity measures including absolute or relative overlap describe how many taxa are shared between habitats
- Beta diversity acts like a similarity score between populations, allowing analysis by sample clustering or, again, by dimensionality reductions such as PCA
- Beta diversity can be measured by simple taxa overlap such as Bray-Curtis dissimilarity

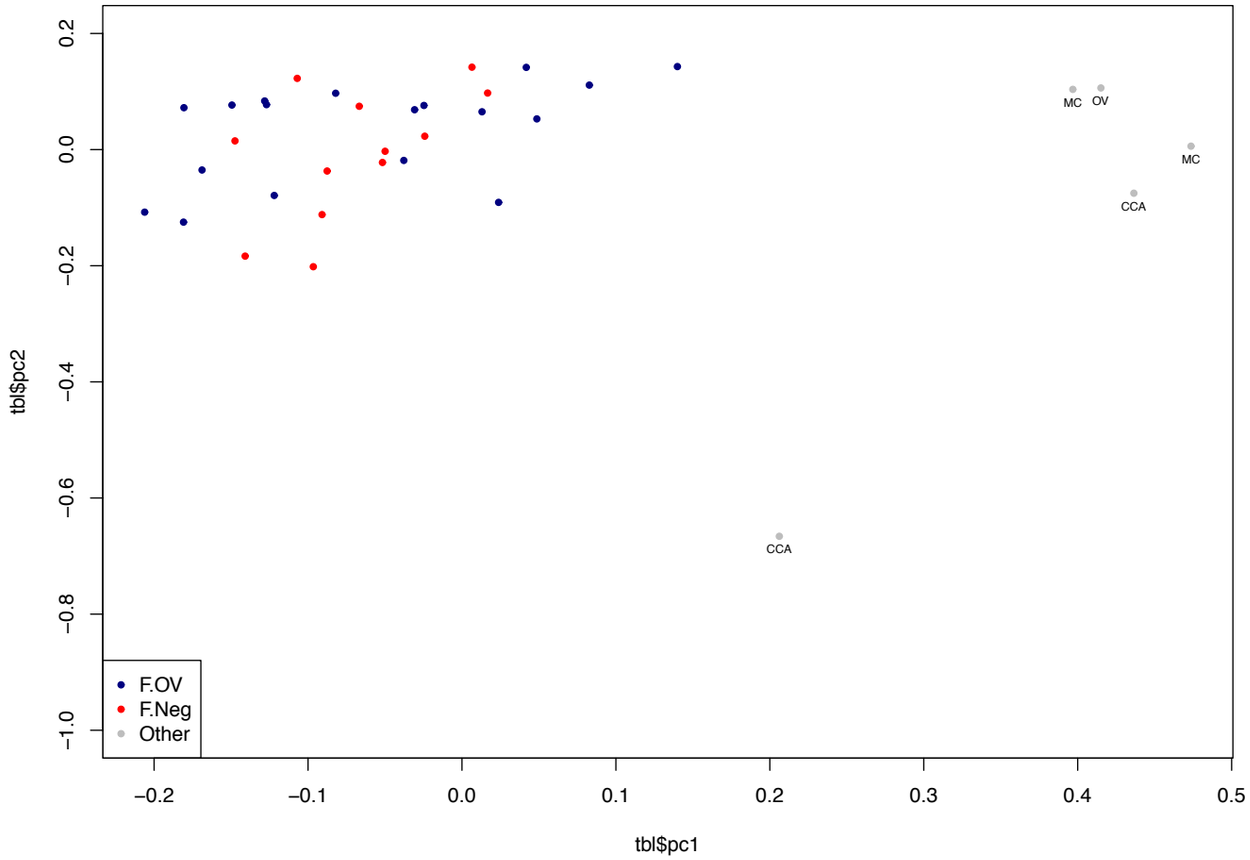
Community Comparison

Unifrac

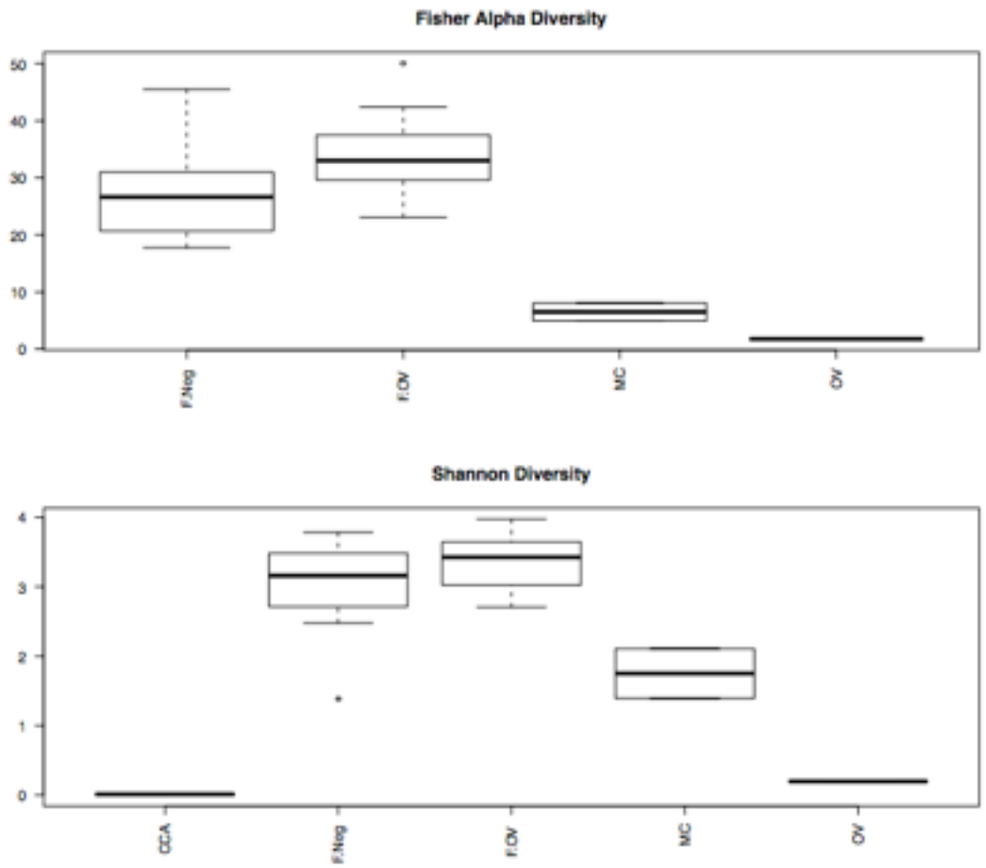
- A distance metric used for comparing biological communities
- It differs from distance metrics (Bray Curtis) as it incorporates phylogenetic distances (tree based) between observed organisms in the computation
- Weighted Unifrac also incorporates taxonomic abundances

Community Comparison

OTU Comparison



PCoA



Diversity

Community Comparison

Statistics

- Metastats is a software used for detecting differentially abundant features (ie phyla or genera)
  - Employs the false discovery rate to improve specificity in high-complexity environments, and separately handles sparsely-sampled features using Fisher's exact test
- LEfSe is an algorithm for high-dimensional biomarker discovery and explanation that identifies genomic features (genes, pathways, or taxa) characterizing the differences between two or more biological conditions (or classes)

16S rRNA Workshop