

CZI-CABANA: Microbiome Bioinformatics with qiime2 Workshop

Online, for the first time!
5-9 October 2020

<https://bit.ly/2HThBcx>

Before we get started, confirm that you have installed the required software for this workshop. See the *Required Software* section of [this page](#).

Then, please review the [QIIME 2 Forum Code of Conduct](#), which we adopt for this workshop.

And if you still have time, check out these QIIME 2 resources:

- User documentation: <https://docs.qiime2.org>
- Technical support (register for a free account if you don't already have one): <https://forum.qiime2.org>
- Follow us on Twitter for announcements: [@qiime2](https://twitter.com/qiime2)
- Developer resources: <https://dev.qiime2.org>
- Source code on GitHub: <https://github.com/qiime2>

<u>Instructor</u>	<u>QIIME 2 Forum</u>	<u>Institutional affiliation</u>
Aeriel Belk	aeriel.belk	Department of Animal Sciences, Colorado State University
Alex Emmons	emmo1	Department of Animal Sciences, Colorado State University
Andrew Sanchez	andrewsanchez	Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, USA
Anthony Simard	Oddant1	Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, USA
Ben Kaehler	BenKaehler	School of Science, University of New South Wales, Canberra, Australia
Chloe Herman	cherman2	Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, USA
Chris Keefe	ChrisKeefe	Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, USA
Emily Borsom	emilyborsom	Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, USA
Emily Cope	Emily_Cope	Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, USA
Evan Bolyen	ebolyen	Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, USA
Greg Caporaso	gregcaporaso	Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, USA
Heather Deel	hdeel	Department of Animal Sciences, Cell and Molecular Biology Special Academic Unit, Colorado State University
Jamie Morton	mortonjt	Center for Computational Biology, Flatiron Institute, Simons Foundation, New York, USA
Jessica Metcalf	jessicalmetcalf	Department of Animal Sciences, Colorado State University
Justine Debelius	jwdebelius	Centre for Translational Microbiome Research, Department of Microbiology, Tumor, and Cancer Biology, Karolinska Institutet, Stockholm, Sweden
Matthew Dillon	thermokarst	Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, USA
Mehrbod Estaki	Mehrbod_Estaki	Department of Pediatrics, University of California San Diego, USA
Mike Robeson	SoilRotifer	Department of Biomedical Informatics, College of Medicine, University of Arkansas for Medical Sciences, Little Rock AR, USA
Nick Bokulich	nicholas_bokulich	Laboratory of Food Systems Biotechnology, Institute of Food, Nutrition, and Health, ETH Zürich, Switzerland
Renato Oliveira	reinator	Environmental Genomics, Instituto Tecnológico Vale, Belém, Pará, Brazil
Yoshiki Vazquez-Baeza	yoshiki	Center for Microbiome Innovation, Jacobs School of Engineering, University of California San Diego, USA

Thanks to our funders...

QIIME 2 project funding

National Cancer Institute: ITCR ([1U24CA248454-01](#))

National Science Foundation ([1565100](#))

Chan-Zuckerberg Initiative

Other funding sources

National Cancer Institute: [Partnership for Native American Cancer Prevention](#) (U54CA143925;

Caporaso Lab)

Alfred P. Sloan Foundation (Caporaso Lab)

RCUK (BB/P027849/1; CABANA)

ERC Horizon 2020 825410 (Justine Debelius)

and to our hosts...

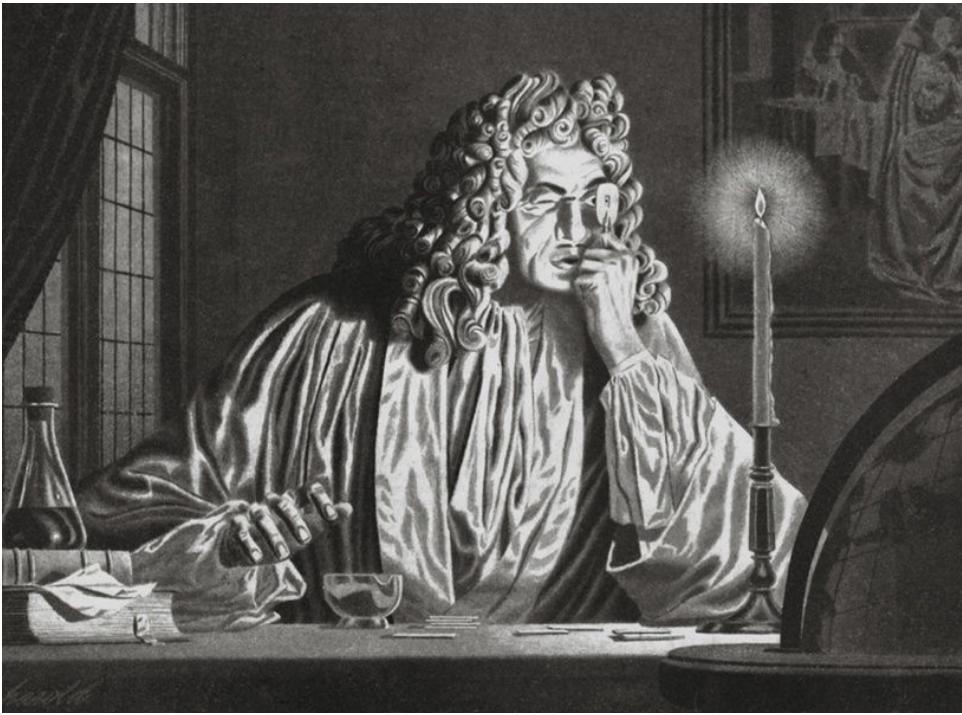
Chan-Zuckerberg Initiative

CABANA

Cath Brooksbank

Guilherme Oliveira

Piraveen Gopalasingam

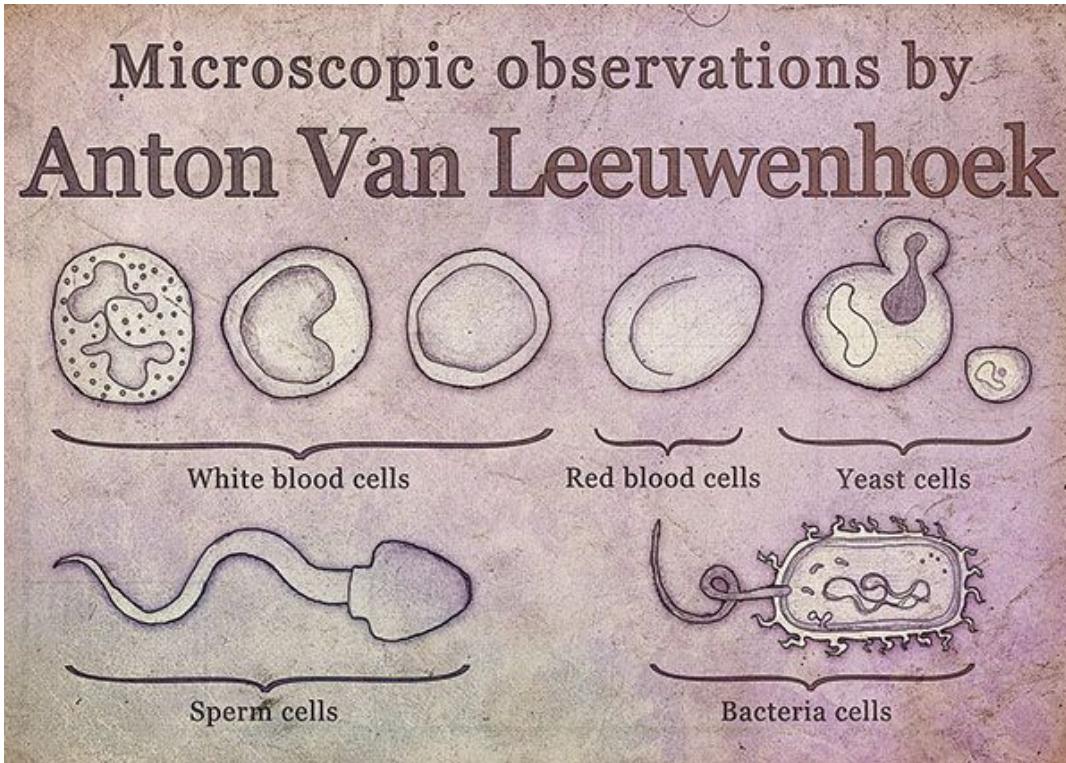


200x magnification and the discovery of
“animalcules”, ~1671 by Antonie van Leeuwenhoek

Image source: Discover Magazine (2015)

van Leeuwenhoek's "animalcules" came to be known as microorganisms or microbes.

"Protozoans"
(single-celled eukaryotes)



Sketches by Taszyn Bailey



Cyanobacteria (i.e., blue-green algae)
and a rotifer (another single-celled
eukaryote)



Images by Lesley Robertson, Delft's Science Centre



200x magnification and the discovery of “animalcules”,
~1671 by van Leeuwenhoek

Image source: Discover Magazine (2015)



Bacillus anthracis in monoculture.

Image source: U.S. Army Medical Research Institute of Infectious Diseases



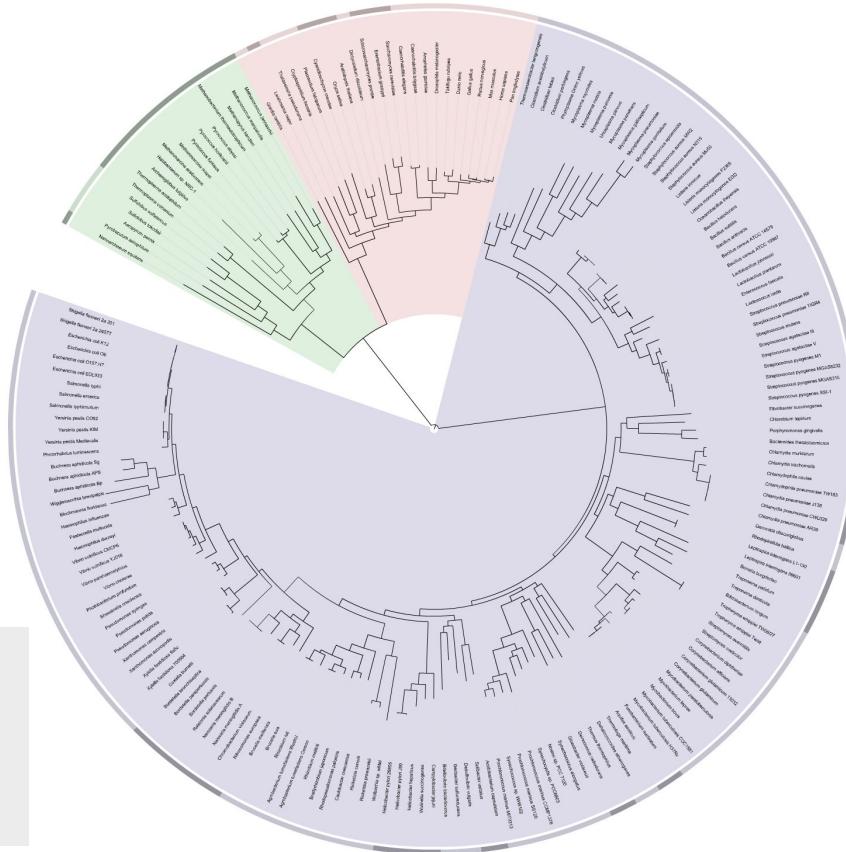
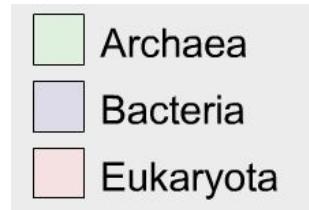
Winogradsky columns
model microbial ecosystems.

Image source: www.hhmi.org/bioInteractive

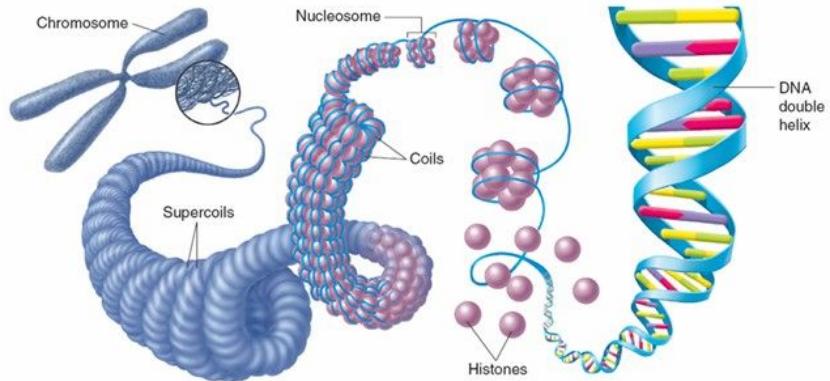
Culture-independent investigation of microbial communities

All cellular life has a shared evolutionary history, and some genes are shared by all organisms.

The sequence of those genes can be used as a *genetic fingerprint* for different organisms.



DNA sequencing



Time



ACCAGGTT

The random accumulation of *mutations* (changes to gene sequences over evolutionary time) gives us information for identifying and comparing organisms.

Time



ACCAGGTT

ACCAGGTT

ACCATGTT

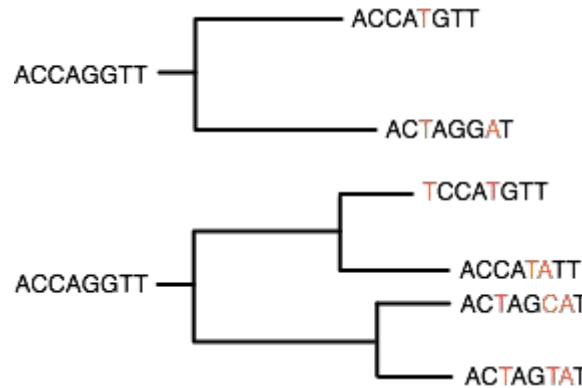
ACTAGGAT

The random accumulation of *mutations* (changes to gene sequences over evolutionary time) gives us information for identifying and comparing organisms.

Time



ACCAAGGTT



The random accumulation of *mutations* (changes to gene sequences over evolutionary time) gives us information for identifying and comparing organisms.

Time



ACCAAGGTT

ACCAAGGTT

```
graph LR; Root[ACCAAGGTT] --> A1[ACCATGTT]; Root --> A2[ACTAGGAT]
```

ACCAAGGTT

```
graph LR; Root[ACCAAGGTT] --> B1[TCCATGTT]; Root --> B2[ACCATATT]; Root --> B3[ACTAGCAT]
```

ACCAAGGTT

```
graph LR; Root[ACCAAGGTT] --> C1[TCAATGTT]; Root --> C2[TCCATGTT]; Root --> C3[ACCATATT]; Root --> C4[ACTAGCAT]
```

Escherichia
Desulfovibrio
Thermus
Thermoplasma
Haloferax

The random accumulation of *mutations* (changes to gene sequences over evolutionary time) gives us information for identifying and comparing organisms.

The *small subunit ribosomal RNA gene* is frequently used to “fingerprint” different microbial organisms.

Why this gene?

- It's ubiquitous.
- Contains regions that are identical across diverse organisms, and regions that are variable across organisms.

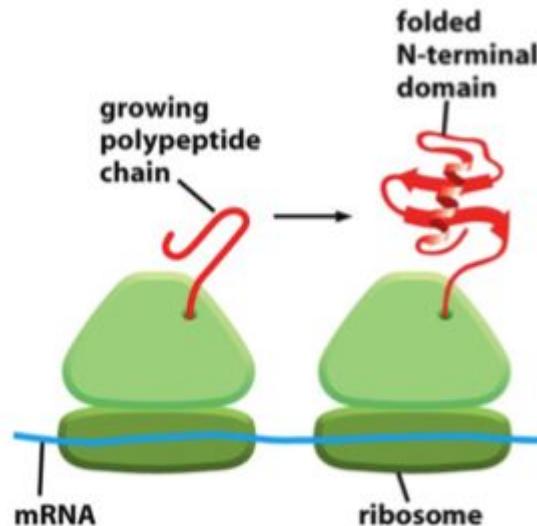


Figure 6-84 Molecular Biology of the Cell (© Garland Science 2006)



200x magnification and the discovery of “animalcules”, ~1671 by van Leeuwenhoek

Image source: Discover Magazine (2015)



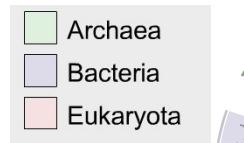
Bacillus anthracis in monoculture.

Image source: U.S. Army Medical Research Institute of Infectious Diseases

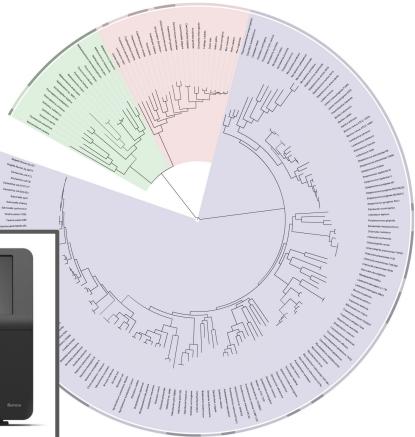


Winogradsky columns model microbial ecosystems.

Image source: www.hhmi.org/bioInteractive



Sequencing of ribosomal RNA and marker genes.



We live in a microbial world.

Coprinus comatus



Zea mays



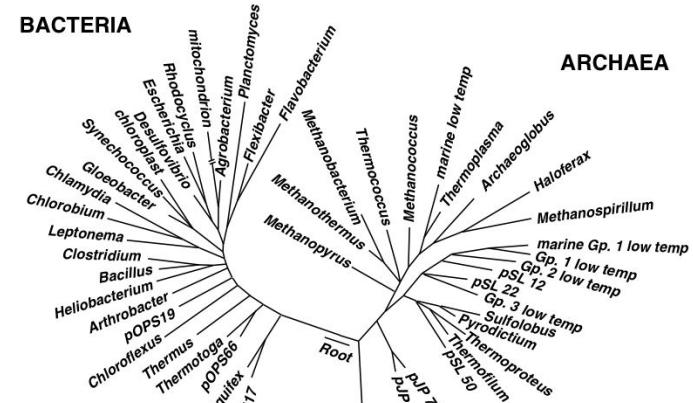
Homo sapiens



Porphyra yezoensis



BACTERIA



ARCHAEA

Image sources:
https://en.wikipedia.org/wiki/Homo_sapiens#/media/File:Akha_cropped_hires.JPG
https://en.wikipedia.org/wiki/Coprinus#/media/File:Coprinus_comatus_fresh.JPG
https://en.wikipedia.org/wiki/Maize#/media/File:Contassell_7095.jpg
https://en.wikipedia.org/wiki/Porphyra#/media/File:Porphyra_yezoensis.jpg

Microbes live in complex and diverse communities, everywhere on Earth.

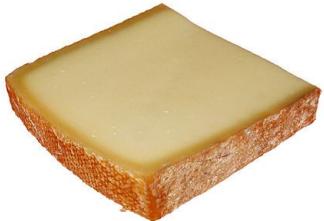


Image sources:

https://en.wikipedia.org/wiki/Homo_sapiens#/media/File:Akha_cropped_hires.JPG
https://en.wikipedia.org/wiki/Maize#/media/File:Cornassel_7095.jpg
<https://en.wikipedia.org/wiki/Bergk%C3%A4se#/media/File:Bergk%C3%A4se2.jpg>

Photo by Greg Caporaso

Our microbiomes impact the efficacy of medical treatment.

Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors

Bertrand Routy,^{1,2,3} Emmanuelle Le Chatelier,⁴ Lisa Derosa,^{1,2,3}
Connie P. M. Duong,^{1,2,5} Maryam Tidjani Alou,^{1,2,3} Romain Daillère,^{1,2,3}
Aurélie Fluckiger,^{1,2,5} Meriem Messaudene,^{1,2} Conrad Rauber,^{1,2,3} Maria P. Roberti,^{1,2,5}
Marine Fidelle,^{1,3,5} Caroline Flament,^{1,2,5} Vichnou Poirier-Colame,^{1,2,5} Paule Opolon,⁶
Christophe Klein,⁷ Kristina Iribarren,^{8,9,10,11,12} Laura Mondragón,^{8,9,10,11,12}
Nicolas Jacquemet,^{1,2,3} Bo Qu,^{1,2,3} Gladys Ferrere,^{1,2,3} Céline Clémenson,^{1,13}
Laura Mezquita,^{1,14} Jordi Remon Masip,^{1,14} Charles Naltet,¹⁵ Solenne Brosseau,¹⁵
Coureche Kaderbhai,¹⁶ Corentin Richard,¹⁶ Hira Rizvi,¹⁷ Florence Levenez,⁴
Nathalie Galleron,⁴ Benoit Quinquis,⁴ Nicolas Pons,⁴ Bernhard Ryffel,¹⁸
Véronique Minard-Colin,^{1,19} Patrick Gonin,^{1,20} Jean-Charles Soria,^{1,14} Eric Deutsch,^{1,13}
Yohann Loriot,^{1,3,14} François Ghiringhelli,¹⁶ Gérard Zalcman,¹⁵
François Goldwasser,^{9,21,22} Bernard Escudier,^{1,14,23} Matthew D. Hellmann,^{24,25}
Alexander Eggermont,^{1,2,14} Didier Raoult,²⁶ Laurence Albiges,^{1,3,14}
Guido Kroemer,^{8,9,10,11,12,27,28*} Laurence Zitvogel^{1,2,3,5,*}

Good bacteria help fight cancer

Resident gut bacteria can affect patient responses to cancer immunotherapy (see the Perspective by Jobin). Routy et al. show that antibiotic consumption is associated with poor response to immunotherapeutic PD-1 blockade. They profiled samples from patients with lung and kidney cancers and found that nonresponding patients had low levels of the bacterium *Akkermansia muciniphila*. Oral supplementation of the bacteria to antibiotic-treated mice



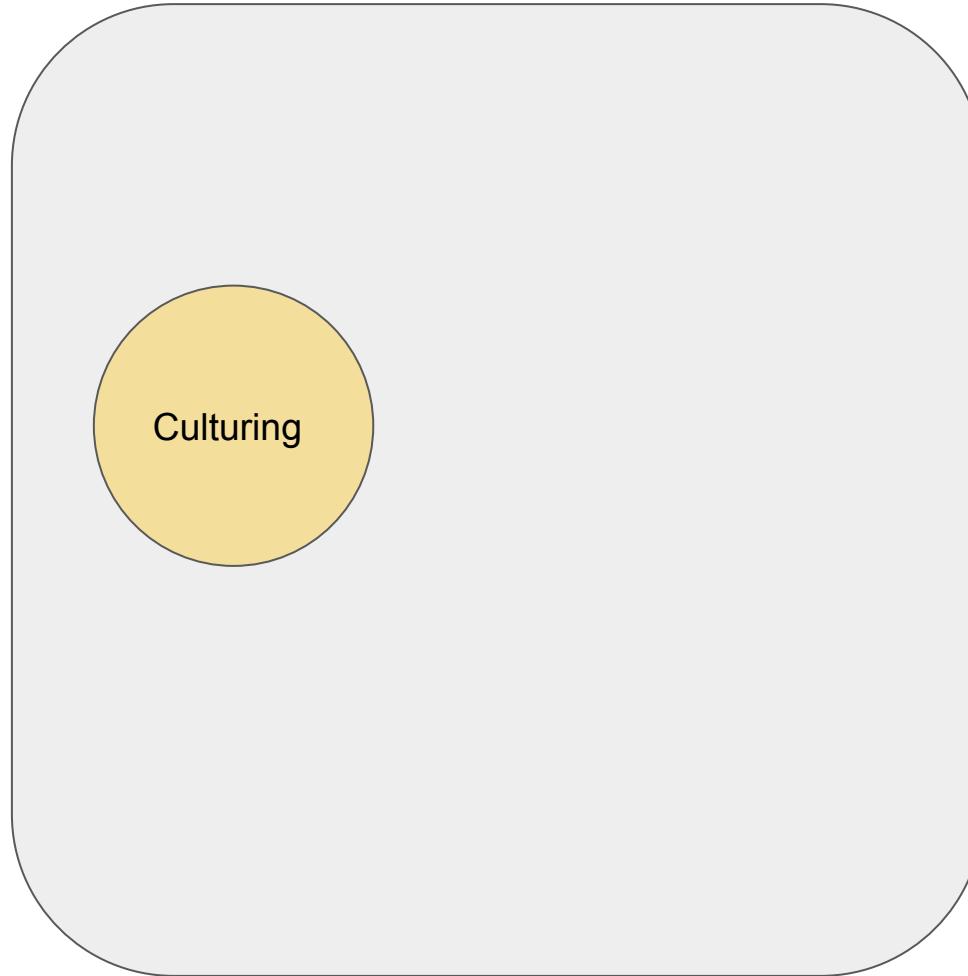
Microbes can help us reduce our impact on the Earth by composting our food waste. And, maybe even by degrading pollution...



Understanding microbiomes may lead to new food varieties and more sustainable crops.

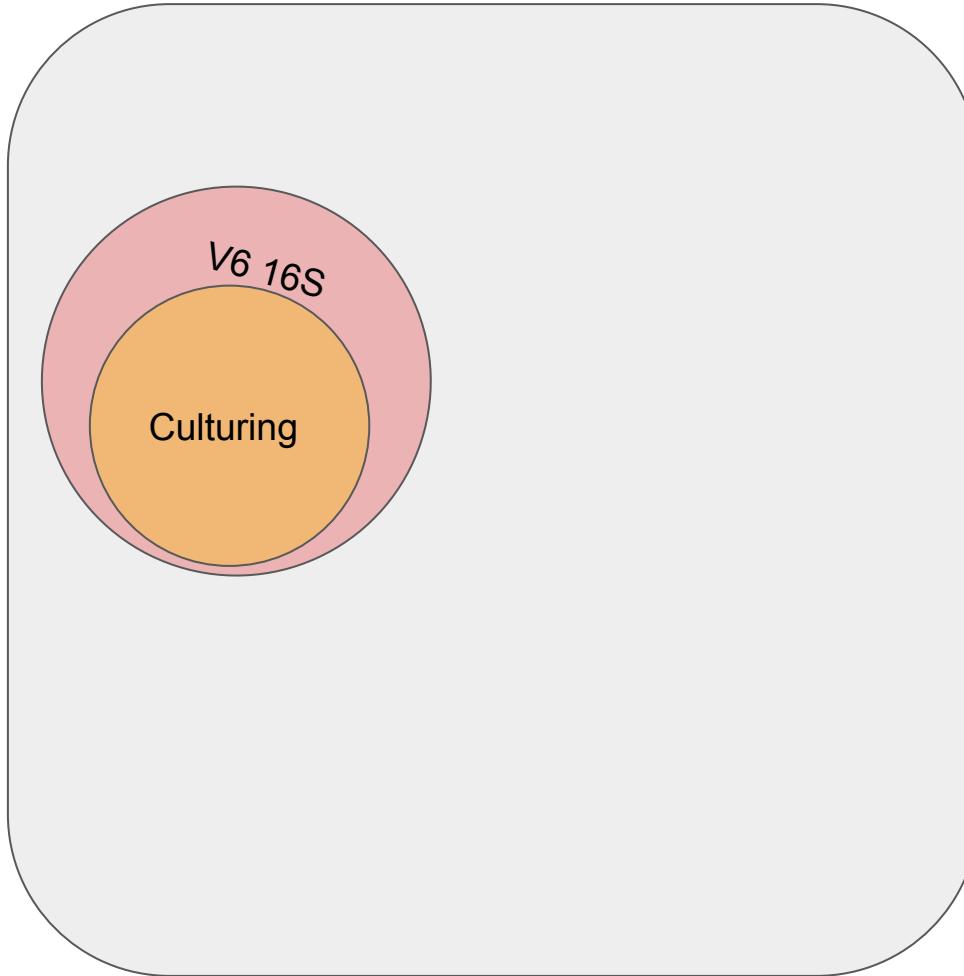


We illuminate
different regions of
the microbial world
(represented in grey)
with different
technologies
(represented in other
colors).



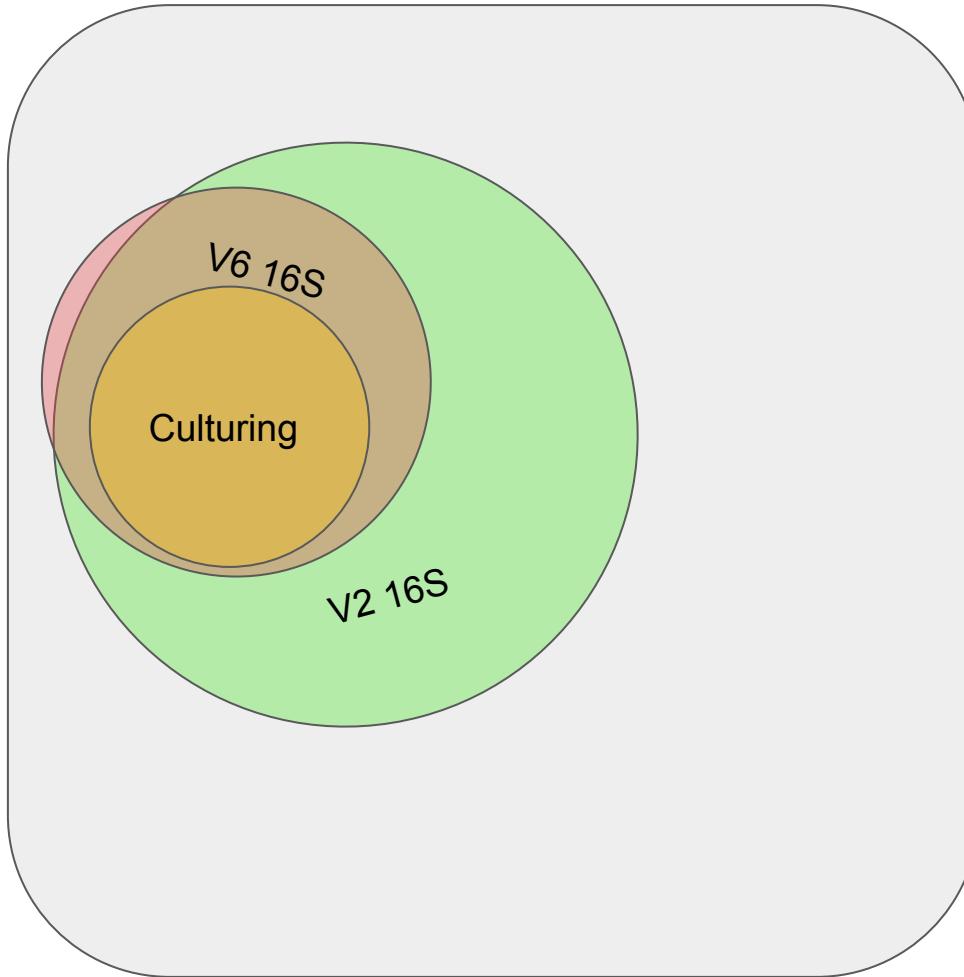
(These figures are not intended to adequately represent scale.)

We illuminate
different regions of
the microbial world
(represented in grey)
with different
technologies
(represented in other
colors).



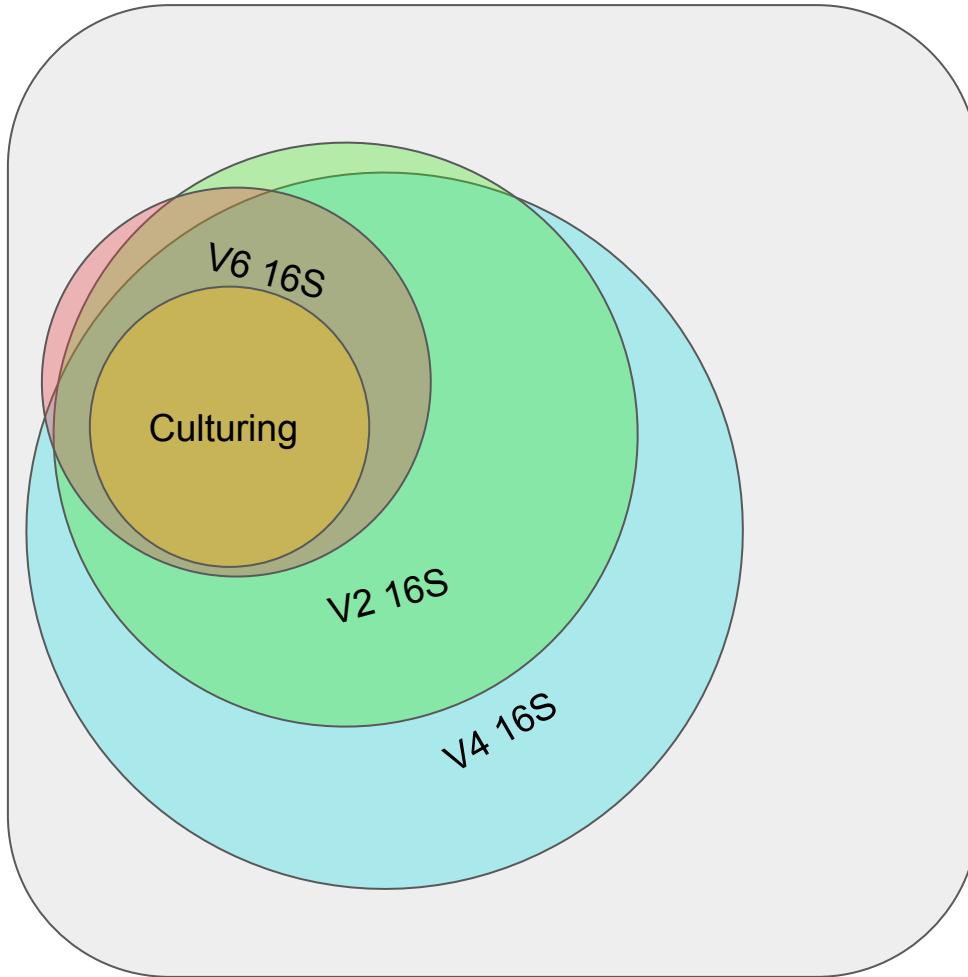
(These figures are not intended to adequately represent scale.)

We illuminate
different regions of
the microbial world
(represented in grey)
with different
technologies
(represented in other
colors).



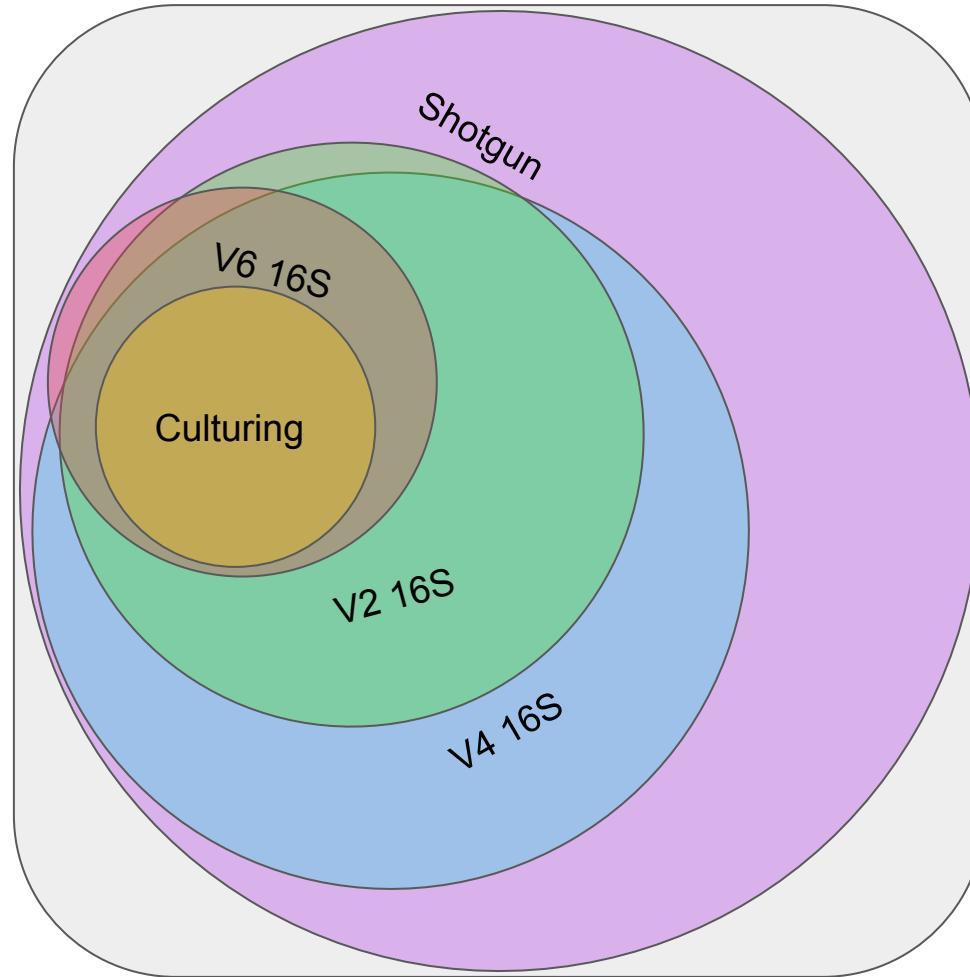
(These figures are not intended to adequately represent scale.)

We illuminate
different regions of
the microbial world
(represented in grey)
with different
technologies
(represented in other
colors).



(These figures are not intended to adequately represent scale.)

We illuminate different regions of the microbial world (represented in grey) with different technologies (represented in other colors).



(These figures are not intended to adequately represent scale.)

Geographic Information Systems layers:

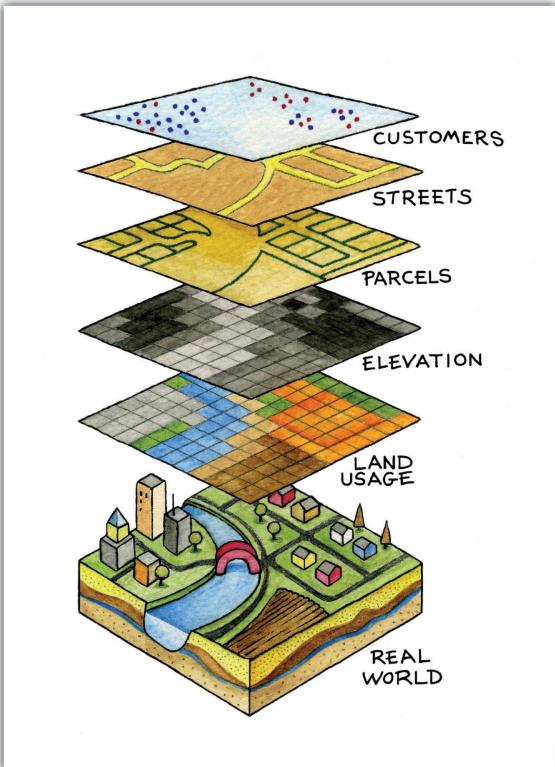


Image source: [Essentials of Geographic Information Systems, v. 1.0](#)
by Jonathan Campbell and Michael Shin

Geographic Information Systems layers:

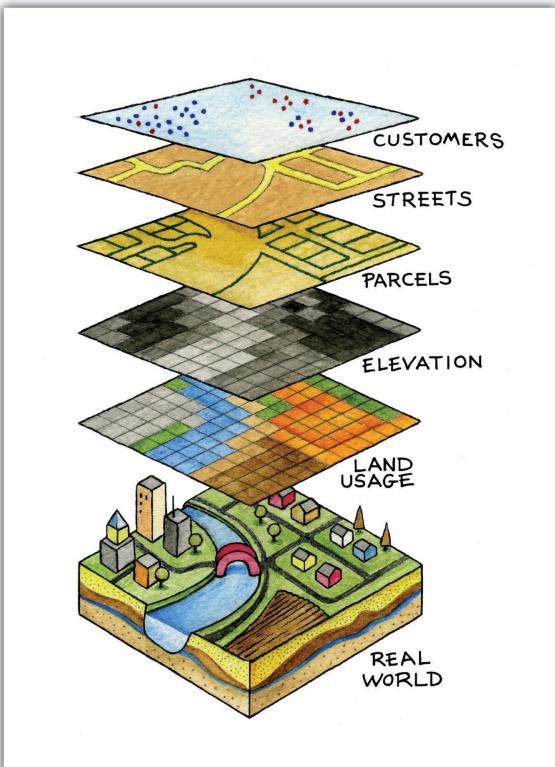


Image source: [Essentials of Geographic Information Systems, v. 1.0](#)
by Jonathan Campbell and Michael Shin

Microbiome layers:

Taxonomy:

- Bacteria and archaea (16S rRNA surveys)
- Fungus and other eukaryotes (ITS and 18S rRNA surveys)
- Phage and other viruses (shotgun surveys)

Functional potential via shotgun metagenome surveys

Functional activity:

- Metatranscriptome
- Metaproteome
- Metabolome

I think there is a world market for
maybe five computers.

- Thomas Watson, Chairman and CEO of IBM 1914–1956

I think there is a world market for maybe five computers.

- Thomas Watson, Chairman and CEO of IBM 1914–1956

It is difficult to make predictions, especially about the future.

- Journal of the Royal Statistical Society (1956), “Proceedings of the Meeting”, [Speaker: Bradford Hill]



200x magnification and the discovery of “animalcules”, ~1671 by van Leeuwenhoek

Image source: Discover Magazine (2015)



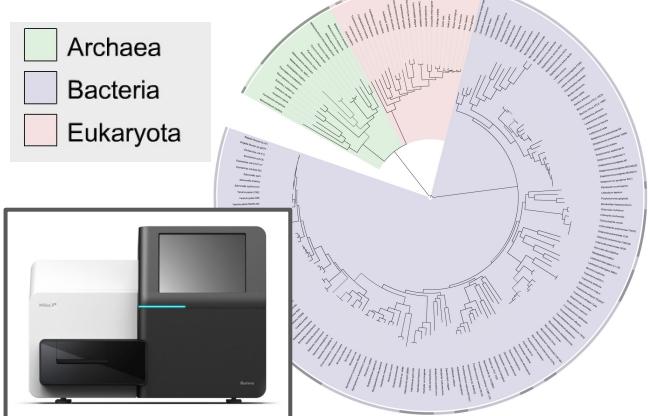
Bacillus anthracis in monoculture.

Image source: U.S. Army Medical Research Institute of Infectious Diseases

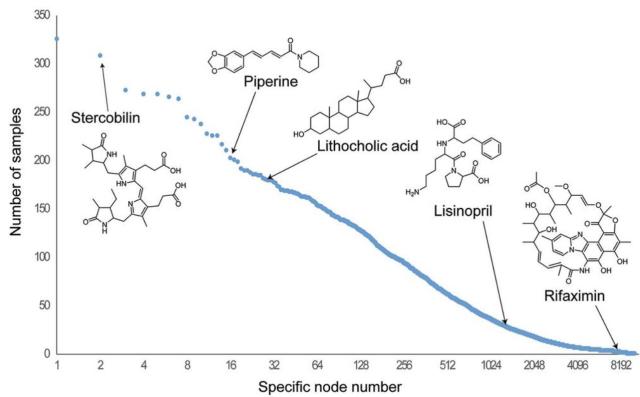


Winogradsky columns model microbial ecosystems.

Image source: www.hhmi.org/biointeractive

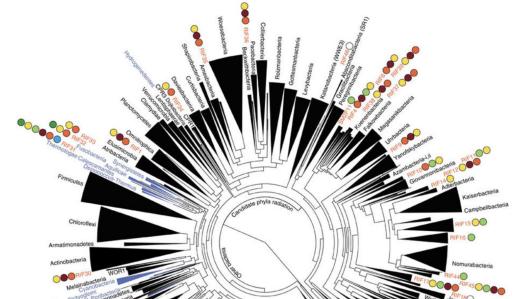


Sequencing of ribosomal RNA and marker genes.



Mass spectrometry to study small molecules.

Image source: McDonald et al (2018) DOI: 10.1128/mSystems.00031-18



Highly multiplexed bacterial genome sequencing.

Image source: Anantharaman et al (2016) DOI: 10.1038/ncomms1329



nature
biotechnology

Correspondence | Published: 24 July 2019

Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2

Evan Bolyen, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A.

Al-Ghalith, Harriet Alexander, Eric J. Alm, Manimozhiyan Arumugam, Francesco Asnicar, Yang Bai,

Jordan E. Bisanz, Kyle Bittinger, Asker Brejnrod, Colin J. Brislawn, C. Titus Brown, Benjamin J.

Callahan, Andrés Mauricio Caraballo-Rodríguez, John Chase, Emily K. Cope, Ricardo Da Silva,

Christian Diener, Pieter C. Dorrestein, Gavin M. Douglas, Daniel M. Durall, Claire Duvall, Christian F.

Edwardson, Madeleine Ernst, Mehrobd Estaki, Jennifer Fouquier, Julia M. Gauglitz, Sean M. Gibbons,

Deanna L. Gibson, Antonio Gonzalez, Kestrel Gorlick, Jiarong Guo, Benjamin Hillmann, Susan

Holmes, Hannes Holste, Curtis Huttenhower, Gavin A. Huttley, Stefan Janssen, Alan K. Jarmusch,

Lingjing Jiang, Benjamin D. Kaebler, Kyo Bin Kang, Christopher R. Keefe, Paul Keim, Scott T. Kelley,

Dan Knights, Irina Koester, Tomasz Kosciolek, Jorden Kreps, Morgan G. I. Langille, Joslynn Lee, Ruth

Ley, Yong-Xin Liu, Erikka Loftfield, Catherine Lozupone, Massoud Maher, Clarisse Marotz, Bryan D.

Martin, Daniel McDonald, Lauren J. Mciver, Alexey V. Melnik, Jessica L. Metcalf, Sydney C. Morgan,

Jamie T. Morton, Ahmad Turan Naimey, Jose A. Navas-Molina, Louis Felix Nothias, Stephanie B.

Orchianian, Talima Pearson, Samuel L. Peoples, Daniel Petras, Mary Lai Preuss, Elmar Pruesse, Lasse

Buur Rasmussen, Adam Rivers, Michael S. Robeson II, Patrick Rosenthal, Nicola Segata, Michael

Shaffer, Arron Shiffer, Rashmi Sinha, Se Jin Song, John R. Spear, Austin D. Swafford, Luke R.

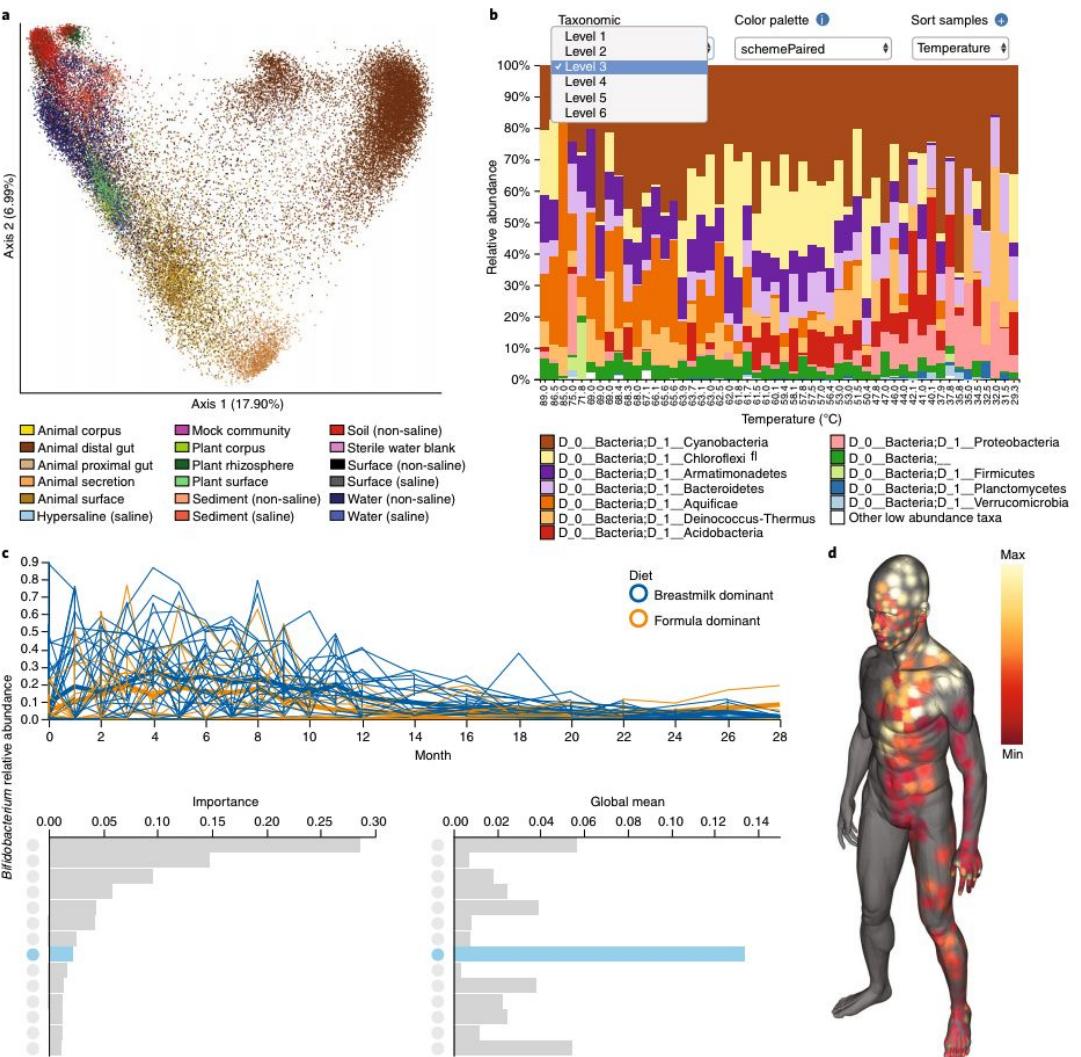
Thompson, Pedro J. Torres, Pauline Trinh, Anupriya Tripathi, Peter J. Turnbaugh, Sabah Ul-Hasan,

Justin J. van der Hooft, Fernando Vargas, Yoshiki Vázquez-Baeza, Emily Vogtmann, Max von

Hippel, William Walters, Yunhu Wan, Mingxun Wang, Jonathan Warren, Kyle C. Weber, Charles H. D.

Williamson, Amy D. Willis, Zhenjiang Zech Xu, Jesse R. Zaneveld, Yilong Zhang, Qiyun Zhu, Rob

Knight & J. Gregory Caporaso - Show fewer authors



Over 23,000 QIIME citations

(Source: Google Scholar, 4 Oct 2020)

Structure, function and diversity of the healthy human microbiome

The Human Microbiome Project Consortium

Affiliations | Contributions | Corresponding author

Nature 486, 207–214 (14 June 2012) | doi:10.1038/nature11234

Received 02 November 2011 | Accepted 16 May 2012 | Published online 13 June 2012

Intestinal microbiome is related to lifetime antibiotic use in Finnish pre-school children

Katri Korppela, Anne Salonen, Lauri J. Virta, Riina A. Kekkonen, Kristoffer Forslund, Peer Bork & Willem M. de Vos

Affiliations | Contributions | Corresponding author

Nature Communications 7, Article number: 10410 | doi:10.1038/ncomms10410

Mx1 reveals innate pathways to antiviral resistance and lethal influenza disease

Padmini S. Pillai¹, Ryan D. Molony¹, Kimberly Martinod², Huiping Dong³, Iris K. Pang¹, Michal C. Tai^{1,*}, Angel G. Solis¹, Piotr Bielecki¹, Subhasis Mohanty³, Mark Trentalange⁴, Robert J. Homer⁵, Richard A. Flavell^{1,8}, Denisa D. Wagner², Ruth R. Montgomery⁶, Albert C. Shaw³, Peter Staeheli⁷, Akiko Iwasaki^{1,8,†}

+ Author Affiliations

↔ Corresponding author. Email: akiko.iwasaki@yale.edu

† Present address: Institute of Stem Cell Biology and Regenerative Medicine, School of Medicine, Stanford, CA 94305, USA.

Science 22 Apr 2016:

Vol. 352, Issue 6284, pp. 463–466

DOI: 10.1126/science.aaf3926

Cell Host & Microbe

Volume 19, Issue 5, 11 May 2016, Pages 731–743

Resource

Genetic Determinants of the Gut Microbiome in UK Twins

Julia K. Goodrich¹, Emily R. Davenport¹, Michelle Beaumont², Matthew A. Jackson², Rob K. Ober³, Tim D. Spector², Jordana T. Bell², Andrew G. Clark¹, Ruth E. Ley^{1,5}, ▲, ■

Marine mammals harbor unique microbiota by and yet distinct from the sea

Elisabeth M. Bik, Elizabeth K. Costello, Alexandra D. Switzer, Benjamin J. Callahan, Susan P. Holmes, Randall S. Wells, Kevin P. Carlin, Eric D. Jensen, Stephanie Venn-Watson & David A. Relman

Affiliations | Contributions | Corresponding author

Nature Communications 7, Article number: 10516 | doi:10.1038/ncomms10516

Received 04 August 2015 | Accepted 18 December 2015 | Published 03 February 2016

Microbial community assembly and metabolic function during mammalian corpse decomposition

Jessica L. Metcalf^{1,2,*}, Zhenjiang Zech Xu², Sophie Weiss³, Simon Lux^{4,5}, Will Van Treuren⁶, Embrettie R. Hyde⁶, Se Jin Song^{1,2}, Amnon Amir², Peter Larsen^{4,7}, Naseer Sangwan^{4,7,8}, Daniel Haarmann⁹, Greg C. Humphrey⁹, Gail Ackermann², Luke R. Thompson², Christian Lauber¹⁰, Alexander Bibat¹¹, Catherine Nicholas¹¹, Matthew J. Geber¹¹, Joseph F. Petrosino¹², Sasha C. Reed¹³, Jack A. Gilbert^{4,5,7,8,14}, Aaron M. Lynne⁹, Sibyl R. Bucheli⁹, David O. Carter¹⁵, Rob Knight^{2,16,*}

+ Author Affiliations

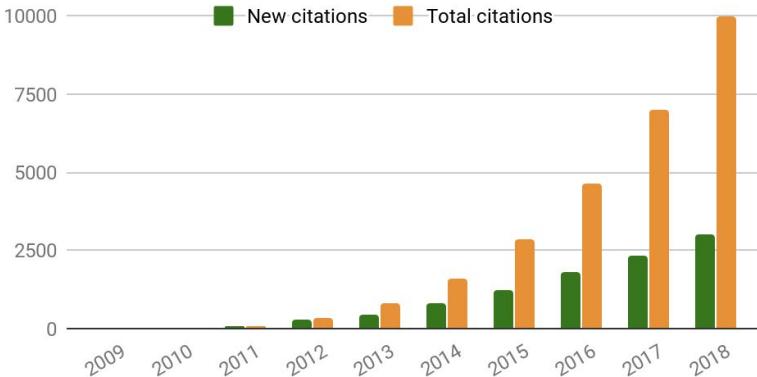
↔ Corresponding author. E-mail: robknight@ucsd.edu (R.K.); jessica.metcalf@colorado.edu (J.L.M.)

Science 08 Jan 2016:

Vol. 351, Issue 6269, pp. 158–162

DOI: 10.1126/science.aad2546

New citations and total citations by year (source: Web of Science)



The ISME Journal (2016) 10, 1308–1322; doi:10.1038/ismej.2015.221

Wind and sunlight shape microbial diversity in surface waters of the North Pacific Subtropical Gyre
OPEN

Jessica A Bryant^{1,2}, Frank O Aylward^{2,3}, John M Eppley^{2,3}, David M Karl^{2,3}, Matthew J Church^{2,3} and Edward F DeLong^{1,2,3}

Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients

V. Gopalakrishnan^{1,2,*}, C. N. Spencer^{2,3,*}, L. Nezi^{3,7}, A. Reuben¹, M. C. Andrews¹, T. V. Karpinets³, P. A. Prieto^{1,1}, D. Vicente^{1..}

* See all authors and affiliations

Science 02 Nov 2017:

eaan4236

DOI: 10.1126/science.aan4236

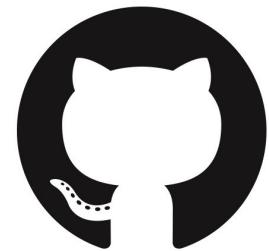


Peer Reviewed
← see details

Completely Free and Open Source

- You have a *licensed right* to use, redistribute, and even modify QIIME 2
 - (Consider [contributing to the QIIME 2 project](#) via GitHub Pull Requests!)
- You can audit every line of code
- [Tested on every code change](#)
 - Unit tests, code coverage, linting
 - Tutorial commands automatically tested
- World-wide [network of collaborators](#)

Find the code at: <https://github.com/qiime2/>



QIIME 2 has succeeded QIIME 1 as of January 2018, and QIIME 1 is no longer supported! Learn more about what that means [here](#).

Installing QIIME 2



Version: 2020.8 ▾

Table of Contents

- Getting started
- What is QIIME 2?
- Core concepts
- Installing QIIME 2
 - Natively installing QIIME 2
 - Installing QIIME 2 using Virtual Machines
 - Recommendations
 - QIIME 2 Core 2020.8 distribution
- Tutorials
- Interfaces
- Plugins
- Semantic types
- Community
- Data resources
- Supplementary resources
- User Glossary
- Citing QIIME 2

Quick search

Natively installing QIIME 2

This guide describes how to natively install the QIIME 2 Core 2020.8 distribution.

Miniconda

Installing Miniconda

Miniconda provides the `conda` environment and package manager, and is the recommended way to install QIIME 2. Follow the [Miniconda instructions](#) for downloading and installing Miniconda. You may choose either Miniconda2 or Miniconda3 (i.e. Miniconda Python 2 or 3). QIIME 2 will work with either version of Miniconda. It is important to follow all of the directions provided in the [Miniconda instructions](#), particularly ensuring that you run `conda init` at the end of the installation process, to ensure that your Miniconda installation is fully installed and available for the following commands.

Updating Miniconda

After installing Miniconda and opening a new terminal, make sure you're running the latest version of `conda`:

```
conda update conda
```

Installing wget

```
conda install wget
```

Install QIIME 2 within a conda environment

Once you have Miniconda installed, create a `conda` environment and install the QIIME 2 Core 2020.8 distribution within the environment. We **highly** recommend creating a *new* environment specifically for the QIIME 2 release being installed, as there are many required dependencies that you may not want added to an existing environment. You can choose whatever name you'd like for the environment. In this example, we'll name the environment `qiime2-2020.8` to indicate what QIIME 2 release is installed (i.e. 2020.8).

Instructions

macOS/OS X (64-bit)

Linux (64-bit)

Windows Subsystem for Linux (64-bit)

From the above tabs, please choose the installation instructions that are appropriate for your platform.

Options for installing QIIME 2:

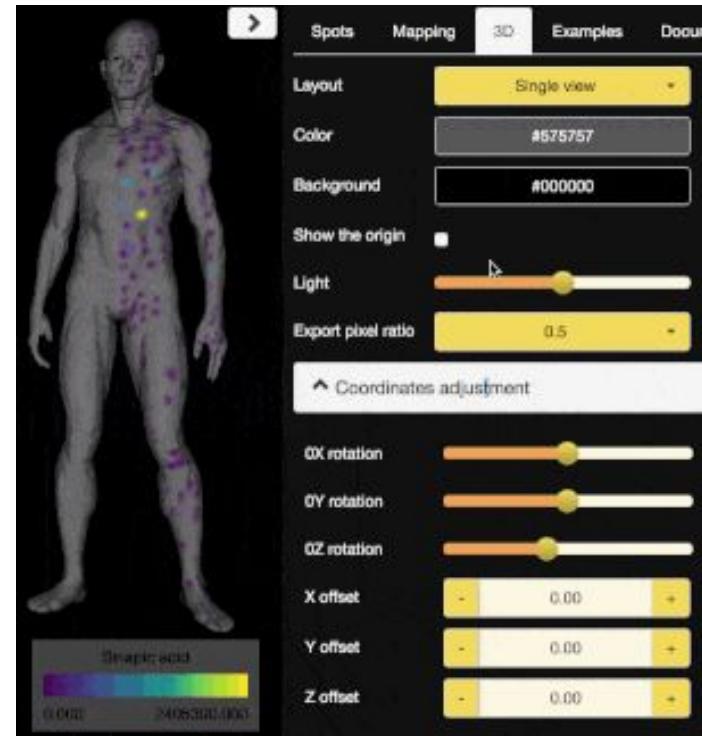
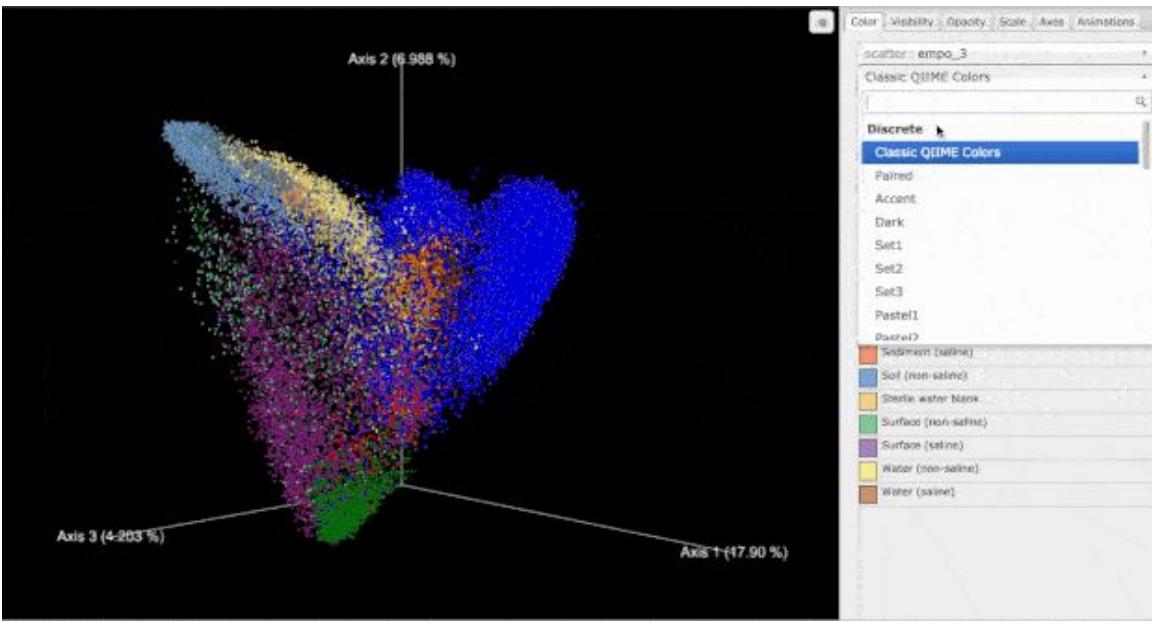
- Natively on your computer with Anaconda
- Virtual Machine on your computer with VirtualBox or Docker
- Virtual Machine on Amazon Web Services



High-level features (these attract users to system)

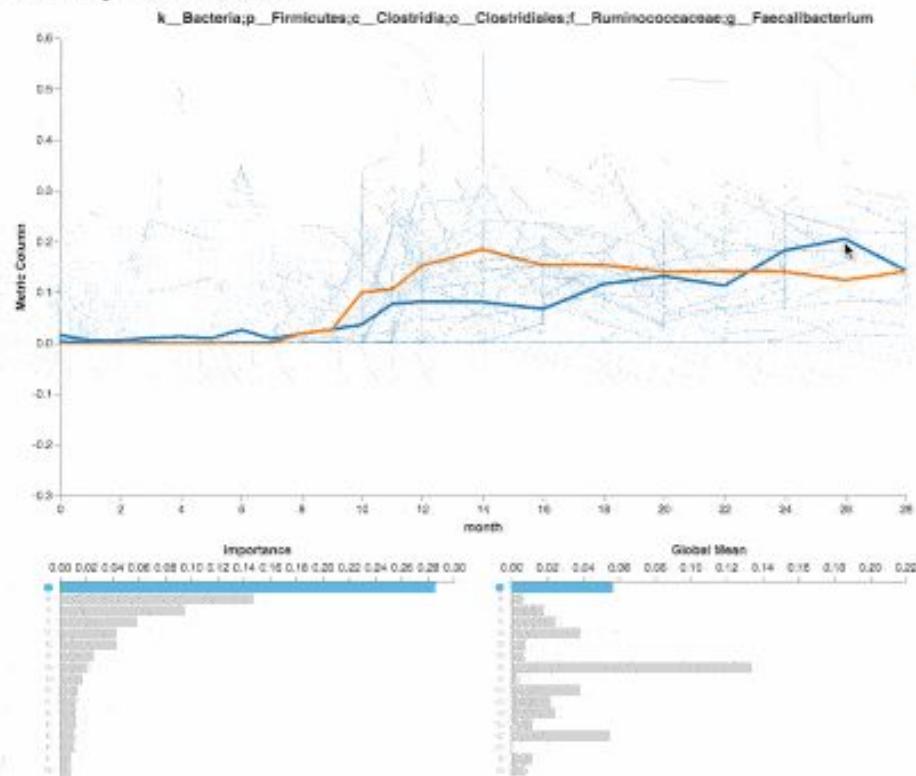
- **Latest and greatest microbiome bioinformatics methods and visualizations.**
- **Accessibility** through accurate, detailed, and interesting documentation and well-designed interfaces.
- **A community** of microbiome scientists, developers, and bioinformaticians.

QIIME 2: latest methods + interactive visualizations



QIIME 2: latest methods + interactive visualizations

Volatility Control Chart



Plot Controls

Export as TSV Save as SVG Save as PNG

[View Source](#) [Open in Vega Editor](#)

Click the individual group value labels in the legend to toggle their visibility in the displayed plot.

Group column

antiesposedell

Metric column

k_Bacteria;p_Firmicutes;

Color schema

category10

Show error

bars

Show global

mean

Show global control limits

+/- 2x and 3x standard deviations from global mean

Mean line width

3

Spaghetti line width

0.5

Mean line opacity

1

Spaghetti line opacity

0.5

Mean scatter size

292

Spaghetti scatter size

50

Mean scatter opacity

0

Spaghetti scatter opacity

0

Feature Stats Subplot Controls

The bar charts show raw abundance statistics for all features



<https://forum.qiime2.org>

Please read our [Code of Conduct](#) when joining.

all categories ► all tags ► Categories Latest New (6) Unread (1) Top Bookmarks My Pos

Category

User Support

Post to this category if you need help understanding output produced while running QIIME 2. Examples of this include help understanding plots/labels, techniques that are used in QIIME 2, etc. Posts in this category will be triaged by a QIIME 2 Moderator and responded to promptly.

Technical Support

Post to this Category if you are experiencing a technical difficulty while running QIIME 2. Examples of difficulties include installation errors, help deciphering error messages, etc. Posts in this category will be triaged by a QIIME 2 Moderator and responded to promptly.

Community Plugin Support

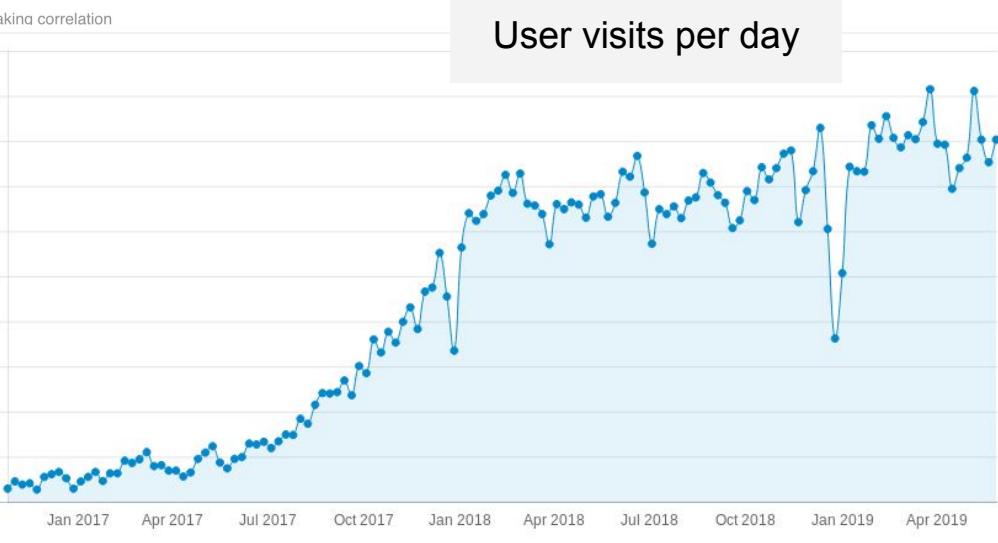
Post to this category if you have a question about a plugin (bug report, technical detail, etc.). Contributions are distributed in the QIIME 2 Core Distribution, so we are not we are planning on moving away from "Distribution," where all plugins are...

QIIME2中文帮助文档 (Chinese Manual)

Community Translations

In-progress

Install



QIIME 2 workshops

<https://workshops.qiime2.org>

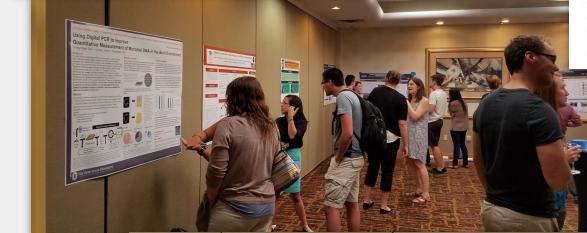
Upcoming Workshops

Title	Location	Workshop Dates
Microbiome Bioinformatics with QIIME 2 - Online!	Online	Oct. 5, 2020 - Oct. 9, 2020
An Introduction to QIIME 2	Online (via World of Microbiome)	Nov. 4, 2020 - Nov. 4, 2020
Microbiome Bioinformatics with QIIME 2	Online (via FAES at the National Institutes of Health)	Jan. 4, 2021 - Jan. 8, 2021

Past Workshops

Title	Location	Workshop Dates
Introduction to microbiome study design and analysis	Puerto Rico	Aug. 1, 2020 - Aug. 1, 2020
Microbiome Bioinformatics with QIIME 2	Bethesda, Maryland	Jan. 8, 2020 - Jan. 10, 2020
Microbiome Bioinformatics with QIIME 2 Workshop	Fort Collins, Colorado (USA)	Nov. 25, 2019 - Nov. 26, 2019
Microbiome Bioinformatics with QIIME 2 Workshop (not open to the public)	University of Wyoming	Oct. 24, 2019 - Oct. 25, 2019
Microbiome Bioinformatics with QIIME 2	Bangkok, Thailand	Sept. 11, 2019 - Sept. 12, 2019
QIIME 2 @ One Health Summer School	University of Bern, Switzerland	Aug. 14, 2019 - Aug. 14, 2019
Strategies and Techniques for Analyzing Microbial Population Structures (STAMPS); includes a QIIME 2 session	Woods Hole, MA, USA	July 29, 2019 - July 29, 2019

Las Vegas Workshop, June 2017



Register for the June 2017 *Microbiome Bioinformatics with QIIME 2* workshop in Las Vegas by 11:59pm Pacific Time on May 5th for a chance to win one of three limited-edition QIIME 2 cross-stitches. Be the envy of your lab by owning this beautiful piece of microbiome bioinformatics art!

<https://workshops.qiime.org/qiime-2-workshop-2017-06-21>



Fine art by Matthew Dillon (@thermokarst).



Low-level features (these get our users hooked)

- **Decentralized provenance tracking** automates bioinformatics record keeping facilitating reproducibility.
- **Multiple user interfaces.** The same functionality is accessible through graphical interface, command line interface, and API, which target different types of users.
- **Plugin architecture** allows the software to keep pace with the field. Any developer can create and distribute a QIIME 2 plugin.

Retrospective data provenance tracking (or “What did you do 5 months ago?”)

notes.txt

```
echo "core_diversity_analyses.py -i  
/home/caporaso/analysis/atacama-7may2013/slout_r1/or_otus/otu_table_mc2_w_tax_no_pynast_failures.  
biom -o /home/caporaso/analysis/atacama-7may2013/slout_r1/or_otus/cd_16662/ -e 16662 -m  
/home/caporaso/analysis/atacama-7may2013/map.txt -ao 26 -t  
/home/caporaso/analysis/atacama-7may2013/slout_r1/or_otus/rep_set.tre -c  
SiteName,Depth,ExtractGroupNo,TransectID,Vegetation" | qsub -keo -N ata-cd
```

The above failed during OTU category significance (see the log file for the error - maybe one of these categories has a value that is observed only once?), so re-running without that step for now...

```
echo "core_diversity_analyses.py -i  
/home/caporaso/analysis/atacama-7may2013/slout_r1/or_otus/otu_table_mc2_w_tax_no_pynast_failures.  
biom -o /home/caporaso/analysis/atacama-7may2013/slout_r1/or_otus/cd_16662/ -e 16662 -m  
/home/caporaso/analysis/atacama-7may2013/map.txt -ao 26 -t  
/home/caporaso/analysis/atacama-7may2013/slout_r1/or_otus/rep_set.tre -c  
SiteName,Depth,ExtractGroupNo,TransectID,Vegetation --suppress_otu_category_significance  
--recover_from_failure" | qsub -keo -N ata-cd
```

```
echo "pick_open_reference_otus.py -i /home/caporaso/analysis/atacama-7may2013/slout_r2/seqs.fna  
-r /data/gg_13_5_otus/rep_set/97_otus.fasta -o  
/home/caporaso/analysis/atacama-7may2013/slout_r2/or_otus/ -ao 28 -p  
/home/caporaso/analysis/atacama-7may2013/uc_fast_params.txt" | qsub -keo -N or-3
```

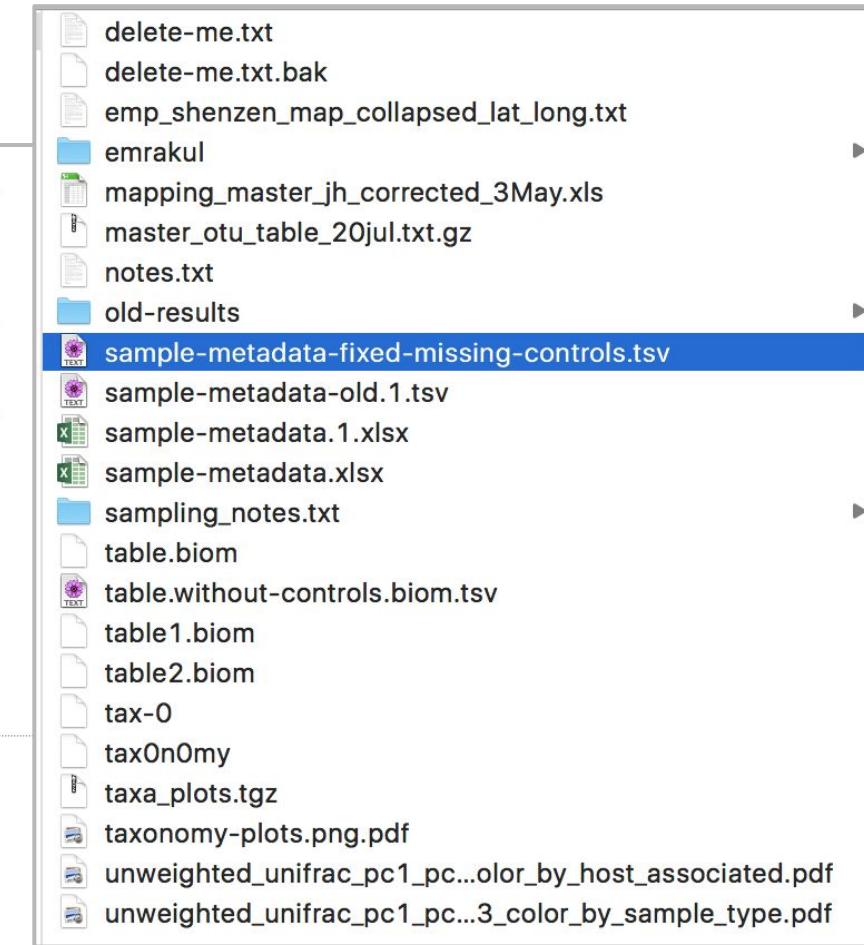
Katy's analysis on first and second sequencing runs

You have read access to all files in /home/caporaso/analysis/atacama-7may2013.

The split_libraries_fastq.py input files you need are:

sequences:

```
/home/caporaso/analysis/atacama-7may2013/2014.04.30/Undetermined_S0_L0  
01_R1_001.fastq.gz
```



QIIME 2 integrated data provenance ensures reproducibility

(try this out by clicking the *Provenance* tab [here](#))

qiime2view

File: taxa-bar-plots.qzv

Download: SVG (bars) SVG (legend) CSV

Hover over the plot to learn more

Relative Abundance

Sample

Taxonomic Level

- Level 1
- Level 2
- Level 3
- Level 4
- Level 5
- Level 6
- Level 7

Color Palette

Sort Samples By

schemeAccent

k_Bacteria.p_Firmicutes

Ascending

qiime2view

File: taxa-bar-plots.qzv

Visualization Peak Provenance

Provenance Graph

```
graph TD; seqs --> sequences; sequences --> reference_taxonomy; reference_taxonomy --> demultiplexed_seqs; reference_taxonomy --> reference_reads; demultiplexed_seqs --> reads_classifier; reference_reads --> reads_classifier; reads_classifier --> tab_prime[tab']; taxonomy --> tab_prime;
```

Action Details

execution:

- uuid: "3897fb5c-55ed-46b1-a48d-ae0651d2b597"

runtime:

- start: 2017-09-28T21:14:34.374Z
- end: 2017-09-28T21:23:05.935Z
- duration: "8 minutes, 31 seconds, and 561708 microsecond s"

action:

- type: "method"
- plugin: "environment;plugins:dada2"
- action: "denoise_single"

inputs:

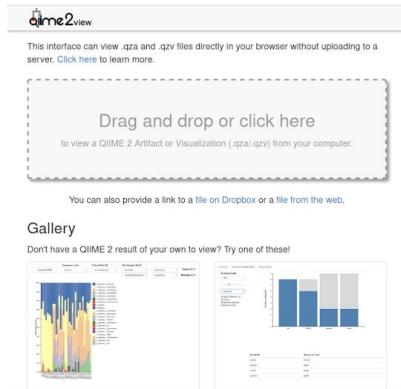
- 0:
 - demultiplexed_seqs: "ce7e102e-4b8c-455c-b2af-a7fb342b7fa1"

parameters:

- 0:
 - trunc_len: 120
- 1:
 - trim_left: 0
- 2:
 - max_ee: 2
- 3:
 - trunc_q: 2
- 4:
 - chimera_method: "consensus"
- 5:
 - min_fold_parent_over_abundance: 1
- 6:
 - n_threads: 1

Choose the interface that meets your needs

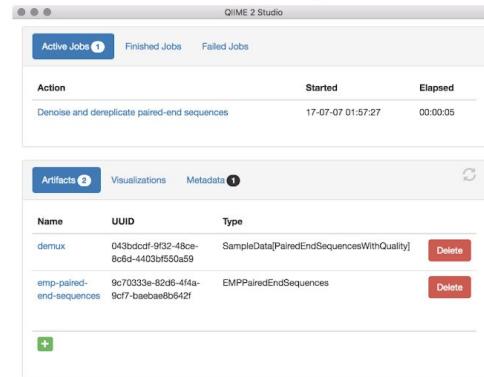
a) QIIME 2 View



Related Software:

Galaxy

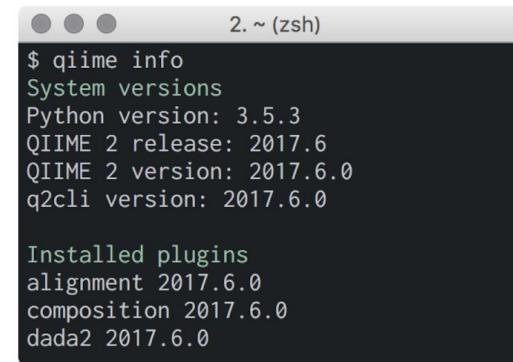
b) QIIME 2 Studio prototype



Related Software:

Galaxy
EBI Metagenomics Portal
QIITA
NIH Nephele

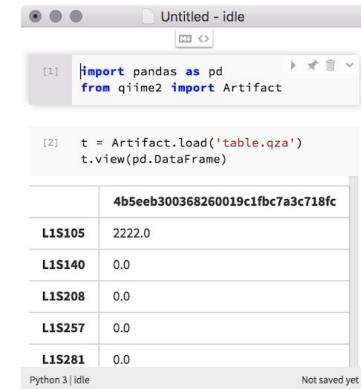
c) QIIME 2 CLI



Related Software:

Mothur
QIIME 1

d) QIIME 2 Artifact API



Related Software:
phyloseq

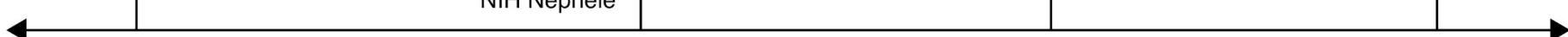
Data analyst (clinician, policy maker, research subject)

Cancer researchers and other domain scientists

Power users

Data scientists

COMPUTATIONAL SOPHISTICATION



QIIME 2 View (<https://view.qiime2.org>): a read-only web interface for viewing results without having QIIME 2 installed

- Reads QIIME 2 outputs (QZAs and QZVs)
- No installation required
- Easily share links using Dropbox with collaborators
- No uploading (your data stays on your computer)!

The screenshot shows the QIIME 2 View web interface. At the top, it says "qiime2view" and provides instructions: "This interface can view .qza and .qzv files directly in your browser without uploading to a server. [Click here](#) to learn more." Below this is a dashed box containing the text "Drag and drop or click here to view a QIIME 2 Artifact or Visualization (.qza/.qzv) from your computer." Further down, it says "You can also provide a link to a file on Dropbox or a [file from the web](#)." A section titled "Gallery" shows two examples: a phylogenetic tree and a bar chart. The phylogenetic tree is labeled "Duchesne_2016" and the bar chart is labeled "Duchesne_2016". Both have dropdown menus for "Select Sample ID" and "Select Feature ID".

Anyone can create and distribute a plugin.

- They define *all* bioinformatics analysis functionality.
- They're Python 3 "method annotations" that QIIME 2 interprets.
- They can wrap methods not written in Python 3 (e.g., DADA2 is written in R, and mafft is a binary).
- The QIIME 2 Library (<https://library.qiime2.org>) is the primary site for discovering and disseminating QIIME 2 plugins.

```
8
9 import qiime2.plugin
10 from q2_types.per_sample_sequences import (
11     SequencesWithQuality, PairedEndSequencesWithQuality)
12 from q2_types.sample_data import SampleData
13 from q2_types.feature_data import FeatureData, Sequence
14 from q2_types.feature_table import FeatureTable, Frequency
15
16 import q2_dada2
17
18 plugin = qiime2.plugin.Plugin(
19     name='dada2',
20     version=q2_dada2.__version__,
21     website='http://benjneb.github.io/dada2/',
22     package='q2_dada2',
23     description='This QIIME 2 plugin wraps DADA2 and supports '
24         'sequence quality control for single-end and paired-end '
25         'reads using the DADA2 R library.'),
26     short_description='Plugin for sequence quality control with DADA2.',
27     citations=qiime2.plugin.Citations.load('citations.bib', package='q2_dada2')
28 )
29
30
31 plugin.methods.register_function(
32     function=q2_dada2.denoise_single,
33     inputs={'demultiplexed_seqs': SampleData[SequencesWithQuality | 
34                                                 PairedEndSequencesWithQuality]},
35     parameters={'trunc_len': qiime2.plugin.Int,
36                 'trim_left': qiime2.plugin.Int,
37                 'max_ee': qiime2.plugin.Float,
38                 'trunc_q': qiime2.plugin.Int,
39                 'chimera_method': qiime2.plugin.Str %
40                     qiime2.plugin.Choices(_CHIM_OPT),
41                 'min_fold_parent_over_abundance': qiime2.plugin.Float,
42                 'n_threads': qiime2.plugin.Int,
43                 'n_reads_learn': qiime2.plugin.Int,
44                 'hashed_feature_ids': qiime2.plugin.Bool},
45     outputs=[('table', FeatureTable[Frequency]),
46              ('representative_sequences', FeatureData[Sequence])],
47     input_descriptions={
48         'demultiplexed_seqs': ('The single-end demultiplexed sequences to be '
49                               'denoised.')
50     },
51     parameter_descriptions={
52         'trunc_len': ('Position at which sequences should be truncated due to '
53                       'decrease in quality. This truncates the 3\' end of the '
54                       'sequences.')
55     }
56 )
```



<https://library.qiime2.org>

Home Plugins About Resources ▾

Add new plugin

DEICODE

0.2.3

(pronounced /de.ko.de/) Robust Aitchison PCA for sparse omics datasets, linking specific features to beta-diversity ordination through the use of compositional biplots.



mmvec

v1.0.1

A software package for learning microbe-metabolite interactions.



q2-aldex2

0.1.1

Compositional differential abundance analysis. ALDEX2 provides a framework that encompasses essentially all high-throughput sequencing data types by modelling the data as a log-ratio transformed probability distribution rather than as counts.



q2-breakaway

1.0

`breakaway` is the premier package for statistical analysis of microbial diversity. `breakaway` implements the latest and greatest estimates of richness, as well as the most commonly used estimates. The `breakaway` philosophy is to estimate diversity, to put error bars on diversity estimates, and to perform hypothesis tests for diversity that use those error bars.



q2-clawback

0.0.3

Assembles taxonomic weights to increase classification accuracy with q2-feature-classifier. Can download data from Qita or use your data.



q2-coordinates

2018.11

A qiime2 plugin supporting methods for geographic mapping of qiime2 artifact data or metadata.



q2-dbotu

2018.4.2

q2-feature-classifier

2019.1

q2-fragment-insertion

2019.1

Chan
Zuckerberg
Initiative CZI project description
(source: NAU News)

What's next for QIIME 2?

QIIME 2 will help you to analyze and integrate your microbiome multi-omics data

Microbiome layers:

Taxonomy:

- Bacteria and archaea (16S rRNA surveys)
- Fungus and other eukaryotes (ITS and 18S rRNA surveys)
- Phage and other viruses (shotgun surveys)

Functional potential via shotgun metagenome surveys

Functional activity:

- Metatranscriptome
- Metaproteome
- Metabolome

Provenance “replay”: you’ll be able to generate executable code from data provenance, making it easier to re-run analyses that you (or others) have run before

qiime2view

File: taxa-bar-plots.qzv

Visualization Peek Provenance

Provenance Graph

Action Details

▼ execution:
uid: "3897fb5c-55ed-46b1-a48d-ae0651d2b597"
▼ runtime:
start: 2017-09-28T21:14:34.374Z
end: 2017-09-28T21:23:05.935Z
duration: "8 minutes, 31 seconds, and 561708 microsecond s"
▼ action:
type: "method"
plugin: "environment:plugins:dada2"
action: "denoise_single"
▼ inputs:
▼ 0:
demultiplexed_seqs: "ce7e102e-4b8c-455c-b2af-a7fb342b7fa1"
▼ parameters:
▼ 0:
trunc_len: 120
▼ 1:
trim_left: 0
▼ 2:
max_ee: 2
▼ 3:
trunc_q: 2
▼ 4:
chimera_method: "consensus"
▼ 5:
min_fold_parent_over_abundance: 1
▼ 6:
n_threads: 1

```
# This file was generated from QIIME 2 Provenance. Cool!
#
# your_pipeline.py
# -----
# def your_pipeline(ctx, table, phylogeny, sampling_depth, metadata,
#                  n_jobs_or_threads=1):
#     faith_pd = ctx.get_action('diversity_lib', 'faith_pd')
#     unweighted_unifrac = ctx.get_action('diversity_lib', 'unweighted_unifrac')
#     weighted_unifrac = ctx.get_action(
#         'diversity_lib',
#         'weighted_unifrac')
#     pcoa = ctx.get_action('diversity', 'pcoa')
#     emperor_plot = ctx.get_action('emperor', 'plot')
#     core_metrics = ctx.get_action('diversity', 'core_metrics')
#
#     cr = core_metrics(table=table, sampling_depth=sampling_depth,
#                       metadata=metadata, n_jobs=n_jobs_or_threads)
#
#     faith_pd_vector, = faith_pd(table=cr.rarefied_table,
#                                 phylogeny=phylogeny)
#
#     dms = []
#     dms += unweighted_unifrac(table=cr.rarefied_table, phylogeny=phylogeny,
#                               threads=n_jobs_or_threads)
#     dms += weighted_unifrac(table=cr.rarefied_table,
#                            phylogeny=phylogeny,
#                            threads=n_jobs_or_threads)
#
#     pcoas = []
#     for dm in dms:
#         pcoas += pcoa(distance_matrix=dm)
#
#     plots = []
#     for pcoa in pcoas:
#         plots += emperor_plot(pcoa=pcoa, metadata=metadata)
#
#     return (
#         cr.rarefied_table, faith_pd_vector, cr.observed_features_vector,
#         cr.shannon_vector, cr.evenness_vector, *dms,
#         cr.jaccard_distance_matrix, cr.bray_curtis_distance_matrix,
#         *pcoas, cr.jaccard_pcoa_results, cr.bray_curtis_pcoa_results,
#         *plots, cr.jaccard_emperor, cr.bray_curtis_emperor)
```

your_pipeline.py 7:19 LF UTF-8 Python 6x9 GitHub Git (0) 2 updates

New ways to use QIIME 2, through Galaxy and CWL interfaces, and an R API.



([Work In Progress](#))

([Ready for Testing](#))



Our R API is funded and in the planning stages
(in the meantime, try [QIIME2R](#)).

And a lot more!

 **QIIME 2 is now funded by the National Cancer Institute and lots of exciting new things are coming! 🎉**

Announcements

 **gregcaporaso** Leader 6 Aug 11

I'm extremely excited to announce that QIIME 2 is now funded by the National Cancer Institute [13](#) through the Informatics Technology for Cancer Research [2](#) program! This five-year project, entitled *Advancing our Understanding of Cancer and the Human Microbiome with QIIME 2*, provides stable support for our core development team so we can keep building and supporting QIIME 2 for you. It will also enable many new features for QIIME 2 that stand to improve your microbiome research projects.

For folks interested in the details, I provide some information below on what's to come. But first, I want to thank the entire QIIME 2 community for your support. This grant, and QIIME 2 as a whole, would not be possible without our development and support teams, our third party developers, and our user community. I look forward to continuing to work together to advance microbiome research and improve human health, environmental health and sustainability, and all of the incredible projects that folks are now using QIIME 2 for which we never could have imagined when we got started.

Thank you, thank you, thank you, as always, for your interest in and support of QIIME! 

Now, onto the details...

Aug 11
1 / 6
Aug 11

8d ago

Read this full post [here](#).

The Caporaso Lab is currently hiring software engineers, post-doctoral scholars, and graduate students to join the QIIME 2 team!

Watch the [QIIME 2 Forum job board](#) and/or [@qiime2 on Twitter](#) for announcement of these positions (and post your own job listings for free).



QIIME 2 types and formats

<https://bit.ly/2HThBcx>

We've all gotten emails like this...

New | Delete | Archive | Junk | Sweep | Move to | Categories | ...

Inbox Filter ▾

Alan R Dale
PMI servers are updated and available 9:37 AM
All Windows and LINUX servers are back online. All serv...
Yesterday

Evan Thomas Bolyen URGENT: Analysis needed by Tuesday 2PM Sat 12/15
Hi Matt, Hope everything is going well with you. I was ju...

Lyft Ride Receipt Your ride with Arlady on December 15 Sat 12/15
Thanks for riding with Arlady! December 15, 2018 at 10:2...

Last week

Lyft Ride Receipt Your ride with Paa on December 14 Fri 12/14
Thanks for riding with Paa! December 14, 2018 at 8:46 P...

Lyft Ride Receipt Your ride with Michael on December 14 Fri 12/14

URGENT: Analysis needed by Tuesday 2PM

Evan Thomas Bolyen Yesterday, 2:56 PM Matthew Ryan Dillon ▾

 data.zip 282 bytes

Download

Action Items

Hi Matt,
Hope everything is going well with you. I was just wondering if you could analyze this data (attached) really quick?
I need it by Tuesday at 2pm. There's a few files in here, but I just need you to check into the Sample11EByTotal one.
Thanks so much!
-Evan

	A	B	C
s1	0	15	10
s2	10	5	0
s3	1	5	4

We've all gotten emails like this...

New | Delete | Archive | Junk | Sweep | Move to | Categories | ...

Inbox Filter ▾ URGENT: Analysis needed by Tusday 2PM

Alan R Dale
PMS servers are updated :
All Windows and LINUX ser

Yesterday

Evan Thomas Boley
URGENT: Analysis needed
Hi Matt, Hope everything is

Lyft Ride Receipt
Your ride with Arlady on I
Thanks for riding with Arlac

Last week

Lyft Ride Receipt
Your ride with Paa on December 14
Fri 12/14
Thanks for riding with Paa! December 14, 2018 at 8:46 P...

Lyft Ride Receipt
Your ride with Michael on December 14
Fri 12/14

	A	B	C
s1	0	15	10
0	5	0	
1	5	4	

We are missing
important context!

I need it by Tuesday at 2pm. There's a few files in here, but I just need you to check into the Sample11EByTotal one.
Thanks so much!
-Evan

ze this data (attached) really quick?

Impacts from missing context in email

Scenario 1:

Attempt to interpret results but do it incorrectly by running invalid commands.

Impacts from missing context in email

Scenario 1:

Attempt to interpret results but do it incorrectly by running invalid commands.

incorrect conclusions

Impacts from missing context in email

Scenario 2:

Go back to Evan to get more details.

Impacts from missing context in email

Scenario 2:

Go back to Evan to get more details.

miss the deadline

Impacts from missing context in email

Scenario 3:

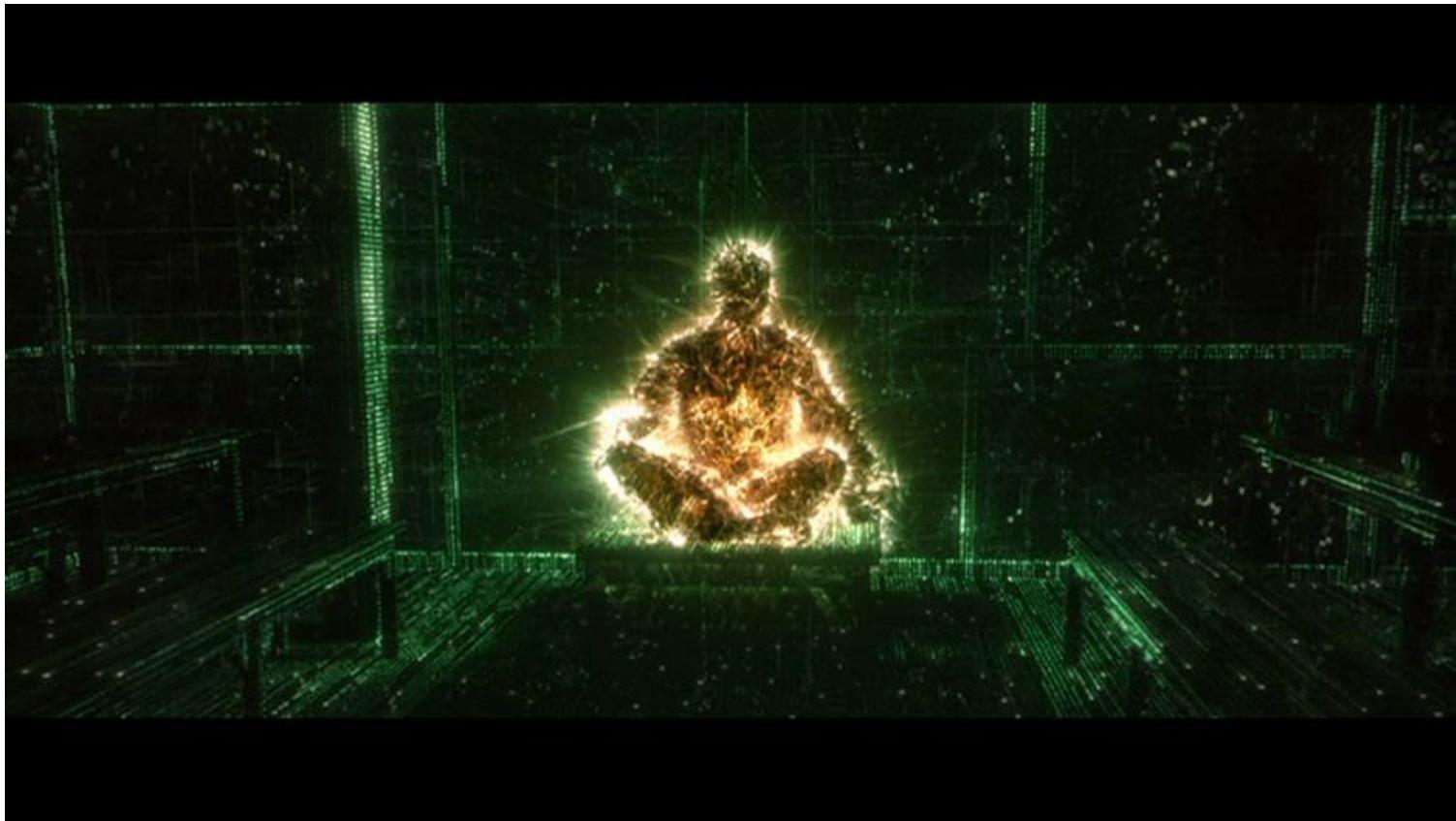
Do what you usually do with these data.

Impacts from missing context in email

Scenario 3:

Do what you usually do with these data.

not using the latest techniques available



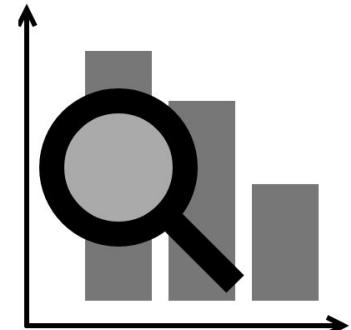
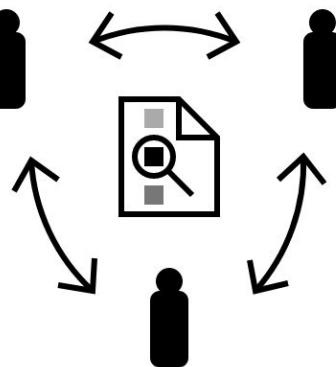
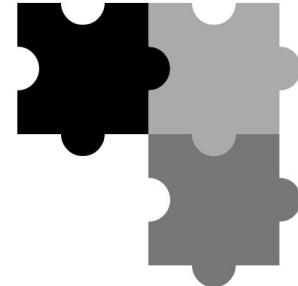
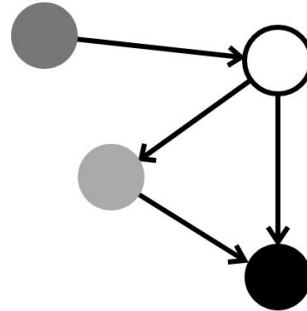
Still from the movie "The Matrix Revolutions"

Computers are not clever - they do *exactly* as they are instructed, no more and no less.



Still from the movie "The Matrix Revolutions"

What kinds of opportunities are created when we are able to provide more context to the computer?



Prevent you from providing the wrong input to a command.

Scenario 1:

Attempt to interpret results but do it incorrectly by running invalid commands.

incorrect conclusions

Scenario 2:

Go back to Evan to get more details.

The contextual details
are encoded in the
data.

miss the deadline

A search engine where you can ask “what can I do with this type of data?”

Scenario 3:

Do what you usually do with these data.

not using the latest techniques available

Context in QIIME 2 via Semantic Types

Semantic Types are a vocabulary for describing what the data “is”

FeatureTable[Frequency]

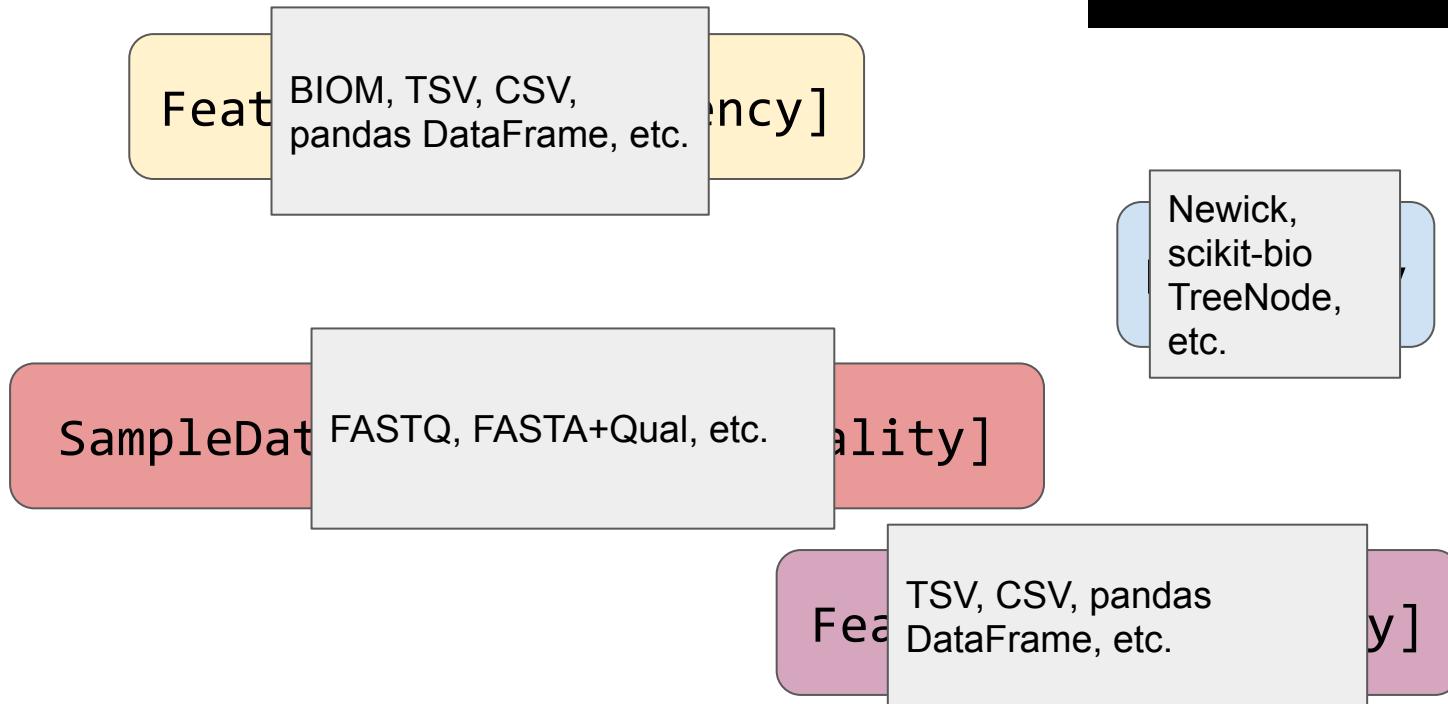
Phylogeny

SampleData[SequencesWithQuality]

FeatureData[Taxonomy]

Data Formats in QIIME 2

Data Formats are a way that data can be represented

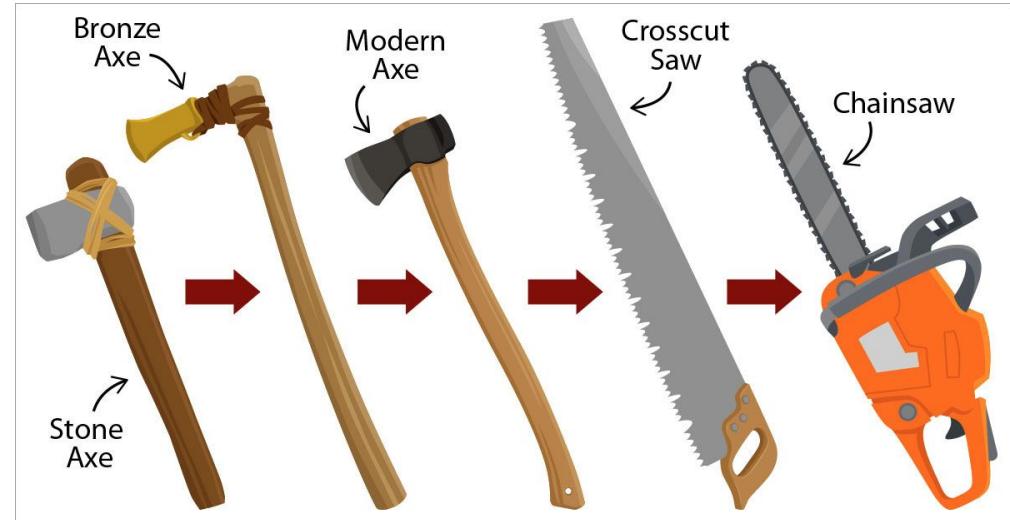


Semantic Types keep the *formatting* separate from the *meaning*

Formats can change (but the semantics remain)

- Formats improve over time (BIOM 1.0.0 vs BIOM 2.1.0)
- Backwards compatible - data you generate today will work with future versions of QIIME 2

HandTool [Woodcutting]

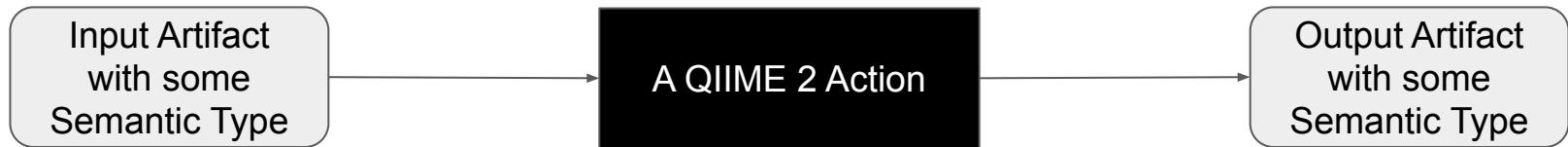


<http://devicedesigns.com/work/tool-evolution>

What are some *formats* that might represent this type?

There is NO “one/right way”
to do a QIIME 2 analysis

How commands work in QIIME 2



QIIME Zipped Artifact

QIIME 2 uses Artifacts and Visualizations

- Where did my FASTA file go?
 - It's inside a ZIP file

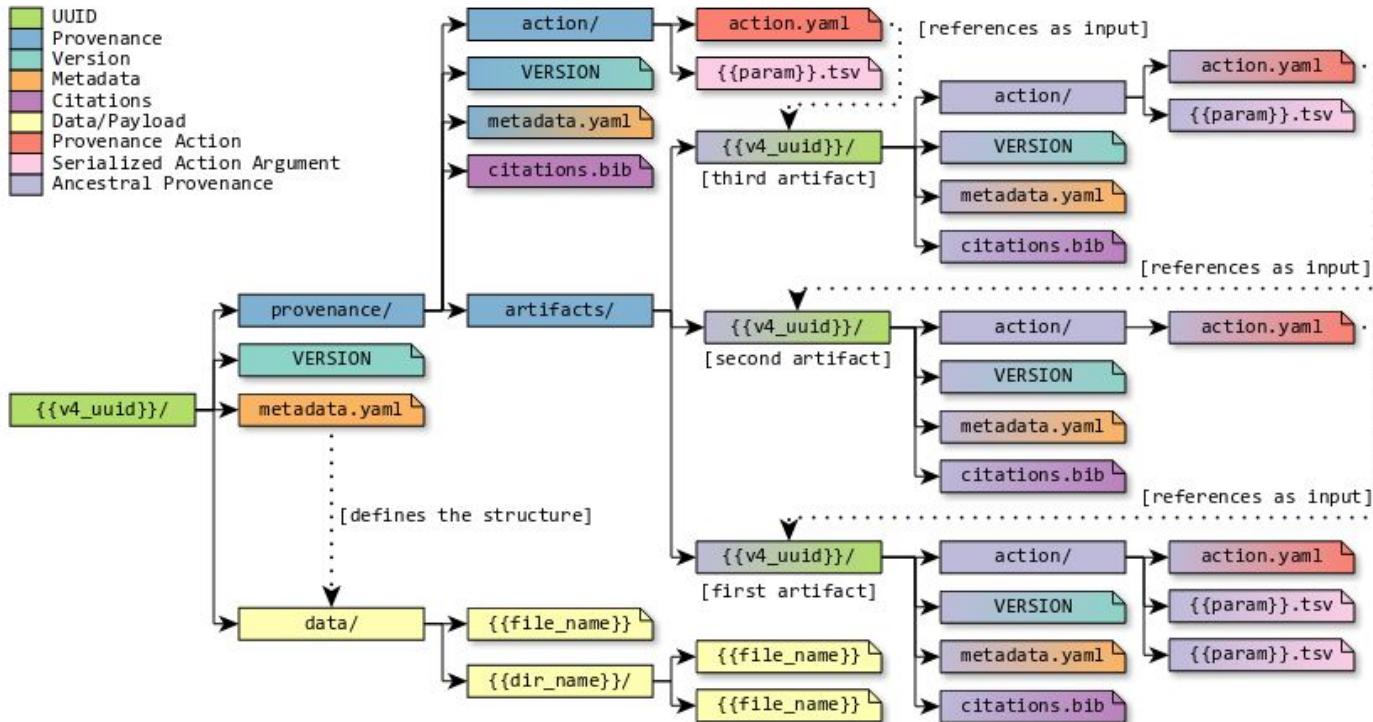
.qza and .qzv files are just zip files

- We aren't inventing new formats
 - A QIIME 2 result can be "used" on almost any computer
- ZIP64 (OS X by default only knows ZIP32, but other OSes work well)
 - Use `qiime tools extract` if unzipping isn't working.

Learn More:

<https://dev.qiime2.org/latest/storing-data/>

Archive Structure



Sample Metadata

<https://bit.ly/2HThBcx>

Performing a microbiome study

Study design
(power, randomization)
(metadata--standard + specific to your study)

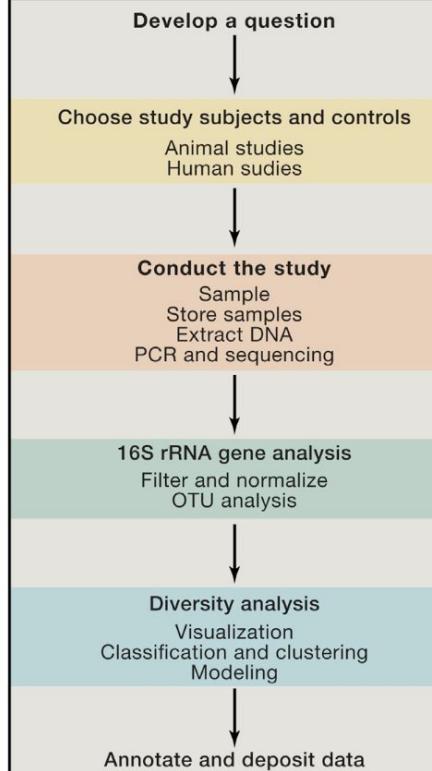
Sample collection
(method, shipping/storage)
lab protocols
(include neg/pos controls)

Data processing

Data analysis

For more on these topics, see:

- [Conducting a Microbiome Study](#), by Goodrich et al. 2014
- [Reagent Contamination](#), Salter et al. 2014
- [Storage effects](#), by Song et al. 2016
- [Microbiome Quality Control \(MBQC\)](#), by Sinha et al. 2017
- [MIMARKS](#), by Yilmaz et al. 2011
- [KatharoSeq low biomass workflow](#), by Minich et al. 2017



Sample Information

Sequence Data

Feature Table

Summary Statistics

Other resources:

- [Earth Microbiome Project website](#)
- [Human Microbiome Project website](#)
- [American Gut Project website](#)

Sample metadata: what to include

Things you think may affect your hypothesis

- Main Exposure*
- Potential biological confounders
- Potential technical confounders

*carefully consider temporality in a case-control paradigm

Sample metadata: what to include

Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIS) specifications.

Yilmaz et al. Nature Biotechnology 29, 415–420 (2011).

Specification projects	MIGS					MIMS	MIMARKS		New checklists
Checklists	EU	BA	PL	VI	ORG	metagenomes	survey	specimen	e.g., pan-genomes
Shared descriptors	collection date, environmental package, environment (biome), environment (feature), environment (material), geographic location (country and/or sea, region), geographic location (latitude and longitude), investigation type, project name, sequencing method, submitted to INSDC								
Checklist-specific descriptors	assembly, estimated size, finishing strategy, isolation and growth condition, number of replicons, ploidy, propagation, reference for biomaterial			target gene					
Applicable environmental packages (measurements and observations)	Air Host-associated Human-associated Human-oral Human-gut Human-skin Human-vaginal			Microbial mat/biofilm Miscellaneous natural or artificial environment Plant-associated Sediment Soil Wastewater/sludge Water					

Sample Metadata: how to format it

id	Description
s1	pd1 c1 m1-7
s2	pd2 c4 m13-7

id	donor	cage	mouse	day
#q2:types	categorical	categorical	categorical	numeric
s1	PD1	c1	m1	7
s2	PD2	c2	m13	7

Anything else to worry about?

Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics

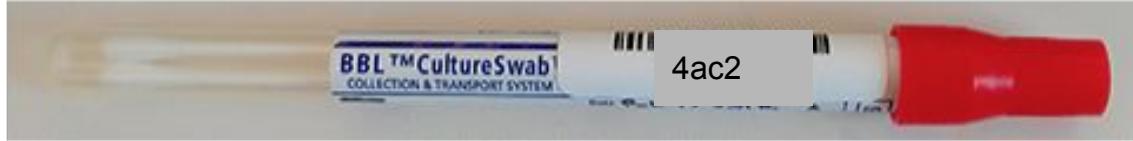
Barry R Zeeberg[†], Joseph Riss[†], David W Kane, Kimberly J Bussey, Edward Uchio, W Marston Linehan, J Carl Barrett and John N Weinstein 

[†] Contributed equally

BMC Bioinformatics 2004 5:80 | DOI: 10.1186/1471-2105-5-80 | © Zeeberg et al; licensee BioMed Central Ltd. 2004

Received: 05 March 2004 | Accepted: 23 June 2004 | Published: 23 June 2004

Collect samples and metadata



sample-metadata.tsv				
sample_name	donor_id	genotype	mouse_id	days_post_transplant
4ac2	hc_1	wild type	m1	7
e375	hc_1	wild type	m1	14
4gd8	hc_1	susceptible	m2	7
9872	pd_1	wild type	m3	7

Choose your sample ids carefully! See [cual-id](#) (Chase et al., 2015) for help with this.

FAQ: Metadata

1. Do I really need it?
2. Where does it come from?
3. Are there any hard requirements?

FAQ: Metadata

1. Do I really need it?

Yes!

2. Where does it come from?

Hopefully its collected at the same time that you collect your samples

3. Are there any hard requirements?

TSV format

QIIME 2 requires a sample identifier column of the form

{‘#SampleID’, ‘sample_name’, ‘#Sample ID’ “id”, ‘sampleid’ ‘sample-id’, ‘sampleid’}

Sample metadata: what to include

Things you think may affect your hypothesis

- Main Exposure*
- Potential biological confounders
- Potential technical confounders

Categoricals and continuous variables are better than long strings

id	Description
s1	pd1 c1 m1-7
s2	pd2 c4 m13-7

VS

id	donor	cage	mouse	day
s1	PD1	c1	m1	7
s2	PD2	c2	m13	7

*rarely the microbiome in a case-control paradigm

Command Line Refresher

<https://bit.ly/2HThBcx>

What you'll be refreshed on

- File system concepts: relative and absolute paths, your home directory, the root directory
- Important built-in commands
- Shortcuts: tab, arrow up, arrow down, CTRL-A, CTRL-E

File paths specify the location of a file on your computer.

Directory paths specify the location of a *directory* (also called a *folder*) on your computer.

These *paths* can either be:

- *relative* to your current location on the file system, or
- *absolute*, meaning they are not relative to your location on the file system but rather fully specified with respect to the root of the file system.

The *root directory* is a special directory that contains all of the other files and directories on the computer. The root directory is specified as a / at the beginning of a path.

Absolute paths are specified starting with the root directory.
The following are absolute paths:

/

/home/greg

/home/greg/chicken-photos/barbara.png

The location specified by an absolute path doesn't change meaning, regardless of what directory you're currently in on the file system.

The location specified by relative paths does change meaning depending on your current location. The following are relative paths:

chicken-photos/

chicken-photos/barbara.png

Notice that these do not start with a /

Your *home directory* is an important directory on the file system. This is generally where you will store your files. It is also the directory you will be in when you log into the server.

On Linux based systems the absolute path to your home directory is:

/home/username/

where **username** stands for your user in your machine.

You may see your home directory abbreviated in a few ways:

~

\$HOME

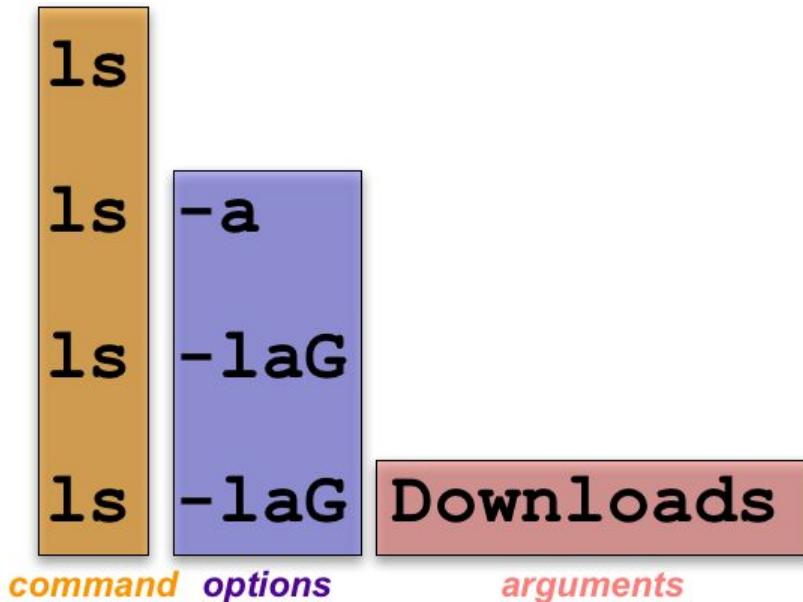
Since these are abbreviations of an absolute path, they work like absolute paths.



Important built-in commands

pwd	print working (i.e., current) directory
ls	list directory contents
cd	change directory
mkdir	make directory
nano	open a text editor
head <i>filepath</i>	print the first ten lines of a file to the terminal
tail <i>filepath</i>	print the last ten lines of a file to the terminal
less <i>filepath</i>	interactively print contents of a file to the terminal
tree	visualize directories, recursively
mv <i>old-path new-path</i>	move (or rename) a directory or a file
cp <i>existing-path new-path</i>	copy a directory or a file
rm <i>filepath</i>	remove a file - be careful, this is permanent!
rm <i>-r directory-path</i>	remove a directory - be careful, this is permanent!

Anatomy of a command



Some more advanced commands that we'll use in this workshop

`curl` or `wget` download a file from the Internet given its location (i.e., it's URL)

`qiime` use the QIIME 2 program through its command line interface

Resources for practicing or learning more.

Learning the command line (or even learning to program) is empowering and will increase your value to potential employers.

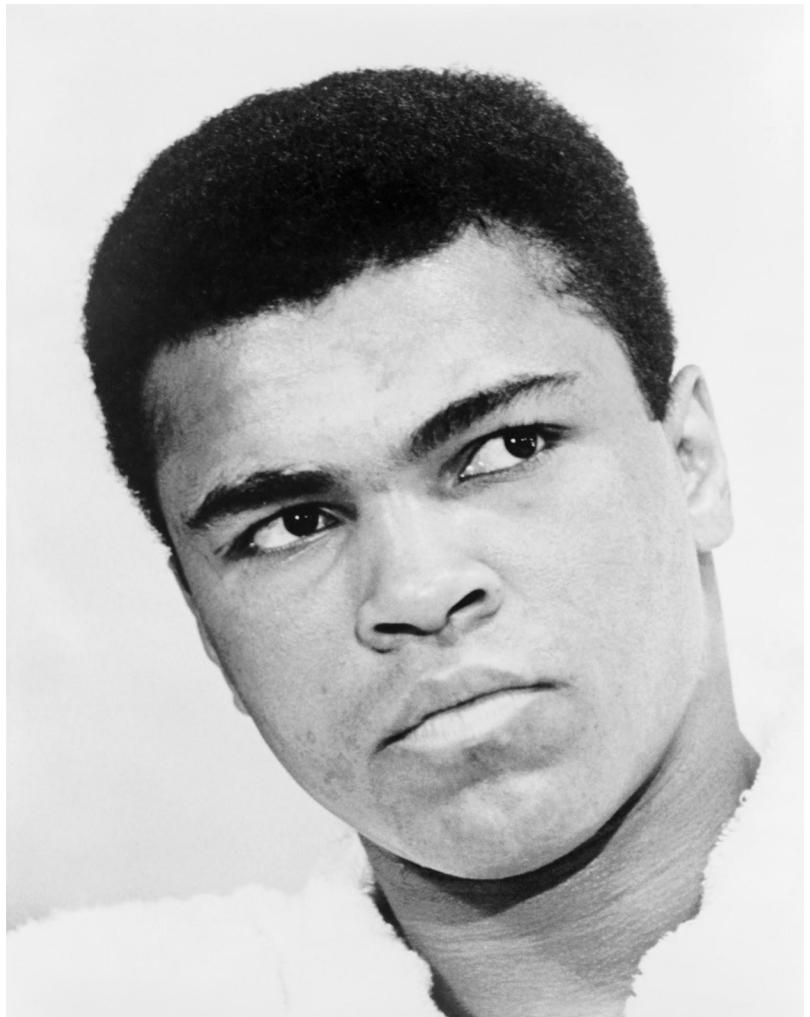
You can learn to program! Like learning to play a musical instrument, it just takes time.

Use lessons from [The Carpentries](#) (such as [their Unix Shell lesson](#)) or attend one of their many workshops.

See [Josylnn Lee's lesson](#) (developed for a QIIME 2 workshop).

Study background and information

<https://bit.ly/2HThBcx>



Left: World Journal Tribune photo by Ira Rosenberg; Right: Thomas Atilla Lewis

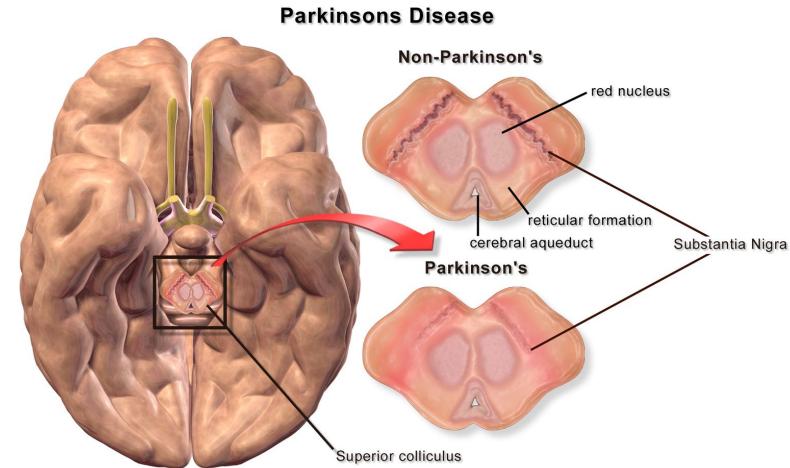
Parkinson's Disease

1 million Americans

96% diagnosed after 50

Motor Impairment

Non motor Symptoms





Gut Microbiota Are Related to Parkinson's Disease and Clinical Phenotype

Filip Schepersjans, MD, PhD,^{1*} Velma Aho, MSc, BA,² Pedro A. B. Pereira, MSc,² Kaisa Koskinen, PhD,² Lars Paulin, MSc,² Eero Pekkonen, MD, PhD,¹ Elena Haapaniemi, MD, PhD,¹ Seppo Kaakkola, MD, PhD,¹ Johanna Eerola-Rautio, MD, PhD,¹ Marjatta Pohja, MD, PhD,¹ Esko Kinnunen, MD, PhD,³ Kari Murros, MD, PhD,¹ and Petri Auvinen, PhD²



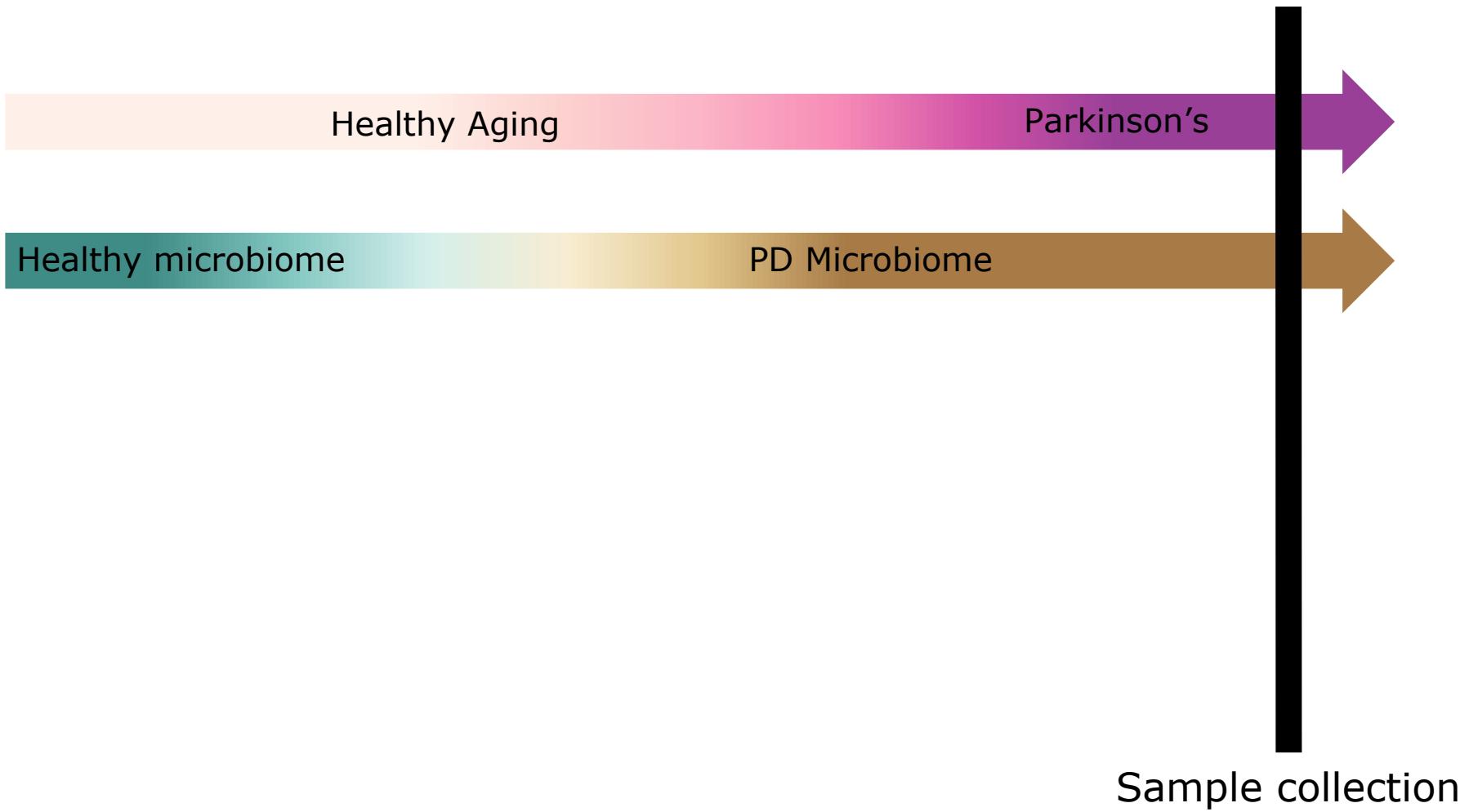
Colonic Bacterial Composition in Parkinson's Disease

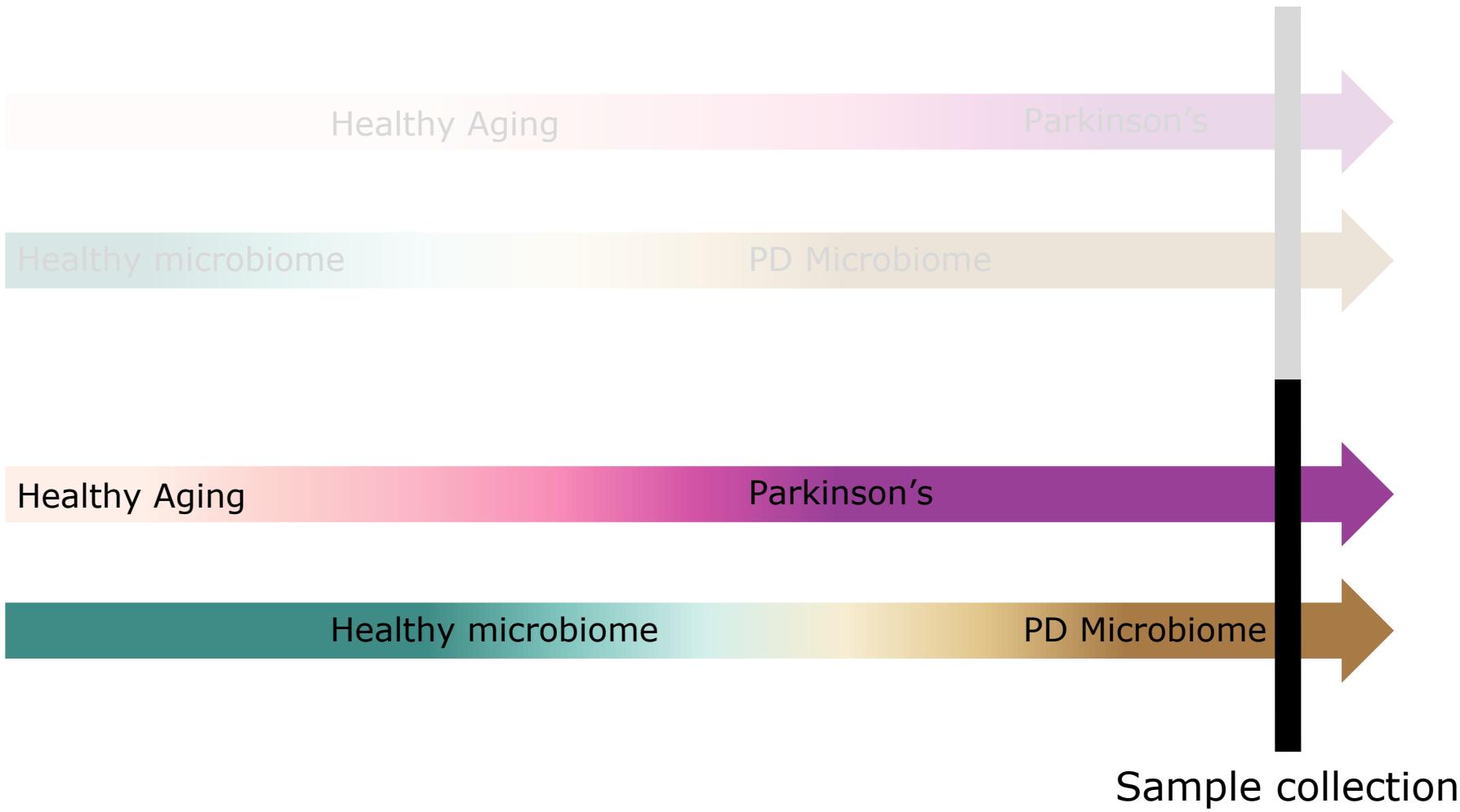
Ali Keshavarzian, MD,^{1,6,7,8*} Stefan J. Green, PhD,^{3,4} Phillip A. Engen, BS,¹ Robin M. Voigt, PhD,¹ Ankur Naqib, BS,³ Christopher B. Forsyth, PhD,^{1,5} Ece Mutlu, MD,¹ and Kathleen M. Shannon, MD²

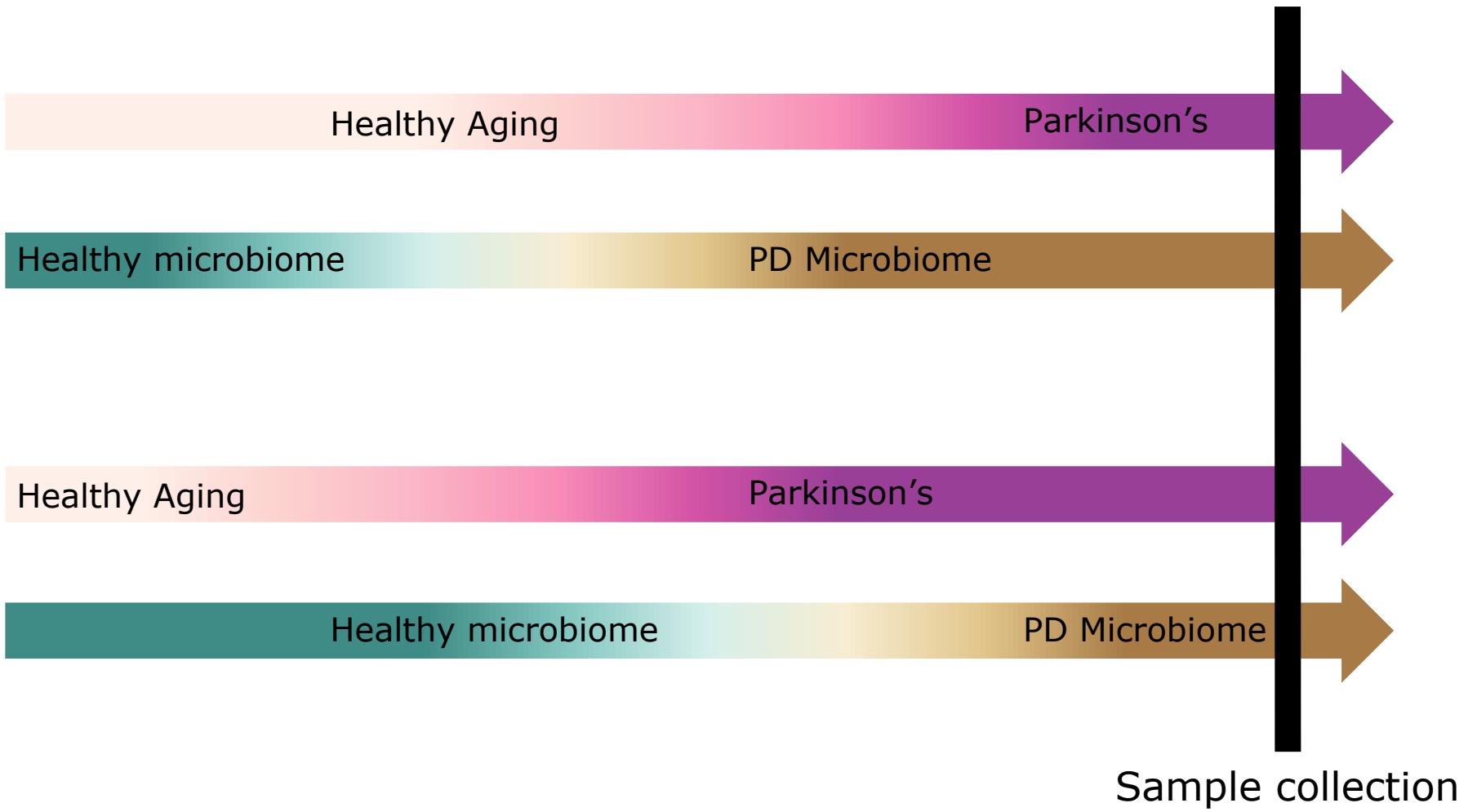


Parkinson's Disease and Parkinson's Disease Medications Have Distinct Signatures of the Gut Microbiome

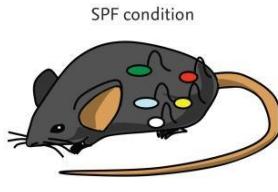
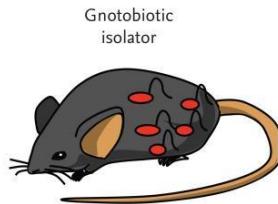
Erin M. Hill-Burns, PhD,¹ Justine W. Debelius, PhD,² James T. Morton, BS,³ William T. Wissemann, BA,¹ Matthew R. Lewis, MS,¹ Zachary D. Wallen, MS,¹ Shyamal D. Peddada, PhD,⁴ Stewart A. Factor, DO,⁵ Eric Molho, MD,⁶ Cyrus P. Zabetian, MD, MS,⁷ Rob Knight, PhD,^{2,3,8} and Haydeh Payami, PhD^{1,9*}





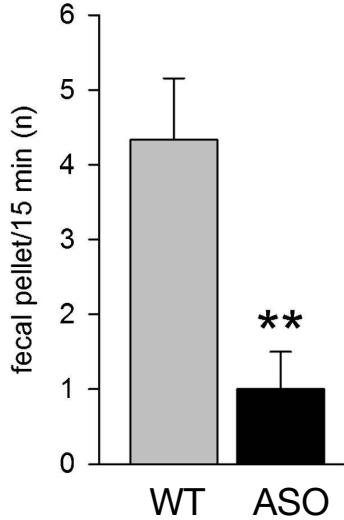
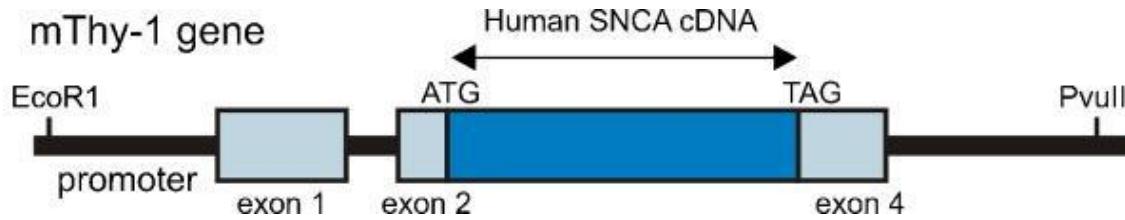
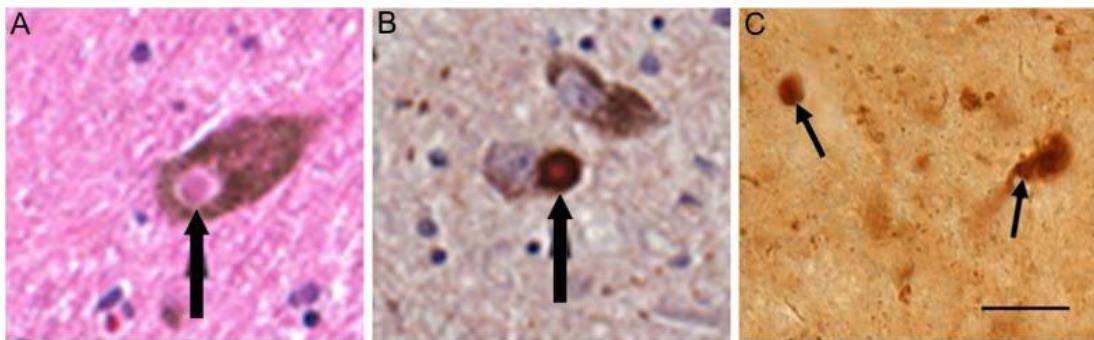


A system for testing microbes



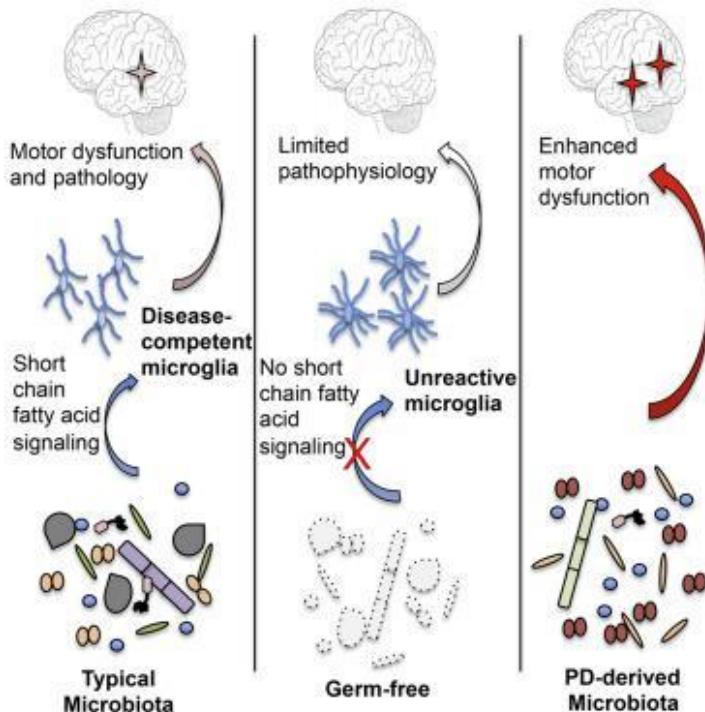
Mouse model of PD

Thy1-aSyn
(ASO)



Rockenstein et al (2002). *J Neurosci Res.* **68**:568-678;
Chesselet et al (2012) *Neurotherapeutics* **9**:296-314

Original Study



Volume 167, Issue 6, 1 December 2016, Pages 1469-1480.e12

Article

Gut Microbiota Regulate Motor Deficits and Neuroinflammation in a Model of Parkinson's Disease

Timothy R. Sampson ¹ Justine W. Debelius ², Taren Thron ¹, Stefan Janssen ², Gauri G. Shastri ¹, Zehra Esra Ilhan ³, Collin Challis ¹, Catherine E. Schretter ¹, Sandra Rocha ⁴, Viviana Grardinaru ¹, Marie-Francoise Chesselet ⁵, Ali Keshavarzian ⁶, Kathleen M. Shannon ^{7, 9}, Rosa Krajmalnik-Brown ³, Pernilla Wittung-Stafshede ⁴, Rob Knight ^{2, 8}, Sarkis K. Mazmanian ^{1, 10}

[Show more](#)

12 donors (6 Parkinsons patients, 6 HC)
2 genetic backgrounds
4 timepoints per mouse

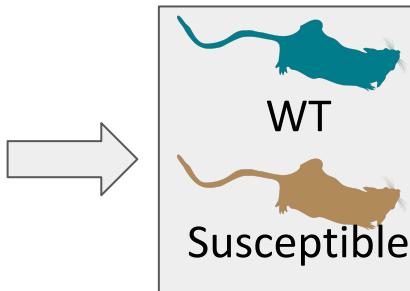
Humanized Mice Experiment



Healthy
Control



PD Patient

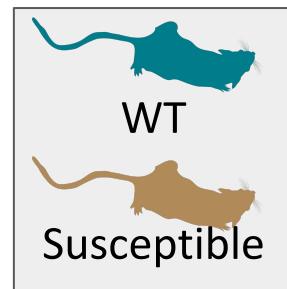


x 3

12 donors: 6 healthy, 6 with parkinson's disease

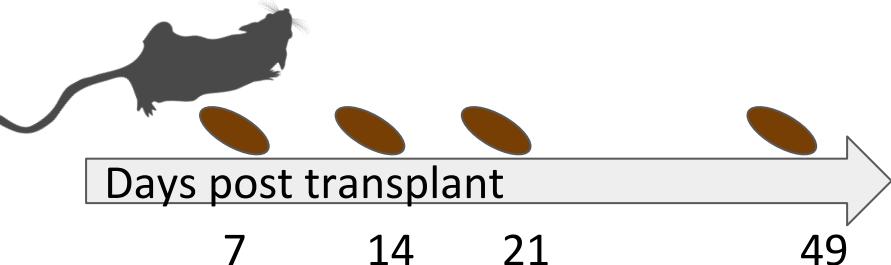
2 genotypes: susceptible and PD

3 cages per donor with mixed mice



x 3

4 time points per mouse



Our tutorial

Hypothesis:

The genetic background of a recipient mouse shapes its fecal community

Data:

2 donors (1 PD, 1 healthy)

6 mice/donor (3 susceptible, 3 WT/donor)

4 timepoints/mouse over 7 weeks

Importing, demultiplexing

<https://bit.ly/2HThBcx>



Importing and Demultiplexing

Mehrbod Estaki

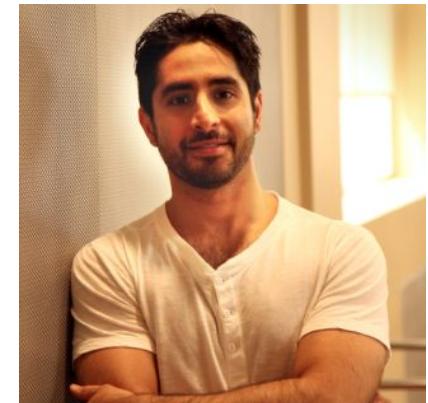
Postdoctoral Researcher

Knight Lab

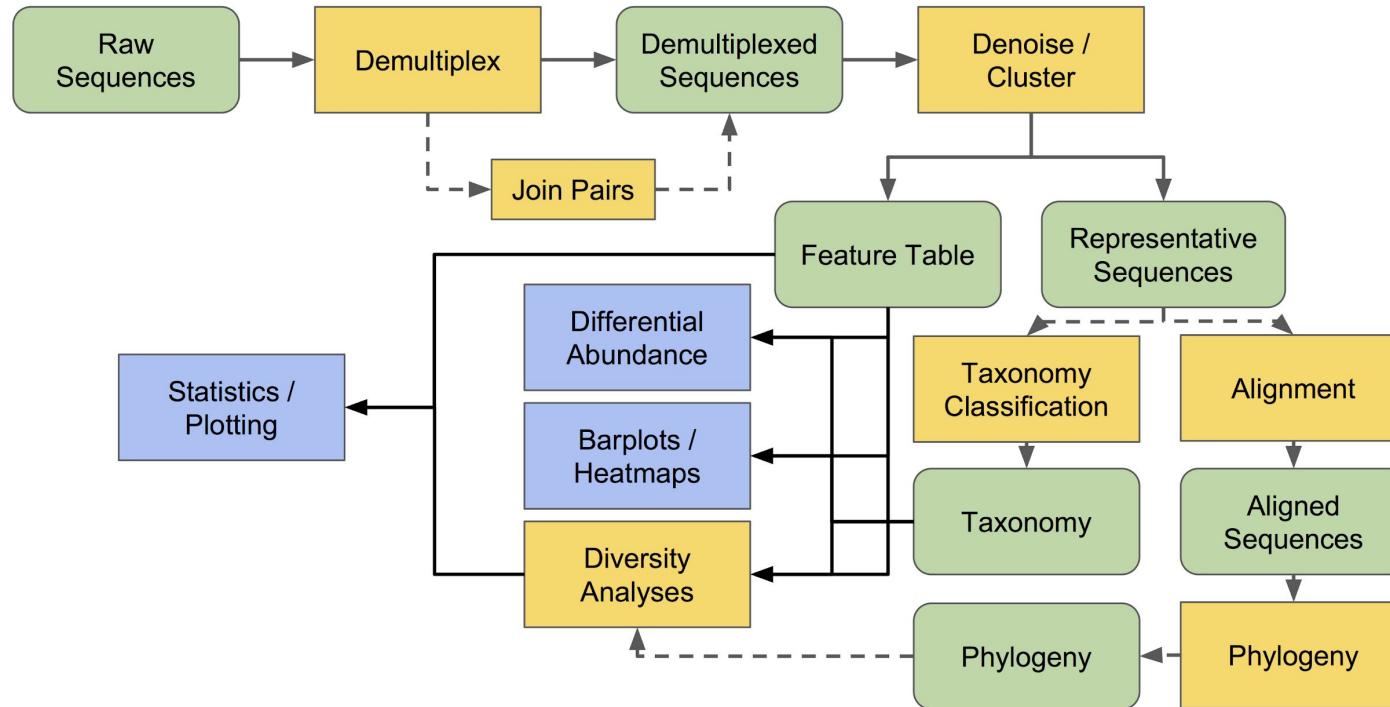
University of California San Diego

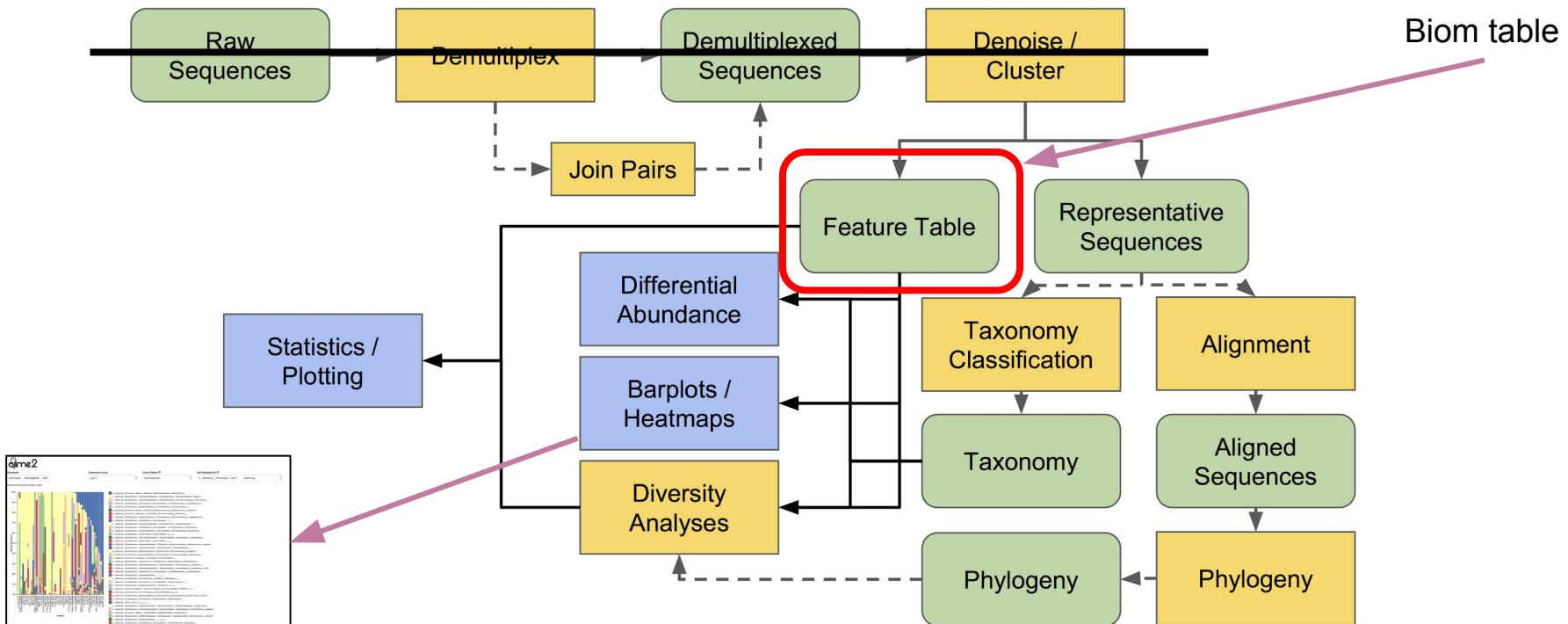
 @MehrbodEstaki

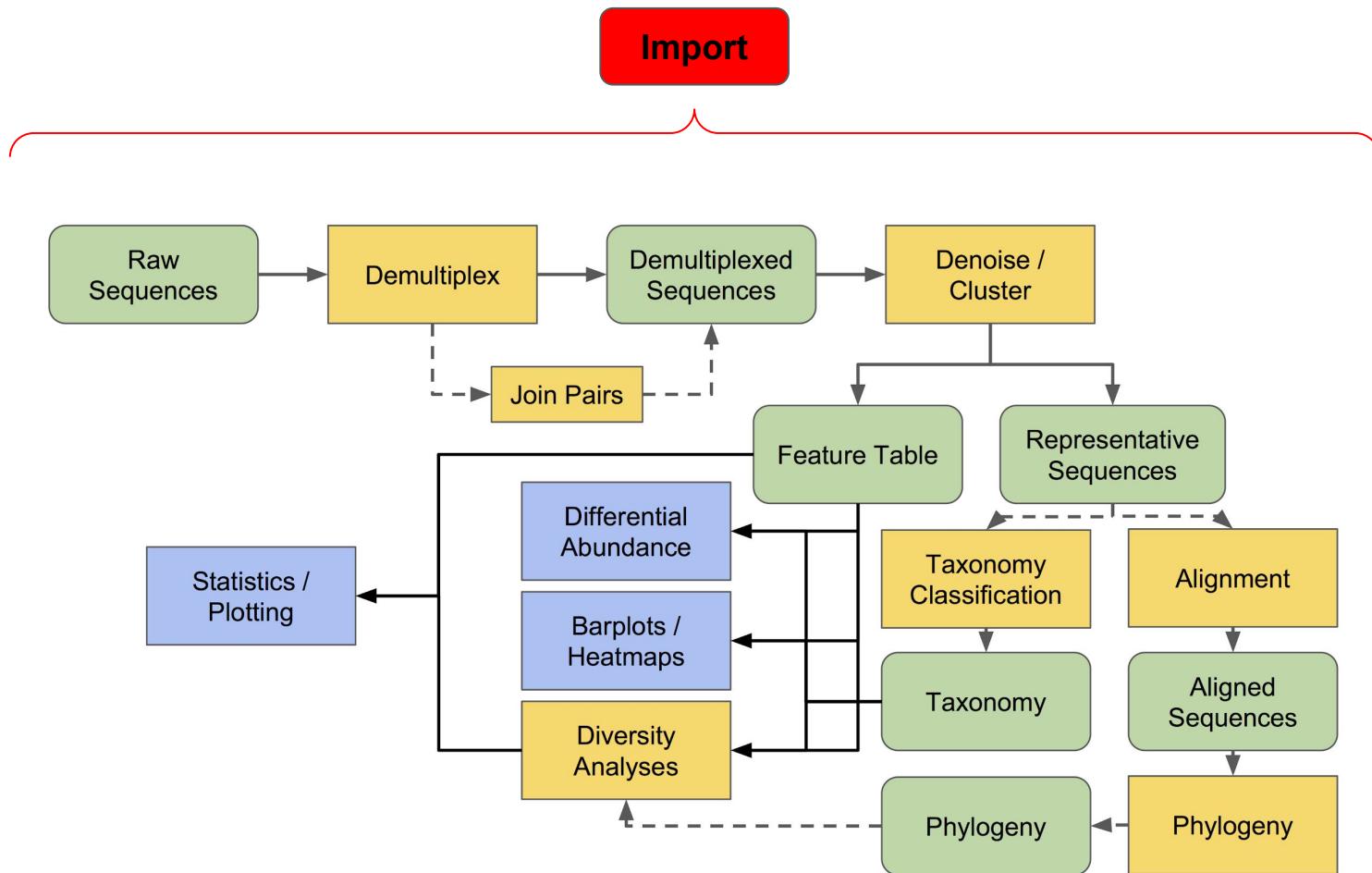
 <https://mestaki.wordpress.com>

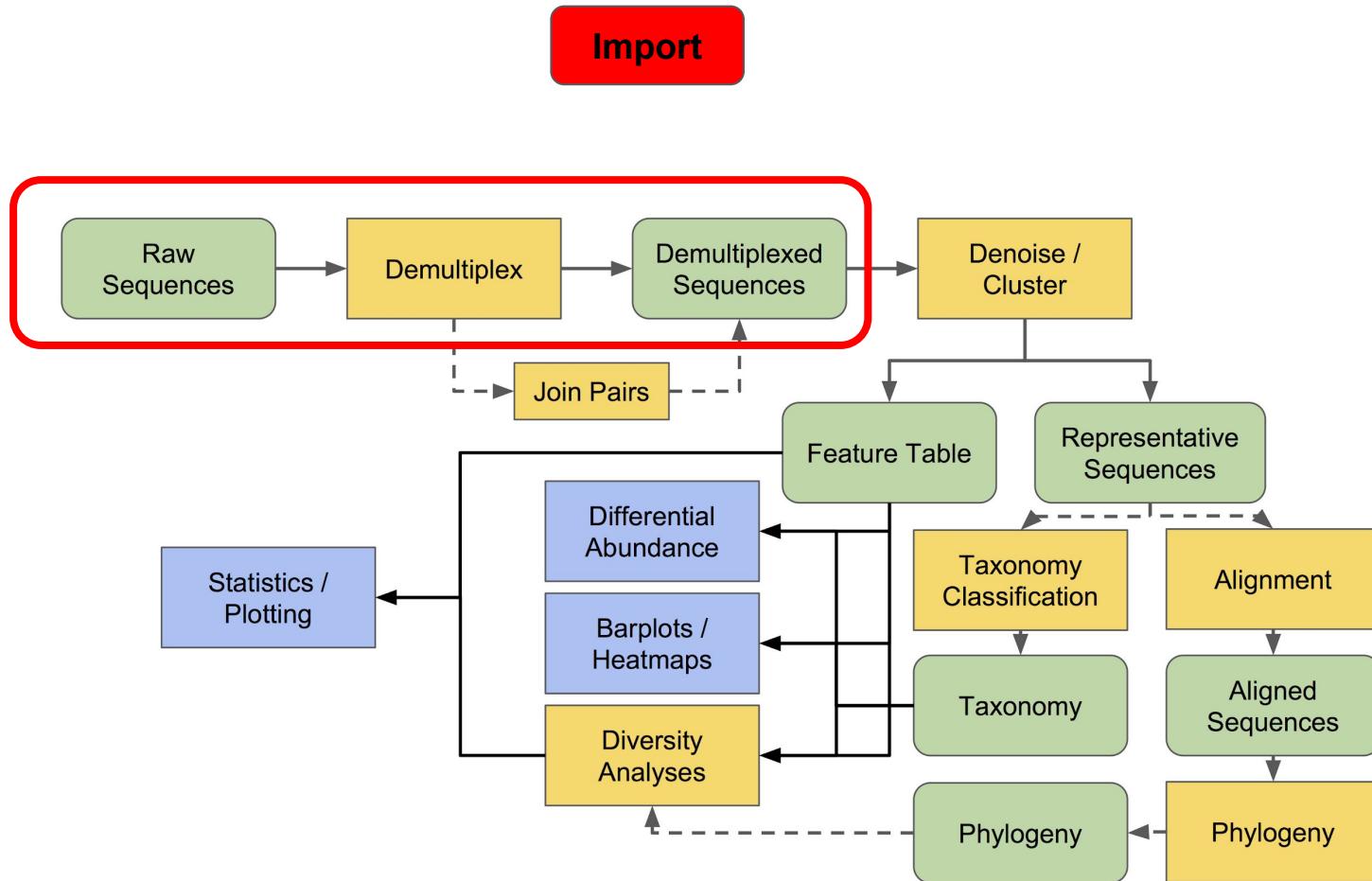


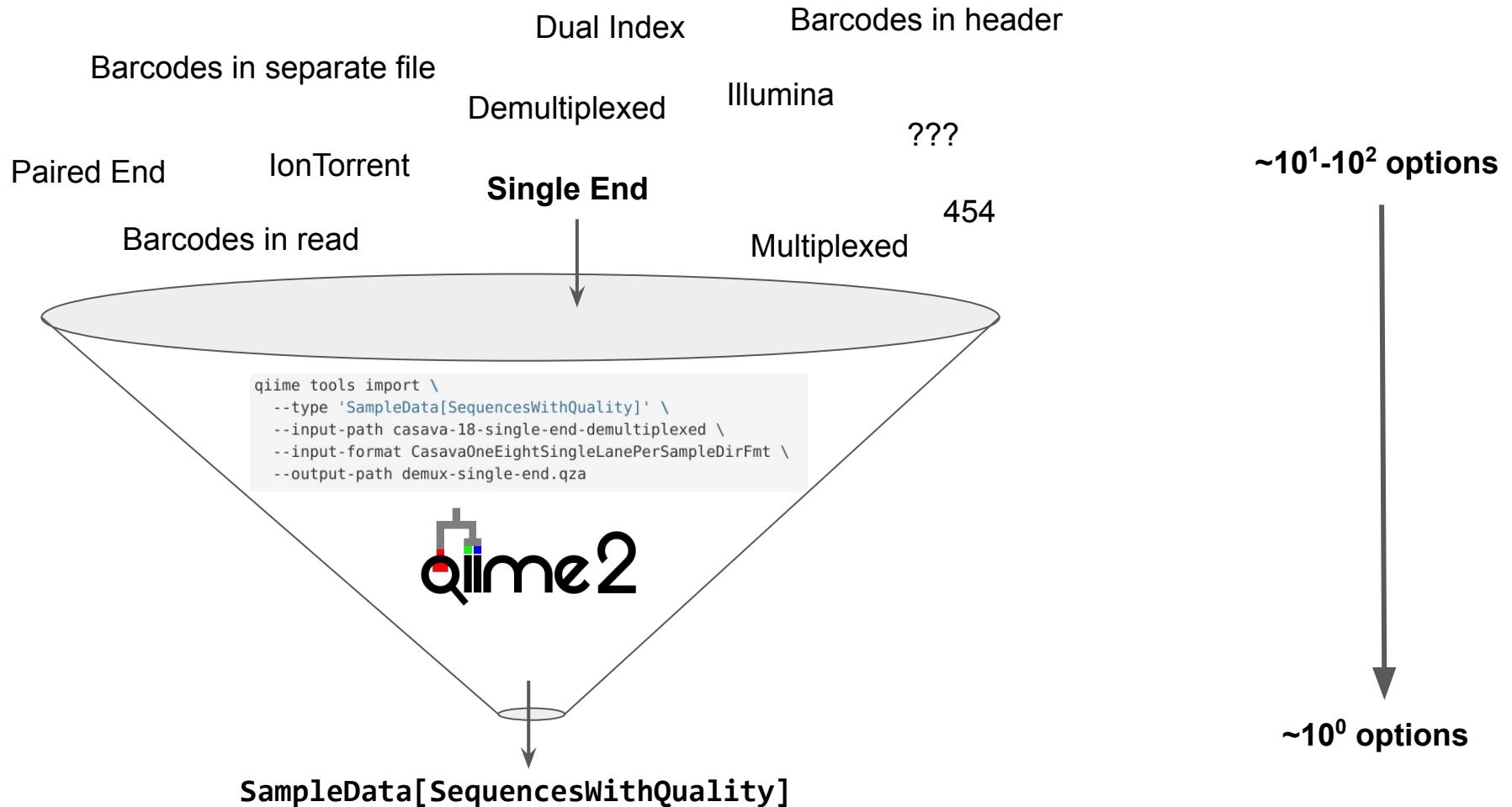
UC San Diego











Importable types

Bowtie2Index
DeblurStats
DistanceMatrix
EMPPairedEndSequences
EMPSingleEndSequences
ErrorCorrectionDetails
FeatureData[AlignedSequence]
FeatureData[Conditional]
FeatureData[Differential]
FeatureData[Importance]
FeatureData[PairedEndSequence]
FeatureData[Sequence]
FeatureData[Taxonomy]
FeatureTable[Balance]
FeatureTable[Composition]
FeatureTable[Frequency]
FeatureTable[PercentileNormalized]
FeatureTable[PresenceAbsence]
FeatureTable[RelativeFrequency]
Hierarchy
MultiplexedPairedEndBarcodeInSequence
MultiplexedSingleEndBarcodeInSequence
PCoAResults
Phylogeny[Rooted]
Phylogeny[Unrooted]
Placements
QualityFilterStats
RawSequences
SampleData[AlphaDiversity]
SampleData[BooleanSeries]
SampleData[ClassifierPredictions]
SampleData[DADA2Stats]

SampleData[FirstDifferences]
SampleData[JoinedSequencesWithQuality]
SampleData[LogRatios]
SampleData[PairedEndSequencesWithQuality]
SampleData[Probabilities]
SampleData[RegressorPredictions]
SampleData[SequencesWithQuality]
SampleData[Sequences]
SampleData[SongbirdStats]
SampleEstimator[Classifier]
SampleEstimator[Regressor]
SeppReferenceDatabase
TaxonomicClassifier
UchimeStats

Importable formats

AlignedDNAFASTAFormat
AlignedDNASEquencesDirectoryFormat
AlphaDiversityDirectoryFormat
AlphaDiversityFormat
BIOMV100DirFmt
BIOMV100Format
BIOMV210DirFmt
BIOMV210Format
BooleanSeriesDirectoryFormat
BooleanSeriesFormat
Bowtie2IndexDirFmt
CasavaOneEightLanelessPerSampleDirFmt
CasavaOneEightSingleLanePerSampleDirFmt
ConditionalDirFmt
ConditionalFormat
DADA2StatsDirFmt
DADA2StatsFormat
DNAFASTAFormat
DNASEquencesDirectoryFormat
DeblurStatsDirFmt
DeblurStatsFmt
DifferentialDirectoryFormat
DifferentialFormat
DistanceMatrixDirectoryFormat
EMPPairedEndCasavaDirFmt
EMPPairedEndDirFmt
EMPSingleEndCasavaDirFmt
EMPSingleEndDirFmt
ErrorCorrectionDetailsDirFmt
FastqGzFormat
FirstDifferencesDirectoryFormat
FirstDifferencesFormat
HeaderlessTSVTaxonomyDirectoryFormat
HeaderlessTSVTaxonomyFormat
ImportanceDirectoryFormat
ImportanceFormat
LSMatFormat
LogRatiosDirFmt
LogRatiosFormat
MultiplexedPairedEndBarcodeInSequenceDirFmt
MultiplexedSingleEndBarcodeInSequenceDirFmt
NewickDirectoryFormat
NewickFormat
OrdinationDirectoryFormat
OrdinationFormat
PairedDNASEquencesDirectoryFormat
PairedEndFastqManifestPhred33
PairedEndFastqManifestPhred33V2
PairedEndFastqManifestPhred64
PairedEndFastqManifestPhred64V2
PlacementsDirFmt
PlacementsFormat
PredictionsDirectoryFormat
PredictionsFormat
ProbabilitiesDirectoryFormat
ProbabilitiesFormat
QIIME1DemuxDirFmt
QIIME1DemuxFormat
QualityFilterStatsDirFmt
QualityFilterStatsFmt
SampleEstimatorDirFmt
SeppReferenceDirFmt
SingleEndFastqManifestPhred33
SingleEndFastqManifestPhred33V2
SingleEndFastqManifestPhred64
SingleEndFastqManifestPhred64V2
SingleLanePerSamplePairedEndFastqDirFmt
SingleLanePerSampleSingleEndFastqDirFmt
SongbirdStatsDirFmt
SongbirdStatsFormat
TSVTaxonomyDirectoryFormat
TSVTaxonomyFormat
TaxonomicClassifierTemporaryPickleDirFmt
UchimeStatsDirFmt
UchimeStatsFmt

Importing and Demultiplexing Sequence Data Quick Reference

Community Contributions  Tutorials 



Nicholas_Bokulich  Leader

Mar 11

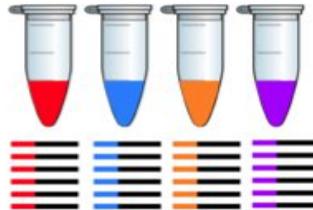
The following is a “pocket guide” to determining the appropriate methods for importing and demultiplexing FASTA/FASTQ sequences (primarily from marker-gene sequencing experiments). Many samples are often “multiplexed” (pooled and sequenced together) on a single sequencing run. So the first step to analyzing these data is to [demultiplex](#)  the data. Often, demultiplexing is performed automatically by sequencing centers/services and users. How can you tell if you have demultiplexed data? If you have one sequence file (or pair of files) per sample. The following steps are meant as a guide to determine appropriate steps and tutorials for importing (and demultiplexing) sequence data.

If you find errors in this guide, or want to provide steps for importing or demultiplexing other formats, get in touch!

- Do you have **demultiplexed FASTQ sequences**? (i.e., one file per sample)
 - Yes
 - Are the sequences in CASAVA 1.8 format? (see descriptions of [CASAVA 1.8](#)  and other [FASTQ formats](#)  for details)
 - Yes: [use the CASAVA](#)  format to import
 - No: [Use the appropriate Manifest format](#)  to import
 - I don't know!  [Use the appropriate Manifest format](#)  to import
 - No
 - If you followed the [Earth Microbiome Project \(EMP\) protocol](#) : import as an EMP format — [single-end](#) or [paired-end](#)  — and demux with the corresponding method in q2-demux...
 - If you have a **separate barcodes/index FASTQ file**: import as an EMP format: [single-end](#) or [paired-end](#)  — and demux with the corresponding method in q2-demux...

Extract DNA, isolate and amplify the rRNA from all samples using barcoded PCR, and sequence.

Barcoded per-sample
rRNA

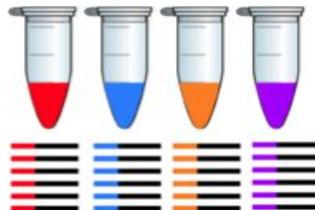


Track per-sample
barcodes (e.g., in
spreadsheet)

sample-metadata.tsv	
SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

Extract DNA, isolate and amplify the rRNA from all samples using barcoded PCR, and sequence.

Barcoded per-sample rRNA



Track per-sample barcodes (e.g., in spreadsheet)

Pool and sequence samples



sample-metadata.tsv	
SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

x 2 if paired-end

sequences.fastq(.gz)

@HWI-6X_9267:1:1:25:1051
GACGAAGGTGACGACCCTTGCTCGAACATGGGCATAAGCGCGTAGGTG
GCTTGGTAAGTCCATGGTGAATCCCTCGGCTAACCGAGGAACCTG
+
abaaaaaa^`a_]^\` ``a`^`]]]^`a[VXGX`z_\`^`a^SYOZVV
SVYGYVDXOZVT\TITBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI-6X_9267:1:1:25:267
TACGTATGGGGCAAGCGTTATCCGGAAATTTTCCCTAAACACCTCCCTACCTC
GTGGCTTAAGGCAGGGTTAACGGAAAT+
aa^`[_ ^`^`_ ^`[^`[^`_ ZZ[^`
XUWWURZUYY]XXRZRNVTRNTWUUU^
@HWI-6X_9267:1:1:25:609
TACGTAGGGGGCAAGCGTTATCCGGATT
GATGGACAAGTCTGATGTGAAAGGCTGG
+
aaab`aaa`aaaaaaaaaaaaaaaaaaaa^aa
[] [I^`azz^WW^`_ `Z_Z_T]XY^`^`^`Z
@HWI-6X_9267:1:1:25:519
GACGGAGGATGCAAGTGTATCCGGAAAT
GTTTACTAACGCAACTGTAAATCTTGA
+
abaaaaaa`aaaaaa`aaaaaaaa`^`aa
WY]_Z_XX\[\]]]^`[\XTVX`T_V
@HWI-6X_9267:1:1:25:1109
TACGGAGGGTGCAGCGTTAACCGGAAT
GTTAGGTAAGTCAGATGTGAAAGCCCCG
+
aaaba^`a^N_`\`^`a_a]Zaa^`^`Z`
^RVH_PHOWZM[PTRPTRYUBBBBBBBB
barcodes.fastq(.gz)

@HWI-6X_9267:1:1:25:1051
AACGCAC
+
bbbbbbb
@HWI-6X_9267:1:1:25:267
AAGAGAT
+
bbbbbbb
@HWI-6X_9267:1:1:25:609
AACGCAC
+
bbbbbbb
@HWI-6X_9267:1:1:25:519
ACAGCAG
+
bbbbbbb
@HWI-6X_9267:1:1:25:1109
ACAGCTA
+
bbbbbbb
@HWI-6X_9267:1:1:25:434
ACACGAG
+

Anatomy of FASTQ files

Based on FASTA

sequences.fastq(.gz)

>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLEEEQIAEKEAFSLFDKDGDTITTKELGTVMRSLGQNPTAELODMINEVDADGNGTID
FPEFLTMARKMKTDSSEEIREAFRVDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYESEFVQMMTAK*

>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNYYGSYLYSETWNTGIMLLLITMATAFMGYLPWGQMSFWGATVITNLSAIPIYGINTLV
EWIWGGFSVDKATLNRFFAFHIFILPFTMVALAGVHLTFHETGSNNPLGLTSDSKIPFHPYTTIKDFLG
LLTLIPLLALLSPDMLGDPDNHMPADPLNTPHLIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTWTWIGSQPVYPTIIGQMASILYFSIIILAFPLIAGX
IENY

@HWI-6X_9267:1:1:25:1109

TACGGAGGGTGCAGCGTTAACGGAATTACTGGCGTAAAGCGTACGTAGGC GG
TTAGGTAAGTCAGATGTGAAAGCCCCGGGCTCCACCTGGGAATGG

+

aaaba^`a^N_`_``a_a]Zaa^^\Z`[M]a`[VYa^_X^_Z]NZ\`]TY_]_^RVH_PHOWZM[
PTRPTRYUBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

@HWI-6X_9267:1:1:25:1051

GACGAAGGTGACGACC GTT GCT CGGAATCACTGGGCATAAAGCGCGCGTAGGTGG
CTTGGTAAGTCCATGGTGAAATCCCTGGCTAACCGAGGA ACTG

+

abaaaaaa^`a_]^`\``a^`^`]]^`a[VXGX^`Z_\\`_`^`a^`SYOZVVSVYGYVDXOZVT\TI
TBB

Anatomy of FASTQ files

sequences.fastq(.gz)

@HWI-6X_9267:1:1:25:1109

TACGGAGGGTGCAGCGTTAACGGAATTACTGGCGTAAAGCGTACGTAGGCGGTTAG
GTAAGTCAGATGTGAAAGCCCCGGGCTCCACCTGGGAATGG

+

aaaba^`a^N_`_\``a_a]Zaa^^\Z`[M]a`[VY^a_X^a_Z]NZ\`]TY\]_^RVH_
PHOWZM[PTRPTRYUBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

@HWI-6X_9267:1:1:25:1051

GACGAAGGTGACGACCGTTGCTCGGAATCACTGGCATAAAGCGCGTAGGTGGCTTG
GTAAGTCCATGGTGAAATCCCTCGGCTAACCGAGGAACTG

+

abaaaaa^`a_]^`\\``a^`^`]]]^`a[VXGX``Z_\\`_`a^SYOZVVSVYGY
VDXOZVT\TITBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

Sequencer identifier

Anatomy of FASTQ files

sequences.fastq(.gz)

@HWI-6X_9267:1:1:25:1109

TACGGAGGGTGCAGCGTTAACGGAATTACTGGGCGTAAAGCGTAGGCGGTAG
GTAAGTCAGATGTGAAAGCCCCGGGCTCCACCTGGGAATGG

+

aaaba^`a^N_`_\``a_a]Zaa^^\Z`[M]a`[VY^a_X^a_Z]NZ\`]TY\]_^RVH_
PHOWZM[PTRPTRYUBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

@HWI-6X_9267:1:1:25:1051

GACGAAGGTGACGACCGTTGCTCGGAATCACTGGCATAAAGCGCGTAGGTGGCTTG
GTAAGTCATGGTGAAATCCCTCGGCTAACCGAGGAACTG

+

abaaaaa^`a_]^`\\``a^`^`]]]^`^`a[VXGX``Z_\\`\\`^`a^SYOZVVSVYGY
VDXOZVT\TITBBB

actual sequence -or-
barcode sequence

Anatomy of FASTQ files

+, placeholder line

sequences.fastq(.gz)

@HWI-6X_9267:1:1:25:1109

TACGGAGGGTGCAGCGTTAACGGAATTACTGGGCGTAAAGCGTAGGCGGTAG
GTAAGTCAGATGTGAAAGCCCCGGGCTCCACCTGGGAATGG

+

aaaba^`a^N_`_\``a_a]Zaa^^\Z`[M]a`[VY^a_X^a_Z]NZ\`]TY\]_^RVH_
PHOWZM[PTRPTRYUBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

@HWI-6X_9267:1:1:25:1051

GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGTAGGTGGCTTG
GTAAGTCCATGGTGAAATCCCTCGGCTAACCGAGGAACTG

+

abaaaaa^`a_]^`\\``a^`^`]]]^`^`a[VXGX``Z_\\`_`^`a^SYOZVVSVYGY
VDXOZVT\TITBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

Anatomy of FASTQ files

sequences.fastq(.gz)

@HWI-6X_9267:1:1:25:1109

TACGGAGGGTGCAGCGTTAACGGAATTACTGGGCGTAAAGCGTAGGCGGTTAG
GTAAGTCAGATGTGAAAGCCCCGGGCTCCACCTGGGAATGG

+

aaaba^`a^N_`_\``a_a]Zaa^^\z`[M]a`[VY a^_X^_Z]NZ\`]TY\]_^RVH_
PHOWZM[PTRPTRYUBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

@HWI-6X_9267:1:1:25:1051

GACGAAGGTGACGACCGTTGCTCGGAATCACTGGCATAAAGCGCGTAGGTGGCTTG
GTAAGTCCATGGTGAAATCCCTCGGCTAACCGAGGAACTG

+

abaaaaa^`^`a_]^`^`\\``a^`^`]]]^`^`a[VXGX^`^`Z_\\`^`\\`^`a^SYOZVVSVYGY
VDXOZVT\TITBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

Quality score of sequences

Quality/Phred scores

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Quality Score Encoding

In FASTQ files, quality scores are encoded into a compact form, which uses only 1 byte per quality value. In this encoding, quality score is represented as the character with an ASCII code equal to its value + 33. The following table demonstrates relationship between the encoding character, its ASCII code, and the quality score represented.



When Q-score binning is in use, the subset of Q-scores applied by the bins is displayed.

Table 2. ASCII Characters Encoding Q-scores 0-40

Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score
!	33	0	6	54	21
*	34	1	7	55	22
#	35	2	8	56	23
\$	36	3	9	57	24
%	37	4	:	58	25
&	38	5	:	59	26
'	39	6	<	60	27
(40	7	=	61	28
)	41	8	>	62	29
*	42	9	?	63	30
+	43	10	@	64	31
,	44	11	A	65	32
-	45	12	B	66	33
.	46	13	C	67	34
/	47	14	D	68	35
0	48	15	E	69	36
1	49	16	F	70	37
2	50	17	G	71	38
3	51	18	H	72	39
4	52	19	I	73	40
5	53	20			

Multiplexed data

sample-metadata.tsv

SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

barcodes.fastq(.gz)

@HWI-6X_9267:1:1:25:1051

AACGCAC

+

Bbbbbbb

@HWI-6X_9267:1:1:25:267

AAGAGAT

+

bbbbbbb

@HWI-6X_9267:1:1:25:609

AACGCAC

+

Bbbbbbb

@HWI-6X_9267:1:1:25:519

ACAGCAG

+

bbbbbbb

@HWI-6X_9267:1:1:25:1109

ACAGCTA

+

bbbbbbb

sequences.fastq(.gz)

@HWI-6X 9267:1:1:25:1051

GACGAAGGTGACGACCCTTGCTCGGAATCACTGGGCATAAACGCGCGTAGGTG
GCTTGGTAAGTCATGGTGAATCCCTCGGCTCAACCGAGGAAC

+

abaaaaaa`^`a]^\\``a ``]]]^`a[VXGX``z \\\\ ^`a^SYOZVV
SVYGYVDXOZVT\TITBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

@HWI-6X 9267:1:1:25:267

TACGTATGGGCAGCGTTATCCGGAATTATTGGCGTAAAGAGTGCCTAGGTG
GTGGCTTAAGCGCAGGGTTAACGCAATGGCTAACATTGTTCTC

+

aa``^` ``^` ``[``^` ``^` ZZ [``^` ``^` ZUZ] WUZXYU[Q [X] UVXVN [
XUWWURZUYU] XXRZRNVRTNTWUUU^VJVOMIHQU\URRN [BBB

@HWI-6X 9267:1:1:25:609

TACGTAGGGGCAAGCGTTATCCGATTACTGGGTGAAAGGGAGCGTAGACG
GATGGACAAGTCTGATGTGAAAGGCTGGGCTAACCCCAGGACGG

+

aaab`aaa`aaaaaaaaaaaaaaaaaaaaaaYQ``^]a]\a`a] ``]Z_
[] [I ``^` aZZ^WW^ ``^` ZZ T] XY ``^` ``^` ZX\ZJS [W [V ``^` HOVYTET

@HWI-6X 9267:1:1:25:519

GACGGAGGTGCAAGTGTATCCGGAATTCACTGGCGTAAAGCGCTGTAGGTG
GTTTACTAACGTCAACTGTTAACCTGGCTAACCTCGAAATCG

+

abaaaaaaaaaaaaa\aaaaaaaa``^` aa aaaa^Z [ZY^aa`U[``^`]YZ]
WY] Z XX\\[]]^` ``^` [XTVX] T_VZ [] ZXVXYFX_VYJWWZL

@HWI-6X 9267:1:1:25:1109

TACGGAGGGTCCGAGCGTTAACCGGAATTACTGGCGTAAAGCGTACGTAGGCG
GTTAGGTAAAGTCAGATGTGAAAGCCCCGGGCTCCACCTGGGATGG

+

aaaba``^` a^N ``^` ``^` a]Zaa``^` Z` [M] a` [VY a^ X^ Z` NZ` ``^` TY\] _
^` RVH_PHOWZM [PTRPTRYUBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

Demultiplexing

sample-metadata.tsv

SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

barcodes.fastq(.gz)

@HWI-6X_9267:1:1:25:1051

AACGCAC
+
Bbbbbbb

@HWI-6X_9267:1:1:25:267

AAGAGAT
+
bbbbbbb

@HWI-6X_9267:1:1:25:609

AACGCAC
+
Bbbbbbb

@HWI-6X_9267:1:1:25:519

ACAGCAG
+
Bbbbbbb

@HWI-6X_9267:1:1:25:1109

ACAGCTA
+
Bbbbbbb

sequences.fastq(.gz)

@HWI-6X_9267:1:1:25:1051

GACGAAGGTGACGACCCTTGCTCGGAATCACTGGGCATAAACGCGCGTAGGTG
GCTTGGTAAGTCCATGGTGAATCCCTCGGCTCAACCGAGGAACTG
+
abaaaaaa^`a]^` ``a ^`]]]^` `a [VXGX ``z \\ \\ ^` a^SYOZVV
SVYGVVDXOZVT\TITBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

@HWI-6X_9267:1:1:25:267

TACGTATGGGCAAGCGTTATCCGGAATTATTGGCGTAAAGAGTCGCTAGGTG
GTGGCTTAAGCGCAGGGTTAACGCAATGGCTTAACATTGTTCTC
+
aa^` [^` ^` ^` [^` [ZUZ] WUZXYU[Q[X] UVXVN[
XUWWURZUYU] XXRZRNVRTNTWUUU^VJVOMIHQU\URRN[BBB

@HWI-6X_9267:1:1:25:609

TACGTAGGGGCAAGCGTTATCCGGAATTACTGGGTGAAAGGGAGCGTAGACG
GATGGACAAGTCTGATGTGAAAGGCTGGGCTAACCCCGGGACGG
+
aaab`aaa`aaaaaaaaaaaaaaaaaaaaaaYQ^` ^` a]\a`a] ``]Z_
[] [I^`aZ^WW^ `^` Z T] XY^`^` ZX\ZJS[W[V^`HOVYTET

@HWI-6X_9267:1:1:25:519

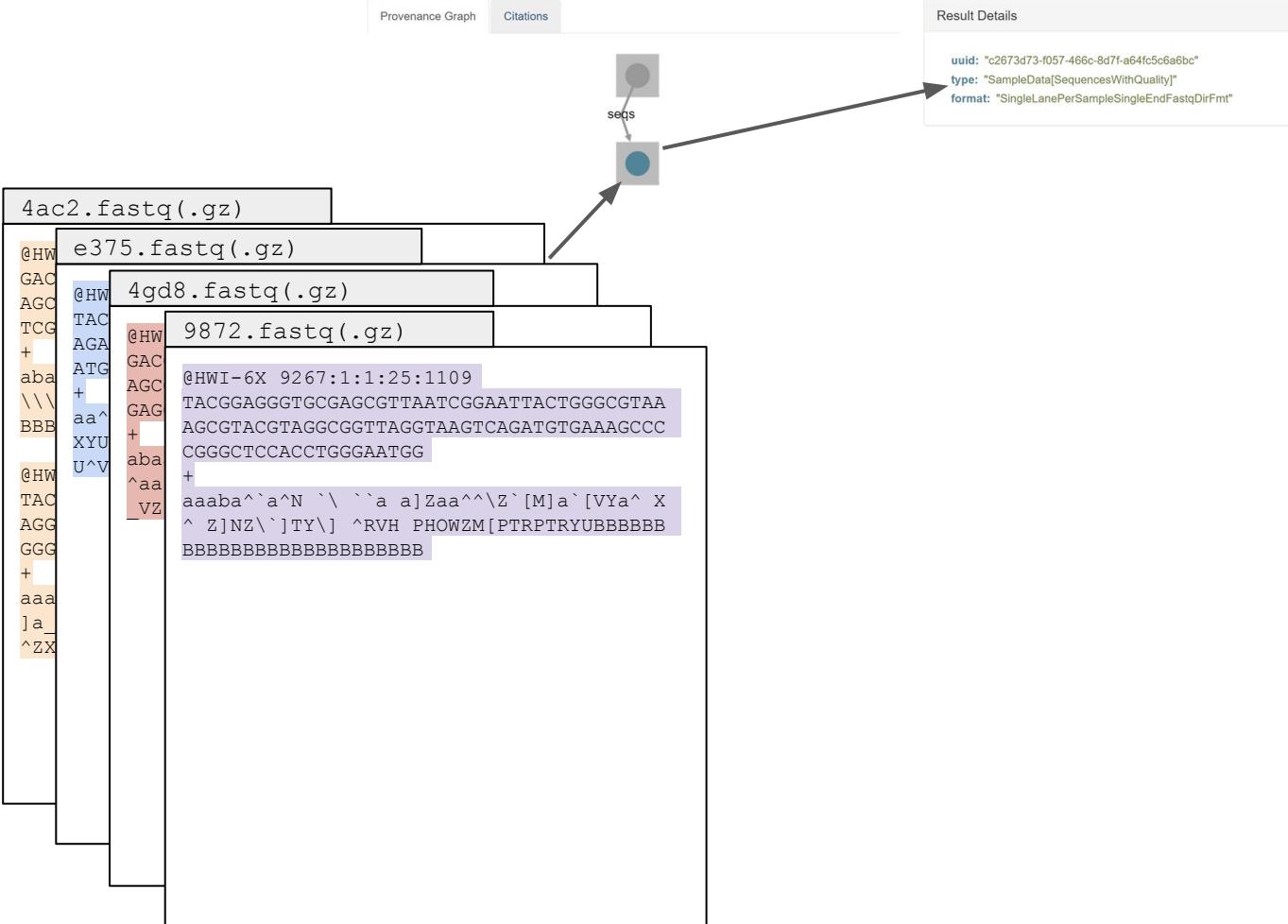
GACGGAGGTGCAAGTGTATCCGGAATCACTGGCGTAAAGCGTCTGTAGGTG
GTTTACTAACGTCAACTGTTAACCTTGAGGCTAACCTCGAAATCG
+
abaaaaaaaaaaaaaa\aaaaaaaa``aa aaaa^Z [ZY^aa`U[^`][YZ]
WY]] Z XX\\[]]]^` [\XTVX` T_VZ[]ZXVXYFX_VYJWWZL

@HWI-6X_9267:1:1:25:1109

TACGGAGGGTCCGAGCGTTAACCGGAATTACTGGCGTAAAGCGTACGTAGGCG
GTTAGGTAAAGTCAGATGTGAAAGCCCCGGGCTCCACCTGGGATGG
+
aaaba^`a^N ```a a]Zaa^`^` [M]a` [VY a^ X^ Z` NZ`] TY\]
^RVH_PHOWZM[PTRPTRYUBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

Demultiplexed

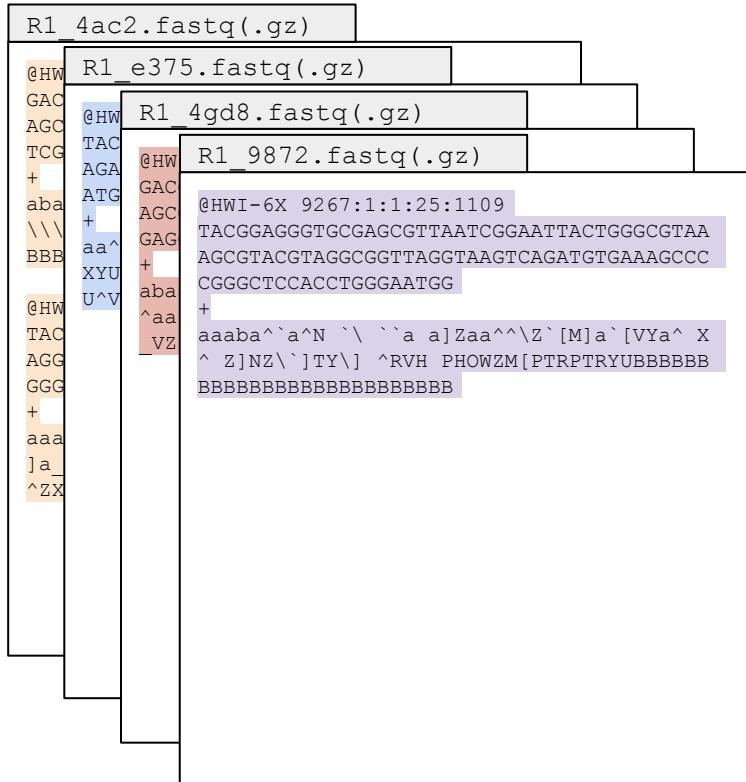
sample-metadata.tsv
SampleID
4ac2
e375
4gd8
9872



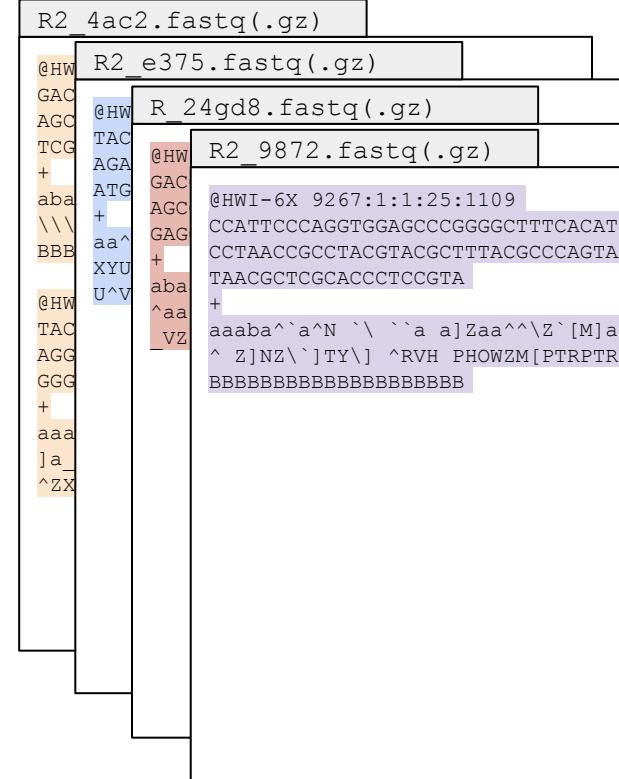
Demultiplexed (paired)

sample-metadata.tsv
SampleID
4ac2
e375
4gd8
9872

Forward reads (R1)

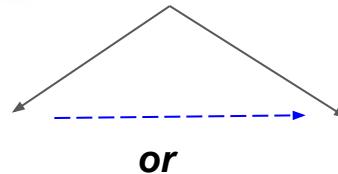


Reverse reads (R2)



Multiplexed sequence data

sample-metadata.tsv	
SampleID	BarcodeSequence
4ac2	AACGCAC
e375	<p>sequences.fastq(.gz)</p> <pre>@HWI-6X_9267:1:1:25:1051 GAGGAAGGTGACGACCCTGCTCGGAATCACTGGGCATAAAGCGCGCTAGGTG GCTTGGTAAGTCCA + abaaaaaa^__a]^ SVYGYVDXOZVT\T @HWI-6X_9267:1 TACGTATGGGCCAA GTGGCTTAAGCGCA + aa^^[___^__^_ ^_ XUWWURZUYY]XXR @HWI-6X_9267:1 TACGTAGGGCCAA GATGGACAAGTCTG + aaab`aaa`aaaaa [] [I^^aZZ^WW^ @HWI-6X_9267:1 GACGGAGGATGCAA GTTTACTAAGTCAA</pre>
4gd8	<p>barcodes.fastq(.gz)</p> <p>TG</p> <pre>@HWI-6X_9267:1:1:25:1051 AACGCAC + bbbbbbb @HWI-6X_9267:1:1:25:267 AAGAGAT + bbbbbbb @HWI-6X_9267:1:1:25:609 AACGCAC + bbbbbbb @HWI-6X_9267:1:1:25:519 ACAGCAG + bbbbbbb</pre>
9872	



Demultiplexed sequence data

4ac2.fastq(.gz)
e375.fastq(.gz)
4gd8.fastq(.gz)
9872.fastq(.gz)
<pre>@HWI-6X 9267:1:1:25:1109 TACGGAGGTGCGAGCGTTATCGGAATTACTGGCGTAA AGCGTACGTAGGCCTTAGGTAAGTCAGATGTGAAAGCCC CGGGCTCACCTGGGAATGG + aaaba^`a^N `\\ ``a a]Zaa^`^Z`[M]a`[VY`^ X ^ Z]NZ`^]TY\] ^RVH PHOWZM[PTRPTRYUBBBBBBB BBBBBBBBBBBBBBBBBBBBBB</pre>



End of Importing and Demultiplexing lecture

Mehrbod Estaki

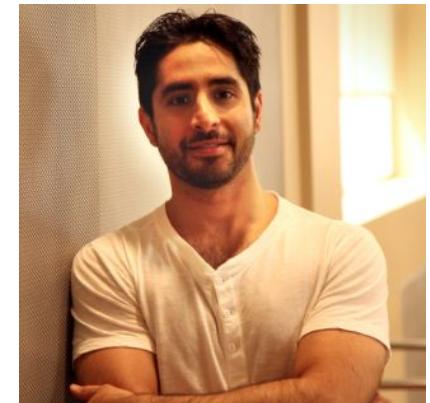
Postdoctoral Researcher

Knight Lab

University of California San Diego

 @MehrbodEstaki

 <https://mestaki.wordpress.com>



UC San Diego

Denoising

<https://bit.ly/2HThBcx>



Denoising/Clustering

Mehrbod Estaki

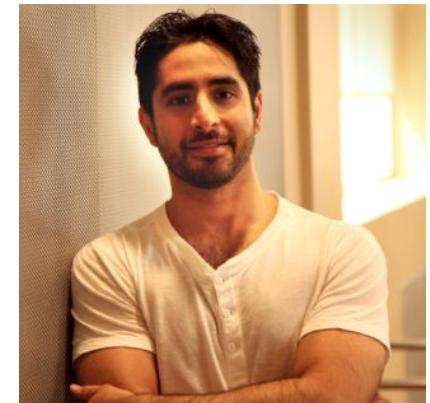
Postdoctoral Researcher

Knight Lab

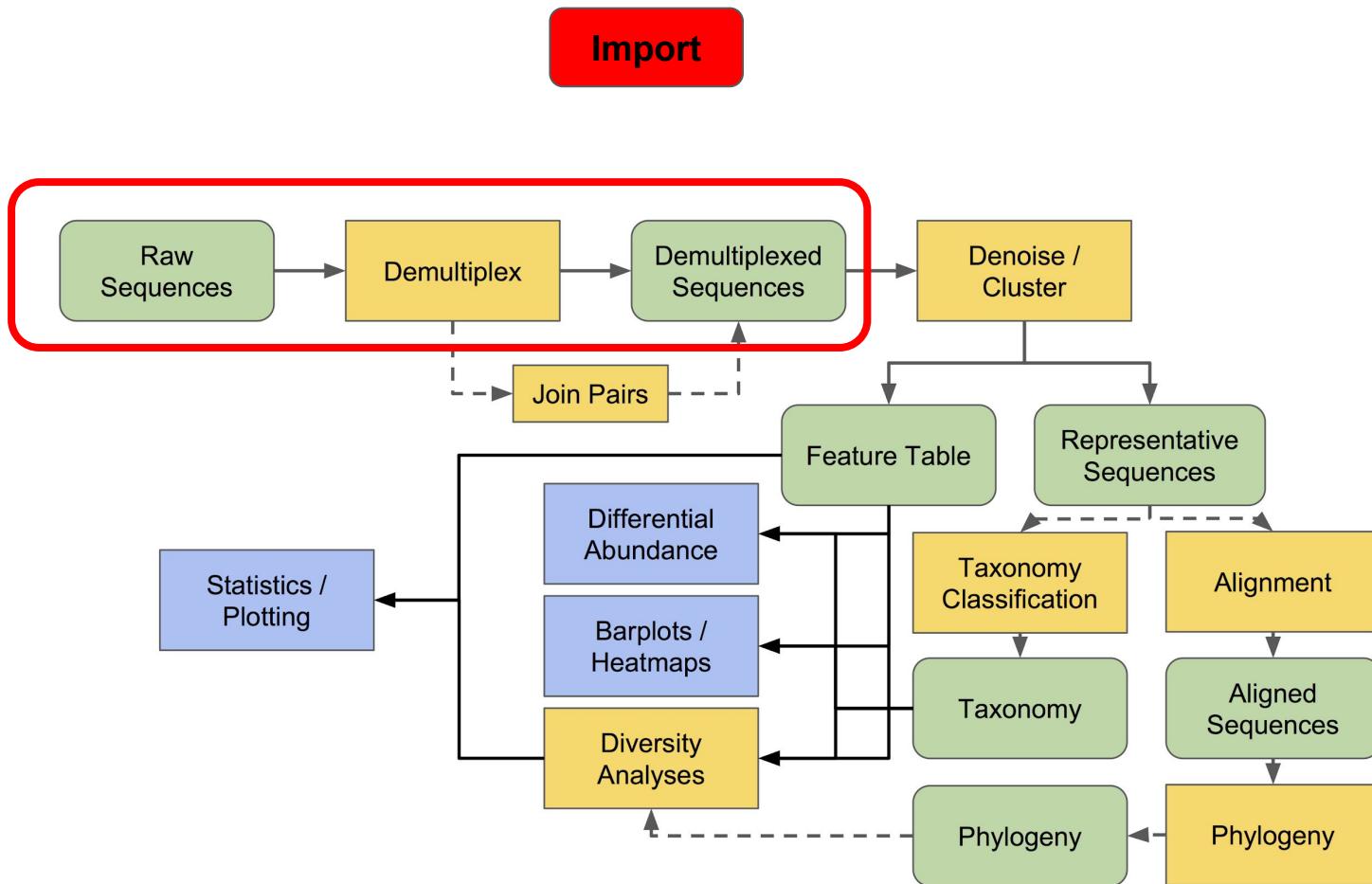
University of California San Diego

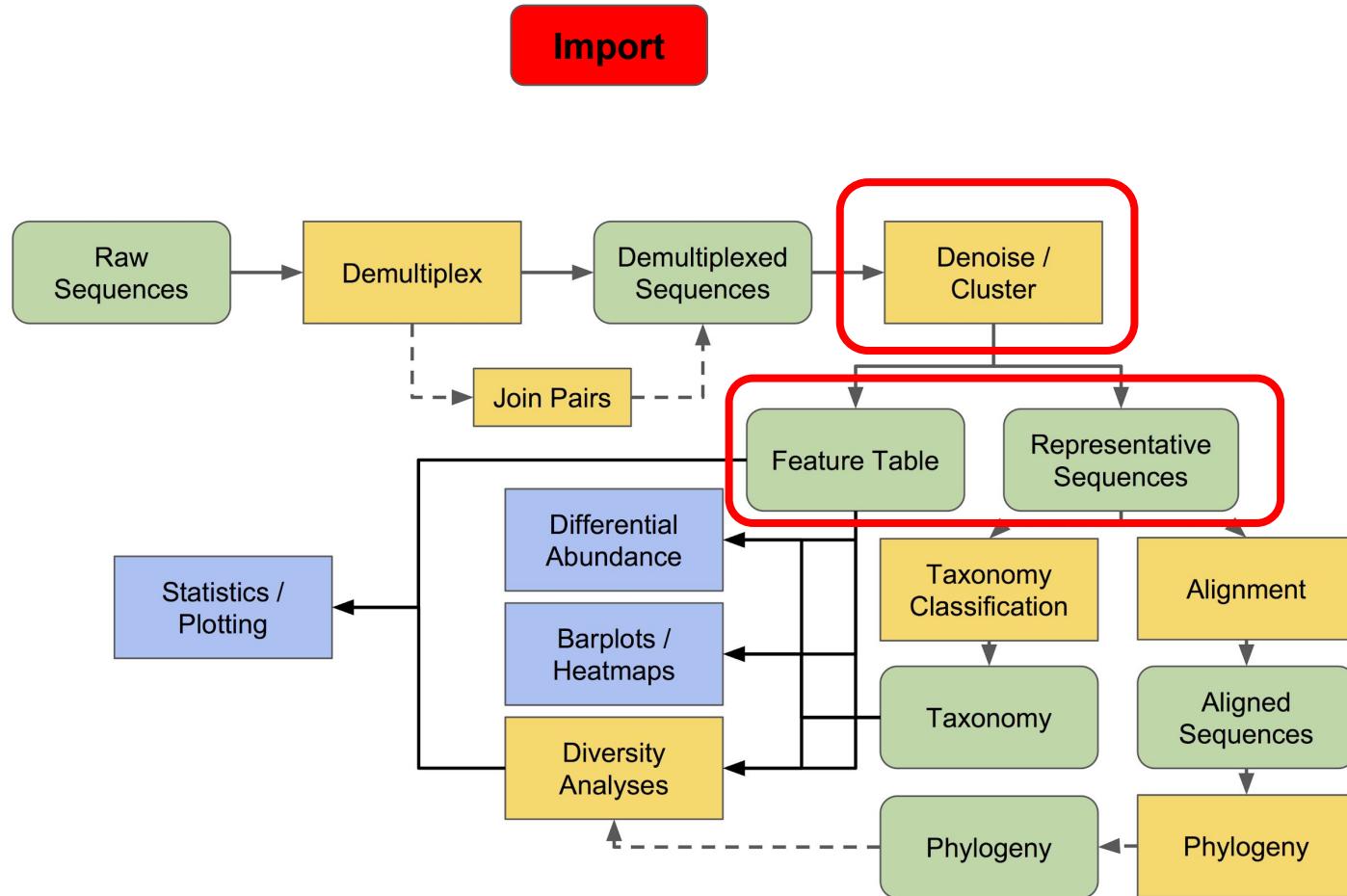
 @MehrbodEstaki

 <https://mestaki.wordpress.com>



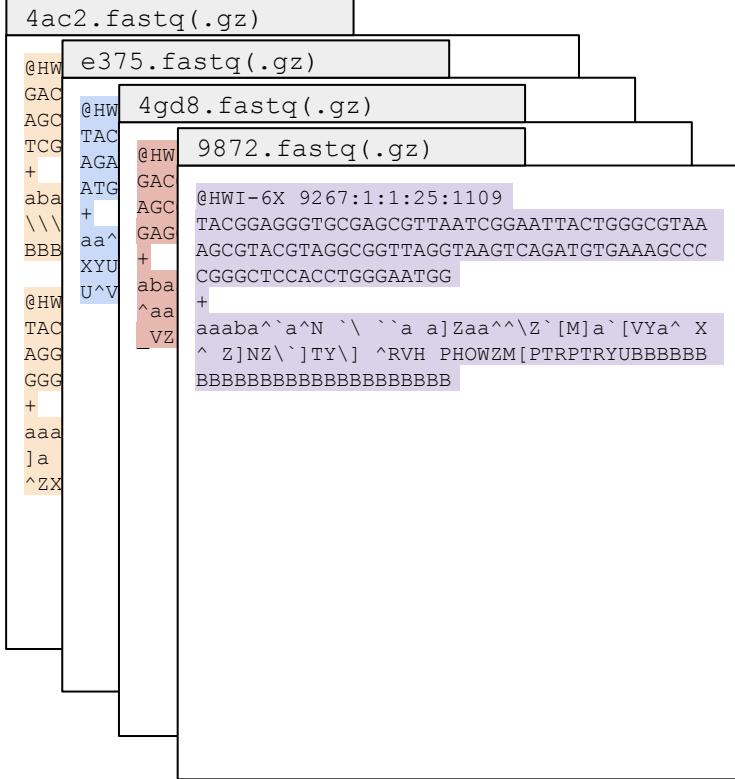
UC San Diego





Output of Denoising/Clustering

SampleData[SequencesWithQuality]



Output of Denoising/Clustering

SampleData [SequencesWithQuality]

4ac2.fastq(.gz)

e375.fastq(.gz)

GAC

AGC

TAC

AGA

+

aba

\\\

aa^

BBC

XYU

U^V

@HW

TAC

AGG

GGG

+

aaa

]a

^ZX

4gd8.fastq(.gz)

9872.fastq(.gz)

@HWI-6X 9267:1:1:25:1109

TACGGAGGGTGCAGCGTTAACCGAAATTACTGGCGTAA

AGCGTACGTAGCGGTTAGGTAAGTCAGATGTGAAAGCCC

CGGGCTCACCTGGGAATGG

+

aaaba^`a^N `\\ ``a a]Zaa^^^\\Z`[M]a`[VY a^ X

^ Z]NZ\\`]TY\\] ^RVH PHOWZM[PTRPTRYUBBBBBB

BBBBBBBBBBBBBBBBBBBBBB

	Feature 1	feature2	feature3	feature4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0



Output of Denoising/Clustering

SampleData [SequencesWithQuality]

4ac2.fastq(.gz)

e375.fastq(.gz)

GAC

AGO

TAC

AGA

+

aba

\\\

BBC

XYU

U^V

@HW

TAC

AGG

GGG

+

aaa

]a

^ZX

4gd8.fastq(.gz)

9872.fastq(.gz)

@HWI-6X 9267:1:1:25:1109

TACGGAGGGTGCAGCGTTAACCGAAATTACTGGCGTAA
AGCGTACGTAGCGGTTAGGTAAGTCAGATGTGAAAGCCC
CGGGCTCACCTGGGAATGG

aaaba^`a^N `\\ ``a a]Zaa^^^\\Z` [M]a` [VY a^ X
^ Z]NZ\\`]TY\\] ^RVH PHOWZM[PTRPTRYUBBBBBB
BBBBBBBBBBBBBBBBBBBBBB

	feature1	feature2	feature3	feature4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

FeatureData [Sequence]

>feature5
GACGAAGGTGACGACCCTGCTCGGAATCACTGGGCATAAGCGCCGTAGGTG
GCTTGGTAAGTCCATGGTAAATCCCTGGCTAACCGAGGAACGT

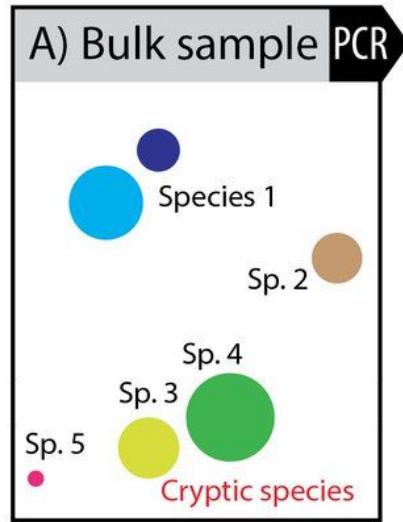
>feature4
TACGTAGGGGCAAGCGTTATCCGGATTACTGGGTGAAAGGGAGCGTAGACG
GATGGACAAGTCTGATGTGAAAGGCTGGGCTAACCCCGGGACGG

>feature2
TACGTATGGGCAAGCGTTATCCGAATTATTGGCGTAAAGAGTGCCTAGGTG
GTGGCTTAAGCGCAGGGTTAAGGCAATGGCTTAACATTGTTCTC

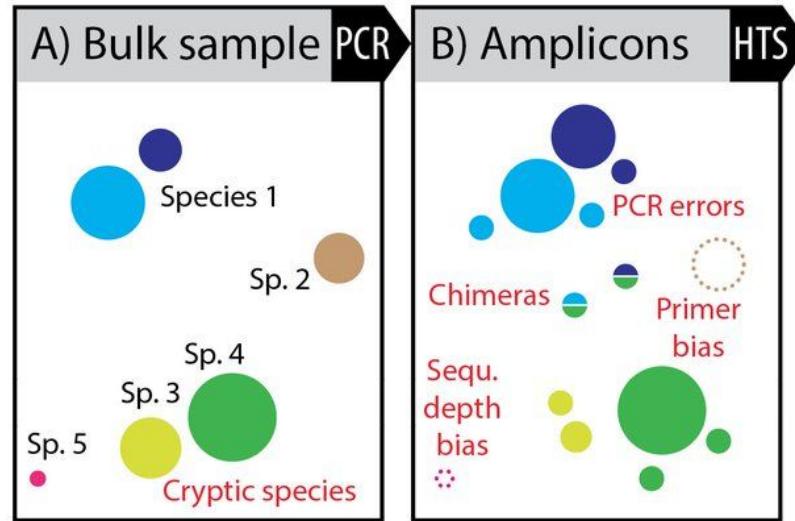
>feature1
GACGGAGGATGCAAGTGTATCCGAATCACTGGCGTAAAGCGCTGTAGGTG
GTTTACTAAGTCAACTGTTAAATCTGAGGCTAACCTCGAAATCG

>feature3
TACGGACGCTGCCGACCGTTAACCGAAATACTGGCGTAAAGCGCTACCTAGGGC

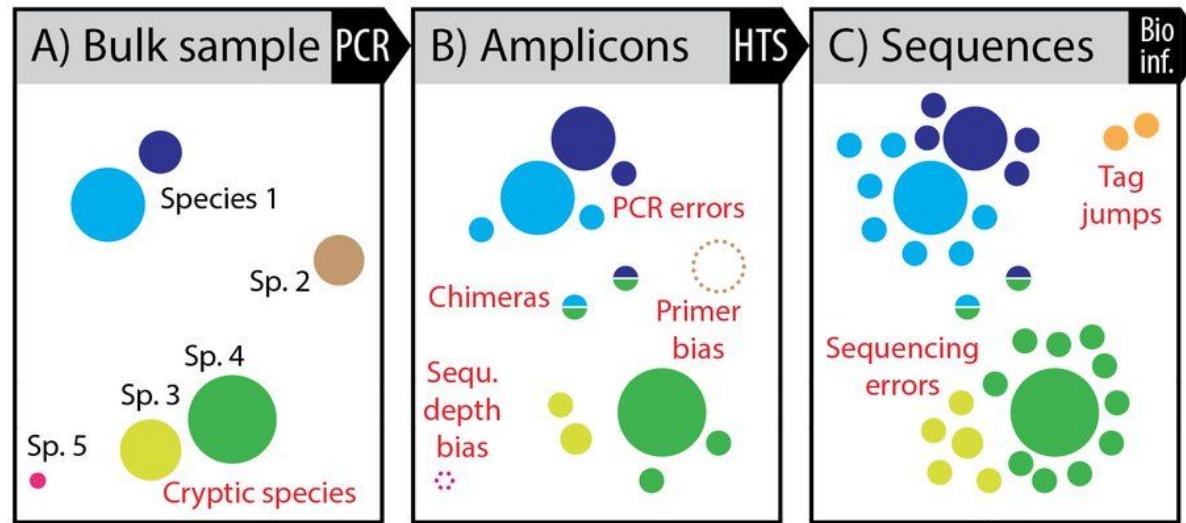
Denoising and Clustering



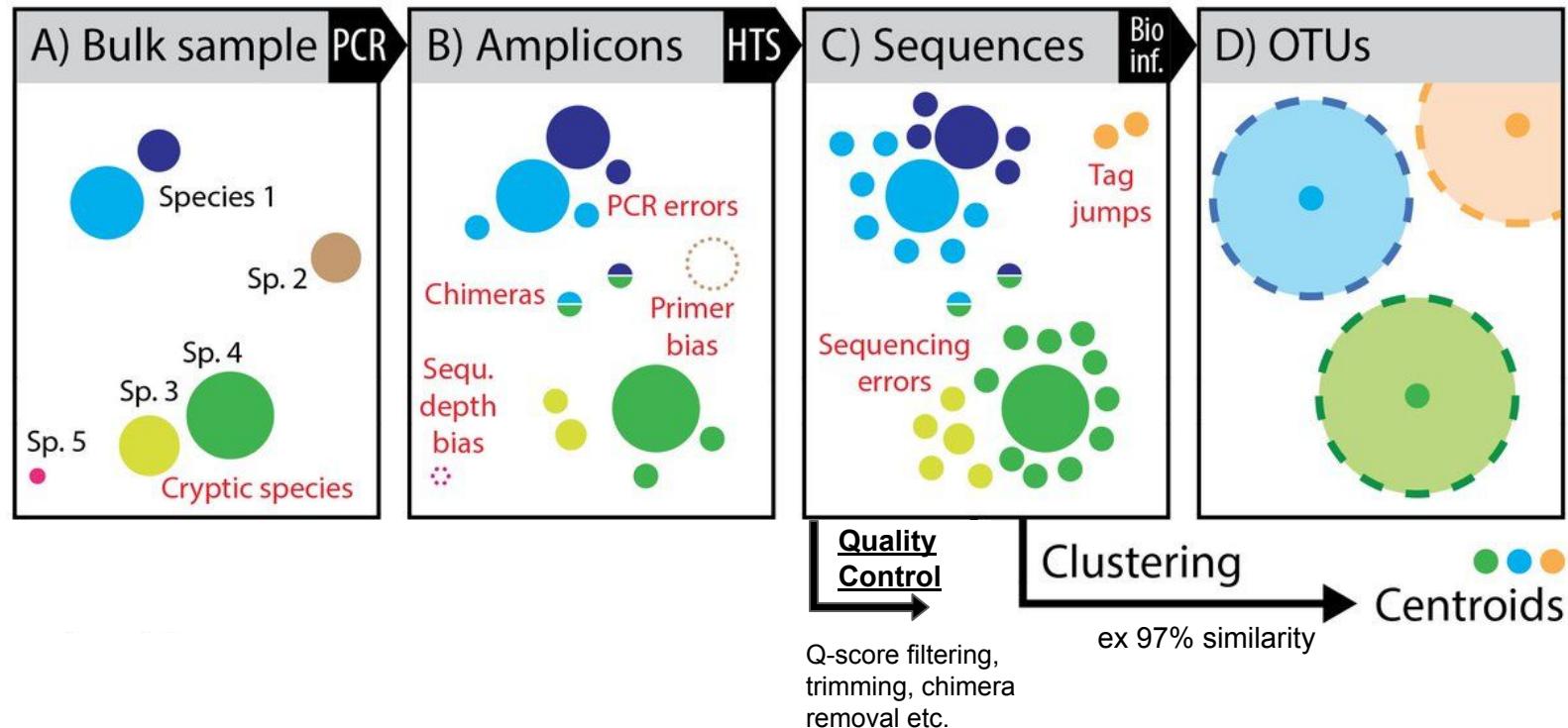
Denoising and Clustering



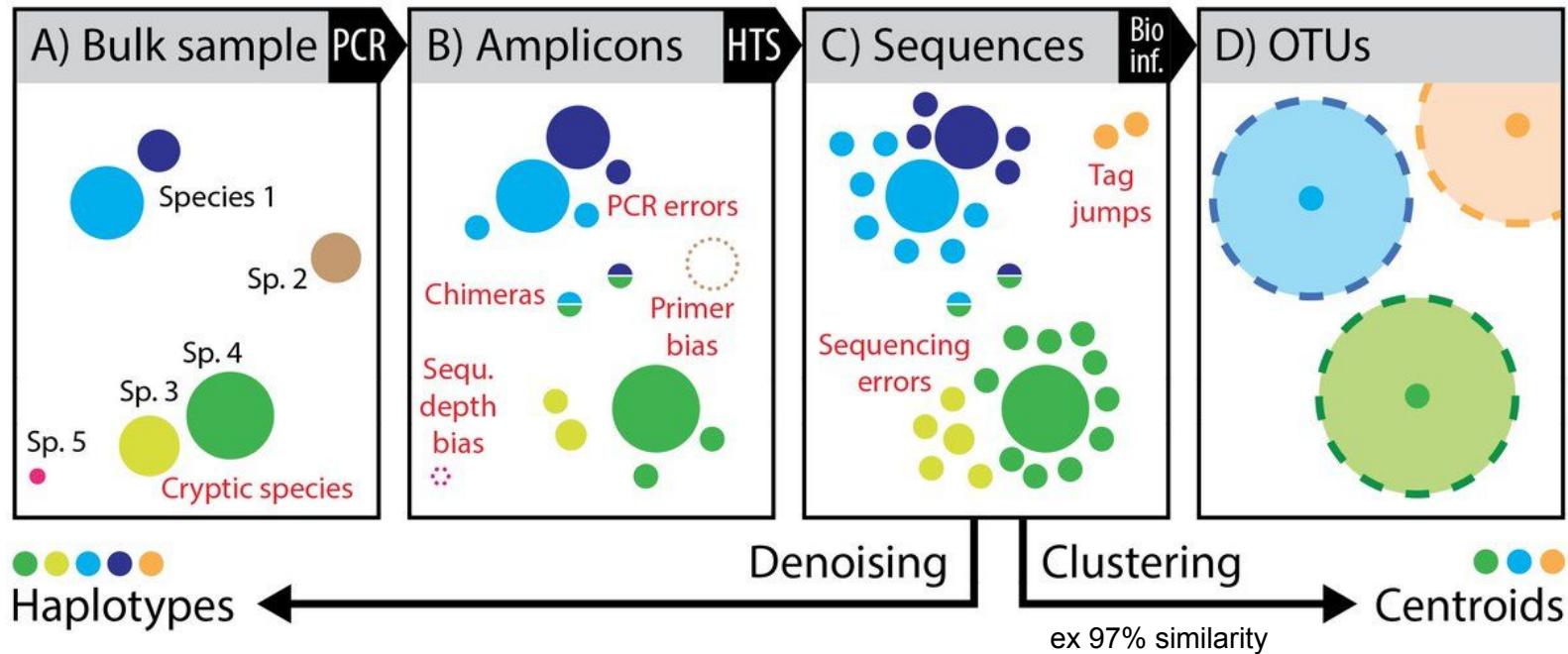
Denoising and Clustering



Denoising and Clustering



Denoising and Clustering



Terminology

Denoisers:

[DADA2](#) -> Amplicon Sequence Variant (ASV) [q2-dada]

[Deblur](#) -> sub-OTU (sOTU) [q2-deblur]

[MED](#) -> ASV

[Unoise3](#) -> zero-radius OTU (zOTU)

Clustering:

Q2-vsearch

Usearch

q2-dbOTU

..many more



Operational Taxonomic Unit (OTU)

ESV?

Open Access | Published: 21 July 2017

Exact sequence variants should replace operational taxonomic units in marker-gene data analysis

Benjamin J Callahan , Paul J McMurdie & Susan P Holmes

The ISME Journal 11, 2639–2643(2017) | Cite this article



Terminology

Denoisers:

DADA2 -> Amplicon Sequence Variant (ASV) [q2-dada]

Deblur -> sub-OTU (sOTU) [q2-deblur]

MED -> ASV

Unoise3 -> zero-radius OTU (zOTU)

Clustering:

Q2-vsearch

Usearch

q2-dbOTU

..many more

Operational Taxonomic Unit (OTU)



OTUs vs. ASVs

GGCGAGCGTT
GGCGAGCGGT
GGACGGCGTT
GGACGGCGTT
GGACGGCGTT
GGACGGCGTT
GGACGGCTTT
GGACGGCTTT
GGACGGCTTT
GGACGGCTTT
GGACGGCTGT
GGACGGCTGT

90% OTU Clustering

FeatureTable[Frequency]		
	OTU1	OTU2
4ac2	100	79
e375	88	35
4gd8	86	51
9872	12	87

Amplicon Sequence Variants
(100% OTU Clustering)

FeatureTable[Frequency]					
	SV1	SV2	SV3	SV4	SV5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

You can do both!

Denoise with
q2-dada2
q2-deblur



ASV/
sOTU

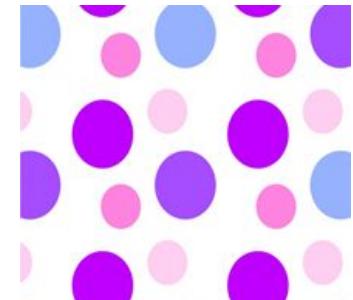
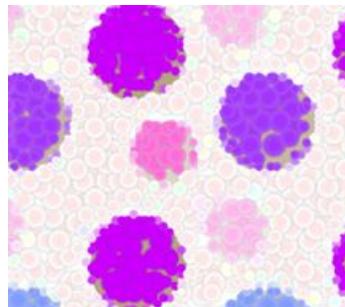
AGCGTTGTCCCGGAATTATT
GGGCGTAAAGGGCACCGCAGGCCGG
TCTTCTAAGTCTGATGTGAAATCCC
ACGGCTTAACCGTGGAGGGTCATT
GGAAACTGGAGGACTTGAGTGCAG
AAGAGGAGAGTGGATTCCACG

5f2bcbed298ca6cfdbdce9a1ac0188cb

Cluster with
q2-vsearch
q2-dbotu

OTUs

OTU 1257
OTU 1258
OTU 1261
...
OTU 1690





End of Denoising/Clustering lecture

Mehrbod Estaki

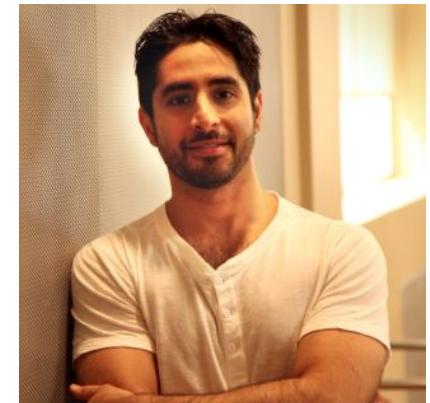
Postdoctoral Researcher

Knight Lab

University of California San Diego

 @MehrbodEstaki

 <https://mestaki.wordpress.com>



UC San Diego

Phylogenetic Reconstruction

<https://bit.ly/2HThBcx>

Phylogenetic Reconstruction

Microbiome Bioinformatics with





Michael S. Robeson II, Ph.D.
University of Arkansas for Medical Sciences
Department of Biomedical Informatics



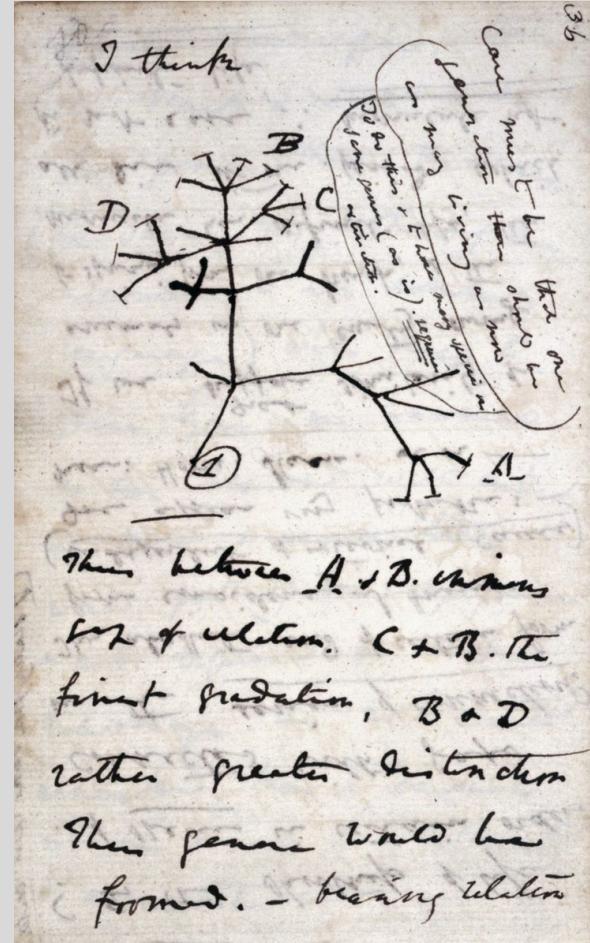
Photo credit: Michael Robeson

Okay, grab some coffee.
We're going to chat about trees!

A brief history of microbial phylogenetics.

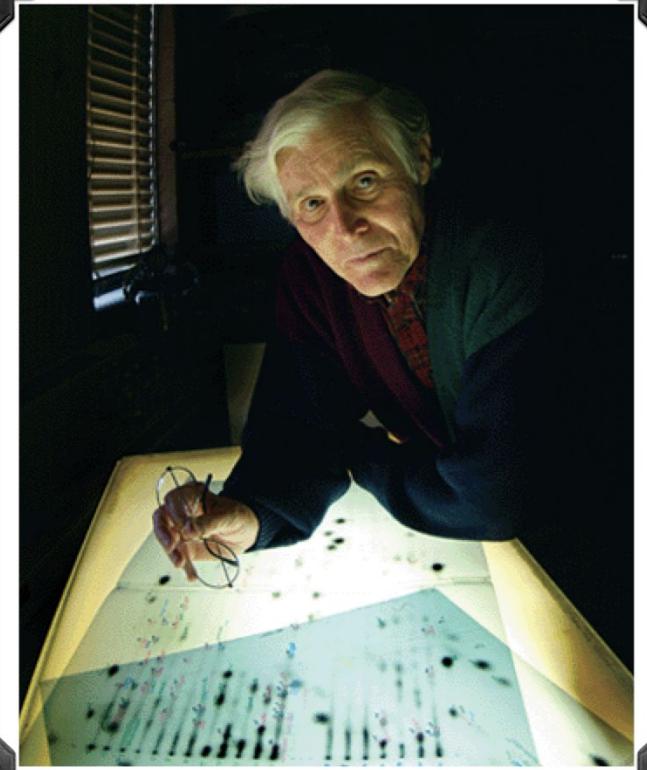
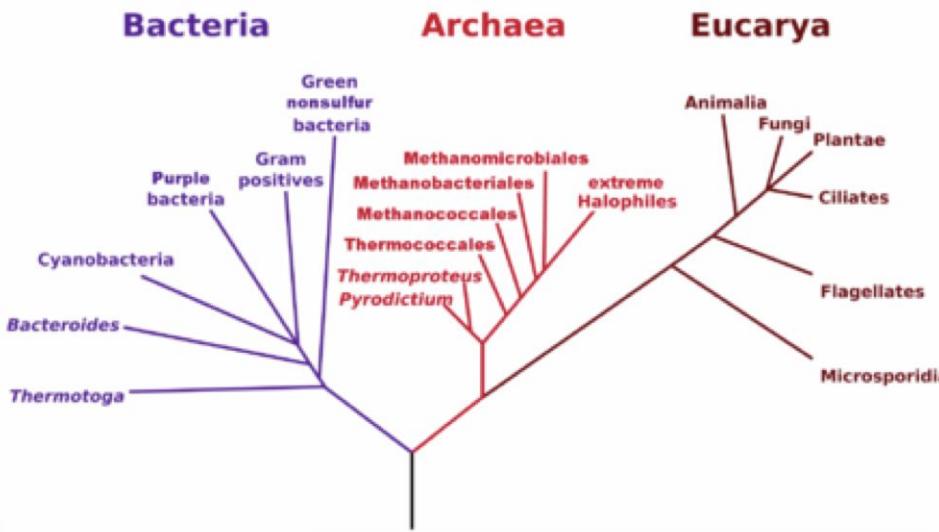
phylé / phylon = tribe, clan
genetikós = origin, source

An approach to investigate the historical evolutionary relationships and relatedness among groups of organisms or populations.



From the very beginning microbiologists have cared about quantitative ways to compare the relationships between microbes.

Phylogenetic Tree of Life



Carl Woese
(July 15, 1928 – December 30, 2012)

- Woese C, Fox G (1977). "Phylogenetic structure of the bacteria domain: the primary kingdoms." PNAS. 74 (11): 5088–90
- Woese CR, Kandler O, Wheelis ML. (1990). "Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya." PNAS 87 (12): 4576–79.

Carl Woese proposed rRNA sequences as an ideal comparative marker.

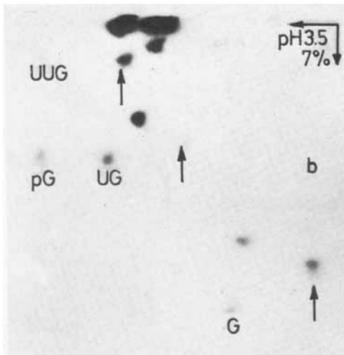
- *“Phylogenetic relationships cannot be reliably established in terms of non-comparable properties.”*
- *“A comparative approach that can measure degree of difference in comparable structures is required... To determine relationships covering the entire spectrum of extant living systems, one optimally needs a molecule of appropriately broad distribution.”*
- *“However, ribosomal RNA ... is a component of all self-replicating systems; it is readily isolated; and its sequence changes but slowly with time-permitting the detection of relatedness among very distant species.”*

Table 1. Association coefficients (S_{AB}) between representative members of the three primary kingdoms

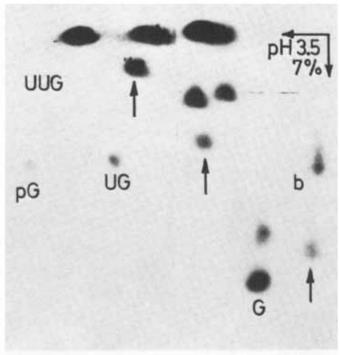
	1	2	3	4	5	6	7	8	9	10	11	12	13	
Eukaryota	1. <i>Saccharomyces cerevisiae</i> , 18S	—	0.29	0.33	0.05	0.06	0.08	0.09	0.11	0.08	0.11	0.11	0.08	0.08
	2. <i>Lemna minor</i> , 18S	0.29	—	0.36	0.10	0.05	0.06	0.10	0.09	0.11	0.10	0.10	0.13	0.07
	3. L cell, 18S	0.33	0.36	—	0.06	0.06	0.07	0.07	0.09	0.06	0.10	0.10	0.09	0.07
Bacteria	4. <i>Escherichia coli</i>	0.05	0.10	0.06	—	0.24	0.25	0.28	0.26	0.21	0.11	0.12	0.07	0.12
	5. <i>Chlorobium vibrioforme</i>	0.06	0.05	0.06	0.24	—	0.22	0.22	0.20	0.19	0.06	0.07	0.06	0.09
	6. <i>Bacillus firmus</i>	0.08	0.06	0.07	0.25	0.22	—	0.34	0.26	0.20	0.11	0.13	0.06	0.12
	7. <i>Corynebacterium diphtheriae</i>	0.09	0.10	0.07	0.28	0.22	0.34	—	0.23	0.21	0.12	0.12	0.09	0.10
	8. <i>Aphanocapsa 6714</i>	0.11	0.09	0.09	0.26	0.20	0.26	0.23	—	0.31	0.11	0.11	0.10	0.10
	9. Chloroplast (<i>Lemna</i>)	0.08	0.11	0.06	0.21	0.19	0.20	0.21	0.31	—	0.14	0.12	0.10	0.12
	10. <i>Methanobacterium thermoautotrophicum</i>	0.11	0.10	0.10	0.11	0.06	0.11	0.12	0.11	0.14	—	0.51	0.25	0.30
	11. <i>M. ruminantium</i> strain M-1	0.11	0.10	0.10	0.12	0.07	0.13	0.12	0.11	0.12	0.51	—	0.25	0.24
	12. <i>Methanobacterium</i> sp., Cariaco isolate JR-1	0.08	0.13	0.09	0.07	0.06	0.06	0.09	0.10	0.10	0.25	0.25	—	0.32
Archaea (archaeabacteria)	13. <i>Methanosa</i> cina barkeri	0.08	0.07	0.07	0.12	0.09	0.12	0.10	0.10	0.12	0.30	0.24	0.32	—

“The table provided the first gene sequence-based quantitative assessment of phylogenetic (evolutionary) relationships between representatives of the major known kinds of organisms.”

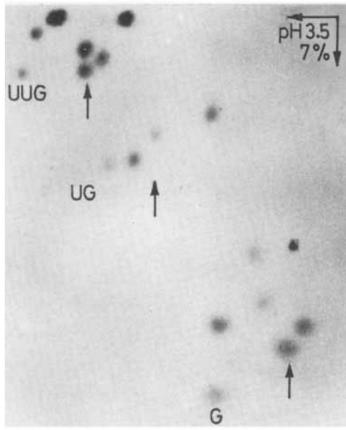
A (T4-1)



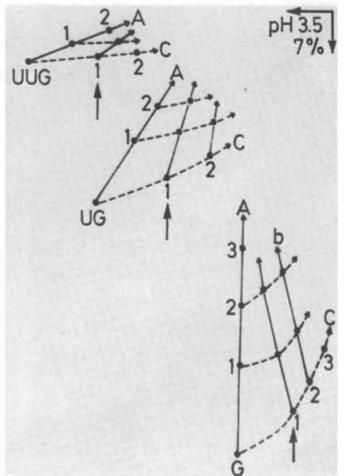
B (T4-8)



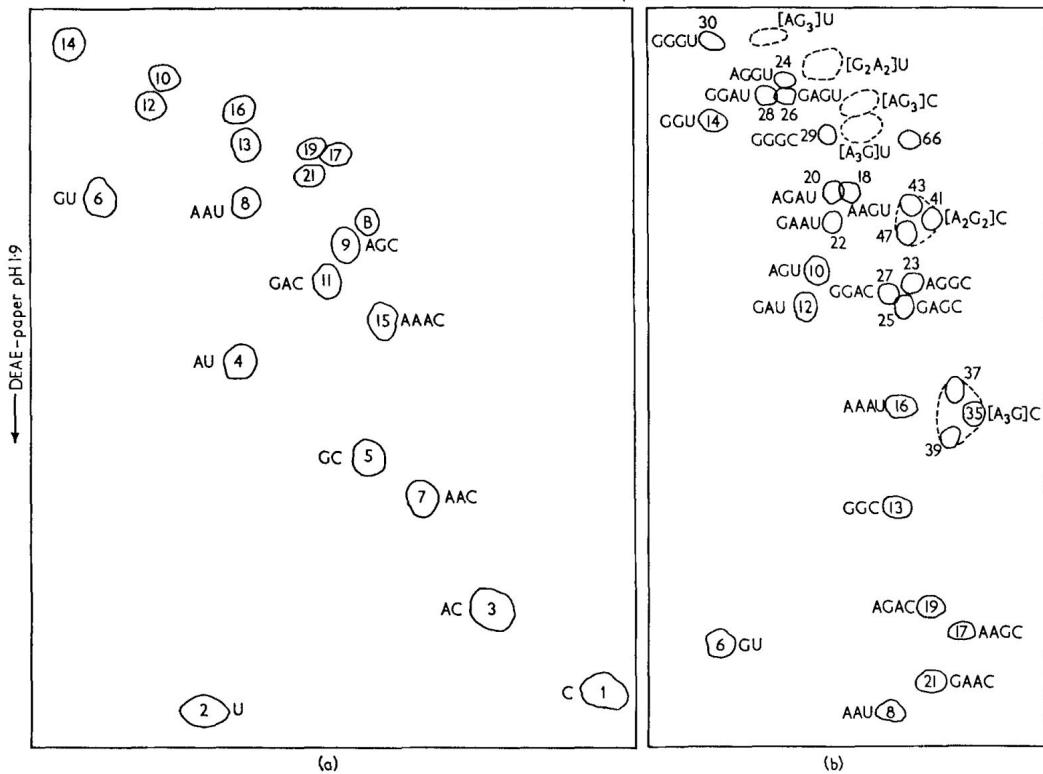
C (T4-10)



D



← Cellulose acetate pH 3.5



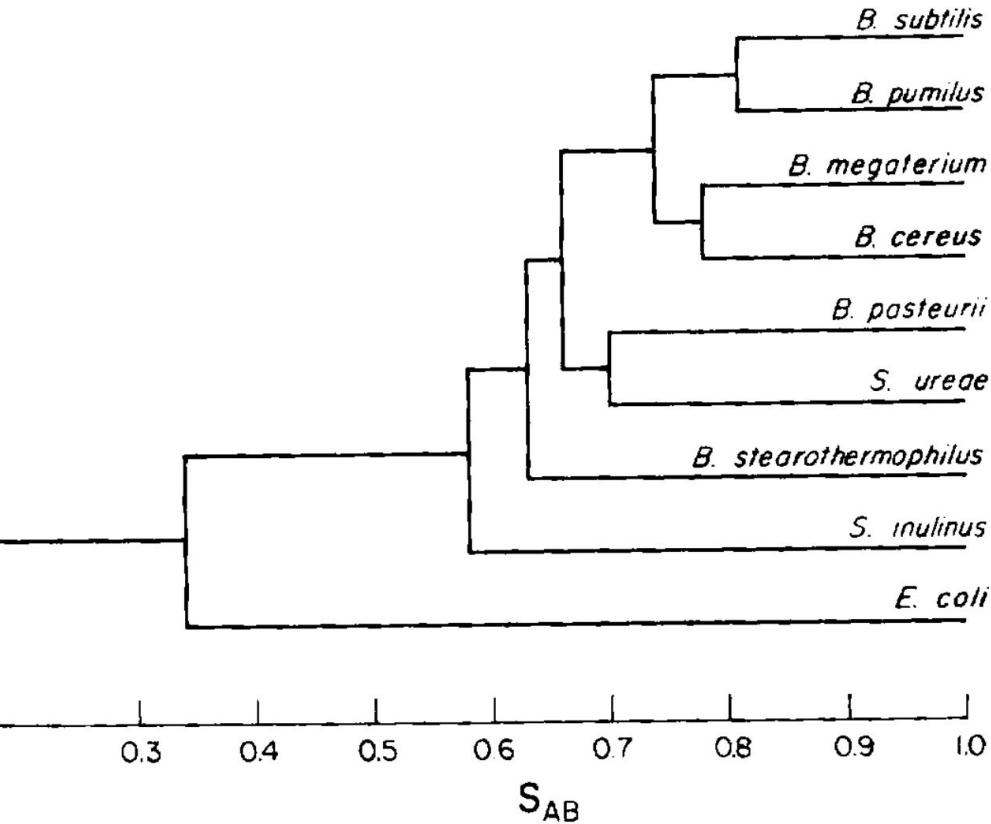
(a)

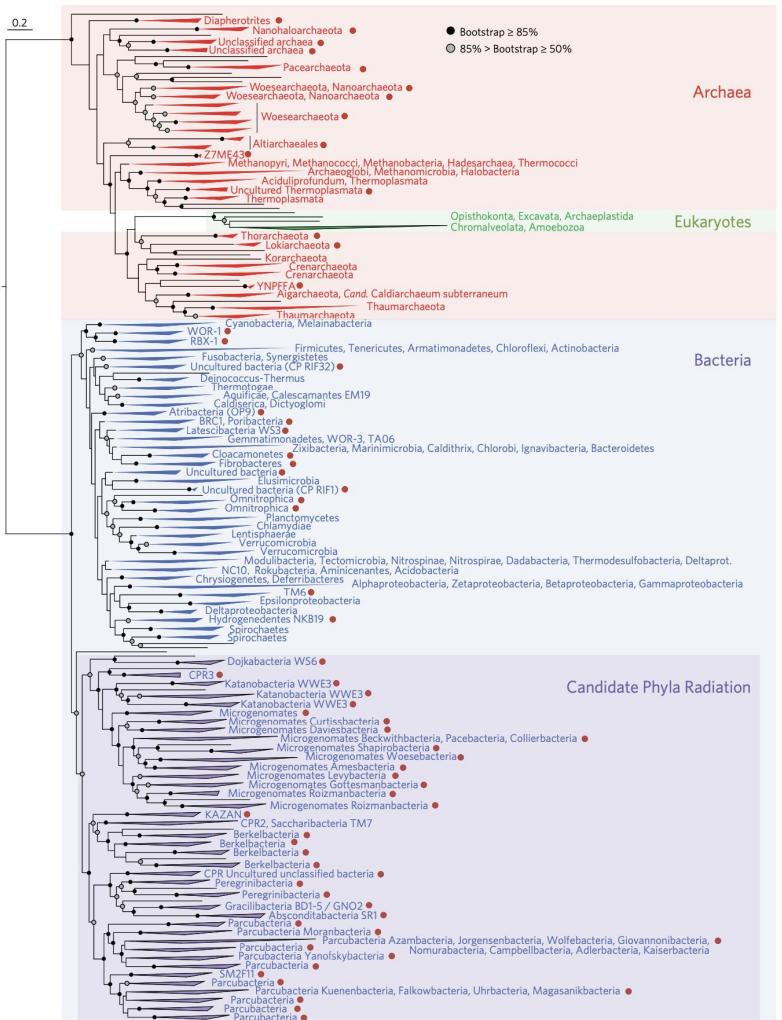
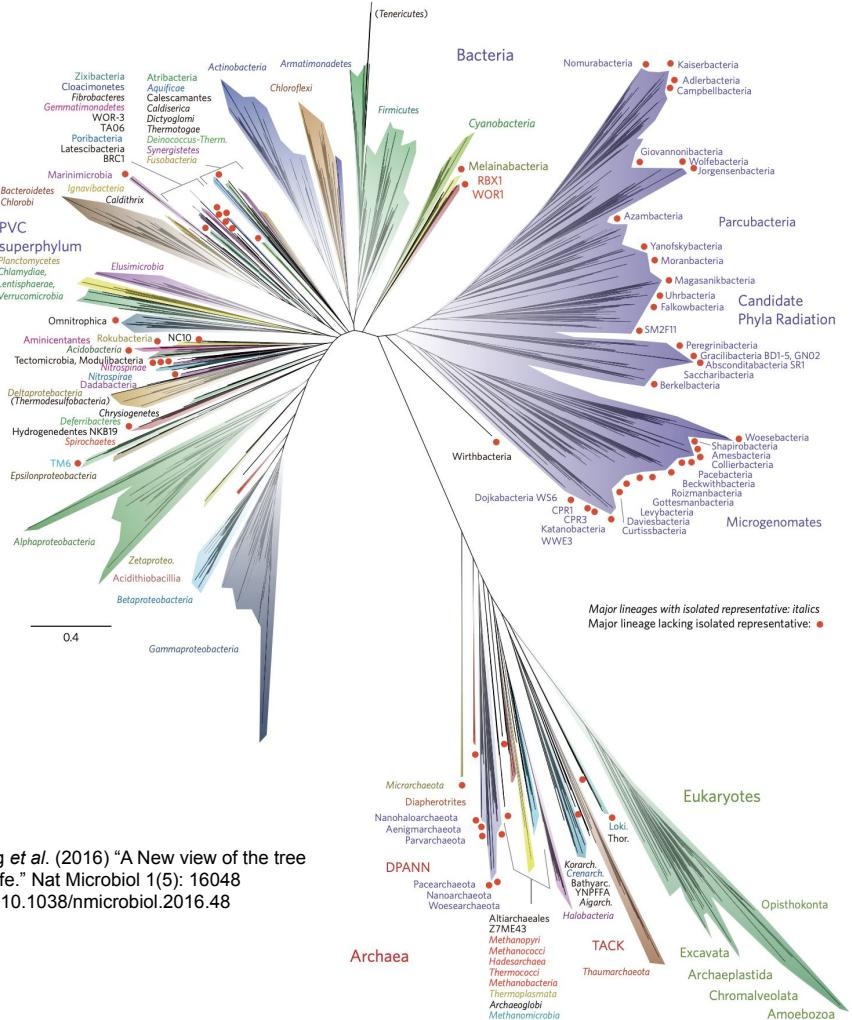
(b)

- Sanger et al. (1965) "A Two-Dimensional Fractionation Procedure for Radioactive Nucleotides." *Journal of Molecular Biology* 14 (1): 303.
- Ikemura et al. (1975) "Two-Dimensional Polyacrylamide-Gel Electrophoresis for Purification of Small RNAs Specified by Virulent Coliphages T4, T5, T7 and BF23." *European Journal of Biochemistry / FEBS* 51 (1): 117-27.
- Uchida et al. (1974) "The Use of Ribonuclease U2 in RNA Sequence Determination." *Journal of Molecular Evolution* 3 (1): 63-77.

TABLE 3. Oligonucleotide families^a

Families	<i>B. subtilis</i>	<i>B. pumilus</i>	<i>B. megaterium</i>	<i>B. cereus</i>	<i>B. pasteurii</i>	<i>S. ureae</i>	<i>B. stearothermophilus</i>	<i>S. inulinus</i>
Complete families								
1.	AAUACG AAUCGG	1 0	1 0	1 0	1 0	1 0	1 0	- + -
2.	UAACUG UUACUG CAACUG	1 0 0	1 0 0	1 0 0	1 0 0	1 0 0	0 0 1	- + -
3.	CAAACAG AACAG ^b	0 1	0 1	1 0	1 0	1 0	1 0	- + -
4.	AACAAG ACAAG ^b	1 0	1 0	1 0	1 0	1 0	0 1	+ - +
5.	CACUCGG CAUUCGG	1 0	1 0	1 0	1 0	1 0	1 0	- + -
6.	UCCACG UCCAUG	1 0	1 0	1 0	0 1	1 0	1 0	- + -
7.	CCCCCG CCCACG	1 0	1 0	0 1	0 1	0 1	0 1?	- - -
8.	UUCUCAG UUCCCG CUCUCAG	1 0 0	1 0 0	1 0 0	0 1 0	0 1 0	0 0 1	- - -
9.	UAACCUG CAACCUG	1 0	1 0	0 1	1 0	0 1	0 1	- + -
10.	pUAUUAUG pUCUUAUG pUUUAUUG pUUUAUCG pUUUUUUCG pUUCUJUG pCUUUUUG	0 0 0 0 0 0 0	0 0 0 0 0 0 0	0 0 0 0 0 0 0	1 0 1 0 0 0 0	0 1 0 0 0 0 0	0 0 0 0 0 0 0	- - - - -
11.	ACUUUCUG ACUCUCUG CUCUCUG CUUUCUG	0 1 0 0	0 1 0 0	0 0 0 0	1 0 0 1	0 0 0 1	0 0 1 0	- + - + -
12.	ACAAACCG ACAACCCG	1 0	1 0	1 0	1 0	1 0	0 1	- + -
13.	CUCAACCG CUUAACCG	1 0	1 0	1 0	0 1	1 0	1 0	- + -
14.	AACACCAG AAUACCAG	1 0	1 0	1 0	1 0	1 0	1 0	- + -
15.	CCACACUG CCACAUUG	1 0	1 0	1 0	1 0	1 0	1 0	- + -





This is nice and all... but of what practical purpose is knowing the phylogenetic relatedness of microbial taxa?

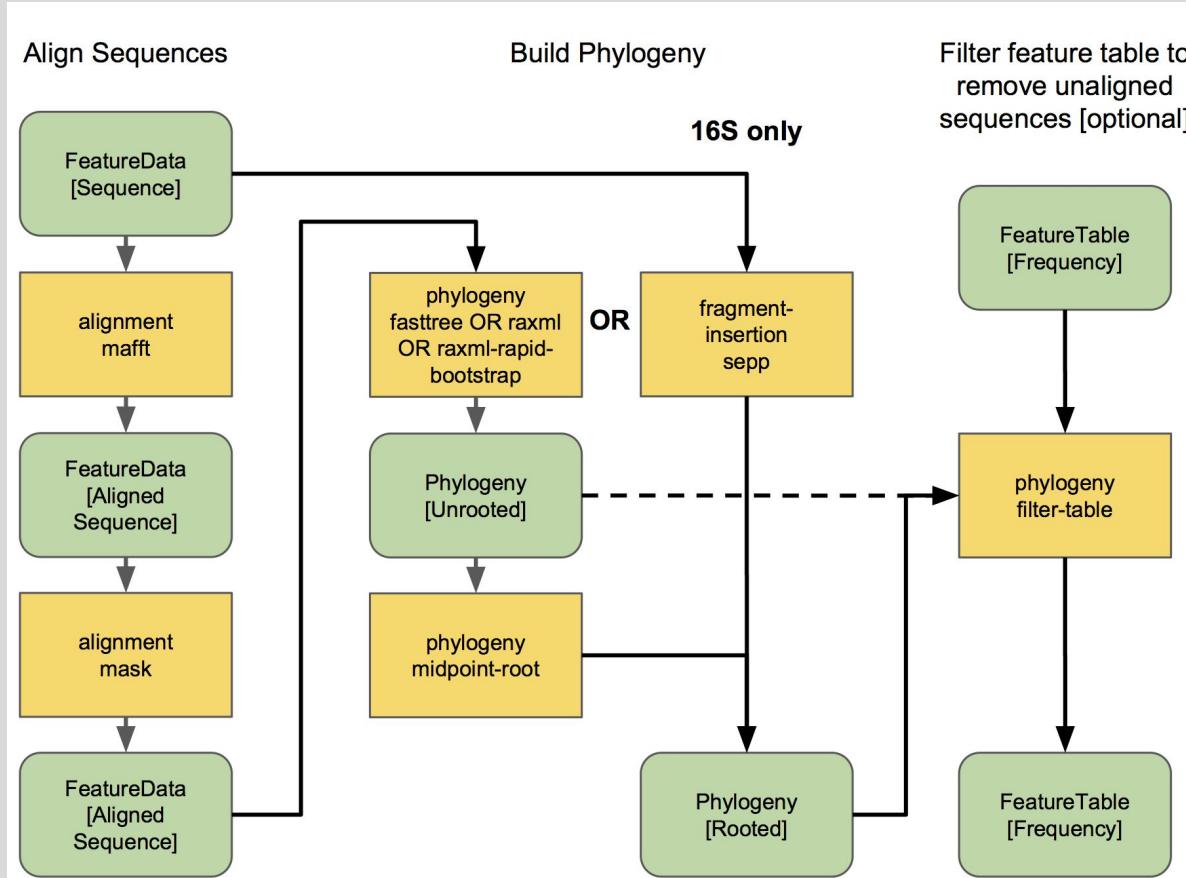
Unknown organisms that are within phylogenetic proximity to known organisms allows us to better infer their properties. For example, we might be able to infer or predict the following features:

- antibiotic-susceptibility patterns
- the nature of the DNA-replication machinery
- metabolism/ biosynthetic pathways
- niche
- regulatory mechanisms
- determine if free-living or symbiotic?
- etc...

*Knowing these features can improve our ability to cultivate microorganisms that have been previously regarded as recalcitrant.**

*Cross et al. (2018) "Insights into the Evolution of Host Association through the Isolation and Characterization of a Novel Human Periodontal Pathobiont, Desulfobulbus Oralis." mBio 9(2). <https://doi.org/10.1128/mBio.02061-17>.

Phylogenetic Reconstruction



Multiple options for phylogenetic reconstruction in QIIME 2

`qiime alignment mafft`

`qiime alignment mask`

`qiime phylogeny fasttree`

`qiime phylogeny raxml (-rapid-bootstrap)`

`qiime phylogeny iqtree (-ultrafast-bootstrap)`

`qiime phylogeny midpoint-root`

`qiime phylogeny align-to-tree-mafft- (fasttree|raxml|iqtree)`

`qiime fragment-insertion spp`

De novo phylogenies.

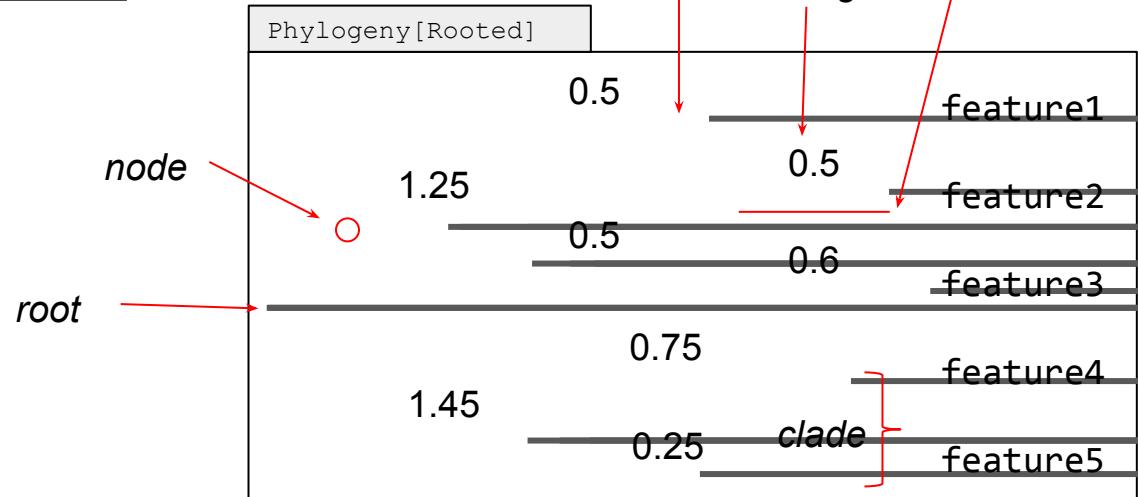
Phylogenetic reconstruction of observed sequences.

FeatureData [Sequence]

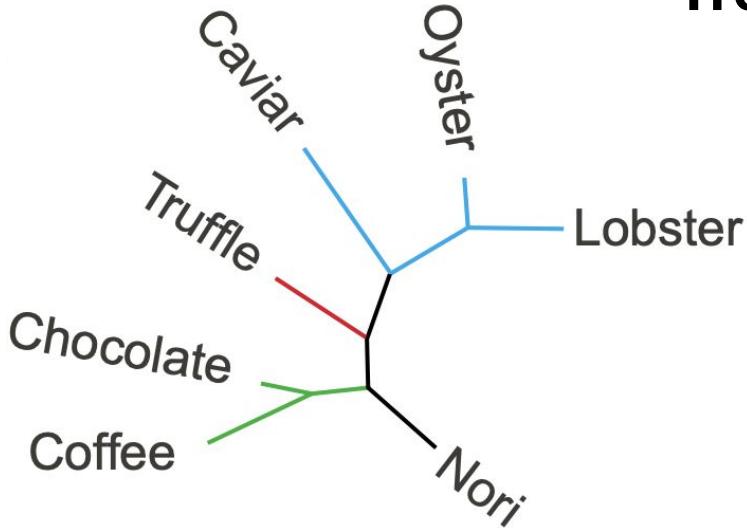
```
>feature5  
GACGAAGGTGACGCCGTTGCTCGGAATCACTGGGCATAAAGCGCGTAGGTGGCTTGGTAAGTCATGGTGA  
ATCCCTCGGCTCAACCGAGGAACGT  
>feature4  
TACGTAGGGGCAAGCGTTATCCGGATTACTGGGTGAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAA  
AGGCTGGGGCTCAACCCGGGACGG  
>feature2  
TACGTATGGGGCAAGCGTTATCCGGATTATTGGGCGTAAAGAGTGCGTAGGTGGCTTAAGCGCAGGGTTA  
AGGCAATGGCTTAACATTGTTCTC  
>feature1  
GACGGAGGATGCAAGCTTATCCGAATCACTGGGCTAAAGCGCTGTAGGTGGTTACTAAGTCAACTGTTA  
ATCTGAGGCTCAACCTCGAAATCG  
>feature3  
TACGGAGGGTGCAGCGTTAACGGAATTACTGGGCGTAAAGCGTACGTAGGCAGTTAGGTAAGTCAGATGTGAA  
AGCCCCGGCTCACCTGGGAATGG
```

Align sequences,
filter highly variable
(i.e., randomly
evolving) positions,
and build
phylogenetic tree.

branch
branch length
tips/leaves

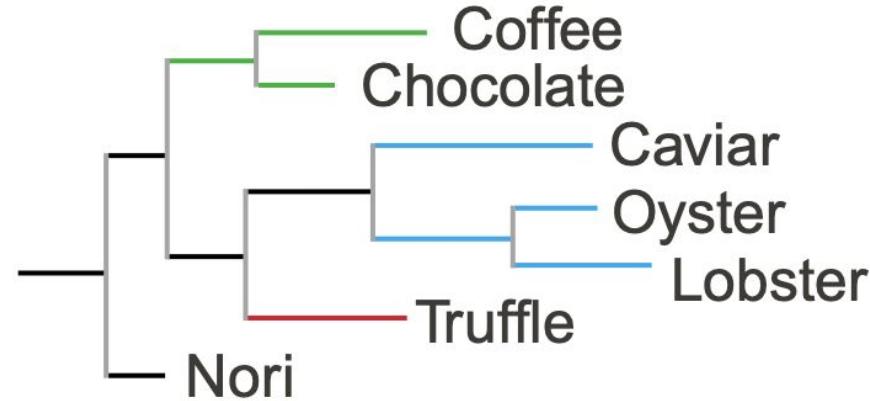


Tree structure



Unrooted Tree

- difficult to infer ancestry
- no directionality of ancestral relationships
- no implied hierarchy / groupings



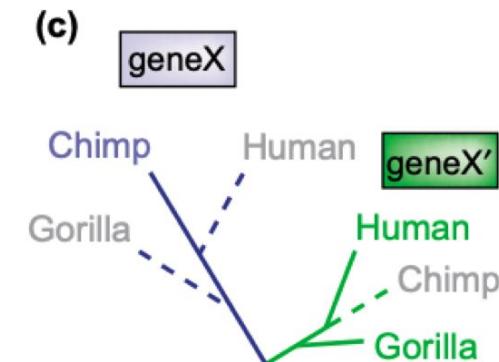
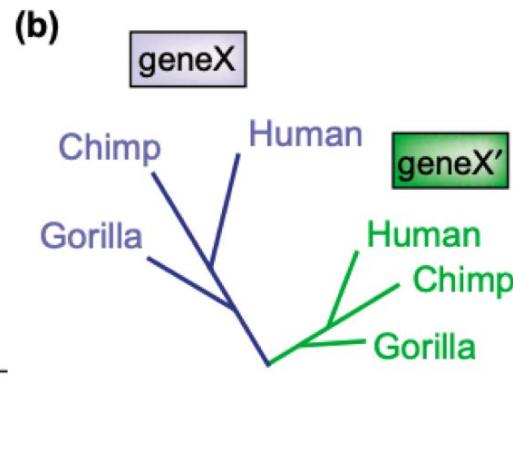
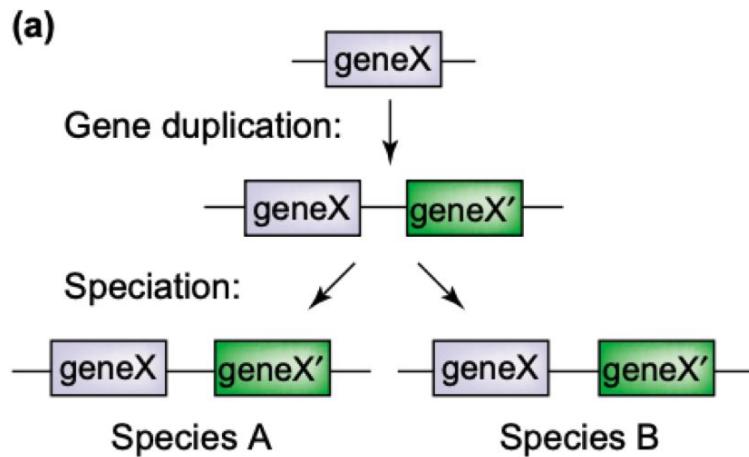
Rooted Tree

- implies the order of branching of the tree; “*who shares a more recent common ancestor with whom*”
- directionality of ancestral relationships
- imposed hierarchy / groupings

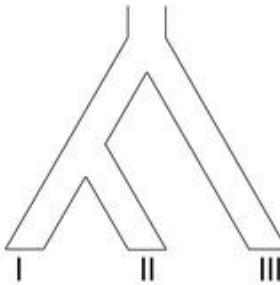
Homologues: sequences that have common origins but may or may not share common activity.

Orthologs: homologous sequences in different genomes produced by speciation. Tend to have similar function.

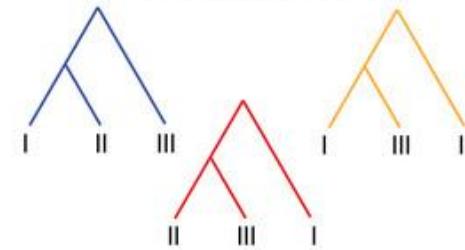
Paralogs: homologous sequences in the same genome produced by gene duplication. Tend to have different functions.



A. Species Relationships: ((I, II), III)



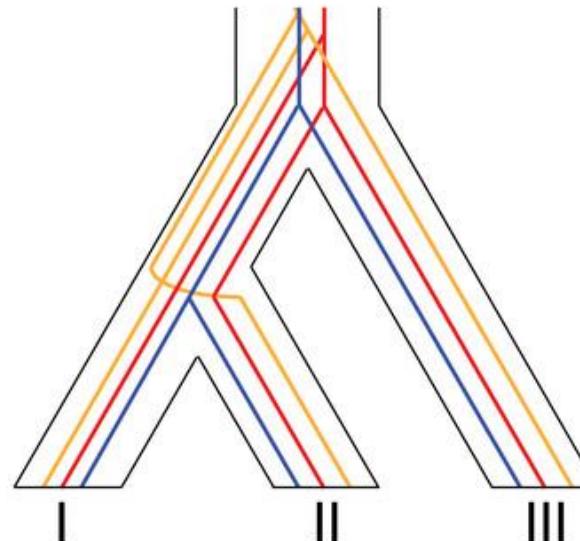
B. Possible Gene Trees



Discordance

*Gene trees do not
always correspond
with the species tree!*

C. Different Gene Trees within Species Tree



De novo alignment algorithms

There are many algorithms.

In a nutshell: how do we decide where to place “indels” (insertions / deletions) to best represent positional homology (orthologous sequences)?

TTCATA
TGCTCGTA

↓

T--TCATA
TGCTCGTA

Dynamic programming matrix:

		(sequence y)								
		0	1	2	3	4	5	6	7	8 = N
		T	G	C	T	C	G	T	A	
i	0	0	-6	-12	-18	-24	-30	-36	-42	-48
	1 T	-6	5	-1	-7	-13	-19	-25	-31	-37
	2 T	-12	-1	3	-3	-2	-8	-14	-20	-26
	3 C	-18	-7	-3	8	2	3	-3	-9	-15
	4 A	-24	-13	-9	2	6	0	1	-5	-4
	5 T	-30	-19	-15	-4	7	4	-2	6	0
	M = 6 A	-36	-25	-21	-10	1	5	2	0	11

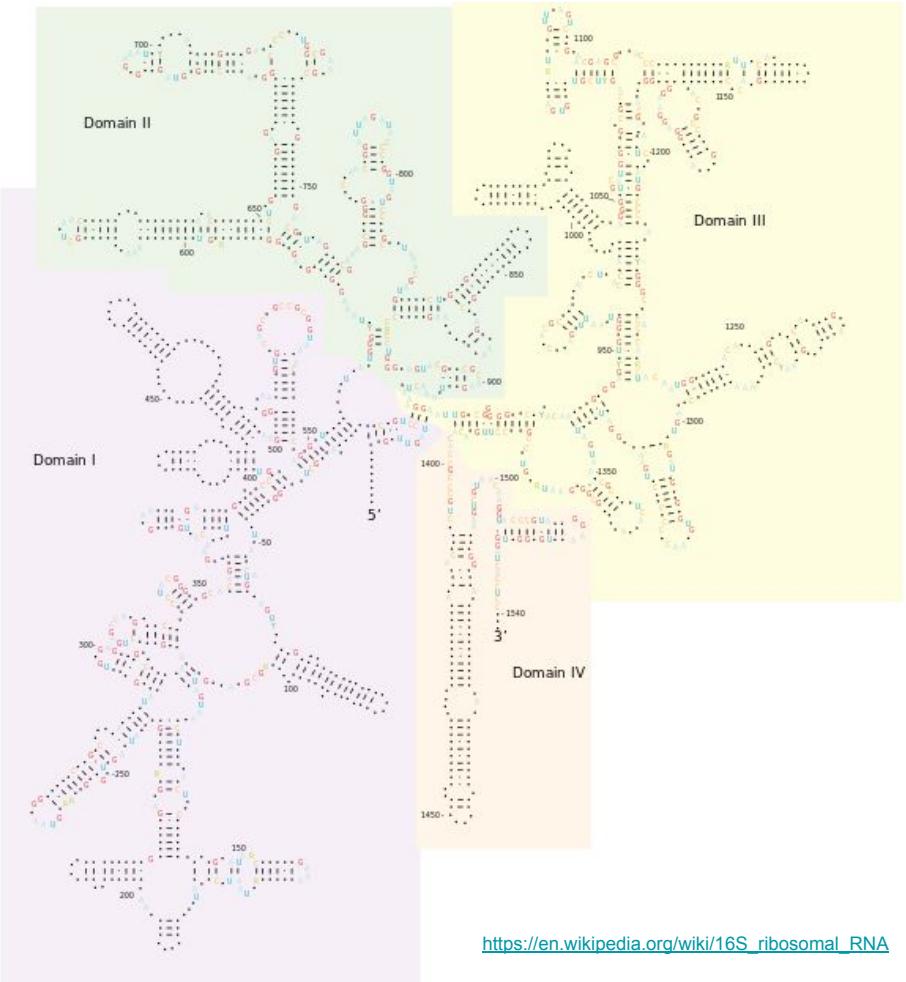
Optimum alignment scores 11:

T - - T C A T A
T G C T C G T A
+5 -6 -6 +5 +5 -2 +5 +5

Reducing alignment ambiguity

Reference-based alignments:

- Align sequences based on a profile / reference alignment
 - e.g. SINA, infernal, NAST, etc...
 - incorporate secondary structure
- Can easily merge different alignments together



Template: ATAC-----GTA-AC-----GTA---C---G-T-AC-GG
Candidate: CACGTTAACGTCGTACCCGG

↓
(A) Find pair-wise alignment

NAST (Nearest Alignment Space Termination)

Template: ATACGT-A-ACGTACGTAC--GG
Candidate: C-ACGTTAACGT-CGTACCCGG

↓
(B) Re-introduce template spacing

Template: ATAC-----GT-A-AC-----GTA---C---G-T-AC--GG
Candidate: C-AC-----GTTAAC----GT----C---G-T-ACCCGG

↓
(C) Identify template-extending insertions

Template: ATAC-----GT-A-AC-----GTA---C---G-T-AC--GG
Candidate: C-AC-----GTTAAC----GT----C---G-T-ACCCGG
 α ← → β

↓
(D) Search for nearest alignment spaces
(hyphens) in candidate

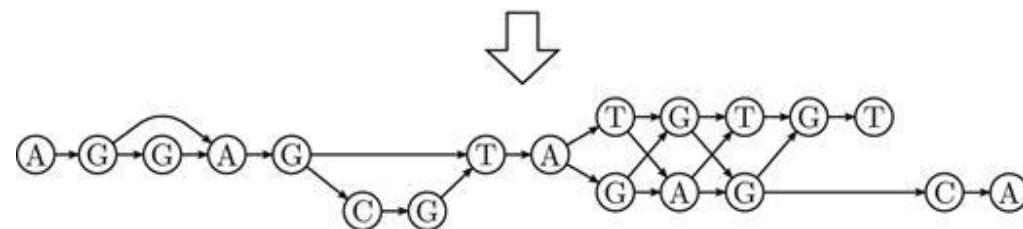
↓
(E) Gap at nearest search terminus in
candidate removed

↓
(F) Gap in template removed

Template: ATAC-----GT-AAC-----GTA---C---G-T-AC-GG
Candidate: C-AC-----GTTAAC----GT----C---G-TACCGG

SINA (SILVA Incremental Aligner)

A G G A G - - T A T A T G T - -
A G - A G C G T A T C T G T - -
A G G A G - - T A G C G G T - -
A G G A G - - T A G A G - - C A



- DeSantis *et al.* (2006) "NAST: A Multiple Sequence Alignment Server for Comparative Analysis of 16S rRNA Genes." *Nucleic Acids Research* 34 (Web Server): W394–99.
- Caporaso *et al.* (2010) "PyNAST: A Flexible Tool for Aligning Sequences to a Template Alignment." *Bioinformatics* 26 (2): 266–67.
- Pruesse *et al.* (2012) "SINA: Accurate High-Throughput Multiple Sequence Alignment of Ribosomal RNA Genes." *Journal of Gerontology* 28 (14): 1823–29.

TTGCAGTTGATACT**GAT**ATCTT-
CTGCGTTCTGAAC**GGT**GACTA-
-CGCTTGAAACT**TTTAA**CTTG
TTGCAGTTGATACT**GAT**GTCTT-
TTGCAGTTGAAACT**GCAGT**CTT-
TTGCATTTCATACT**GGT**CGCTA-
·
1111111111111111**000101111**



Apply mask

TTGCAGTTGATACTGACTT-
CTGCGTTCTGAAC**GG**GCTA-
-CGCTTGAAACTGACTTG
TTGCAGTTGATACT**GG**CTT-
TTGCAGTTGAAACT**GG**CTT-
TTGCATTTCATACT**GC**CTA-

Why “mask” alignment columns?

- Proposed by David Lane (1991)
 - “Lane mask”
- Remove errors introduced by alignment heuristics
 - i.e. incorrect statements of positional homology.
- Eliminate phylogenetically uninformative / misleading sites
- Remove repetitive / homopolymeric regions (TATATATA / AAAAAAAA)
- These obfuscate phylogenetic inference.

*** Reminder: Sequences can be aligned differently by different programs, and the resulting inferred phylogenies can differ substantially! Often additional algorithmic and/or manual curation is required. Can affect masking!*

Caveats to masking alignment columns.

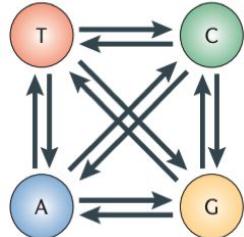
- hard-to-align regions that contain phylogenetically useful information may be rejected due to a lack of confidence in the alignments
 - Is your mask too aggressive?
 - Masking increases the similarity of the sequences, potentially making them identical.
 - Potential loss of phylogenetic resolution.
 - Might detract from the benefits gained by using Exact Sequence Variants?

Goal: Balance out these opposing factors in such a way that enables you to best address your research questions. There is no one-size fits all!

JC69

K80

HKY85



Substitution Model

PyCUT
PurAG

(a) Step 1
Assemble pseudo-datasets, repeat 1000 times

Replicate 1

1562314951
seqA CTCCGTTTC
seqB TTGGTTATT
seqC TTCCGTAATT

Replicate 2

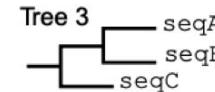
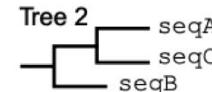
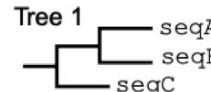
5234924418
seqA TCGTTCTCG
seqB TGGTAGTTT
seqC TCGAACAAATG

Replicate 3

5607718907
seqA TCAGGCGTAG
seqB TCAAATGAAA
seqC TCAGGTGAAG

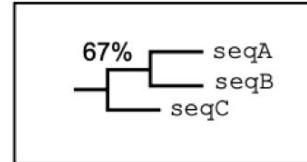
etc

(b) Step 2
Build trees for each
pseudo-dataset
to give 1000 trees



etc

(c) Step 3
Tabulate results
(strict consensus tree)



Bootstrap consensus tree

Bootstrapping

Dataset

0123456789
seqA ACCGTTCGGT
seqB ATGGTTACAGA
seqC ATCGATCGGA

Reference based phylogenies.



AMERICAN
SOCIETY FOR
MICROBIOLOGY



RESEARCH ARTICLE

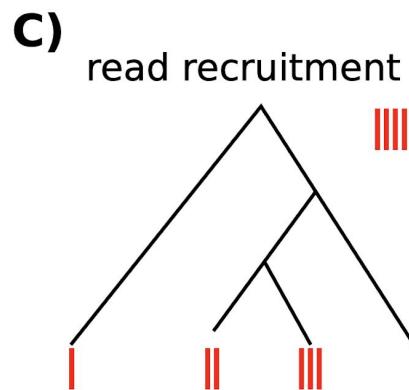
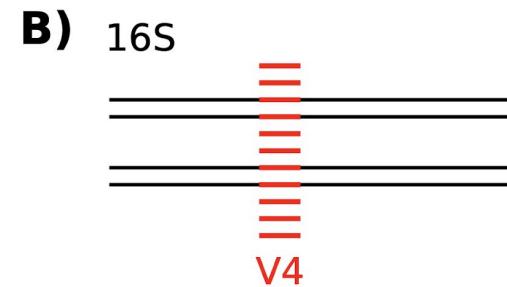
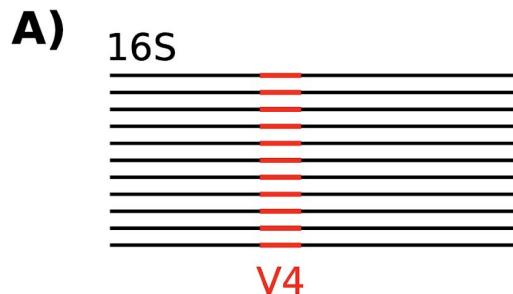
Ecological and Evolutionary Science



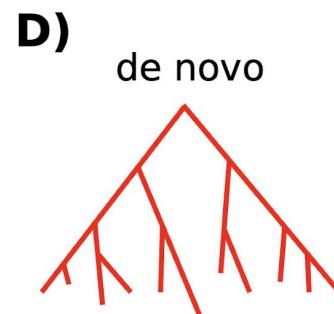
Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information

Stefan Janssen,^a Daniel McDonald,^a Antonio Gonzalez,^a Jose A. Navas-Molina,^b Lingjing Jiang,^d Zhenjiang Zech Xu,^a Kevin Winker,^c Deborah M. Kado,^d Eric Orwoll,^e Mark Manary,^f Slavash Mirarab,^{g,h} Rob Knight^{a,b,g}

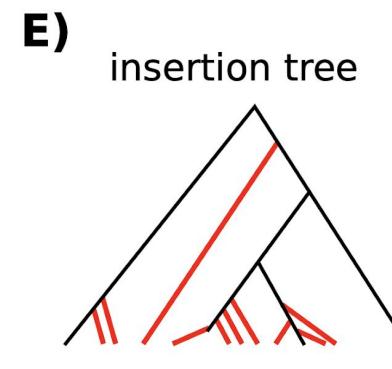
<https://github.com/biocore/q2-fragment-insertion>



pro: reference phylogeny
con: losing sOTUs



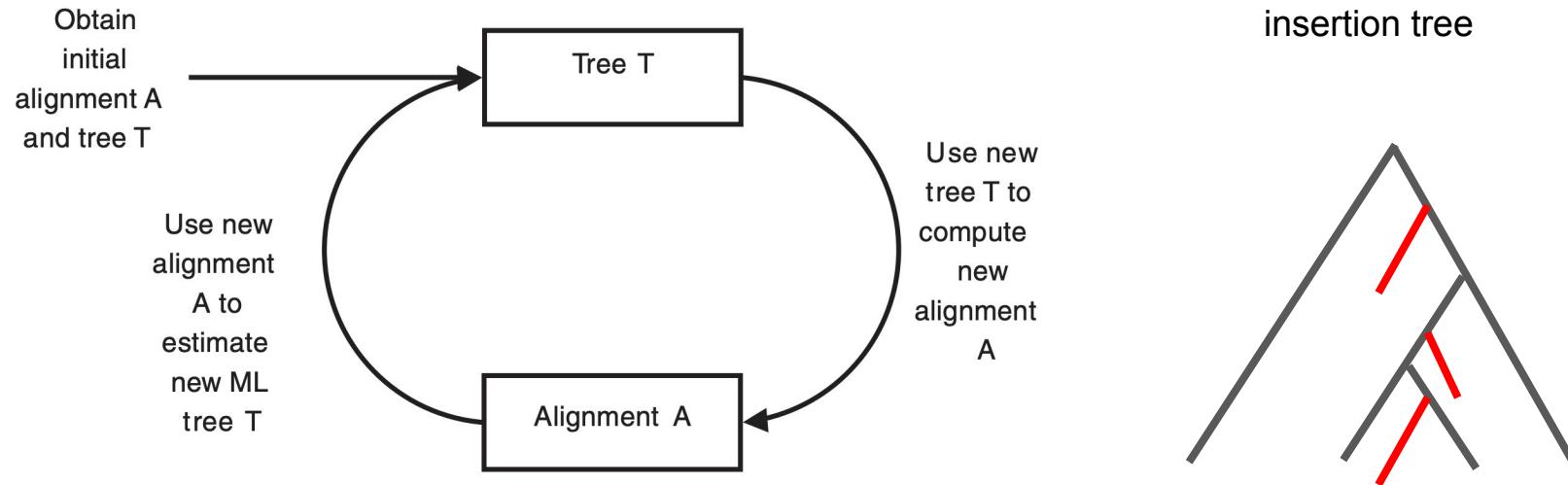
con: no reference
pro: keep all sOTUs



pro: reference phylogeny
pro: keep most sOTUs

SATé == Simultaneous Alignment and Tree Estimation

SEPP == SATé-Enabled Phylogenetic Placement



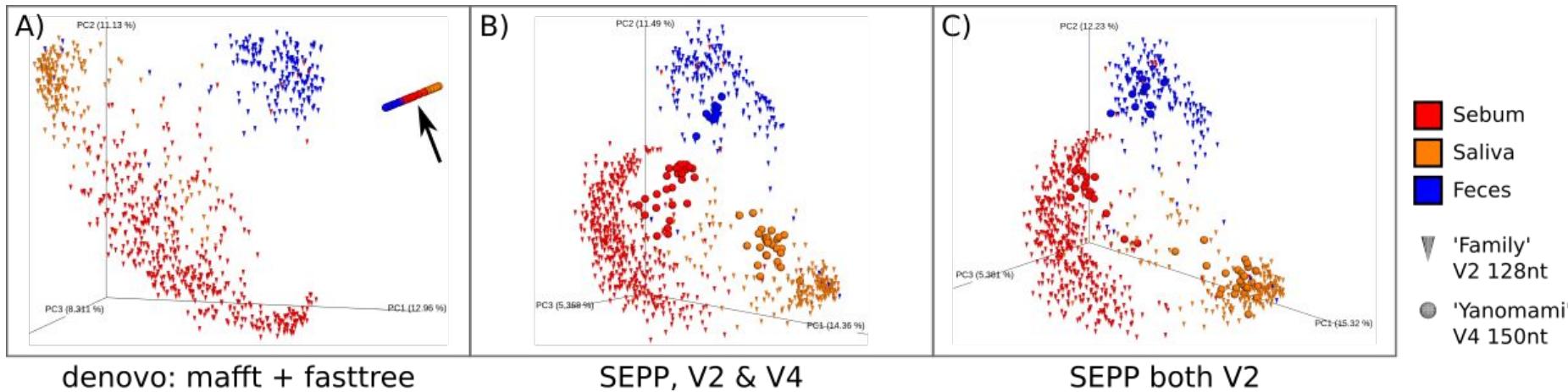
Liu *et al.* 2009. "Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees." *Science* 324(5934): 1561–64.

Liu *et al.* (2012) "SATe-II: Very Fast and Accurate Simultaneous Estimation of Multiple Sequence Alignments and Phylogenetic Trees." *Systematic Biology* 61(1): 90–106.

Mirarab *et al.* (2012) "SEPP: SATé-Enabled Phylogenetic Placement." In *Biocomputing* 247–58. World Scientific.

Janssen *et al.* (2018) "Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information." *mSystems* 3(3): e00021–18.

meta analysis



two studies, different variable regions

SEPP / fragment insertion reference files are available on the [Data resources](#) page on the QIIME 2 website.

Table of Contents

- Getting started
- What is QIIME 2?
- Core concepts
- Installing QIIME 2
- Tutorials
- Interfaces
- Plugins
- Semantic types
- Community
- Data resources
- Supplementary resources
- User Glossary
- Citing QIIME 2

SEPP reference databases

The following databases are intended for use with q2-fragment-insertion, and are constructed directly from the [SEPP-Refs](#) project.

- Silva 128 SEPP reference database (MD5: `7879792a6f42c5325531de9866f5c4de`)
- Greengenes 13_8 SEPP reference database (MD5: `9ed215415b52c362e25cb0a8a46e1076`)

Checkout the phylogeny tutorials at:

<https://docs.qiime2.org/2020.6/tutorials/phylogeny/>

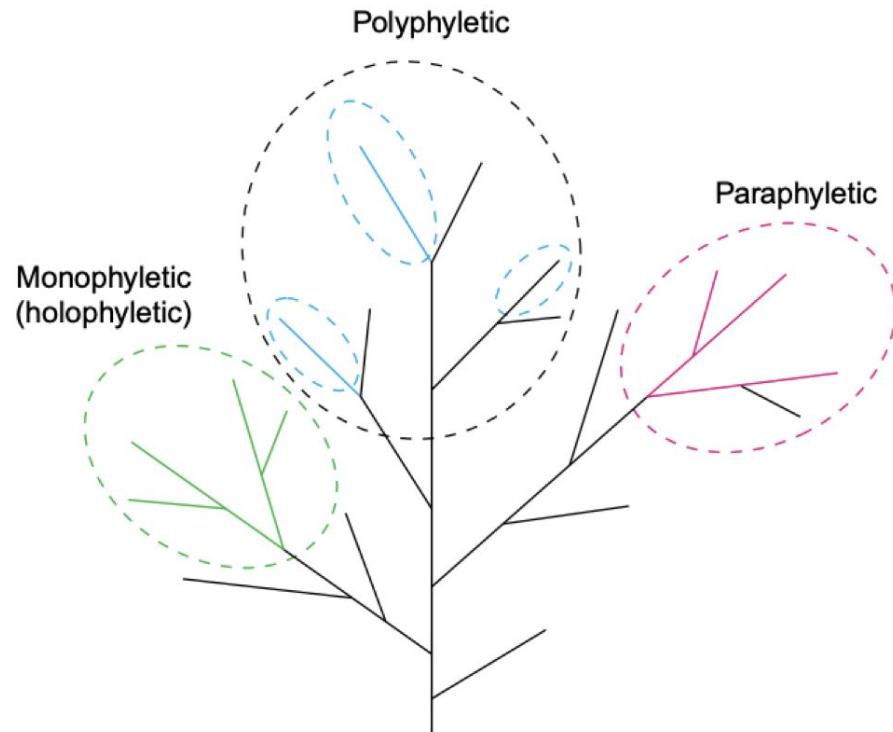
<https://github.com/qiime2/q2-fragment-insertion>

```
qiime alignment --help
```

```
qiime phylogeny --help
```

```
qiime fragment-insertion --help
```

May all your trees be fully resolved... or nearly so!

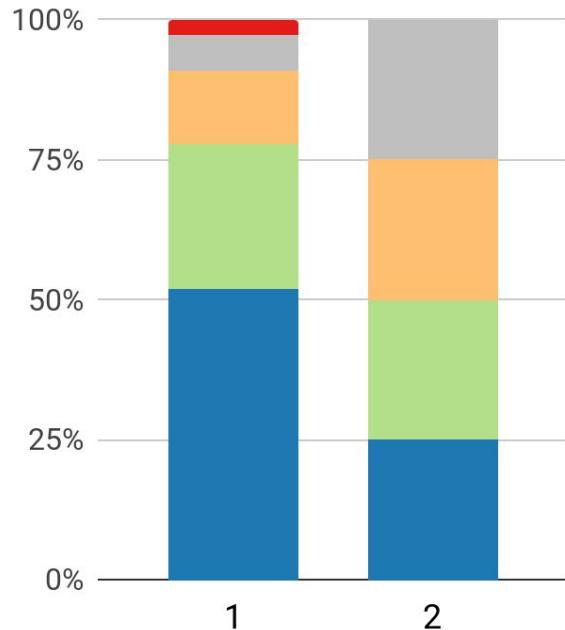


Rarefaction

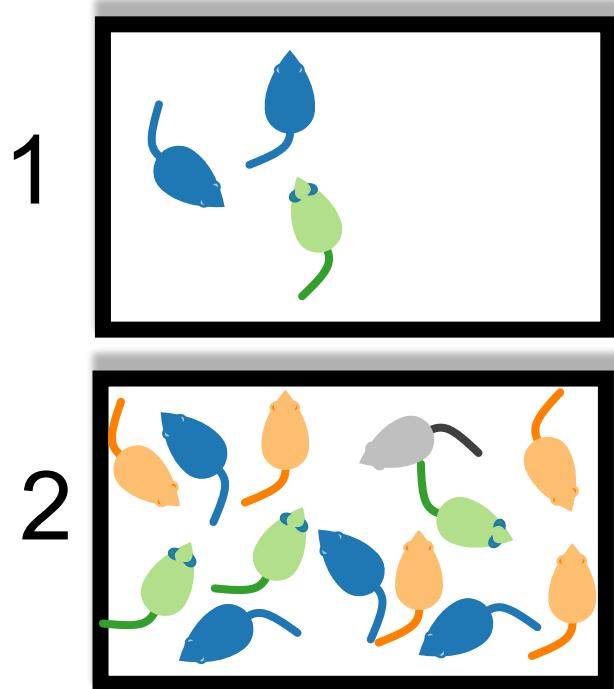
<https://bit.ly/2HThBcx>

A theoretical sampling problem

Actual distribution



Observed distribution



Does anything concern you about this table?

FeatureTable[Frequency]					
	feature1	feature2	feature3	feature4	feature5
4ac2	84	1	73	198	2
e375	24	2	44	176	1
4gd8	11	0	10	30	0
9872	0	0	25	2	0

Diversity metrics in ordinations are often impacted by the total frequency observed in samples, such that in this example 9872 (wild type genotype) might look more similar to 4gd8 (susceptible genotype) than to e375 (wild type genotype).

FeatureTable[Frequency]					
	feature1	feature2	feature3	feature4	feature5
4ac2	84	1	73	198	2
e375	24	2	44	176	1
4gd8	11	0	10	30	0
9872	0	0	25	2	0

	Total frequency
4ac2	358
e375	247
4gd8	51
9872	27

This is most commonly handled by *rarefaction*, which is currently* a necessary evil. Frequencies are subsampled without replacement until all samples have the same total. Samples with fewer sequences than your *even sampling depth* will be filtered out of the feature table.

FeatureTable[Frequency]					
	feature1	feature2	feature3	feature4	feature5
g345	11	1	10	29	0
c5d7	4	0	7	40	0
f6ee	11	0	10	30	0
efd3	0	0	0	0	0

	Total frequency
g345	51
c5d7	51
f6ee	51
efd3	0

* A good project would be developing diversity metrics that are not sensitive to total frequency.

α -diversity Metrics

<https://bit.ly/2HThBcx>

sample-metadata.tsv

sample-metadata.tsv	
SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

table.qza

FeatureTable[Frequency]					
	feature 1	feature 2	feature 3	Feature 4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

rooted-tree.qza
Phylogeny [Rooted]



Diversity and statistical analyses

Comparing microbial communities

How similar or different are the things in the community from each other?

Alpha diversity

How similar or different are two communities?

What specifically leads to differences between the communities?

What features of microbiomes differ?

Community richness (often referred to as alpha diversity)



Image source:
<http://miriadna.com/desktopwalls/images/max/Field-of-yellow-tulips.jpg>



Image source:
<https://imgflip.com/mememplate/62338435/Flower-garden>

Comparing microbial communities

How similar or different are the things in the community from each other?

Alpha diversity

How similar or different are two communities?

Beta diversity

What specifically leads to differences between the communities?

What features of microbiomes differ?

Community composition (often referred to as beta diversity)



Image source:
<https://shawnacoronado.com/front-lawn-vegetable-garden-design/>



Image source:
<https://imgflip.com/mememplate/62338435/Flower-garden>

Comparing microbial communities

How similar or different are the things in the community from each other?

Alpha diversity

How similar or different are two communities?

Beta diversity

What specifically leads to differences between the communities?

Taxonomic profiling, differential

What features of microbiomes differ?

Community composition (often referred to as beta diversity)



Image source:
<https://shawnacoronado.com/front-lawn-vegetable-garden-design/>



Image source:
<https://imgflip.com/mememplate/62338435/Flower-garden>

Why incorporate phylogeny in a diversity metric?

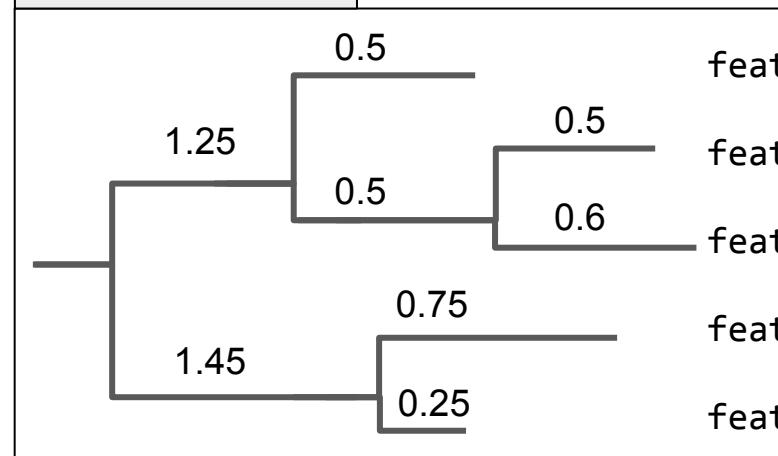
FeatureTable[PresenceAbsence]

	feature1	feature2	feature3	feature4	feature5
4ac2	1	1	1	0	0
e375	0	1	1	1	0

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	25	30	15	0	0
e375	0	17	33	25	0

Phylogeny[Rooted]



Why incorporate phylogeny in a diversity metric?

FeatureTable[PresenceAbsence]

	feature1	feature2	feature3	feature4	feature5
4ac2	1	1	1	0	0
e375	0	1	1	1	0

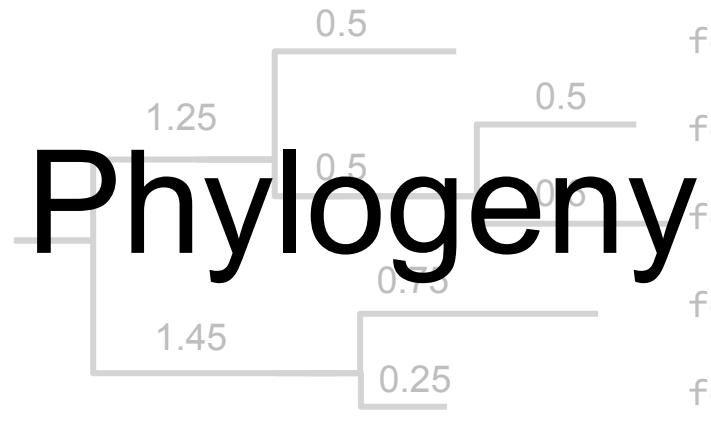
Richness

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	25	30	15	0	0
e375	0	17	33	25	0

Evenness

Phylogeny[Rooted]



Observed OTUs (or Observed Species): non-phylogenetic, alpha diversity metric measuring richness

FeatureTable [Frequency]		feature1	feature2	feature3	feature4	feature5
4ac2		25	30	15	0	0
e375		0	17	33	25	0



SampleData [AlphaDiversity]	
	Observed OTUs
4ac2	
e375	

Count the number of different features in a sample.

Observed OTUs (or Observed Species): non-phylogenetic, alpha diversity metric measuring richness

FeatureTable [Frequency]		feature1	feature2	feature3	feature4	feature5
	4ac2	25	30	15	0	0
e375	0	17	33	25	0	



SampleData [AlphaDiversity]	
	Observed OTUs
4ac2	
e375	

Count the number of different features in a sample.

Observed OTUs (or Observed Species): non-phylogenetic, alpha diversity metric measuring richness

FeatureTable [Frequency]		feature1	feature2	feature3	feature4	feature5
	4ac2	1	1	1	0	0
e375	0	1	1	1	1	0



SampleData [AlphaDiversity]	
	Observed OTUs
4ac2	
e375	

Count the number of different features in a sample.

Observed OTUs (or Observed Species): non-phylogenetic, alpha diversity metric measuring richness

FeatureTable [Frequency]		feature1	feature2	feature3	feature4	feature5
	4ac2	1	1	1	0	0
	e375	0	1	1	1	0



SampleData [AlphaDiversity]		
	Observed OTUs	
	4ac2	3
	e375	3

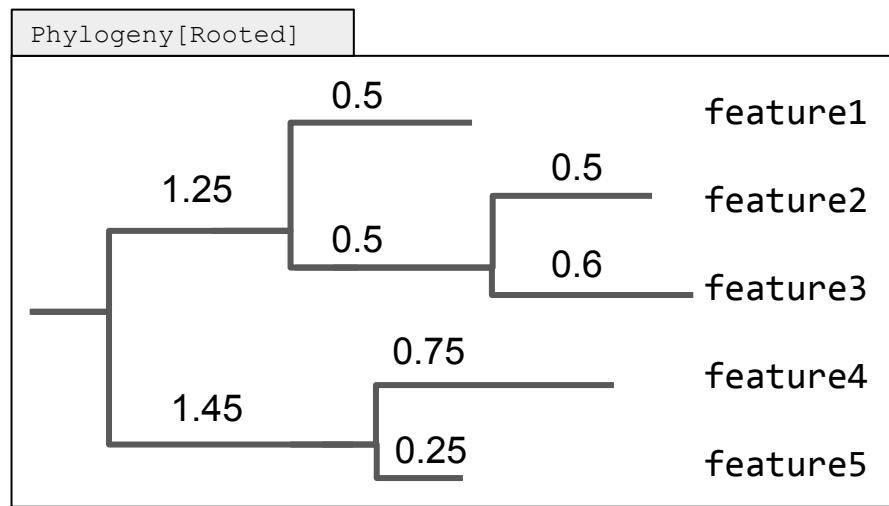
Count the number of different features in a sample.

Why incorporate phylogeny in a diversity metric?

FeatureTable [Frequency]					
	feature1	feature2	feature3	feature4	feature5
4ac2	25	30	15	0	0
e375	0	17	33	25	0

→

SampleData [AlphaDiversity]	
	Observed OTUs
4ac2	3
e375	3



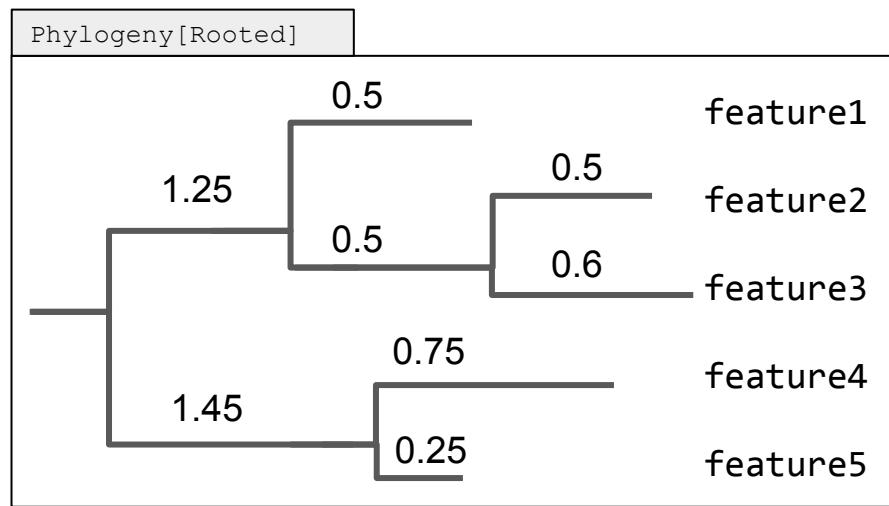
FeatureData [Taxonomy]	
	Domain
feature1	Bacteria
feature2	Bacteria
feature3	Bacteria
feature4	Archaea
feature5	Archaea

Why incorporate phylogeny in a diversity metric?

FeatureTable [Frequency]					
	feature1	feature2	feature3	feature4	feature5
4ac2	25	30	15	0	0
e375	0	17	33	25	0

→

SampleData [AlphaDiversity]	
	Observed OTUs
4ac2	3
e375	3



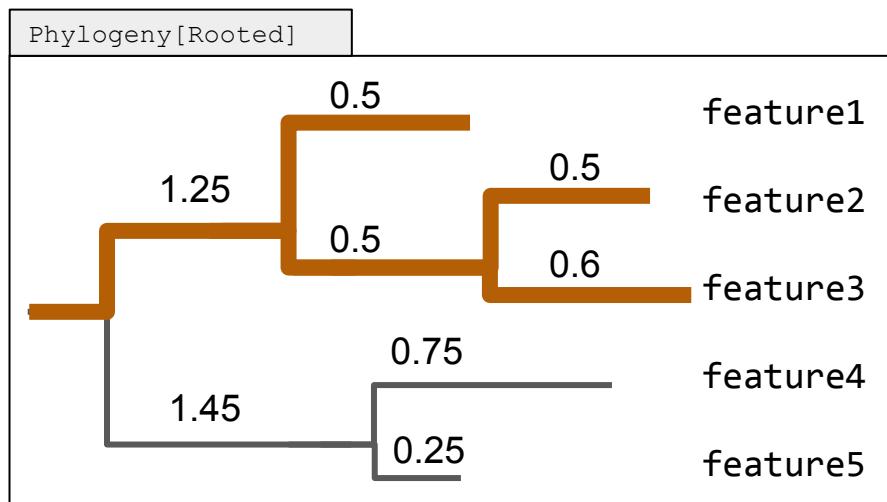
FeatureData [Taxonomy]	
	Domain
feature1	Bacteria
feature2	Bacteria
feature3	Bacteria
feature4	Archaea
feature5	Archaea

Faith's Phylogenetic Diversity

FeatureTable [Frequency]		feature1	feature2	feature3	feature4	feature5
	4ac2	25	30	15	0	0
	e375	0	17	33	25	0

→

SampleData [AlphaDiversity]	
	Faith's PD
4ac2	3.35
e375	



Sum of branch length covered by a sample.

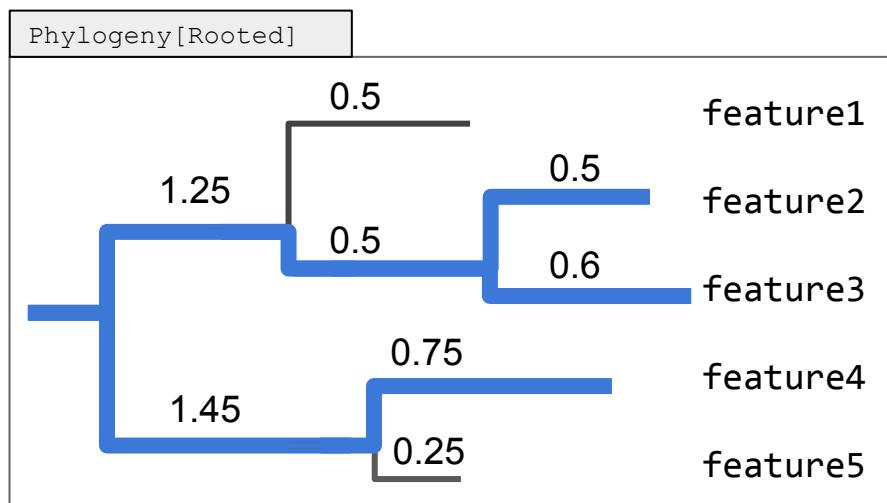
Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biological Conservation. 61:1-10.

Faith's Phylogenetic Diversity

FeatureTable [Frequency]					
	feature1	feature2	feature3	feature4	feature5
4ac2	25	30	15	0	0
e375	0	17	33	25	0

→

SampleData [AlphaDiversity]	
	Faith's PD
4ac2	3.35
e375	5.05



Sum of branch length covered by a sample.

Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biological Conservation. 61:1-10.

Shannon Diversity Index:

non-phylogenetic, alpha diversity metric measuring richness and evenness

$$H' = - \sum_{i=1}^s p_i \ln p_i$$

FeatureTable [Frequency]					
	feature1	feature2	feature3	feature4	feature5
4ac2	25	30	15	0	0
e375	0	17	33	25	0



SampleData [AlphaDiversity]	
	Shannon
4ac2	
e375	

Shannon Diversity Index:

non-phylogenetic, alpha diversity metric measuring richness and evenness

$$H' = - \sum_{i=1}^s p_i \ln p_i$$

FeatureTable [Frequency]		feature1	feature2	feature3	feature4	feature5
4ac2		25	30	15	0	0
e375		0	17	33	25	0



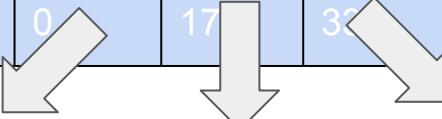
SampleData [AlphaDiversity]	
	Shannon
4ac2	
e375	

Shannon Diversity Index:

non-phylogenetic, alpha diversity metric measuring richness and evenness

$$H' = - \sum_{i=1}^s p_i \ln p_i$$

FeatureTable [Frequency]					
	feature1	feature2	feature3	feature4	feature5
4ac2	25	30	15	0	0
e375	0	17	32	25	0



SampleData [AlphaDiversity]	
	Shannon
4ac2	
e375	

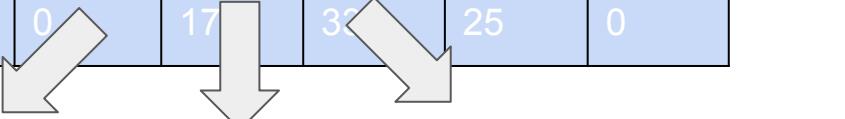
$$H' = - (0.375(-1.030) + 0.429(-0.847) + 0.214 (-1.540))$$

Shannon Diversity Index:

non-phylogenetic, alpha diversity metric measuring richness and evenness

$$H' = - \sum_{i=1}^s p_i \ln p_i$$

FeatureTable [Frequency]					
	feature1	feature2	feature3	feature4	feature5
4ac2	25	30	15	0	0
e375	0	17	32	25	0



SampleData [AlphaDiversity]	
	Shannon
4ac2	1.061
e375	

$$H' = - (0.375(-1.030) + 0.429(-0.847) + 0.214 (-1.540))$$

Shannon Diversity Index:

non-phylogenetic, alpha diversity metric measuring richness and evenness

The diagram illustrates the process of calculating alpha diversity (Shannon Index) from a FeatureTable [Frequency].

FeatureTable [Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	25	30	15	0	0
e375	0	17	33	25	0

SampleData [AlphaDiversity]

	Shannon
4ac2	1.061
e375	1.064

An arrow points from the FeatureTable to the SampleData table, indicating the transformation.

Pielou's Evenness Index: non-phylogenetic, alpha diversity metric measuring evenness

$$H' = - \sum_{i=1}^S p_i \ln p_i$$

$$J' = \frac{H'}{H'_{max}}$$

FeatureTable [Frequency]					
	feature1	feature2	feature3	feature4	feature5
4ac2	25	30	15	0	0
e375	0	17	33	25	0



SampleData [AlphaDiversity]	
	Shannon
4ac2	1.061
e375	1.064

Pielou's Evenness Index: non-phylogenetic, alpha diversity metric measuring evenness

$$H' = - \sum_{i=1}^S p_i \ln p_i$$

$$J' = \frac{H'}{H'_{max}}$$

FeatureTable [Frequency]					
	feature1	feature2	feature3	feature4	feature5
4ac2	25	30	15	0	0
e375	0	17	33	25	0



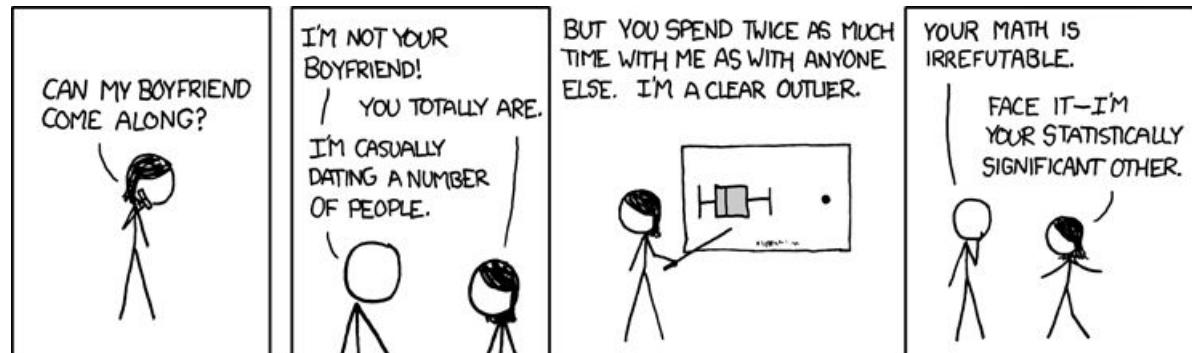
SampleData [AlphaDiversity]	
	Pielou_e
4ac2	1.000
e375	1.003

4 good metrics to know

- Observed Features
 - Richness
 - Number of different things present
- Faith's Phylogenetic Diversity
 - Phylogenetic richness
 - Cumulative evolutionary history
- Shannon Diversity
 - Richness and evenness (weighted by abundance)
- Pielou's Evenness
 - Only evenness (feature distribution)

Statistical and Display Properties of Alpha Diversity

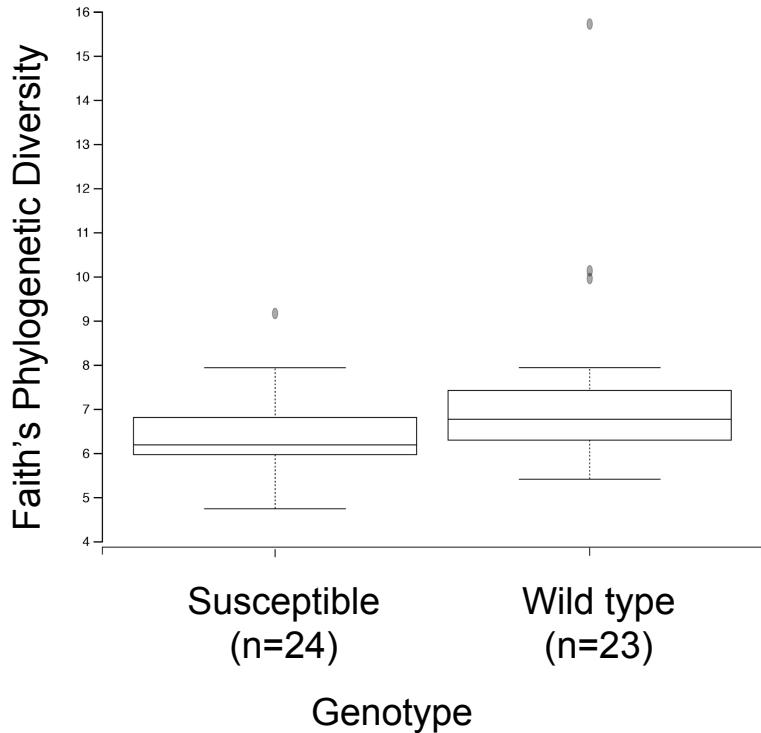
- Univariate and continuous
- Generally treated as non-normal, but always worth trying
- Fits in most of your favorite models and visualizations for continuous data!!!!



"Boyfriend". [XKCD](#) (593).

Alpha diversity comparison

- visually
 - distribution comparison plots (discrete)
 - scatter plots (continuous)
- statistically
 - Kruskal-Wallis (discrete data)
 - Spearman correlation (continuous)
 - Regression (when asymptotically normal)



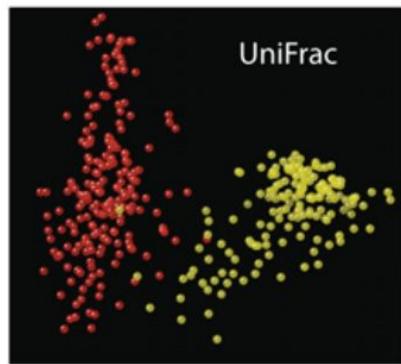
β -diversity Metrics

<https://bit.ly/2HThBcx>

We measure Beta Diversity using distance metrics...

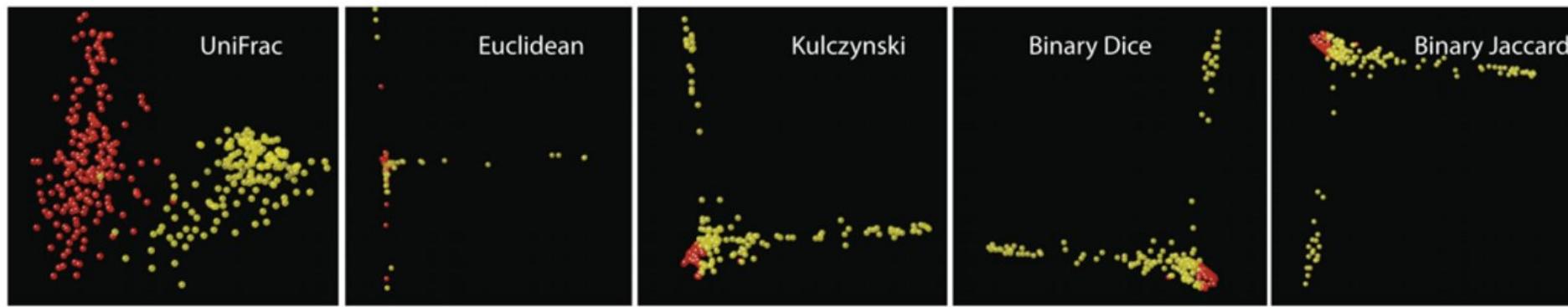
...But what is a distance?

Selecting a metric



- Vertebrate Gut
- Free living

Selecting a metric



- Vertebrate Gut
- Free living

Commonly used distance metrics for microbiome data:

- Jaccard
- Bray-Curtis
- Unweighted Unifrac
- Weighted Unifrac
- Generalized Unifrac

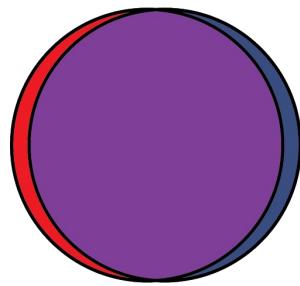
Distance Axioms

For all x_i , x_j and x_k

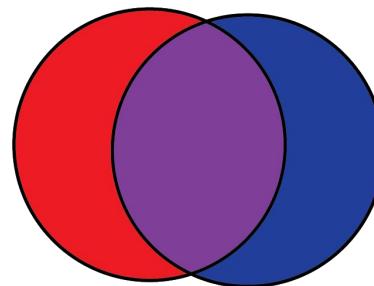
1. $d(x_i, x_j) \geq 0.$
2. $d(x_i, x_j) = 0$, iff x_i is equal to $x_j.$
3. $d(x_i, x_j) = d(x_j, x_i).$
4. $d(x_i, x_j) \leq d(x_i, x_k) + d(x_k, x_j).$

Jaccard distance (1908)

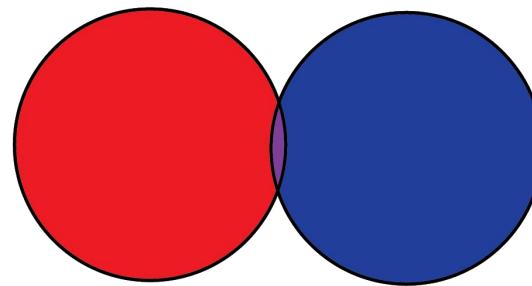
Fraction of unique features, regardless of abundance.



$$d_j \approx 0$$



$$d_j \approx 0.5$$



$$d_j \approx 1$$

Jaccard distance:

a qualitative, non-phylogenetic beta diversity metric

FeatureTable [Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

$$J(A, B) = 1 - \frac{A \cap B}{A \cup B}$$



DistanceMatrix

	4ac2	e375	4gd8	9872
4ac2				
e375				
4gd8				
9872				

Jaccard distance:

a qualitative, non-phylogenetic beta diversity metric

FeatureTable [Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

$$J(A, B) = 1 - \frac{A \cap B}{A \cup B}$$



	4ac2	e375	4gd8	9872
4ac2	0.0			
e375		0.0		
4gd8			0.0	
9872				0.0

Jaccard distance:

a qualitative, non-phylogenetic beta diversity metric

FeatureTable [Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

$$J(A, B) = 1 - \frac{A \cap B}{A \cup B}$$



DistanceMatrix

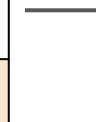
	4ac2	e375	4gd8	9872
4ac2	0.0			
e375		0.0		
4gd8			0.0	
9872				0.0

Jaccard distance:

a qualitative, non-phylogenetic beta diversity metric

FeatureTable [Frequency]		feature1	feature2	feature3	feature4	feature5
		42	0	37	99	1
4ac2		12	1	22	88	0
e375		25	3	23	86	0
4gd8		0	0	87	12	0
9872						

$$J(A, B) = 1 - \frac{A \cap B}{A \cup B}$$



	4ac2	e375	4gd8	9872
4ac2	0.0			
e375		0.0		
4gd8			0.0	
9872				0.0

Jaccard distance:

a qualitative, non-phylogenetic beta diversity metric

FeatureTable [Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

$$J(A, B) = 1 - \frac{A \cap B}{A \cup B}$$



DistanceMatrix

	4ac2	e375	4gd8	9872
4ac2	0.0			
e375	0.4	0.0		
4gd8	0.4	0.0	0.0	
9872	0.5	0.5	0.5	0.0

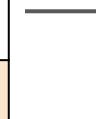
Jaccard distance:

a qualitative, non-phylogenetic beta diversity metric

FeatureTable [Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

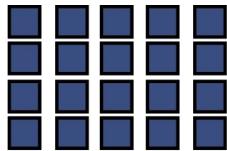
$$J(A, B) = 1 - \frac{A \cap B}{A \cup B}$$



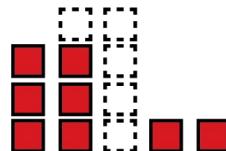
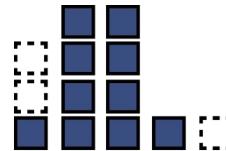
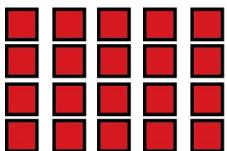
	4ac2	e375	4gd8	9872
4ac2	0.0	0.4	0.4	0.5
e375	0.4	0.0	0.0	0.5
4gd8	0.4	0.0	0.0	0.5
9872	0.5	0.5	0.5	0.0

Bray-Curtis distance (1948)

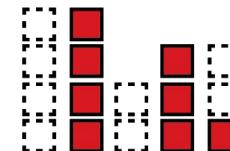
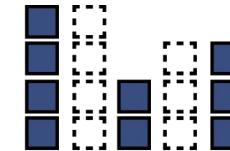
Fraction of overabundant counts.



$$d_{bc} = 0$$



$$d_{bc} \approx 0.5$$



$$d_{bc} \approx 1$$

$$\frac{0 \times \boxed{}}{20 \times \boxed{\text{Red}} + 20 \times \boxed{\text{Blue}}} \quad \boxed{}$$

$$\frac{8 \times \boxed{}}{8 \times \boxed{\text{Red}} + 10 \times \boxed{\text{Blue}}} \quad \boxed{}$$

$$\frac{16 \times \boxed{}}{8 \times \boxed{\text{Red}} + 9 \times \boxed{\text{Blue}}} \quad \boxed{}$$

Bray-Curtis distance:

a quantitative, non-phylogenetic beta diversity metric

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

$$BC(A, B) = \frac{\sum_i |X_{iA} - X_{iB}|}{\sum_i (X_{iA} + X_{iB})}$$

X_{iA} : frequency of feature i in sample A

	4ac2	e375	4gd8	9872
4ac2	0.0			
e375		0.0		
4gd8			0.0	
9872				0.0

Bray-Curtis distance:

a quantitative, non-phylogenetic beta diversity metric

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

$$BC(A, B) = \frac{\sum_i |X_{iA} - X_{iB}|}{\sum_i (X_{iA} + X_{iB})}$$

X_{iA} : frequency of feature i in sample A

	4ac2	e375	4gd8	9872
4ac2	0.0			
e375	0.19	0.0		
4gd8	0.15	0.07	0.0	
9872	0.65	0.69	0.70	0.0

Bray-Curtis distance:

a quantitative, non-phylogenetic beta diversity metric

FeatureTable[Frequency]

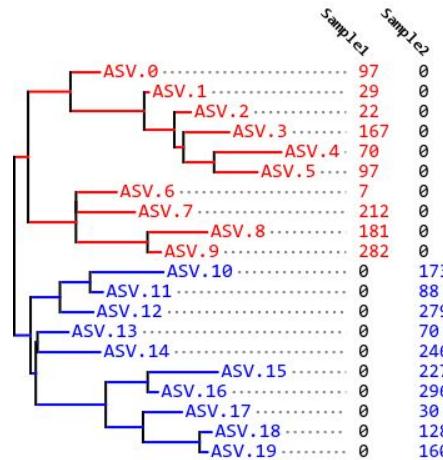
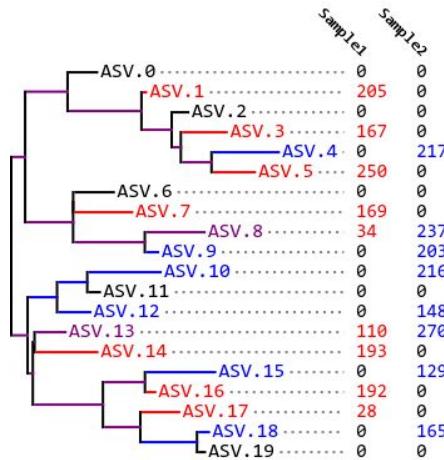
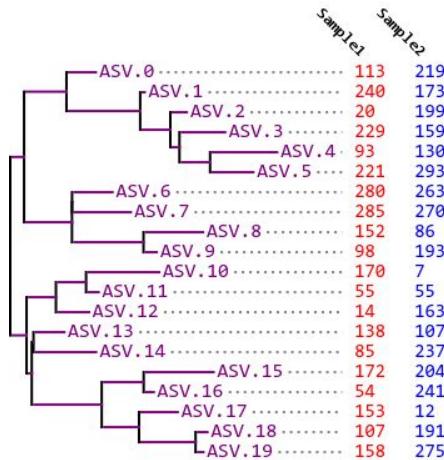
	feature1	feature2	feature3	feature4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

$$BC(A, B) = \frac{\sum_i |X_{iA} - X_{iB}|}{\sum_i (X_{iA} + X_{iB})}$$

X_{iA} : frequency of feature i in sample A

	4ac2	e375	4gd8	9872
4ac2	0.0	0.19	0.15	0.65
e375	0.19	0.0	0.07	0.69
4gd8	0.15	0.07	0.0	0.70
9872	0.65	0.69	0.70	0.0

Unweighted UniFrac distance: a qualitative, phylogenetic beta diversity metric

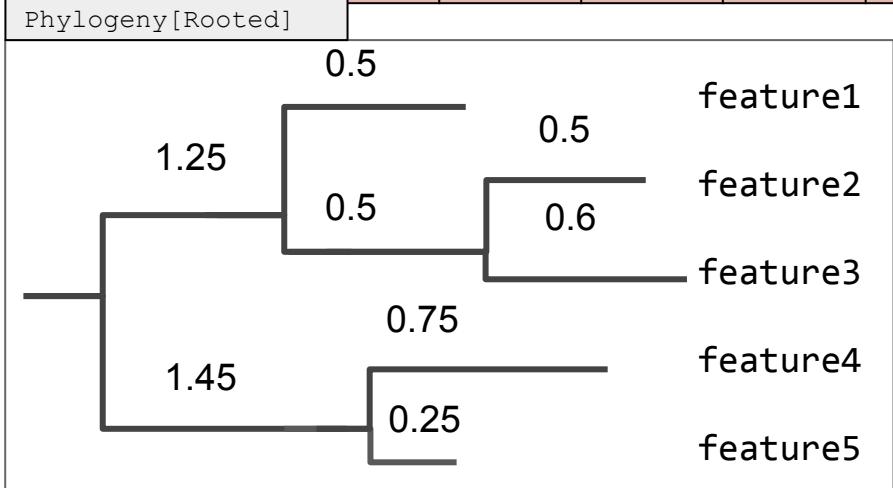


$$UU(A, B) = \frac{\text{sum of unique branch length}}{\text{sum of observed branch length}}$$

Unweighted UniFrac

a qualitative, phylogenetic beta diversity metric

	feature1	feature2	feature3	feature4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0



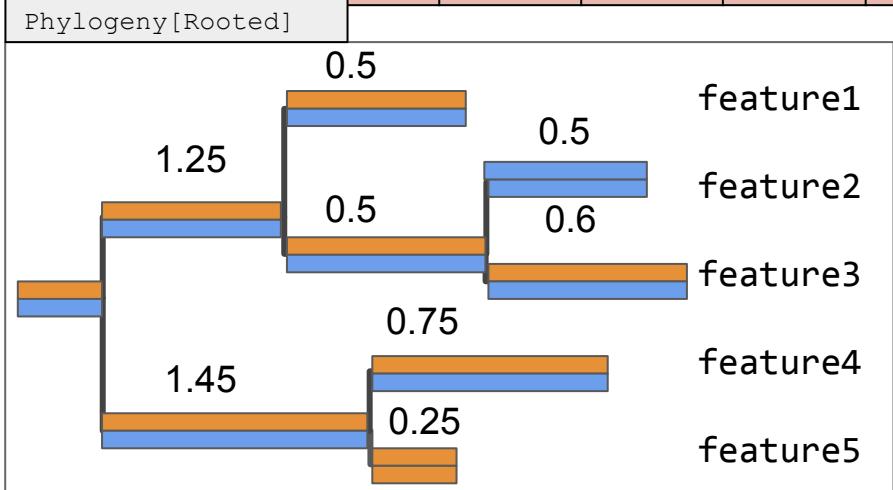
DistanceMatrix

	4ac2	e375	4gd8
4ac2	0.0		
e375		0.0	
4gd8			0.0

Unweighted UniFrac

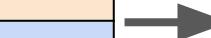
a qualitative, phylogenetic beta diversity metric

	feature1	feature2	feature3	feature4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0



DistanceMatrix

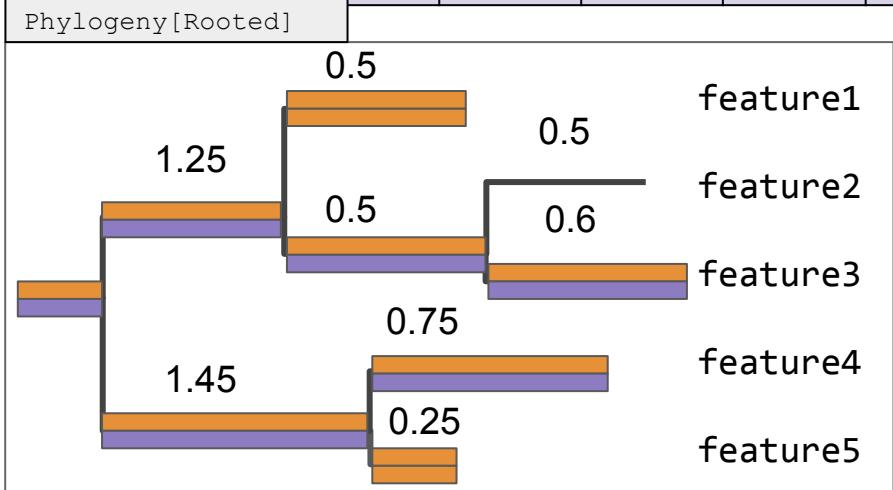
	4ac2	e375	4gd8
4ac2	0.0		
e375	0.13	0.0	
4gd8			0.0



Unweighted UniFrac

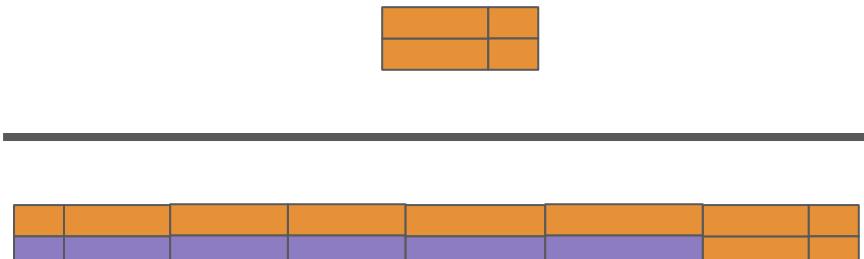
a qualitative, phylogenetic beta diversity metric

	feature1	feature2	feature3	feature4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
9872	0	0	87	12	0



DistanceMatrix

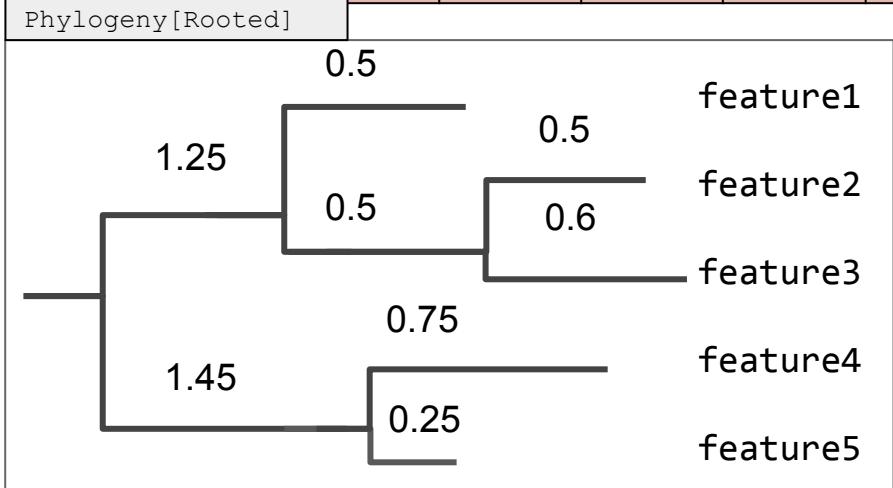
	4ac2	e375	4gd8
4ac2	0.0		
e375	0.13	0.0	
4gd8	0.14		0.0



Unweighted UniFrac

a qualitative, phylogenetic beta diversity metric

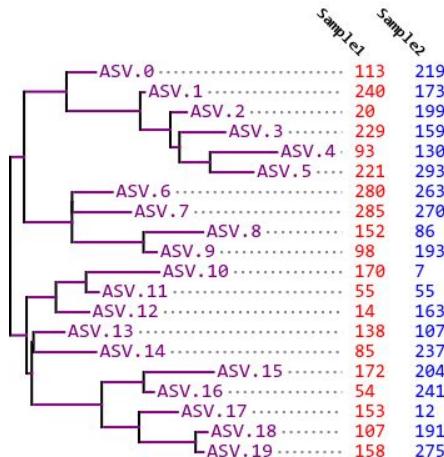
	feature1	feature2	feature3	feature4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0



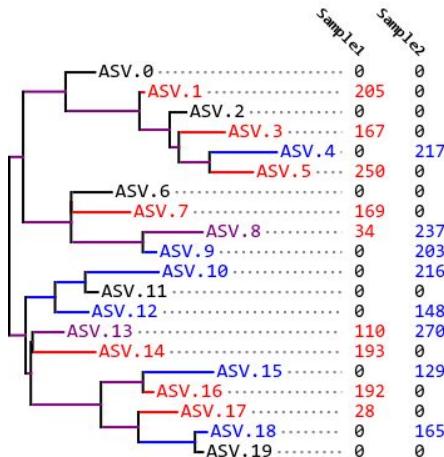
DistanceMatrix

	4ac2	e375	4gd8
4ac2	0.0	0.13	0.14
e375	0.13	0.0	0.18
4gd8	0.14	0.18	0.0

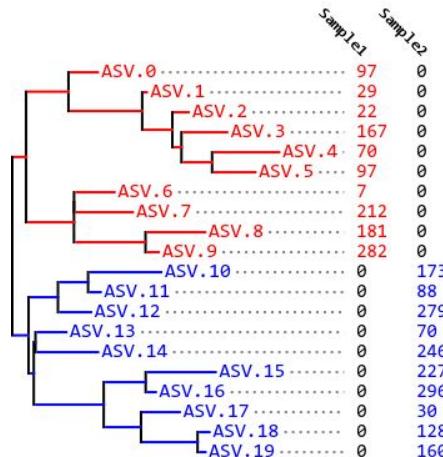
Weighted UniFrac distance: a qualitative, phylogenetic beta diversity metric



Unweighted UniFrac = 0.000000
Weighted UniFrac = 0.118186



Unweighted UniFrac = 0.526659
Weighted UniFrac = 0.560364



Unweighted UniFrac = 1.000000
Weighted UniFrac = 1.000000

$$UU(A, B) = \frac{\text{sum of unique branch length}}{\text{sum of observed branch length}}$$

Generalized UniFrac

<https://docs.qiime2.org/2020.8/plugins/available/diversity/beta-phylogenetic/>

[Generalized UniFrac Paper](#)

Performs UniFrac distance calculations that are in-between weighted and unweighted

Weighted UniFrac has biases towards the most abundance taxa

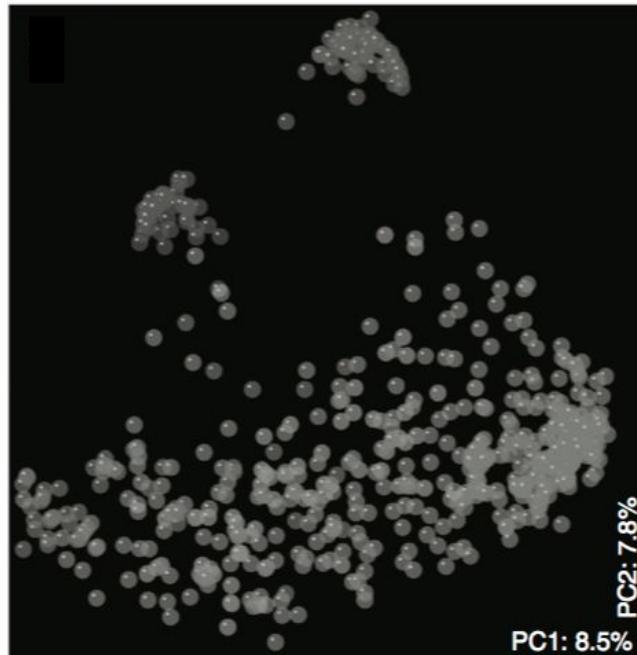
Unweighted UniFrac has biases towards the rare taxa

Generalized UniFrac ($\alpha=0.5$) provides a happy medium, although it is still suggest you explore all your options to gain biological insights

Fantastic matrices and where to see them

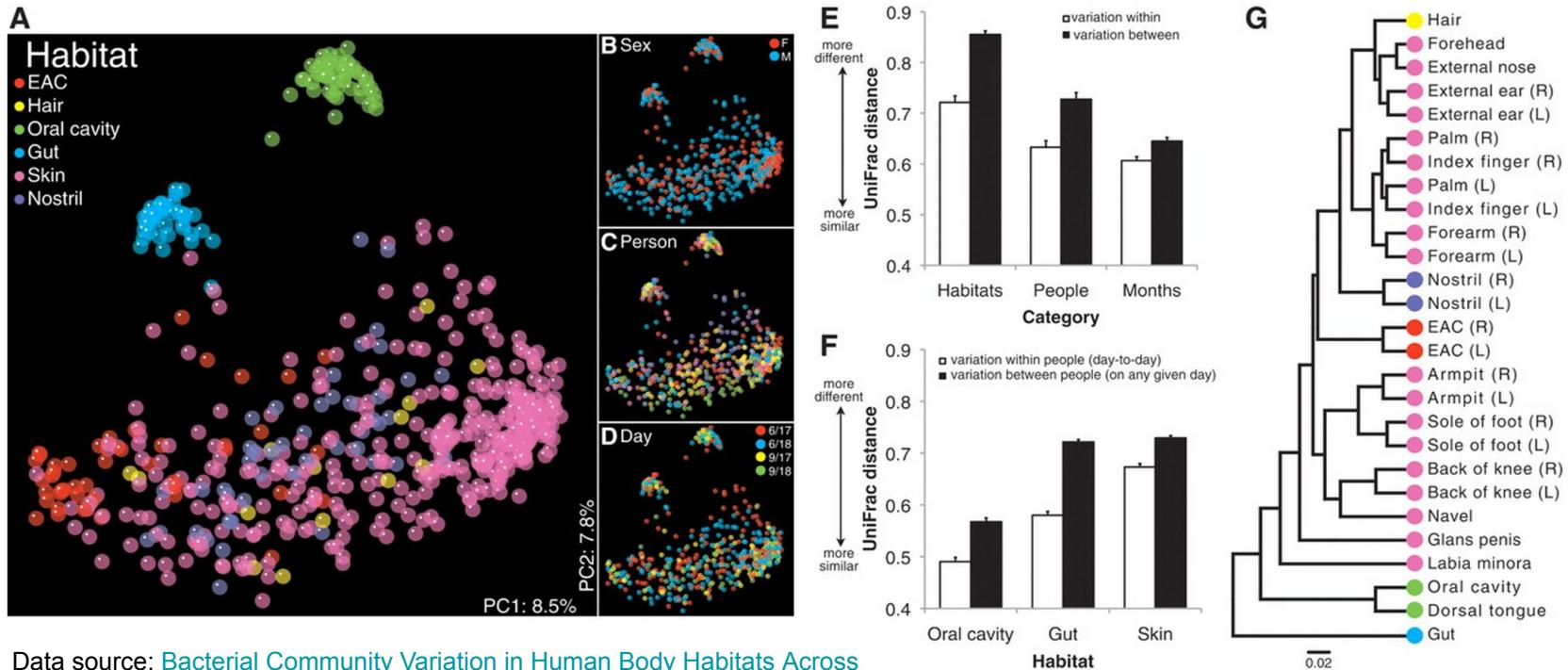
Principal Coordinates Analysis

- Widely adopted in microbiome research.



Knight, Rob et al. "Unlocking the potential of metagenomics through replicated experimental design." *Nature biotechnology* 30.6 (2012): 513-520. 237

Various techniques are applied to interpret distance matrices

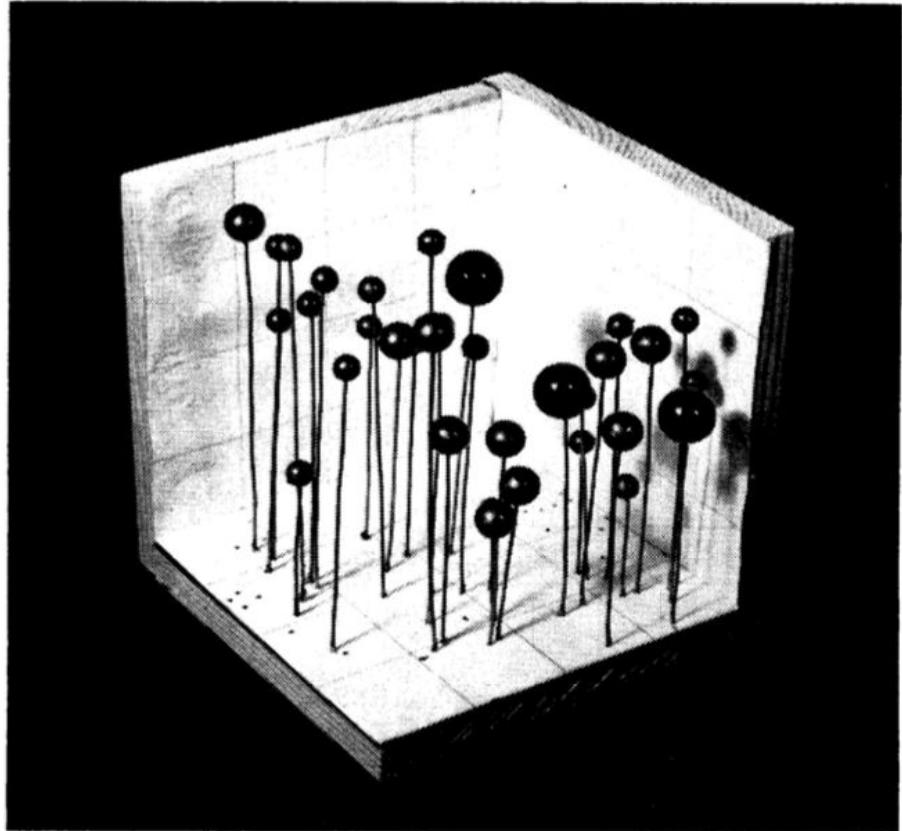


Data source: [Bacterial Community Variation in Human Body Habitats Across Space and Time](#). Costello et al. *Science* (2009)

If you're interested in "horseshoe effects" in ordination results, see: <http://msystems.asm.org/content/2/1/e00166-16>

Data Dioramas

- We've come a long way since the 50's!
- Way easier these days



Bray, J Roger, and John T Curtis. "An ordination of the upland forest communities of southern Wisconsin." *Ecological monographs* 27.4 (1957): 325-349.

Statistical Tests for Beta Diversity

- Categorical univariate: Permanova, Adonis, Permdisp
 - Pairwise testing possible
- Continuous univariate: Mantel
- Multivariate: Adonis

Summary

Different Metrics tell you different things

Unweighted Metrics ("Qualitative")

- Presence/Absence of an OTU
- More sensitive to rare OTUs
- Ex: Jaccard, Unweighted UniFrac

Weighted Metrics ("Quantitive")

- Considers relative abundance (composition)
- More sensitive to abundant taxa
- Ex: Bray-Curtis, Canberra, Weighted UniFrac

Different Metrics tell you different things

Taxonomic

- Assume everything is equally dissimilar
- More likely to see differences based on close relatives
- Shannon; Observed ASVs; Bray Curtis,
Euclidian

Phylogenetic

- Take into account similarity based on shared evolution
- Better for scaling the differences which are seen
- PD Whole Tree; UniFrac

FAQ

- What is the best distance metric?
 - Different metrics show different properties of the data
 - No one single metric is better than the rest
- How do I know what metadata category is the most important?
 - Use a statistical test (to find most important category)
 - Visualizations are exploratory
 - Use custom --p-custom-axis in emperor plugin
 - Biplots are another powerful option (`qiime diversity pcoa-biplot`, [DEICODE](#), or [PhyloSeq](#) in R)
- Can I use other metrics?
 - Yes, look at `--p-metric` in the [beta diversity command](#).
 - Check out this [forum post](#) for more information!

Taxonomic Classification

<https://bit.ly/2HThBcx>

Taxonomic assignment of observed sequences.

FeatureData [Sequence]

```
>feature5
GACGAAGGTGACGCCGTTGCTCGGAATCACTGGGCATAAAGCGCCGTAGGTGGCTTGGTAAGTCATGGTGA
ATCCCTCGGCTCAACCGAGGAACCTG
>feature4
TACGTAGGGGCAAGCGTTATCGGATTTACTGGGTGAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAA
AGGCTGGGCTCAACCCGGGACGG
>feature2
TACGTATGGGCAAGCGTTATCGGAATTATTGGGCGTAAAGAGTGCCTAGGTGGCTTAAGCGCAGGGTTA
AGGAATGGCTTAACTATTGTTCTC
>feature1
GACGGAGGATCCAAGTGTATCGGAATCACTGGGCATAAGCGCTGTAGGTGGTTACTAAGTCAACTGTTAA
ATCTTGAGGCTCAACCTCGAAATCG
>feature3
TACGGAGGGTGCAGCGTTAACCGAATTACTGGGCGTAAAGCGTACGTAGGCCTTAGGTAAGTCAGATGTGAA
AGCCCCGGGCTCCACCTGGGAATGG
```

Taxonomic assignment of observed sequences.

Reference Database
Silva, Greengenes, etc.

FeatureData [Sequence]

```
>feature5  
GACGAAGGTGACGCCGTTGCTCGGAATCACTGGGCATAAAGCGCCGTAGGTGGCTTGGTAAGTCATGGTAA  
ATCCCTCGGCTCAACCGAGGAACCTG  
>feature4  
TACGTAGGGGCAAGCGTTATCCGATTACTGGGTGAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAA  
AGGCTGGGCTCAACCCGGGACGG  
>feature2  
TACGTATGGGCAAGCGTTATCCGAATTATGGGCGTAAAGAGTGCCTAGGTGGCTTAAGCGCAGGGTTA  
AGGAATGGCTTAACATTGTTCTC  
>feature1  
GACGGAGGATCCAAGTGTATCCGAATCACTGGGCCTAAAGCGCTGTAGGTGGTTACTAAGTCACGTAA  
ATCTTGAGGCTCAACCTCGAAATCG  
>feature3  
TACGGAGGGTGCAGCGTTATCGGAATTACTGGGCGTAAAGCGTACGTAGGCCTTAGGTAAGTCAGATGTGAA  
AGCCCCGGGCTCCACCTGGGATGG
```

FeatureData [Sequence]

```
>reference-sequence-1  
TTGAAGGTGGACGCCGTTGCTCGGAATCACTGGGCATAAAGCGCCGTAGGTGGCTTGGTAAGTCACATGGT  
GACTCAACCGAGGAACCTGAAAGTGAAGGTGGGACGCCGTTGCTCGGAATCACTGGGCATAAAGCGCCGTAGGT  
GCTTGGTAAGTCACATGGTACTCAACCGAGGAACCTGAA
```

```
>reference-sequence-2  
AACGTAGGCAAGCGTTATCCGATTACTGGGTGAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAAAGG  
CTGGGCTCAACCCGGGCTTGCCTCGGAATCACTGGGCATAAAGCGCCGTAGGT
```

FeatureData [Taxonomy]

> T A T > T A > T A >	reference-sequence-1 Bacteria; Proteobacteria; Gammaproteobact reference-sequence-2 Bacteria; Bacteroidetes; Flavobacteria; F reference-sequence-3 Bacteria; Proteobacteria; Deltaproteobact reference-sequence-4 Archaea; Euryarchaeota; DSEG; 104A5
---	--

Taxonomic assignment of observed sequences.

Reference Database
Silva, Greengenes, etc.

FeatureData [Sequence]

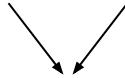
```
>feature5
GACGAAGGTGACGCCGTTGCTCGGAATCACTGGGCATAAAGCGCCGTAGGTGGCTTGGTAAGTCATGGTCAA
ATCCCTCGGCTCAACCGAGGAACCTG
>feature4
TACGTAGGGGCAAGCGTTATCCGATTACTGGGTGAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAA
AGGCTGGGCTCAACCCGGACGG
>feature2
TACGTATGGGCAAGCGTTATCCGAATTATGGGCGTAAAGAGTGCCTAGGTGGCTTAAGCGCAGGGTTA
AGGAATGGCTTAACCTATTGTTCTC
>feature1
GACGGAGGATCCAAGTGTATCCGAATCACTGGGCCTAAAGCGCTGTAGGTGGTTACTAACGTAACTGTTAA
ATCTTGAGGCTCAACCTCGAAATCG
>feature3
TACGGAGGGTGCAGCGTTATCGGAATTACTGGGCTAAAGCGTACGTAGGCCTTAGGTAAGTCAGATGTGAA
AGCCCCGGGCTCCACCTGGGATGG
```

FeatureData [Sequence]

```
>reference-sequence-1
TTGAAGGTGGACGCCGTTGCTCGGAATCACTGGGCATAAAGCGCCGTAGGTGGCTTGGTAAGTCACATGGT
GACTCAACCGAGGAACCTGAAAGTGAAGGTGGACGCCGTTGCTCGGAATCACTGGGCATAAAGCGCCGTAGGTG
GCTTGGTAAGTCACATGGTACTCAACCGAGGAACCTGAA
>reference-sequence-2
AACGTAGGCAGCGTTATCCGATTACTGGGTGAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAAAGG
CTGGGCTCAACCCGGACGGCTTGTGCTCGGAATCACTGGGCATAAAGCGCCGTAGGTG
```

FeatureData [Taxonomy]

reference-sequence-1	Bacteria; Proteobacteria; Gammaproteobact
reference-sequence-2	Bacteria; Bacteroidetes; Flavobacteria; F
reference-sequence-3	Bacteria; Proteobacteria; Deltaproteobact
reference-sequence-4	Archaea; Euryarchaeota; DSEG; 104A5



Compare observed sequences to annotated reference sequences to make taxonomic assignments.



FeatureData [Taxonomy]

feature5	Bacteria; Proteobacteria
feature4	Bacteria; Proteobacteria
feature2	Bacteria; Bacteroidetes; Flavobacteria; Flavobacteriales
feature1	Bacteria; Proteobacteria
feature3	Bacteria; Proteobacteria; Deltaproteobacteria

???

Sequence Alignment

Query

GACGAAG

ACTGA**GACGAAG**AATGTGCTGAT
ATGTGTGCTGTGATGTCTGTGTA
TGTGTATGGCTGTGATGCTGATA

GACGAAAGGTGACGACCGTTGCTC

GACGAAG

ACGAAGG

CGAAGGT

GAAGGTG

AAGGTGA

AGGTGAC

GGTGACG

K-mer decomposition for
feature extraction

A cloud-shaped container holds several DNA sequence fragments. An arrow points from the cloud to a rectangular box labeled "Machine Learning Classifier". A second arrow points downwards from the classifier box to a list of taxonomic names.

GACGAAG
ACGAAGG
CGAAGGT
GAAGGTG
AAGGTGA
AGGTGAC
GGTGACG

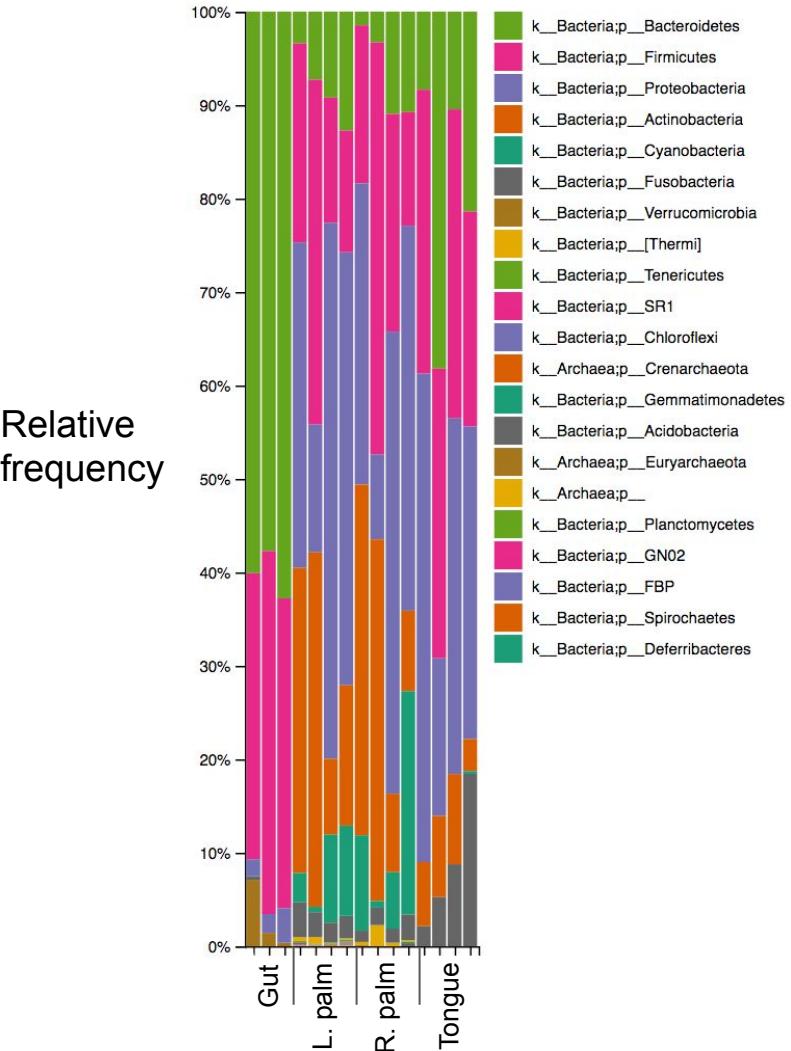
Machine
Learning
Classifier

Bacteria; Proteobacteria; Deltaproteobacteria;
Desulfuromonadales; Geobacteraceae; Geobacter;

Visualizing taxonomic profiles

Interactive barplots support:

- Taxonomic level selection
 - Multi-level sorting
 - Filtering
 - Coloring
 - Exporting plots (SVG) and raw data



Classification - Alignment methods

Consensus approach:

- [classify-consensus-blast](#)
- [classify-consensus-vsearch](#)

Your sequence	Your ref. database	BLAST top hits	consensus
Kingdom	Bacteria	Bacteria	✓
Phyla	Proteobacteria	Proteobacteria	✓
Class	Gammaproteobacteria	Gammaproteobacteria	✓
Order	Legionellales	Legionellales	✓
Genus	Legionellaceae	Legionellaceae	✓
Species	Legionella	---	✗

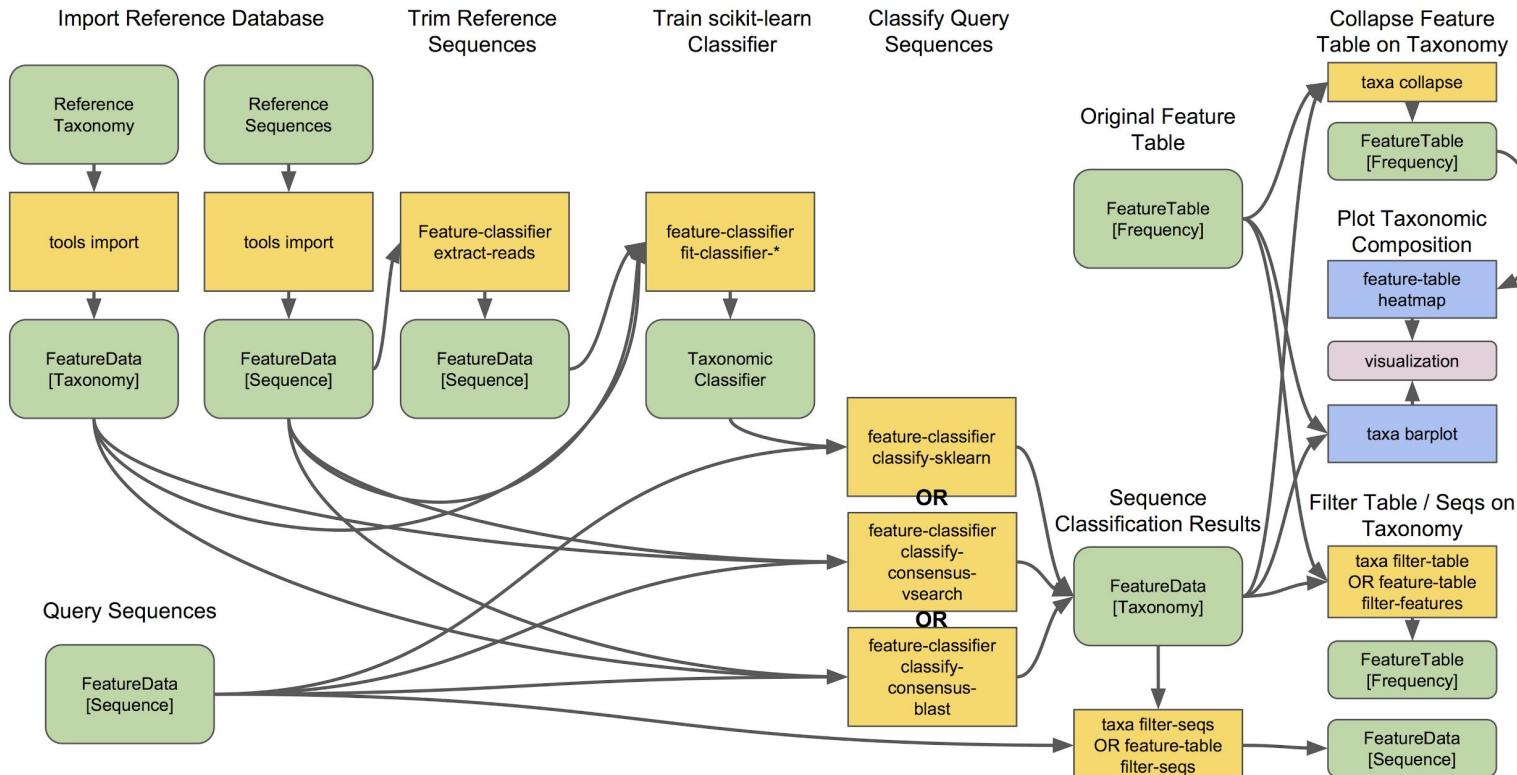
Taxonomies from the past?

You can import tables and **ALSO** their taxonomies (from QIIME 1.x.x datasets).

```
qiime tools import \  
  --input-path my-table.biom \  
  --type FeatureData[Taxonomy] \  
  --source-format BIOMV210Format \  
  --output-path my-table.qza
```

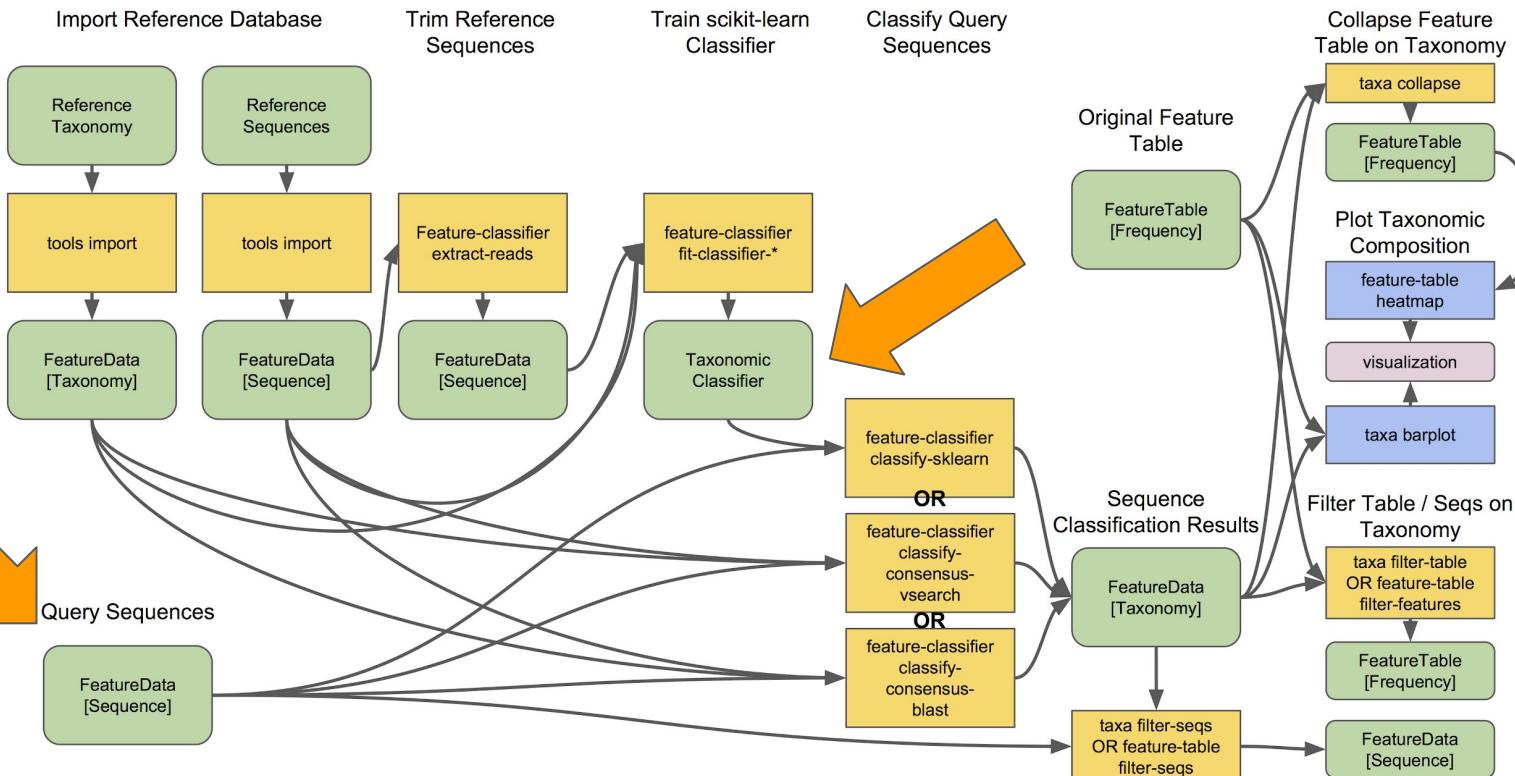
The source format may need to be BIOMV100Format depending on the version of QIIME used to process the dataset.

Taxonomy Classification



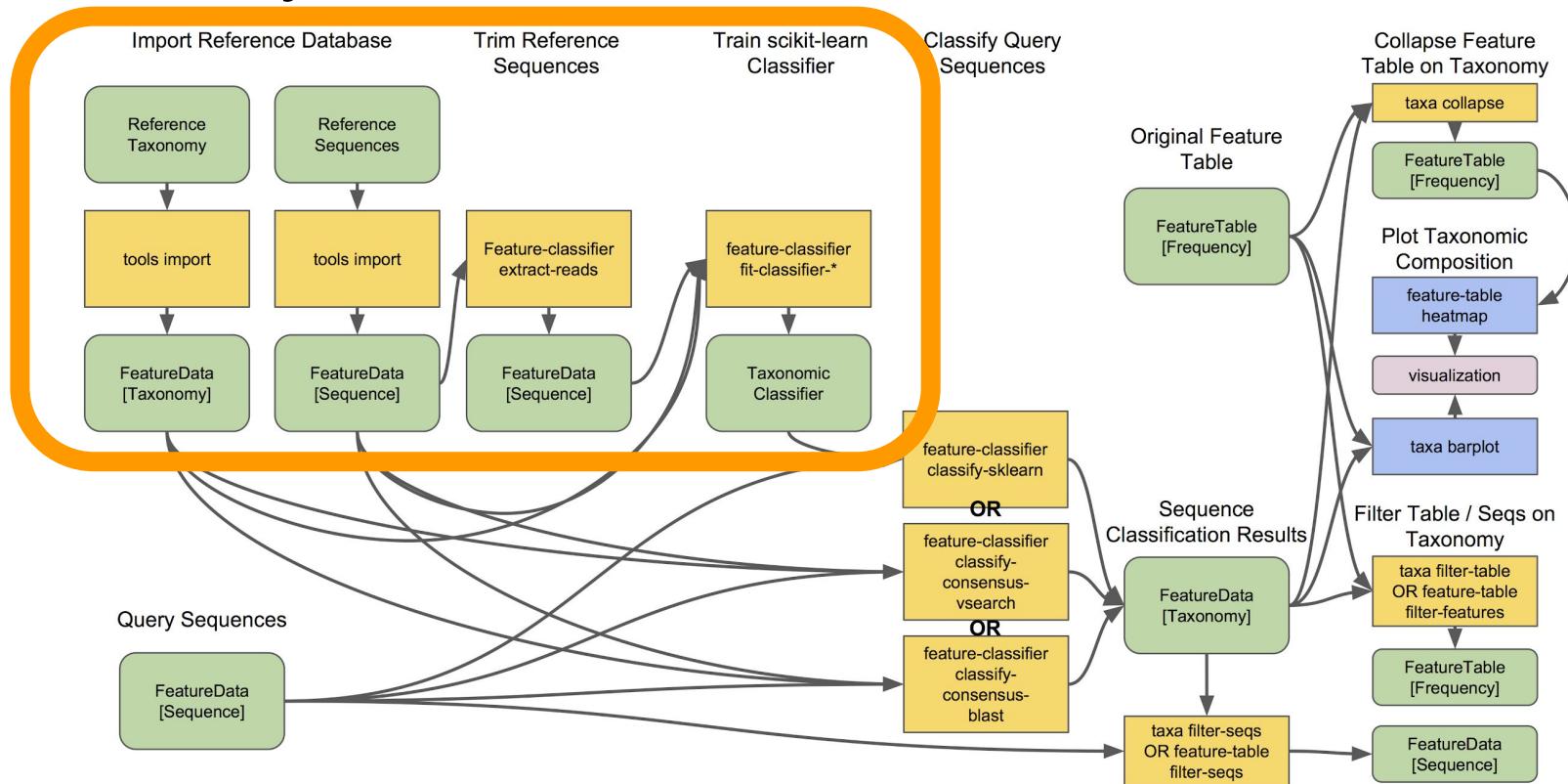
Source: Overview of Taxonomy Classification in QIIME 2

Taxonomy Classification



Source: Overview of Taxonomy Classification in QIIME 2

Taxonomy Classification



Source: Overview of Taxonomy Classification in QIIME 2

Training your own reference DB tutorial

- Data resources

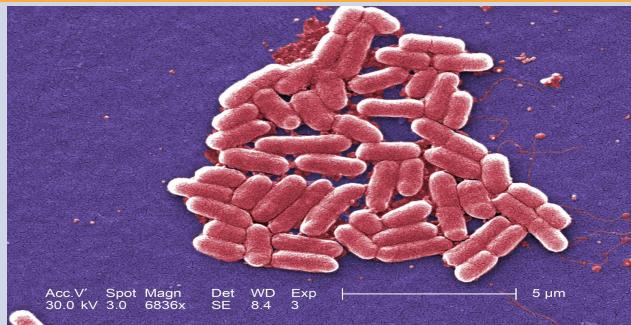
Danger

Pre-trained classifiers that can be used with `q2-feature-classifier` currently present a security risk. If using a pre-trained classifier such as the ones provided here, you should trust the person who trained the classifier and the person who provided you with the qza file. This security risk will be addressed in a future version of `q2-feature-classifier`.

- Training feature classifiers

<https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0470-z>

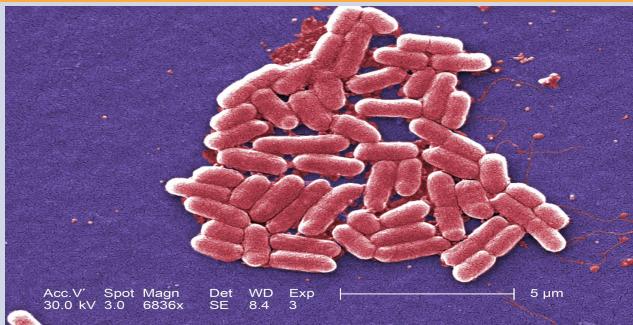
Ideal 16S



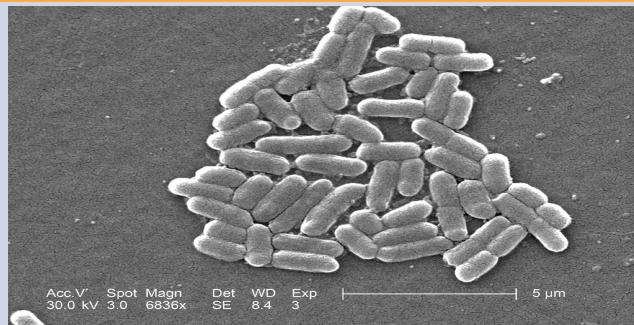
Kingdom	Bacteria
Phylum	Proteobacteria
Class	Gammaproteobacteria
Order	Enterobacteriales
Family	Enterobacteriaceae
Genus	<i>Escherichia</i>
Species	<i>coli</i>
Strain	O157:H7

E. coli: <http://media-3.web.britannica.com/eb-media//87/141087-050-24850517.jpg>

Ideal 16S



Real 16S



Kingdom	Bacteria	Bacteria
Phylum	Proteobacteria	Proteobacteria
Class	Gammaproteobacteria	Gammaproteobacteria
Order	Enterobacteriales	Enterobacteriales
Family	Enterobacteriaceae	Enterobacteriaceae
Genus	<i>Escherichia</i>	---
Species	<i>coli</i>	OTU 2445338
Strain	O157:H7	--

E. coli: <http://media-3.web.britannica.com/eb-media//87/141087-050-24850517.jpg>

Lactobacillus helveticus or *Lactobacillus hamsteri*?

- First 90 nucleotides of 16S v4 are exactly the same:

TACGTAGGTGGCAAGCGTTGCCGGATTATTGGGCGTAAAGCGAGCGCAG
GCAGGAAGAATAAGTCTGATGTGAAAGCCCTCGGCTTAACCGGGGAAAGT

Lactobacillus helveticus or *Lactobacillus hamsteri*?

- First 90 nucleotides of 16S v4 are exactly the same:

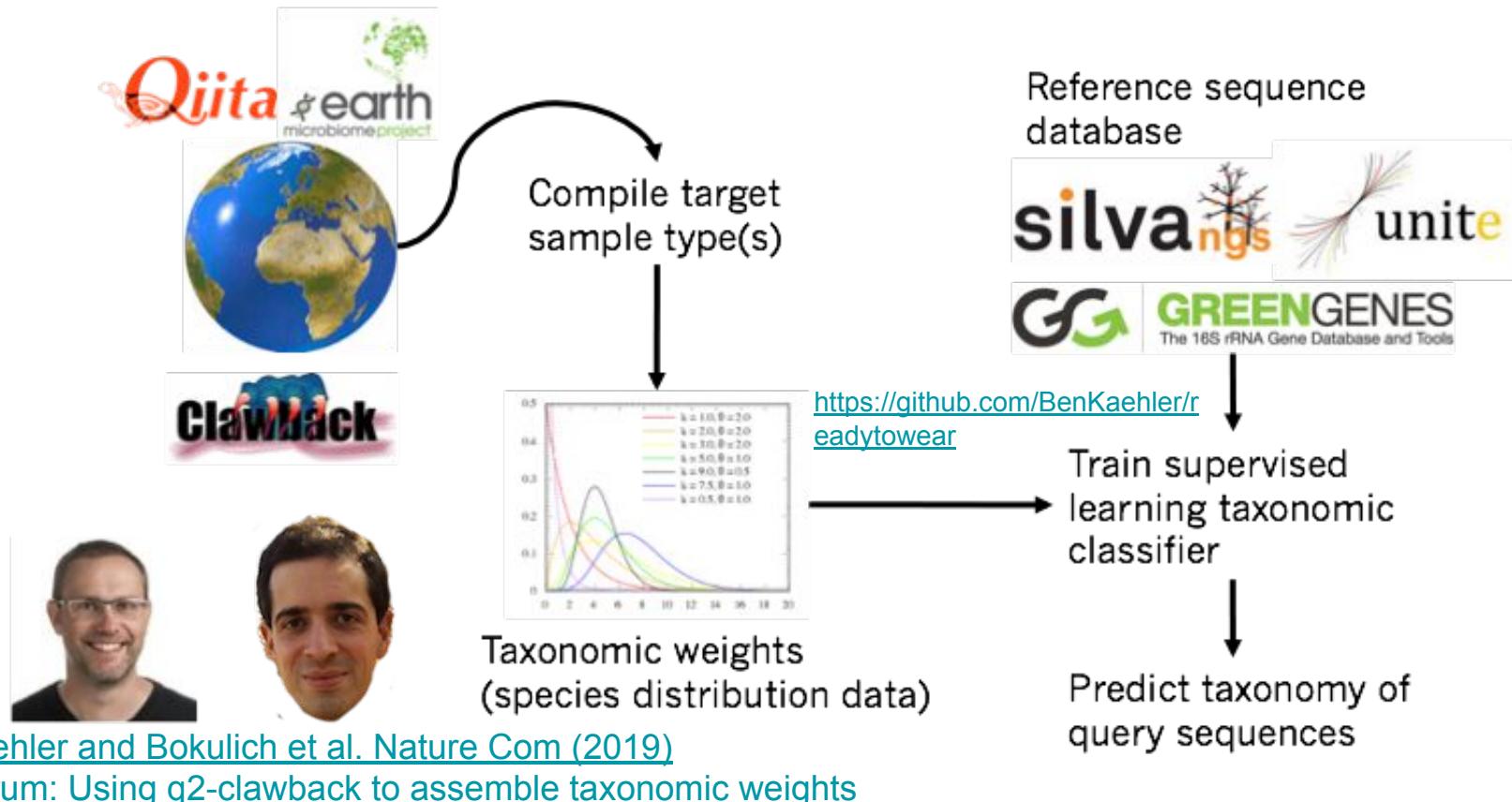
TACGTAGGTGGCAAGCGTTGCCGGATTATTGGGCGTAAAGCGAGCGCAG
GCAGGAAGAATAAGTCTGATGTGAAAGCCCTCGGCTTAACCGGGGAAAGT

- *L. helveticus* accounted for ~21% of all reads across 3,921 human vaginal samples[†]
- *L. hamsteri* isolated from a hamster gut in 1987[‡]
- If you saw this sequence in a human vaginal sample, would you call it *L. helveticus* or *L. hamsteri*?

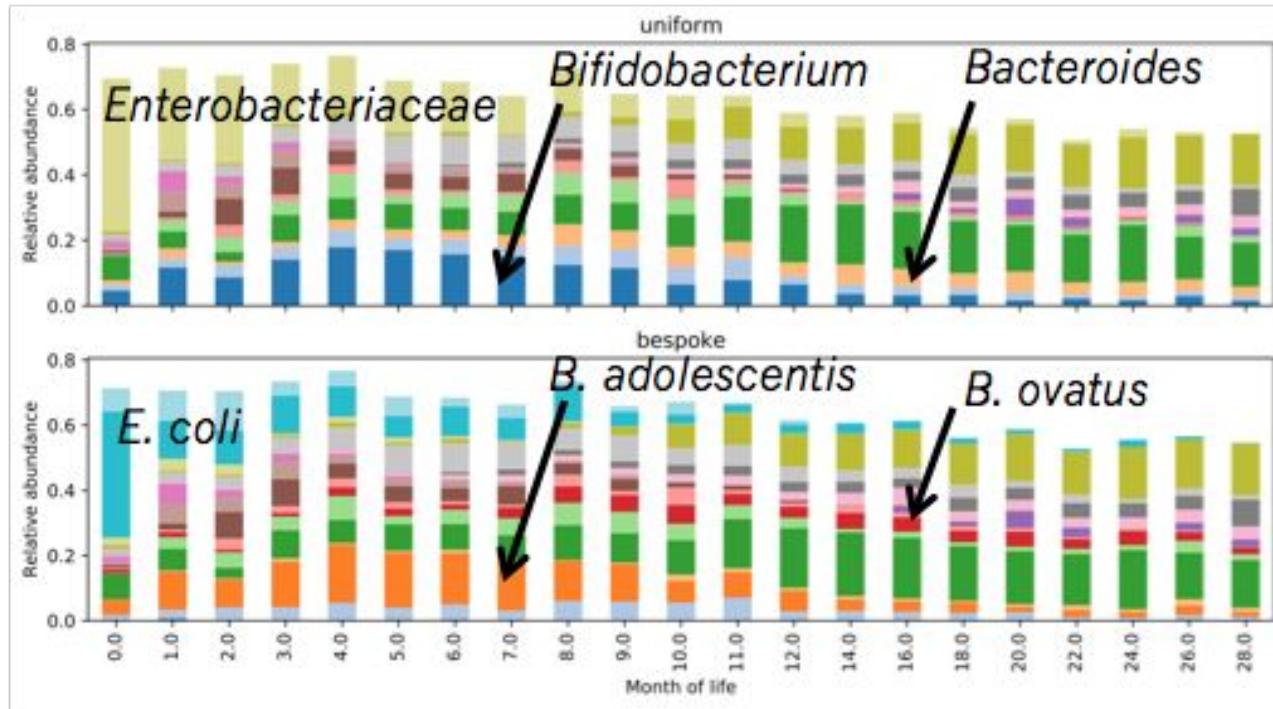
[†]data from [Fettweis et al. Nature Prec \(2010\)](#)

[‡][Mitsuoka and Fujisawa, Proc. Japan Acad. Ser B \(1987\)](#)

Habitat information improves classification



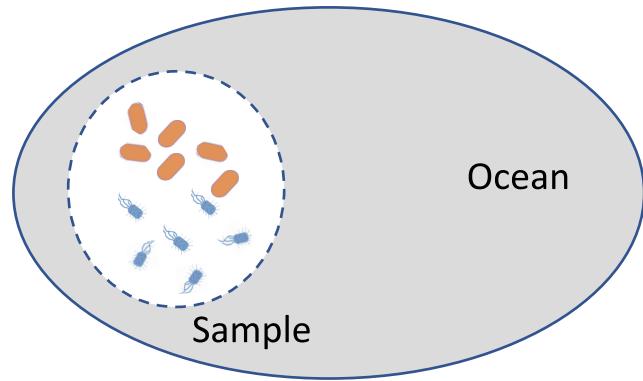
Taxonomic weights always improve accuracy

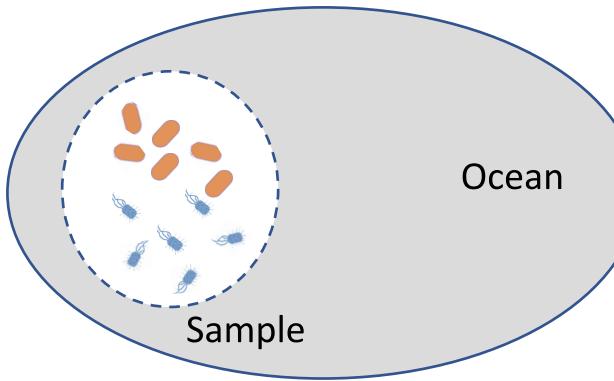
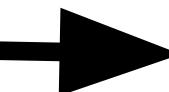
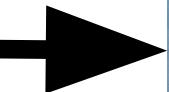


Differential Abundance Testing

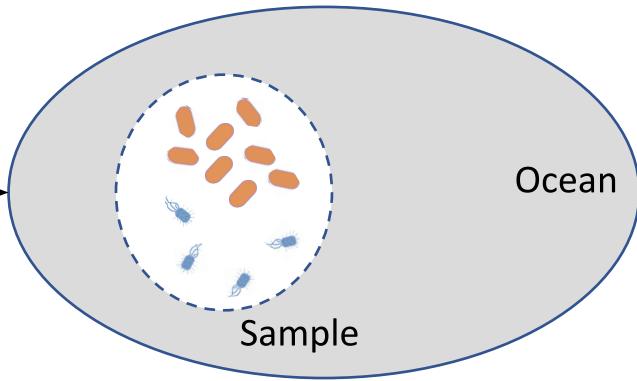
<https://bit.ly/2HThBcx>



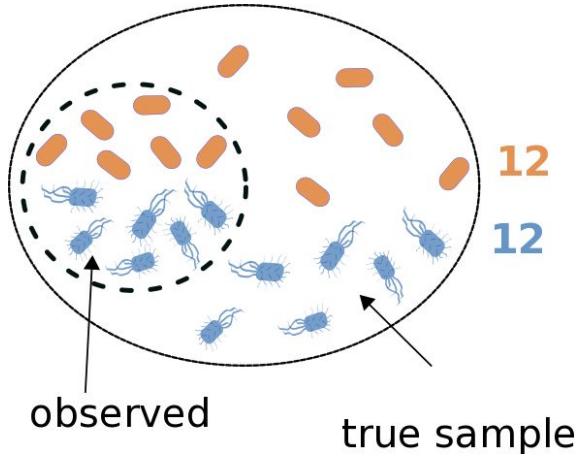




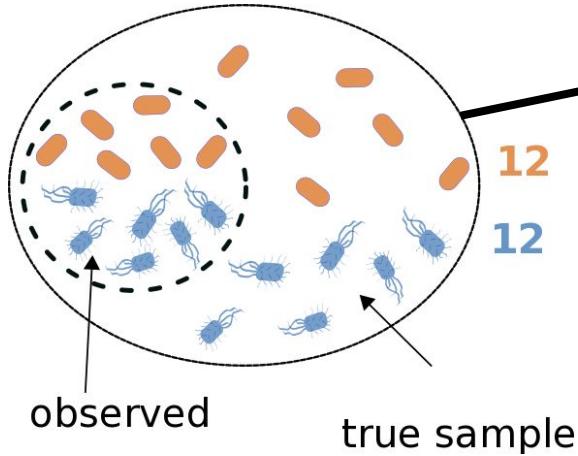
?



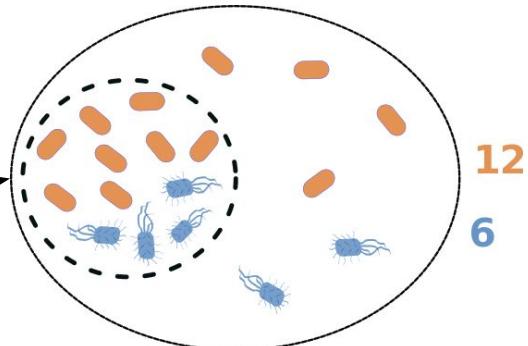
Before oil spill (B)



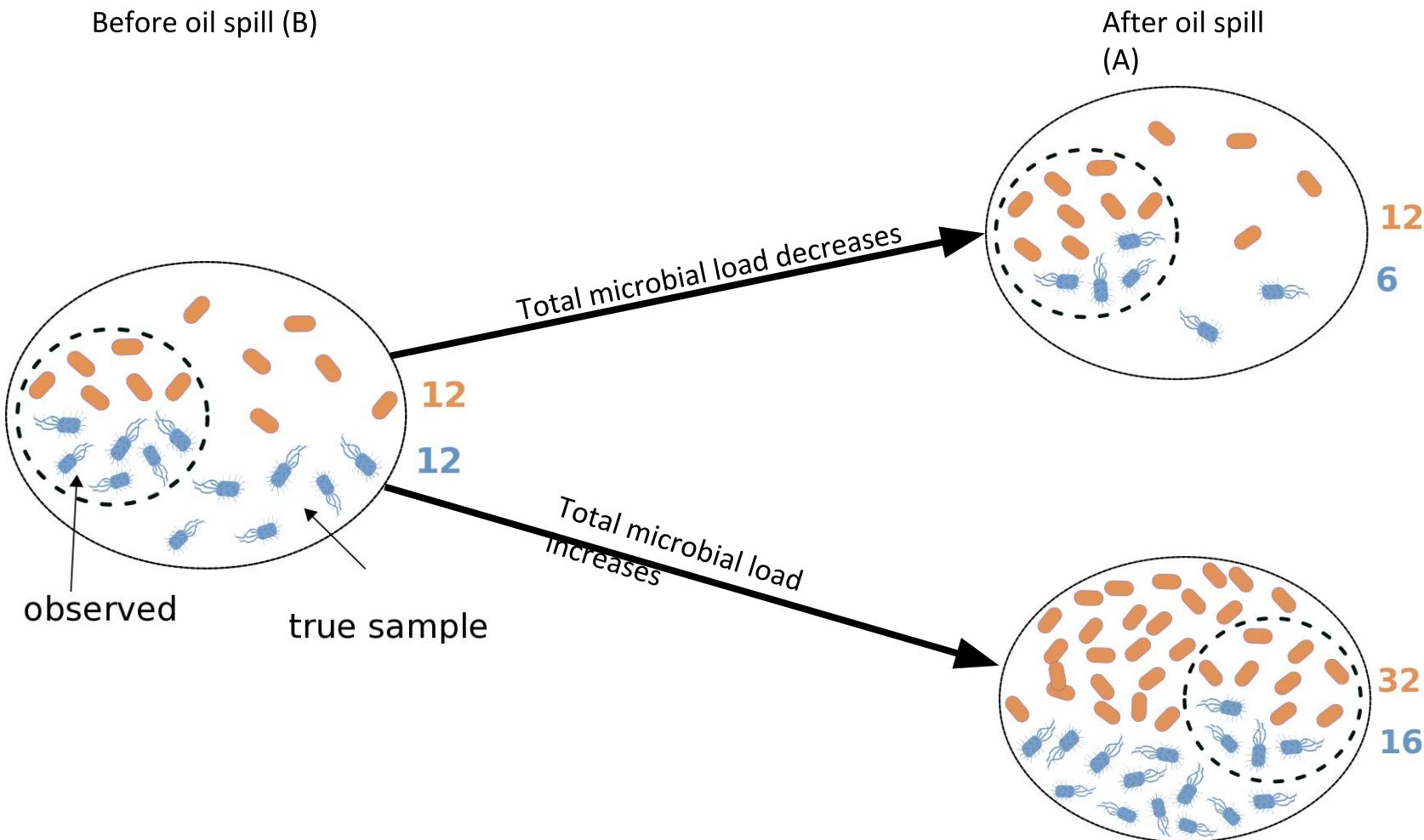
Before oil spill (B)



After oil spill
(A)



Total microbial load decreases



Methods for Measuring Microbial Concentrations

qPCR	Tkacz et al Mircrobiome, 2017
Flow cytometry	Vandeputte et al Nature, 2017
Spike-in	Stammler et al 2016

$$\text{Total} = (\text{cells} / \text{mL}) \times (\text{mL})$$

The total population bias

- Let A and B be samples with microbial abundances

$$A = (a_1, \dots, a_D)$$

$$B = (b_1, \dots, b_D)$$

- To determine which species change, we want

$$\frac{A}{B} = \left(\frac{a_1}{b_1}, \dots, \frac{a_D}{b_D} \right)$$

The total population bias

- Let A and B be samples with microbial abundances

$$A = (a_1, \dots, a_D)$$

$$B = (b_1, \dots, b_D)$$

- To determine which species change, we want

$$\frac{a_1}{b_1} = 1 \quad a_1 = b_1$$

$$\frac{a_1}{b_1} > 1 \quad a_1 > b_1$$

$$\frac{a_1}{b_1} < 1 \quad a_1 < b_1$$

The total population bias

- We cannot estimate A and B directly, only their proportions

$$A = (a_1, \dots, a_D) = (N_A p_{a_1}, \dots, N_A p_{a_D}) = N_A(p_{a_1}, \dots, p_{a_D})$$

$$B = (b_1, \dots, b_D) = (N_B p_{b_1}, \dots, N_B p_{b_D}) = N_B(p_{b_1}, \dots, p_{b_D})$$

- So if we try to calculate changes, we have a bias

$$\frac{A}{B} = \left(\frac{a_1}{b_1}, \dots, \frac{a_D}{b_D} \right) = \left(\frac{N_A \times p_{a_1}}{N_B \times p_{b_1}}, \dots, \frac{N_A \times p_{a_D}}{N_B \times p_{b_D}} \right)$$

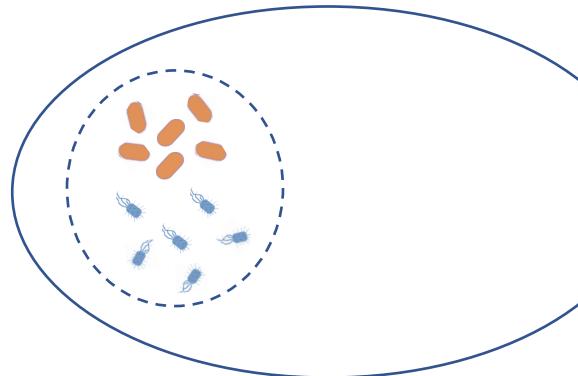
$$= \left(\frac{p_{a_1}}{p_{b_1}}, \dots, \frac{p_{a_D}}{p_{b_D}} \right) \times \frac{N_A}{N_B} = \frac{p_A}{p_B} \times \frac{N_A}{N_B}$$

Solutions

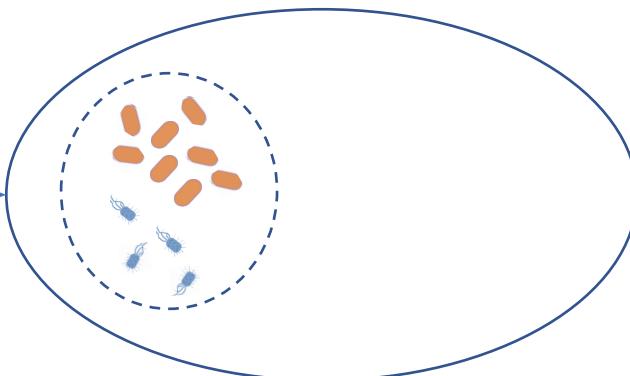
- Need to kill the total population bias
- Solution 1: ratios (i.e. concentrations)

$$\frac{a_1 / a_D}{b_1 / b_D} = \frac{N_A \times p_{a_1} / N_A \times p_{a_D}}{N_B \times p_{b_1} / N_B \times p_{b_D}} = \frac{p_{a_1} / p_{a_D}}{p_{b_1} / p_{b_D}}$$

Before Treatment



After Treatment



?

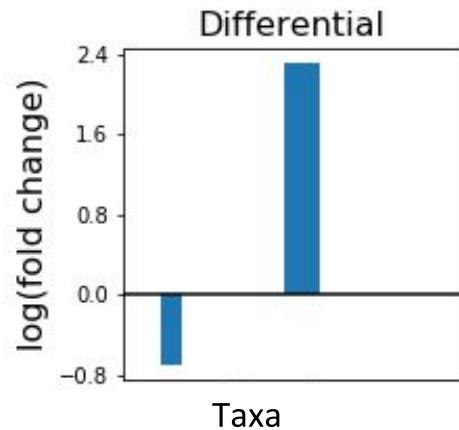
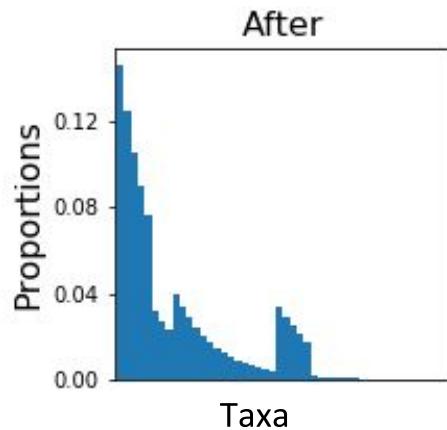
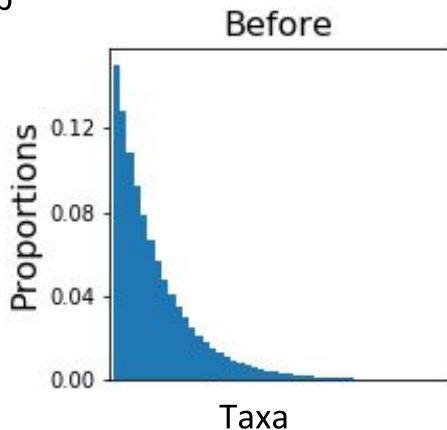
Solutions

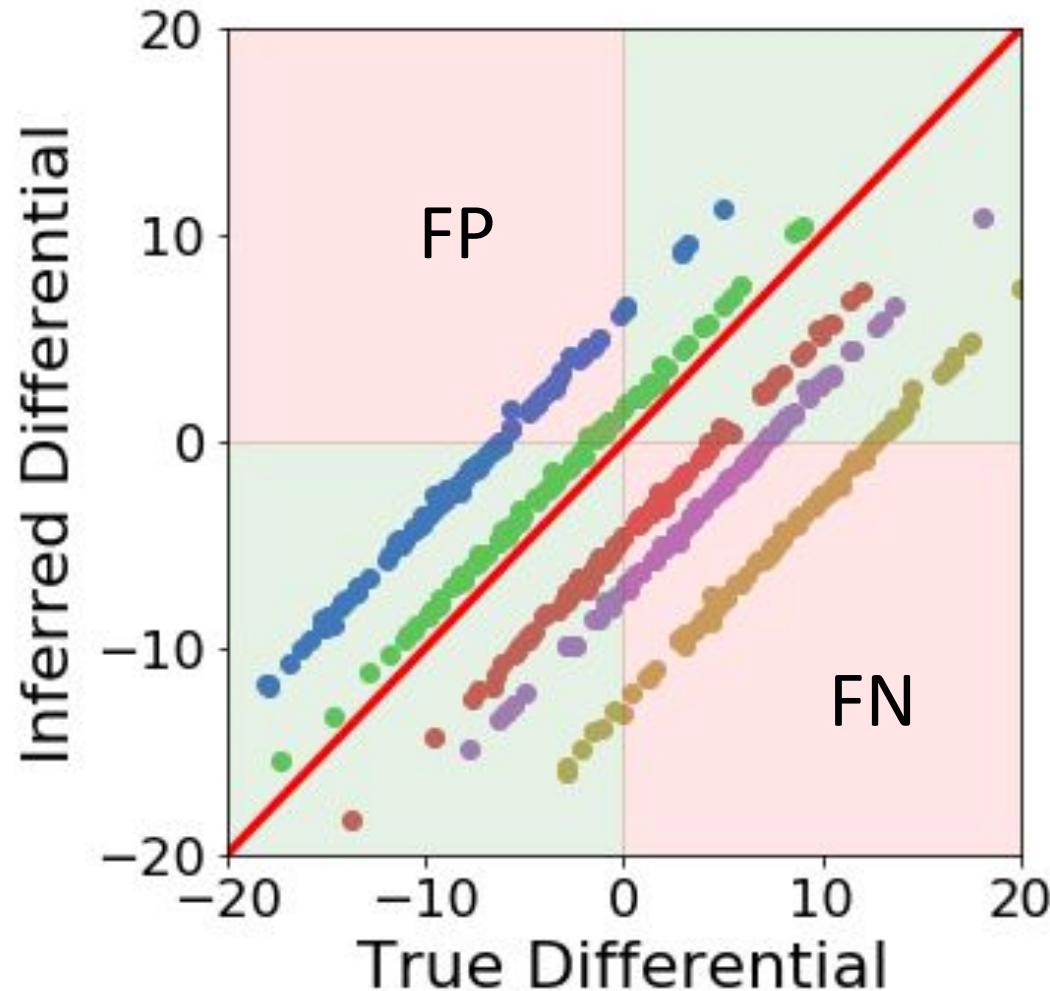
- Need to kill the total population bias
- Solution 2: ranks

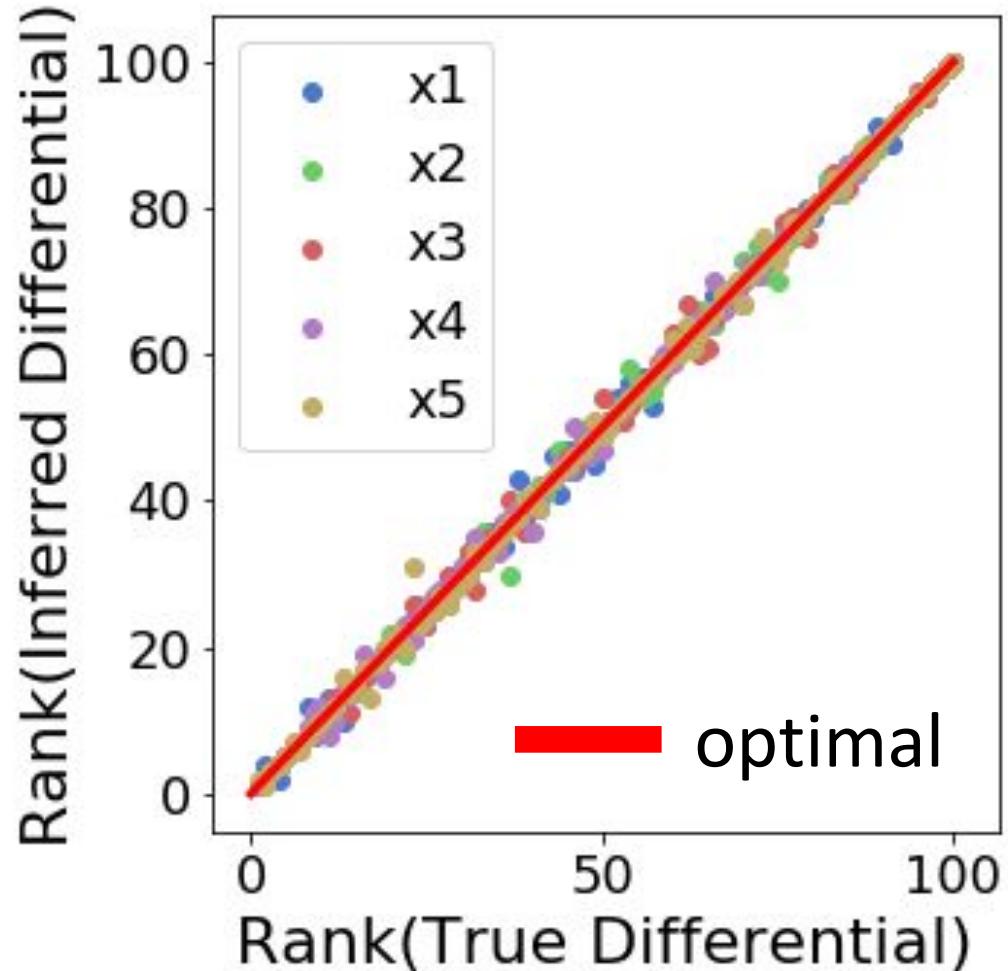
$$\text{rank} \left(\frac{A}{B} \right) = \text{rank} \left(\frac{p_A}{p_B} \times \frac{N_A}{N_B} \right) = \text{rank} \left(\frac{p_A}{p_B} \right)$$

Rank Differentials

b

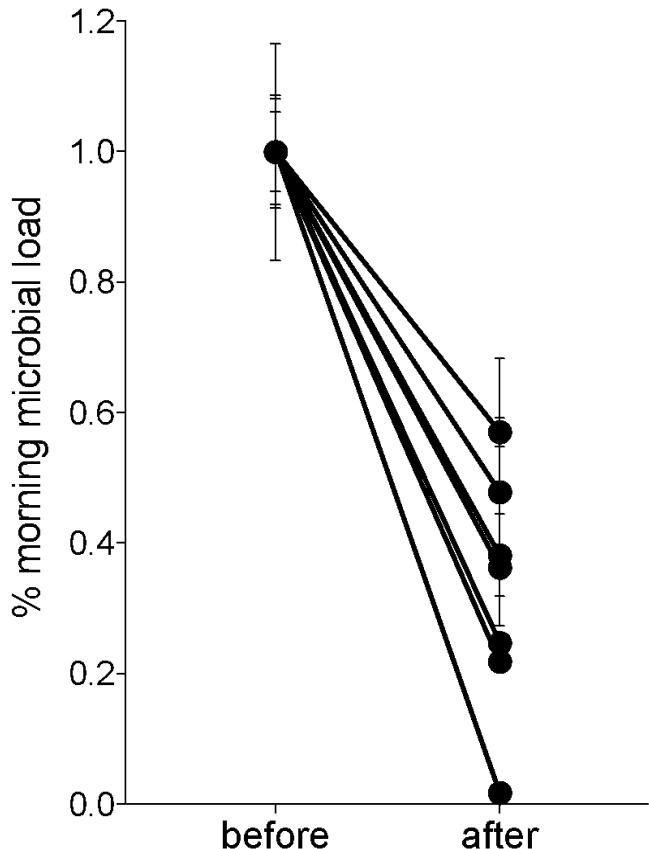


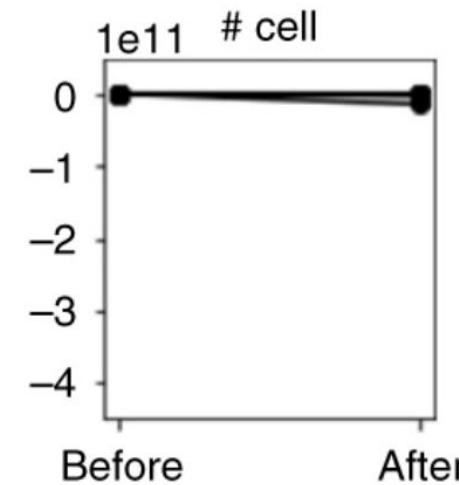
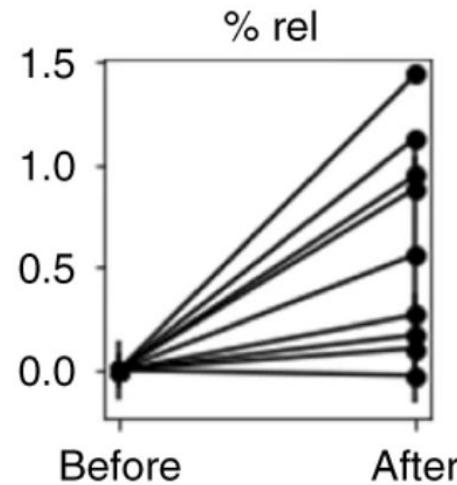
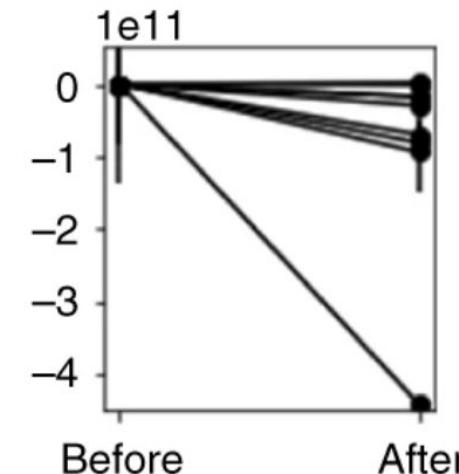
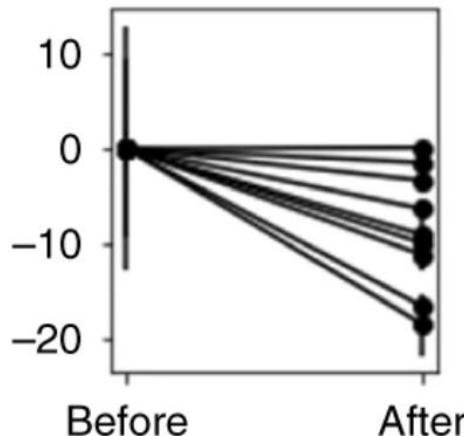




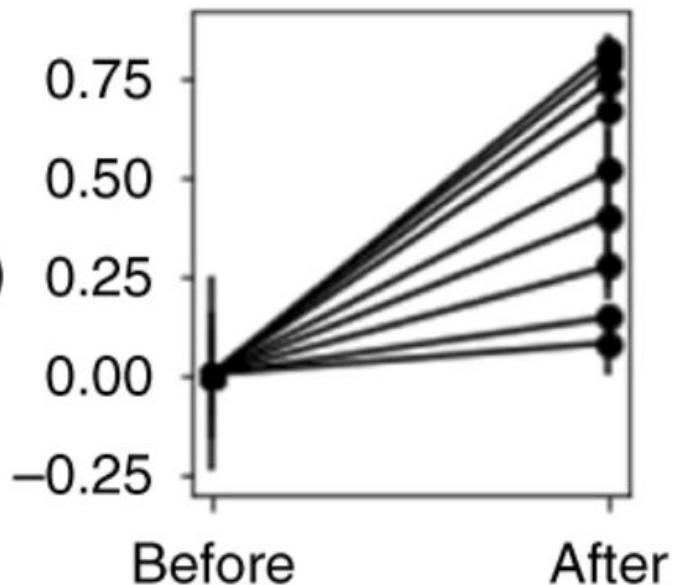
Saliva time series

- Need experimental proof
- Experiment
 - 10 people
 - 9 timepoints
 - 16S + Flow Cytometry
 - Relative vs Absolute abundance

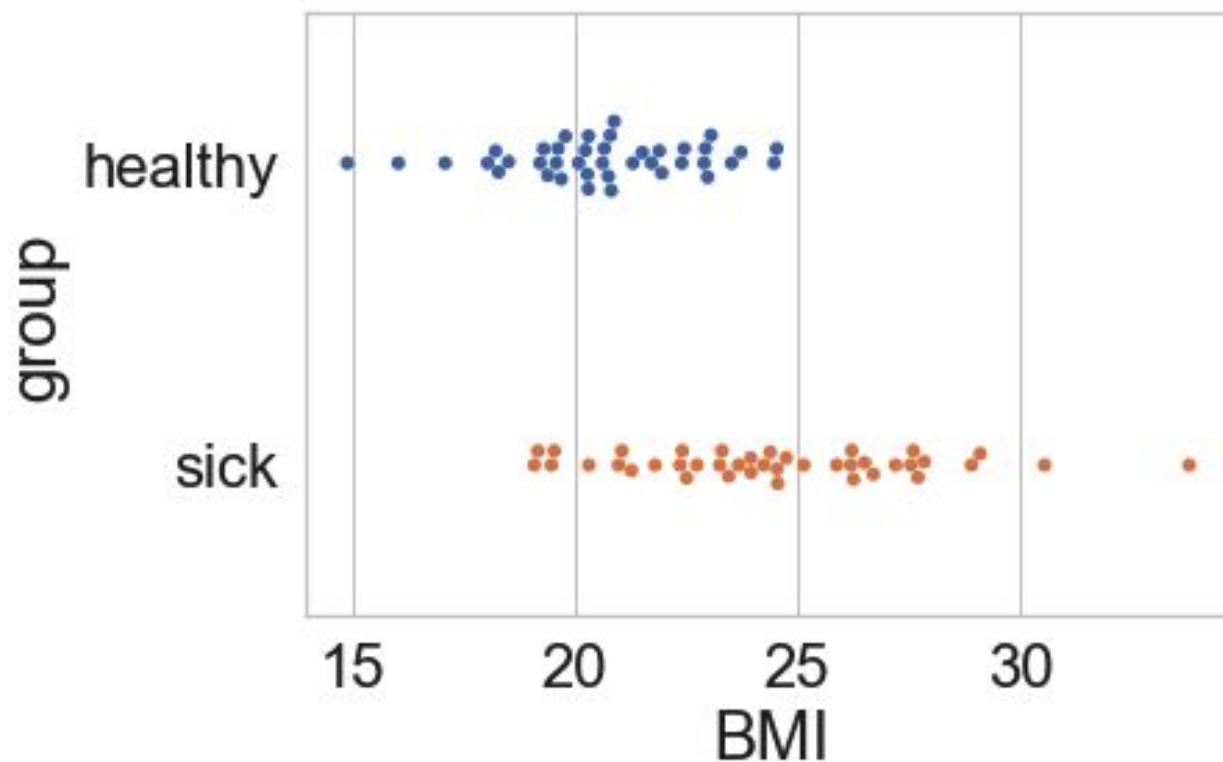


d*Actinomyces**Haemophilus*

$$\log \left(\frac{\text{Actinomyces}}{\text{Haemophilus}} \right)$$



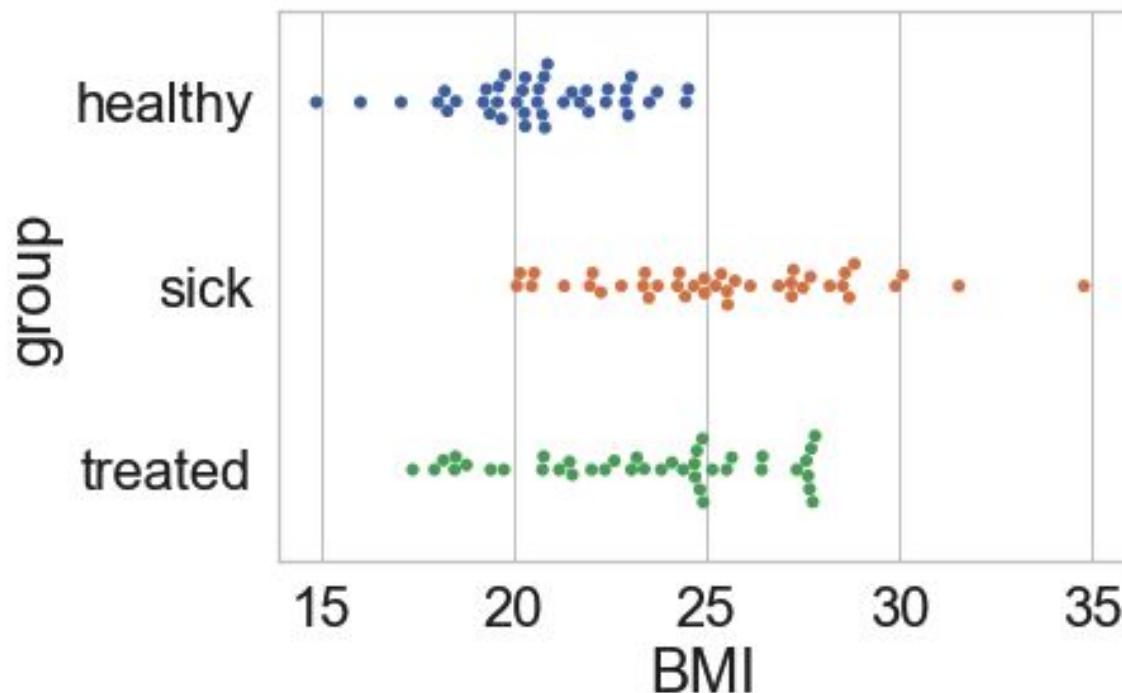
Analysis of Variance



Analysis of Variance

group	BMI
sick	25.887271
healthy	23.735116
sick	22.370807
healthy	23.528105
sick	27.209888
healthy	18.224429
healthy	19.395394
sick	24.561787

Dealing with multiple groups



Univariate response

$$Y = X B$$

$n \times 1$ Response (BMI, age, sex, ...)	$n \times p$ Covariates (gene abundances)	$p \times 1$ Coefficients
--	---	------------------------------

n = number of measurements

p = number of variables measured

Only one variable at a time!

Multivariate response

$$Y = X \cdot B$$

The diagram illustrates the equation $Y = X \cdot B$. It shows three matrices: Y , X , and B . Matrix Y is an $n \times p$ matrix with blue and light blue cells. Matrix X is an $n \times p$ matrix with blue and light blue cells. Matrix B is a $p \times D$ matrix with blue and light blue cells.

$n \times$

1

Respons

(gene
abundances)

n = number of
measurements

p = number of covariates
measured

D = number of variables
measured

$n \times$

p

Covariate

(BMI, age, sex,
...)

$p \times$

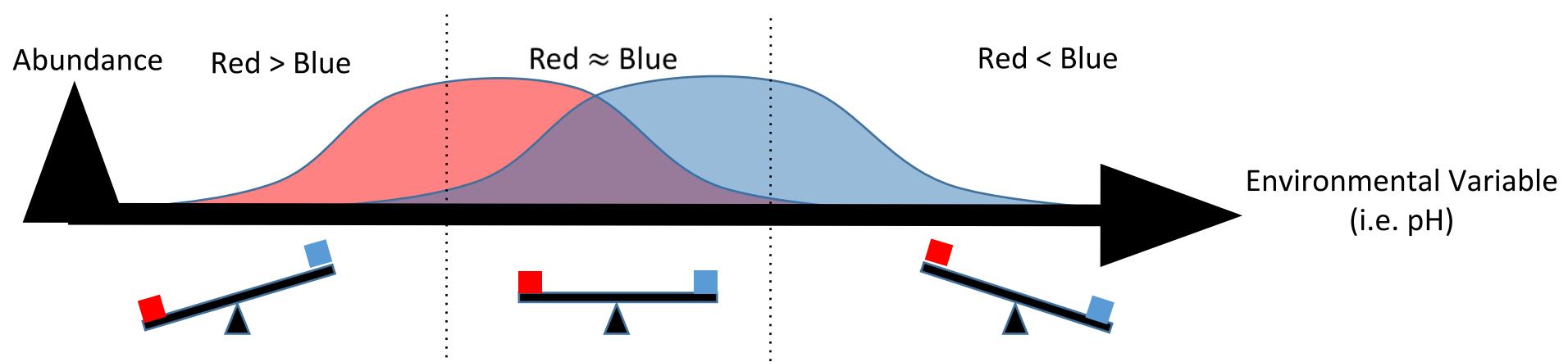
D

Coefficient

s

Can encode categorical
variables

Balances vs Environmental Variables



Differential Abundance

Part 2

Jamie Morton

Overview

- Balances
- Linear Regression



AMERICAN
SOCIETY FOR
MICROBIOLOGY



RESEARCH ARTICLE
Ecological and Evolutionary Science



Balance Trees Reveal Microbial Niche Differentiation

James T. Morton,^{a,b} Jon Sanders,^a Robert A. Quinn,^c Daniel McDonald,^b

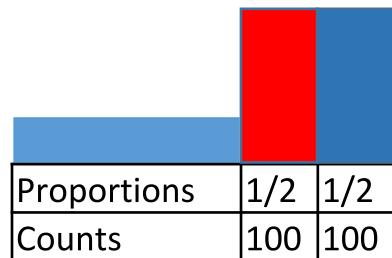
Antonio Gonzalez,^b Yoshiki Vázquez-Baeza,^{a,b} Jose A. Navas-Molina,^{a,b}

Se Jin Song,^a Jessica L. Metcalf,^c Embriette R. Hyde,^b Manuel Lladser,^d

Pieter C. Dorrestein,^e Rob Knight^{a,b}

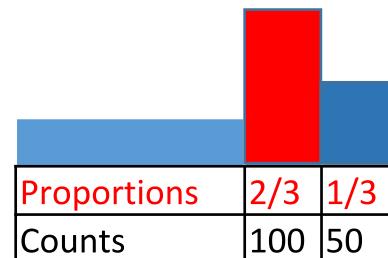
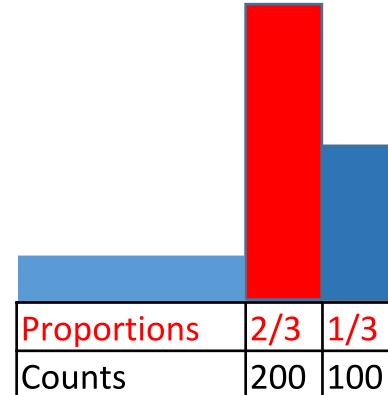
Department of Pediatrics, University of California San Diego, La Jolla, California, USA^a; Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, USA^b; Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy, University of California San Diego, La Jolla, California, USA, and Department of Animal Sciences, Colorado State University, Fort Collins, Colorado, USA^c; Department of Applied Mathematics, University of Colorado Boulder, Boulder, Colorado, USA^d; Pharmaceutical Sciences, University of California San Diego, La Jolla, California, USA^e

Compositionality



Time point 1

Red doubled



Time point 2

- Cannot predict who's growing/dying

Balances

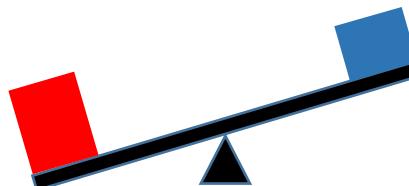


$$balance = \log\left(\frac{100}{100}\right) = 0$$

Time point 1

Red doubled
or
Blue halved

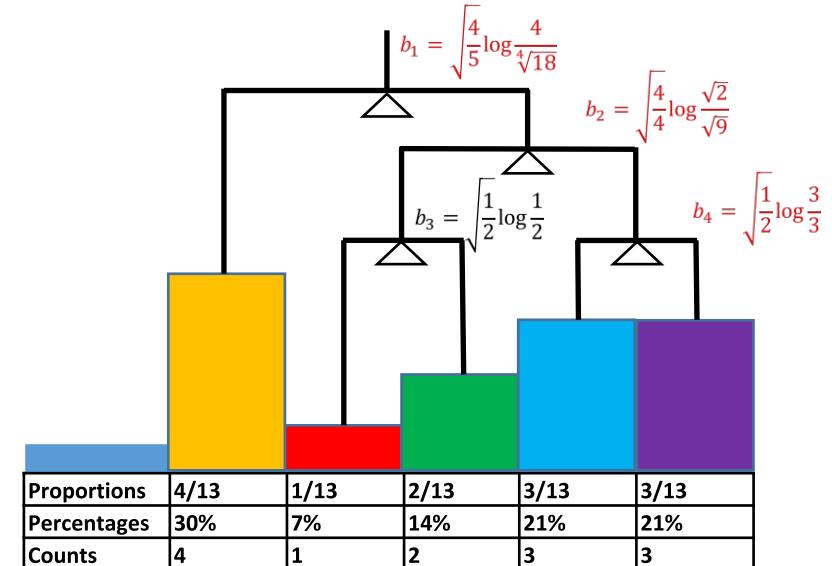
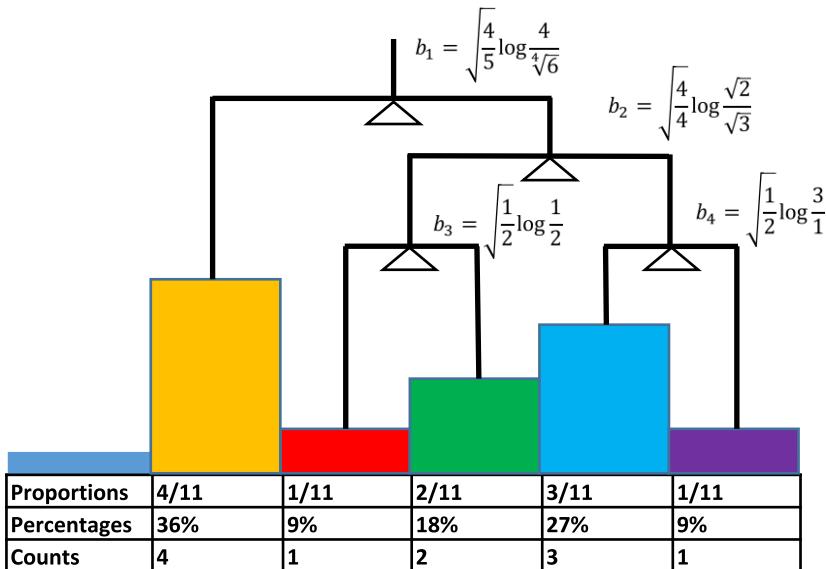
A large black arrow pointing to the right, indicating a transition or comparison between the two time points.



$$balance = \log\left(\frac{100}{50}\right) = \log\left(\frac{200}{100}\right) = \log\left(\frac{2/3}{1/3}\right) = \log 2$$

Time point 2

Balances for multiple proportions



ilr transform
Egozcue (2003)

$$b_i = \sqrt{\frac{|i_L| |i_R|}{|i_L| + |i_R|}} \log \left(\frac{g(i_L)}{g(i_R)} \right)$$

Properties

- Independence
- Scale invariance
- Zero issues – $\log(0)$ is undefined

$$b_i = \sqrt{\frac{|i_L| |i_R|}{|i_L| + |i_R|}} \log\left(\frac{g(i_L + 1)}{g(i_R + 1)}\right)$$

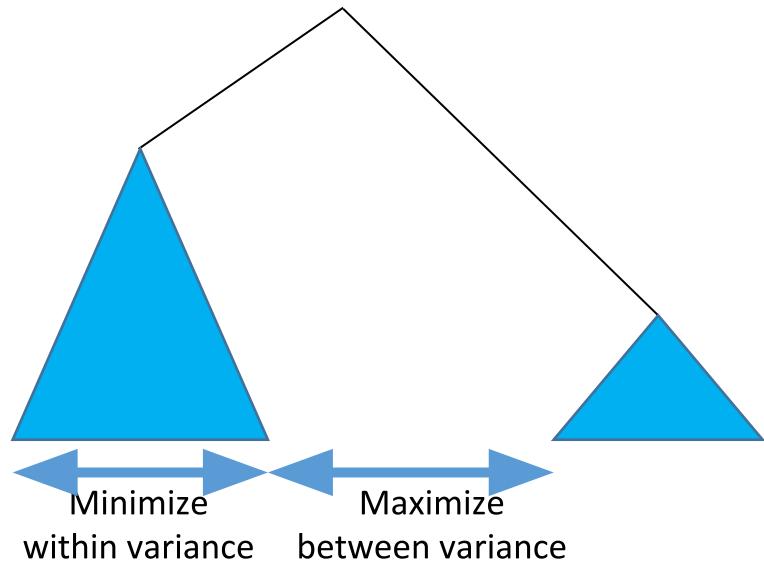
- Can apply standard statistics on balances!

Workflow

- Perform Clustering
- Compute Balances
- Perform Statistical Analysis

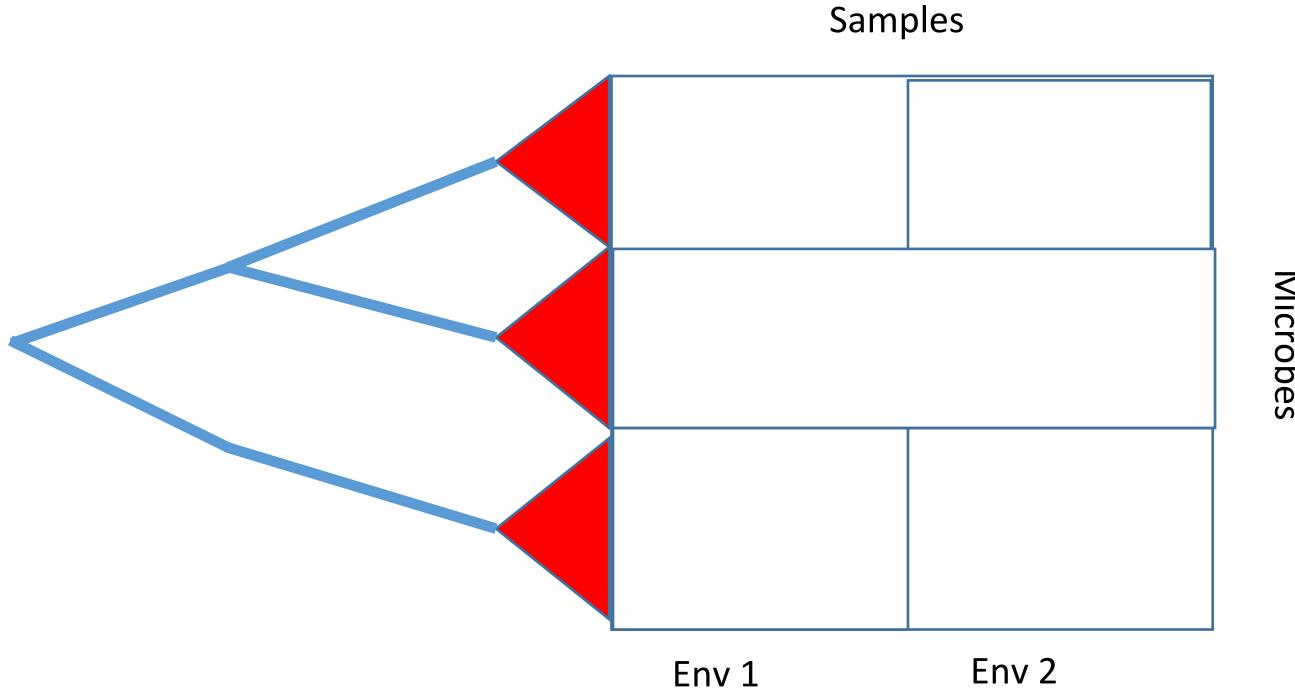
Principal Balances

- Define partitions of microbes
- Ward Hierarchical Clustering
 - Cluster based on abundances
 - Correlated microbes group together



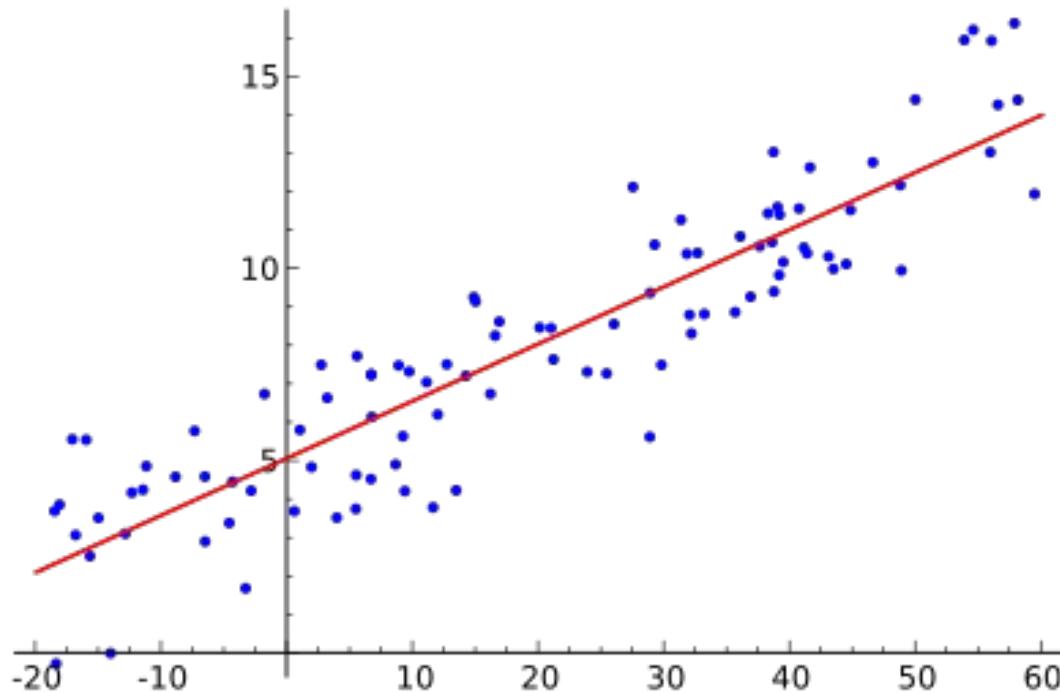
$$d(x, y) = V[\log \frac{x}{y}]$$

Principal Balances



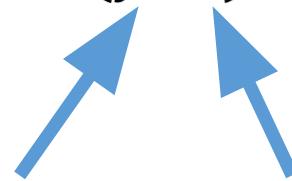
Weapon of Choice : Linear Regression

Linear Regression



ANOVA ⊂ Regression

$ANOVA(y, x)$



Variable

(i.e. $e^{microbe}$
abundance)

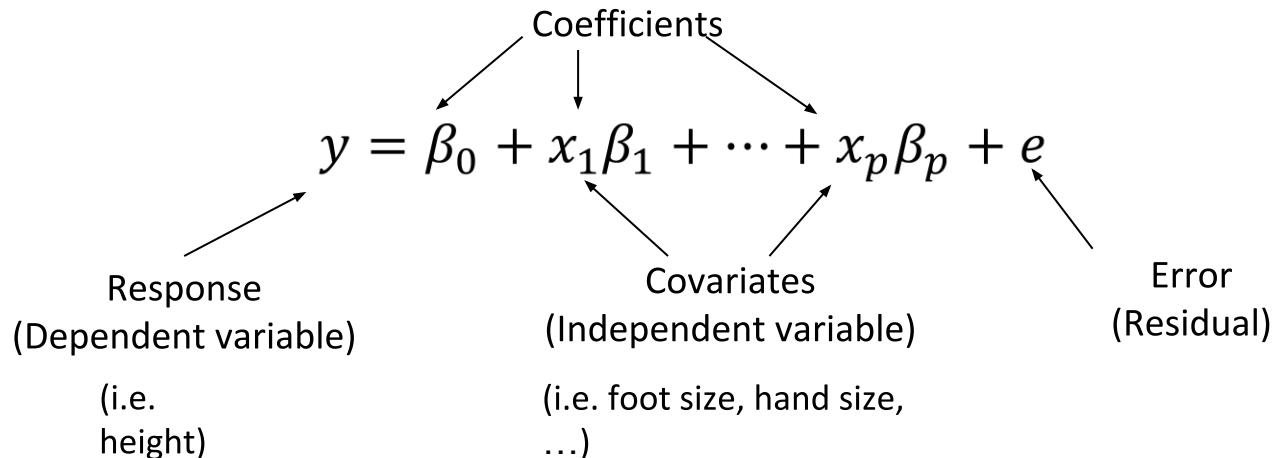
Classes

(i.e. sick vs healthy)

$$y = \beta_0 + x_1\beta_1 + \dots$$

$$d(x, y) = V[\log \frac{x}{y}]$$

Review on Regression



Review on Regression

$$Y = X \beta + e$$

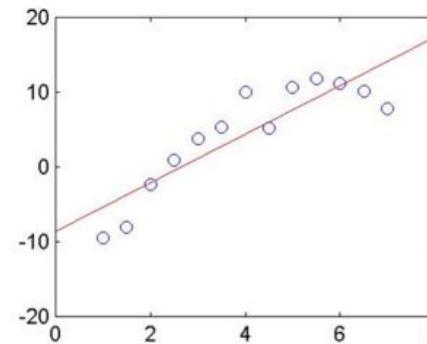
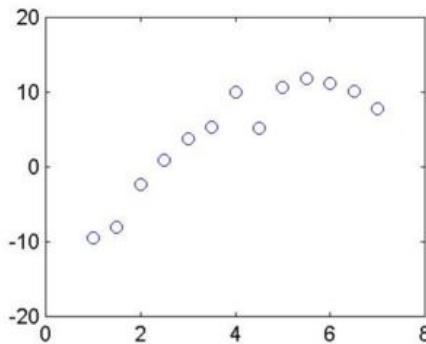
The diagram illustrates the components of a regression equation:

- Y : A vertical vector of $n \times 1$ measurements, represented as a stack of n light blue rectangles.
- $=$: An equals sign.
- X : A matrix of $n \times p$ covariates, represented as a grid of n rows and p columns. The first column contains ones, and the subsequent columns contain varying patterns of light blue and white squares.
- β : A vertical vector of $p \times 1$ coefficients, represented as a stack of p light blue rectangles.
- $+$: A plus sign.
- e : A vertical vector of $n \times 1$ residuals, represented as a stack of n light blue rectangles.

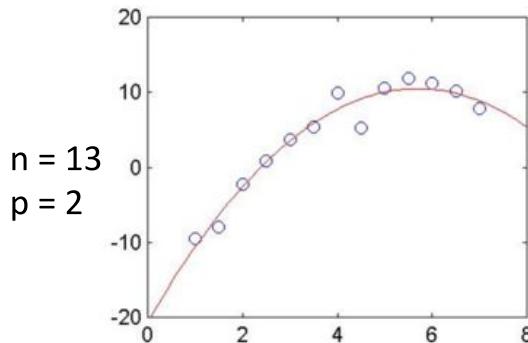
n = number of measurements

p = number of variables measured

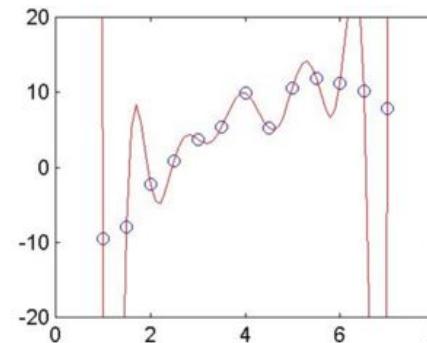
Occams Razor



$n = 13$
 $p = 1$



$n = 13$
 $p = 2$



$n = 13$
 $p = 13$

High dimensional variables

- Genetic data
 - Thousands of variables
 - Hundreds of samples
- Two solutions
 1. Regularization – wisely choose a subset of variables
 1. Multivariate response

High dimensional variables

- Genetic data
 - Thousands of variables
 - Hundreds of samples
- Two solutions
 1. Regularization – wisely choose a subset of variables
 1. Multivariate response

Advantages

- Effectively avoids over parameterization
 - $n \gg p$
 - But always try to run cross validation to confirm
 - Train model on subset of measurements
 - Try to predict the remaining measurements
- Build models with many covariates

Workflow

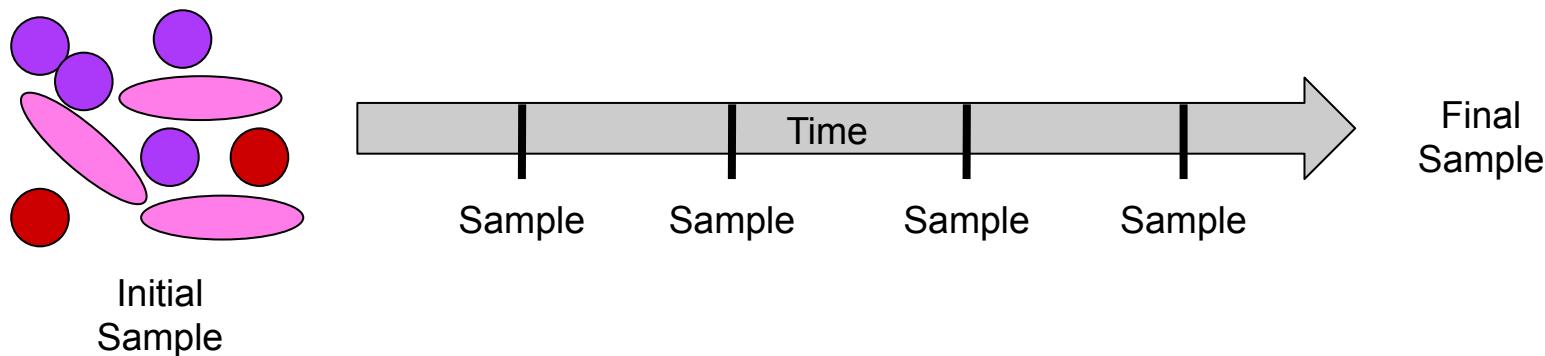
- Perform Clustering
- Compute Balances
- Perform Linear Regression

Longitudinal Studies and Analyses

<https://bit.ly/2HThBcx>

What are longitudinal studies?

- Employ continuous or repeated measures from an initial sample
- Follow individuals (or communities) over a prolonged period of time
- Generally observational in nature, with quantitative and/or qualitative data being collected on any combination of exposures and outcomes
- Reveal complex patterns of change



Why use longitudinal study design?

Advantages

1. Identify and relate events to exposures or treatments
2. Define these exposures with regards to time (How long do they last?, When do they happen?, etc.)
3. Establish sequence of events and patterns
4. Measure individual subject variability against the cohort
5. Provides information on microbial community development, stability, response, and recovery

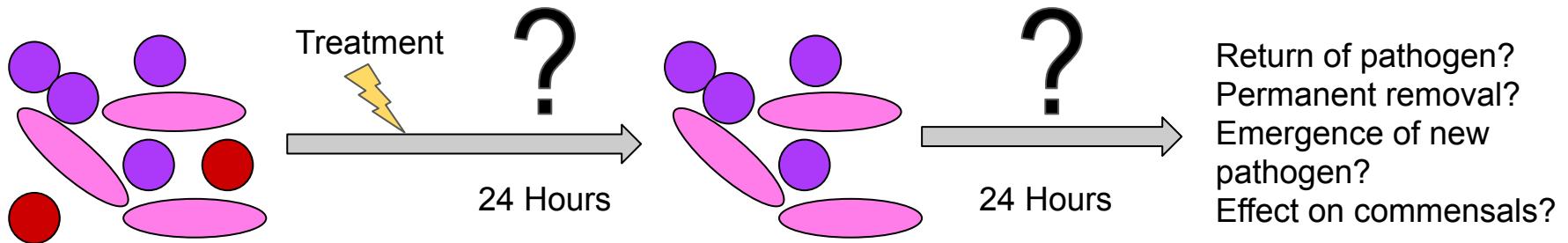
Why use longitudinal study design?

Disadvantages

1. Loss of participants over time - can reduce representative nature of the cohort
2. Difficulty in assessing the impact of exposure to outcome
3. Deeper understanding of statistical testing is needed to avoid inaccurate conclusions
4. Increased time commitment and financial demands

Why use longitudinal study design?

Pre-post study, but what else are you missing?



[Read our COVID-19 research and news.](#)

SHARE

RESEARCH ARTICLE | MICROBIOME



Antibiotics, birth mode, and diet shape microbiome maturation during early life

Nicholas A. Bokulich¹, Jennifer Chung¹, Thomas Battaglia¹, Nora Henderson¹, Melanie Jay^{1,2}, Huilin Li³, Arnon D. Lieber¹, F...

+ See all authors and affiliations

Science Translational Medicine 15 Jun 2016:

Vol. 8, Issue 343, pp. 343ra82

DOI: 10.1126/scitranslmed.aad7121

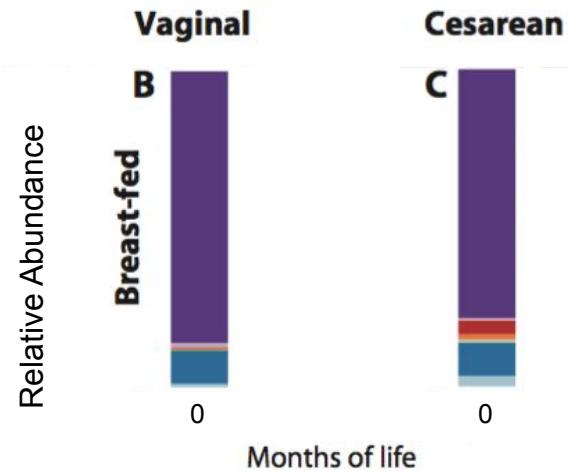
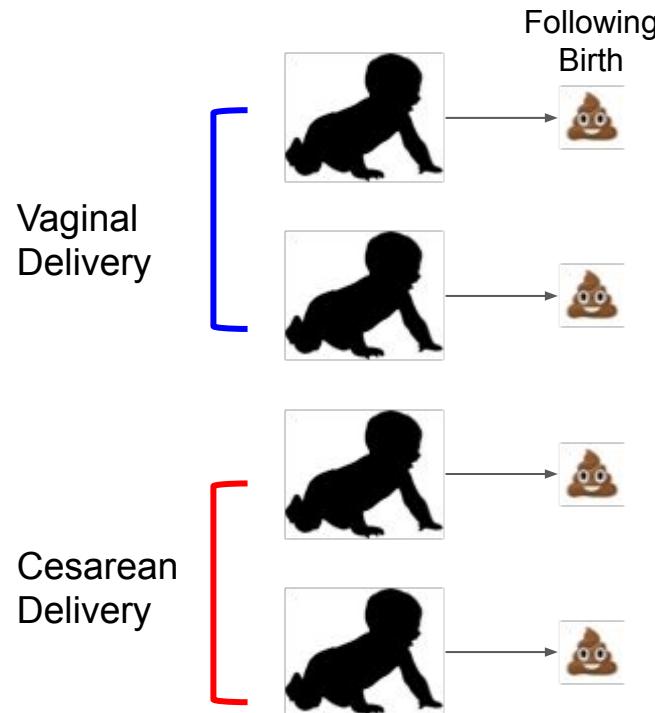


Science
Translational
Medicine

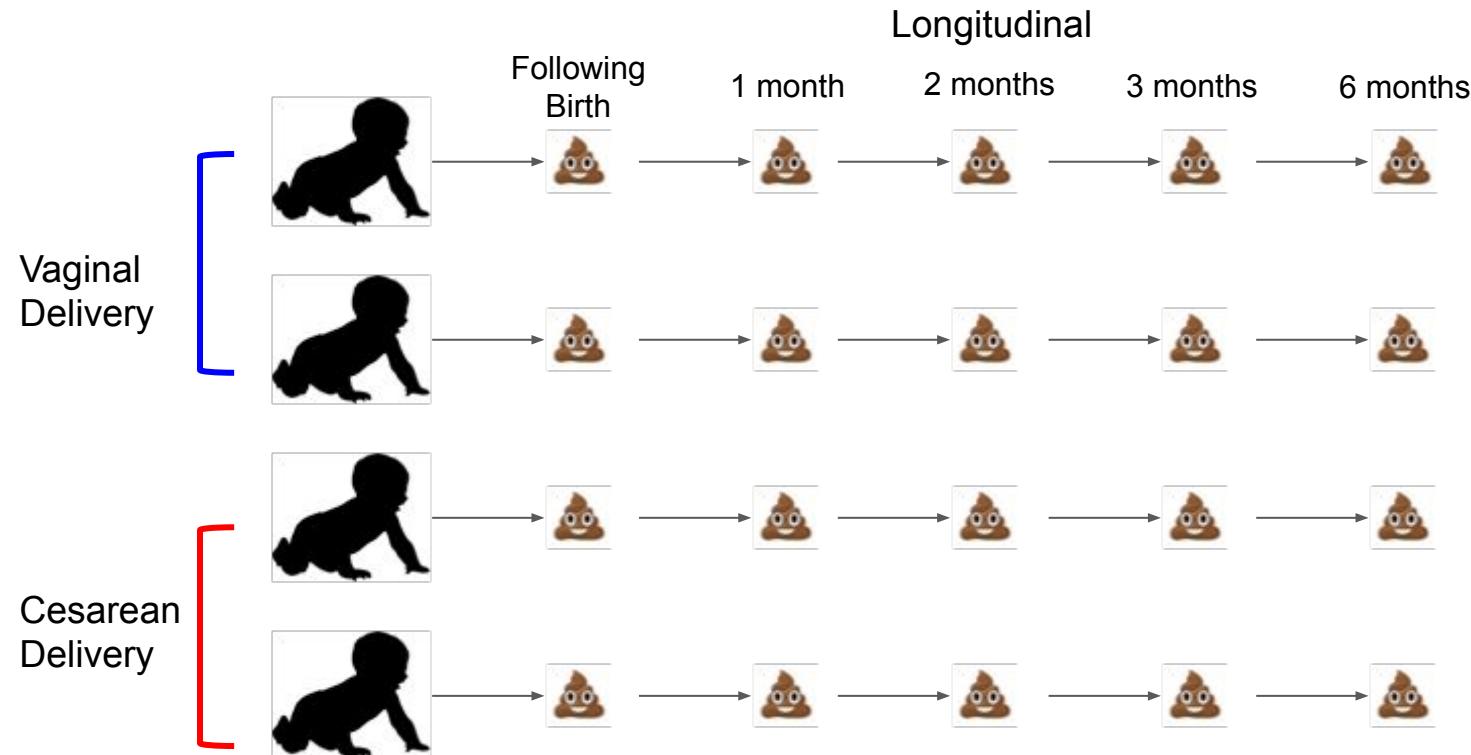
Vol 8, Issue 343
15 June 2016

Table of Contents

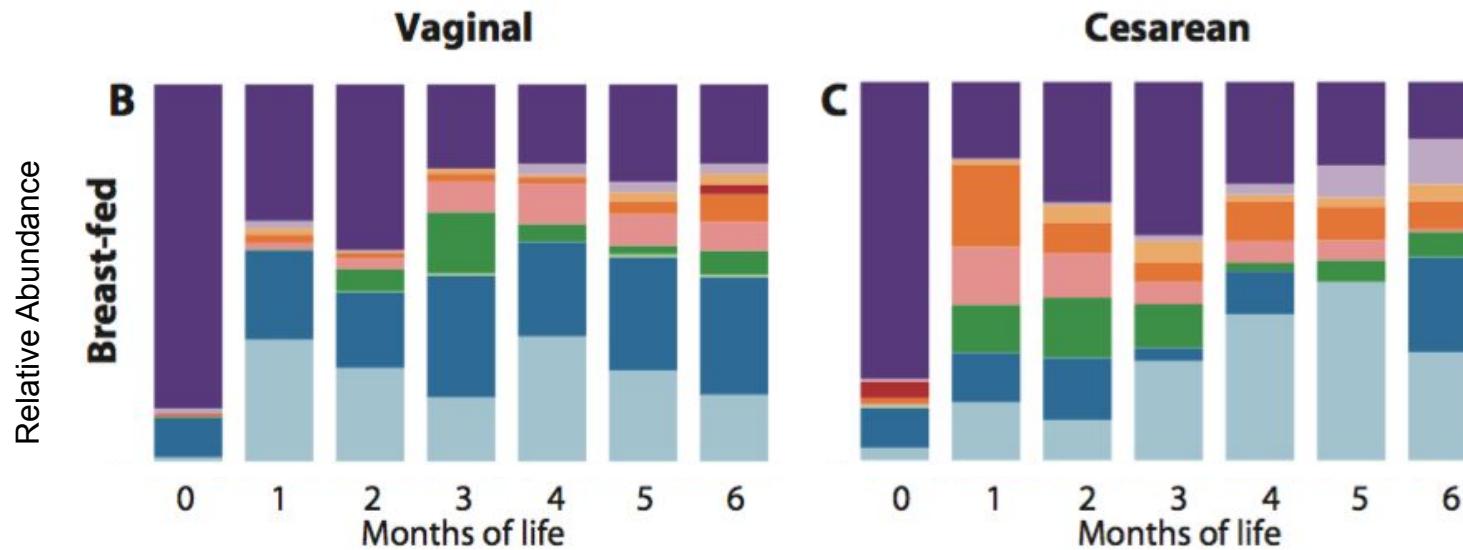
Antibiotics, birth mode, and diet shape microbiome maturation during early life



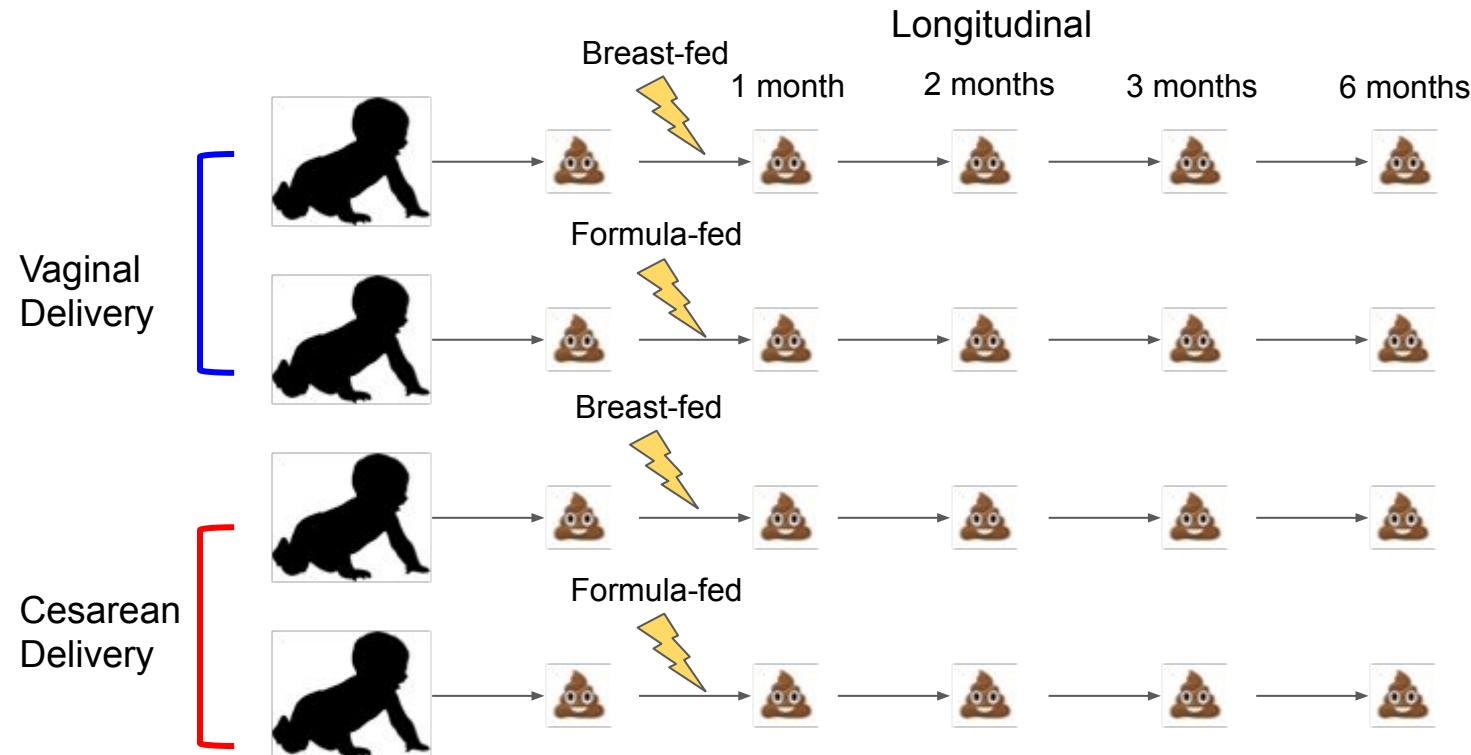
Antibiotics, birth mode, and diet shape microbiome maturation during early life



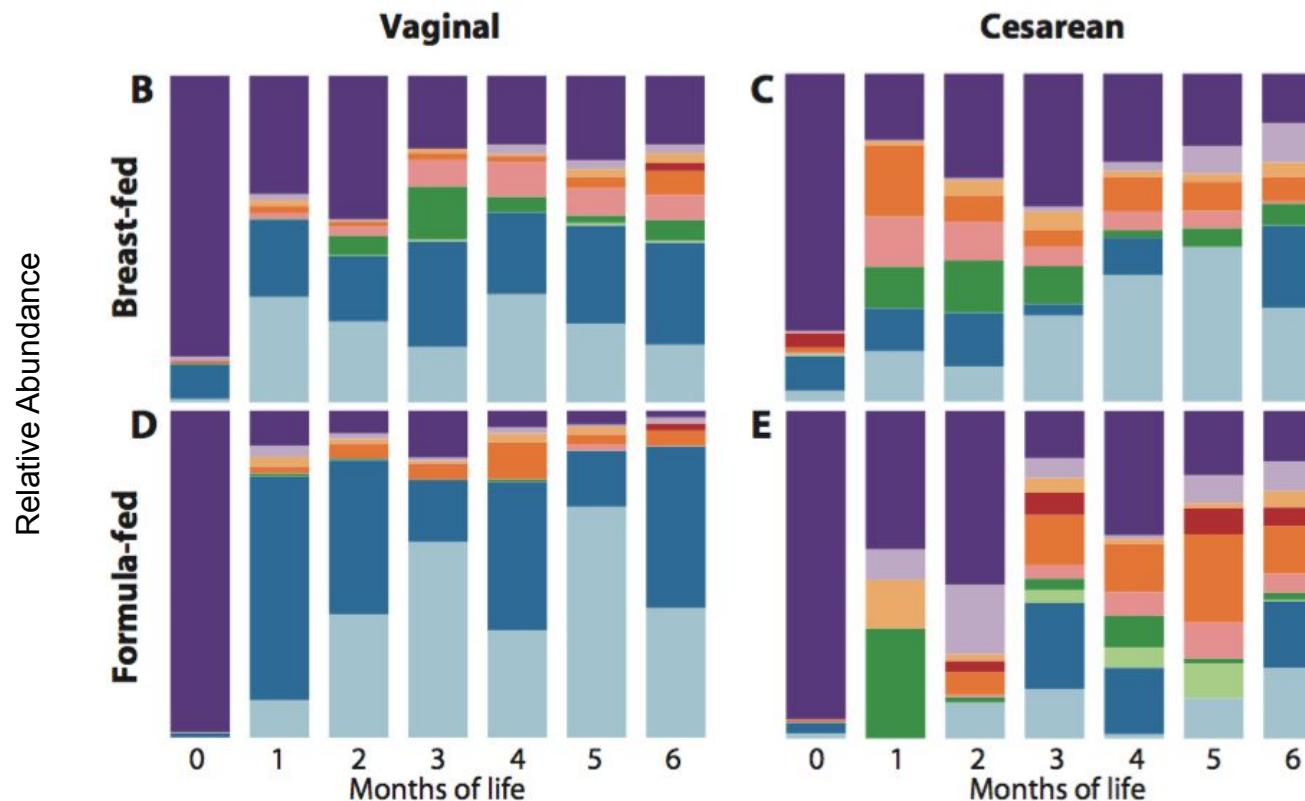
Antibiotics, birth mode, and diet shape microbiome maturation during early life



Antibiotics, birth mode, and diet shape microbiome maturation during early life



Antibiotics, birth mode, and diet shape microbiome maturation during early life



How to tackle a longitudinal study - Design and implementation

- Methods of data collection and recording must be standardized and consistent over time
- Frequency and degree of sampling should vary according to the specific research goals
- Effort should be made to ensure maximal retention of participants
- Implementation of longitudinal research projects can require an extensive amount of time - plan accordingly!
- Perform statistical analyses with care to avoid misrepresentation of data

The screenshot shows the CONSORT website's navigation bar at the top, featuring links for Home, Extensions, Downloads, Examples, Resources (which is highlighted), and About CONSORT. Below the navigation bar, there is a blue horizontal bar containing links for EQUATOR, Protocols (which is highlighted), TIDieR, Glossary, and Useful Links. The main content area has a light gray background and displays the word "SPIRIT" in large, bold, black capital letters. Below "SPIRIT", there is a paragraph of text explaining the purpose of the SPIRIT Initiative. At the bottom of the page, there is a section of text providing more details about the initiative.

SPIRIT

About half of clinical trial protocols lack key information when first submitted for research ethics board (REB) approval. An incomplete protocol impedes transparency and makes it difficult for REBs and other reviewers to understand the trial – leading to time wasted on protocol revisions, and delays in approval and enrolling participants. One third of protocol amendments can be avoided with greater attention to protocol content, saving an average of 6–16 weeks in delays and 4 hours of REB time per trial. Incomplete protocols can also lead to inconsistent trial conduct and in the worst case, trials that produce invalid data.

To address these concerns, the **SPIRIT Initiative** (Standard Protocol Items: Recommendations for Interventional Trials) developed evidence-based guidance for the content of trial protocols. Published in *Annals of Internal Medicine*, *BMJ*, and *Lancet*, SPIRIT aims to provide high-quality standards for trial protocols, while helping to streamline the process from protocol development to REB approval. High-quality protocols can promote transparency, do a better job of protecting trial participants, and help the trial meet its objectives with rigour and efficiency.

Caruana et al. 2015 *J Thorac Dis.*

How to tackle a longitudinal study - Statistical analyses

Factors to consider when performing analyses:

- Account for the intra-individual correlation of measures
- Fixed and dynamic/random effects
- Differences between time intervals; missing data

Common statistical methods used:

- ANOVA and MANOVA
- Mixed-effect regression models
- Generalized estimating equations

If tests commonly used for cross-sectional studies are used, the data will be underutilized, variability will be underestimated, and the likelihood of false negatives will increase.

Editor's Pick Methods and Protocols | Novel Systems Biology Techniques

q2-longitudinal: Longitudinal and Paired-Sample Analyses of Microbiome Data

Nicholas A. Bokulich, Matthew R. Dillon, Yilong Zhang, Jai Ram Rideout, Evan Bolyen, Huilin Li, Paul S. Albert, J. Gregory Caporaso

Mani Arumugam, *Editor*

DOI: 10.1128/mSystems.00219-18



Check for updates

A QIIME2 plugin designed for the appropriate use of longitudinal methods in microbiome studies

- Interactive plotting (volatility plots)
- Linear mixed-effects models
- Paired differences and distances
- Non-metric microbial interdependence testing (NMIT)
- First differences and distances
- Supervised regression for longitudinal feature identification

<https://docs.qiime2.org/2020.6/tutorials/longitudinal/>

Bokulich et al. 2018 *mSystems*

Reanalysis with q2-longitudinal

Science Translational Medicine

[Contents ▾](#)[News ▾](#)[Careers ▾](#)[Journals ▾](#)[Read our COVID-19 research and news.](#)**SHARE**[RESEARCH ARTICLE](#) | MICROBIOME

Antibiotics, birth mode, and diet shape microbiome maturation during early life

Nicholas A. Bokulich¹, Jennifer Chung¹, Thomas Battaglia¹, Nora Henderson¹, Melanie Jay^{1,2}, Huilin Li³, Arnon D. Lieber¹, F...

+ See all authors and affiliations

Science Translational Medicine 15 Jun 2016:

Vol. 8, Issue 343, pp. 343ra82

DOI: 10.1126/scitranslmed.aad7121



**Science
Translational
Medicine**

Vol 8, Issue 343

15 June 2016

[Table of Contents](#)

Volatility

Volatility

- An indication of the temporal stability of a metric over time and between subjects

Microbial volatility

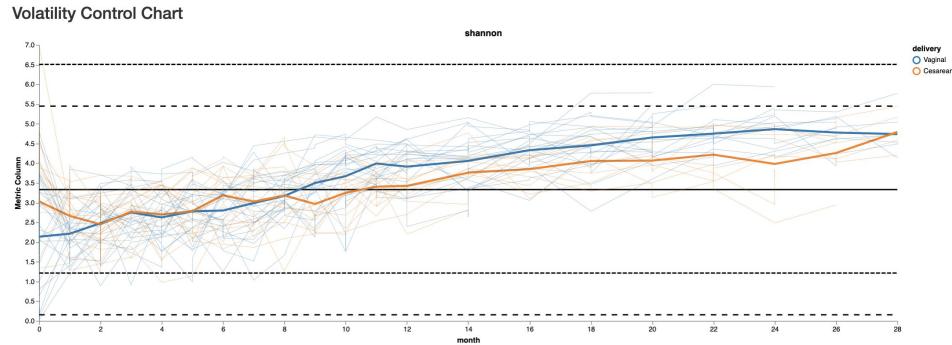
- “the variance in microbial abundance, diversity, or other metrics over time”

Feature volatility

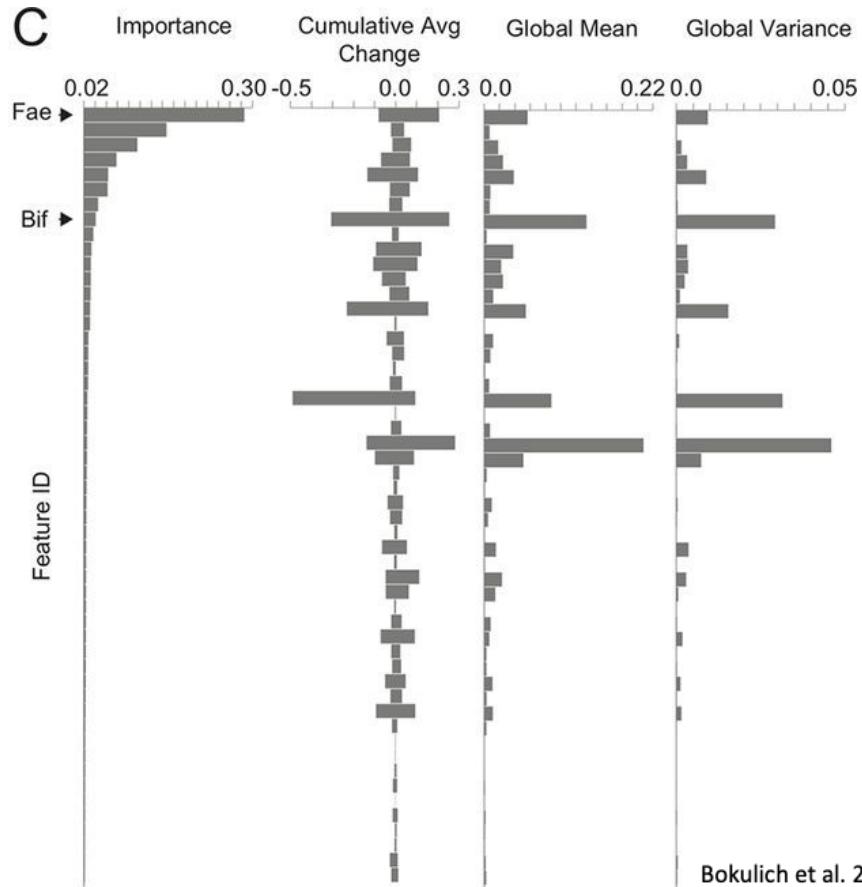
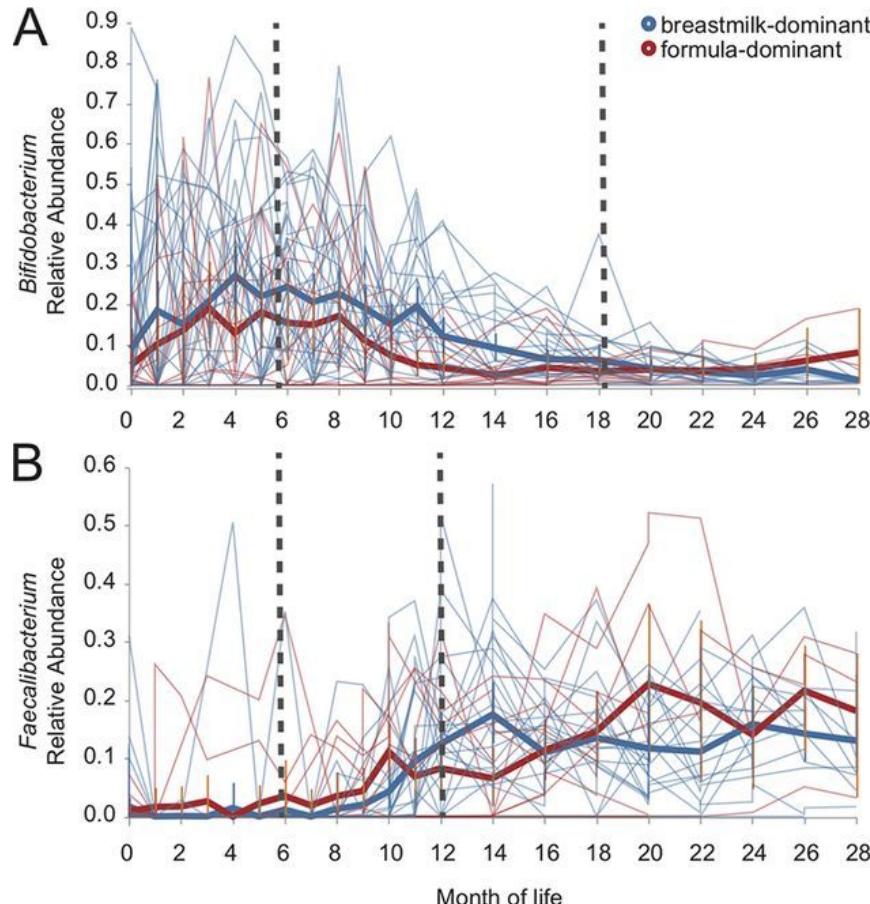
- Exploratory method: Identifies features that are predictive of a state/time point.
- A supervised regression model is used to identify important features (taxa) that can predict a state (e.g., time point in development)
- Unlike other methods, it can be used to identify important low abundant taxa

Volatility Charts

- Combine control charts and spaghetti plots
- Visualization of the change in a variable over time
- Control limits – 3 standard deviations from the mean (dotted line in positive direction)
- Warning limits – 2 standard deviations from the mean (dashed line in positive direction)
- Used to identify observations that are substantially deviating from the mean.



Identification of longitudinally volatile features: ECAM data



Statistical tests q2-longitudinal

- Linear mixed effects
 - LME models “examine the relationship between one or more independent variables (effects) and a single longitudinal response, where observations are made across dependent samples, e.g., in repeated-measures experiments.”
 - Includes random and fixed effects
- ANOVA
- Pairwise distances / differences
 - Wicoxon signed-rank test
 - Kruskal-wallis
 - Mann-whitney U
- Are the relative abundances of *Bifidobacterium* and *Faecalibacterium* impacted by time and subject?

TABLE 1 Linear mixed-effects model results for *Bifidobacterium* relative abundances between 6 and 18 months of life in the ECAM study^a

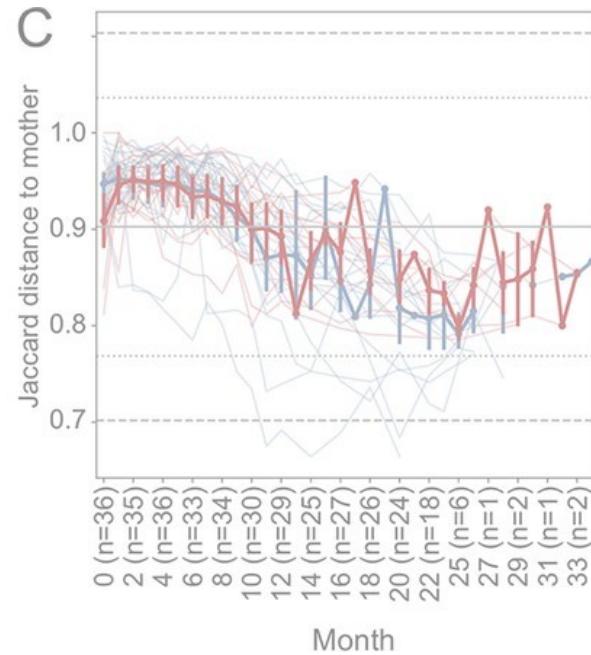
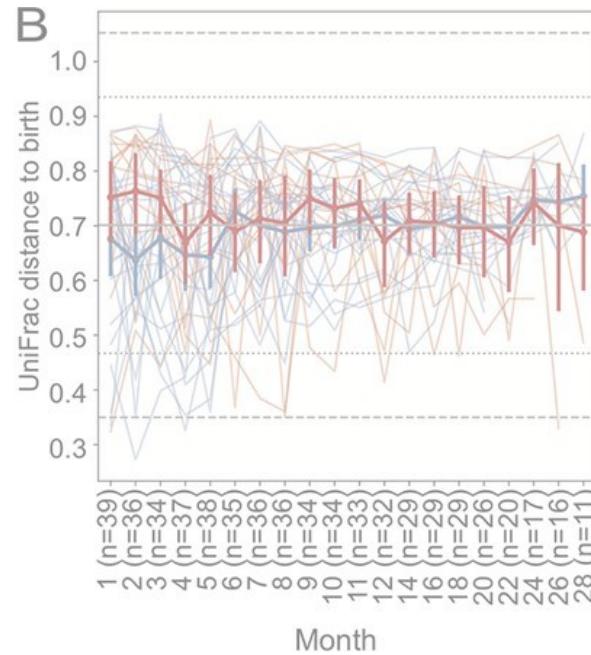
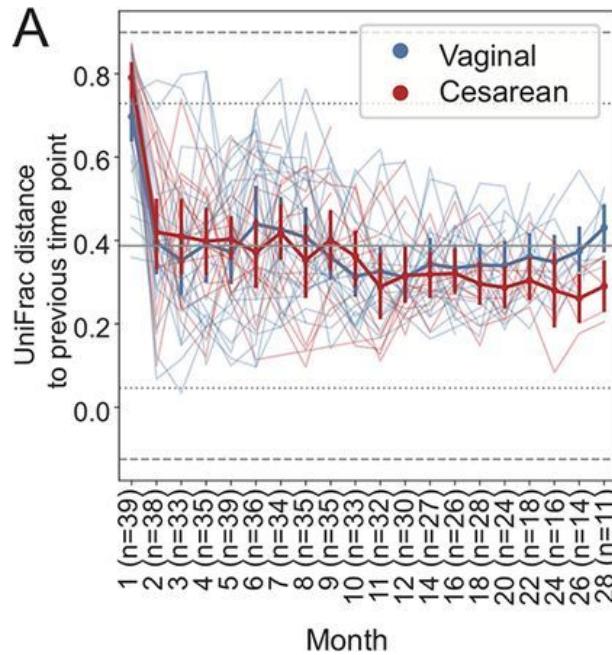
Model	Variable or parameter	Estimate	SE	Z-score	P value
Fixed effects	(Intercept)	0.465	0.101	4.579	<0.001
	Delivery [T.vaginal]	-0.114	0.097	-1.181	0.238
	Diet [T.formula-dominant]	-0.33	0.126	-2.612	0.009
	Sex [T.male]	-0.064	0.086	-0.743	0.457
	Delivery [T.vaginal]:diet [T.fd]	0.46	0.177	2.605	0.009
	Month	-0.015	0.017	-0.904	0.366
Random effects	Intercept (subject ID)	0.037	0.132		
	Slope (change per mo)	0.012	0.02		
	Covariance (intercept, time)	-0.004	0.228		

^aParameter estimate (coefficient), standard error, Z score, and P value for each model parameter. Brackets indicate reference groups for interpreting fixed-effect estimates.

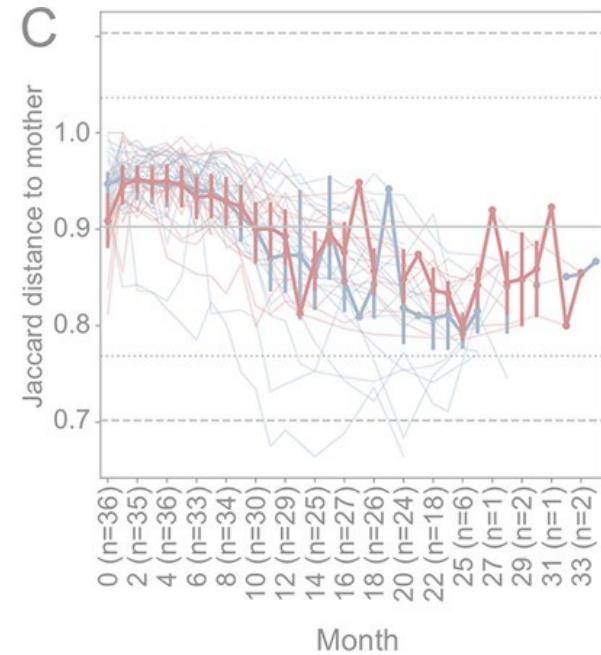
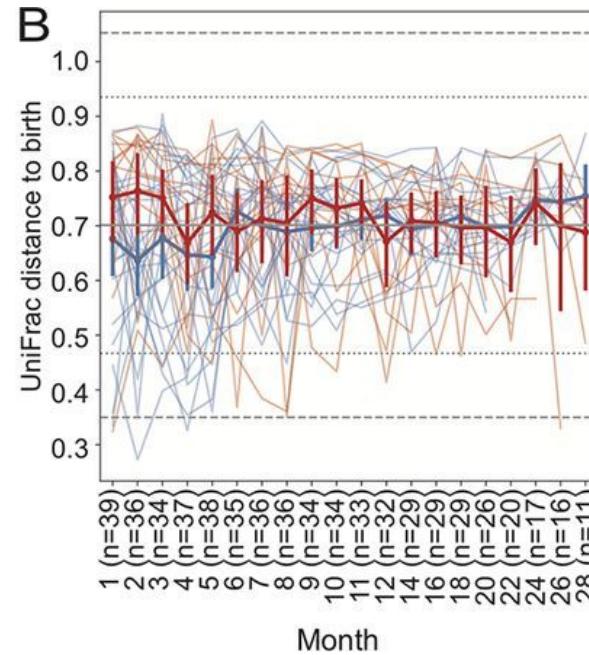
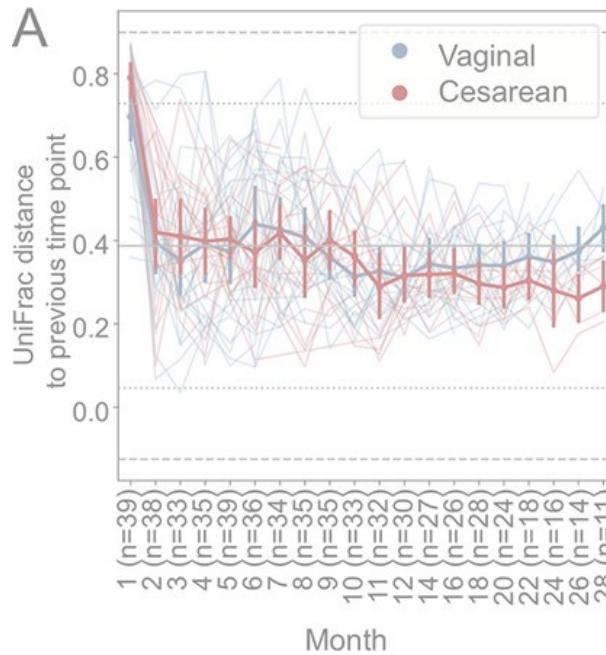
First differences / distances

- Tracks the rate of change over time
- First differences
 - Assesses the magnitude of change in some metadata value of interest (e.g., Shannon's diversity) between successive time points
- First distances
 - identifies the beta diversity distances between successive samples from the same subject
 - LME and volatility plotting cannot be applied directly on a distance matrix
- Both have an optional baseline parameter
 - Calculates differences/distances from a certain state (baseline or other important time point)
 - The baseline could also be a different subject (e.g., a mother compared to an infant)
- Visualized using volatility charts or analyzed with linear mixed effects models

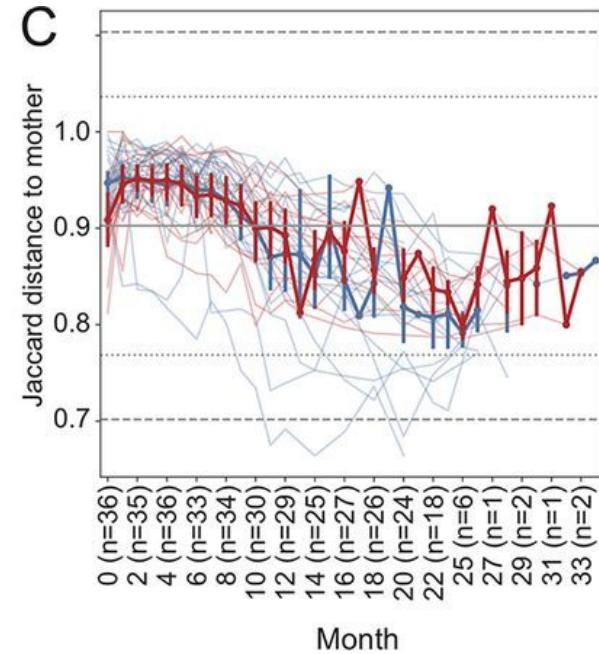
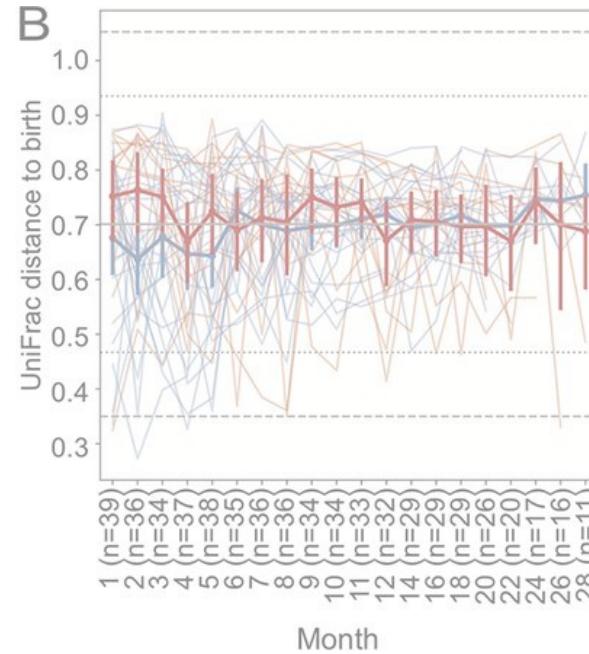
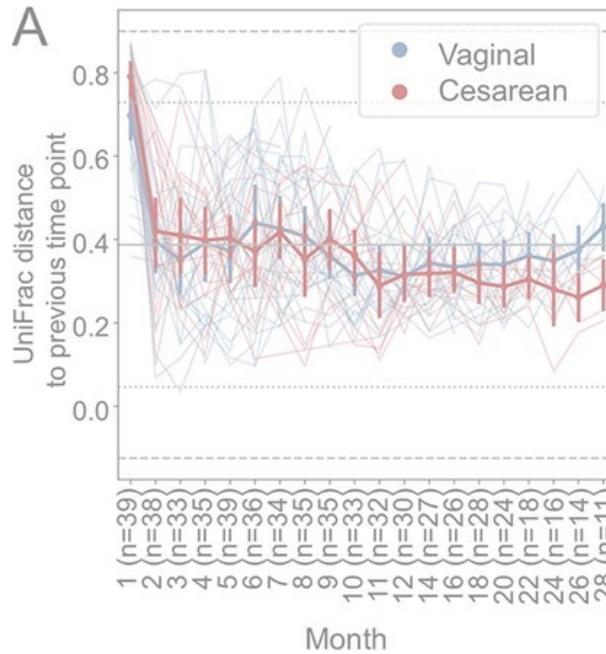
Tracking temporal changes in subjects' beta diversities



Tracking temporal changes in subjects' beta diversities



Tracking temporal changes in subjects' beta diversities



Other methods in q2-longitudinal

- Maturity Index prediction
 - Requires large sample sizes
 - Uses supervised regression
 - Usually includes control and treatment groups
 - Groups must be sampled evenly across time
 - Subramanian et al. 2014
- Non-parametric microbial interdependence test (NMIT)
 - Should have at least 5-6 time points per subject
 - Robust to missing samples
 - Can take a long time to run
 - Zhang et al. 2017

Published: 04 June 2014

Persistent gut microbiota immaturity in malnourished Bangladeshi children

Sathish Subramanian, Sayeeda Huq, Tanya Yatsunenko, Rashidul Haque, Mustafa Mahfuz, Mohammed A. Alam, Amber Benzra, Joseph DeStefano, Martin F. Meier, Brian D. Muegge, Michael J. Barratt, Laura G. VanArendonk, Qunyuan Zhang, Michael A. Province, William A. Petri Jr, Tahmeed Ahmed & Jeffrey I. Gordon 

Nature **510**, 417–421(2014) | Cite this article

4644 Accesses | 462 Citations | 271 Altmetric | Metrics

Received: 12 December 2016 | Revised: 30 May 2017 | Accepted: 10 July 2017

DOI: 10.1002/gepi.22065

RESEARCH ARTICLE

WILEY Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

A multivariate distance-based analytic framework for microbial interdependence association test in longitudinal study

Yilong Zhang¹ | Sung Won Han² | Laura M. Cox³ | Huilin Li^{4,5} 

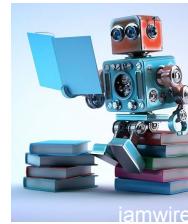
Conclusions

- Longitudinal microbiome studies are invaluable.
 - Understand temporal trends in the microbiome
 - Assess within and between-subject variation across time
- q2-longitudinal plugin supports a range of longitudinal testing methods and interactive visualizations to ease the analysis of longitudinal data
 - [q2-longitudinal tutorial](#)

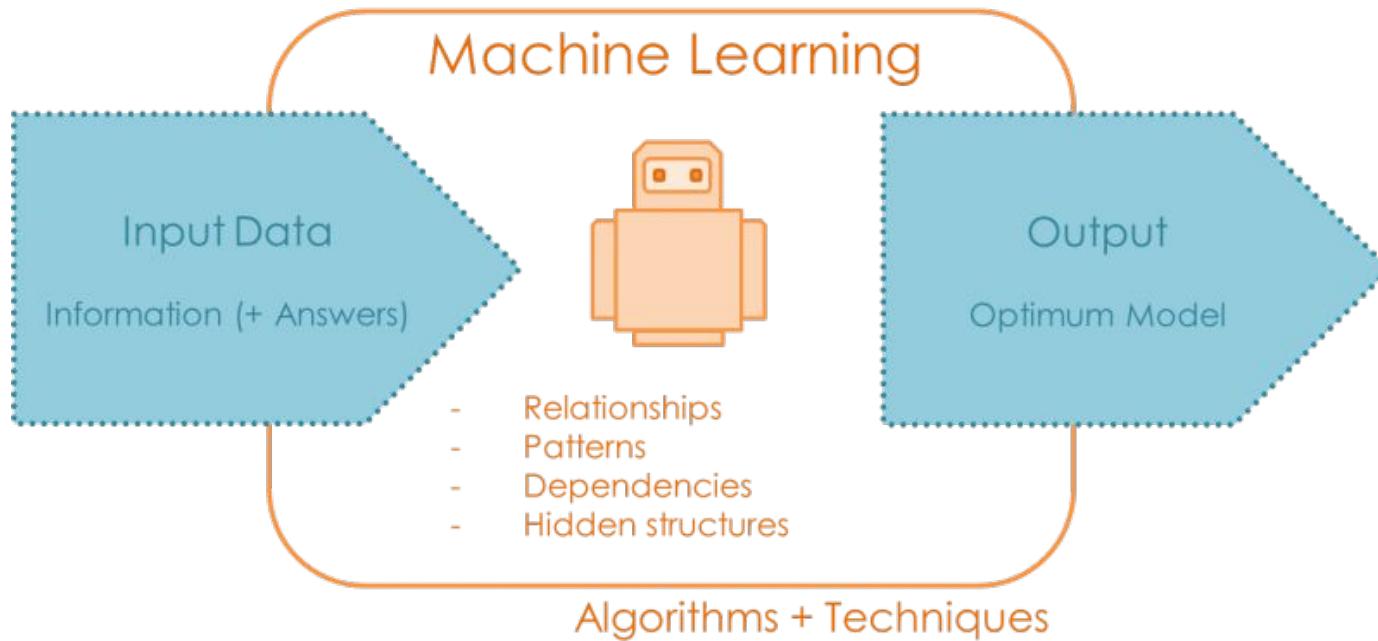
q2-sample-classifier

<https://bit.ly/2HThBcx>

What is machine learning?



iamwire



Supervised Learning:

Predicting values. **Known** targets.

User inputs correct answers to learn from. Machine uses the information to guess new answers.

REGRESSION:

Estimate continuous values
(Real-valued output)

CLASSIFICATION:

Identify a unique class
(Discrete values, Boolean, Categories)

Unsupervised Learning:

Search for structure in data. **Unknown** targets.

User inputs data with undefined answers. Machine finds useful information hidden in data.

Cluster Analysis

Group into sets

Density Estimation

Approximate distributions

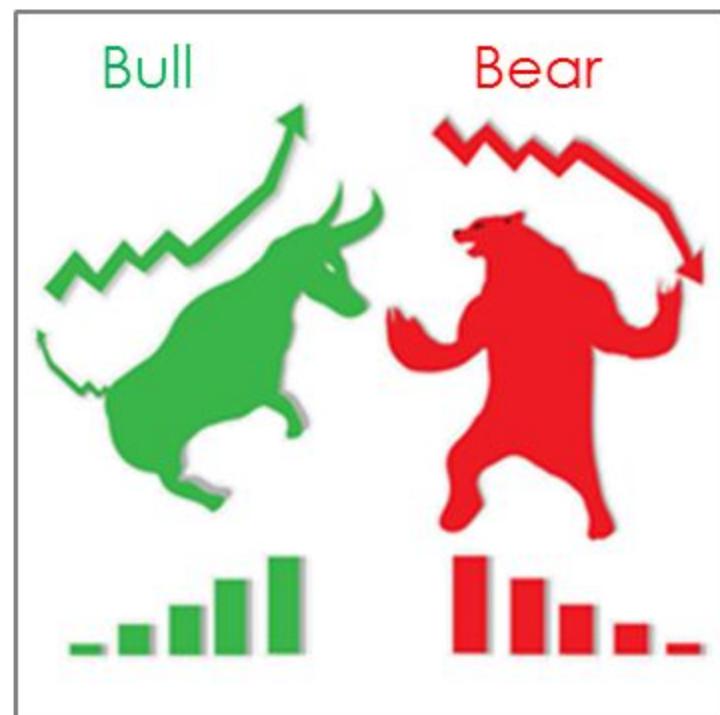
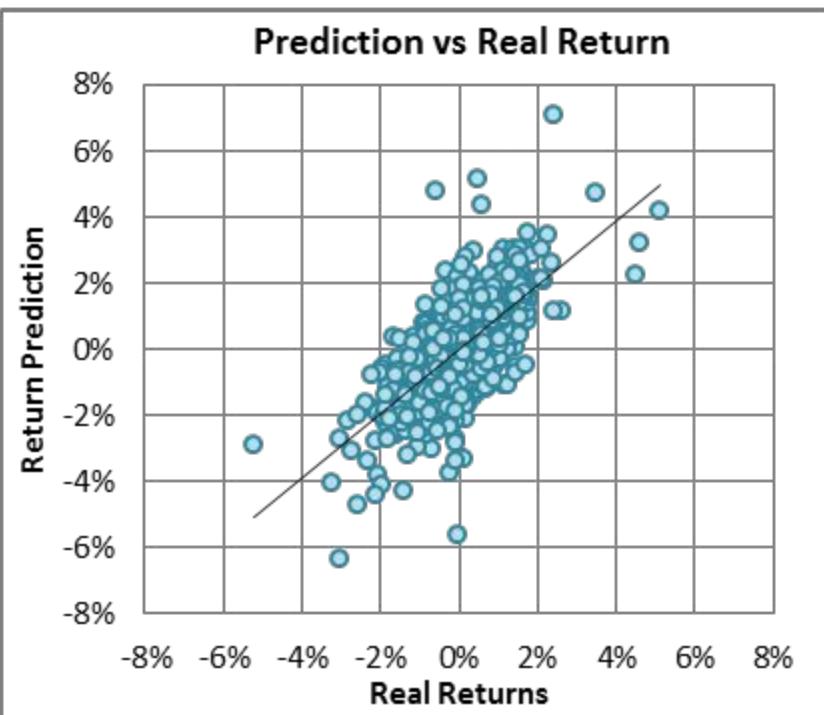
Dimension Reduction

Select relevant variables

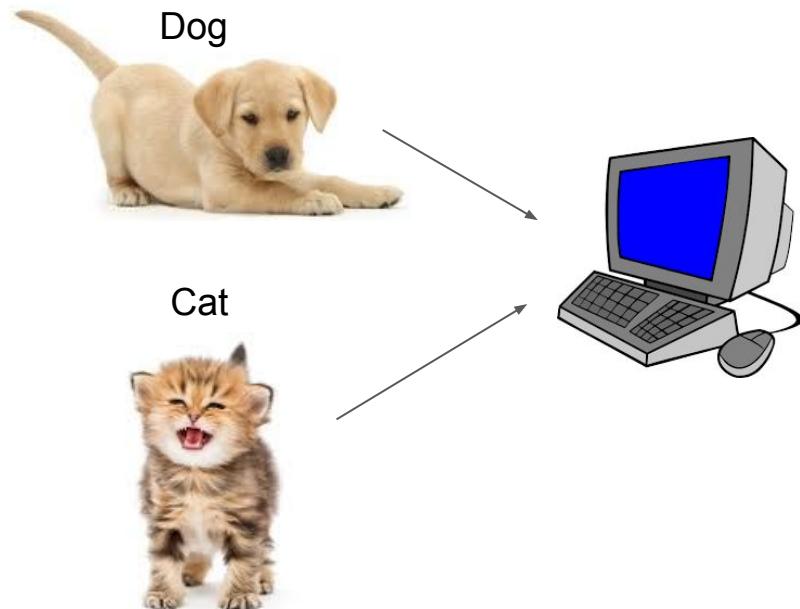
Regression

vs

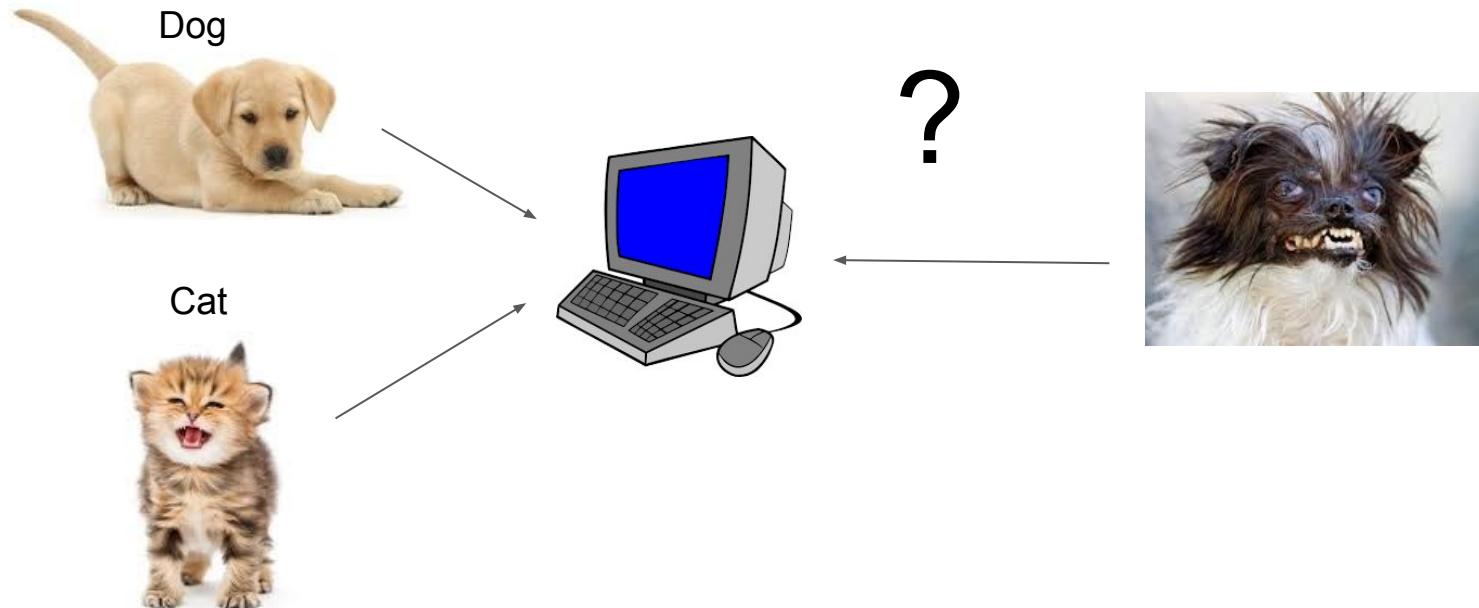
Classification



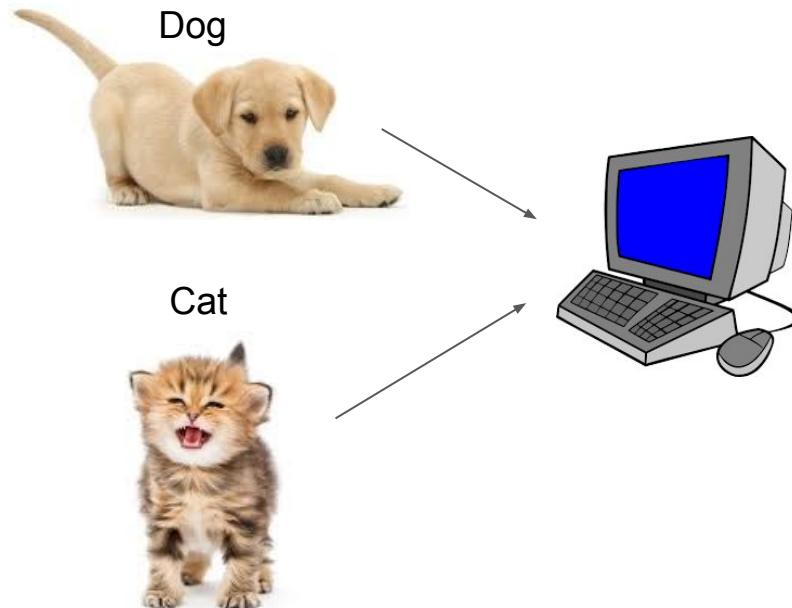
Supervised learning



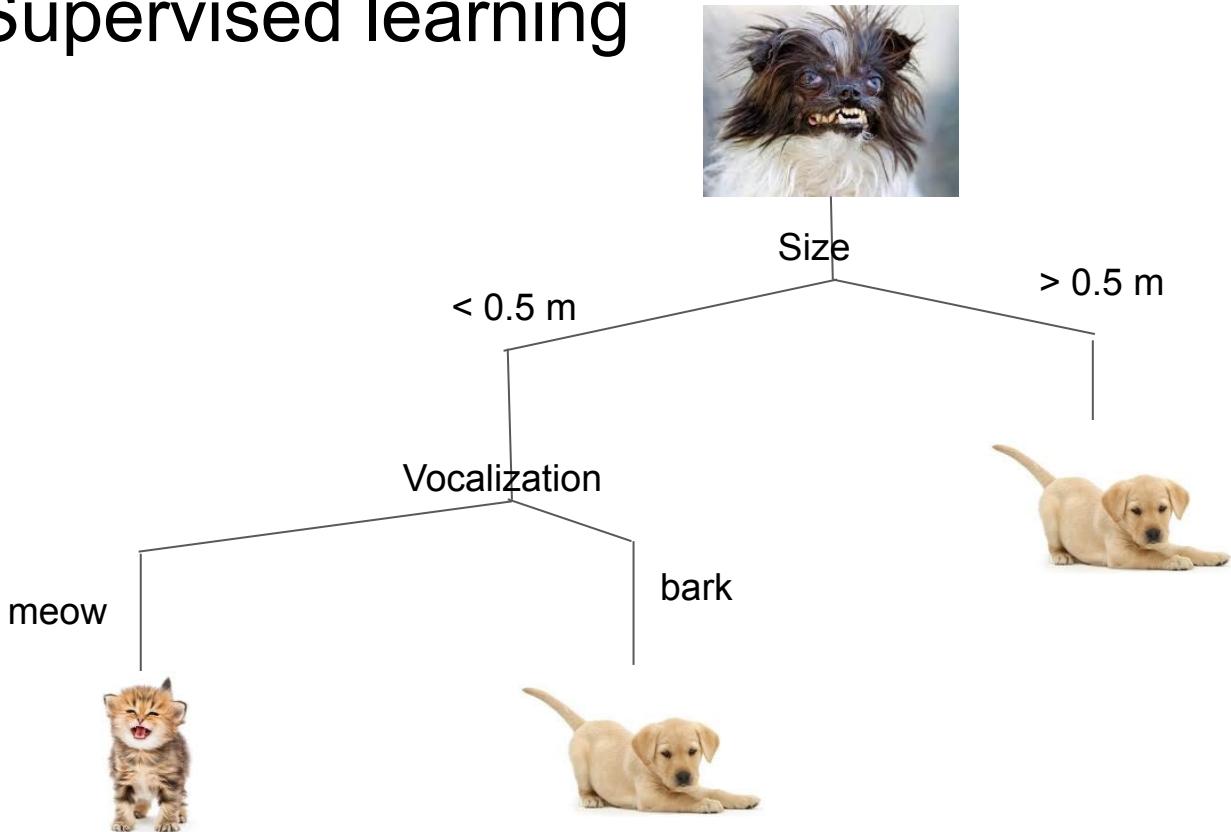
Supervised learning



Supervised learning



Supervised learning

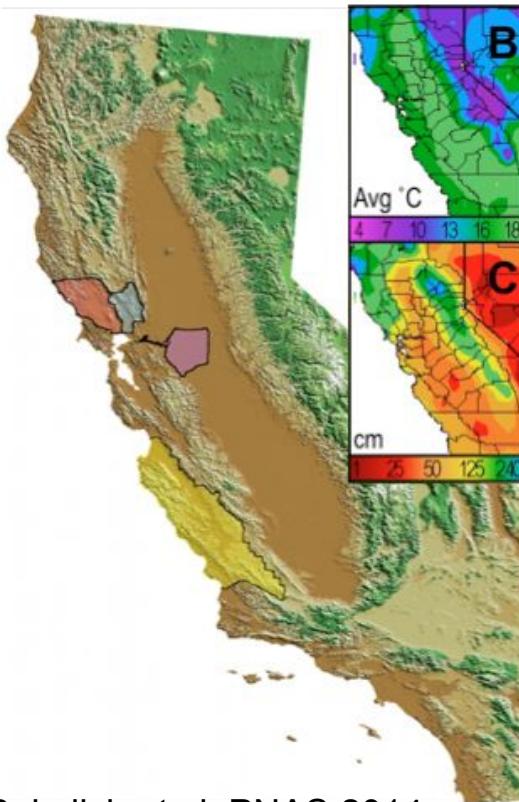


Supervised learning

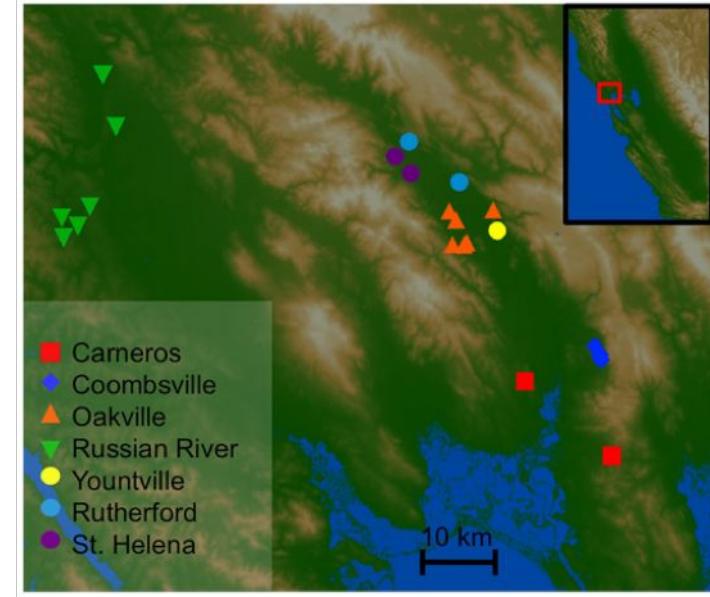
- Do microbiota predict sample metadata or other data?
 - Diagnose disease state - classification
 - Estimate time - regression
 - Metabolites
- What features best predict/distinguish sample states?
 - Taxa
 - Function

<https://docs.qiime2.org/2020.8/tutorials/sample-classifier/>

Supervised learning in action: wine terroir



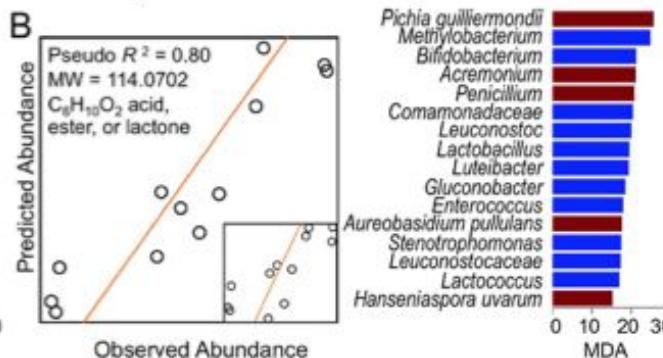
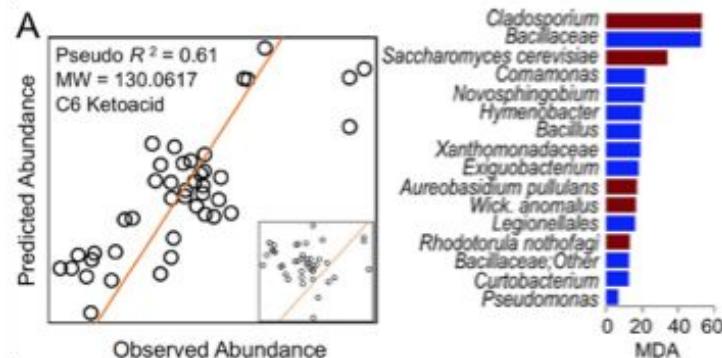
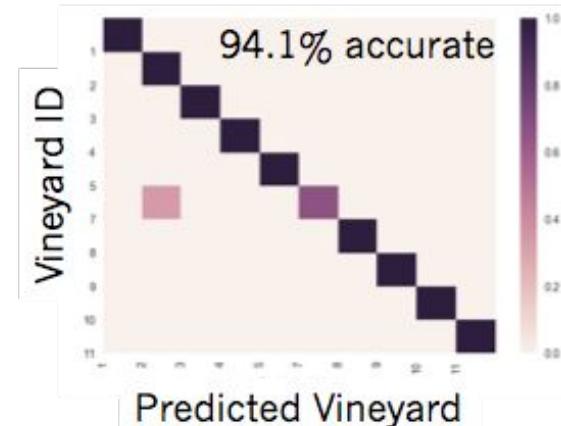
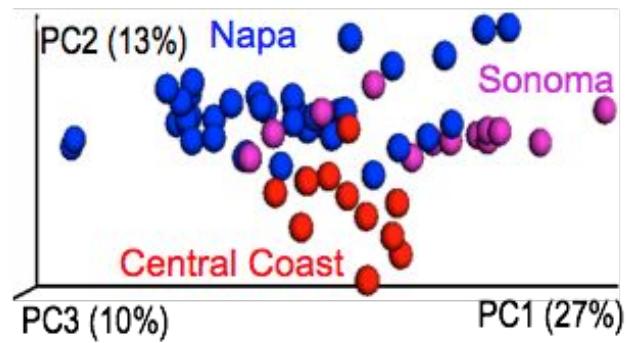
- Terroir: Regional wines, regional sensory traits
- Microbial terroir?
- We collected
 - ~1000 samples
 - 12 wineries
 - 4 growing regions
 - 3 years: 2010-2012
- Goals:
 - Distinguish regions/vineyards
 - Identify regional signatures
 - Predict metabolome



RESEARCH ARTICLE
Associations among Wine Grape Microbiome, Metabolome, and Fermentation Behavior Suggest Microbial Contribution to Regional Wine Characteristics

Nicholas A. Bokulich,^{a,b,*} Thomas S. Collins,^{b,d*} Chad Masarweh,^a Greg Allen,^a Hildegarde Heymann,^b Susan E. Ebeler,^{b,d} David A. Mills^{a,b,c}

Supervised learning in action: wine terroir



Humans differ in their personal microbial cloud

James F. Meadow^{1,2}, Adam E. Alricher^{1,2}, Ashley C. Bateman^{1,2},
Jason Stenson^{1,3}, GZ Brown^{1,3}, Jessica L. Green^{1,2,4} and
Brendan J.M. Bohannan^{1,2}

¹ Biology and the Built Environment Center, University of Oregon, Eugene, OR, USA

² Department of Biology, Institute of Ecology and Evolution, University of Oregon, Eugene, OR, USA

³ Department of Architecture, Energy Studies in Buildings Laboratory, University of Oregon,
Eugene, OR, USA

⁴ Santa Fe Institute, Santa Fe, NM, USA



Lax et al. *Microbiome* (2015) 3:21
DOI 10.1186/s40168-015-0082-9



Microbiome

RESEARCH

Open Access

Forensic analysis of the microbiome of phones and shoes

Simon Lax^{1,2*}, Jarrad T Hampton-Marcell¹, Sean M Gibbons^{1,3}, Geórgia Barguil Colares⁴, Daniel Smith^{1,5},
Jonathan A Eisen^{6,7,8} and Jack A Gilbert^{1,2,3,9,10}

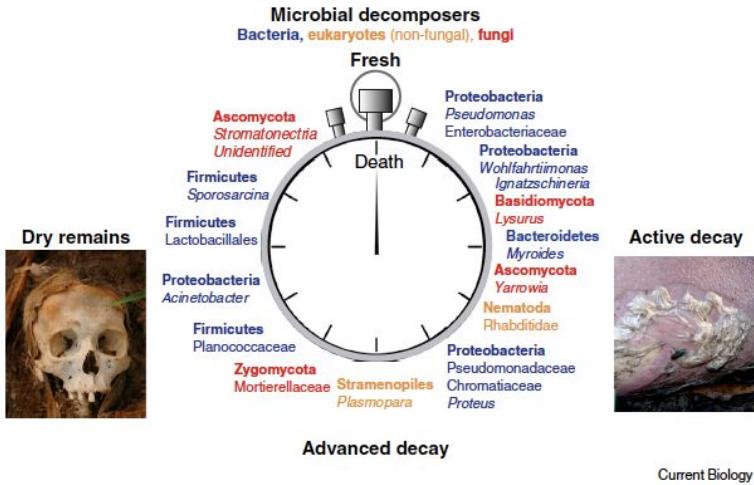
Image credit:

<http://blog.hardydiagnostics.com/wp-content/uploads/2016/02/19dg9kp2dtz8qjpg-768x484.jpg>

MICROBIOME

Microbial community assembly and metabolic function during mammalian corpse decomposition

Jessica L. Metcalf,^{1,2*} Zhenjiang Zech Xu,² Sophie Weiss,³ Simon Lax,^{4,5} Will Van Treuren,⁶ Embrette R. Hyde,² Se Jin Song,^{1,2} Amnon Amir,² Peter Larsen,^{4,7} Naseer Sangwan,^{4,7,8} Daniel Haarmann,⁹ Greg C. Humphrey,² Gail Ackermann,² Luke R. Thompson,² Christian Lauber,¹⁰ Alexander Bibat,¹¹ Catherine Nicholas,¹¹ Matthew J. Gebert,¹¹ Joseph F. Petrosino,¹² Sasha C. Reed,¹³ Jack A. Gilbert,^{4,5,7,8,14} Aaron M. Lynne,⁹ Sibyl R. Bucheli,⁹ David O. Carter,¹⁵ Rob Knight^{2,16*}

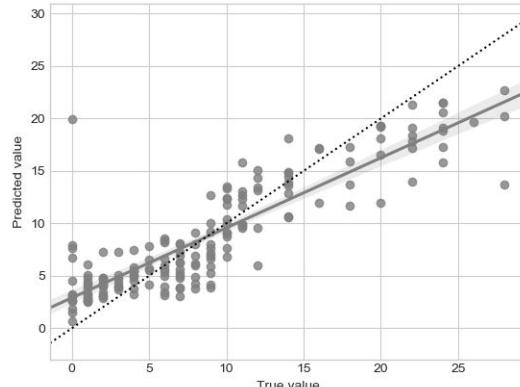
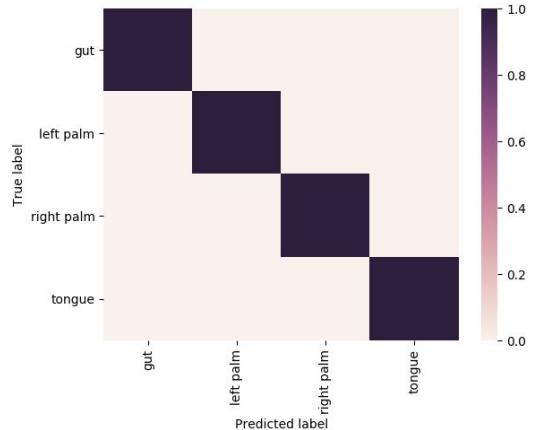
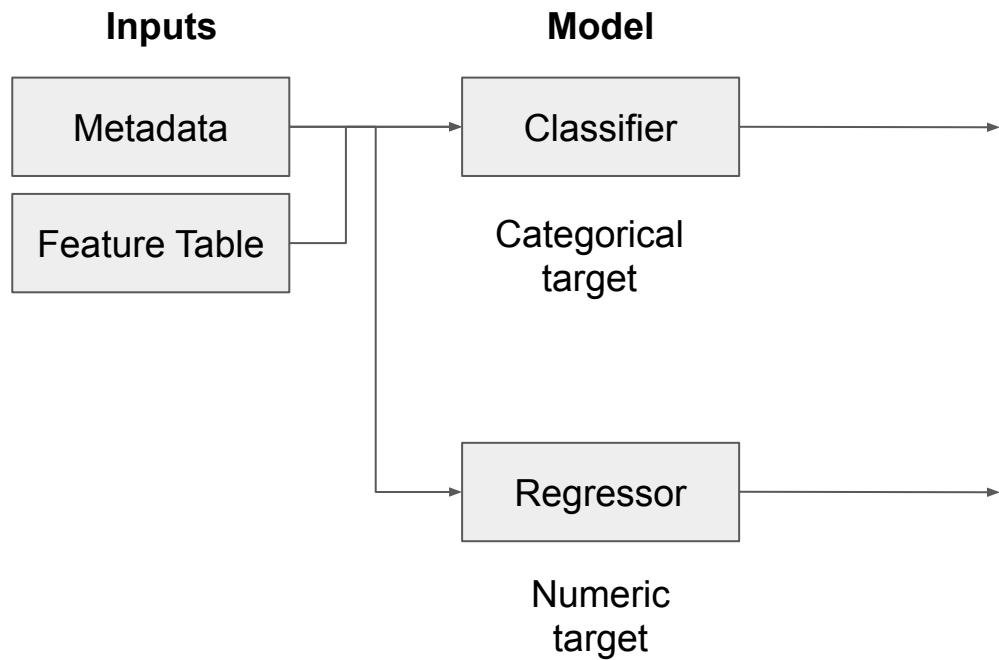


Article

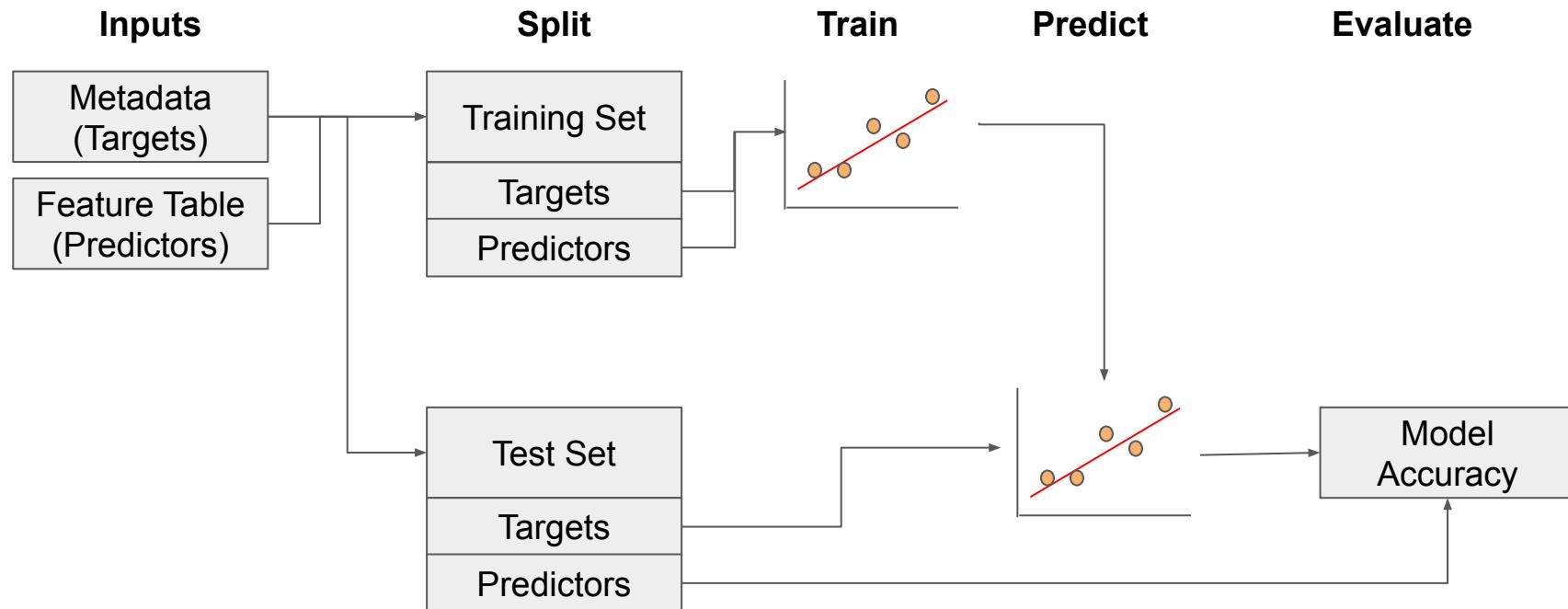
Microbiome Data Accurately Predicts the Postmortem Interval Using Random Forest Regression Models

Ariel Belk ^{1,†}, Zhenjiang Zech Xu ^{2,†}, David O. Carter ³, Aaron Lynne ⁴, Sibyl Bucheli ⁴ , Rob Knight ^{2,5,6} and Jessica L. Metcalf ^{1,*}

Supervised learning



Supervised learning



Sample classifier pipelines

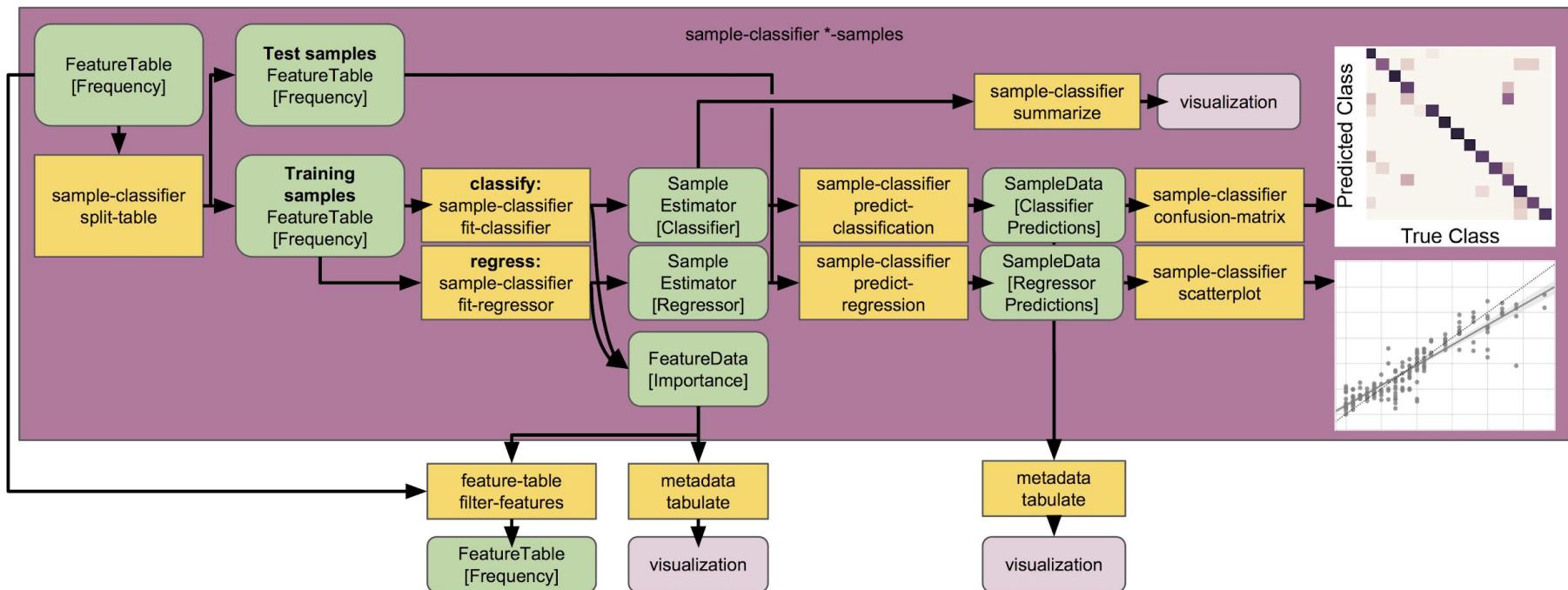
1. Split samples

2. Train model

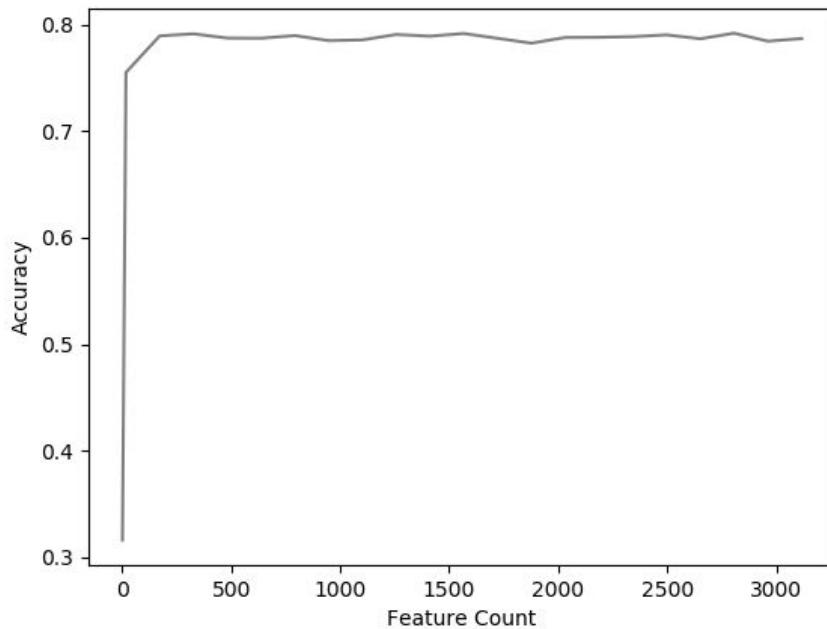
3. Optimization

4. Predict test samples

5. Evaluate accuracy



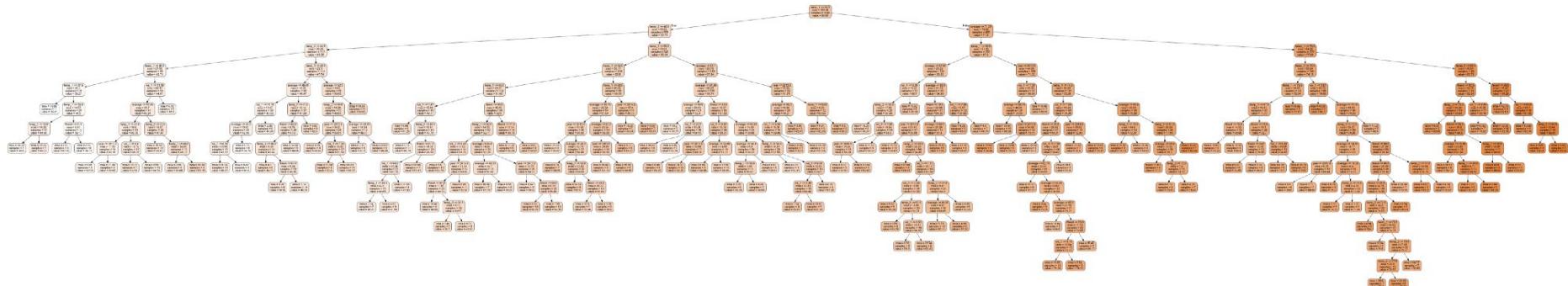
Feature extraction



feature	importance
asv1	8.00E-02
asv2	7.76E-02
asv3	5.94E-02
asv4	5.11E-02
asv5	4.87E-02
...	...

Random Forests

- Series of decision trees, then aggregates them together for final decision
- Advantages:
 - Classification and regression
 - Won't overfit model
 - Easy to pull out relative importance
- Disadvantage:
 - Can get too large, takes time to run
- https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html



Machine learning algorithms

	Linear Regression (lasso, ridge, elastic net)	Logistic Regression	Linear Discriminant Analysis	Naive Bayes	KNN	SVM	Tree (CART)	Random Forest	Boosting Tree
Regression	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes
Classification	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Nonlinearity	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Working in microbiome data	+	+	+	+	++ (using UniFrac distance)	+	+	+++	+++

How is this useful for the mouse tutorial?

- We can determine how predictive the microbiome composition is of other characteristics about a sample.
- In context of the tutorial, the question this is seeking to answer:

Can we use the fecal microbiome to predict if a patient is susceptible to Parkinson's Disease?

Closing notes

<https://bit.ly/2HThBcx>

<u>Instructor</u>	<u>QIIME 2 Forum</u>	<u>Institutional affiliation</u>
Aeriel Belk	aeriel.belk	Department of Animal Sciences, Colorado State University
Alex Emmons	emmo1	Department of Animal Sciences, Colorado State University
Andrew Sanchez	andrewsanchez	Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, USA
Anthony Simard	Oddant1	Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, USA
Ben Kaehler	BenKaehler	School of Science, University of New South Wales, Canberra, Australia
Chloe Herman	cherman2	Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, USA
Chris Keefe	ChrisKeefe	Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, USA
Emily Borsom	emilyborsom	Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, USA
Emily Cope	Emily_Cope	Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, USA
Evan Bolyen	ebolyen	Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, USA
Greg Caporaso	gregcaporaso	Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, USA
Heather Deel	hdeel	Department of Animal Sciences, Cell and Molecular Biology Special Academic Unit, Colorado State University
Jamie Morton	mortonjt	Center for Computational Biology, Flatiron Institute, Simons Foundation, New York, USA
Jessica Metcalf	jessicalmetcalf	Department of Animal Sciences, Colorado State University
Justine Debelius	jwdebelius	Centre for Translational Microbiome Research, Department of Microbiology, Tumor, and Cancer Biology, Karolinska Institutet, Stockholm, Sweden
Matthew Dillon	thermokarst	Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, USA
Mehrbod Estaki	Mehrbod_Estaki	Department of Pediatrics, University of California San Diego, USA
Mike Robeson	SoilRotifer	Department of Biomedical Informatics, College of Medicine, University of Arkansas for Medical Sciences, Little Rock AR, USA
Nick Bokulich	nicholas_bokulich	Laboratory of Food Systems Biotechnology, Institute of Food, Nutrition, and Health, ETH Zürich, Switzerland
Renato Oliveira	reinator	Environmental Genomics, Instituto Tecnológico Vale, Belém, Pará, Brazil
Yoshiki Vazquez-Baeza	yoshiki	Center for Microbiome Innovation, Jacobs School of Engineering, University of California San Diego, USA

Thanks to our funders...

QIIME 2 project funding

National Cancer Institute: ITCR ([1U24CA248454-01](#))

National Science Foundation ([1565100](#))

Chan-Zuckerberg Initiative

Other funding sources

National Cancer Institute: [Partnership for Native American Cancer Prevention](#) (U54CA143925;

Caporaso Lab)

Alfred P. Sloan Foundation (Caporaso Lab)

RCUK (BB/P027849/1; CABANA)

ERC Horizon 2020 825410 (Justine Debelius)

and to our hosts...

Chan-Zuckerberg Initiative

CABANA

Cath Brooksbank

Guilherme Oliveira

Piraveen Gopalasingam

QIIME 2 is a community effort

Thanks to our [plugin developers](#), [QIIME 2 Forum moderators](#), [QIIME 2 developers](#), [QIIME 1 Forum masters](#), [QIIME 1 developers!](#)

Thank you for using and supporting QIIME 2, and for spending the week with us!

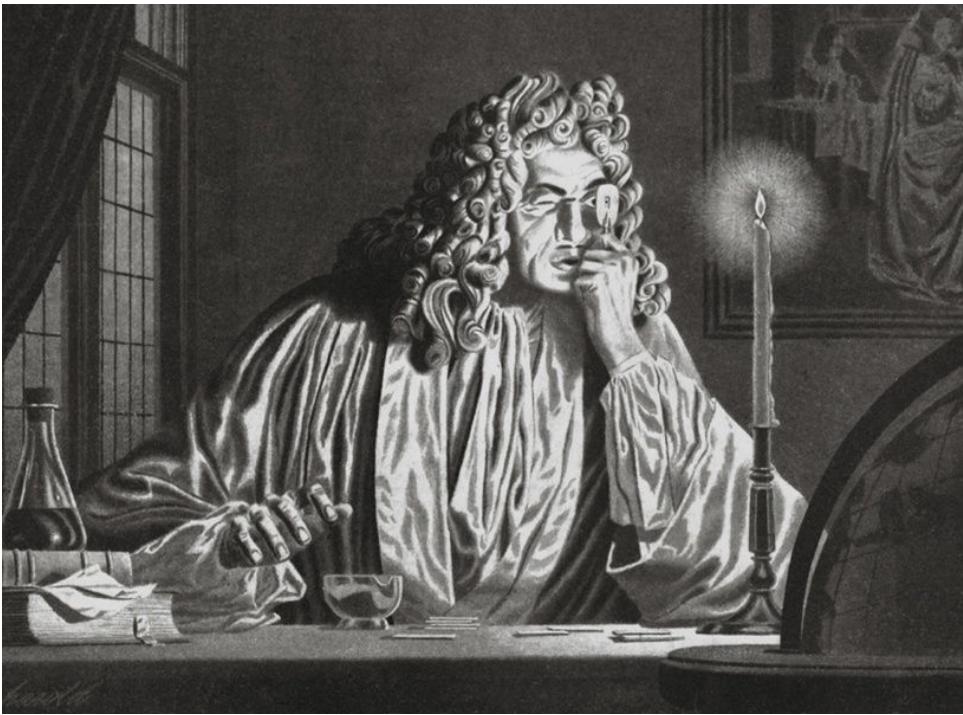


Image source: Discover Magazine (2015)

“[Secretary of the Royal Society] Oldenburg wrote to Leeuwenhoek, asking him to ‘acquaint us with his method of observing, that others may confirm such Observations as these’, and to provide drawings [15]. Leeuwenhoek declined, throughout his life, to give any description of his microscopical methods (‘for reasons best known to himself’, said Hooke [author of *Micrographia*]; though science has hardly resolved the issue of intellectual property since then).”

Lane N. 2015

The unseen world: reflections on Leeuwenhoek (1677) ‘Concerning little animals’.

Phil. Trans. R. Soc. B370: 20140344.

<http://dx.doi.org/10.1098/rstb.2014.0344>



When van Leeuwenhoek died in 1723, the microbial world faded from view for nearly 200 years.

Opinion: open science methods enable us to work together to solve global problems.

Opinion: open science methods enable us to work together to solve global problems.

- Publish in open access journals (if that's not feasible, use pre-print servers to expand access to your work).
- Show your work! (Provide detailed accounts of your bioinformatics methods, such as all successful commands that you ran. Make your QIIME 2 data provenance accessible to your readers, e.g., load [QIIME 2 paper .qzv files from Supp. File 1](#) with [QIIME 2 View](#).)
- Deposit your data and standards compliant metadata in public archives ([“Available upon request”](#): not good enough for microbiome data!).
- Consider [pre-registering your study](#).
- Teach what you know (and what you want to know better!).
- Release your code under open source licenses, such as BSD. Release your educational content under open access licences, such as CC-BY.

What's next?

What's next?

Get involved in the QIIME 2 community!

Please complete the post-workshop survey.
Your suggestions can improve future
workshops.

All video content will be available on YouTube in the first quarter of 2021. Use these to continue your own learning, or take what you learned here back home and teach your own workshop!

Subscribe to the [QIIME 2 YouTube channel](#) for notifications about this and other content.

We are looking for translators of this video content (to add captions) as well as other QIIME 2 educational materials in as many languages as we can.

If you're interested, please complete this form to express your interest: <http://bit.ly/q2-translate>



(<https://library.qiime2.org>)

Primary distribution site for QIIME 2 plugins, and soon to be a lot more, including tutorials, plugin documentation, and install scripts.



Latest Plugins

q2-coremicrobiome 1.0

Qiime2 plugin of COREMIC: CORE
MICrobiome
[<https://doi.org/10.7717/peerj.4395>].
This plugin works with qza files
from QIIME 2.

R

gemelli 0.0.5

Gemelli is a toolbox for running tensor factorization on sparse compositional omics datasets. Gemelli performs unsupervised dimensionality reduction of spatiotemporal microbiome data. The output of gemelli helps to resolve spatiotemporal subject variation and the biological features that separate them.



RESCRIPT 2020.11

REference Sequence annotation and CuRation Pipeline RESCRIPt is a QIIME 2 plugin to support a variety of operations for managing and curating reference sequence databases, DNA/RNA sequence data, and taxonomic data.



Chan
Zuckerberg
Initiative 

[Project announcement \(via NAU News\)](#)

QIIME 2 plugins are publishable as “applications note”-style papers.



METHODS AND PROTOCOLS
Novel Systems Biology Techniques



q2-longitudinal: Longitudinal and Paired-Sample Analyses of Microbiome Data

✉ Nicholas A. Bokulich,^a Matthew R. Dillon,^a Yilong Zhang,^b Jai Ram Rideout,^a Evan Bolyen,^a Huilin Li,^c Paul S. Albert,^d
✉ J. Gregory Caporaso^{a,e}

F1000Research
Publish fast. Openly. Without restrictions.

Version 1. *F1000Res.* 2018; 7: 1418.

Published online 2018 Sep 6. doi: [\[10.12688/f1000research.15704.1\]](https://doi.org/10.12688/f1000research.15704.1)

PMCID: PMC6206612

PMID: 30416717

ITSxpress: Software to rapidly trim internally transcribed spacer sequences with quality scores for marker gene analysis

Adam R. Rivers, Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Software, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing,^{a,1} Kyle C. Weber, Software, Writing – Review & Editing,¹ Terrence G. Gardner, Resources, Writing – Review & Editing,² Shuang Liu, Resources,² and Shalamar D. Armstrong, Resources³

▼ Author information ▶ Article notes ▶ Copyright and License Information [Disclaimer](#)

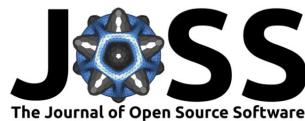
¹Genomics and Bioinformatics Research Unit, USDA Agricultural Research Service, Gainesville, FL, 32608, USA

²Department of Crop and Soil Sciences, North Carolina State University, Raleigh, NC, 27695, USA

³Department of Agronomy, Purdue University, Purdue, IN, 47907, USA

^aEmail: adam.rivers@ars.usda.gov

No competing interests were disclosed.



q2-sample-classifier: machine-learning tools for microbiome classification and regression

Nicholas A Bokulich¹, Matthew R Dillon¹, Evan Bolyen¹, Benjamin D Kaehler², Gavin A Huttley², and J Gregory Caporaso^{1, 3}

¹ The Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA 2 Research School of Biology, Australian National University, Canberra, Australia 3 Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ, USA

DOI: [10.21105/joss.00934](https://doi.org/10.21105/joss.00934)

The Caporaso Lab is currently hiring software engineers, post-doctoral scholars, and graduate students to join the QIIME 2 team!

Watch the [QIIME 2 Forum job board](#) and/or [@qiime2 on Twitter](#) for announcement of these positions (and post your own job listings for free).



QIIME 2 resources:

- User documentation: <https://docs.qiime2.org>
- Technical support (register for a free account if you don't already have one): <https://forum.qiime2.org>
- Follow us on Twitter for announcements: [@qiime2](https://twitter.com/qiime2)
- Developer resources: <https://dev.qiime2.org>
- Source code on GitHub: <https://github.com/qiime2>
- YouTube channel: <http://bit.ly/q2-youtube>

Other microbiome and bioinformatics educational content

mockrobiota: a Public Resource for Microbiome Bioinformatics Benchmarking

Nicholas A. Bokulich, Jai Ram Rideout, William G. Mercurio, Arron Shiffer, Benjamin Wolfe, Corinne F. Maurice, Rachel J. Dutton, Peter J. Turnbaugh, Rob Knight, J. Gregory Caporaso
Josh D. Neufeld, *Editor*

DOI: 10.1128/mSystems.00062-16

- Public repository of mock community datasets
 - Submit a Pull Request to add yours!
- Components:
 - Raw data
 - Sample metadata
 - Expected composition
- Current inventory:
 - 20 16S rRNA studies
 - 5 fungal ITS studies
 - 1 mock metagenome
 - Illumina HiSeq and MiSeq data
 - Even and staggered composition
 - Communities consist of 11-67 strains

An Introduction to Applied Bioinformatics

An Introduction to Applied Bioinformatics (or IAB) is a free, open source interactive text that introduces readers to core concepts of bioinformatics in the context of their implementation and application.



IAB is primarily being developed by Greg Caporaso (GitHub/Twitter: [@gregcaporaso](#)) in the Caporaso Lab at Northern Arizona University. You can find information on the courses I teach on my teaching website and information on my research and lab on [my lab website](#).

Download Local Copy

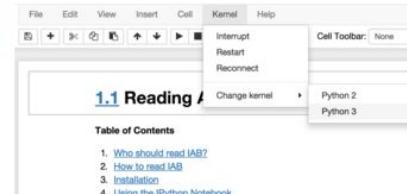
You can also download the latest copy of IAB as Jupyter Notebooks for your own PC.

Download

Read Interactively (recommended)

We're experimenting with Binder to simplify interactively reading IAB.

Note: On opening a notebook in Binder, you'll need to select the *Kernel* menu item, then *Change kernel*, and then *Python 3*. If you don't do this, you will get errors when you try to execute the notebooks. The screenshot to the right illustrates how to change to the Python 3 kernel.



[Read on Binder](#)

Read Statically (easiest)

Here you can find statically published copies of IAB with full output. There are several version to choose from:

- the latest version (you may encounter some instability, but will have the most recent content)
- version 0.1.1 (the most recent stable release)
- version 0.1.0 (the first version)

[Read Latest Online](#)



[readIAB.org](#)

Gut Check: Exploring Your Microbiome

by University of Colorado Boulder & University of Colorado System

i Course Info

UNIVERSITY OF COLORADO BOULDER & UNIVERSITY OF COLORADO SYSTEM

Gut Check: Exploring Your Microbiome

About this Course

Imagine if there were an organ in your body that weighed as much as your brain, that affected your health, your weight, and even your behavior. Wouldn't you want to know more about it? There is such an organ — the collection of microbes in and on your body, your human microbiome.

• Subtitles available in **English**

Log in to enroll in this course

Log In

Instructors**Professor Rob Knight**

Professor

Howard Hughes Medical Institute, and Department of Chemistry & Biochemistry and Computer Science, and BioFrontiers Institute

**Dr. Jessica Metcalf**Senior Research Associate
BioFrontiers Institute**Dr. Katherine Amato**Postdoctoral Research Associate
Department of Anthropology, BioFrontiers Institute

<https://www.coursera.org/learn/microbiome>

Mike Palmer's ordination website:

<http://ordination.okstate.edu/>

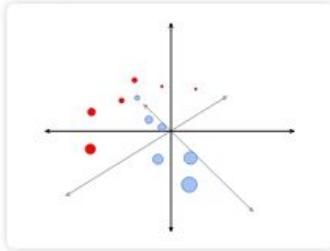
Ordination Methods for Ecologists	JUMP TO: OVERVIEW Ordination Topics Ordination Software Links Ordination Glossary Other Ordination Links Ordination Listserv OSU Botany	
<h2>Ordination Topics</h2>		
<p>Members of the Laboratory for Innovative Biodiversity Research And Analysis (LIBRA) are often available to engage in consulting activities for particular projects, or to offer short courses on ordination methods and the use of CANOCO. For more information, contact Mike Palmer at mike.palmer@okstate.edu.</p> <p>Ordination is a widely-used family of methods which attempts to reveal the relationships between ecological communities. For definitions, go HERE.</p> <p>This ordination web page is designed to address some of the most frequently asked questions about ordination. It is my intention to gear this page towards the student and the practitioner rather than the ordination specialist, so please contact me if the jargon is unintelligible!</p> <p>The ecological literature is filled with papers describing, contrasting, and modifying existing ordination techniques. Then why is an ordination web page needed? My main</p>	<p>General and Reference</p> <ul style="list-style-type: none">• Overview of ordination methods• A Glossary for terms used in Ordination• Milestones in the history of Ordination• Ordination terminology: some confusions• The ideal ordination method• Recommendations for ordination: a key• Suggested references for self-education• Hypothesis-driven and Exploratory Analysis• Ordination links <p>Statistics and Background</p>	<p>Indirect Gradient Analysis</p> <ul style="list-style-type: none">• Distance-based ordination methods• Eigenanalysis-based ordination methods• Principal Components Analysis• Correspondence Analysis• Detrended Correspondence Analysis <p>Direct Gradient Analysis</p>

Welcome to the GUide to STasitical Analysis in Microbial Ecology (GUSTA ME)!

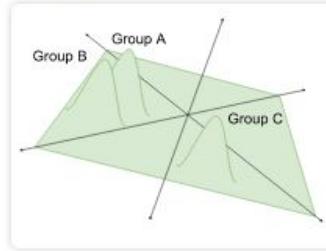
Where would you like to start?

You may start exploring the guide by browsing topics in the sidebar, using the search box at the top right of this page to find a particular method, or by clicking on one of the entry points below...

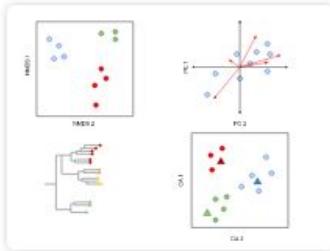
Explore data...



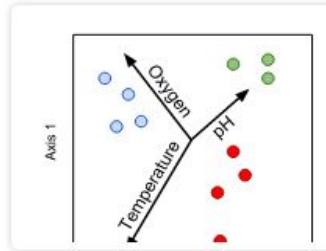
Test a hypothesis...



Browse visualisations...



Explore environmental influence...



License (read this if you're interested in re-using these slides)

These slides were created and arranged by members of the QIIME 1 and QIIME 2 development groups. This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>.

Feel free to use or modify these slides, but please credit the QIIME developers by placing the following attribution information where you feel that it makes sense:

This content was derived from the QIIME 2 educational materials. Visit <https://qiime2.org> to learn about QIIME 2.

Appendix

Special topic: Installing & Importing Data

Installing

<https://docs.qiime2.org/2019.4/install/>

Windows: Docker, Virtualbox, or maybe

<https://docs.microsoft.com/en-us/windows/wsl/install-win10>

Importing sequencing data (FASTQ)

If data is already demultiplexed, import as-is (skip demuxing).

sample-metadata.tsv

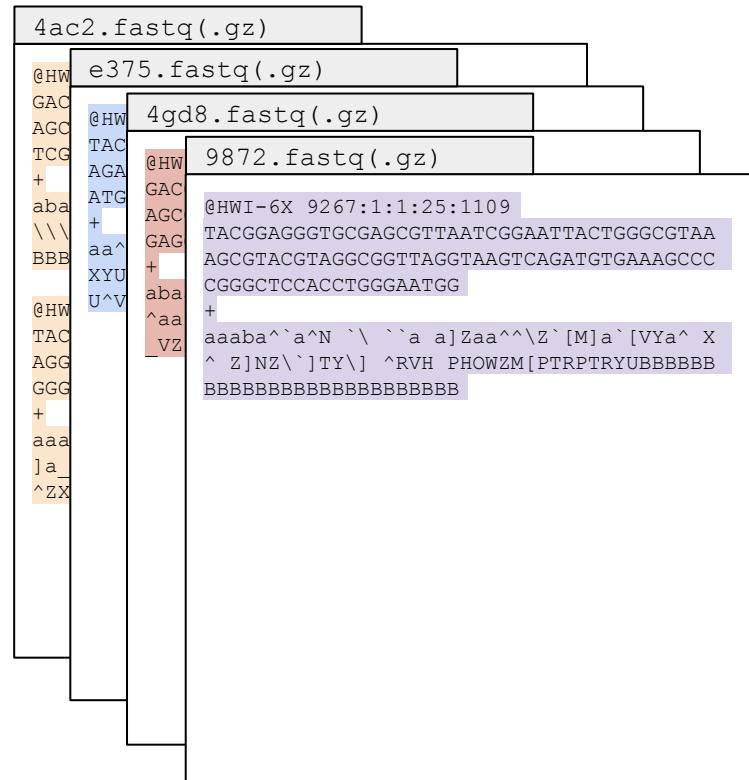
SampleID

4ac2

e375

4gd8

9872



Example: Demuxed

If data is already demultiplexed, import as-is (skip demuxing).

sample-metadata.tsv

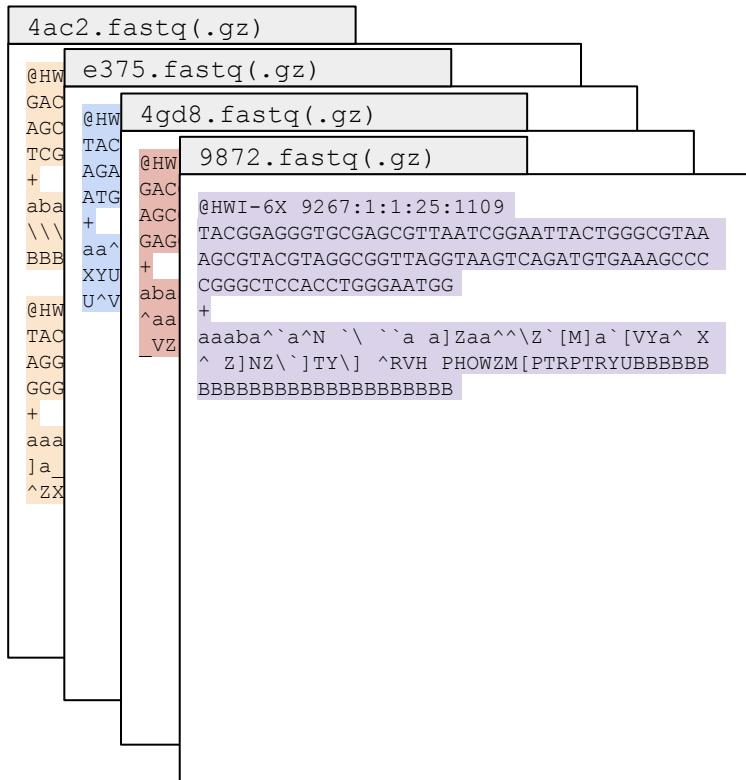
SampleID

4ac2

e375

4gd8

9872

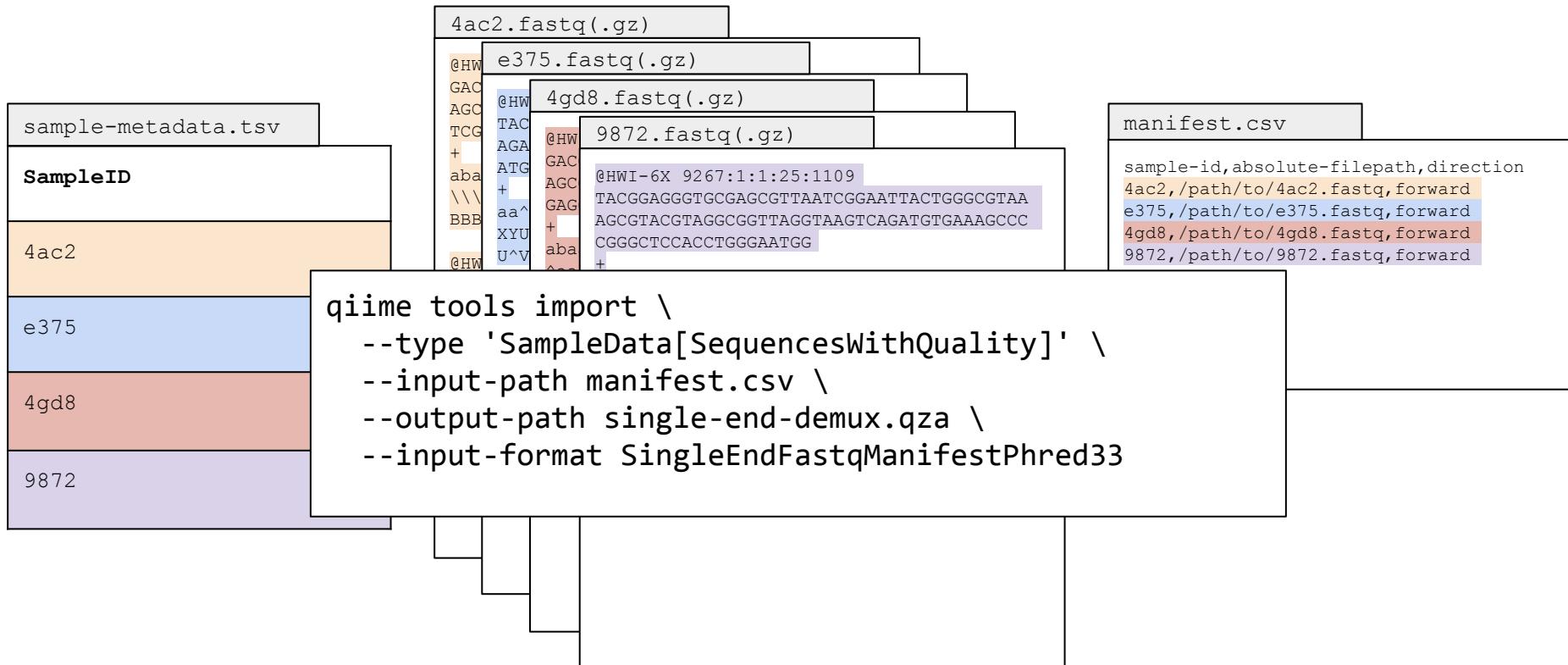


manifest.csv

sample-id	absolute-filepath	direction
4ac2	/path/to/4ac2.fastq	forward
e375	/path/to/e375.fastq	forward
4gd8	/path/to/4gd8.fastq	forward
9872	/path/to/9872.fastq	forward

Example: Demuxed

If data is already demultiplexed, import as-is (skip demuxing).



Example: Demuxed

If data is still multiplexed, assign sequence reads to samples (i.e., *demultiplex* or *demux*).

sample-metadata.tsv

SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

```
barcodes.fastq(.gz)          sequences.fastq(.gz)  
@HWI-6X_9267:1:1:25:1051  @HWI-6X_9267:1:1:25:1051  
AACGCAC                      GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAGCGCGCTAGGTG  
+                                GCTTGGTAAGTCATGGTGAAATCCCTCGGCTAACCGAGGAACTG  
Bbbbbbb                         +  
  
@HWI-6X_9267:1:1:25:1051  abaaaaaa^`a ]^` ``a` ``]]]^` ``a[VXGX``z \\ `` ``a^SYOZVV  
AAGAGAT                      GACGTAGGTG  
+                                BB  
bbbbb                         +  
  
@HWI-6X_9267:1:1:25:1051  TGGCTAGGTG  
AACGCAC                      TC  
+                                Q[X]UVXVN[  
bbbbb                         BB  
  
@HWI-6X_9267:1:1:25:1051  AGCGTAGACG  
ACAGCAG                      GG  
+                                \a`a] ``]z_  
bbbbb                         ET  
  
@HWI-6X_9267:1:1:25:1051  TCTGTAGGTG  
ACAGCTA                      CG  
+                                U[ ``[]YZ]  
bbbbb                         ZL  
  
@HWI-6X_9267:1:1:25:1109  @HWI-6X_9267:1:1:25:1109  
ACAGCTA                      TACGGAGGGTGCAGCGTTAACCGGAATTACTGGCGTAAAGCGCTAGGTAGGCG  
+                                GTTAGGTAAGTCAGATGTGAAAGCCCCGGGCTCACCTGGGAATGG  
bbbbb                         +  
  
aaaba^`a^N `` ``a a]Zaa^``Z` [M]a`[VY a^ X^ Z]NZ` ``]TY\]_  
^RVH_PHOWZM[PTRPTRYUBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB  
+  
bbbbb                         _
```

qiime tools import \
--type EMPSingleEndSequences \
--input-path emp-single-end-sequences \
--output-path seqs.qza

qiime demux emp-single \
--i-seqs seqs.qza \
--m-barcodes-file sample-metadata.tsv \
--m-barcodes-column BarcodeSequence \
--o-per-sample-sequences demux.qza

Example: EMP

If data is still multiplexed, assign sequence reads to samples (i.e., *demultiplex* or *demux*).

sample-metadata.tsv

SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

forward.fastq(.gz)

```
@HWI-6X 9267:1:1:25:1051
AACGCACGACGAAGGTGACGACCGTTGCTCGGAATC
+
Bbbbbbbbabaaaaaa`^`a ]^`\\``a`^`]]]^`^
@HWI-6X 9267:1:1:25:267
AAGAGATTACGTATGGGCAAGCGTTATCCGGATT
+
bbbbbbbbaa^^[ ^ ^ ^ ^[^^[^^__ZZ[^
@HWI-6X 9267:1:1:25:609
AACGCAC TACGTAGGGGCAAGCGTTATCCGGATT
+
bbbbbbbbaaab`aaa`aaaaaaaaaaaaaaaaaaa
@HWI-6X 9267:1:1:25:519
ACAGCAG GACGGAGGATGCAAGTGTATCCGGATT
+
bbbbbbbbaaaaaaaaaaaaaaaaaaaaaaa\aaaaaa
@HWI-6X 9267:1:1:25:1109
ACAGCTA TACGGAGGGTGCAGCGTTAACCGGAATT
+
bbbbbbbbaabaa^`a^N_`\\``a_a]Zaa^^\z`[
```

reverse.fastq(.gz)

```
@HWI-6X 9267:1:1:25:1051
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGC
+
abaaaaaa`^`a ]^`\\``a`^`]]]^`^a[VXGX
@HWI-6X 9267:1:1:25:267
TACGTATGGGCAAGCGTTATCCGGATTATTGGGC
+
aa^^[ ^ ^ ^ ^[^^[^^__ZZ[^[^[^]ZUZ]
@HWI-6X 9267:1:1:25:609
TACGTAGGGGCAAGCGTTATCCGGATTACTGGGT
+
aaab`aaa`aaaaaaaaaaaaaaaaaaaaaaaYQ
@HWI-6X 9267:1:1:25:519
GACGGAGGATGCAAGTGTATCCGGATTACTGGGC
+
abaaaaaaaaaaaaaaaaaaaaaa\aaaaaa``aa_aaa^z
@HWI-6X 9267:1:1:25:1109
TACGGAGGGTGCAGCGTTAACCGGAATTACTGGGC
+
aaaba^`a^N_`\\``a_a]Zaa^^\z`[M]a`[VY
```

Example: Cutadapt

If data is still multiplexed, assign sequence reads to samples (i.e., *demultiplex* or *demux*).

sample-metadata.tsv

SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

forward.fastq(.gz)

```
@HWI-6X 9267:1:1:25:1051  
AACGCACGACGAAGGTGACGACC GTT GCT CGGAATC  
+  
Bbbbbbbaaaaaaa`^`a ] ^\``\``a`^`]]]^`  
@HWI-6X 9267:1:1:25:267
```

AAGAG

+
bbbbb

@HWI-

AACG

+
bbbbb

@HWI-

ACAG

reverse.fastq(.gz)

```
@HWI-6X 9267:1:1:25:1051  
GACGAAGGTGACGACC GTT GCT CGGAATCACTGGGC  
+  
abaaaaaa`^`a ] ^\``\``a`^`]]]^`a[VXGX  
@HWI-6X 9267:1:1:25:267
```

```
qiime tools import \  
--type MultiplexedPairedEndBarcodeInSequence \  
--input-path my-reads/ \  
--output-path seqs.qza
```

```
qiime cutadapt demux-paired \  
--i-seqs seqs.qza \  
--m-forward-barcodes-file sample-metadata.tsv \  
--m-forward-barcodes-column BarcodeSequence \  
--p-error-rate 0 \  
--o-per-sample-sequences demux.qza \  
--o-untrimmed-sequences untrimmed.qza
```

Example: Cutadapt

Notice a pattern?

The goal is to get to a point where your sequences are demultiplexed (i.e. per-sample), that way you can begin QC (DADA2/deblur/OTU clustering/etc).

What about Ion Torrent?

- If reads are demuxed, use the manifest format.
- If reads are muxed, you can most likely make use of the q2-cutadapt -based import & demux approach (although reads are often mixed orientation it seems...).
- If reads are single-end, import using a strategy listed above, then proceed to:
 - DADA2, or
 - OTU-based methods
- If reads are paired-end, you will probably need to import as SampleData[JoinedSequencesWithQuality], since PGM paired-end reads are single-stranded. DADA2 is not an option for you, since it assumes reads aren't joined. Proceed to:
 - [OTU-based methods](#)
- Deblur is Illumina-specific, so no matter what style, IT PGM & Deblur do not mix.
- Stuck? Come visit us at <https://forum.qiime2.org> for more help!

That's nice, but what about
importing other types of data...

How do I learn about types and formats in QIIME 2?

```
qiime tools import --show-importable-formats
```

```
qiime tools import --show-importable-types
```

Let's import a BIOM file (Feature Table)

- Download this file
<https://data.qiime2.org/2018.11/tutorials/importing/feature-table-v100.biom>
- How do we know (or determine) what Semantic Type this is?
- Import
- Summarize

Let's import a Newick file (Phylogeny)

- Download this file
<https://data.qiime2.org/2018.11/tutorials/importing/unrooted-tree.tre>
- How do we know (or determine) what Semantic Type this is?
- Import
- Bonus: open in iTOL

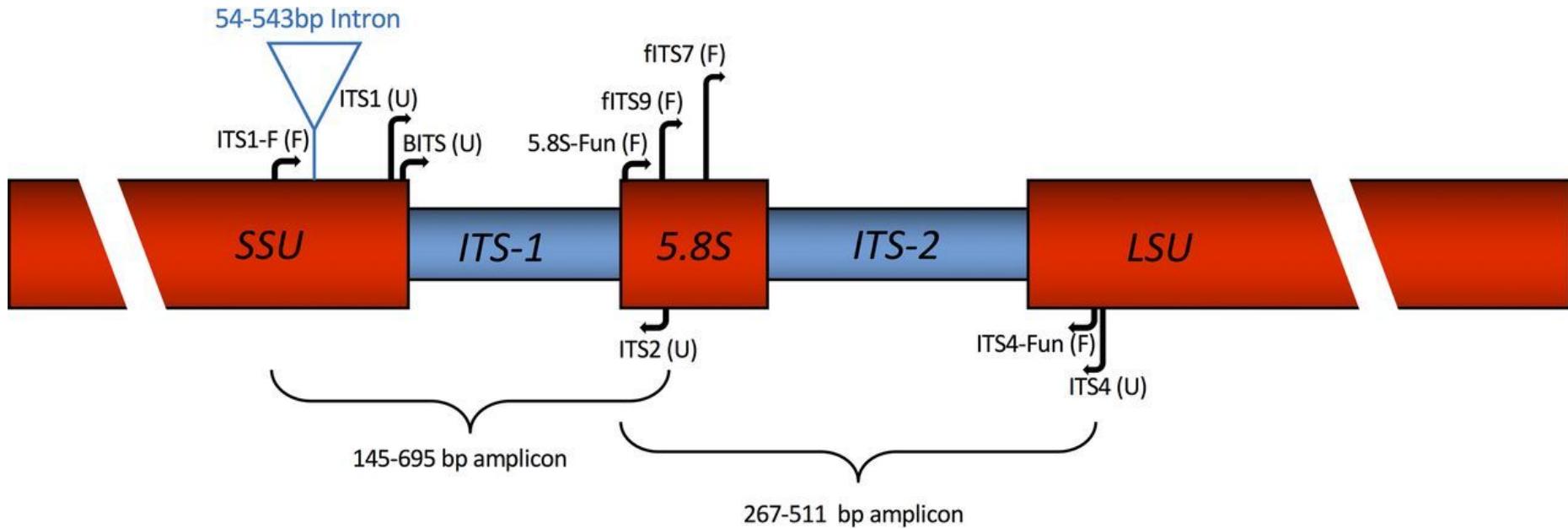
The Fungus Among Us

Instructor: Nick

Time: 930-1020

Slides: <http://bit.ly/q2faes-jan20-slides>
Schedule: <http://bit.ly/q2faes-jan20-schedule>

Fungal ITS



Fungal ITS: length variation

ITS locus	Forward primer	Reverse primer	Amplicon length distribution (mean ± SD)		
			Ascomycota	Basidiomycota	Non-Dikarya ^a
ITS1	BITS	B58S3	183.6 ± 46.8	219.8 ± 56.9	215.0 ± 95.9
	ITS1	ITS2	218.1 ± 54.4	253.3 ± 60.0	201.0 ± 71.8
	NSI1	58A2	357.2 ± 128.9	386.5 ± 105.7	269.1 ± 24.1
	ITS1F_KYO1	ITS2_KYO1	275.3 ± 103.2	285.3 ± 50.1	200.3 ± 54.2
	ITS1F_KYO2	ITS2_KYO2	270.6 ± 90.5	284.5 ± 42.1	216.4 ± 94.4
ITS2	58A1	NLB4	478.8 ± 23.9	528.1 ± 30.3	
	58A2	NLB4	476.8 ± 23.9	525.9 ± 30.4	
	ITS3F_KYO1	ITS4_KYO1	310.5 ± 29.9	362.2 ± 35.3	376.0 ± 57.4
	fITS7f	ITS4	258.5 ± 27.3	309.8 ± 35.6	312.7 ± 47.2
	gITS7f	ITS4	259.9 ± 22.5	307.6 ± 34.7	312.9 ± 47.4
	fITS9f	ITS4	324.4 ± 11.7	354.6 ± 32.9	
Whole ITS	BITS	ITS4	535.8 ± 81.8	618.0 ± 72.6	573.0 ± 132.4
	ITS1	ITS4	547.2 ± 93.0	624.8 ± 79.7	582.4 ± 132.0
	NSI1	NLB4	1,066.6 ± 313.9	927.3 ± 182.6	
	ITS1F_KYO1	ITS4_KYO1	612.4 ± 131.1	664.9 ± 57.8	589.8 ± 155.6

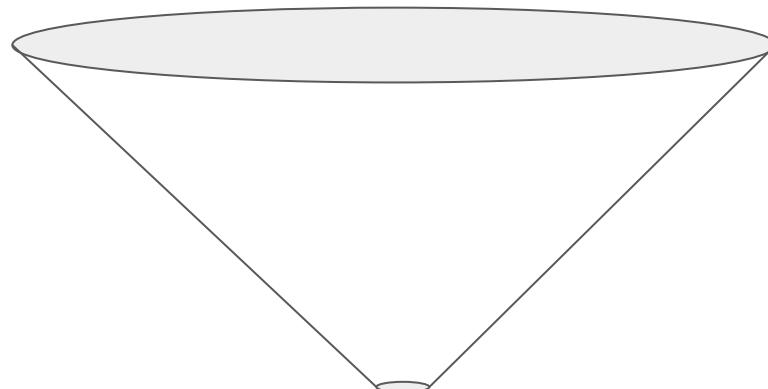
^a Missing values are due to noncoverage of non-Dikarya by that primer.

Fungal ITS Analysis Tutorial

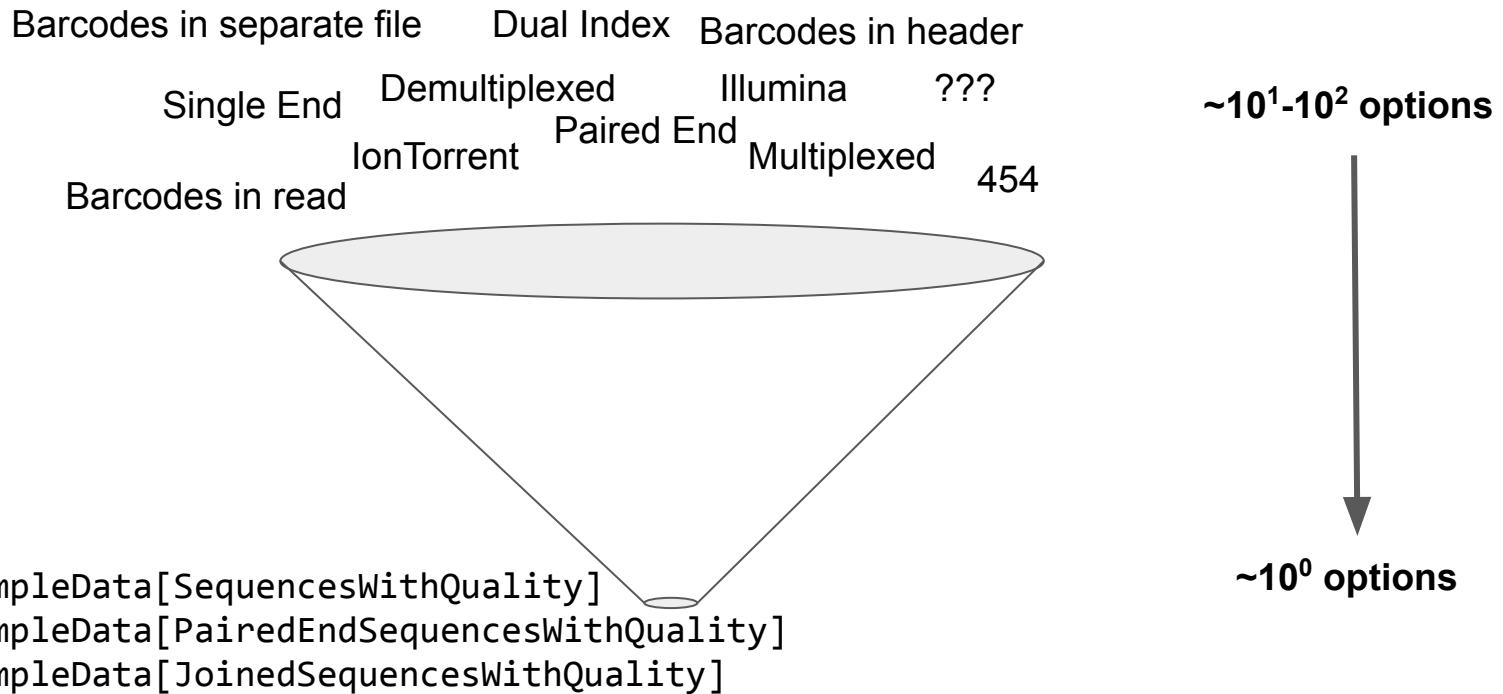
<https://forum.qiime2.org/t/fungal-its-analysis-tutorial/7351>

Opinion: Getting sequence data into QIIME 2 is the hardest part of QIIME 2 (or any software package)

Barcodes in separate file	Dual Index	Barcodes in header
Single End	Demultiplexed	Illumina
	Paired End	??? Multiplexed
IonTorrent		
Barcodes in read		454



Opinion: Getting sequence data into QIIME 2 is the hardest part of QIIME 2 (or any software package)



Opinion: Getting sequence data into QIIME 2 is the hardest part of QIIME 2 (or any software package)

Once your data is in one of these formats, your analysis in QIIME 2 can begin!

`SampleData[SequencesWithQuality]`
`SampleData[PairedEndSequencesWithQuality]`
`SampleData[JoinedSequencesWithQuality]`

Importing sequencing data (FASTQ)

If data is already demultiplexed, import as-is (skip demuxing).

sample-metadata.tsv

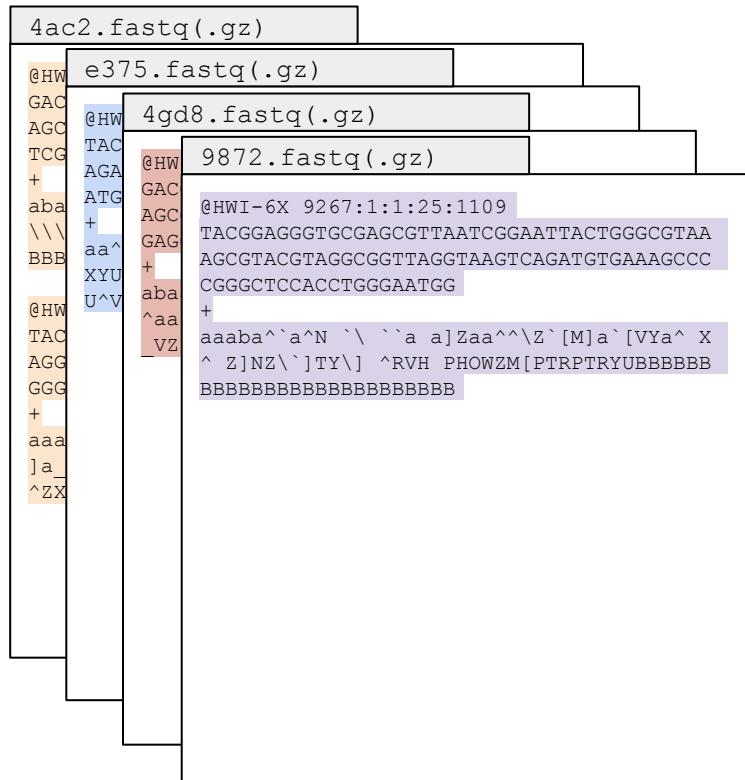
SampleID

4ac2

e375

4gd8

9872



Example: Demuxed

If data is already demultiplexed, import as-is (skip demuxing).

sample-metadata.tsv

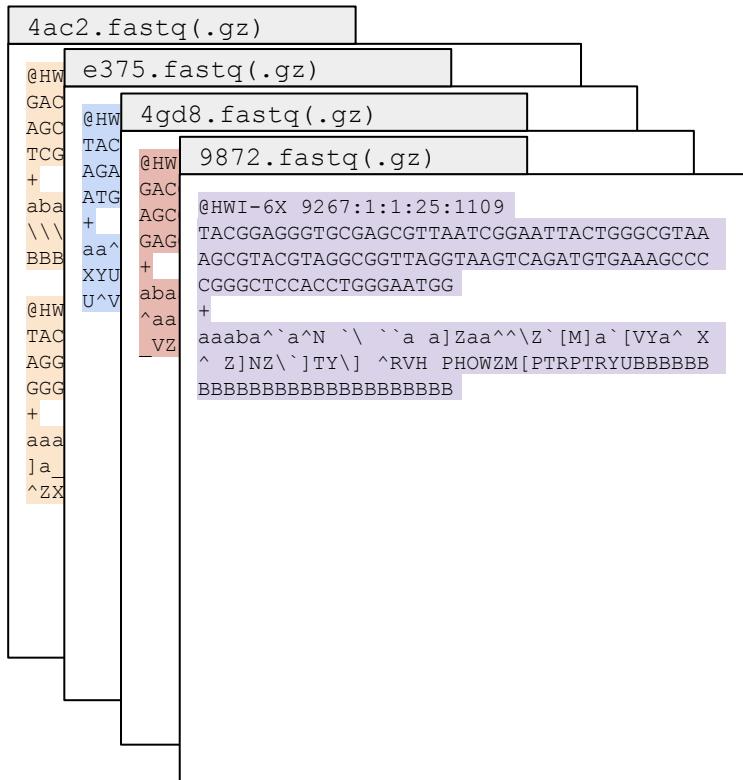
SampleID

4ac2

e375

4gd8

9872

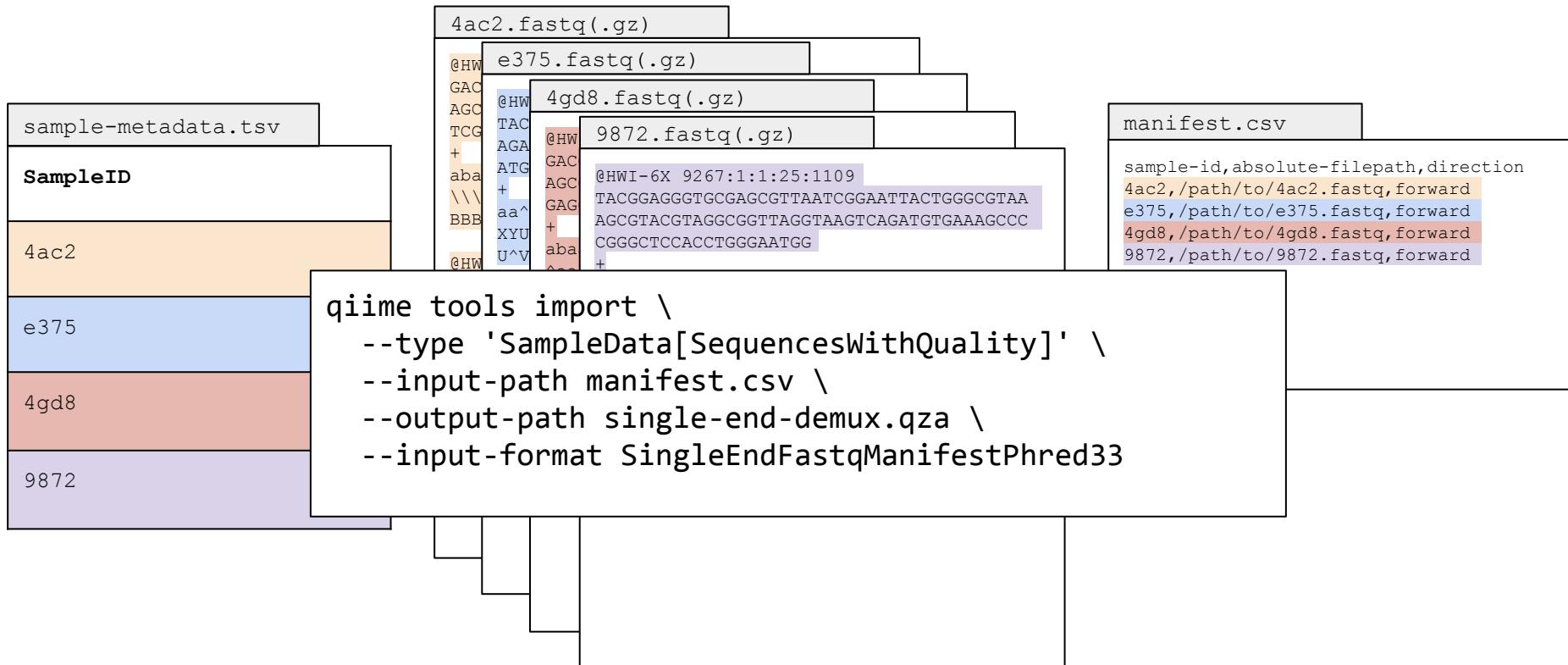


manifest.csv

sample-id	absolute-filepath	direction
4ac2	/path/to/4ac2.fastq	forward
e375	/path/to/e375.fastq	forward
4gd8	/path/to/4gd8.fastq	forward
9872	/path/to/9872.fastq	forward

Example: Demuxed

If data is already demultiplexed, import as-is (skip demuxing).



Example: Demuxed

If data is still multiplexed, assign sequence reads to samples
(i.e., *demultiplex* or *demux*).

sample-metadata.tsv

SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

barcodes.fastq(.gz)

@HWI-6X_9267:1:1:25:1051

AACGCAC

+

Bbbbbbb

@HWI-6X_9267:1:1:25:267

AAGAGAT

+

bbbbbbb

@HWI-6X_9267:1:1:25:609

AACGCAC

+

Bbbbbbb

@HWI-6X_9267:1:1:25:519

ACAGCAG

+

bbbbbbb

@HWI-6X_9267:1:1:25:1109

ACAGCTA

+

bbbbbbb

sequences.fastq(.gz)

@HWI-6X_9267:1:1:25:1051

GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAGCGCGCGTAGGTG
GCTTGGTAAGTCATGGTGAAATCCCTCGGCTAACCGAGGAAC

+

abaaaaaa^`a]^` ``a ``]]]^` ``a [VXGX ``Z \\ \\ ^` a ^SYOZVV
SVGYVDXOZVT\ TITBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

@HWI-6X_9267:1:1:25:267

TACGTATGGGCAGCGTTATCGGAATTATTGGCGTAAAGAGTGCAGTAGGTG
GTGGCTTAAGCGCAGGGTTAACCGCAATGGCTTAACATTGTTCTC

+

aa^`[^` `` ^ [^` `` ZZ [^` `` ZUZ] WUZXYU[Q[X]UVXVN[
XUWWURZUY] XXRZRNVRTNTWUUU^VJVOMIHQU\URRN[BBB

@HWI-6X_9267:1:1:25:609

TACGTAGGGGCAAGCGTTATCGGAATTACTGGGTGTAAAGGGAGCGTAGACG
GATGGACAAGTCTGTGAAAGGCTGGGCTAACCCGGGACGG

+

aaab`aaa`aaaaaaaaaaaaaaaaaaaaaaYQ^` ``]a]\a `a ``]Z_
[] [I^` aZ^WW^ ``ZZ T] XY^` ``ZX\ZJS[W[V^` HOVYTET

@HWI-6X_9267:1:1:25:519

GACGGAGGATGCAAGTGTATCGGAATCACTGGCGTAAAGCGTCTGTAGGTG
GTTACTAAGTCAACTGTTAACCTGAGGCTAACCTCGAAATCG

+

abaaaaaaaaaaaaaa\aaaaaaaa``aa aaaa^Z [ZY^aa`U[``]YZ]
WY] Z XX\\ ``] `` [\XTVX] ``T_VZ[``]ZXVXYFX_ VYJWWZL

@HWI-6X_9267:1:1:25:1109

TACGGAGGTTGCGAGCGTTATCGGAATTACTGGCGTAAAGCGTACGTAGGCG
GTTAGGTTAAGTCAGATGTGAAAGCCCCGGGCTCACCTGGGAATGG

+

aaaba^`a^N `` ``a a]Zaa^` `` [M] a `` [VY a^ X^ Z] NZ ``] TY\]
^RVH_PHOWZM[PTRPTRYUBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

Example: EMP

If data is still multiplexed, assign sequence reads to samples (i.e., *demultiplex* or *demux*).

sample-metadata.tsv

SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

```
barcodes.fastq(.gz)          sequences.fastq(.gz)  
@HWI-6X_9267:1:1:25:1051  @HWI-6X_9267:1:1:25:1051  
AACGCAC                      GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAGCGCGCTAGGTG  
+                                GCTTGGTAAGTCATGGTGAAATCCCTCGGCTAACCGAGGAACTG  
Bbbbbbb                         +  
  
@HWI-6X_9267:1:1:25:1051  abaaaaaa^`a ]^` ``a` ``]]]^` ``a[VXGX``z ``\\ ``^`a^SYOZVV  
AAGAGAT                      GACGTAGGTG  
+                                BB  
bbbbb                         +  
  
@HWI-6X_9267:1:1:25:1051  TGGCTAGGTG  
AACGCAC                      TC  
+                                Q[X]UVXVN[  
bbbbb                         BB  
  
@HWI-6X_9267:1:1:25:1051  AGCGTAGACG  
ACAGCAG                      GG  
+                                \a`a] ``]z_  
bbbbb                         ET  
  
@HWI-6X_9267:1:1:25:1051  TCTGTAGGTG  
ACAGCTA                      CG  
+                                U[ ``[]YZ]  
bbbbb                         ZL  
  
@HWI-6X_9267:1:1:25:1109  @HWI-6X_9267:1:1:25:1109  
ACAGCTA                      TACGGAGGGTGCAGCGTTAACCGGAATTACTGGCGTAAAGCGCTAGGTAGGCG  
+                                GTTAGGTAAGTCAGATGTGAAAGCCCCGGGCTCACCTGGGAATGG  
bbbbb                         +  
  
aaaba^`a^N `` ` ``a a]Zaa^``Z` [M]a`[VY a^ X^ Z]NZ` ``]TY\]_  
^RVH_PHOWZM[PTRPTRYUBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB  
+  
bbbbb                         _
```

qiime tools import \
--type EMPSingleEndSequences \
--input-path emp-single-end-sequences \
--output-path seqs.qza

qiime demux emp-single \
--i-seqs seqs.qza \
--m-barcodes-file sample-metadata.tsv \
--m-barcodes-column BarcodeSequence \
--o-per-sample-sequences demux.qza

Example: EMP

If data is still multiplexed, assign sequence reads to samples (i.e., *demultiplex* or *demux*).

sample-metadata.tsv

SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

forward.fastq(.gz)

```
@HWI-6X 9267:1:1:25:1051
AACGCACGACGAAGGTGACGACCGTTGCTCGGAATC
+
Bbbbbbbbabaaaaaa`^`a ]^`\\``a`^`]]]^`^
@HWI-6X 9267:1:1:25:267
AAGAGATTACGTATGGGCAAGCGTTATCCGGATT
+
bbbbbbbbaa^^[ ^^^ ^ ^[^^[^^ ____ ZZ[^
@HWI-6X 9267:1:1:25:609
AACGCACTACGTAGGGGCAAGCGTTATCCGGATT
+
bbbbbbbbaaab`aaa`aaaaaaaaaaaaaaaaaaa
@HWI-6X 9267:1:1:25:519
ACAGCAGGACGGAGGATGCAAGTGTATCCGGATT
+
bbbbbbbbaaaaaaa`aaaaaa\aaaaaaaa``aa_
@HWI-6X 9267:1:1:25:1109
ACAGCTATACGGAGGGTGCAGCGTTAACCGGAATT
+
bbbbbbbbaaabaa^`a^N_`\\``a_a]Zaa^`^\\z`[
```

reverse.fastq(.gz)

```
@HWI-6X 9267:1:1:25:1051
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGC
+
abaaaaaa`^`a ]^`\\``a`^`]]]^`^a[VXGX
@HWI-6X 9267:1:1:25:267
TACGTATGGGCAAGCGTTATCCGGATTATTGGC
+
aa^^[ ^^^ ^ ^[^^[^^ ____ ZZ[^[^[]ZUZ]
@HWI-6X 9267:1:1:25:609
TACGTAGGGGCAAGCGTTATCCGGATTACTGGGT
+
aaab`aaa`aaaaaaaaaaaaaaaaaaa^aaaaaaaaYQ
@HWI-6X 9267:1:1:25:519
GACGGAGGATGCAAGTGTATCCGGATTACTGGC
+
abaaaaaa`aaaaaa\aaaaaaaa``aa_aaaa^z_
@HWI-6X 9267:1:1:25:1109
TACGGAGGGTGCAGCGTTAACCGGAATTACTGGC
+
aaaba^`a^N_`\\``a_a]Zaa^`^\\z`[M]a`[VY
```

Example: Cutadapt

If data is still multiplexed, assign sequence reads to samples (i.e., *demultiplex* or *demux*).

sample-metadata.tsv

SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

forward.fastq(.gz)

```
@HWI-6X 9267:1:1:25:1051  
AACGCACGACGAAGGTGACGACC GTT GCT CGGAATC  
+  
Bbbbbbbaaaaaaa`^`a ] ^\``\``a`^`]]]^`  
@HWI-6X 9267:1:1:25:267
```

AAGAG

+
bbbbb

@HWI-

AACGC

+
bbbbb

@HWI-

ACAGC

+
bbbbb

@HWI-

ACAGC

+
bbbbb

@HWI-

ACAGC

+
bbbbb

@HWI-

ACAGC

+
bbbbb

@HWI-

ACAGCTA

+
bbbbb

```
qiime tools import \  
--type MultiplexedPairedEndBarcodeInSequence \  
--input-path my-reads/ \  
--output-path seqs.qza
```

```
qiime cutadapt demux-paired \  
--i-seqs seqs.qza \  
--m-forward-barcodes-file sample-metadata.tsv \  
--m-forward-barcodes-column BarcodeSequence \  
--p-error-rate 0 \  
--o-per-sample-sequences demux.qza \  
--o-untrimmed-sequences untrimmed.qza
```

reverse.fastq(.gz)

```
@HWI-6X 9267:1:1:25:1051  
GACGAAGGTGACGACC GTT GCT CGGAATC A C T G G G C  
+  
abaaaaaa`^`a ] ^\``\``a`^`]]]^`a [V X G X  
@HWI-6X 9267:1:1:25:267
```

Example: Cutadapt

FeatureTable[Frequency]

```
qiime tools import \  
  --input-path feature-table-v100.biom \  
  --type 'FeatureTable[Frequency]' \  
  --input-format BIOMV100Format \  
  --output-path feature-table-1.qza
```

Phylogeny[Unrooted]

```
qiime tools import \  
  --input-path unrooted-tree.tre \  
  --output-path unrooted-tree.qza \  
  --type 'Phylogeny[Unrooted]'
```

Getting Data Out of QIIME 2

```
$ qiime tools export \  
>   --input-path table.qza \  
>   --output-path exported_table_dir  
Exported table.qza as BIOMV210DirFmt to directory exported_table_dir
```

Export to multiple formats

```
❸ qiime tools export \  
  --input-path table.qza \  
  --output-format BIOMV100Format \  
  --output-path ./table-v100.biom  
Exported table.qza as BIOMV100Format to file ./table-v100.biom  
(version 3.0.10)
```