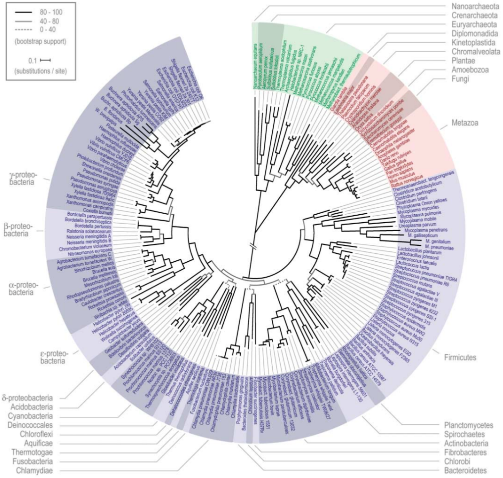
# Metagenomics

Mikhail Dozmorov

2021-04-26

# Tree of life



Ciccarelli, Francesca D., Tobias Doerks, Christian von Mering, Christopher J. Creevey, Berend Snel, and Peer Bork. "Toward Automatic Reconstruction of a Highly Resolved Tree of Life." Science (New York, N.Y.) 311, no. 5765 (March 3, 2006): 1283–87. doi:10.1126/science.1123061.

# Phylogenetic tree

- Three main branches of life (**Bacteria**, **Archaea**, and **Eucarya**).
    - Woese, C. R., O. Kandler, and M. L. Wheelis. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci.
- Higher-order taxonomy to reconcile bacterial taxonomy with rRNA-based phylogeny.
    - Garrity, G. M., J. A. Bell, and D. B. Searles. 2001. Taxonomic outline of the prokaryotes. Bergey's manual of systematic bacteriology, 2nd ed., release 1.0. Springer-Verlag, New York, NY

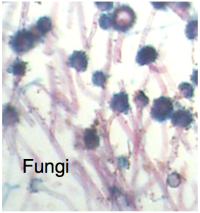
# Microbiome - our second genome

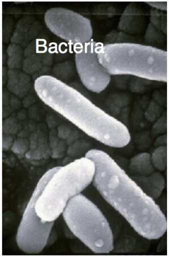
- **Microbiota** - microorganisms—such as bacteria, fungi, viruses, and archaea—present in a community
- **Microbiome** - all of the genetic material of a microbial community sequenced together
- Microbes constitute 90% of the total number of cells associated with our bodies; only the remaining 10% are human cells

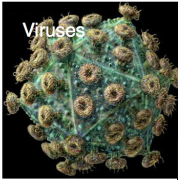
Savage DC. 1977. Microbial ecology of the gastrointestinal tract. Annu. Rev. Microbiol. 31:107–33

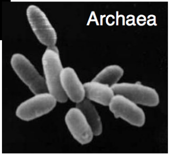
# Why the Human Microbiome?

- Each human cell has the same protein-encoding potential. Microbes are more diverse and dynamic than human genome.
- Human - ~25,000 genes. Human gut microbiome - ~2-3 million genes, typically >160 “species” at any given sample time









Qin, Junjie, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, et al. “A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing.” Nature 464, no. 7285 (March 4, 2010): 59–65. https://doi.org/10.1038/nature08821.

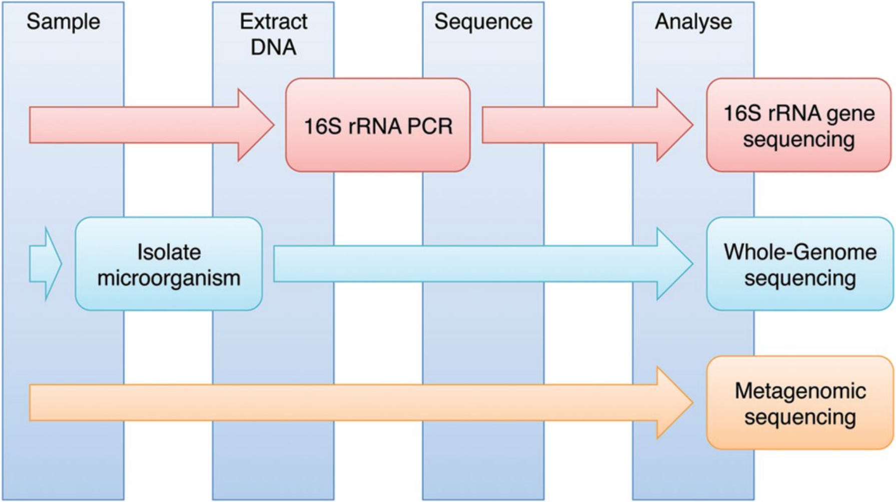
# Human Microbiome

Not all microbes are bad: Beneficial microbes perform functions essential for human health

- Vitamin synthesis
- Digestion
- Education and activation of immune system
- Inhibition of skin colonization by pathogens

**Many microbial-host and microbial-microbial interactions remain unknown**

# How can we analyze the microbiome?



Cox MJ, Cookson WO, Moffatt MF. Sequencing the human microbiome in health and disease. Hum Mol Genet. 2013 Oct 15;22(R1):R88-94. doi: 10.1093/hmg/ddt398. Epub 2013 Aug 13. Review.

# Elucidating the diversity of the human microbiome

- Traditional approaches rely on isolating bacteria in pure culture
- The majority of bacterial species do not grow in culture = “the great plate count anomaly”
- Culturing favors microbial “weeds” - not necessarily the most dominant or influential species
- Excludes microbes that rely on community interactions
- Direct sampling to sequencing is preferrable

# Types of microbiome analysis

1. Environmental clone libraries (functional metagenomics): use of Sanger sequencing (frequently) instead of more cost-efficient next-generation sequencing
2. Amplicon metagenomics (single gene studies, 16s rDNA): next-generation sequencing of PCR amplified ribosomal genes providing a single reference gene–based view of microbial community ecology
3. Shotgun metagenomics: use of next-generation technology applied directly to environmental samples
4. Metatranscriptomics: use of cDNA transcribed from mRNA

# Microbiome analysis questions

- **Who is out there?** Identifying the composition of a microbial community either by using amplicon data for single genes or by deriving community composition from shotgun metagenomic data using sequence similarities.
- **What are they doing?** Using shotgun data (or metatranscriptomic data) to derive the functional complement of a microbial community using similarity searches against a number of databases.
- **Who is doing what?** Based on sequence similarity searches, identifying the organisms encoding specific functions.

Section 1

# Marker gene analysis of microbiome

# Marker gene analysis, 16S ribosomal RNA

- 16S ribosomal RNA (or 16S rRNA) is the component of the 30S small subunit of a prokaryotic ribosome
- The 16S rRNA sequence contains both highly conserved and variable regions.
- Conserved regions allows using universal PCR primers to amplify 16S sequences.
- Variable regions, nine in number (V1 through V9), behave like a molecular clock and are used to classify organisms according to phylogeny

# Marker gene analysis, 16S ribosomal RNA

16S rRNA sequencing has been used to characterize the complexity of microbial communities

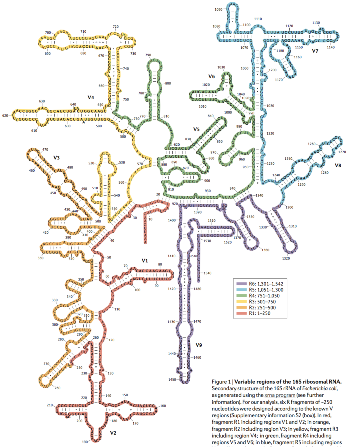
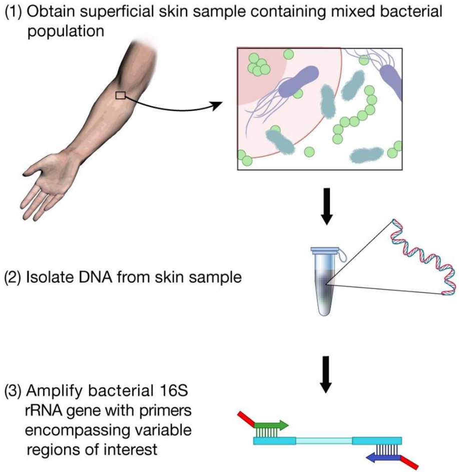


Figure 1 | **Variable regions of the 16S ribosomal RNA.** Secondary structure of the 16S rRNA of *Escherichia coli*, as generated using the xrna program (see Further information). For our analysis, six R fragments of ~250 nucleotides were designed according to the known V regions (Supplementary information S2 (box)). In red, fragment R1 including regions V1 and V2; in orange, fragment R2 including region V3; in yellow, fragment R3 including region V4; in green, fragment R4 including regions V5 and V6; in blue, fragment R5 including regions V7 and V8; and in purple, fragment R6 including region V9.

https://www.nature.com/nrmicro/journal/v12/n9/full/nrmicro3330.html

# Bacterial 16S rRNA sequencing workflow



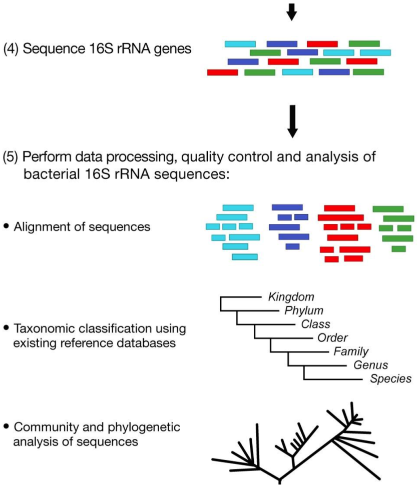
(1) Obtain superficial skin sample containing mixed bacterial population

(2) Isolate DNA from skin sample

(3) Amplify bacterial 16S rRNA gene with primers encompassing variable regions of interest

Kong, Heidi H. "Skin Microbiome: Genomics-Based Insights into the Diversity and Role of Skin Microbes." Trends in Molecular Medicine 17, no. 6 (June 2011): 320–28. https://doi.org/10.1016/j.molmed.2011.01.013.

# Bacterial 16S rRNA sequencing workflow



(4) Sequence 16S rRNA genes

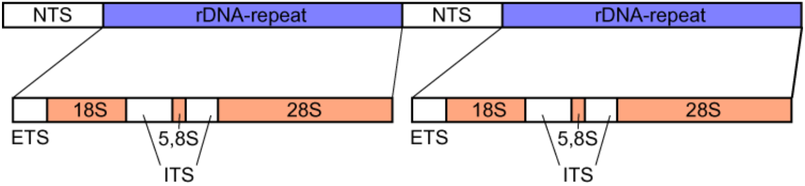
(5) Perform data processing, quality control and analysis of bacterial 16S rRNA sequences:

- Alignment of sequences
- Taxonomic classification using existing reference databases
- Community and phylogenetic analysis of sequences

Kong, Heidi H. "Skin Microbiome: Genomics-Based Insights into the Diversity and Role of Skin Microbes." Trends in Molecular Medicine 17, no. 6 (June 2011): 320–28. https://doi.org/10.1016/j.molmed.2011.01.013.

# Internal Transcribed Spacer marker

- Genes encoding ribosomal RNA and spacers occur in tandem repeats that are thousands of copies long, each separated by regions of non-transcribed DNA termed intergenic spacer (IGS) or non-transcribed spacer (NTS).
- The spacer DNA situated between the small-subunit ribosomal RNA (rRNA) and large-subunit rRNA genes



https://en.wikipedia.org/wiki/Internal_transcribed_spacer

# Other Marker Genes Used

- **Eukaryotic Organisms (protists, fungi)**
    - 18S (http://www.arb-silva.de)
    - ITS (Internal Transcribed Spacer, http://www.mothur.org/wiki/UNITE_ITS_database)
- **Bacteria**
    - CPN60 (Chaperonin 60, http://www.cpndb.ca/cpnDB/home.php)
    - ITS (Martiny, Env Micro 2009)
    - RecA gene (https://en.wikipedia.org/wiki/RecA)
- **Viruses**
    - Gp23 capsid protein for T4-like bacteriophage
    - RdRp (RNA-dependent RNA polymerase) for picornaviruses
    - Faster evolving markers used for strain-level differentiation

Section 2

# 16S clustering

# OTU

**OTUs (Operational taxonomic units)** - groups of sequences that are meaningfully separated from other sequences by hierarchical clustering techniques (independent of phylogenetic inferences) and using strict sequence identity thresholds.

- 16S rRNA gene sequences are routinely assigned to operational taxonomic units (OTUs) that are then used to analyze complex microbial communities.
- The first approach has been referred to as phylotyping (Schloss & Westcott, 2011) or closed-reference clustering (Navas-Molina et al., 2013) - how close OTUs are to the reference sequence
- Reference-based clustering methods suffer when the reference does not adequately reflect the biodiversity of the community. If a large fraction of sequences are novel, then they cannot be assigned to an OTU.

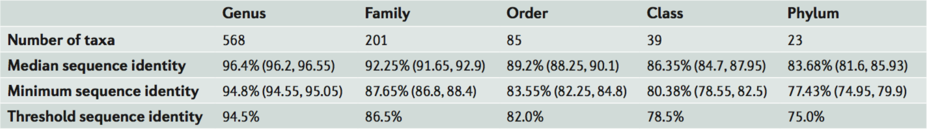
# OTU

- The second approach has been referred to as distance-based (Schloss & Westcott, 2011) or *de novo* clustering (Navas-Molina et al., 2013).
- In this approach, the distance between sequences is used to cluster sequences into OTUs rather than the distance to a reference database.
- The computational cost of hierarchical *de novo* clustering methods scales quadratically with the number of unique sequences.

# OTU

- The third approach, open-reference clustering, is a hybrid of the closed-reference and de novo approaches (Navas-Molina et al., 2013; Rideout et al., 2014).
- Open-reference clustering involves performing closed-reference clustering followed by de novo clustering on those sequences that are not sufficiently similar to the reference.

# Taxonomic thresholds of bacteria and archaea



| | Genus | Family | Order | Class | Phylum |
|---|---|---|---|---|---|
| **Number of taxa** | 568 | 201 | 85 | 39 | 23 |
| **Median sequence identity** | 96.4% (96.2, 96.55) | 92.25% (91.65, 92.9) | 89.2% (88.25, 90.1) | 86.35% (84.7, 87.95) | 83.68% (81.6, 85.93) |
| **Minimum sequence identity** | 94.8% (94.55, 95.05) | 87.65% (86.8, 88.4) | 83.55% (82.25, 84.8) | 80.38% (78.55, 82.5) | 77.43% (74.95, 79.9) |
| **Threshold sequence identity** | 94.5% | 86.5% | 82.0% | 78.5% | 75.0% |

Yarza, Pablo, Pelin Yilmaz, Elmar Pruesse, Frank Oliver Glöckner, Wolfgang Ludwig, Karl-Heinz Schleifer, William B. Whitman, Jean Euzéby, Rudolf Amann, and Ramon Rosselló-Móra. "Uniting the Classification of Cultured and Uncultured Bacteria and Archaea Using 16S RRNA Gene Sequences." Nature Reviews Microbiology 12, no. 9 (August 14, 2014): 635–45. https://doi.org/10.1038/nrmicro3330.

Section 3

# Marker databases

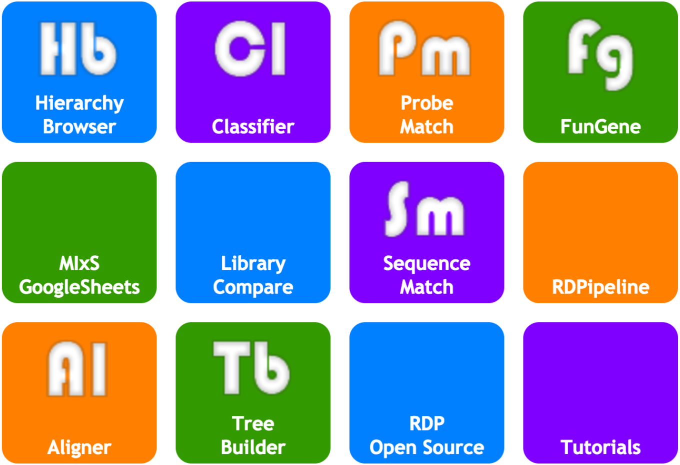
# RDP Database

- RDP provides quality-controlled, aligned and annotated Bacterial and Archaeal 16S rRNA sequences, and Fungal 28S rRNA sequences, and a suite of analysis tools to the scientific community
- RDP Release 11, Update 5, September 30, 2016
- 3,356,809 16S rRNAs, 125,525 Fungal 28S rRNAs

http://rdp.cme.msu.edu/

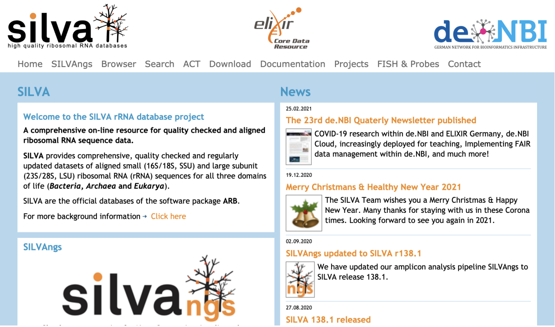
Cole, James R., Qiong Wang, Jordan A. Fish, Benli Chai, Donna M. McGarrell, Yanni Sun, C. Titus Brown, Andrea Porras-Alfaro, Cheryl R. Kuske, and James M. Tiedje. "Ribosomal Database Project: Data and Tools for High Throughput RRNA Analysis." Nucleic Acids Research 42, no. Database issue (January 2014): D633-642. doi:10.1093/nar/gkt1244.

# RDP Tools


Hb
Hierarchy Browser
Cl
Classifier
Pm
Probe Match
Fg
FunGene
MIxS GoogleSheets
Library Compare
Sm
Sequence Match
RDPipeline
Al
Aligner
Tb
Tree Builder
RDP Open Source
Tutorials

# Silva Database (ARB): http://www.arb-silva.de/

- Ribosomal RNA database for all three domains of live, the Bacteria, Archaea (16S/23S), and Eukarya (18S/28S).
- Small subunit (SSU) and two large submunits (LSU) sequences.
- Build a Phylogenetic Tree and calculate branch length.
- Browser, alignment tools, download.



SILVA

Welcome to the SILVA rRNA database project

**A comprehensive on-line resource for quality checked and aligned ribosomal RNA sequence data.**

**SILVA** provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (***Bacteria***, ***Archaea*** and ***Eukarya***).

SILVA are the official databases of the software package **ARB**.

For more background information → Click here

SILVAngs

News

25.02.2021

The 23rd de.NBI Quaterly Newsletter published

COVID-19 research within de.NBI and ELIXIR Germany, de.NBI Cloud, increasingly deployed for teaching, Implementing FAIR data management within de.NBI, and much more!

19.12.2020

Merry Christmas & Healthy New Year 2021

The SILVA Team wishes you a Merry Christmas & Happy New Year. Many thanks for staying with us in these Corona times. Looking forward to see you again in 2021.

02.09.2020

SILVAngs updated to SILVA r138.1

We have updated our amplicon analysis pipeline SILVAngs to SILVA release 138.1.

27.08.2020

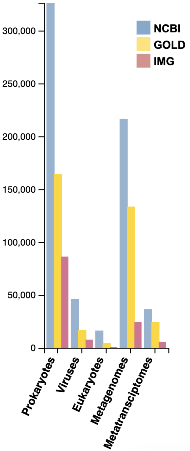
SILVA 138.1 released

Pruesse, E., C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, and F. O. Glockner. “SILVA: A Comprehensive Online Resource for Quality Checked and Aligned Ribosomal RNA Sequence Data Compatible with ARB.” Nucleic Acids Research 35

# Genomes Online Database

- GOLD - a manually curated data management system that catalogs sequencing projects with associated metadata from around the world
- Projects are organized based on a four level classification system: Study, Organism or Biosample, Sequencing Project and Analysis Project.
- As of April 2021, 49,625 Studies, 410,963 Organisms, 133,276 Biosamples, 411,214 Sequencing Projects and 321,376 Analysis Projects.

NCBI Import Tracker



http://www.genomesonline.org

# MetaPhlAn3

- MetaPhlAn (Metagenomic Phylogenetic Analysis) is a computational tool for profiling the composition of microbial communities from metagenomic shotgun sequencing data.
- MetaPhlAn relies on unique clade-specific marker genes identified from ~17,000 reference genomes (~13,500 bacterial and archaeal, ~3,500 viral, and ~110 eukaryotic)
  - Infers the presence and read coverage of clade-specific markers to unequivocally detect the taxonomic clades present in a microbiome sample and estimate their relative abundance
  - Unambiguous taxonomic assignments as the MetaPhlAn markers are clade-specific;
  - Accurate estimation of organismal relative abundance (in terms of number of cells rather than fraction of reads);
  - Species-level resolution for bacteria, archaea, eukaryotes and viruses;
  - Extensive validation of the profiling accuracy on several synthetic datasets and on thousands of real metagenomes.

https://huttenhower.sph.harvard.edu/metaphlan/

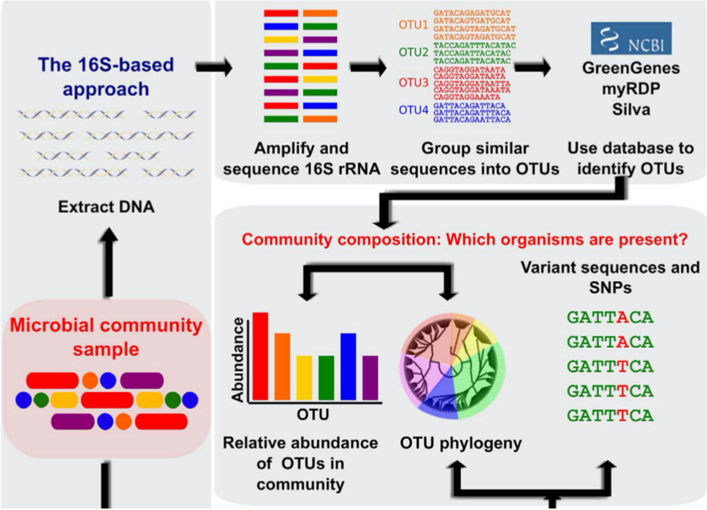
# Why MetaPhlAn?

- Uses “clade-specific” gene markers
    - A clade represents a set of genomes that can be as broad as a phylum or as specific as a species
- Uses ~1 million markers derived from 17,000 genomes
    - ~13,500 bacterial and archaeal, ~3,500 viral, and ~110 eukaryotic
    - Can identify down to the species level (and possibly even strain level)
- Can handle millions of reads on a standard computer within a few minutes
- Main Disadvantage: not all reads are assigned a taxonomic label

# Using MetaPhlan

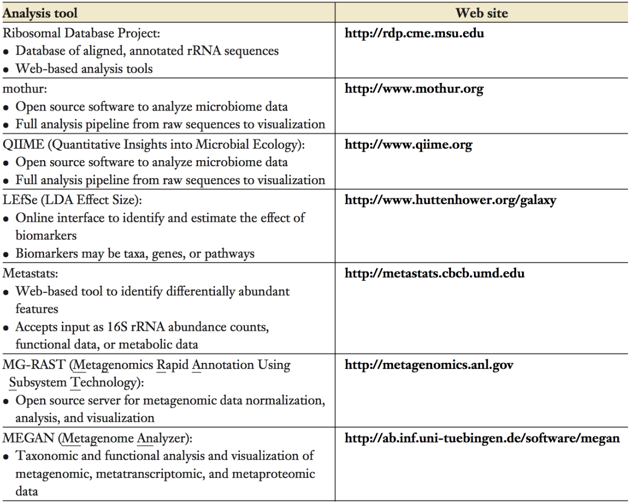
- MetaPhlan uses Bowtie2 for sequence similarity searching (nucleotide sequences vs. nucleotide database)
- Paired-end data can be used directly (but are treated as independent reads)
- Each sample is processed individually and then multiple sample can be combined together at the last step
- Output is relative abundances at different taxonomic levels

# 16S analysis overview



Morgan, Xochitl C., and Curtis Huttenhower. “Chapter 12: Human Microbiome Analysis.” PLoS Computational Biology 8, no. 12 (December 27, 2012): e1002808. https://doi.org/10.1371/journal.pcbi.1002808

# Tools for 16S ribosomal RNA analysis



| Analysis tool | Web site |
|---|---|
| Ribosomal Database Project:<br>• Database of aligned, annotated rRNA sequences<br>• Web-based analysis tools | **http://rdp.cme.msu.edu** |
| mothur:<br>• Open source software to analyze microbiome data<br>• Full analysis pipeline from raw sequences to visualization | **http://www.mothur.org** |
| QIIME (Quantitative Insights into Microbial Ecology):<br>• Open source software to analyze microbiome data<br>• Full analysis pipeline from raw sequences to visualization | **http://www.qiime.org** |
| LEfSe (LDA Effect Size):<br>• Online interface to identify and estimate the effect of biomarkers<br>• Biomarkers may be taxa, genes, or pathways | **http://www.huttenhower.org/galaxy** |
| Metastats:<br>• Web-based tool to identify differentially abundant features<br>• Accepts input as 16S rRNA abundance counts, functional data, or metabolic data | **http://metastats.cbcb.umd.edu** |
| MG-RAST (Metagenomics Rapid Annotation Using Subsystem Technology):<br>• Open source server for metagenomic data normalization, analysis, and visualization | **http://metagenomics.anl.gov** |
| MEGAN (Metagenome Analyzer):<br>• Taxonomic and functional analysis and visualization of metagenomic, metatranscriptomic, and metaproteomic data | **http://ab.inf.uni-tuebingen.de/software/megan** |

http://www.annualreviews.org/doi/10.1146/annurev-genom-090711-163814

Section 4

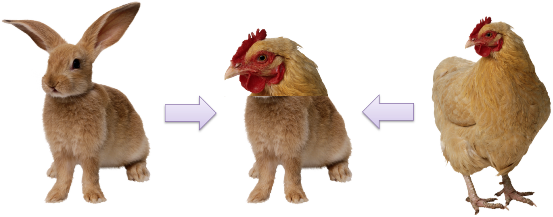
# 16S sequencing issues

# Contamination

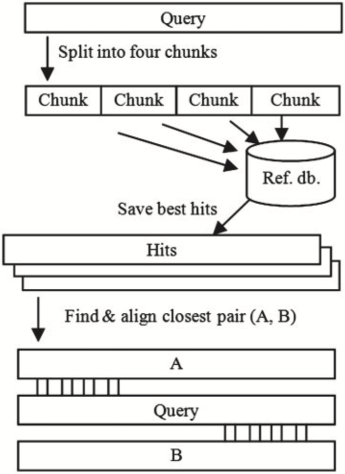
- Extraction process can introduce contamination from the lab
- Reagents may be contaminated with bacterial DNA
- Include **Extraction Negative Control** in your experiments!
- Especially crucial if samples have low DNA yield!
- Host / Environment (metagenomics sequencing):
- Host DNA often ends up in the microbiome sample
- http://hmpdacc.org/doc/HumanSequenceRemoval_SOP.pdf
- Unwanted fractions (e.g. eukaryotes) can be filtered by **cell size-selection** prior to DNA extraction
- Unwanted DNA can be removed by **subtractive hybridization**

# Chimeras - PCR artefacts

- Chimeric sequences that stem from two or more original sequences (the parents of the chimera).
- Incomplete extension of PCR, Template Switching at Conserved Regions
- Chimeras with two segments (bimeras) are most common, multimeras ($>$2 segments) may form at comparable rates
- Undetected chimeras may be misinterpreted as novel species, causing inflated estimates of diversity and spurious inferences of differences between populations.



# UCHIME



https://doi.org/10.1093/bioinformatics/btr381

http://drive5.com/usearch/manual/uchime_algo.html

- UCHIME and UCHIME2 are algorithms for detecting chimeric sequences
- The query sequenceis divided into four chunks, each of which is used to search the reference database.
- The best few hits to each chunk are saved, and the closest two sequences are found by calculating smoothed identity with the query.

Section 5

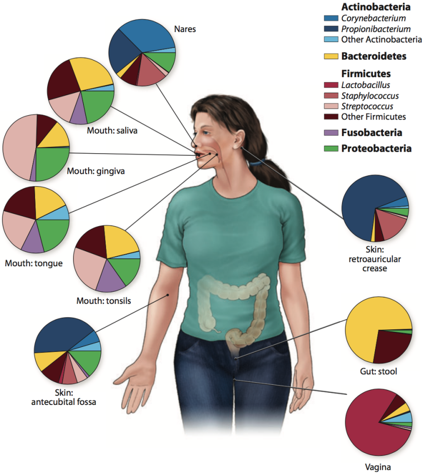
# Metagenomics data

# Human Microbiome Project (HMP)

- A resources to facilitate characterization of the human microbiota to further our understanding of how the microbiome impacts human health and disease
- Characterized the microbial communities from 300 healthy individuals, across several different sites on the human body: nasal passages, oral cavity, skin, gastrointestinal tract, and urogenital tract
- 16S rRNA sequencing

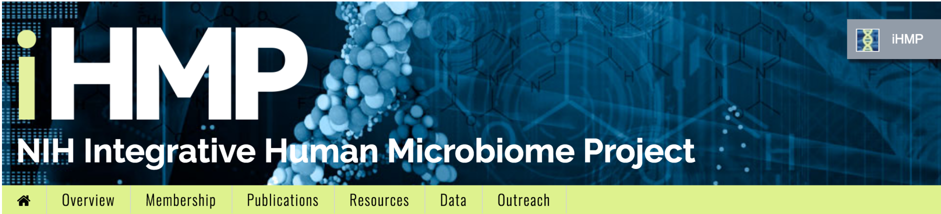
# Human Microbiome Project (HMP)

Longitudinally assess microbial diversity of 250 healthy subjects at 5 major body sites



# Integrative Human Microbiome Project (iHMP)

- Integrative molecular perspectives on microbial activity during dysbiosis - multi-omic data resources
- Pregnancy and Preterm Birth, Inflammatory Bowel Disease, Type 2 diabetes



# iHMP multi-omics data

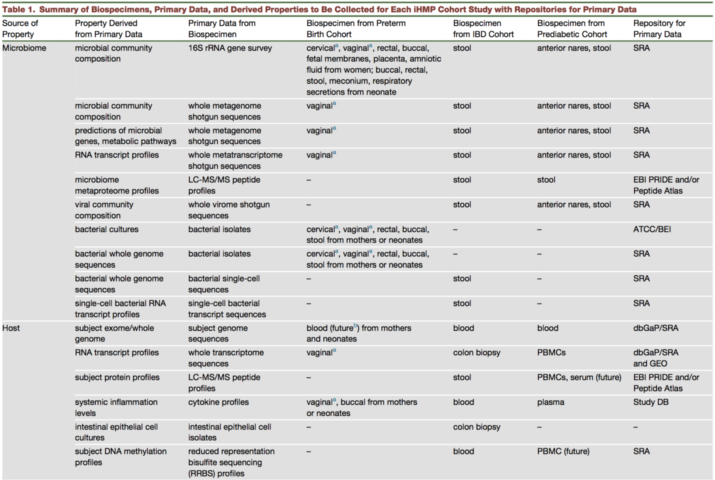


**Table 1. Summary of Biospecimens, Primary Data, and Derived Properties to Be Collected for Each iHMP Cohort Study with Repositories for Primary Data**

| Source of Property | Property Derived from Primary Data | Primary Data from Biospecimen | Biospecimen from Preterm Birth Cohort | Biospecimen from IBD Cohort | Biospecimen from Prediabetic Cohort | Repository for Primary Data |
|---|---|---|---|---|---|---|
| Microbiome | microbial community composition | 16S rRNA gene survey | cervical[a], vaginal[a], rectal, buccal, fetal membranes, placenta, amniotic fluid from women; buccal, rectal, stool, meconium, respiratory secretions from neonate | stool | anterior nares, stool | SRA |
| | microbial community composition | whole metagenome shotgun sequences | vaginal[a] | stool | anterior nares, stool | SRA |
| | predictions of microbial genes, metabolic pathways | whole metagenome shotgun sequences | vaginal[a] | stool | anterior nares, stool | SRA |
| | RNA transcript profiles | whole metatranscriptome shotgun sequences | vaginal[a] | stool | anterior nares, stool | SRA |
| | microbiome metaproteome profiles | LC-MS/MS peptide profiles | – | stool | stool | EBI PRIDE and/or Peptide Atlas |
| | viral community composition | whole virome shotgun sequences | – | stool | anterior nares, stool | SRA |
| | bacterial cultures | bacterial isolates | cervical[a], vaginal[a], rectal, buccal, stool from mothers or neonates | – | – | ATCC/BEI |
| | bacterial whole genome sequences | bacterial isolates | cervical[a], vaginal[a], rectal, buccal, stool from mothers or neonates | – | – | SRA |
| | bacterial whole genome sequences | bacterial single-cell sequences | – | stool | – | SRA |
| | single-cell bacterial RNA transcript profiles | single-cell bacterial transcript sequences | – | stool | – | SRA |
| Host | subject exome/whole genome | subject genome sequences | blood (future[b]) from mothers and neonates | blood | blood | dbGaP/SRA |
| | RNA transcript profiles | whole transcriptome sequences | vaginal[a] | colon biopsy | PBMCs | dbGaP/SRA and GEO |
| | subject protein profiles | LC-MS/MS peptide profiles | – | stool | PBMCs, serum (future) | EBI PRIDE and/or Peptide Atlas |
| | systemic inflammation levels | cytokine profiles | vaginal[a], buccal from mothers or neonates | blood | plasma | Study DB |
| | intestinal epithelial cell cultures | intestinal epithelial cell isolates | – | colon biopsy | – | – |
| | subject DNA methylation profiles | reduced representation bisulfite sequencing (RRBS) profiles | – | blood | PBMC (future) | SRA |

and more.

Integrative HMP (iHMP) Research Network Consortium. “The Integrative Human Microbiome Project: Dynamic Analysis of

# curatedMetagenomicData

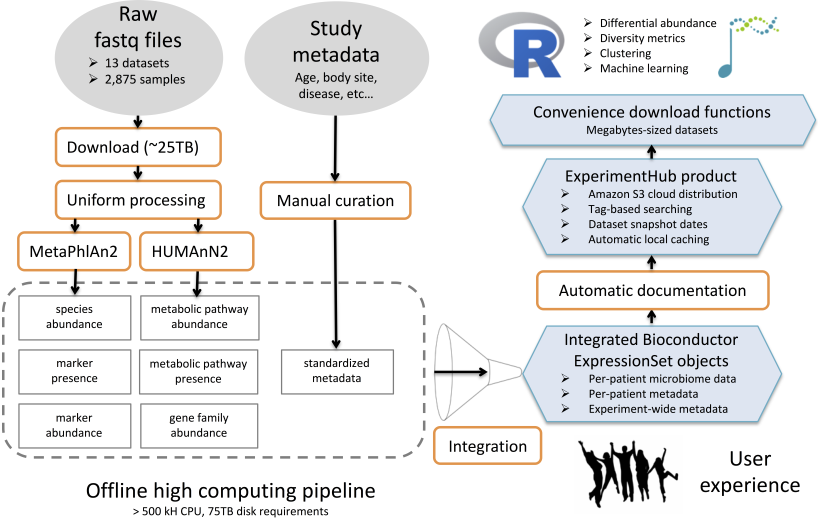
- The curatedMetagenomicData package provides microbial taxonomic, functional, and gene marker abundance for samples collected from different bodysites from thousands of people
- Matched health and socio-demographic data are provided
- Accessible via ExperimentHub

https://bioconductor.org/packages/release/data/experiment/html/curatedMetagenomicData.html, https://waldronlab.github.io/curatedMetagenomicData/

Pasolli, Edoardo, Lucas Schiffer, Audrey Renson, Valerie Obenchain, Paolo Manghi, Duy Tin Truong, Francesco Beghini, et al. "Accessible, Curated Metagenomic Data through ExperimentHub," January 27, 2017. doi:10.1101/103085.

A 6-minute video describing curatedMetagenomicData and giving a simple usage scenario, https://www.youtube.com/watch?v=ZAbaXAQpZPE

# curatedMetagenomicData pipeline



# curatedMetagenomicData

| Dataset | Samples | Citation |
|---|---|---|
| HMP_2012 | 749 | Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature 486, 207–214 (2012). |
| KarlssonFH_2013 | 145 | Karlsson, F. H. et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature 498, 99–103 (2013). |
| LeChatelierE_2013 | 292 | Le Chatelier, E. et al. Richness of human gut microbiome correlates with metabolic markers. Nature 500, 541–546 (2013). |
| LomanNJ_2013_Hi | 44 | Loman, N. J. et al. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. JAMA 309, 1502–1510 (2013). |
| LomanNJ_2013_Mi | 9 | Loman, N. J. et al. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. JAMA 309, 1502–1510 (2013). |
| NielsenHB_2014 | 396 | Nielsen, H. B. et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol. 32, 822–828 (2014). |
| Obregon_TitoAJ_2015 | 58 | Obregon-Tito, A. J. et al. Subsistence strategies in traditional societies distinguish gut microbiomes. Nat Commun 6, 6505 (2015). |
| OhJ_2014 | 291 | Oh, J. et al. Biogeography and individuality shape function in the human skin metagenome. Nature 514, 59–64 (2014). |
| QinJ_2012 | 363 | Qin, J. et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490, 55–60 (2012). |
| QinN_2014 | 237 | Qin, N. et al. Alterations of the human gut microbiome in liver cirrhosis. Nature 513, 59–64 (2014). |
| RampelliS_2015 | 38 | Rampelli, S. et al. Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. Curr. Biol. 25, 1682–1693 (2015). |
| TettAJ_2016 | 97 | Ferretti, P. et al. Experimental metagenomics and ribosomal profiling of the human skin microbiome. Exp. Dermatol. (2016). doi:10.1111/exd.13210 |
| ZellerG_2014 | 156 | Zeller, G. et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol. Syst. Biol. 10, 766 (2014). |

Section 6

# Whole-sample microbiome sequencing

# 16S vs Metagenomics

16S is targeted sequencing of a single gene which acts as a marker for identification

**Pros**

- Well established
- Sequencing costs are relatively cheap (~50,000 reads/sample)
- Only amplifies what you want (no host contamination)

**Cons**

- Primer choice can bias results towards certain organisms
- Usually not enough resolution to identify to the strain level
- Need different primers usually for archaea & eukaryotes (18S)
- Doesn’t identify viruses

# Metagenomics: sequencing all the DNA in a sample

**Pros**

- No primer bias
- Can identify all microbes (euks, viruses, etc.)
- Provides functional information (“What are they doing?”)

**Cons**

- More expensive (millions of sequences needed)
- Host/site contamination can be significant
- May not be able to sequence “rare” microbes
- Complex bioinformatics

# Metagenomics: Who is there?

**Goal:** Identify the relative abundance of different microbes in a sample given using metagenomics

**Problems:**

- Reads are all mixed together
- Reads can be short (∼100bp)
- Lateral gene transfer

**Two broad approaches**

- Binning Based
- Marker Based

# How many reads?

**Initial estimates:**

- 1,000 as a bare minimum, now obsolete, current sequencing procudes enough reads

**Now:**

- Illumina MiSeq generates 2x300 bp paired end for amplicon and bacterial whole-genome sequencing.
- HiSeq generates 200,000,000 reads/lane for metagenomics.
- PacBio for long reads both for complete microbial genome assembly and shotgun metagenomics to scaffold reads.

# LCA: Lowest Common Ancestor

Use all BLAST hits above a threshold and assign taxonomy at the lowest level in the tree which covers these taxa.

Software Examples:

- **MEGAN**: http://ab.inf.uni-tuebingen.de/software/megan6/
  - One of the first metagenomic tools
  - Does functional profiling too!
- **MG-RAST**: https://metagenomics.anl.gov/
  - Web-based pipeline (might need to wait awhile for results)
- **Kraken**: https://ccb.jhu.edu/software/kraken/
  - Fastest binning approach to date and very accurate.
  - Large computing requirements (e.g. >128GB RAM)

# Bacterial genome assembly

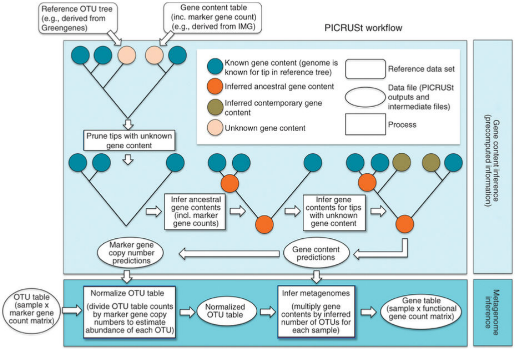
- How to Assemble a Bacterial Genome: Gram-negative is ~6,000,000 base pair
- Shotgun sequence 2x300 bp fragments on Illumina MiSeq at 30-fold redundancy.
- Overlapping reads form large DNA contigs with N50 of ~100 kb.
- Or very low coverage (3-5X) just to define species and strain

# Assemblers (de novo)

- Phrap
- Celera
- Velvet
- SPAdes
- mira
- MaSuRCA
- ALL-PATHS

# PICRUSt: Phylogenetic Investigation of Communities by Reconstruction of Unobserved States

PICRUSt (pronounced "pie crust") predicts metagenome functional content from marker gene (e.g., 16S rRNA) surveys and full genomes.


Reference OTU tree (e.g., derived from Greengenes)
Gene content table (inc. marker gene count) (e.g., derived from IMG)
PICRUSt workflow
Known gene content (genome is known for tip in reference tree)
Inferred ancestral gene content
Inferred contemporary gene content
Unknown gene content
Reference data set
Data file (PICRUSt outputs and intermediate files)
Process
Prune tips with unknown gene content
Infer ancestral gene contents (incl. marker gene counts)
Infer gene contents for tips with unknown gene content
Gene content inference (precomputed information)
Marker gene copy number predictions
Gene content predictions
OTU table (sample x marker gene count matrix)
Normalize OTU table (divide OTU table counts by marker gene copy numbers to estimate abundance of each OTU)
Normalized OTU table
Infer metagenomes (multiply gene contents by inferred number of OTUs for each sample)
Gene table (sample x functional gene count matrix)
Metagenome inference

https://picrust.github.io/picrust/

# Metatranscriptomics - microbial gene expression

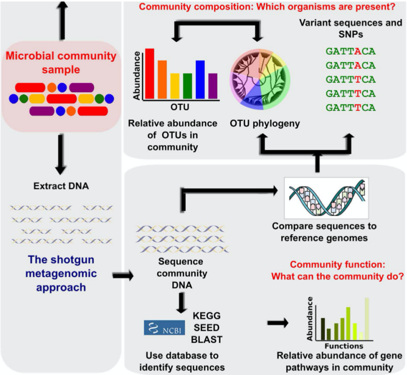
**Challenges:**

- Lack of a polyA signal makes it difficult to isolate bacterial mRNA and resulting in (massive) rRNA contamination
- Environmental microbiome samples lack reference genomes making it difficult to map reads back to their source transcripts
- ~5 million mRNA reads provide 90-95% of expression context in a microbiome
- With kits yielding mRNA read rates of ~25%, this suggests 20 million/sample mRNA

Section 7

# Analysis considerations

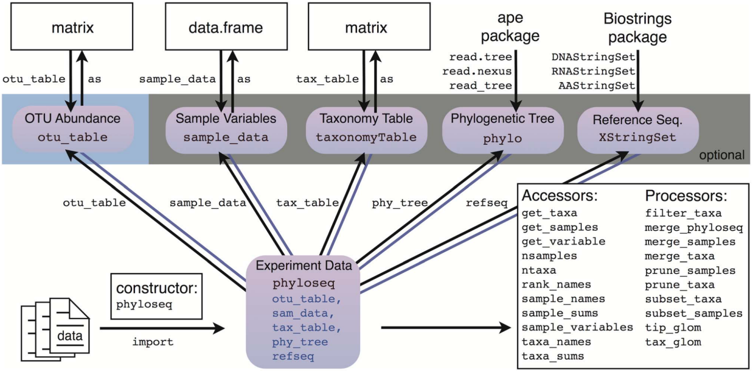
# Bioinformatic methods for functional metagenomics


Community composition: Which organisms are present?
Variant sequences and SNPs
Microbial community sample
Abundance
OTU
GATTACA
GATTACA
GATTTCA
GATTTCA
GATTTCA
Relative abundance of OTUs in community
OTU phylogeny
Extract DNA
The shotgun metagenomic approach
Compare sequences to reference genomes
Sequence community DNA
Community function: What can the community do?
NCBI
KEGG
SEED
BLAST
Abundance
Functions
Use database to identify sequences
Relative abundance of gene pathways in community

Morgan, Xochitl C., and Curtis Huttenhower. “Chapter 12: Human Microbiome Analysis.” PLoS Computational Biology 8, no. 12 (December 27, 2012): e1002808. https://doi.org/10.1371/journal.pcbi.1002808

# phyloseq R package

- High-level analysis of processed metagenomic sequencing data
- Imports various file formats
- Performs clustering, dimensionaliry reduction, visualization, differential analysis



matrix
otu_table
as
data.frame
sample_data
as
matrix
tax_table
as
ape package
read.tree
read.nexus
read_tree
Biostrings package
DNAStringSet
RNAStringSet
AAStringSet

OTU Abundance
otu_table
Sample Variables
sample_data
Taxonomy Table
taxonomyTable
Phylogenetic Tree
phylo
Reference Seq.
XStringSet
optional

otu_table
sample_data
tax_table
phy_tree
refseq

constructor:
phyloseq

data
import

Experiment Data
phyloseq
otu_table,
sam_data,
tax_table,
phy_tree
refseq

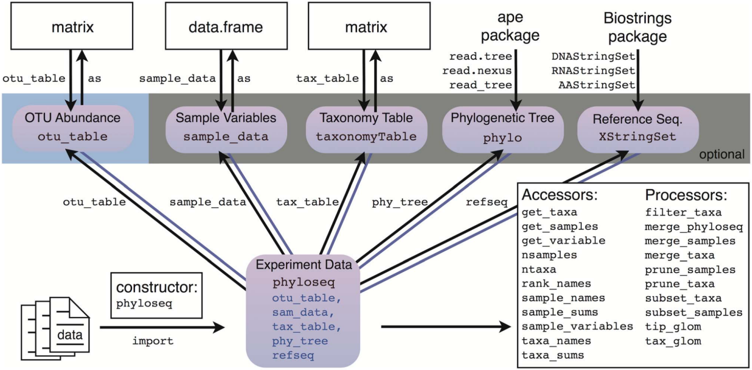
Accessors:
get_taxa
get_samples
get_variable
nsamples
ntaxa
rank_names
sample_names
sample_sums
sample_variables
taxa_names
taxa_sums

Processors:
filter_taxa
merge_phyloseq
merge_samples
merge_taxa
prune_samples
prune_taxa
subset_taxa
subset_samples
tip_glom
tax_glom

https://bioconductor.org/packages/phyloseq/, https://joey711.github.io/phyloseq/

# phyloseq R package

- phyloseq object contains 1) the OTU abundance table, 2) a table of sample data, 3) a table of taxonomic descriptors, and 4) a phylogenetic tree.



https://bioconductor.org/packages/phyloseq/, https://joey711.github.io/phyloseq/

McMurdie, Paul J., and Susan Holmes. “Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of

# Trimming, subsetting, filtering

- Trimming high-throughput phylogenetic sequencing data can be useful, or even necessary, for certain types of analyses.
- Filter taxa, keep most abundant taxa across samples.
- Filter samples, keep samples with sufficient number of reads.

# Normalization

- **Rarefaction** - involves selecting a specified number of samples that is equal to or less than the number of samples in the smallest sample, and then randomly discarding reads from larger samples until the number of remaining samples is equal to this threshold

Amy D. Willis “Rarefaction, Alpha Diversity, and Statistics” Front. Microbiol., 23 October 2019, https://doi.org/10.3389/fmicb.2019.02407

Paul J. McMurdie,Susan Holmes. “Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible” April 3, 2014 https://doi.org/10.1371/journal.pcbi.1003531

# Normalization

- **Total-sum scaling (TSS)** - Divides feature read counts (the number of reads from a particular sample that cluster within the same OTU) by the total number of reads in each sample, i.e., it converts feature counts to appropriately scaled ratios.
  - $M(m,n)$ - a count matrix, where $m$ and $n$ are the number of taxonomic features and samples, respectively
  - $c_{i,j}$ - the number of times taxonomic feature $i$ was observed in sample $j$.
  - $s_j = \sum_i(c_{i,j})$ - sum of counts for sample $i$
  - $c_{i,j}^{norm} = c_{i,j}/s_j$ - Total-sum scaling normalization

# Normalization

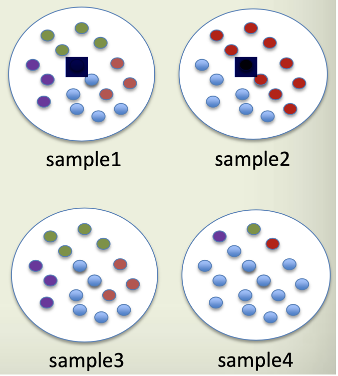
- **Cumulative-sum scaling (CSS)** - raw counts are divided by the cumulative sum of counts up to a percentile determined using a data-driven approach, e.g., the 75th percentile of each sample's nonzero count distribution.
  - $q_j^l$ - $l^{th}$ quantile of sample $j$
  - $s_j^l = \sum_{i|c_{i,j} \leq q_j^l} (c_{i,j})$ - the sum of counts for sample $j$ up to the $l^{th}$ quantile
  - $c_{i,j}^{norm} = (c_{i,j}/s_j^l)N$ - Cumulative-sum scaling, $N$ is an appropriately chosen normalization constant

Paulson, Joseph N, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. "Differential Abundance Analysis for Microbial Marker-Gene Surveys." Nature Methods 10, no. 12 (September 29, 2013): 1200–1202. https://doi.org/10.1038/nmeth.2658.

# Alpha diversity

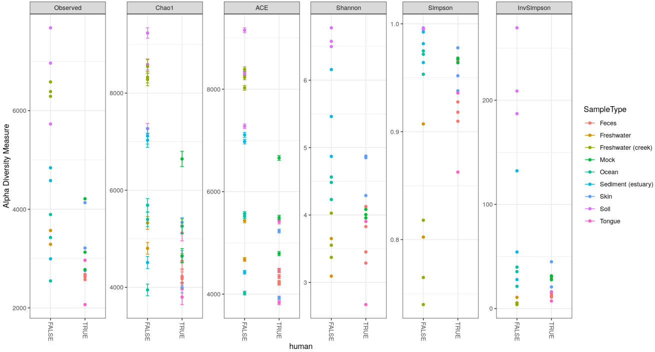
Diversity of microbes in a single sample.

Count of different microbes (OTU count)
**Richness** - Richness is a measure of number of species present in a sample.
Distribution of different microbes
**Evenness** - Evenness is a measure of relative abundance of different species that make up the richness in that area



# Alpha diversity

Commonly used diversity metrics - **Observed** (measure richness only) - **Chao1** (measures richness and evenness) - **Shannon** $H = -\sum_{i=1}^{S}(p_i ln(p_i))$ - **Simpson** $D = \sum_{i=1}^{S} p_i^2$, where $p_i$ is the fraction of total species comprised by species $i$



https://bioconductor.org/packages/release/bioc/vignettes/phyloseq/inst/doc/phyloseq-analysis.html

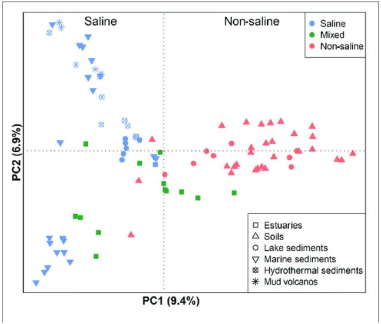
# Beta diversity

- Variation of microbial communities between samples
- Requires the definition of distance between samples
  - **Bray–Curtis dissimilarity** - based on differences in microbial abundances between two samples (e.g., at species level). $BC_{ij} = \frac{S_i + S_j - 2C_{ij}}{S_i + S_j}$, where $S_x$ are the number of species in population $x$, and $C_{ij}$ is the total number of species at the location with the fewest species. Ranges from 0 (samples share the same species at exact the same abundances) to 1 (samples have complete different species abundances).
  - **Jaccard distance** - based on presence or absence of species (different in microbial composition between two samples), does not include abundance information. Ranges from 0 (samples share exact the same species) to 1 (samples have no species in common)
  - **UniFrac** - sequence distances (phylogenetic tree), based on the fraction of phylogenetic branch length that is shared between two samples or unique to one or the other sample. Weighted and unweighted variants include/do not include abundance information, respectively.

http://www.metagenomics.wiki/pdf/definition/alpha-beta-diversity

# Ordination methods

- Microbiome "landscaping", dimensionality reduction
- Many methods, e.g., Principal Coordinates Analysis (PCoA), Multidimensional scaling (MDS), non-metric Multi-Dimensional Scaling (NMDS), Canonical correspondence analysis (CCA), t-SNE



O. Paliy V. Shankar. "Application of multivariate statistical techniques in microbial ecology" 20 January 2016 https://doi.org/10.1111/mec.13536

https://www.researchgate.net/publication/318491697 Global Biogeographic Analysis of Methanogenic Archaea

# Differential abundance

- Microbiome sequencing, unlike RNA-seq, is very sparse - most OTUs in marker-gene studies are rare (that is, absent from a large number of samples).
- This sparsity is due to both biological and technical phenomena: some organisms are found in only a small percentage of samples, whereas others are simply not detected owing to insufficient sequencing depth.
- These phenomena can lead to strong biases when data sets are scaled for comparison and when sequence read counts are tested for significant differences.

# Differential abundance

- Solution: a zero-inflated Gaussian (ZIG) distribution mixture model that accounts for biases in differential abundance testing resulting from undersampling of the microbial community
- The components of the mixture model correspond to normally distributed log abundances in each group of interest: for example, case or control, and a spike-mass at 0 indicating absence of the feature owing to undersampling
- The model estimates the probability that an observed zero is generated from the detection distribution due to undersampling or from the actual absense of the feature. EM algorithm.
- `metagenomeSeq` - Statistical analysis for sparse high-throughput sequencing, https://bioconductor.org/packages/metagenomeSeq/

Paulson, Joseph N, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. "Differential Abundance Analysis for Microbial Marker-Gene Surveys." Nature Methods 10, no. 12 (September 29, 2013): 1200–1202. https://doi.org/10.1038/nmeth.2658.

# References

https://github.com/mdozmorov/Microbiome_notes

https://github.com/stevetsa/awesome-microbes