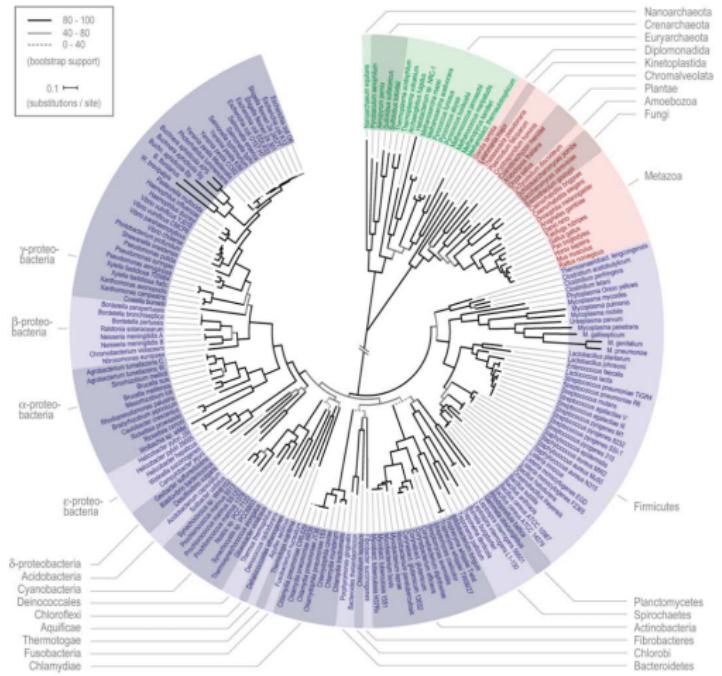


# Metagenomics

Mikhail Dozmorov

Spring 2018

# 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

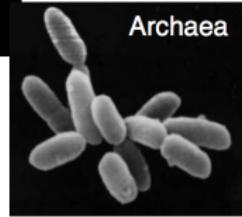
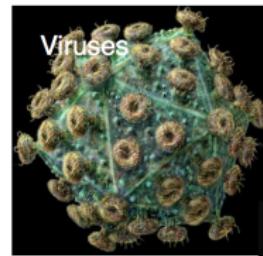
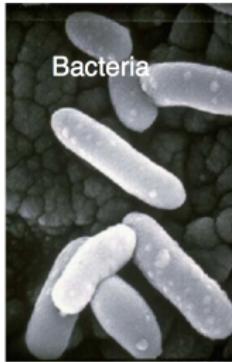
# Microbiome - our second genome

- **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>

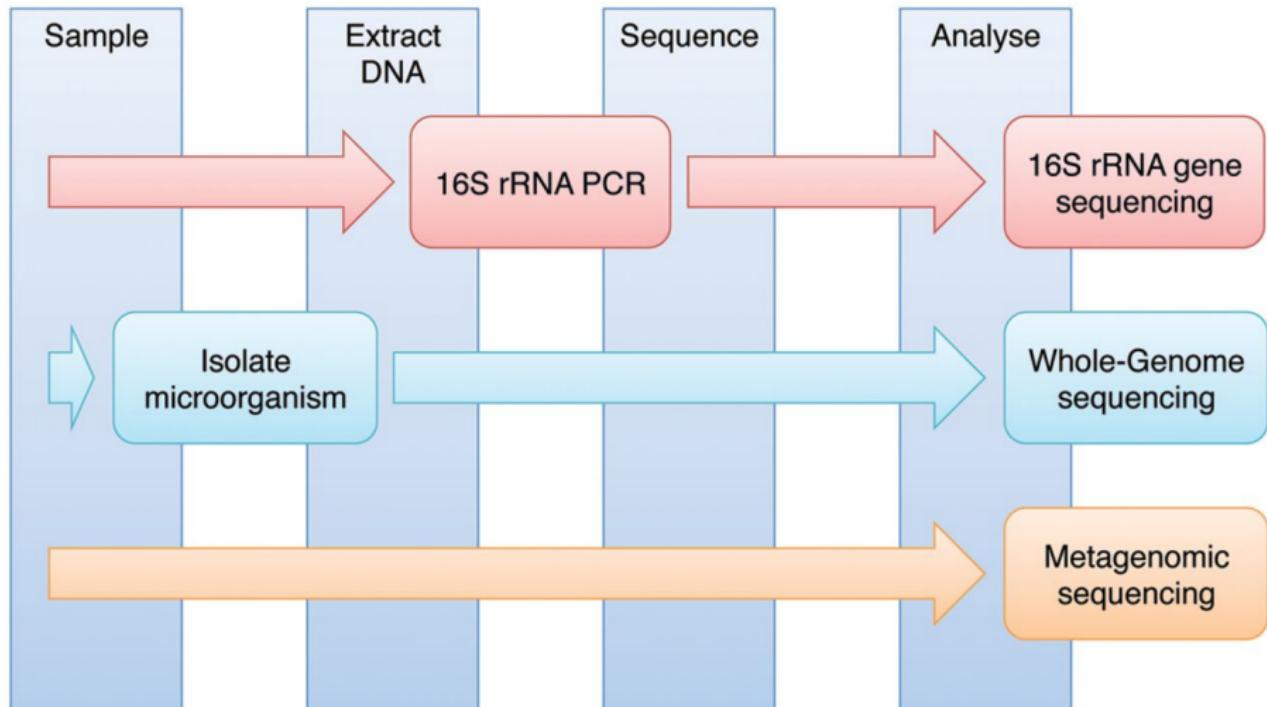
# Human Microbiome

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 preferable

# 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.

# 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 allow 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

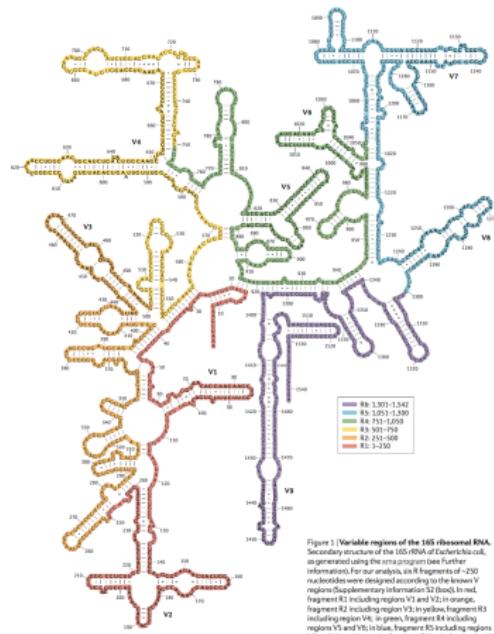
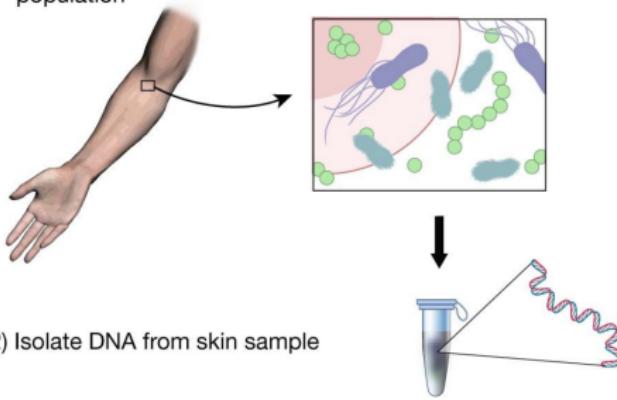


Figure 1 | Variable regions of the 16S ribosomal RNA. Secondary structure of the 16S rRNA of *Escherichia coli*, interconnecting the rRNA genes (see figure 1 for more information). The total length of the rRNA molecule is ~250 nucleotides were designed according to the known V regions and their variability in different bacteria. In red, fragment R1 including regions V1 and V2; in orange, fragment R2 including regions V3 and V4; in light blue, fragment R3 including region V5; in green, fragment R4 including regions V6 and V7; in blue, fragment R5 including regions V7 and V8; in purple, fragment R6 including region V9.

# Bacterial 16S rRNA sequencing workflow

(1) Obtain superficial skin sample containing mixed bacterial population

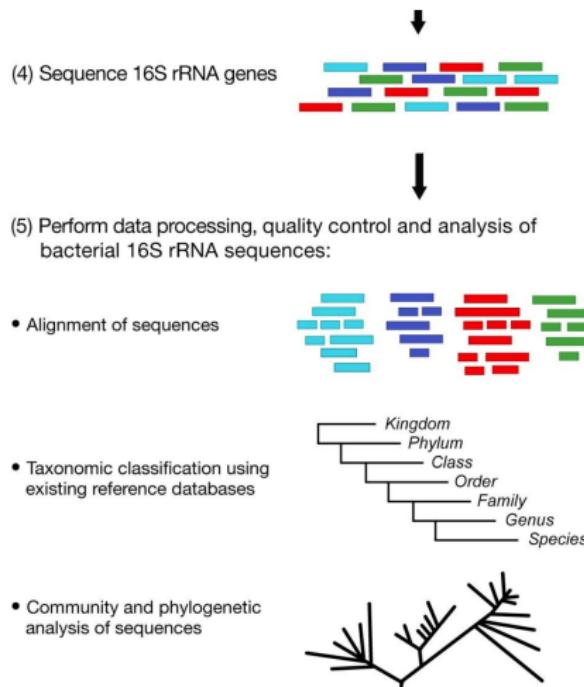


(2) Isolate DNA from skin sample

(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>.

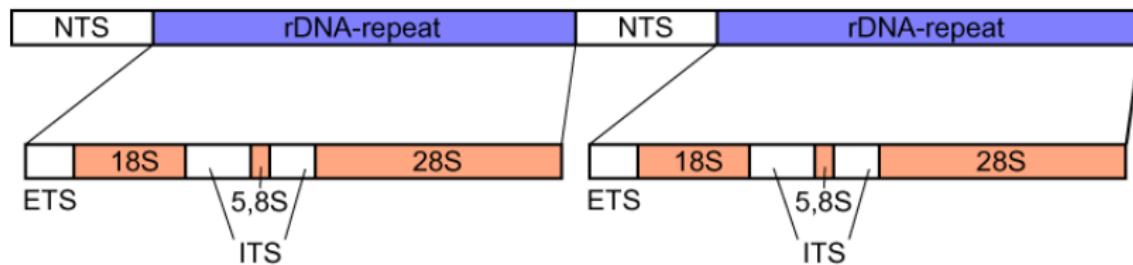
# Bacterial 16S rRNA sequencing workflow



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](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](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

# 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 phlyotyping (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
<b>Number of taxa</b>	568	201	85	39	23
<b>Median sequence identity</b>	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)
<b>Minimum sequence identity</b>	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)
<b>Threshold sequence identity</b>	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>.

# Marker databases

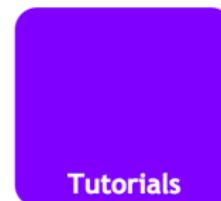
# RDP Database

- 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



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

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

  
high quality ribosomal RNA databases

Home SILVAngs Browser Search Aligner Download Documentation Projects FISH & Probes Contact

### 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](#)

### News

12.09.2017 [Registration for the Primer 2 Paper Workshop now open](#)  
 The next SILVA Primer 2 Paper workshop takes place from November 6 to 10 in Bremen. If you are interested please register quickly since space is limited to 11 participants.

24.08.2017 [MISAG and MIMAG standards published](#)  
 Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea published in Nature Biotechnology

24.07.2017 [ARB SILVA review published](#)  
 Our ARB SILVA review is now published as part of the de.NBI Special Issue: 25 years of serving the community with ribosomal RNA gene reference databases and tools.

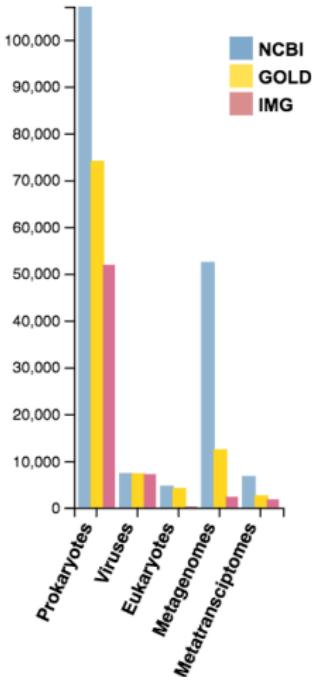
01.06.2017 [The 8th de.NBI Quaterly Newsletter published](#)  
 Highlights: The integration of German and the de.NBI

  
Check out our new service for Next Generation Amplicon data

# 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 January 2017, 26,117 Studies, 239,100 Organisms, 15,887 Biosamples, 97,212 Sequencing Projects and 78,579 Analysis Projects.

<http://www.genomesonline.org>



# SpecI

Species identification tool based on 40 universal phylogenetic marker genes.

[Home](#)[Overview](#)[Marker Genes](#)[Downloads](#)[Citation](#)

# specI

A species identification tool

<http://vm-lux.embl.de/~mende/specI/>

# MetaPhlAn2

- 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
- Eukaryotic and viral quantification, in addition to bacteria and archaea
- MetaPhlAn2 relies on ~1M unique clade-specific marker genes identified from ~17,000 reference genomes (~13,500 bacterial and archaeal, ~3,500 viral, and ~110 eukaryotic)

<http://segatalab.cibio.unitn.it/tools/metaphlan2/>

# 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

# Tools for 16S ribosomal RNA analysis

Analysis tool	Web site
Ribosomal Database Project: <ul style="list-style-type: none"> <li>• Database of aligned, annotated rRNA sequences</li> <li>• Web-based analysis tools</li> </ul>	<a href="http://rdp.cme.msu.edu">http://rdp.cme.msu.edu</a>
mothur: <ul style="list-style-type: none"> <li>• Open source software to analyze microbiome data</li> <li>• Full analysis pipeline from raw sequences to visualization</li> </ul>	<a href="http://www.mothur.org">http://www.mothur.org</a>
QIIME (Quantitative Insights into Microbial Ecology): <ul style="list-style-type: none"> <li>• Open source software to analyze microbiome data</li> <li>• Full analysis pipeline from raw sequences to visualization</li> </ul>	<a href="http://www.qiime.org">http://www.qiime.org</a>
LEfSe (LDA Effect Size): <ul style="list-style-type: none"> <li>• Online interface to identify and estimate the effect of biomarkers</li> <li>• Biomarkers may be taxa, genes, or pathways</li> </ul>	<a href="http://www.huttenhower.org/galaxy">http://www.huttenhower.org/galaxy</a>
Metastats: <ul style="list-style-type: none"> <li>• Web-based tool to identify differentially abundant features</li> <li>• Accepts input as 16S rRNA abundance counts, functional data, or metabolic data</li> </ul>	<a href="http://metastats.cbcb.umd.edu">http://metastats.cbcb.umd.edu</a>
MG-RAST (Metagenomics Rapid Annotation Using Subsystem Technology): <ul style="list-style-type: none"> <li>• Open source server for metagenomic data normalization, analysis, and visualization</li> </ul>	<a href="http://metagenomics.anl.gov">http://metagenomics.anl.gov</a>
MEGAN (Metagenome Analyzer): <ul style="list-style-type: none"> <li>• Taxonomic and functional analysis and visualization of metagenomic, metatranscriptomic, and metaproteomic data</li> </ul>	<a href="http://ab.inf.uni-tuebingen.de/software/megan">http://ab.inf.uni-tuebingen.de/software/megan</a>

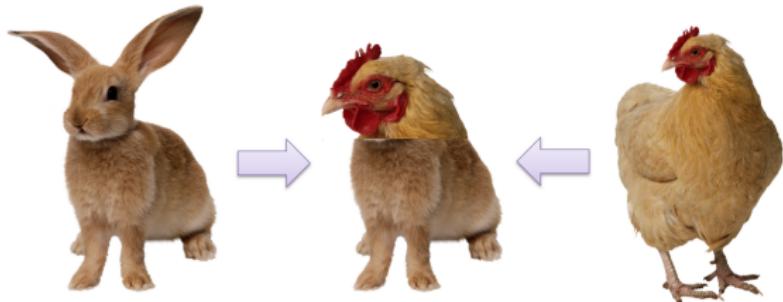
# 16S sequencing issues

# 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](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.



# Greengenes

Greengenes - a Chimera-Checked 16S rRNA Gene Database



## Download Area

Below are links to three download areas:

- You can download The Greengenes Database files from May, 2013.
- You can download The Greengenes Database files from October, 2012.
- You can download The Greengenes Database files from August, 2012.
- **PhyloChip Datasets Downloads** : A link to datasets created with the PhyloChip 16S rRNA microarray.

## Special Collections

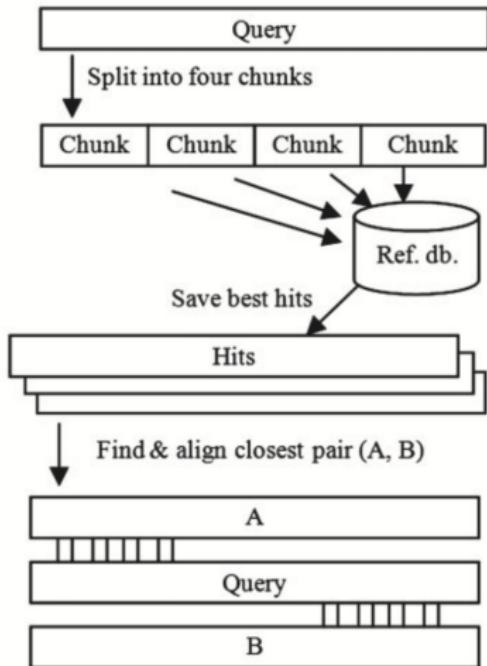
- Check out our first Special Collection: [Conlan et al., 2012, PLoS ONE](#)

<http://greengenes.lbl.gov/Download/>



The Greengenes Database by [The Greengenes Database Consortium](#) is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License.

# UCHIME



- UCHIME and UCHIME2 are algorithms for detecting chimeric sequences
- The query sequence is 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.

<https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr381>

[http://drive5.com/usearch/manual/uchime\\_algo.html](http://drive5.com/usearch/manual/uchime_algo.html)

# Metagenomics data

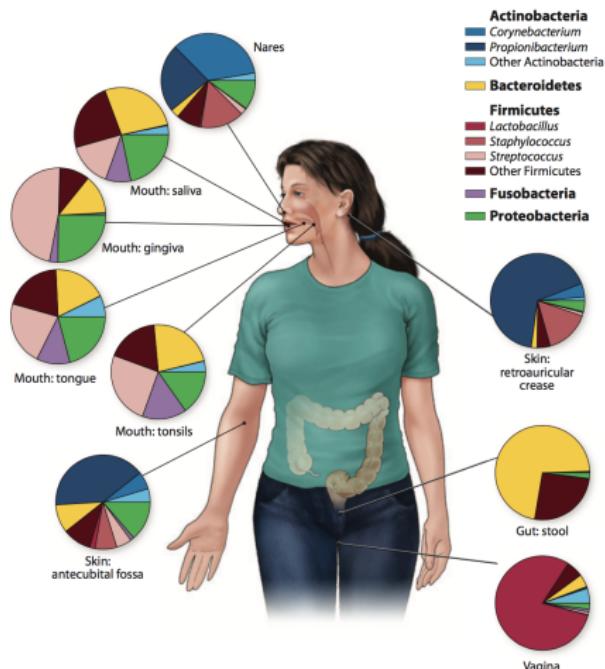
# Human Microbiome Project (HMP)

- 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



The image shows the homepage of the NIH Integrative Human Microbiome Project (iHMP). The header features the iHMP logo in large white letters on a dark blue background with a molecular structure graphic. Below the logo, the text "NIH Integrative Human Microbiome Project" is displayed. A navigation bar at the bottom includes links for Home, Overview, Membership, Publications, Resources, Data, and Outreach.

**iHMP**  
NIH Integrative Human Microbiome Project

Home Overview Membership Publications Resources Data Outreach

# 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 <sup>a</sup> , vaginal <sup>b</sup> , 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 <sup>b</sup>	stool	anterior nares, stool	SRA
	predictions of microbial genes, metabolic pathways	whole metagenome shotgun sequences	vaginal <sup>b</sup>	stool	anterior nares, stool	SRA
	RNA transcript profiles	whole metatranscriptome shotgun sequences	vaginal <sup>b</sup>	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 <sup>a</sup> , vaginal <sup>b</sup> , rectal, buccal, stool from mothers or neonates	–	–	ATCC/BEI
	bacterial whole genome sequences	bacterial isolates	cervical <sup>a</sup> , vaginal <sup>b</sup> , 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 <sup>c</sup> ) from mothers and neonates	blood	blood	dbGaP/SRA
	RNA transcript profiles	whole transcriptome sequences	vaginal <sup>b</sup>	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 <sup>b</sup> , 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.

# curatedMetagenomicData

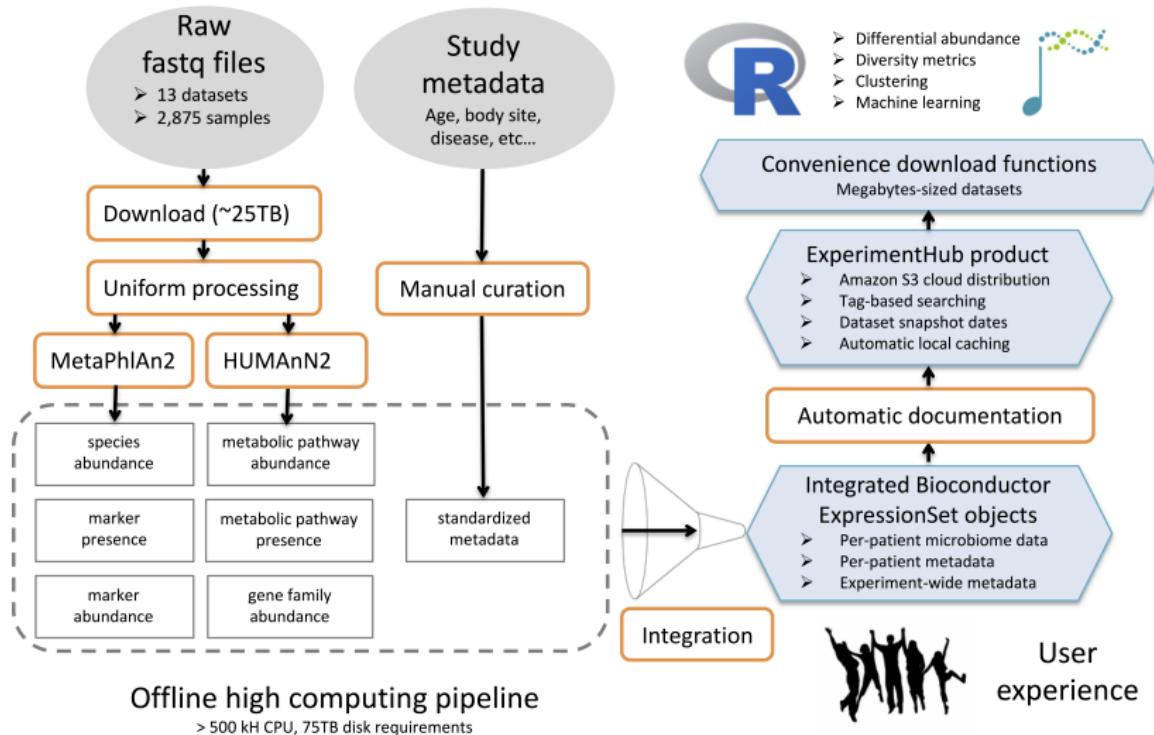
- The curatedMetagenomicData package provides microbial taxonomic, functional, and gene marker abundance for samples collected from different bodysites from thousands of people
- Matched health and socio-demographic data are provided
- Accessible via ExperimentHub

<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. <i>Nature</i> 486, 207–214 (2012).
KarlssonFH_2013	45	Karlsson, F. H. et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. <i>Nature</i> 498, 99–103 (2013).
LeChatelierE_2013	292	Le Chatelier, E. et al. Richness of human gut microbiome correlates with metabolic markers. <i>Nature</i> 500, 541–546 (2013).
LomanNJ_2013	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. <i>JAMA</i> 309, 1502–1510 (2013).
LomanNJ_2013	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. <i>JAMA</i> 309, 1502–1510 (2013).
NielsenHB_2014	86	Nielsen, H. B. et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. <i>Nat. Biotechnol.</i> 32, 822–828 (2014).
Obregon_TitoA_2015	58	Obregon-Tito, A. J. et al. Subsistence strategies in traditional societies distinguish gut microbiomes. <i>Nat Commun</i> 6, 6505 (2015).
OhJ_2014	291	Oh, J. et al. Biogeography and individuality shape function in the human skin metagenome. <i>Nature</i> 514, 59–64 (2014).
QinJ_2012	363	Qin, J. et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. <i>Nature</i> 490, 55–60 (2012).
QinN_2014	237	Qin, N. et al. Alterations of the human gut microbiome in liver cirrhosis. <i>Nature</i> 513, 59–64 (2014).
RampelliS_2015	38	Rampelli, S. et al. Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. <i>Curr. Biol.</i> 25, 1682–1693 (2015).
TettAJ_2016	97	Ferretti, P. et al. Experimental metagenomics and ribosomal profiling of the human skin microbiome. <i>Exp. Dermatol.</i> (2016). doi:10.1111/exd.13210
ZellerG_2014	156	Zeller, G. et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. <i>Mol. Syst. Biol.</i> 10, 766 (2014).

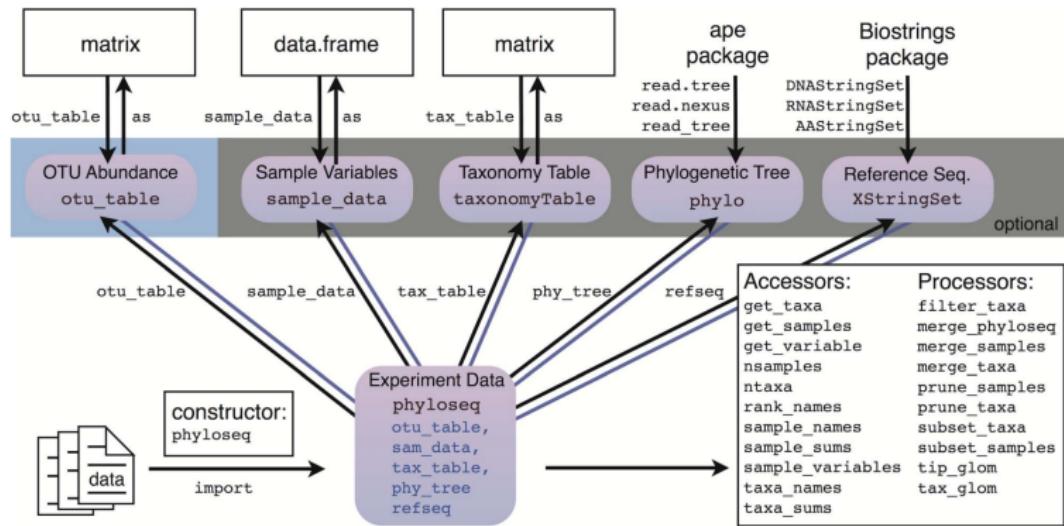
# MetaPhlAn: Metagenomic Phylogenetic Analysis

- MetaPhlAn 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 3,000 reference genomes from IMG/M: Integrated Microbial Genomes & Microbiomes
- BLAST search
- Used by the Human Microbiome project

<http://huttenhower.sph.harvard.edu/metaphlan/>

# phyloseq R package

- High-level analysis of processed metagenomic sequencing data
- Support for various file formats, clustering, dimensionality reduction, visualization
- Differential analysis



# 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 produces 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

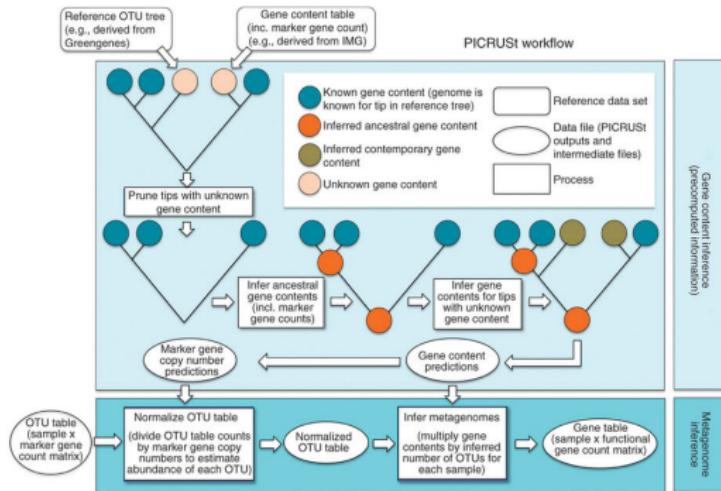
# Velvet (Zerbino and Birney, 2008)

- Works in base-space and color-space
- Good for small genomes
- Agnostic of read length

- ① Construct k-mer hash
- ② Build De Bruijn graph
- ③ Simplify graph
- ④ Resolve Tips and Bubbles

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

PICRUSt (pronounced “pie crust”) is a bioinformatics software package designed to predict metagenome functional content from marker gene (e.g., 16S rRNA) surveys and full genomes.



# STAMP - metagenomics visualization software

- STAMP is a graphical software package that provides statistical hypothesis tests and exploratory plots for analysing taxonomic and functional profiles.
- It supports tests for comparing pairs of samples or samples organized into two or more treatment groups. Effect sizes and confidence intervals are provided to allow critical assessment of the biological relevancy of test results.
- A user-friendly graphical interface permits easy exploration of statistical results and generation of publication-quality plots.

<http://kiwi.cs.dal.ca/Software/STAMP>

Parks, Donovan H., Gene W. Tyson, Philip Hugenholtz, and Robert G. Beiko. "STAMP: Statistical Analysis of Taxonomic and Functional Profiles." *Bioinformatics* (Oxford, England) 30, no. 21 (November 1, 2014): 3123–24.  
<https://doi.org/10.1093/bioinformatics/btu494>.

# 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

# Analysis considerations

# 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  $i$
  - $c_{i,j}^{\text{norm}} = c_{i,j}/s_j$  - 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>.

# Differential analysis

- 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.

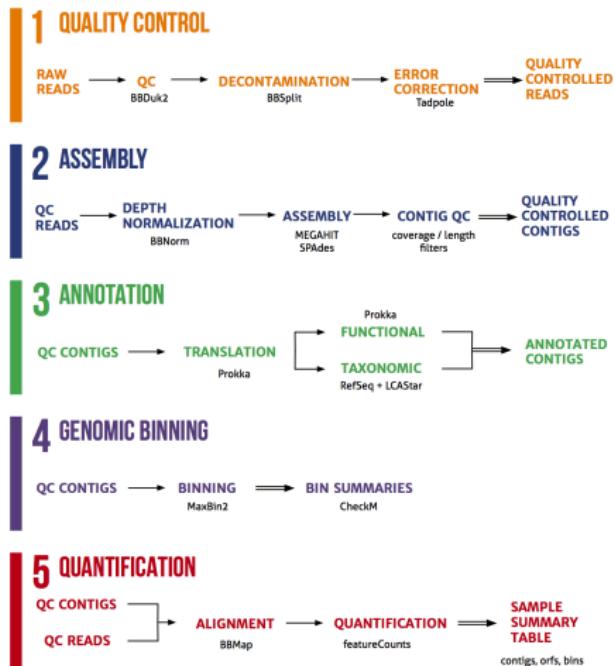
# Differential analysis

- 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 absence of the feature. EM algorithm.
- `metagenomeSeq` - Statistical analysis for sparse high-throughput sequencing, <https://bioconductor.org/packages/release/bioc/html/metagenomeSeq.html>

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>.

# Latest development

ATLAS - a framework for assembly, annotation, and genomic binning of metagenomic and metatranscriptomic data.



<https://github.com/pnnl/atlas>

<https://pnnl-atlas.readthedocs.io/en/latest/>