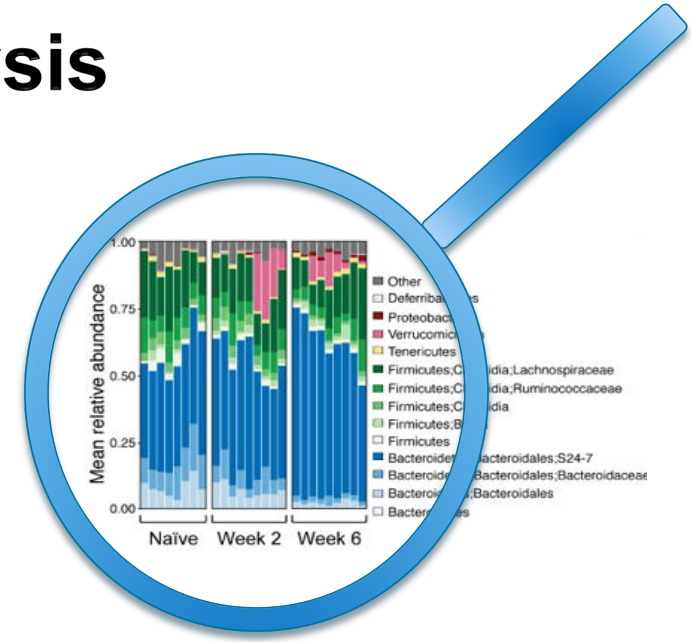
National Institute of Allergy and Infectious Diseases

Microbiome Analysis

RML Workshop
Sept 12, 2019



NIAID

National Institute of Allergy and Infectious Diseases

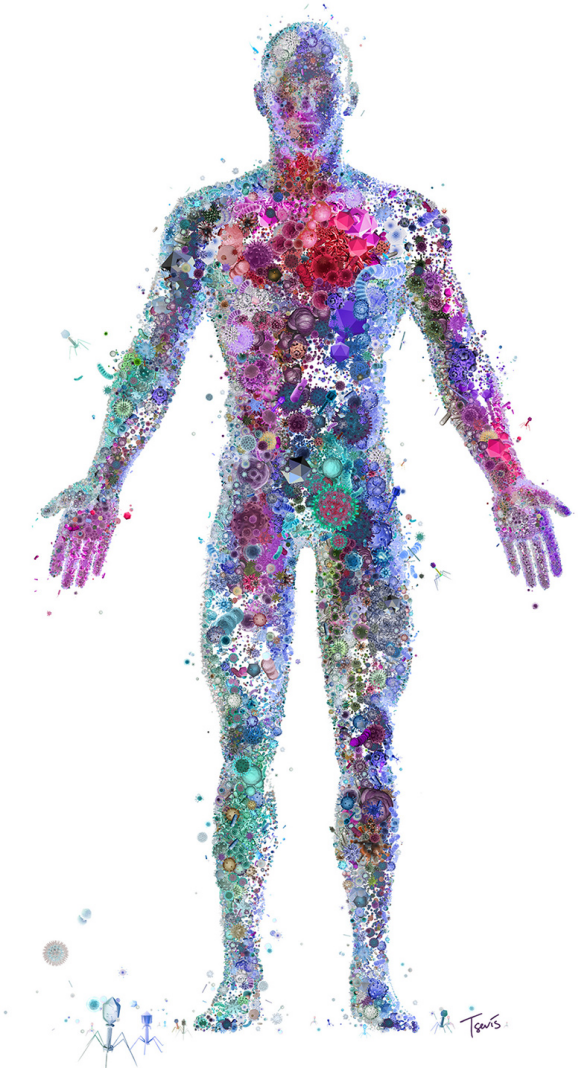
Dr. Mariam Quiñones
Office of Cyber Infrastructure and Computational Biology

Topics

1. Background on importance of microbiome
2. Tools and technologies for studying the microbiome
3. Data analysis workflow for metagenomics
4. Demo on Nephele (amplicon and WGS)

Why study the microbiome?

To understand the role of the microbiome in health and disease



Microbiome

"to signify the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space and have been all but ignored as determinants of health and disease"
- Joshua Lederberg (2001)

The human microbiome is essential to human health

1. Extracting energy from food
2. Producing essential vitamins
3. Modulating our immune system
4. Regulating our glucose levels and metabolism
5. Protecting us against disease causing microbes

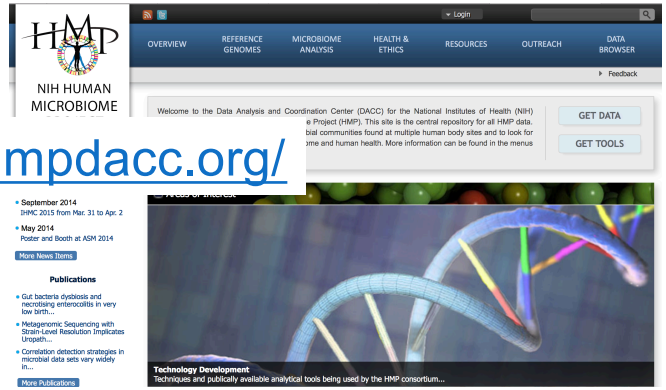
Adapted from https://twitter.com/SeresTX

The Human Microbiome Project

Funded by the NIH Common Fund, FY2007-2016

HMP Phase 1: 2007-2012

- Characterize the human microbiome's role in health and disease
- Create protocols, develop new computational approaches, and create a Data Analysis & Coordinating Center

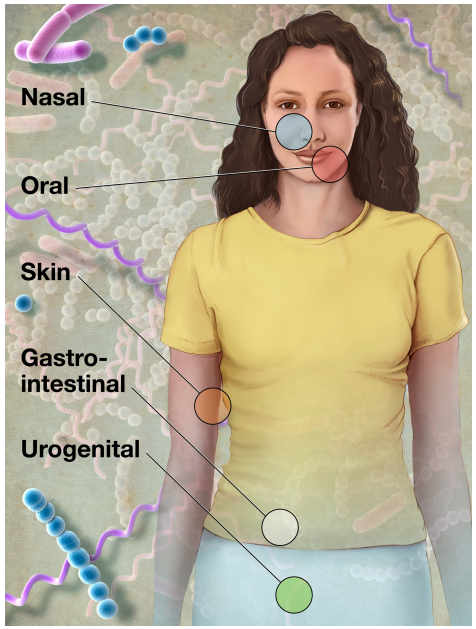


https://hmpdacc.org/

What has the HMP1 found?

1. increased risk of cardiovascular disease due to microbial metabolic byproducts (Koeth et al., 2013)
2. taxonomic composition of the microbiomes between subjects can differ significantly
3. microbiomes of healthy subjects may share similarities in their metabolic pathways

Phase I: 242 “Healthy Human Subjects”



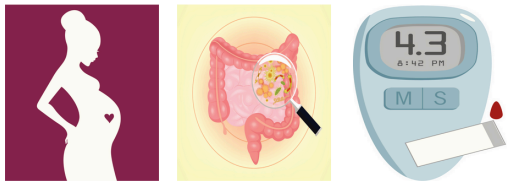
Data available at http://hmpdacc.org

iHMP

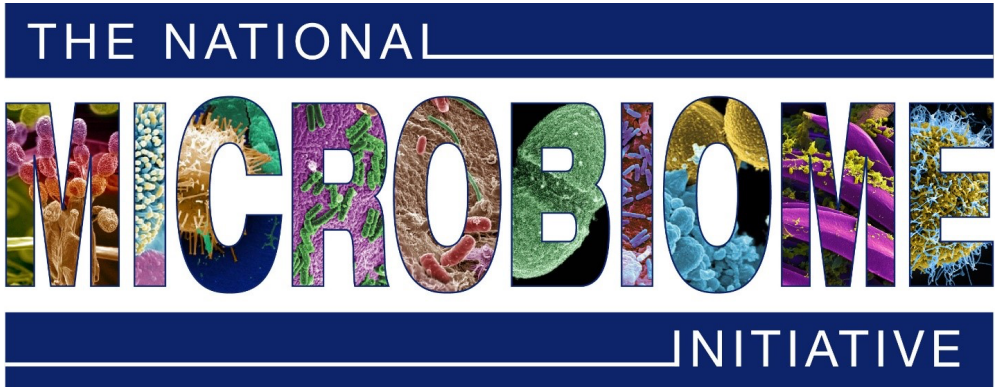
Phase II “Integrative HMP”

Explore biological properties from both the microbiome and host from three different cohort studies of microbiome-associated conditions using multiple "omics" technologies.

- Pregnancy and Pre-term birth
- Inflammatory bowel disease
- Prediabetes



Data available at http://hmp2.org/resources/data_browser.php



Started on
May 13, 2016

The National Microbiome Initiative (NMI) aims to advance understanding of microbiome behavior and enable protection and restoration of healthy microbiome function. In a year-long fact-finding process, scientists from Federal agencies, academia, and the private sector converged on three recommended areas of focus for microbiome science, which are now the goals of the NMI:

- Supporting **interdisciplinary research** to answer fundamental questions about microbiomes in diverse ecosystems.
- **Developing platform technologies** that will generate insights and help share knowledge of microbiomes in diverse ecosystems and enhance access to microbiome data.
- Expanding the microbiome workforce through citizen science, **public** engagement, and **educational** opportunities

How do we study the microbiome?

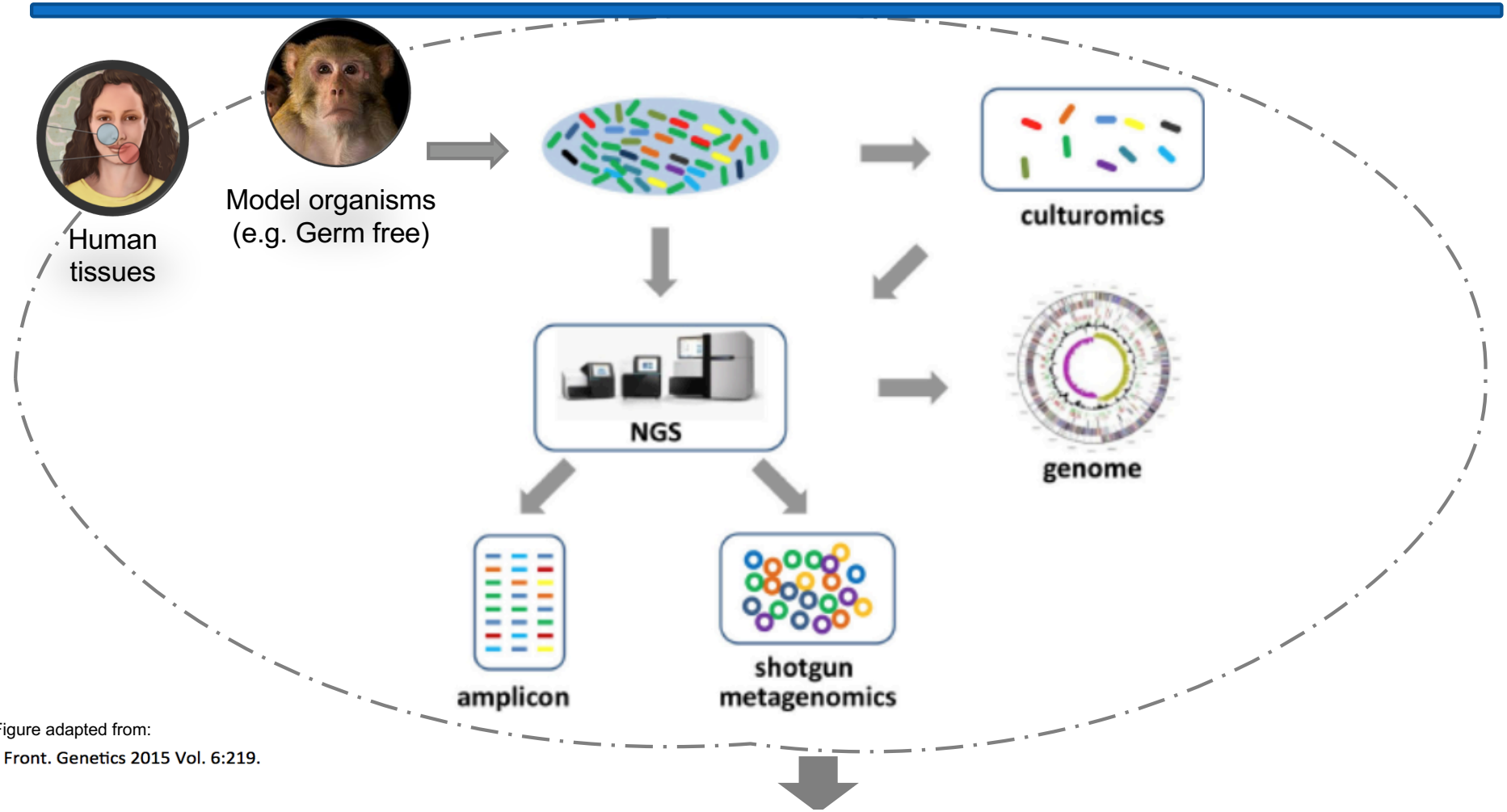


Figure adapted from:
Ji Front. Genetics 2015 Vol. 6:219.

How else could we study the microbiome?

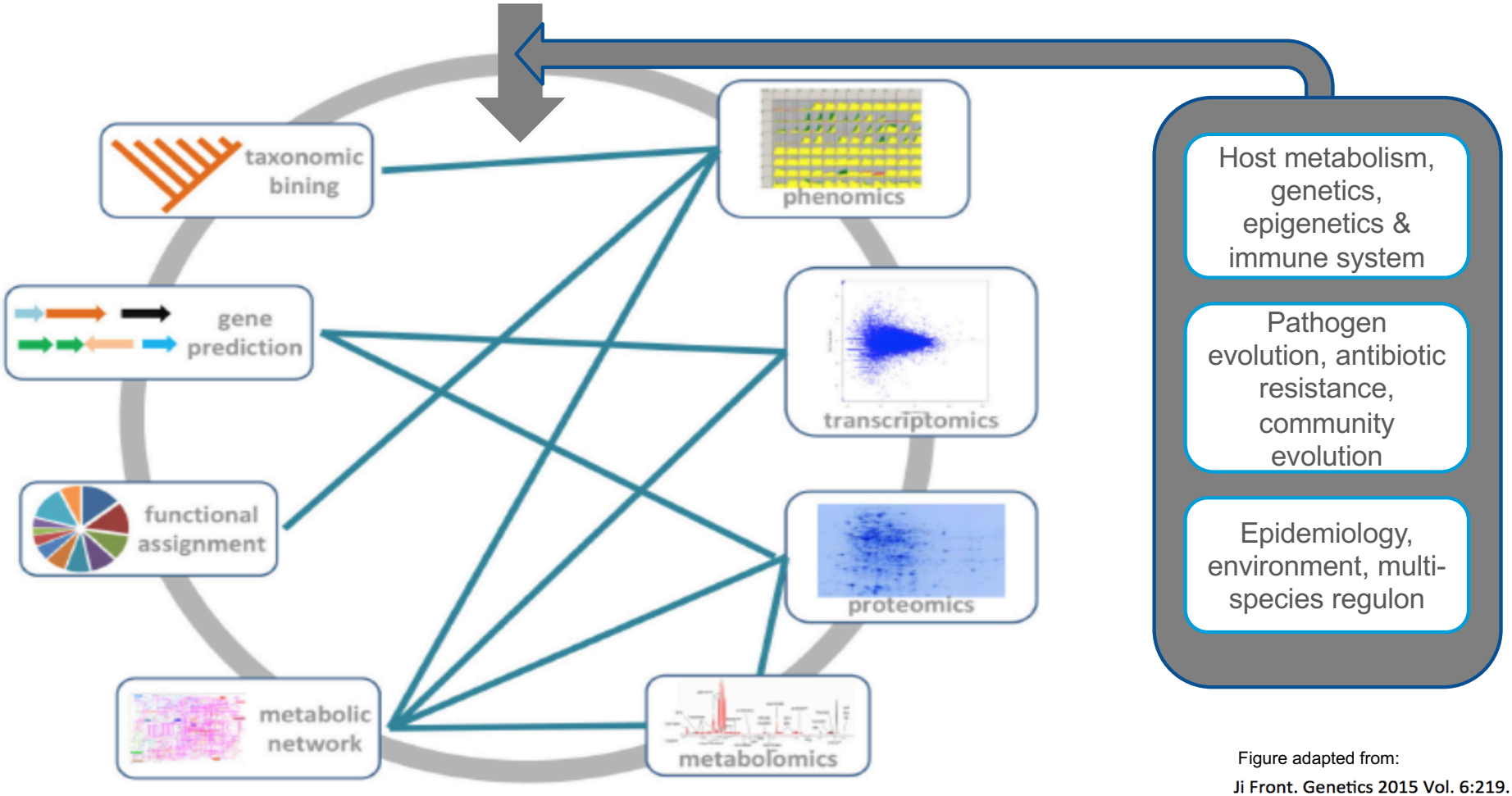


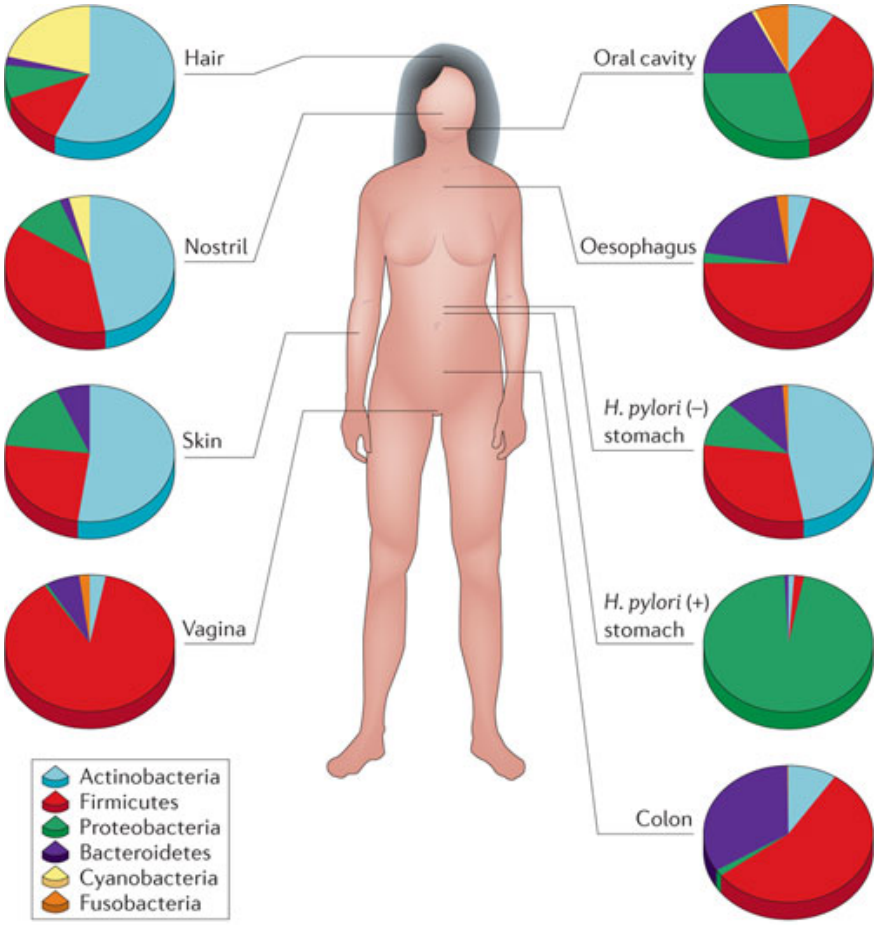
Figure adapted from:

Ji Front. Genetics 2015 Vol. 6:219.

Commonly used OMICs to study the microbiome

Metagenomics

The analysis of the collective microbial genomes in the sample. It usually involves sequencing the variable regions of the rRNA gene or the whole genome to characterize microbial communities.



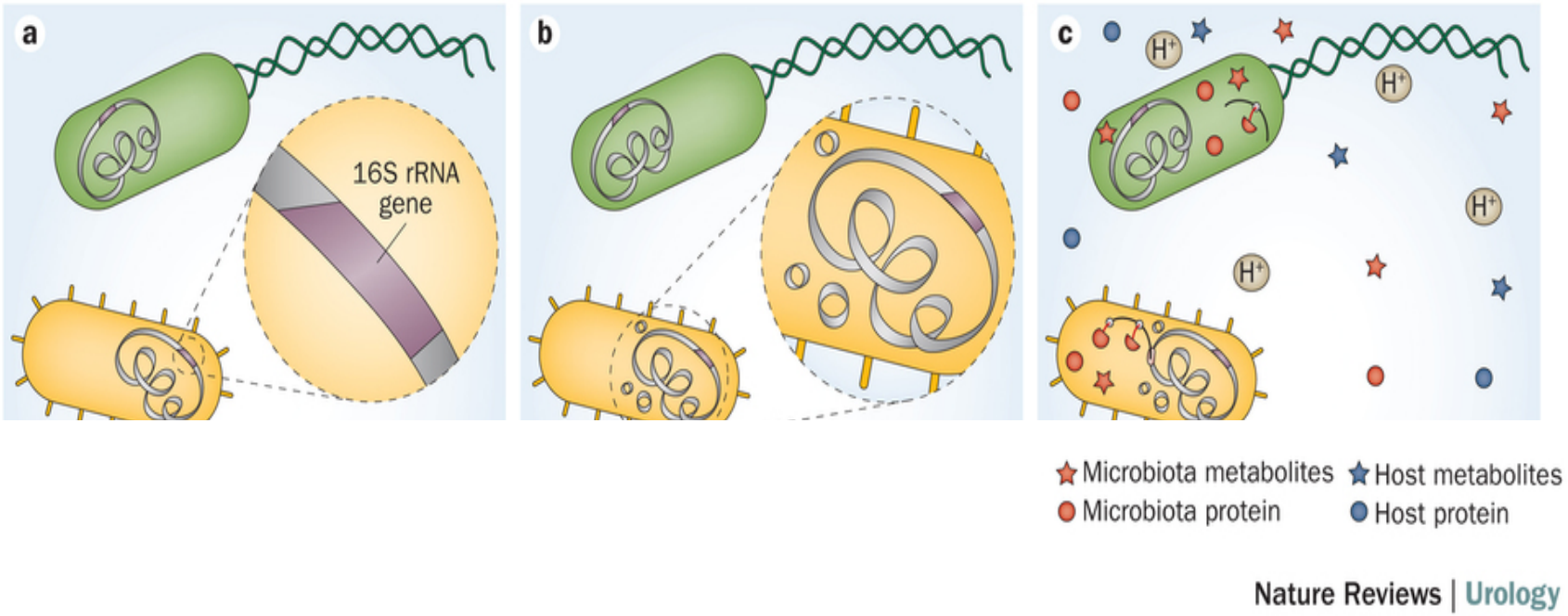
http://www.nature.com/nrg/journal/v13/n4/fig_tab/nrg3182_F1.html

Nature Reviews | Genetics

https://doi.org/10.1186/s40168-015-0094-5

Metagenomics (Definitions)

Microbiota Metagenome Microbiome

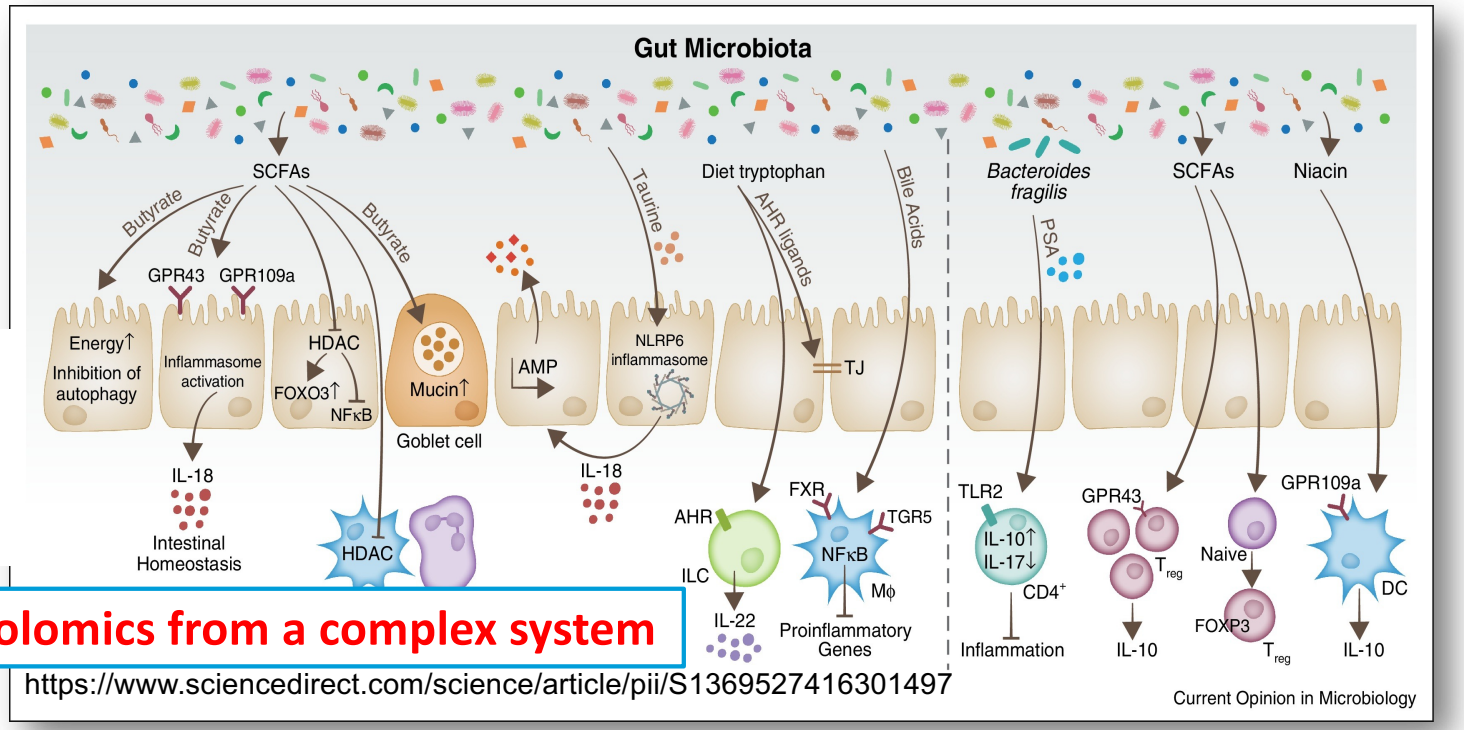


Metabolomics = The analytical approaches used to determine the metabolite profile(s) in any given strain or single tissue. The resulting census of all metabolites present in any given strain or single tissue is called the *metabolome*

Eat fiber → microbes produce Butyrate (anti-inflammatory)

~500 cell types in human body with distinct metabolomes.

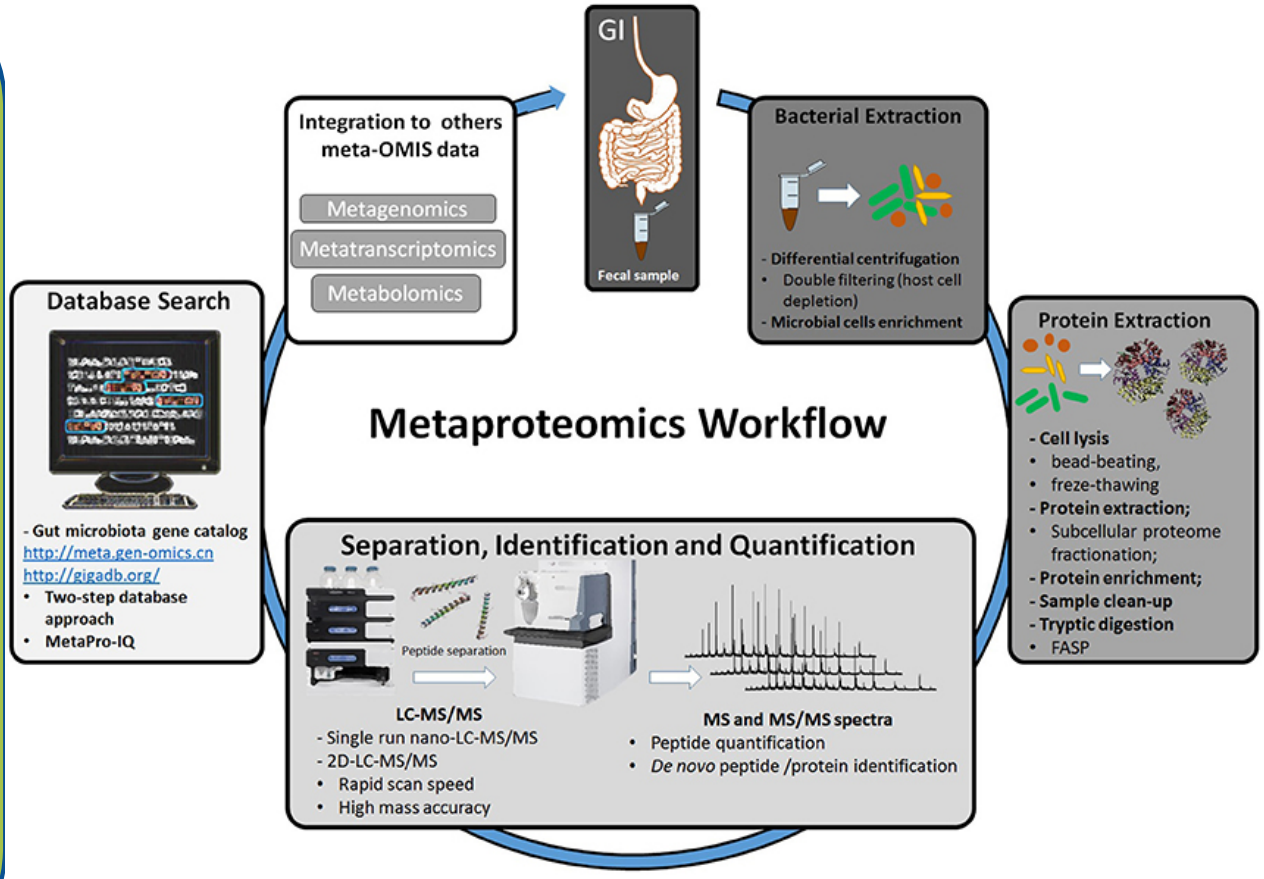
- J. Nicholson



Metabonomics – metabolomics from a complex system

Metaproteomics

The large-scale characterization of the entire protein complement of environmental or clinical samples at a given point in time. The method indiscriminately identifies proteins from the microbiota and the host/environment



https://www.frontiersin.org/articles/10.3389/fchem.2017.00004/full

Metatranscriptomics
The analysis of the suite of expressed RNAs (meta-RNAs).This approach provides information on the regulation and expression profiles of complex microbiomes.

Manuscript example:
Increased virulence of the oral microbiome in oral squamous cell carcinoma revealed by metatranscriptome analyses.
Int J Oral Sci v.10(4); 2018 Dec

“The expression of putative virulence factors in the oral communities associated with OSCC showed that activities related to capsule biosynthesis, flagellum synthesis and assembly, chemotaxis, iron transport, haemolysins and adhesins were upregulated at tumour sites.”

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6232154/

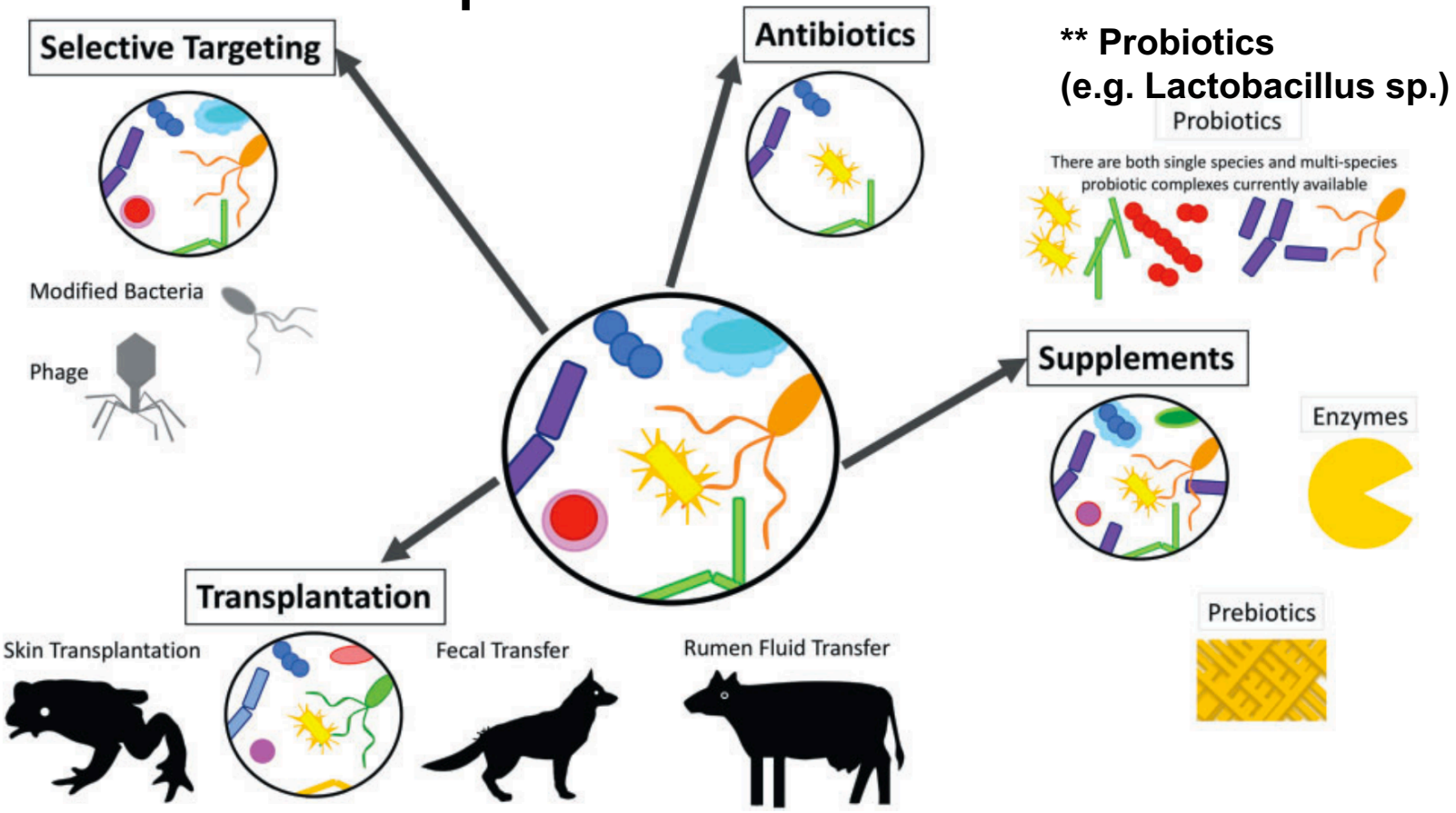
Machine Learning

Gut Microbiota Offers Universal Biomarkers across Ethnicity in Inflammatory Bowel Disease Diagnosis and Infliximab Response Prediction Zhou et. Al, 2018

- Multiple cohorts were used but gut microbial alteration patterns in IBD were similar among Chinese and Westerners.
- **87.5% and 79.1% prediction accuracy in Crohn's disease (CD) and ulcerative colitis (UC) patients respectively**
- Certain microbes, mainly **Clostridiales**, predicted the treatment effectiveness with 86.5% accuracy alone and 93.8% accuracy in combination with calprotectin levels and Crohn's disease activity index (CDAI).
- Random forest classification

What's next?

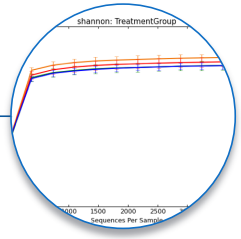
Could we manipulate the microbiome to improve health?



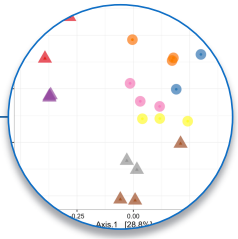
Song SJ, Knight R, et. al 2019. Experimental Biology and Medicine

Let’s now start with Metagenomics Data Analysis

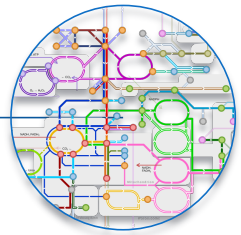
Metagenomics (Questions)



How diverse?



How communities differ?



What are their Functional capabilities?

Bacteria, Viruses, Fungi, Worms…

What is the composition of the total microbiota?

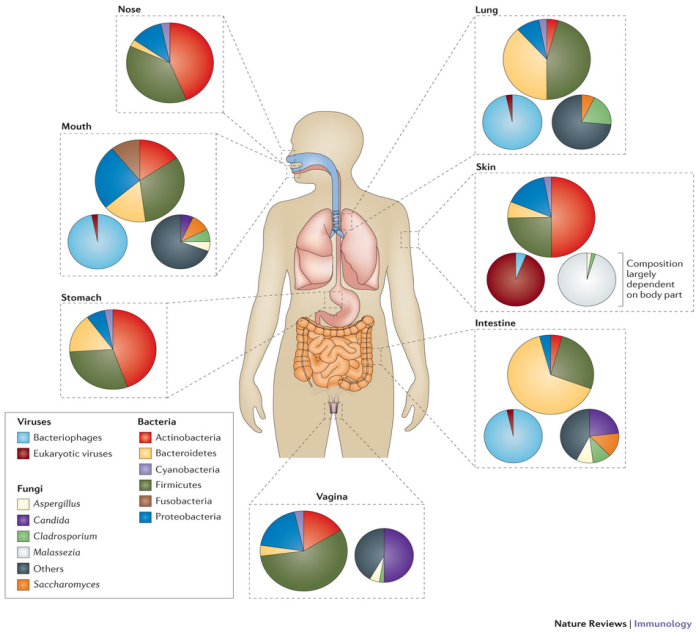
Bacteria : at least 1 microbial cell/human cell.

Viruses : 7 – 10 viral particles/bacterial cell (mostly phages).

Fungi : less than 0.1% of total microbiota; in the skin 70% of all eukaryote OTUs are fungi.

Parasites/Protozoa : Entamoeba and Blastocystis are common in the gut. Overall, amount of protozoa is very low (usually less than fungi) but genome size is large.

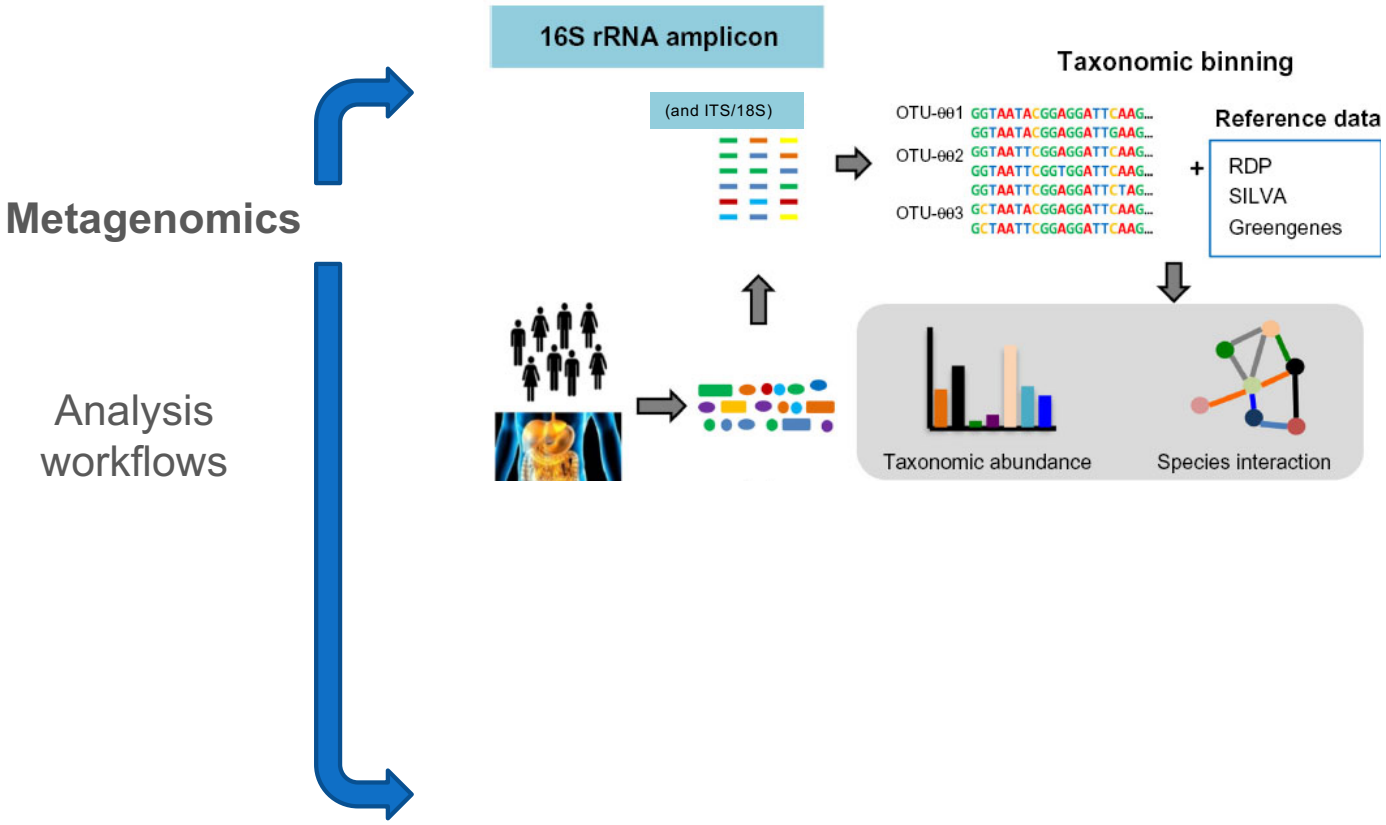
Archea: In gut, archea account for up to 11.5% of total microbiota. In other tissues such as skin, it is not found.



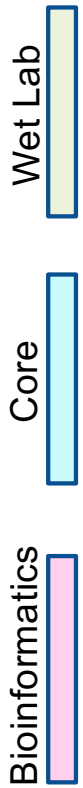
20,000 human genes
2-20 million microbial genes -R. Knight

http://www.nature.com/nri/journal/v14/n12/full/nri3769.html

Metagenomics Data Analysis Workflow



A few factors that influence data analysis workflow

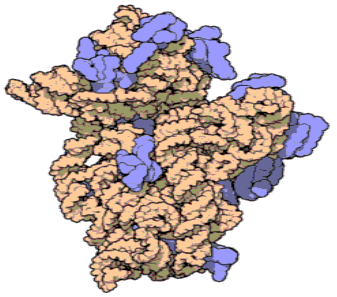


- Good experimental design with enough replicates (e.g. mice per cage and multiple cages)
- Standardization of sample collection and handling
- Select appropriate protocol (e.g. low bacterial biomass)

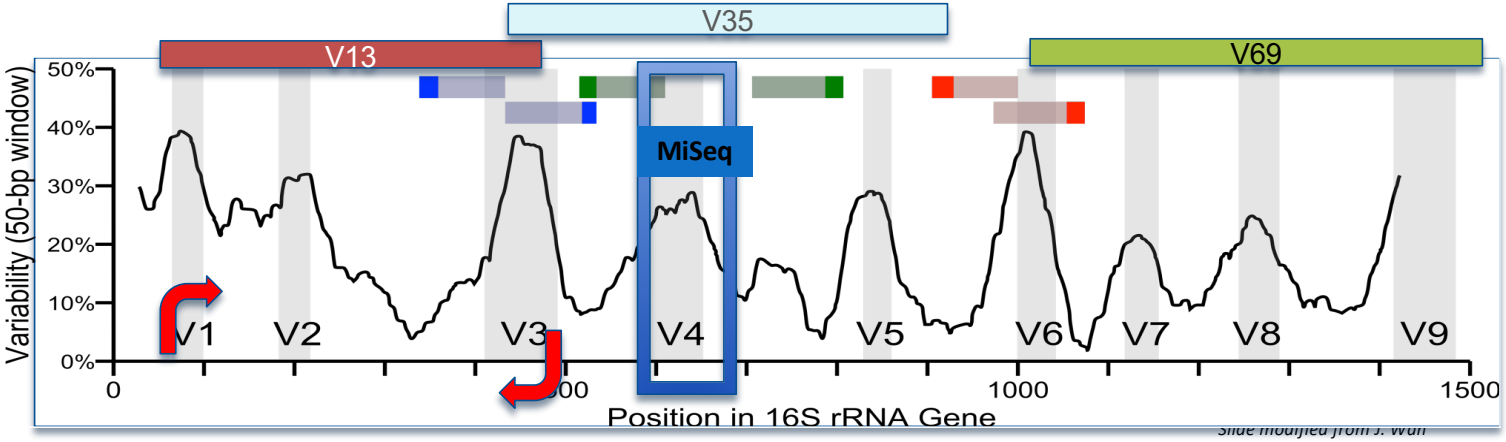
- Choice of sequencing platform and expected error rates
- Amplicon versus Whole Genome sequencing
- Selection of primer pair, barcodes and library preparation
- Use of standards (e.g. mock community) and controls

- Data pre-processing and quality control analysis
- Availability of reproducible pipeline
- Multi-omics data integration methods
- Unification in vocabulary use to describe samples

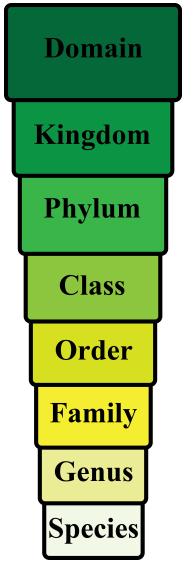
16S rRNA Variable Regions



- Part of the 30S subunit of the prokaryotic ribosome
- Widely conserved (bacteria, archaea)
- 9 hypervariable regions, flanked by conserved sequences
- See http://themicrobiome.com/media/16S_viewer.cfm



Taxonomic classification



King
Phillip
Came
Over
From
Great
Spain

National Institute of Allergy and Infectious Diseases

NIAID

Databases for 16S and fungal 28S,ITS

- www.arb-**silva**.de/
- 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***)
- Alignment is curated

- http://greengenes.secondgenome.com/
- 16S rRNA database and taxonomy
- Large collection with species level taxonomy

- https://rdp.cme.msu.edu
- aligned and annotated Bacterial and Archaeal 16S rRNA sequences,
- Fungal 28S rRNA sequences
- Bayesian classifier
- Classifier for Fungal ITS sequences

16S rRNA survey analysis is not always appropriate

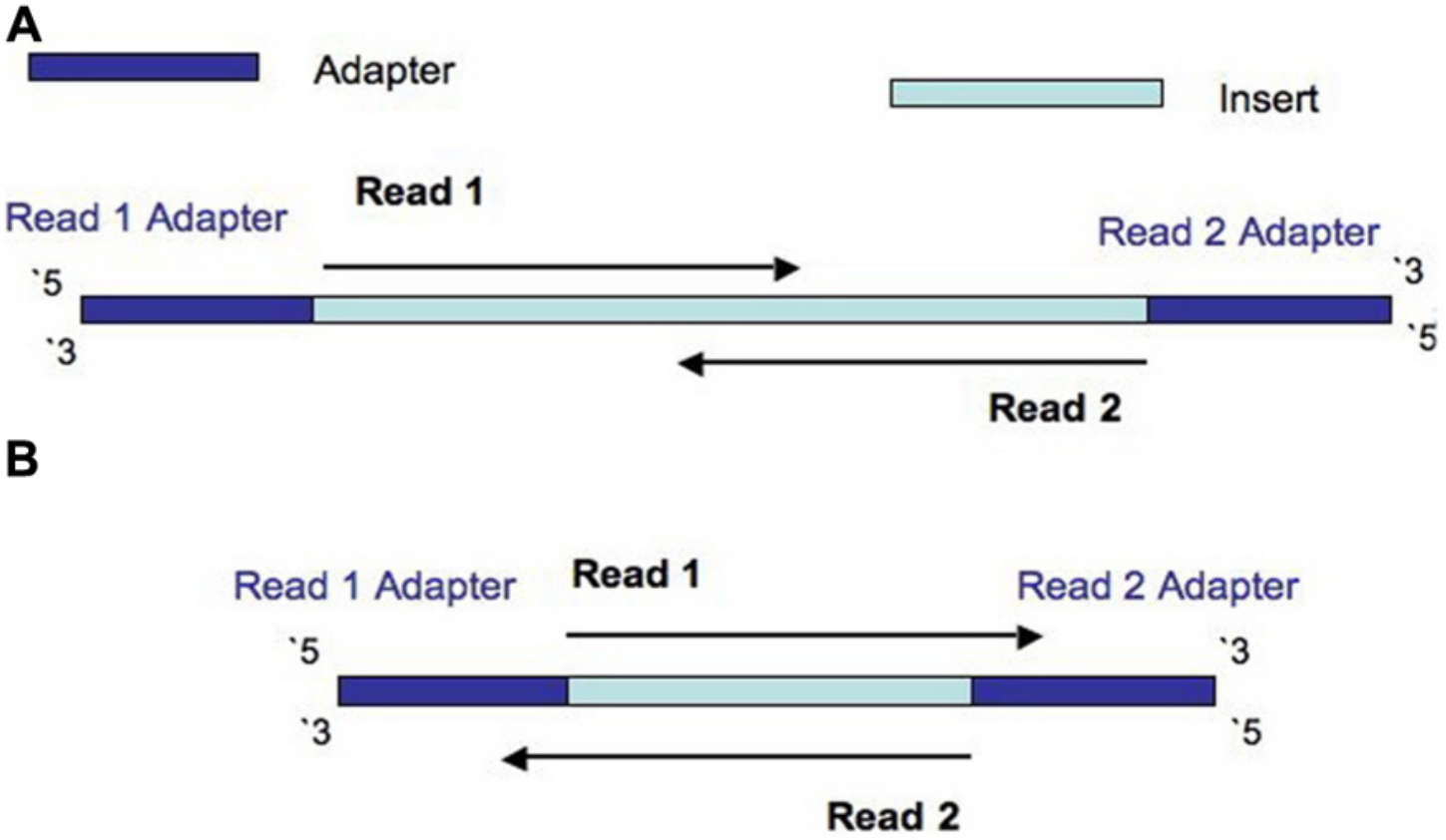
Risks

- Selection of variable region could influence which taxa can be resolved
- Taxonomy assignments need a comprehensive well curated reference database

Potential mitigation approach

- New alternatives (PacBio, minION) for sequencing more of the 16S gene
- Use of alternative databases
- Perform whole genome sequencing

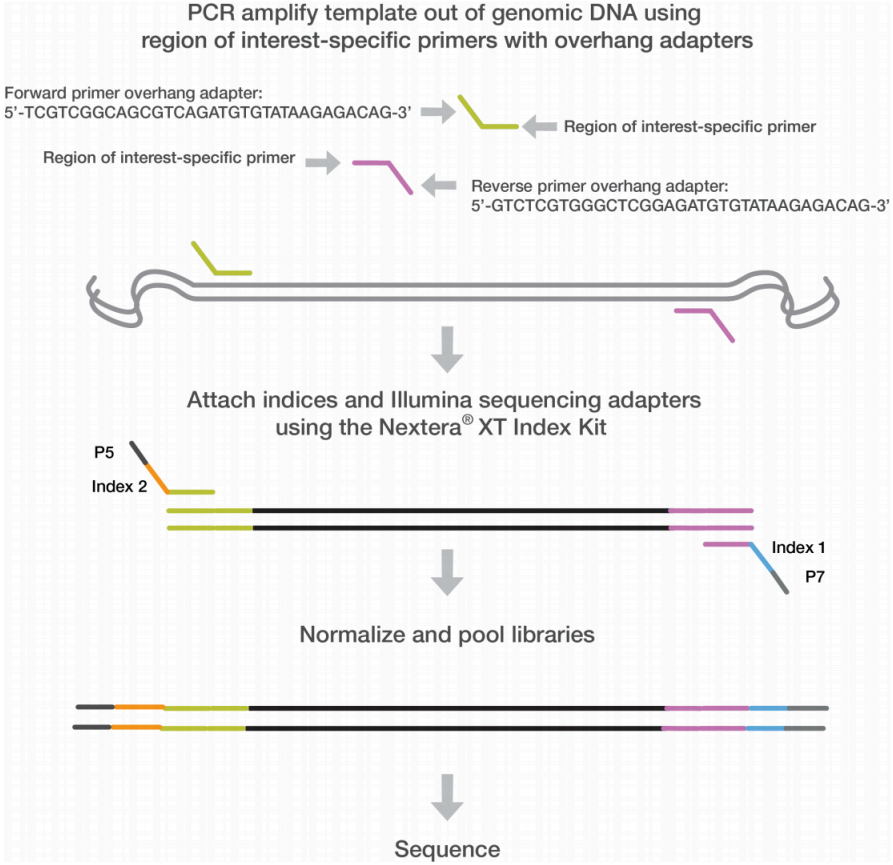
When pairs overlap well .. as in MiSeq V4, processing is easier



These reads are typically joined, a consensus read is created with new quality scores assigned.

Let’s understand primer design

Figure 1 16S V3 and V4 Amplicon Workflow

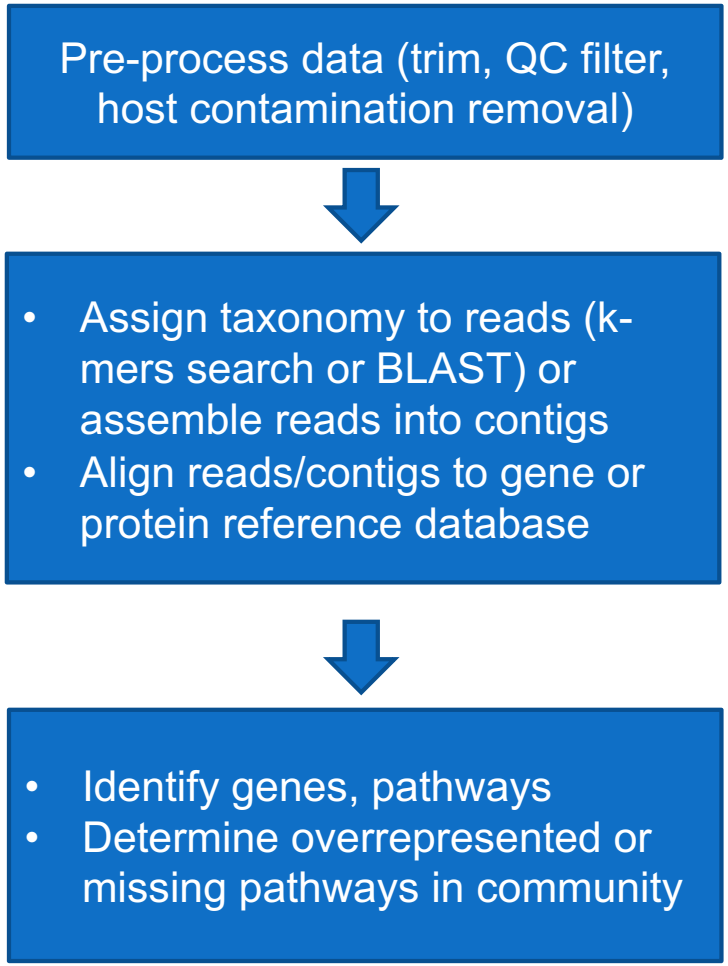


https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf

Typical steps for Pre-processing of data

- Examine quality
  - FASTQC - Common tool for quality assesment
- Remove adapters and Trim low quality ends (not absolutely required)
  - Trimmomatic, TrimGalore, TagCleaner
- Join pairs
  - FLASH, fastq-join, panda-seq, PEAR
- Dereplicate, pre-cluster

Whole genome sequencing data analysis workflow

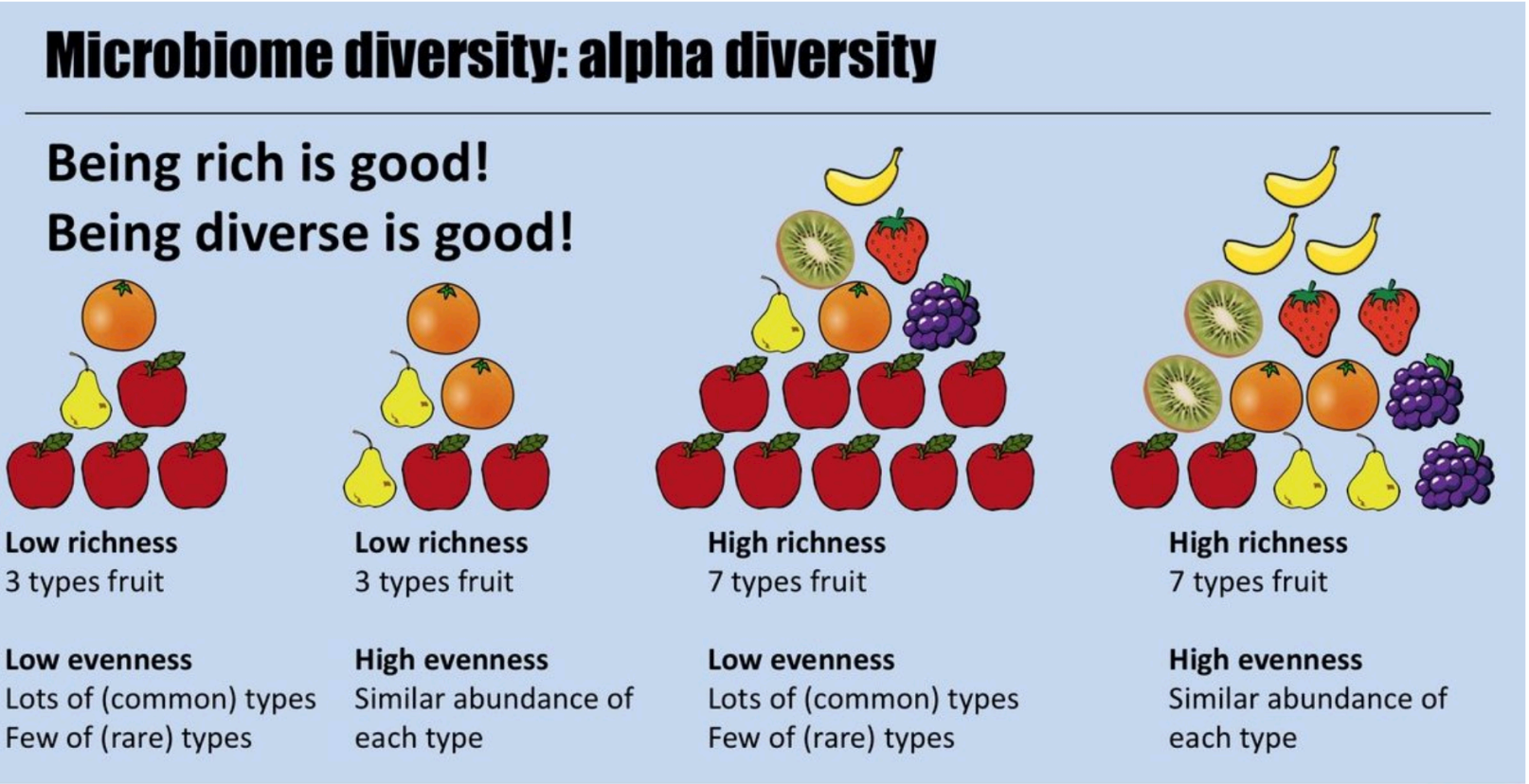


*Database could be of marker genes only

Software/Databases for Microbiome Analysis

- **Nephele (https://nephele.niaid.nih.gov)**
- mothur (mothur.org) - full 16S analysis suite
- DADA2 – 16S analysis
- QIIME & QIIME2 (qiime2.org) - full 16S analysis suite
- Phyloseq – R package for community analysis and visualization starting from biom file
- MG-RAST server (metagenomics.anl.gov) - 16S and WGS pipelines
- Kraken2 https://ccb.jhu.edu/software/kraken2/ - kmer taxonomy classification
- Centrifuge - https://ccb.jhu.edu/software/centrifuge/ - metagenomic classification
- BioBakery (bitbucket.org/biobakery/biobakery) – WGS pipeline & MetaPhlAn
- IMG (img.jgi.doe.gov/imgm_hmp) - DOE Joint Genome Institutes; genome annotation
- PATRIC (patricbrc.org) – genome annotation and comparative genomics
- MEGAN5 (http://ab.inf.uni-tuebingen.de/software/megan5/) – taxonomy from BLAST
- UPARSE - http://drive5.com/usearch/manual/uparse_pipeline.html) – clustering algorithm
- More tools listed @ HMP DACC: http://www.hmpdacc.org/tools_protocols/tools_protocols.php

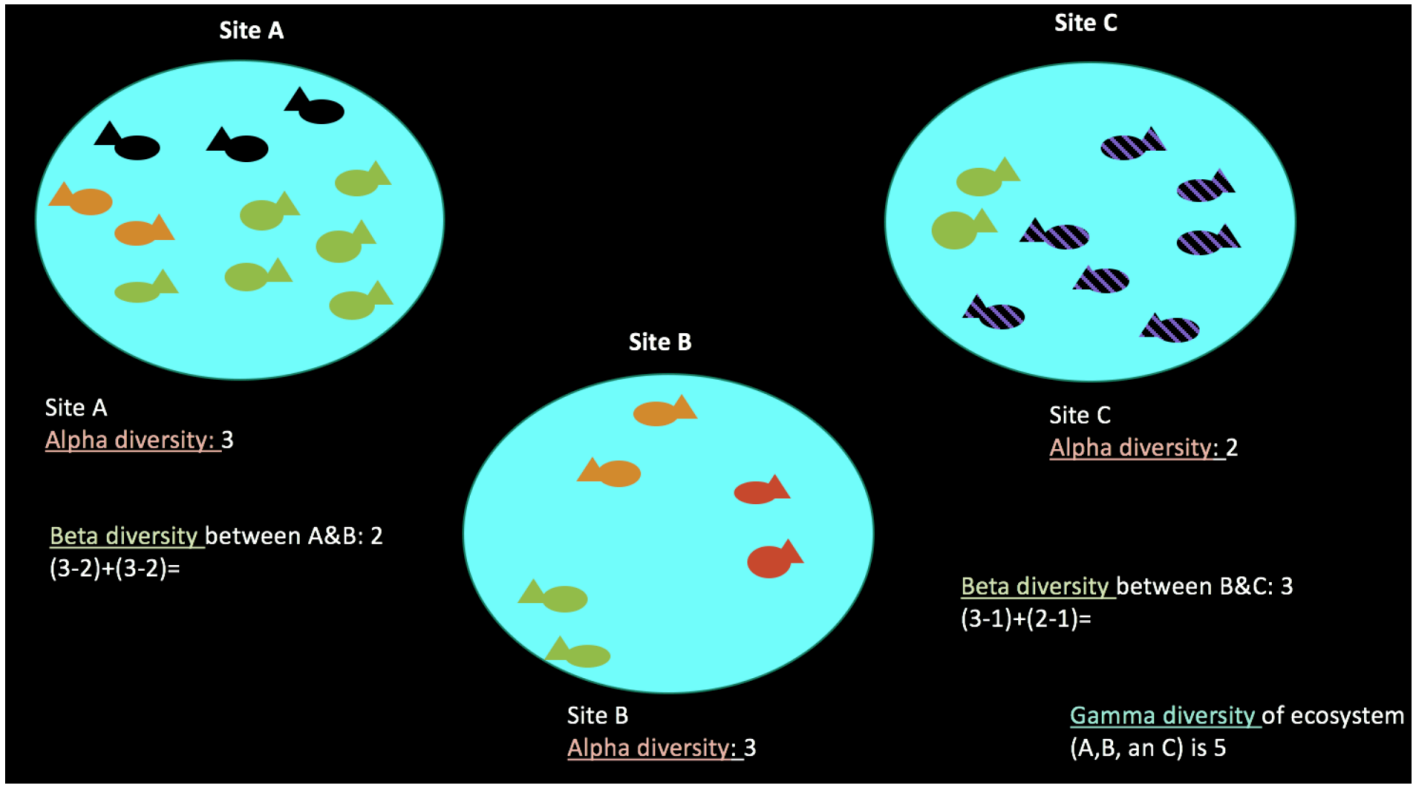
Alpha Diversity: Measures “How many” and/or “How different”



Alpha diversity metrics can look at richness, evenness, or both within a sample. Source: Finotello. Briefings Bioinformatics. 2016. https://awbrooks19.github.io/vmi_microbiome_bootcamp/rst/4_concepts_of_community_analysis.html

Beta Diversity:

Measure of microbiome community similarity structure between samples



Beta diversity calculates how similar two total ecosystems are. Source: Kate M. Socratic.org.

Bray Curtis – measures presence or absence of species and also abundance

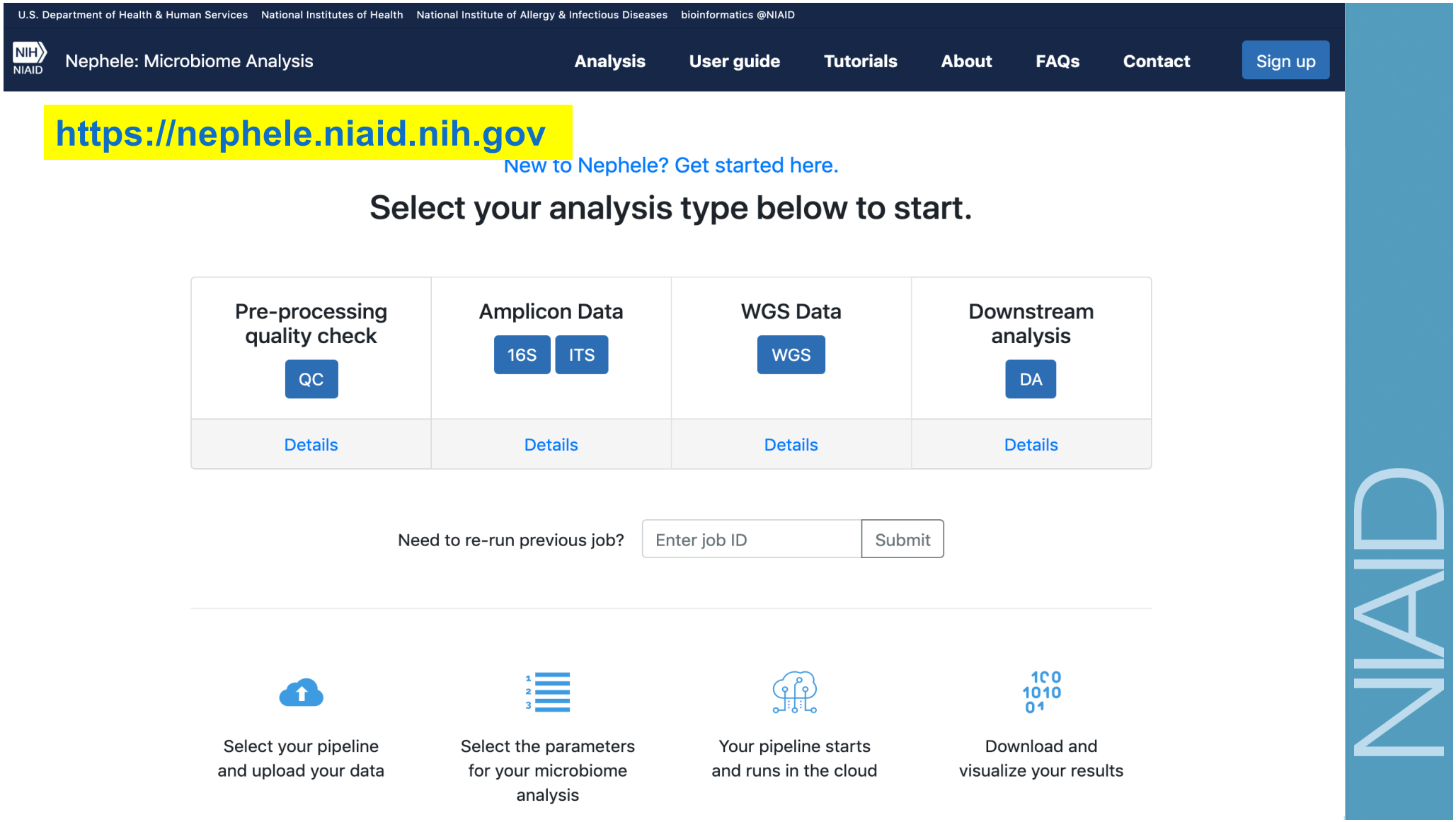
Unifrac – incorporates sequence distances (phylogenetic tree) and calculates fraction of branch length that is shared by two samples - weighted or unweighted

and then do Statistical Analyses

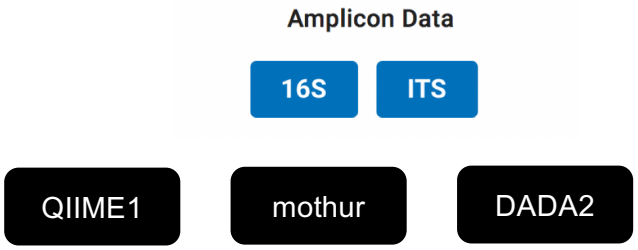
http://cgenome.net/wiki/index.php/Calypso

Method	Description
Network analysis	Correlation network showing co-occurring and mutual exclusive taxa
WGCNA	Weighted correlation network analysis (WGCNA) can be used for finding clusters (modules) of highly correlated taxa, for relating modules to external sample traits (using eigengene network methodology), and for calculating module membership measures.
PCoA, PCA, DCA, NMDS	Unsupervised ordination methods used for data clustering and the identification of outliers
Anosim, PERMDISP2	Supervised univariate methods for identifying significant associations between community composition and a single explanatory variable
RDA, Adonis, CCA	Supervised multivariate method for identifying significant associations between microbial community composition and multiple explanatory variables
Partial least squares regression (PLS)	Multivariate method used to identify taxa associated with multiple explanatory variables
LDA Effect Size (LEfSe)	Identifies features (e.g. genes, pathways, or proteins) characterizing the differences between two or more biological conditions
(Paired) T-test, (nested) Anova, Bayes T-test, Bayes Anova, logistic regression, (paired) Wilcoxon rank test, Kruskal-Wallis test	Identify taxa significantly differentially abundant between sample groups
DESeq2, ANCOM, ALDEx2	Methods specifically developed for counts data. Used for identifying taxa significantly differentially abundant between sample groups.
Multiple linear regression	Identify significant associations between individual taxa and multiple explanatory variables
Support Vector Machine (SVM)	Examine if microbial community composition is predictive of an outcome of interest
Step-wise regression, LASSO regularized regression, random forest	Feature selection methods used for identifying a subset of relevant taxa predictive of an outcome of interest
Shannon index, richness, evenness, Chao 1, ACE, Fisher's Alpha, Simpson index	Quantify microbial alpha diversity
mcpHill	Assess microbiome diversity on multiple indices simultaneously
Jaccard, Bray-Curtis, Yue & Clayton, Chao	Calculate pairwise distances of microbial community profiles
Rarefaction analysis	Estimate coverage of microbial diversity by sequence data
Square root, log, asinh, and centered log ratio transformation; total sum (TSS) and quantile normalization	Data transformation and normalization to render data suitable for analysis by standard statistical procedures

Live demo on Nephele


U.S. Department of Health & Human Services National Institutes of Health National Institute of Allergy & Infectious Diseases bioinformatics @NIAID
NIH NIAID
Nephele: Microbiome Analysis
Analysis
User guide
Tutorials
About
FAQs
Contact
Sign up
https://nephele.niaid.nih.gov
New to Nephele? Get started here.
Select your analysis type below to start.
Pre-processing quality check
QC
Amplicon Data
16S
ITS
WGS Data
WGS
Downstream analysis
DA
Details
Details
Details
Details
Need to re-run previous job?
Enter job ID
Submit
Select your pipeline and upload your data
Select the parameters for your microbiome analysis
Your pipeline starts and runs in the cloud
Download and visualize your results
NIAID

Nephele's Main Data Analysis Pipelines



- ✓ Preprocessing (e.g. join pairs, chimera removal)
- ✓ Taxonomic assignments (SILVA, HOMD, Greengenes)
- ✓ Alpha and Beta Diversity, PCoA
- ✓ Heatmaps, barplots, tables

- ✓ Trim and filter reads
- ✓ Taxonomic assignment with MetaPhlAn2
- ✓ Functional and strain profiling with HUMAnN2 & StrainPhlAn
- ✓ Summary visualizations of species and functional gene abundance
- ✓ PCoA ordination of species composition
- ✓ Plots of functional feature detection vs sequencing depth

https://nephele.niaid.nih.gov/user_guide_pipes/

Additional Pipelines

Pre-processing quality check

QC

- ✓ Input: Mapping file + fastq
- ✓ FastQC quality check
- ✓ Trim primers and/or adapters
- ✓ Trim reads based on quality
- ✓ Filter reads on quality
- ✓ Merge read pairs
- ✓ Summary graph of QC steps

Downstream analysis

DA

- ✓ Input: mapping and biom file
- ✓ Based on QIIME2
- ✓ Outputs a table describing features
- ✓ Diversity core-metrics (alpha / beta)
- ✓ Diversity alpha group significance
- ✓ Taxonomy barplot

https://nephele.niaid.nih.gov/user_guide_pipes/

Required input files:
1- The “mapping” file to describe samples
2- Input read files

Sample ID

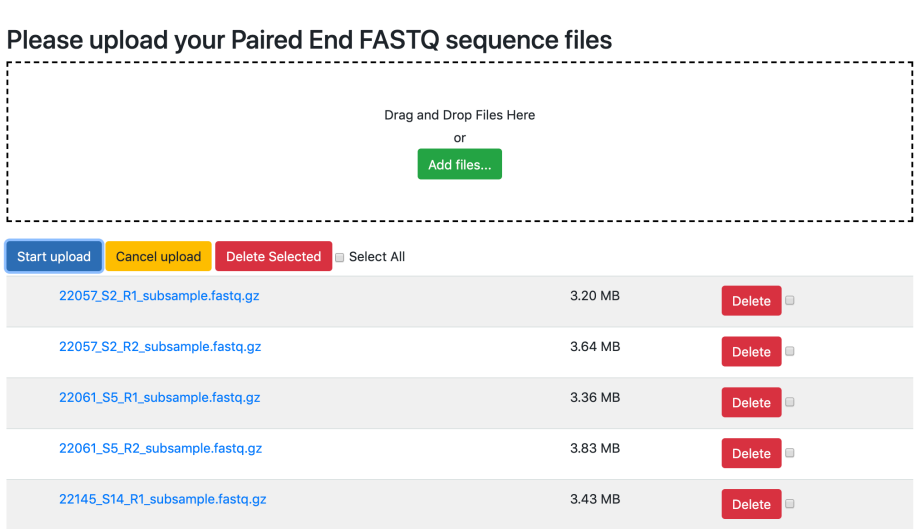
Sequence file names

Metadata is used to describe groups

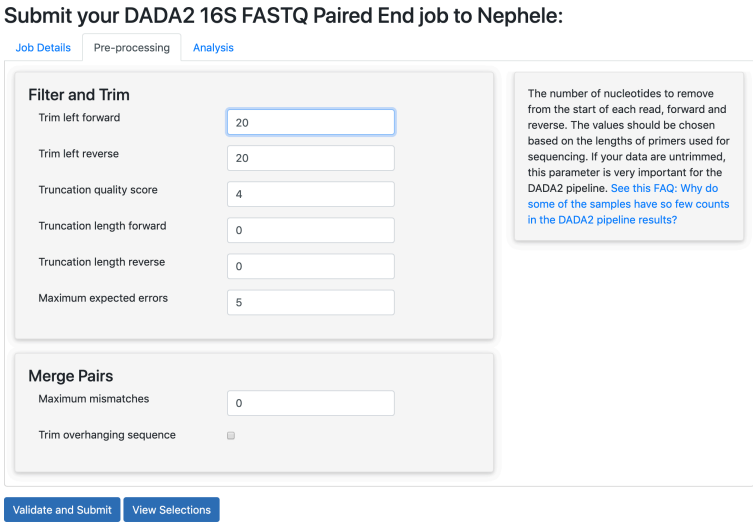
#SampleID	BarcodeSequence	LinkerPrimerSequence	ForwardFastqFile	ReverseFastqFile	TreatmentGroup	OtherDescription	Description
Naive1			Hi1_S17_L001_R1_001.fastq	Hi1_S17_L001_R2_001.fastq	Naive	GroupA	FecalFlush
Naive2			Hi2_S18_L001_R1_001.fastq	Hi2_S18_L001_R2_001.fastq	Naive	GroupA	FecalFlush
Naive3			Hi3_S19_L001_R1_001.fastq	Hi3_S19_L001_R2_001.fastq	Naive	GroupA	FecalFlush
Naive4			Hi4_S20_L001_R1_001.fastq	Hi4_S20_L001_R2_001.fastq	Naive	GroupA	FecalFlush
Naive5			Hi5_S21_L001_R1_001.fastq	Hi5_S21_L001_R2_001.fastq	Naive	GroupA	FecalFlush
Naive6			Hi6_S22_L001_R1_001.fastq	Hi6_S22_L001_R2_001.fastq	Naive	GroupB	FecalFlush
Naive7			Hi7_S23_L001_R1_001.fastq	Hi7_S23_L001_R2_001.fastq	Naive	GroupB	FecalFlush
Naive8			Hi8_S24_L001_R1_001.fastq	Hi8_S24_L001_R2_001.fastq	Naive	GroupB	FecalFlush
Naive9			Hi9_S25_L001_R1_001.fastq	Hi9_S25_L001_R2_001.fastq	Naive	GroupB	FecalFlush
Naive10			Hi10_S26_L001_R1_001.fastq	Hi10_S26_L001_R2_001.fastq	Naive	GroupB	FecalFlush
Res1			Hi11_S27_L001_R1_001.fastq	Hi11_S27_L001_R2_001.fastq	Resveartrol	GroupA	FecalFlush
Res2			Hi12_S28_L001_R1_001.fastq	Hi12_S28_L001_R2_001.fastq	Resveartrol	GroupA	FecalFlush
Res3			Hi13_S29_L001_R1_001.fastq	Hi13_S29_L001_R2_001.fastq	Resveartrol	GroupA	FecalFlush
Res4			Hi14_S30_L001_R1_001.fastq	Hi14_S30_L001_R2_001.fastq	Resveartrol	GroupA	FecalFlush
Res5			Hi15_S31_L001_R1_001.fastq	Hi15_S31_L001_R2_001.fastq	Resveartrol	GroupA	FecalFlush
Res6			Hi16_S32_L001_R1_001.fastq	Hi16_S32_L001_R2_001.fastq	Resveartrol	GroupB	FecalFlush
Res7			Hi17_S33_L001_R1_001.fastq	Hi17_S33_L001_R2_001.fastq	Resveartrol	GroupB	FecalFlush
Res8			Hi18_S34_L001_R1_001.fastq	Hi18_S34_L001_R2_001.fastq	Resveartrol	GroupB	FecalFlush
Res9			Hi19_S35_L001_R1_001.fastq	Hi19_S35_L001_R2_001.fastq	Resveartrol	GroupB	FecalFlush
Res10			Hi20_S36_L001_R1_001.fastq	Hi20_S36_L001_R2_001.fastq	Resveartrol	GroupB	FecalFlush

Simple to use interface

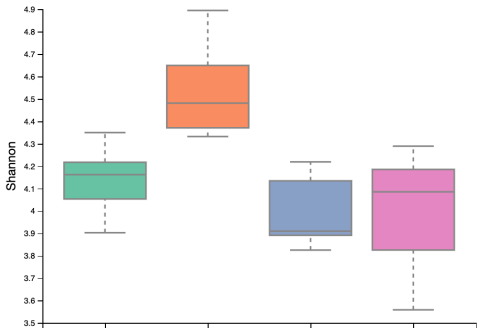
File Uploads

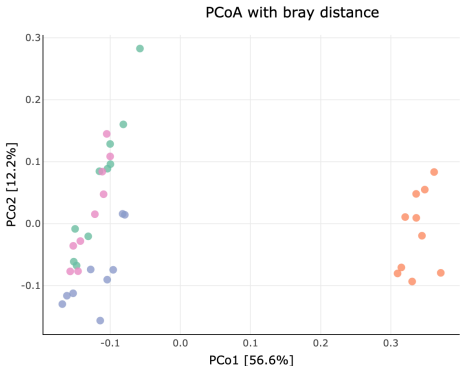


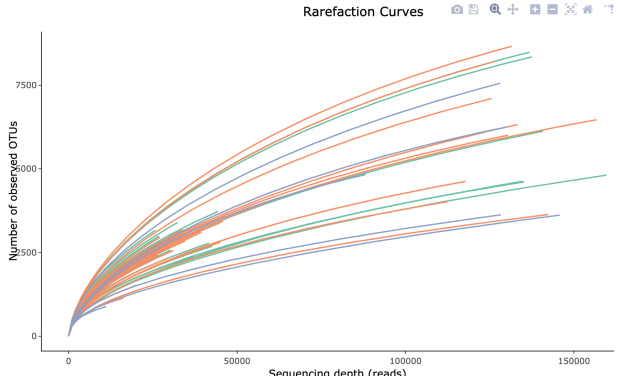
Parameter Selection



Nephele's interactive outputs

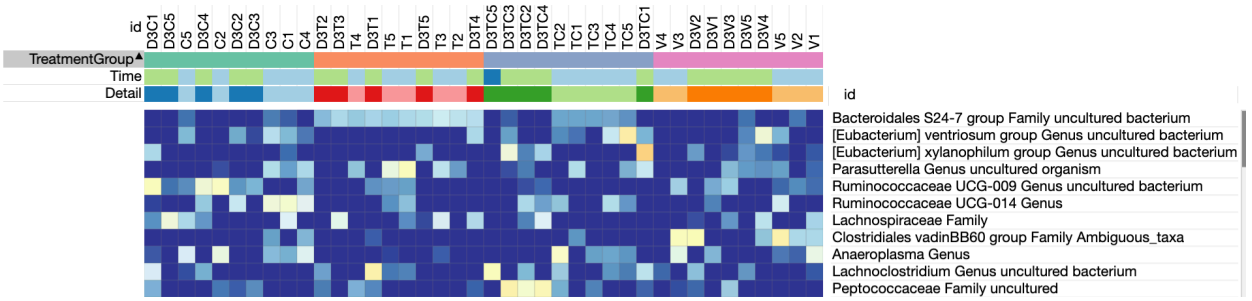




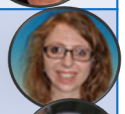
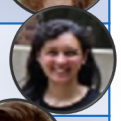


Num samples: 40
Num observations: 12,847
Total count: 5,173,018
Table density (fraction of non-zero values): 0.137

Counts/sample summary:
Min: 4,524.000
Max: 346,487.000
Median: 114,578.500
Mean: 129,325.450
Std. dev.: 76,361.883
Sample Metadata Categories: None provided
Observation Metadata Categories: taxonomy



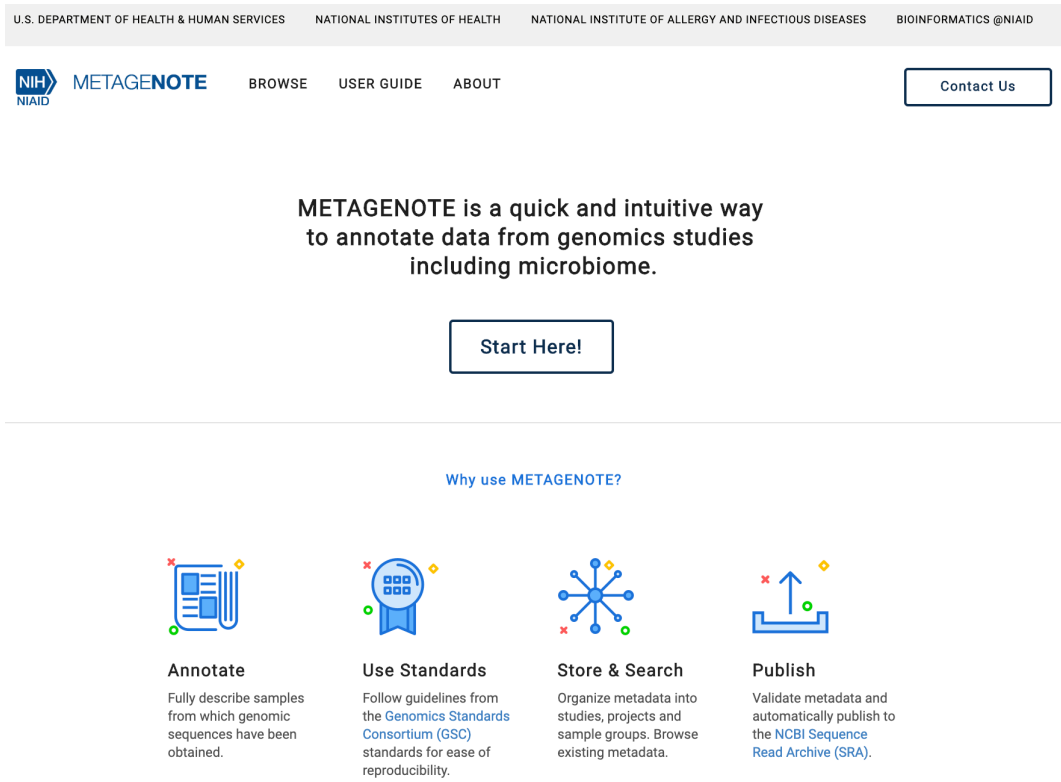
Nephele: Recent Citations

Publications	Year	User
The Impact of Anthelmintic Treatment on Human Gut Microbiota Based on Cross-Sectional and Pre- and Postdeworming Comparisons in Western Kenya. https://www.ncbi.nlm.nih.gov/pubmed/31015324	2019	Alice Easton (NIAID/NIH) 
Novel Cardinium strains in non-marine ostracod (Crustacea) hosts from natural populations https://www.sciencedirect.com/science/article/pii/S1055790318303518	2019	Isa Schön (Brussels, Belgium) 
Multi-Method Characterization of the Human Circulating Microbiome https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6345098/	2019	Emma Whittle (Keele University, UK) 
Experimental Microbial Dysbiosis Does Not Promote Disease Progression in SIV-Infected Macaques https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6129204/	2019	Alexandra M. Ortiz (NIAID/NIH) 
Diet composition and gut microbiome of 0-group European plaice Pleuronectes platessa L. - Strong homogeneity and subtle spatial and temporal differences https://www.sciencedirect.com/science/article/pii/S1385110117303660	2019	Franz M.Heindler (Leuven, Belgium) 
Experimentally induced metamorphosis in highly regenerative axolotl (ambystoma mexicanum) under constant diet restructures microbiota. https://www.ncbi.nlm.nih.gov/pubmed/30030457	2018	Yıldırım Süleyman (Istanbul, Turkey) 
Composition of the intestinal microbiota in extended-spectrum β-lactamase-producing Enterobacteriaceae carriers and non-carriers in Thailand https://www.sciencedirect.com/science/article/pii/S0924857918303686?via%3Dihub#!	2018	Pipat Piewngam (Bangkok, Thailand & NIH) 
Analysis of oral bacterial communities: comparison of HOMI*NGS* with a tree-based approach implemented in QIIME https://www.tandfonline.com/doi/full/10.1080/20002297.2019.1586413	2018	Robert J. Palmer (NIDCR/NIH)

Metadata annotation in metagenomics is important

learn about METAGENOTE

https://metagenote.niaid.nih.gov



National Institute of Allergy and Infectious Diseases

NIAID

Thank you

Bioinformatics and Computational Biosciences Branch
Office of Cyber Infrastructure and Computational Biology

National Institute of Allergy and Infectious Diseases

NIAID

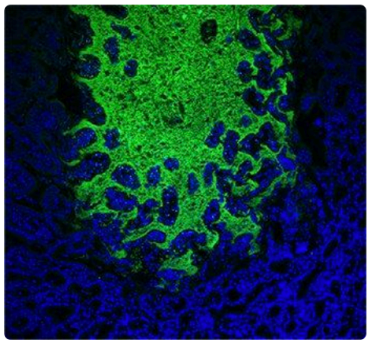
Other NIAID resources for microbiome

NIAID
Microbiome Program

Dr. Yasmine Belkaid

NIAID Microbiome Program

The NIAID Microbiome Program is a collaborative effort to explore the metaorganism using existing and new facilities and personnel. The program has developed a microbiome sequencing facility with bioinformatics support and a gnotobiotic mouse facility, and is currently expanding microbiology resources. The program has supported over 35 groups across the National Institutes of Health (NIH) in their research related to the exploration of the metaorganism.



Visualization and localization of commensal bacteria by Fluorescence in situ Hybridization (FISH) in the small intestine of Toxoplasma gondii-infected mice.

Credit: NIAID

Contact Information

- Xing He

Main Areas of Focus

- To promote the exploration of the metaorganism in order to address questions of relevance to NIAID's mission in both experimental and clinical settings
- To address research topics such as:
  - The role of microbiota in the development and function of the immune system
  - How changes to the microbiota promote the emergence or speed of multidrug-resistant pathogens
  - How dysregulation of host-microbiota interactions leads to chronic inflammatory diseases
  - How pathogens, their expressed genes, and their metabolites directly affect their own proliferation and/or immune signaling networks to exacerbate disease

Services

- Sequencing
- Gnotobiotic Animal Facility
- Microbiology

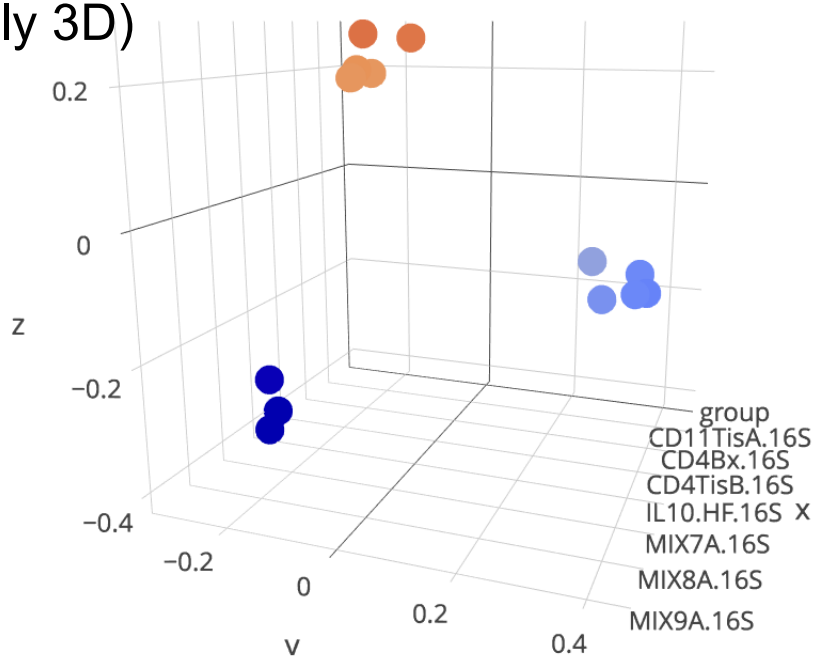
Germ-free animal models are good for studying the impact of one or a few microbes

1. Germ-free (free of all microbes)
2. Gnobiotic (defined flora)
3. Specific pathogen free (free of specific pathogens)

NIAID's gnotobiotic mice facility

A few web resources for displaying data from this demo

Phinch http://phinch.org/ (taxonomy) – now a desktop app
Treeview http://etetoolkit.org/treeview/ (trees)
Plot.ly https://plot.ly/ (any plot, specially 3D)



National Institute of Allergy and Infectious Diseases

NIAID