

# **Microbiome Analysis with QIIME2**





# Table of Contents

---

## Home

- 
- Course Overview 8

---

## QIIME2 Introduction

- 
- Lesson 1: Toward fully reproducible microbiome multi-omics bioinformatics with QIIME 2 10
  - Presentation description 10
  - Meet our presenter 10
  - Slides 10
  - Link to the recording 11

---

## Metadata and importing

- 
- Lesson 2: Getting Started with QIIME2 12
  - Lesson Objectives 12
  - DNAlexus 12
  - Getting started with DNAlexus 12
  - What is QIIME 2? 15
  - Multiple ways to use QIIME2 15
  - Some useful linux commands 15
  - QIIME2 Installation 16
  - Using QIIME2 on Biowulf 16

---

● What is amplicon sequencing?	16
● Things to know about 16S rRNA:	17
● Other target genes of interest:	17
● The data used in this tutorial	17
● Getting started with QIIME2	18
● QIIME2 Artifacts	18
● Qiime2view	19
● Metadata formatting	19
● Examining the metadata	20
● Data import	20
● Importing raw fastq files	20
● Importing our example data	21
● Summary of imported data	22
● Import ASV table	22
● Provenance Tracking	23
● QIIME2 forum	23

---

## Generating a feature table / feature data

---

● Lesson 3: Creating a feature table	24
● Lesson Objectives	24
● Primer trimming	24
● Using qiime cutadapt trim-paired	24
● OTU Clustering vs Denoising	25
● What do we mean by OTU Clustering / OTU picking?	25
● Methods for OTU clustering include	26
● Clustering methods on QIIME 2	27

---

● Denoising	27
● Denoising methods on QIIME2	28
● Preparing to Denoise (Hands on)	28
● Denoising stats	30
● Feature table and feature data summary information	30
● References	32

---

## Filtering, taxonomy, and phylogeny

● Lesson 4: Feature table filtering, taxonomic classification, and phylogeny	33
● Learning objectives	33
● Filtering	33
● Methods of filtering	33
● Metadata based filtering	34
● Feature filtering	34
● Taxonomy	35
● Alignment based taxonomy consensus classifiers:	35
● Machine learning approach	36
● Taxonomic based filtering	38
● Visualizing our taxonomy	38
● Phylogeny	39

---

## Microbial diversity, alpha rarefaction, alpha diversity

● Lesson 5: Microbial diversity, alpha rarefaction, alpha diversity	41
● Learning Objectives	41
● What is alpha diversity?	41

---

● Beta diversity (More on this in Lesson 6)	43
● Rarefaction	43
● What is rarefaction?	44
● Selecting a read depth to rarefy	44
● Core metrics phylogenetic	46
● Alpha diversity comparison	47
● Q2-longitudinal	48
● Optional filtering of samples	49

---

## Beta Diversity

● Lesson 6 .	50
● Learning Objectives	50
● Beta diversity	50
● Distance and dissimilarity metrics	51
● Beta rarefaction	52
● Ordination methods	52
● Generating a PCoA and UMAP in QIIME2	53
● PCoA	53
● UMAP	53
● Statistics	54

---

## Course Wrap-Up

● Lesson 7: Course Wrap-Up	56
● Learning Objectives	56
● QIIME 2 on Biowulf	56

---

● Review	57
● What have we done?	57
● Other plugins of interest	58
● Differential abundance testing	58
● Methods in QIIME 2	58
● ANCOM example	58
● Core microbiome	59
● Random forest regression and classification	59
● Random Forest example	60
● Other notable plugins	60
● Exporting results	61
● Working in R	61
● Struggling with command line?	61

---

## Practice

<b>Lesson 2: Data Import</b>	<b>63</b>
● Practice Lesson 2	63
● Download the sequences and import for further processing with the QIIME2 platform.	
● Step 1: Get the run info from the SRA	63
● Step 2: Download the data	64
● Step 3: Create the manifest	65
● Step 4: Import	65
● Summarize import	66
<b>Lesson 3: Denoising</b>	<b>68</b>
● Practice Lesson 3	68

Lesson 4: Filtering, classification, phylogeny	71
● Practice Lesson 4	71
Lesson 5: Alpha diversity	74
● Lesson 5 Practice	74
Lesson 6: Beta diversity	77
● Practice Lesson 6	77

---

## Getting the Data

● Getting the Data	79
--------------------	----

---

## References

References	81
------------	----

● References for the main content	81
-----------------------------------	----

---

## Additional Resources

Further readings and tutorials	83
--------------------------------	----

● Additional Resources	83
● The QIIME 2 docs and forum	83
● Related readings	83
● Linux help	83
● Other microbiome analysis platforms / tools	83



# Course Overview

This course was designed to teach the basics of targeted amplicon data processing and analysis using the QIIME2 (<https://qiime2.org/>) platform. Attendees will learn how to format data and metadata, import data, demultiplex sequences, trim sequences, denoise and classify sequences, and conduct basic analyses including measures of alpha and betadiversity.

Content from this course was inspired by the QIIME 2 Cancer Microbiome Intervention Tutorial (<https://docs.qiime2.org/jupyterbooks/cancer-microbiome-intervention-tutorial/index.html>) provided by the QIIME2 developers. While much of the code used from that tutorial will be integrated into these lessons, this course does not seek to replicate tutorial materials already available on the QIIME2 website (See the QIIME2 youtube playlist (<https://www.youtube.com/playlist?list=PLbVDKwGpb3XmvnTrU40zHRT7NZWWVNUp>) for detailed lessons).

This course will include seven 1 - 1.25 hour lectures followed by a 45 minute optional practice session. Lessons will be on Mondays and Wednesdays from 1 - 2:15 pm.

Lesson topics:

**Lesson 1:** Toward fully reproducible microbiome multi-omics bioinformatics with QIIME 2 (Oct 19th)

*Lesson 1 will not include a hands on component, but rather will include an introduction to QIIME2 by guest speaker, Greg Caporaso, a leading developer of the QIIME2 platform.*

Recording link: <https://cbiit.webex.com/cbiit/ldr.php?RCID=c44af836ce9dd1631b58ccc285880a56>

**Lesson 2:** Preparing the data, data import, and demultiplexing (Oct 24th)

Recording link: <https://cbiit.webex.com/cbiit/ldr.php?RCID=d7fca7b1d6b3441c445770782762c712>

**Lesson 3:** Trimming, read joining and quality filtering, OTU clustering / denoising (Oct 26th)

Recording link: <https://cbiit.webex.com/cbiit/ldr.php?RCID=354bc1fb92d6288a0c00ed0a7a1c777>

**Lesson 4:** Taxonomic classification, phylogeny, feature table filtering (Oct 31st)

Recording link: [https://cbiit.webex.com/cbiit/ldr.php?  
RCID=b6ca607256a4e8c461487b40362fc51c](https://cbiit.webex.com/cbiit/ldr.php?RCID=b6ca607256a4e8c461487b40362fc51c)

**Lesson 5:** Alpha diversity (Nov 2nd - CANCELLED). Lesson taught in combination with Lesson 6.

**Lesson 6:** Beta diversity (Nov 7th)

Recording link: [https://cbiit.webex.com/cbiit/ldr.php?  
RCID=b667466d9ae5f273cf273b6ecc7f93d1](https://cbiit.webex.com/cbiit/ldr.php?RCID=b667466d9ae5f273cf273b6ecc7f93d1)

**Lesson 7:** Course Wrap-up (Additional analyses to consider, feature table export, qiime2R import) (Nov 9th)

Recording link: [https://cbiit.webex.com/cbiit/ldr.php?  
RCID=4d90e55b0adaebe54bff0ed5030783f3](https://cbiit.webex.com/cbiit/ldr.php?RCID=4d90e55b0adaebe54bff0ed5030783f3).

---

Course requirements:

Who can take this course?

There are no prerequisites to take this course. However, learners should have basic unix skills (e.g., know how to navigate directories, copy, move, and download files from the web). This course is open to NCI researchers interested in using the QIIME2 platform to process and analyze microbiome data.

What materials are needed to take this course?

To participate in this course, you will need a computer, a reliable internet connection, and a web browser. All classes and help sessions will be held virtually through Webex. This class will be taught using the GOLD learning environment on the DNAexus platform. Learners will need to sign up for a DNAexus account and send their user name to ncibtep@nih.gov.

# Lesson 1: Toward fully reproducible microbiome multi-omics bioinformatics with QIIME 2

Lesson 1 does not include a hands on component, but rather includes an introduction to QIIME2 by guest speaker, Dr. Greg Caporaso, a leading developer of the QIIME2 platform.

## Presentation description

The QIIME platform, including QIIME 1 and QIIME 2, has been extensively applied in microbiome research, repeatedly making analyses that were challenging or impossible into routine tasks. While QIIME began as a marker gene (e.g., 16S, ITS, ...) analysis platform, microbiome research is transitioning toward multi-omics. With funding from NCI's Informatics Technology for Cancer Research program, QIIME 2 is transitioning to become a microbiome multi-omics analysis platform. In this talk, Dr. Caporaso will introduce QIIME 2, including current work on expanding beyond marker gene analysis. He will also discuss QIIME 2's retrospective data provenance tracking system, and how it can help you to get help with your bioinformatics analyses and ensure that your work is reproducible. Dr. Caporaso will describe the ways that QIIME 2 can be used, including through the Galaxy graphical user interface, a command line interface, or a Python 3 API. Full support for using QIIME 2 through these different interface types ensures that using QIIME 2 will be accessible and convenient for you, regardless of your computational background. Finally, He will present on QIIME 2's extensive educational and technical support resources, so that you can start learning QIIME 2 as quickly as possible.

## Meet our presenter

Greg Caporaso, PhD

- Professor at Northern Arizona University
- A microbiome expert with 100+ related publications
- Lead developer of the QIIME 2 Platform
- Visit his lab website at <https://caporasolab.us>

## Slides

Lecture slides can be found [here](https://bit.ly/3CMrNfl) (<https://bit.ly/3CMrNfl>).

## Link to the recording

The recording can be accessed here (<https://cbiit.webex.com/cbiit/lxr.php?RCID=c44af836ce9dd1631b58ccc285880a56>).

# Lesson 2: Getting Started with QIIME2

## Lesson Objectives

- Obtain sequence data and sample metadata
- Import data and metadata
- Discuss other useful QIIME2 features including view QIIME2, provenance tracking, and the QIIME2 forum.

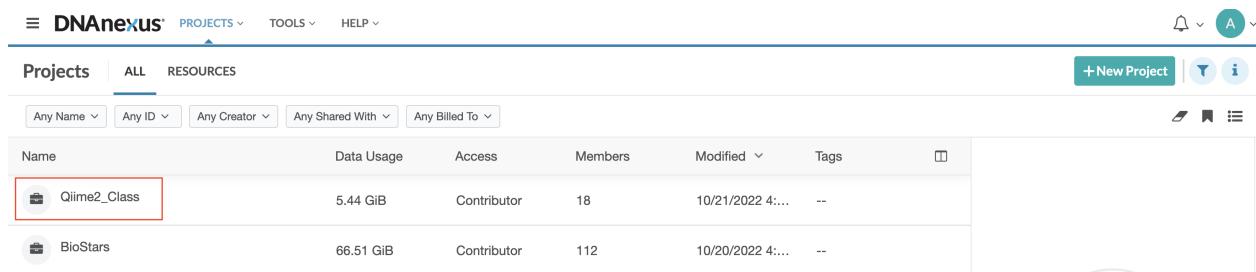
## DNAAnexus

DNAAnexus provides a secure cloud based platform for the analysis and sharing of next generation sequencing data. This class will use a pre-built teaching environment, the GOLD platform, which includes all of the software needed installed and ready to go.

### Getting started with DNAAnexus

#### Step 1: Login to DNAAnexus

**Step 2:** Once you login, you should see the *Projects* page. If you have used DNAAnexus previously, you may see more than one project listed. If this is your first time using DNAAnexus, you will only see the project name for this course listed, **Qiime2\_Class**. Double click on Qiime2\_Class.



The screenshot shows the DNAAnexus web interface. At the top, there is a navigation bar with 'PROJECTS' selected. Below the navigation is a search bar with dropdowns for 'Any Name', 'Any ID', 'Any Creator', 'Any Shared With', and 'Any Billed To'. A 'New Project' button is located in the top right corner. The main area is titled 'Projects' and contains a table with two rows. The first row has a red border and lists 'Qiime2\_Class' as the project name, '5.44 GiB' as data usage, 'Contributor' as access, '18' as members, '10/21/2022 4...' as modified date, and '--' as tags. The second row lists 'BioStars' as the project name, '66.51 GiB' as data usage, 'Contributor' as access, '112' as members, '10/20/2022 4...' as modified date, and '--' as tags. There are also icons for edit, delete, and details for each project row.

Name	Data Usage	Access	Members	Modified	Tags
Qiime2_Class	5.44 GiB	Contributor	18	10/21/2022 4...	--
BioStars	66.51 GiB	Contributor	112	10/20/2022 4...	--

**Step 3:** Once you double click on the Qiime2\_Class project, you will see a project directory containing multiple subdirectories and files. Select (double click) on **Microbiome\_Class.html**.

The screenshot shows the DNAnexus web interface. At the top, there's a navigation bar with 'PROJECTS', 'TOOLS', and 'HELP'. Below it, the project name 'Qiime2\_Class' is selected. The main area has tabs for 'SETTINGS', 'MANAGE' (which is active), 'MONITOR' (with a notification count of 1), and 'VISUALIZE'. On the left, a sidebar shows a hierarchical tree of project contents: 'Qiime2\_Class' (expanded) contains 'Applications', 'Params', 'Sessions', and 'TEST'. On the right, a list of files is shown under 'All Projects > Qiime2\_Class'. The files listed are 'Applications' (Folder), 'Params' (Folder), 'Sessions' (Folder), 'TEST' (Folder), and 'Microbiome\_Class.html' (File). The 'Microbiome\_Class.html' file is highlighted with a red box.

**Step 4:** The *Microbiome\_Class.html* file will open the GOLD platform application, and you will see a screen that looks like this:

Welcome to GOLD, an online learning platform presented by BTEP  
Use the links below to a) login to your account and b) view files in the [public] folder

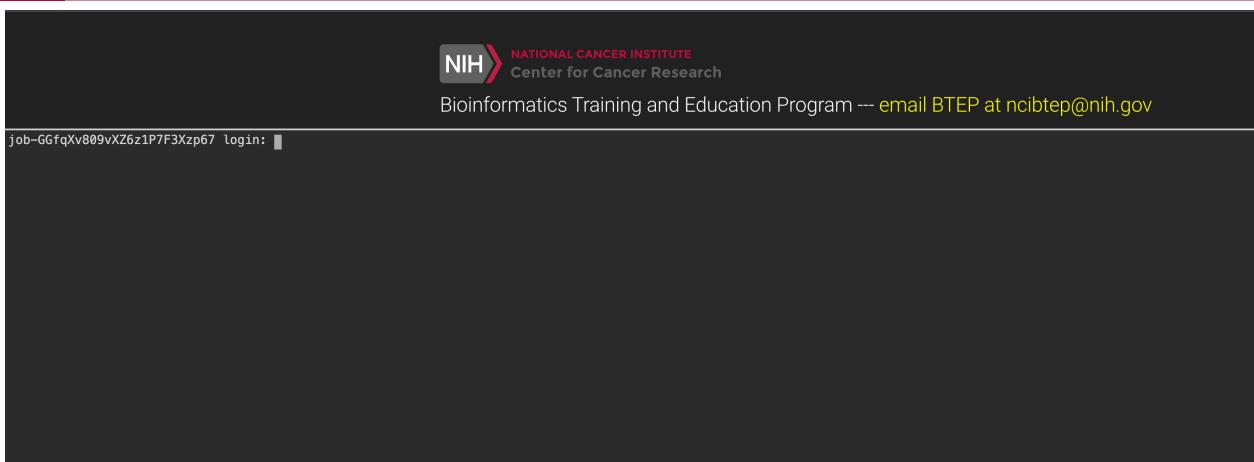
Peter Fitzgerald	Amy Stone lake	Carl McIntosh	Desiree Tilto	Alex Emmons	Joe Wu
peter	amy	carl	des	alex	joe
Files	Files	Files	Files	Files	Files

User	Login	View
First Last	Name	Files
First Last	Name	Files
First Last	Name	Files

-Find your name and select from the Login column

At the top of the page you will see the instructors pictures and logins. You will need to find your name (First and Last) in the table below the instructors. Once you find your name click on the link associated with your name in the login column. **The name that you see in the login column will serve as your username in step 5.**

**Step 5:** The login link will open a terminal with a prompt to login. Login with your username (See step 4) and password (to be distributed in class).



**Step 6:** Once you login at the terminal, you will see the following page:

A screenshot of a web-based course documentation for "Microbiome Analysis with QIIME2". The top navigation bar includes the NIH logo, the title "Microbiome Analysis with QIIME2", and a search bar. On the left, a sidebar lists course topics: Home, QIIME2 Introduction, Metadata and Importing, Generating a feature table / feature data, Filtering, taxonomy, and phylogeny, Microbial diversity, alpha rarefaction, alpha diversity. The main content area is titled "Course Overview" and contains a description of the course goals and its relationship to the QIIME 2 Cancer Microbiome Intervention Tutorial. A red arrow points from the sidebar to the "Command prompt; QIIME 2 activated at login" message in the terminal below. Another red arrow points from the bottom of the sidebar to a double-headed arrow indicating where to drag the screen bar to resize the windows.

The course documentation is accessible at the top of the page and can be dragged up or down for viewing. The command line terminal accounts for the rest of the page. You may need to resize the screen to see the command prompt.

Now you should be logged onto the GOLD platform and ready for class.

**Ending your DNAlexus session:** if you are finished with the GOLDsystem for the day, logout using

```
exit
```

# What is QIIME 2?

A powerful, extensible, and decentralized microbiome analysis package with a focus on data and analysis transparency. QIIME 2 enables researchers to start an analysis with raw DNA sequence data and finish with publication-quality figures and statistical results. --- 2016-2021, QIIME 2 development team.

The plugin architecture of QIIME 2 enables the platform to easily evolve with the latest developments in the field. See the [core plugins](https://docs.qiime2.org/2022.8/plugins/) (<https://docs.qiime2.org/2022.8/plugins/>) as of QIIME 2-2022.8 and the [latest plugins](https://library.qiime2.org/) (<https://library.qiime2.org/>).

## Multiple ways to use QIIME2

### Q2studio

- Graphical user interface

### Q2cli

- Command line interface
- The classic way to use qiime2

### Artifact API

- Python 3 application programming interface (API) for QIIME 2
- Use with Jupyter notebook
- Recommended for advanced users

### Galaxy

The recommended use of qiime2 and the most common use is via command line (q2cli), which we will be using for this course. The q2cli is particularly powerful if you are working with big data. See the additional resources, if you need a brief linux refresher.

## Some useful linux commands

- `pwd` (print working directory)
- `ls` (list)
- `nano` (basic editor for creating small text files)
- `rm` (remove files)
- `mkdir` (make a directory)
- `cd` (change directory)
- `mv` (rename or move files)
- `less` (view files)
- `man` (manual)

- cp (copy)

## QIIME2 Installation

You will not need to install QIIME2 for this course series. If you would like to install QIIME2 on your local computer, there are detailed installation instructions on the [QIIME2 website \(\*https://docs.qiime2.org/2022.8/install/\*\)](https://docs.qiime2.org/2022.8/install/).

## Using QIIME2 on Biowulf

There are also versions of qiime2 available as modules on Biowulf, NIH's high performance computing system.

To see available versions use

```
module avail qiime
```

The default version on Biowulf is qiime2-2021.4, and the latest installed version is qiime2-2022.2.

Also, check out the QIIME2 Biowulf help page (<https://hpc.nih.gov/apps/QIIME.html#:~:text=QIIME2%20on%20Biowulf&text=QIIME%20is%20a%20powerful,quality%20figures>)

If you are interested in a reproducible workflow to use on Biowulf, Samantha Chill, a bioinformatician with CCBR, created a workflow that is readily available from github ([https://github.com/CCBR/BETP\\_microbiome\\_2022](https://github.com/CCBR/BETP_microbiome_2022)).

## What is amplicon sequencing?

The QIIME2 platform can be used for different types of -omics data. For this course, we will be focusing on targeted amplicon sequencing of the 16S rRNA gene.

The 16S rRNA gene (~1500 bp) codes for a ribosomal RNA of the small ribosomal subunit of the prokaryotic ribosome (30S). Ribosomes are made up of proteins and RNAs and are important for translation (protein synthesis from mRNA). The 16S rRNA is highly conserved among bacteria and archaea due to the importance of their function. Within conserved regions of 16S rRNA, there are nine hypervariable regions (V1-V9), and these regions are used for establishing phylogenetic relationships useful for taxonomic classification. See the following figure from Fukuda et al. 2016 ([https://www.jstage.jst.go.jp/article/juoeh/38/3/38\\_223/\\_article](https://www.jstage.jst.go.jp/article/juoeh/38/3/38_223/_article)).

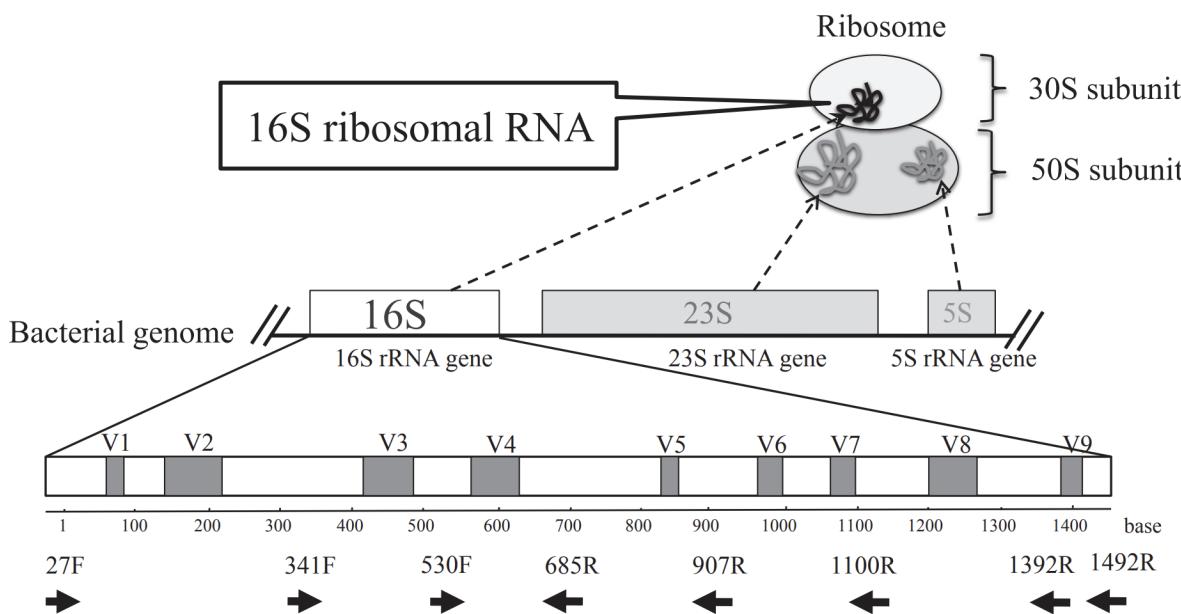


Image from: Fukuda K, Ogawa M, Taniguchi H, Saito M. Molecular Approaches to Studying Microbial Communities: Targeting the 16S Ribosomal RNA Gene. J UOEH. 2016 Sep;38(3): 223-32. doi: 10.7888/juoeh.38.223. PMID: 27627970.

## Things to know about 16S rRNA:

1. Found in all bacteria and archaea. Also present in mtDNA and chloroplasts (See the [endosymbiotic theory](https://en.wikipedia.org/wiki/Symbiogenesis) (<https://en.wikipedia.org/wiki/Symbiogenesis>)).
2. Multiple copies per genome
3. Taxonomy based on 16S rRNA has been extremely popular, and databases are continuously growing.
4. Classification resolution at the Genus level or higher.

## Other target genes of interest:

18S rRNA - Microbial eukaryotes  
 18S rRNA (ITS1, ITS2) - Fungi  
 COI of mtDNA (cytochrome c oxidase) - Animals

## The data used in this tutorial

This course will use code and data from the QIIME2 Cancer Microbiome Intervention tutorial (<https://docs.qiime2.org/jupyterbooks/cancer-microbiome-intervention-tutorial/index.html>) from the QIIME 2 website. The data used herein were published in Liao et al. 2021 (<https://www.nature.com/articles/s41597-021-00860-8>) and Taur et al. 2018 ([https://www.science.org/doi/10.1126/scitranslmed.aap9489?url\\_ver=Z39.88-2003&rfr\\_id=ori:rid:crossref.org&rfr\\_dat=cr\\_pub%3dpubmed](https://doi.org/10.1126/scitranslmed.aap9489?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%3dpubmed)). In particular,

Taur et al. 2018, will be the focus of the data analysis steps or post-denoising steps (See the cancer tutorial from QIIME 2). This research focuses on reestablishing the gut microbiome using auto-FMT (a fecal transplant using the patient's preserved gut microbiome) following allogeneic hematopoietic stem cell transplantation. More on this later.

## Getting started with QIIME2

QIIME2 is a platform for the processing and analysis of microbiome sequencing data. A general amplicon workflow in QIIME2 may look like the following:

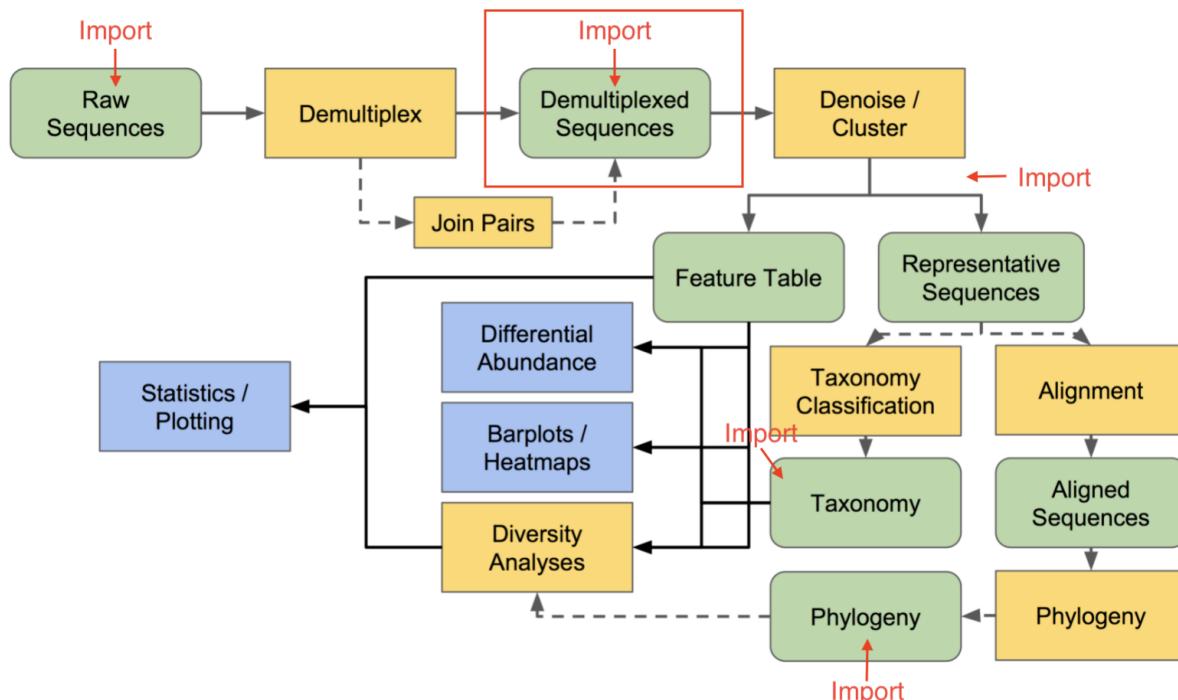


Image adapted from QIIME2 documentation (Conceptual overview of QIIME2) (<https://docs.qiime2.org/2022.8/tutorials/overview/#let-s-get-oriented-flowcharts>)

The first step is to import the data as a QIIME2 artifact (a .qza file). Data can be imported at most stages in the workflow. The red arrows highlight only a few possible objects one would be interested in importing. We are starting in the red box.

## QIIME2 Artifacts

Before we get started, let's briefly discuss the two main file types used with QIIME2. These are .qza files and .qzv files.

.qza - The QIIME2 artifact file. These contain data.

.qzv - The QIIME2 visualization file. These contain visualizations that can be viewed using QIIME 2 View (<https://view.qiime2.org/>).

These are zipped files that contain provenance and other information in addition to data. Each artifact has a unique identifier so that you can easily track provenance.

You can simply use `unzip` to access the data, or use `qiime tools export`. Check out the QIIME2 export tutorial (<https://docs.qiime2.org/2022.8/tutorials/exporting/>) for more information on exporting data and visualizations.

## Qiime2view

We will use `Qiime2view` (<https://view.qiime2.org/>) frequently throughout this course, and you will use it frequently in the future if you plan to use QIIME2 in your research. This is a great tool for exploring QIIME2 visualizations. You can drag and drop files from your local computer or visualize a file from the web. This allows you to easily share results with collaborators without requiring software installations.

## Metadata formatting

For any next generation sequencing experiment, you will need sample information (sample metadata) to make sense of your data. The key to a good study is to collect good metadata. You should minimally have all of the information required to investigate your hypotheses. Also, check out the `MIMARKS` (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3367316/table/T1/>) recommendations for depositing data in NCBI and other data repositories.

### QIIME2 requirements for sample metadata

- In tsv (tab separated) format
- Include a `SampleID` column as the first column.
- Missing data is represented by empty cells, not NAs
- Supports categorical and numeric data (may include a row with `#q2 : types` of either categorical or numeric)
- rows that begin with `#` are ignored.
- Whitespace is ignored.

`Keemei` (<https://keemei.qiime2.org/>) is a particularly nice metadata validation plugin on google chrome, if your data is available in google sheets.

For more detailed information on QIIME 2 metadata, see the QIIME2 metadata tutorial (<https://docs.qiime2.org/2022.8/tutorials/metadata/>).

**Note on Excel:** Excel will also guess column types at import. This has resulted in incorrect gene names in data (E.g., `SEPT4` becoming `4-Sept`) ([https://www.nature.com/articles/d41586-021-02211-4?utm\\_source=Nature+Briefing&utm\\_campaign=f3302f10ba-briefing-dy-20210818&utm\\_medium=email&utm\\_term=0\\_c9dfd39373-f3302f10ba-45948134](https://www.nature.com/articles/d41586-021-02211-4?utm_source=Nature+Briefing&utm_campaign=f3302f10ba-briefing-dy-20210818&utm_medium=email&utm_term=0_c9dfd39373-f3302f10ba-45948134)). It can also lead to altered changes in sample names, if sample names are numeric.

## Examining the metadata

Let's take a look at the metadata associated with QIIME 2 Cancer Microbiome Intervention tutorial.

```
qiime metadata tabulate \  
  --m-input-file /data/sample-metadata.tsv \  
  --o-visualization metadata-summary.qzv
```

This command allows us to interactively explore the metadata.

If we simply want to get some basic information. We can try out one of the QIIME 2 utilities.

```
qiime tools inspect-metadata /data/*tsv
```

This gives us the column names, types, and the dimensions of the data.

## Data import

As mentioned previously, the first step of any QIIME 2 analysis will be to import the data. Each type of data will be stored in its own QIIME2 artifact. For example, sample metadata, ASV / OTU tables, representative sequences, taxonomy, will each be located in a different qza file. This will make more sense as we begin to work through the data. Check out the [QIIME2 Importing Data](#) (<https://docs.qiime2.org/2021.8/tutorials/importing/#>) tutorial for examples on how to import different types of data.

In following the QIIME2 Cancer Microbiome Intervention Tutorial, this course will use [Liao et al. 2021](#) (<https://www.nature.com/articles/s41597-021-00860-8>) and [Taur et al. 2018](#) ([https://www.science.org/doi/10.1126/scitranslmed.aap9489?url\\_ver=Z39.88-2003&rfr\\_id=ori:rid:crossref.org&rfr\\_dat=cr\\_pub%3dpubmed](https://www.science.org/doi/10.1126/scitranslmed.aap9489?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%3dpubmed)).

## Importing raw fastq files

We will import a small subset of fastq files to demonstrate initial sequence processing steps. The main questions we need to answer to import our raw data are:

1. Is the data demultiplexed or multiplexed?

Often you will receive your data back from a sequencing facility already demultiplexed, meaning the sequences have been separated by sample into individual files. In this case, you will have a `.fastq` file per sample (or two if paired end). Conversely, you may need to demultiplex, or split the reads by sample, yourself. In this case, you will often have three files containing the barcodes, the forward reads, and the reverse reads. QIIME 2 has protocols for handling demultiplexed or multiplexed data.

## 2. Is the data paired-end or single-end?

We also will need to know whether our data is paired-end, includes forward and reverse reads, or single-end. NOTE: If using paired-end sequencing, make sure the paired end chemistry is sufficient for read overlap. For example 250 PE chemistry will result in almost complete overlap of V4.

Check out [this forum post](https://forum.qiime2.org/t/importing-and-demultiplexing-sequence-data-quick-reference/14002) (<https://forum.qiime2.org/t/importing-and-demultiplexing-sequence-data-quick-reference/14002>) for raw data import guidance.

## Importing our example data

In our example data, the sequences are paired-end demultiplexed data.

Raw fastq files are currently in a directory named `/data/data_to_import`. QIIME2 has specific functions for importing specific types of raw sequencing data. There are protocols for EMP data (multiplexed and demultiplexed), other multiplexed fastq data, Casava 1.8 demultiplexed data (format: `SampleID_BarcodelD_L001_R1_001.fastq.gz`), and all other types of demultiplexed fastq data using a fastq manifest.

All import steps use `qiime tools import` but they vary in the command options (`--type` and `input-format`)

If you run

```
qiime tools import --help
```

You will see there are options to view the arguments for `types` and `formats`.

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

We know we have paired end fastq sequence files, so the type that works best for that appears to be `SampleData[PairedEndSequencesWithQuality]`.

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

We can tell by our file names (e.g., FMT.0093C\_46\_L001\_R2\_001.fastq.gz) that we have Casava data, so we can select `CasavaOneEightSingleLanePerSampleDirFmt` for the importable format.

The only additional information we need to provide include the `--input-path`, which is where our data is located (i.e., `/data/data_to_import`) and the `--output-path`, where we want the results to be stored. Let's name our imported demultiplexed sequence artifact `demuxsequences.qza`.

Let's import

```
qiime tools import \  
  --type 'SampleData[PairedEndSequencesWithQuality]' \  
  --input-format CasavaOneEightSingleLanePerSampleDirFmt \  
  --input-path /data/data_to_import \  
  --output-path demuxsequences.qza
```

See the [moving pictures tutorial](https://docs.qiime2.org/2022.8/tutorials/moving-pictures/) (<https://docs.qiime2.org/2022.8/tutorials/moving-pictures/>) for an example of importing multiplexed EMP sequences.

Our output (`demuxsequences.qza`) is demultiplexed sequences ready for denoising or OTU clustering.

## Summary of imported data

Following import, we want to check our sequence quality and the number of sequences per sample (read depth). This can be done using `qiime demux summarize`.

```
qiime demux summarize \  
  --i-data demuxsequences.qza \  
  --o-visualization demuxsequences-summary.qzv
```

Let's move this file to `public` so that we can view it on [view.qiime2.org](https://view.qiime2.org) (<https://view.qiime2.org/>).

```
mv demuxsequences-summary.qzv public/
```

We will return to these results in Lesson 3, to determine the parameters for denoising.

## Import ASV table

To demonstrate that you can import at later stages in the workflow, let's import a feature table (e.g. ASV count matrix). To import an ASV / OTU table, the table has to be in `.biom` format.

Luckily, there is nice documentation ([https://biom-format.org/documentation/biom\\_conversion.html](https://biom-format.org/documentation/biom_conversion.html)) for converting a tab-delimited file to a biom file.

Note: A feature table is the equivalent of an OTU / ASV table in QIIME2.

```
qiime tools import \
--input-path /data/feature-table.biom \
--type 'FeatureTable[Frequency]' \
--input-format BIOMV210Format \
--output-path featuretable_ex.qza
```

## Provenance Tracking

Every qiime2 artifact includes provenance information, which includes things like the unique ID of the artifact(s) used as input, the format, type, method and action, run time, etc.

You can check the uuid (universally unique identifier) of an artifact at any time using `qiime tools peek filename`.

## QIIME2 forum

Lastly, the [QIIME2 forum](https://forum.qiime2.org/) (<https://forum.qiime2.org/>) is a fantastic resource to get help with qiime2 plugins or questions related to your workflow or research design. There is also a "**best-of-the-forum**" (<https://forum.qiime2.org/tag/best-of-the-forum>) tag, which is worth a peruse.

# Lesson 3: Creating a feature table

## Lesson Objectives

- Check for primers
- Generate an ASV count table and representative sequence file
- Understand the difference between OTU picking and denoising

The two primary files that will be used throughout any microbiome analysis are the

1. feature table (OTU / ASV count data)
2. feature data (representative sequences).

These will be generated using either an OTU clustering method or a denoising method. The goal is to end up with counts of features, whether these be OTUs or ASVs (ESVs, zOTUs, etc.). Ideally, these features represent an organism or species of organisms.

### But first...

Depending on your library preparation protocol, you may or may not have primers in your sequences. Non-biological sequences (e.g., adapters, primers, linker pads, etc.) can pose problems for OTU clustering and denoising (e.g., by interfering with [chimera](https://www.drive5.com/usearch/manual/chimeras.html) (<https://www.drive5.com/usearch/manual/chimeras.html>) identification).

Primers are generally not found in sequences when using the [EMP protocol](https://earthmicrobiome.org/protocols-and-standards/16s/) (<https://earthmicrobiome.org/protocols-and-standards/16s/>) due to the use of custom sequencing primers, but most other library prep strategies will result in primers in the sequences.

We can use `qiime cutadapt` to trim primers. If you are unsure whether to expect primers or not, it's a good idea to use `qiime cutadapt` to check for the presence of primers.

## Primer trimming

### Using `qiime cutadapt trim-paired`

We will use `qiime cutadapt trim-paired` because we are working with paired-end reads. We will also use more than one thread (`--p-cores`). The region sequenced was V4-V5 with the forward primer, AYTGGGYDTAAAGNG, and the reverse primer, CCGTCAATTYHTTTRAGT, which will fill the `--p-front-f` and `--p-front-r` parameters respectively. We will also specify that we want the command to run in `--verbose` mode so that we can save the output from cutadapt to a log for our records.

```
qiime cutadapt trim-paired \
--i-demultiplexed-sequences demuxsequences.qza \
--p-cores 8 \
--p-front-f AYTGGGYDTAAAGNG \
--p-front-r CCGTCAATTYHTTTRAGT \
--verbose \
--o-trimmed-sequences demux-trimmed.qza | tee cutadaptresults.log
```

{{Sdet}}

What is `tee`?{{Esum}}

`tee` (<https://www.geeksforgeeks.org/tee-command-linux-example/>) allows us to display output as standard output (to the screen) and written to file.

{{Edet}}

For the most part, it seems that the primers are not in the sequences, aside from a 4 bp match. I wouldn't worry about these 4 bp unless something seems off in later analyses. From the overall length of the sequences, ~220 bp forward and ~217 bp reverse, it seems that the primers have likely been trimmed.

## OTU Clustering vs Denoising

Once we have removed non-biological sequences, we can proceed to OTU clustering or denoising.

### What do we mean by OTU Clustering / OTU picking?

An OTU is an operational taxonomic unit. This is derived by binning sequences at a certain threshold of similarity. This threshold is generally set at 97%, which is associated with a species level assignment.

Clustering can

1. eliminate variation added by sequencing error
2. avoid splitting counts from one organism, as bacteria have more than one 16S rRNA gene (average 4 copies), which may differ.
3. give you an idea of "species" counts; though, there may be important phenotypic differences between strains of a species.

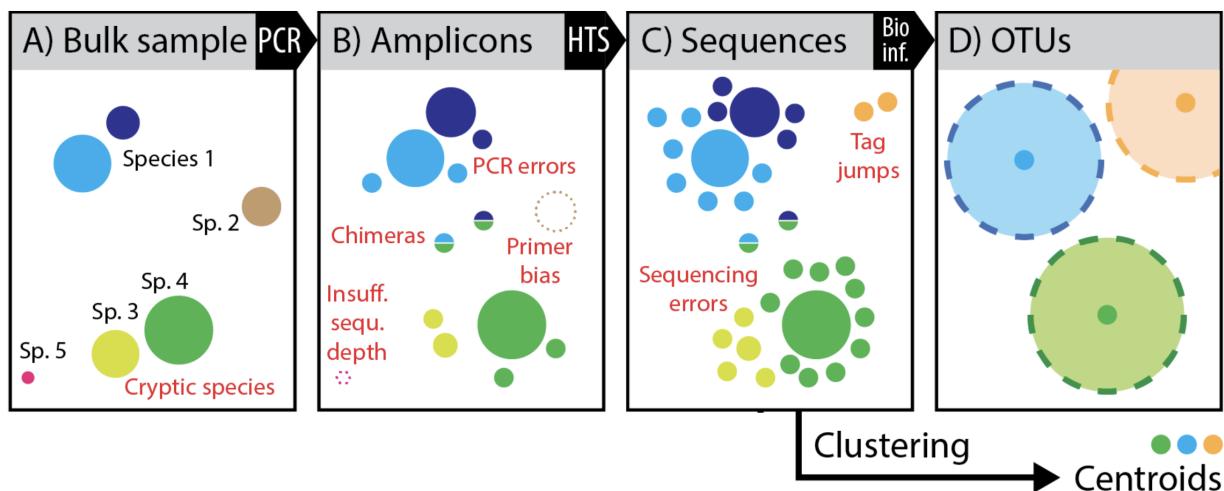


Figure modified slightly from Elbrecht V, Vamos EE, Steinke D, Leese F. 2018. Estimating intraspecific genetic diversity from community DNA metabarcoding data. PeerJ 6:e4644 <https://doi.org/10.7717/peerj.4644> Supplemental Figure 1.

This image nicely outlines the biases and errors that accumulate from sample collection to sequencing and how those errors culminate in OTUs. First, with amplification you may have some primer bias, misrepresentation by sequencing depth, and chimera formation (B) then with high throughput sequencing we include sequencing error complicated by things like tag jumping in metabarcode studies. What we end up with if we cluster into OTUs is a cluster of sequences (represented usually by a single dominant sequence) considered to be a microbial species, but in reality this species could be multiple species and erroneous sequences clustered together. For example, the yellow and green species were lumped together. Also, only the representative sequence of the group will later be used for taxonomic classification and tree generation.

## Methods for OTU clustering include

### 1. Closed reference OTU picking

- uses a reference database to cluster OTUs against
- reads that fail to cluster with a reference sequence due to real variation or sequence error variation are dropped
- subject to biases or errors in databases; lose any newly recovered taxa
- Can be used to compare studies if the same reference is used

### 2. De novo OTU picking

- creates clusters of sequences based only on the observed sequences (no reference database)
- Cannot be compared across studies

Open reference OTU picking (combination of 1 and 2)

- a combination of closed-reference and de novo
- sequences are clustered based on a reference
- sequences that fail to cluster to a reference sequence are clustered using a de novo method
- Cannot be compared across studies

Some potential reference databases include Greengenes, SILVA, and UNITE (for ITS fungi).

## Clustering methods on QIIME 2

1. q2-dbotu (<https://library.qiime2.org/plugins/q2-dbotu/4/>)
2. q2-vsearch (<https://docs.qiime2.org/2022.8/tutorials/otu-clustering/?highlight=otu>)

## Denoising

The field has moved to denoising sequences rather than OTU clustering. In a denoising approach, the exact biological sequence is inferred and noise is removed from the dataset via error correction. This is generally done using some type of error modeling, dependent on sequence quality information. More details [here](https://www.zymoresearch.com/blogs/blog/microbiome-informatics-otu-vs-asv) (<https://www.zymoresearch.com/blogs/blog/microbiome-informatics-otu-vs-asv>).

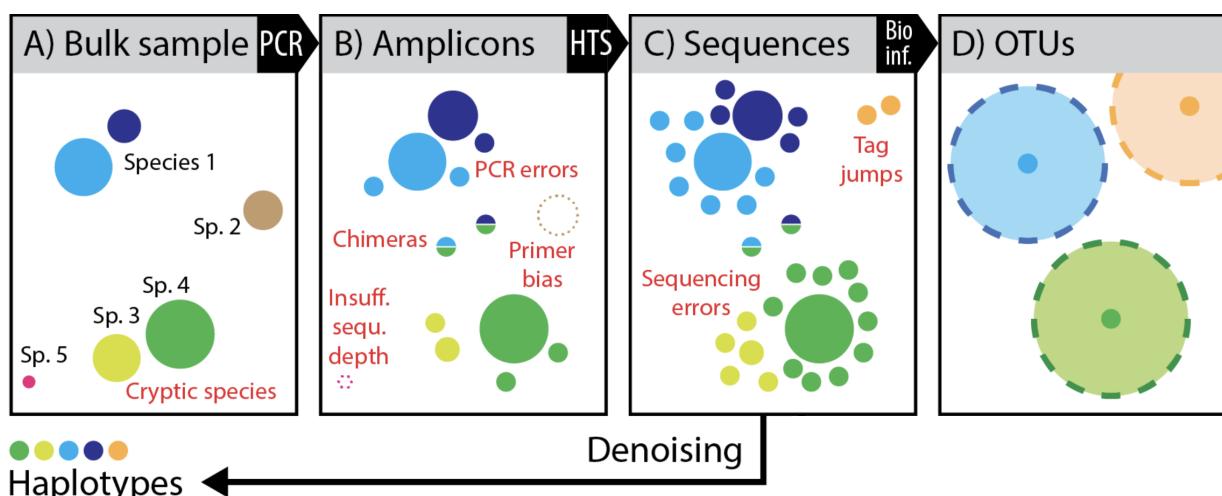


Figure modified slightly from Elbrecht V, Vamos EE, Steinke D, Leese F. 2018. Estimating intraspecific genetic diversity from community DNA metabarcoding data. PeerJ 6:e4644 <https://doi.org/10.7717/peerj.4644> Supplemental Figure 1.

Back to our figure from Elbrecht et al. 2018, we can see that we get much greater resolution of the original diversity, when we remove sequences impacted by error. Though, we could have inflated diversity in the case of a single organism having multiple 16S rRNA copies with 1-2 nucleotide differences.

## Denoising methods on QIIME2

The two methods used for denoising on QIIME 2 include:

**DADA2** (<https://www.nature.com/articles/nmeth.3869>)

- Uses a run specific error profile
- Unclear how an incomplete run profile would impact results
- There is a **method** (<https://docs.qiime2.org/2022.8/plugins/available/dada2/denoise-ccs/>) available for Pacbio CCS sequences

**Deblur** (<https://msystems.asm.org/content/2/2/e00191-16>)

- Not run specific, uses a fixed model
- Does not include an inherent read joining step (will drop reverse reads)
- **Can read join before denoising** (<https://docs.qiime2.org/2022.8/tutorials/read-joining/>)

Related methods include:

**Minimum Entropy Decomposition (MED)** (<https://www.nature.com/articles/ismej2014195>)

**Unoise3** ([https://www.drive5.com/usearch/manual/cmd\\_unoise3.html](https://www.drive5.com/usearch/manual/cmd_unoise3.html))

Note: Each of these methods calls the resulting features something different (e.g., ASVs, zOTUs, sOTUs, etc.).

For a comparison of DADA2, Deblur, and Unoise3, see *Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches* (<https://peerj.com/articles/5364/>).

For an interesting discussion on OTUs and ASVs, see this (<https://forum.qiime2.org/t/to-cluster-or-not-to-cluster/10022/6>) QIIME 2 forum post.

Also...You can do both denoising and OTUclustering! See **qiime vsearch cluster-features-open-reference** (<https://docs.qiime2.org/2022.8/plugins/available/vsearch/cluster-features-open-reference/>)

## Preparing to Denoise (Hands on)

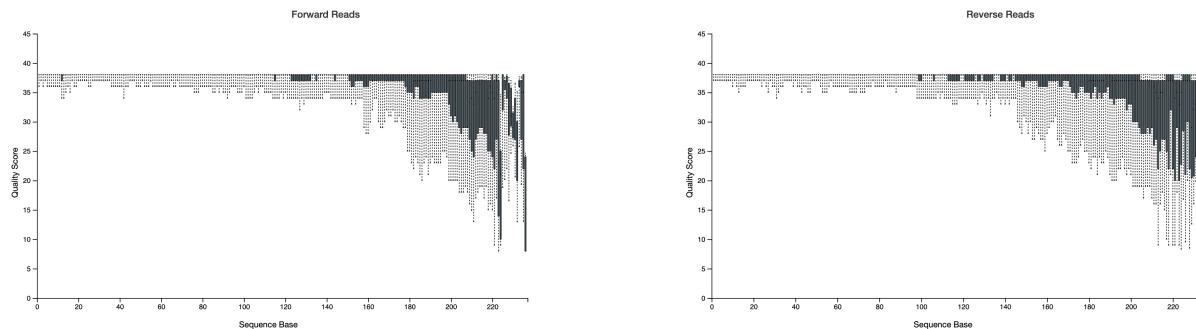
Now that we know what we mean by denoising, let's apply it to our data. We will use **DADA2**, which seems to be the more popular method. To use DADA2, we need to make some decisions regarding the quality of our data. This means we need to refer back to our output from **qiime demux summarize**. In particular, we want to trim where the reads (forward and reverse) drop in quality. We can also trim any low quality bases from the beginning of the reads. However, we need to be wary of the length of our sequences (forward and reverse), the size of our target amplicon, and the general size of the overlap (between forward and reverse for merging). If we trim too much, we could impact our ability to merge the reads. Our primers are 563F and 926R, targeting V4-V5, so we are expecting a 363 bp amplicon. Because we are using 250 PE sequencing we can use the following to calculate approximate overlap:

(forward read) + (reverse read) - (length of amplicon) = overlap

$250 + 250 - 363 = 137$  bp overlap (Note: If primers were removed, it would not impact the overlap, so I'm going to ignore the current lengths in our summary report at the moment)

(See some explanations here (<https://forum.qiime2.org/t/merging-quality-control-and-overlapping/12618/2>) and here (<https://forum.qiime2.org/t/questions-about-v3-v4-primers-for-16s-rRNA-amplicon-sequencing-and-calculating-overlap/20250>).

We need to trim based on quality but also recognize that sequences with lengths less than our truncation lengths will be removed, so we also need to consider our length table.



## Demultiplexed sequence length summary

### Forward Reads

Total Sequences Sampled	10000.0
2%	215 nts
9%	216 nts
25%	217 nts
<b>50% (Median)</b>	<b>220 nts</b>
75%	222 nts
91%	223 nts
98%	223 nts

### Reverse Reads

Total Sequences Sampled	10000.0
2%	213 nts
9%	213 nts
25%	215 nts
<b>50% (Median)</b>	<b>217 nts</b>
75%	219 nts
91%	220 nts
98%	223 nts

Notice that these plots are based on a subsampling of 10,000 reads, so they are simply a representation of our data. Based on these results, we will need to trim at a maximum of ~217 bp (forward) and ~215 bp (reverse) if we want to retain a majority of our sequences. In the Cancer Microbiome tutorial, they suggest truncating sequences when the twenty-fifth percentile quality score drops below 30. This is a bit conservative; I would consider lowering this to 25. The cutoffs are fairly arbitrary and will be data dependent. Normally, I would trim at ~221 bp for the forward and ~213 bp for the reverse, but due to the length inconsistencies, let's go a bit shorter to avoid losing shorter reads (207 bp Forward; 204 bp Reverse). Even with this lower cut off, our reads should overlap just fine. I will also set `--p-n-threads` to 2 or more to improve the speed. Threads should be set based on your computational resources.

```
time qiime dada2 denoise-paired \
--i-demultiplexed-seqs demuxsequences.qza \
--p-trunc-len-f 207 \
--p-trunc-len-r 204 \
--p-n-threads 2 \
--o-representative-sequences asv-sequences.qza \
--o-table feature-table.qza \
--o-denoising-stats dada2-stats.qza
```

(With 8 threads, this took 13 minutes, so we will run this at the beginning of the lesson.)

## Denoising stats

Let's take a look at our DADA2 stats. This will give us an idea about the number of reads filtered at various steps.

```
qiime metadata tabulate \
--m-input-file dada2-stats.qza \
--o-visualization dada2-stats-summ.qzv
```

Upon examining our denoising stats, we can see that > 80% of reads passed the initial input filter; 75-92% of reads were merged, and 70-92% or reads were non-chimeric. If you see heavy read loss here, you may want to refer back to your quality information and adjust denoising parameters.

## Feature table and feature data summary information

Finally, we want to obtain summary information from our feature table and feature data (representative sequences). Our feature table includes count data of our ASVs in each sample, while the feature data provides the sequence for each ASV.

```
qiime feature-table summarize \
--i-table feature-table.qza \
--m-sample-metadata-file /data/sample-metadata.tsv \
--o-visualization feature-table-summ.qzv
qiime feature-table tabulate-seqs \
--i-data asv-sequences.qza \
--o-visualization asv-sequences-summ.qzv
```

Overview    Interactive Sample Detail    Feature Detail

### Table summary

Metric	Sample
Number of samples	41
Number of features	658
Total frequency	1,532,349

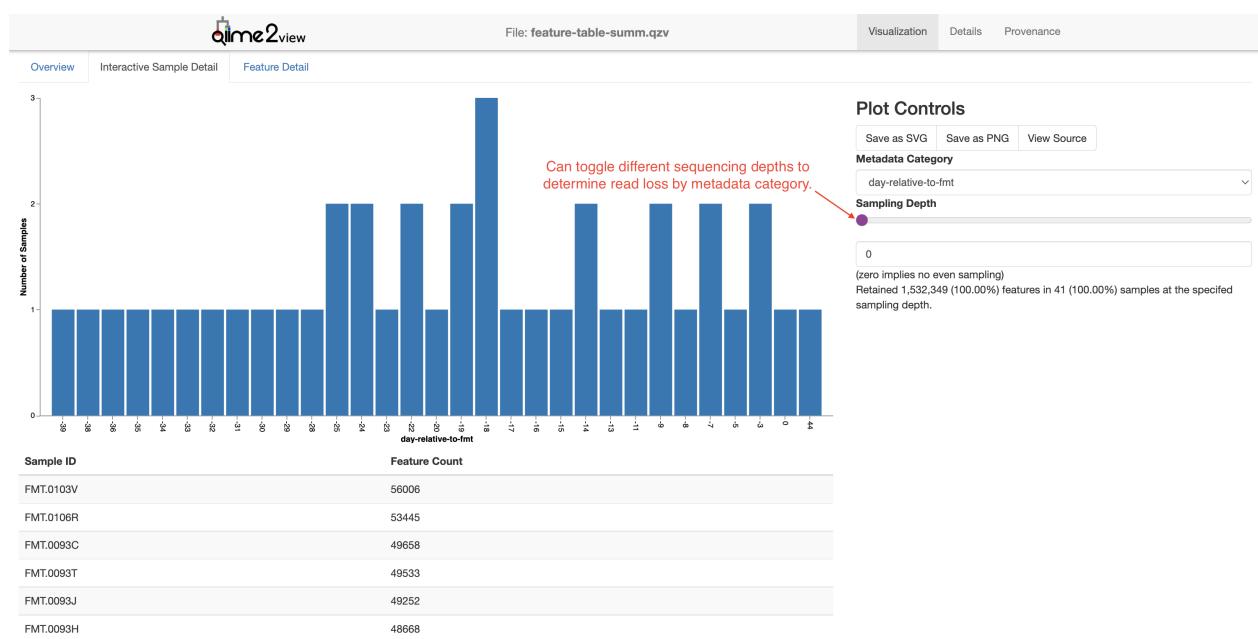
### Frequency per sample

	Frequency
Minimum frequency	16,946.0
1st quartile	31,754.0
Median frequency	37,825.0
3rd quartile	45,010.0
Maximum frequency	56,006.0
Mean frequency	37,374.365853658535

Frequency per sample detail ([csv](#) | [html](#))

Download as PDF

This provides nice information about the number of samples, the number of features, the number of reads per sample, the number of features per sample, etc.



In the *Interactive Sample Detail* you can obtain additional information about how modifying the sequencing depth (i.e., rarefaction) impacts sample loss. (More on rarefaction later.)

We can also look at a summary of our representative sequences. Unless you modified denoising parameters, feature ids will be hashed (<https://medium.com/swlh/the-difference-between-encoding-encryption-and-hashing-878c606a7aff>). The hash is always the same for identical sequences, so this is not an issue for merging runs. The cool thing about this interactive summary is that you can click on the sequence, which will take you to NCBI's blastn.

We can also see that our read lengths match nicely with our expected amplicon size without primers (~330 bp).

Important note: DADA2 and Deblur are not length agnostic, meaning features with different sequencing lengths will be considered different features. Thus, we are generally fairly careful when choosing method parameters, especially if merging several sequencing runs. The same parameters should be applied across all runs. For DADA2, it is important that runs be denoised individually and then merged.

## References

<https://www.zymoresearch.com/blogs/blog/microbiome-informatics-otu-vs-asv> www.drive5.com  
Elbrecht V, Vamos EE, Steinke D, Leese F. 2018. Estimating intraspecific genetic diversity from community DNA metabarcoding data. PeerJ 6:e4644 <https://doi.org/10.7717/peerj.4644>

# Lesson 4: Feature table filtering, taxonomic classification, and phylogeny

## Learning objectives

- learn how to apply different types of filtering to your ASV table and representative sequence data.
- classify your ASVs.
- Generate a phylogenetic tree.

Now that we have imported and denoised, let's move on to feature table filtering, taxonomic classification, and phylogenetic tree construction. For this lesson, we will use a much larger feature table that contains the entire data set from Taur et al. 2018 (See the QIIME2 Cancer Microbiome Tutorial). This will make the analysis more interesting.

We are going to skip a few filtering steps compared with the QIIME2 tutorial. There are great filtering functions in QIIME2. We will apply some of these, but you should have a look at the filtering tutorial (<https://docs.qiime2.org/2022.8/tutorials/filtering/>). We will work directly with the autoFMT study and will skip the filtering step that returns only this group ('--p-where 'autoFmtGroup IS NOT NULL').

## Filtering

### Methods of filtering

- Frequency based filtering, contingency based filtering, metadata based filtering
- `qiime feature-table filter-samples`
  - Filter samples based on total number of sequences (e.g., `--p-min-frequency 1000`)
  - Filter samples with a minimal number of features (e.g., `--p-min-features 10`)
  - Metadata based filtering:
    - Retain only samples present in the metadata (e.g., `--m-metadata-file samples-to-keep.tsv`)
    - Retain samples based on a metadata description (e.g., `--p-where "[subject]='subject-1'"`). Uses SQLite WHERE-clause syntax (Use IN, AND, OR, AND NOT).

- `qiime feature-table filter-features`
- - Remove features (ASVs) with low abundances after summing across all samples (e.g., `--p-min-frequency 10`)
  - Remove features found in only a small number of samples (e.g., `--p-min-samples 2`)
- Taxonomy based filtering (`qiime taxa filter-table`, `qiime taxa filter-seqs`)
  - MORE ON THIS BELOW
- Distance matrix filtering (`qiime diversity filter-distance-matrix`)

*The filtering you apply will depend on your data and your questions. I tend to filter out low abundant taxa and depending on my question I may also filter taxa with low prevalence.*

## Metadata based filtering

We are going to eliminate some samples based on the timing of collection. We are interested in only the samples collected between ten days prior to their hematopoietic cell transplantation and seventy days post transplantation.

To do this, let's use the `--p-where` option of `qiime feature-table filter-samples`

```
qiime feature-table filter-samples \
--i-table /data/autofmt-table.qza \
--m-metadata-file /data/sample-metadata.tsv \
--p-where 'DayRelativeToNearestHCT BETWEEN -10 AND 70' \
--o-filtered-table filtered-table-1.qza
```

Notice the SQLite syntax with key words `BETWEEN` and `AND`. If confused by the syntax, check out the filtering tutorial for other examples.

The next two steps will primarily focus on reducing run times. You may or may not want to include them in your own analysis workflow.

## Feature filtering

We can filter features observed in only a single sample.

```
qiime feature-table filter-features \
--i-table filtered-table-1.qza \
--p-min-samples 2 \
--o-filtered-table filtered-table-2.qza
```

Unless we explicitly remove them, removed features will remain in our FeatureData (Representative Sequences). This isn't necessarily a problem, but eliminating them can reduce

computation time when classifying our sequences and generating a phylogenetic tree. So, we will do that here. Also, because we are starting from a larger data set in this part of the tutorial, we will need a new, more comprehensive representative features table.

Let's use `wget` to get the new representative sequences.

```
wget \
-O 'rep-seqs.qza' \
'https://docs.qiime2.org/jupyterbooks/cancer-microbiome-interventional-analysis/qza-representative-sequences.qza'
```

```
qiime feature-table filter-seqs \
--i-data rep-seqs.qza \
--i-table filtered-table-2.qza \
--o-filtered-data filtered-sequences-1.qza
```

Let's summarize our newly filtered tables.

```
qiime feature-table summarize \
--i-table filtered-table-2.qza \
--m-sample-metadata-file /data/sample-metadata.tsv \
--o-visualization filtered-table-summ.qzv
qiime feature-table tabulate-seqs \
--i-data filtered-sequences-1.qza \
--o-visualization filt-asv-sequences-summ.qzv
```

## Taxonomy

After denoising, we have unique ASVs or sequences, which in itself can be quite revealing. However, often we want to classify our sequences to know more about the organisms represented by our ASVs, particularly regarding evolutionary relationships. In QIIME2, we can classify our representative sequences using the `q2-feature-classifier` plugin (<https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0470-z>) and generate a phylogenetic tree via de novo or reference based methods.

Alignment based taxonomy consensus classifiers:

**BLAST+** (<https://pubmed.ncbi.nlm.nih.gov/20003500/>) - local sequence alignment followed by consensus taxonomy classification

**VSEARCH** (<https://pubmed.ncbi.nlm.nih.gov/27781170/>) - global sequence alignment followed by consensus taxonomy classification

These essentially align sequences to references and take the top matches (maxaccepts) above some threshold of similarity (perc\_identity) and then assign taxonomy by how well the top matches agree (min\_consensus).

**maxaccepts** - the maximum number of hits to keep for each query

**perc\_identity** - the percentage of similarity below which a match is rejected

**min\_consensus** - the minimum fraction of assignments that must match the top hits from maxaccepts to be accepted as consensus assignment

A nice description of these can be found in [Bokulich, N.A., Kaehler, B.D., Rideout, J.R. et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 6, 90 \(2018\). <https://doi.org/10.1186/s40168-018-0470-z> \(<https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0470-z>\).](#)

## Machine learning approach

The recommended method involves applying a [scikit-learn \(<http://scikit-learn.org/>\)](#) multinomial naive Bayes classifier. This uses machine learning to train a classifier on a reference database and then use the trained classifier to classify your ASVs.

These classifiers work best if they are trained on your target region of interest ([Werner et al. 2012 \(<https://pubmed.ncbi.nlm.nih.gov/21716311/>\)](#), in this case V4-V5. While there are some pre-trained classifiers in the QIIME 2 [Data Resources \(<https://docs.qiime2.org/2021.11/data-resources/>\)](#), the classifier used here was created according to the [Training feature classifiers with q2-feature-classifier tutorial \(<https://docs.qiime2.org/2022.8/tutorials/feature-classifier/>\)](#) using Greengenes v 13.8.

{}{Sdet}}

The code to train the classifier{}{Esum}}

```
qiime tools import \
--type 'FeatureData[Sequence]' \
--input-path 99_otus.fasta \
--output-path 99_otus.qza
```

```
qiime tools import \
--type 'FeatureData[Taxonomy]' \
--input-format HeaderlessTSVTaxonomyFormat \
--input-path 99_otu_taxonomy.txt \
--output-path ref-taxonomy.qza
```

```
qiime feature-classifier extract-reads \
--i-sequences 99_otus.qza \
```

```
--p-f-primer AYTGGGYDTAAAGNG \
--p-r-primer CCGTCAATTYHTTTRAGT \
--p-min-length 100 \
--p-max-length 400 \
--p-n-jobs 10 \
--o-reads ref-seqs.qza
```

```
qiime feature-classifier fit-classifier-naive-bayes \
--i-reference-reads ref-seqs.qza \
--i-reference-taxonomy ref-taxonomy.qza \
--o-classifier gg-13-8-99-563-926-nb-classifier.qza
```

{{Edet}}

Note: Files to create a train classifier are not available in DNAexus. This code is simply for example.

We will use our trained classifier to annotate our ASVs.

```
qiime feature-classifier classify-sklearn \
--i-classifier /data/gg-13-8-99-563-926-nb-classifier.qza \
--i-reads filtered-sequences-1.qza \
--o-classification taxonomy.qza
```

And generate a summary of the results:

```
qiime metadata tabulate \
--m-input-file taxonomy.qza \
--o-visualization taxonomy.qzv
```

Note: You can increase classification accuracy by providing taxonomic weights for certain common sample types (See QIIME 2 clawback (<https://www.nature.com/articles/s41467-019-12669-6#Sec7>))

## Taxonomic based filtering

By targeting 16S rRNA, we want to target bacteria and archaea. Therefore, we can exclude sequences that are unexpected such as those from chloroplasts or mitochondria. By setting `--p-include p__`, we are retaining only sequences annotated at a minimum to the phylum level. Note: this will look different depending on the database used. Greengenes specifically uses the following format for annotations: `k__;p__;c__;o__;f__;g__;s__`. Also, `--p-mode contains` ensures that search terms are case insensitive (e.g., mitochondria versus Mitochondria).

```
qiime taxa filter-table \
--i-table filtered-table-2.qza \
--i-taxonomy taxonomy.qza \
--p-mode contains \
--p-include p__ \
--p-exclude 'p__,Chloroplast,Mitochondria' \
--o-filtered-table filtered-table-3.qza
```

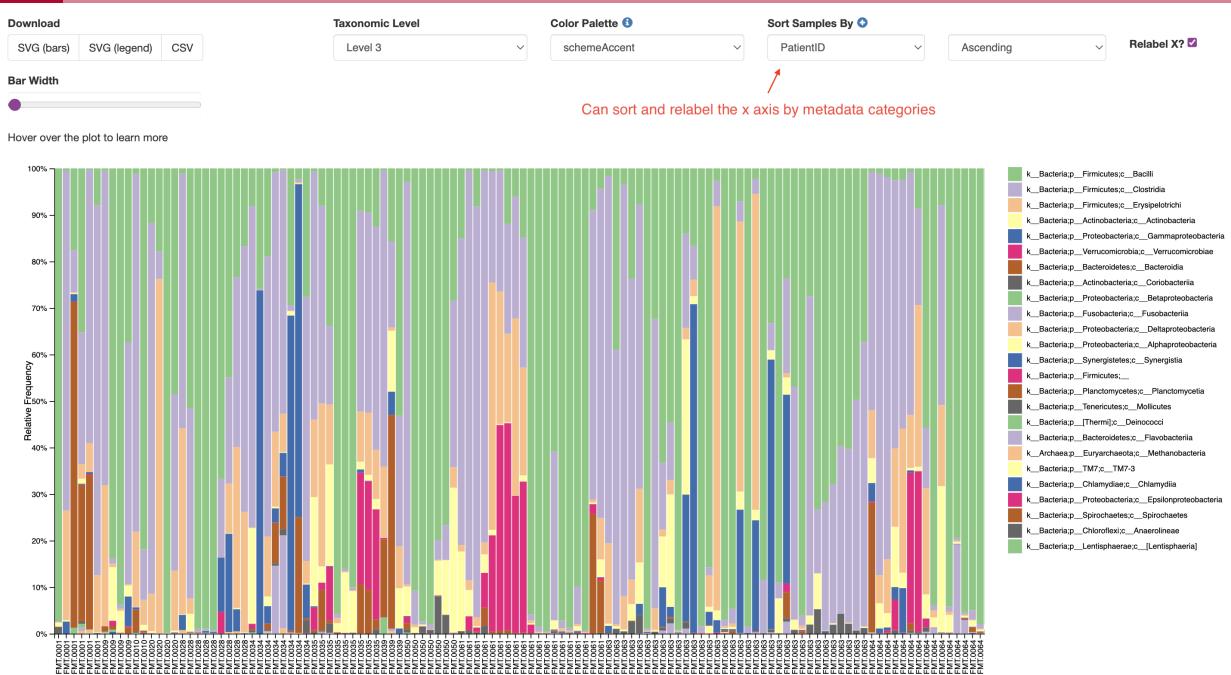
Let's also remove these from our feature data to save computational time later.

```
qiime feature-table filter-seqs \
--i-data filtered-sequences-1.qza \
--i-table filtered-table-3.qza \
--o-filtered-data filtered-sequences-2.qza
```

## Visualizing our taxonomy

We can visualize sample by sample taxonomic composition using a stacked bar plot generated with `qiime taxa barplot`. Let's take a look.

```
qiime taxa barplot \
--i-table filtered-table-3.qza \
--i-taxonomy taxonomy.qza \
--m-metadata-file /data/sample-metadata.tsv \
--o-visualization taxa-bar-plots-1.qzv
```



## Phylogeny

In addition to classifying our organisms, we also want to reconstruct their phylogenetic relationships by generating a phylogenetic tree. We often assume that phylogenetic closeness can elucidate commonalities in phenotypic properties / functions, so it is worth examining. Moreover, we will need these trees later for alpha and beta diversity measures.

We can generate a phylogenetic tree using:

- de novo methods - constructed without a reference database
- reference based methods - uses a reference database
- **SEPP fragment insertion** (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5904434/>) will insert sequences into a reference phylogeny, with the caveat that sequences not at least 75% similar to any sequence in the tree is discarded. (This is a nice solution for meta analyses, but can be computationally intensive.)

For more details, refer to the **QIIME 2 documentation** (<https://docs.qiime2.org/2022.8/tutorials/phylogeny/?highlight=phylogeny>).

We will perform a de novo method here, using a pipeline available in the q2-phylogeny plugin (qiime phylogeny align-to-tree-mafft-fasttree).

This will run through the following steps:

- generates a multiple sequence alignment
- masks the alignment (removes errors, uninformative sites, repetitive regions, which can lose some resolution)

- builds a tree [de novo methods output an unrooted tree, which lacks directionality (no inferred ancestry)]
- root the tree (a rooted tree is required for alpha and beta diversity metrics) - the unrooted tree is rooted at its midpoint

Example of unrooted vs rooted (at midpoint):

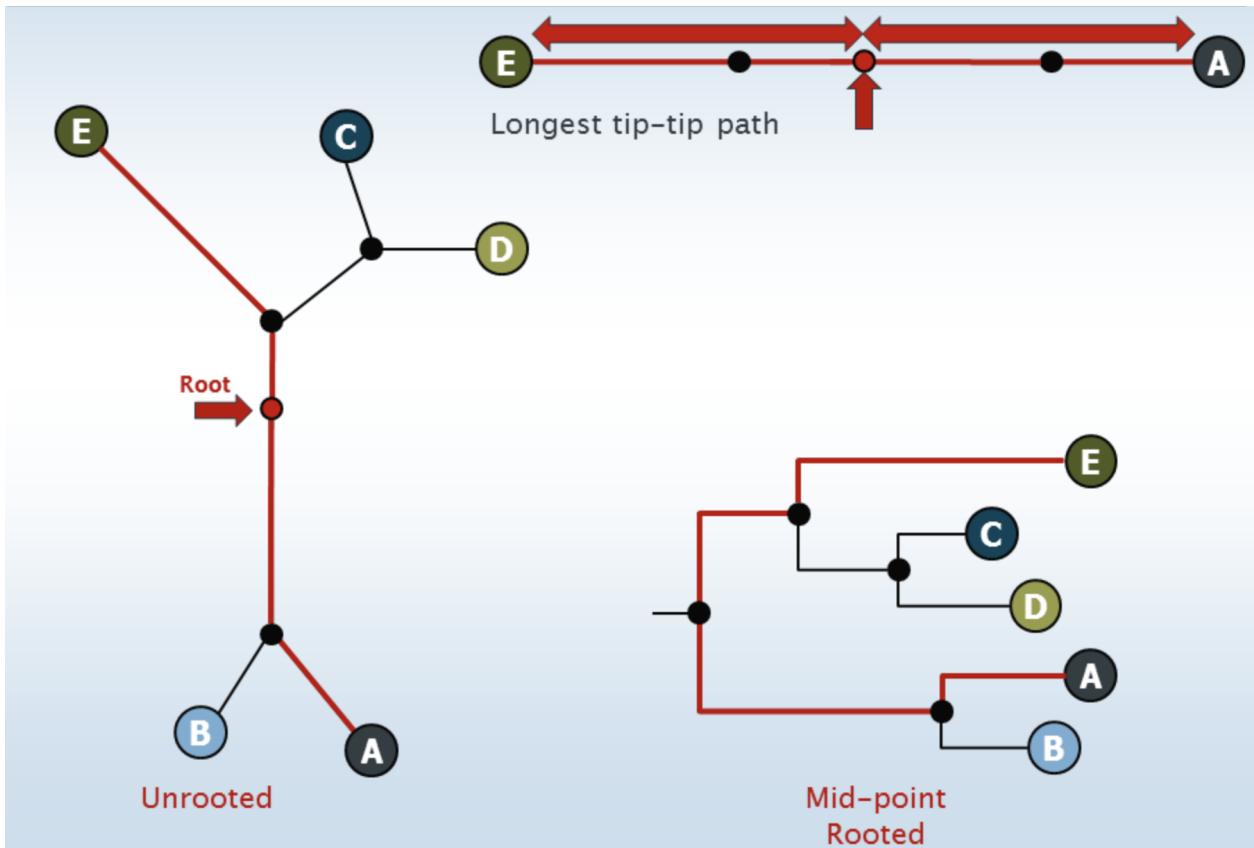


Image from *The Cabbages of Doom, How to root a phylogenetic tree* (<http://cabbagesofdoom.blogspot.com/2012/06/how-to-root-phylogenetic-tree.html>)

Let's run the pipeline:

```
qiime phylogeny align-to-tree-mafft-fasttree \
--i-sequences filtered-sequences-2.qza \
--output-dir phylogeny-align-to-tree-mafft-fasttree
```

There is no visual output to view here.

# Lesson 5: Microbial diversity, alpha rarefaction, alpha diversity

## Learning Objectives

1. Understand the difference between alpha and beta diversity
2. Introduce several alpha diversity metrics
3. Understand what rarefaction is and why it is important
4. Introduce the debate regarding rarefaction and other methods of normalization

Often many questions related to the microbiome center on ecological diversity. How many and what types of microbes are in a sample and how does this compare to other samples? Microbial diversity is often sensitive to environmental disturbances (e.g., trauma, medication, sanitation, diet, etc.).

There are two primary types of diversity explored in microbial ecology: alpha diversity and beta diversity.

## What is alpha diversity?

Alpha diversity is within sample diversity. When exploring alpha diversity, we are interested in the distribution of microbes within a sample or metadata category. This distribution not only includes the number of different organisms (richness) but also how evenly distributed these organisms are in terms of abundance (evenness).

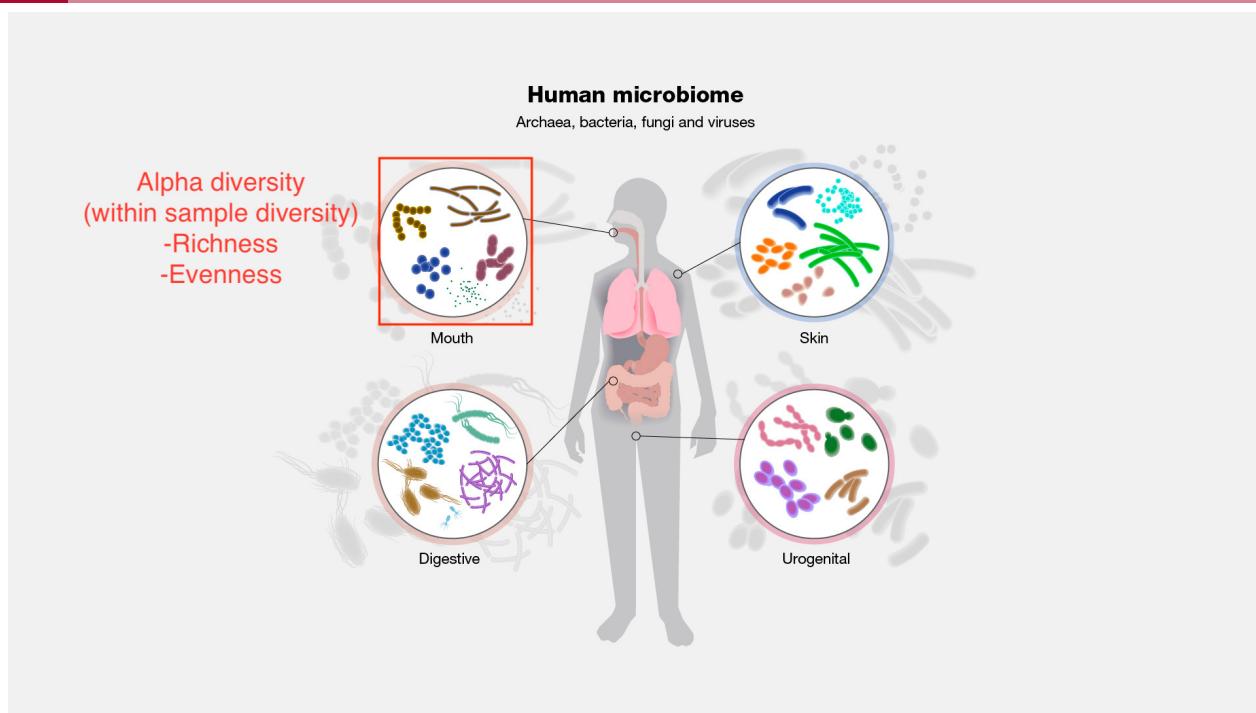


Image modified from <https://www.genome.gov/genetics-glossary/Microbiome> (<https://www.genome.gov/genetics-glossary/Microbiome>)

In addition, some diversity metrics include a phylogenetic component (i.e., Faith's phylogenetic diversity). The logic behind a phylogenetic metric is that a sample comprised of some number of highly related organisms, for example, all from the same genus, is not as diverse as a sample comprised of organisms with greater phylogenetic distances (for example, organisms from different phyla or even different domains).

Alpha diversity methods include information on either richness, evenness, or both. Here are a few examples:

**Richness:** High richness equals more ASVs or more phylogenetically dissimilar ASVs in the case of Faith's PD.

Observed OTUs/ASVs

Faith's PD (Sum of branch lengths)

**Both:** Diversity increases as richness and evenness increase.

Simpson's Dominance or Gini-Simpson (biased toward dominant species, meaning it is impacted more by evenness; values 0-1)

Shannon (treats rare and abundant more equitably than Gini-Simpson; values mostly from 0-10, typically 1-3.5)

**Evenness:** High values suggest more equal numbers between species.

Pileou's Evenness - calculated from Shannon (values 0-1)

Simpson's Evenness

In QIIME2, alpha diversity metrics are computed using scikit-bio; here is a link (<http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.html>) to the metrics with included descriptions.

Additionally, here (<https://www.davidzeleny.net/anadat-r/doku.php/en:div-ind>) is a nice resource comparing some prominent alpha diversity indices.

It is good practice to report more than one metric, since each metric can be interpreted slightly differently.

## Beta diversity (More on this in Lesson 6)

Beta diversity is between sample diversity. This is useful for answering the question, how different are these microbial communities?

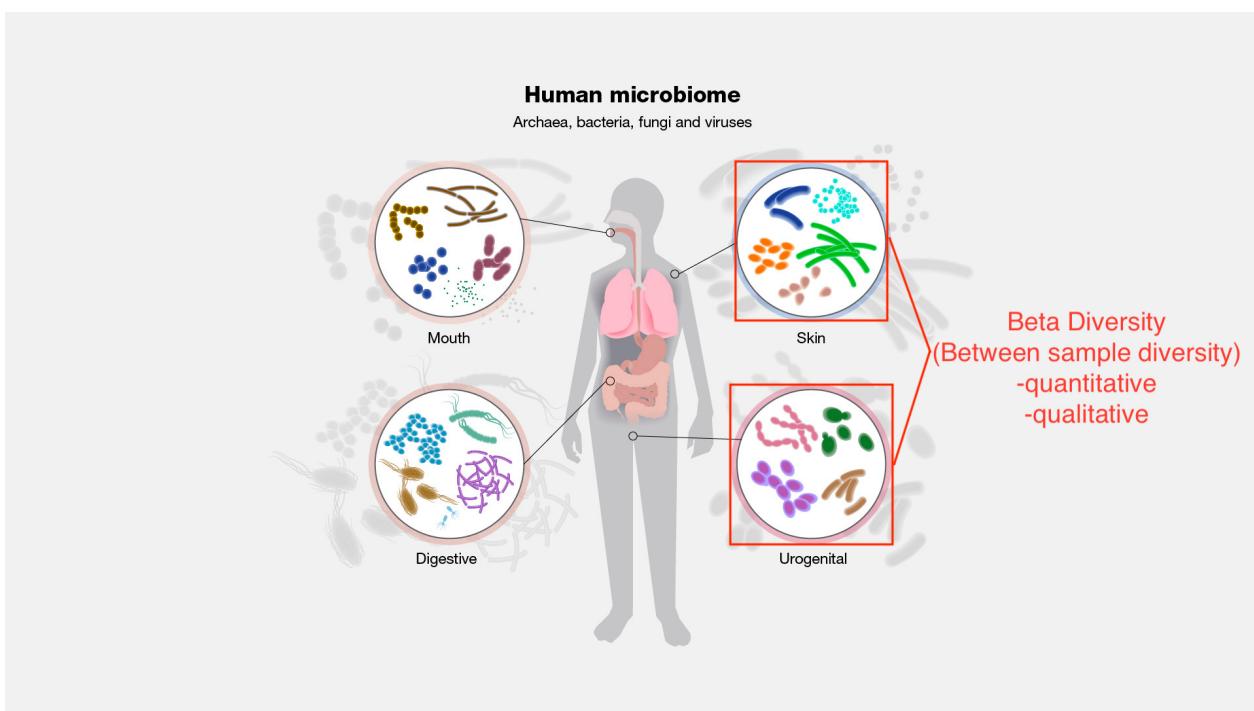


Image modified from <https://www.genome.gov/genetics-glossary/Microbiome> (<https://www.genome.gov/genetics-glossary/Microbiome>)

We will look into specific metrics of beta diversity in Lesson 5.

## Rarefaction

To rarefy or not to rarefy?

Feature tables are composed of sparse and compositional (<https://www.frontiersin.org/articles/10.3389/fmicb.2017.02224/full>) data. Measuring microbial diversity using 16S rRNA sequencing is dependent on sequencing depth. By chance, a sample that is more deeply sequenced is more likely to exhibit greater diversity than a sample with a low sequencing depth.

While there is a debate in the field about the most appropriate way to normalize data (See [here](https://pubmed.ncbi.nlm.nih.gov/24699258/) (<https://pubmed.ncbi.nlm.nih.gov/24699258/>) and [here](https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-017-0237-y) (<https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-017-0237-y>)) prior to downstream analyses, rarefaction is still used as a primary method for correcting differences in read depth prior to diversity analyses and is the included method used in the core diversity metrics pipeline in QIIME 2.

Note: you may want to skip rarefaction if library sizes are fairly even. Rarefaction is more beneficial when there is a greater than ~10x difference in library size (Weiss et al. 2017).

## What is rarefaction?

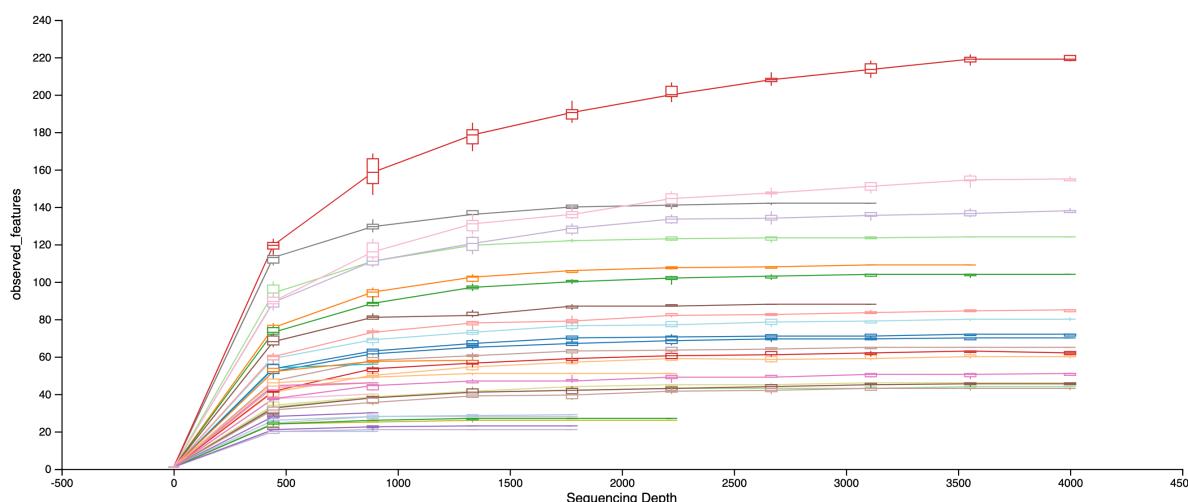
Rarefaction is the process of subsampling reads without replacement to a defined sequencing depth, thereby creating a standardized library size across samples. Any sample with a total read count less than the defined sequencing depth used to rarefy will be discarded. Post-rarefaction all samples will have the same read depth. How do we determine the magic sequencing depth at which to rarefy? We typically use an alpha **rarefaction curve**.

## Selecting a read depth to rarefy

A rarefaction curve

plot[s] the number of counts sampled (rarefaction depth) vs. the expected value of species diversity. --- Weiss et al. 2017 (<https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-017-0237-y>)

Let's take a look at an alpha **rarefaction curve** (<https://view.qiime2.org/?type=html&src=https%3A%2F%2Fdocs.qiime2.org%2F2021.8%2Fdata%2Ftutorials%2Fmoving-pictures%2Falpha-rarefaction.qzv>).



Demo plot from [view.qiime2.org](https://view.qiime2.org/) (<https://view.qiime2.org/>).

As the sequencing depth increases, you recover more and more of the diversity observed in the data. At a certain point (read depth), diversity will stabilize, meaning the diversity in the data has been fully captured. This point of stabilization will result in the diversity measure of interest plateauing.

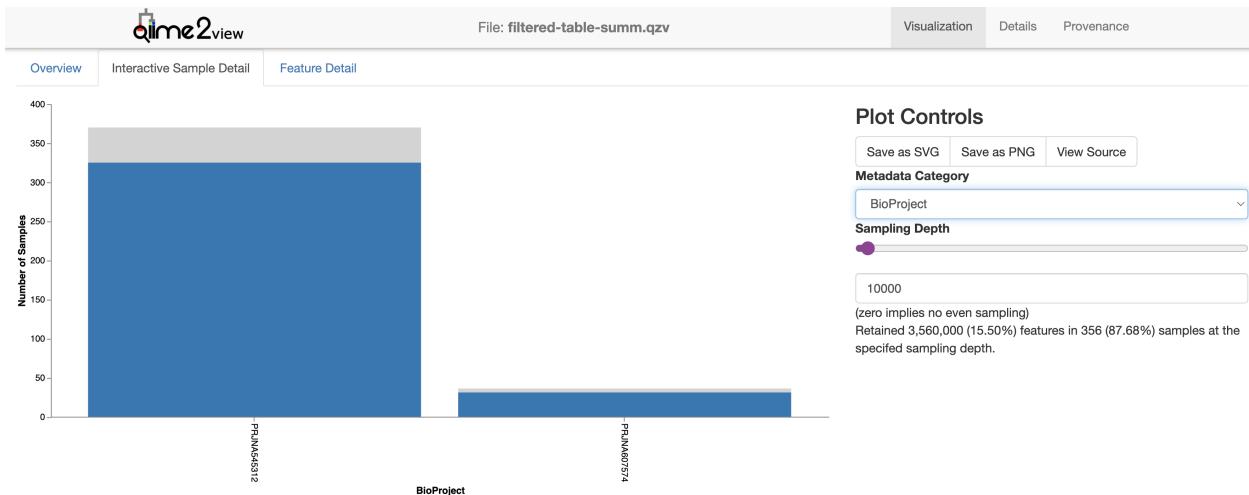
Let's create an alpha rarefaction plot.

```
qiime diversity alpha-rarefaction \
--i-table filtered-table-3.qza \
--i-phylogeny /data/cancer_data_catchup/phylogeny-align-to-tree-ma \
--m-metadata-file /data/sample-metadata.tsv \
--p-max-depth 33000 \
--o-visualization alpha-rarefaction-plot.qzv
```

You can adjust the metrics produced as well as other parameters regarding the number of sequencing depths to include, the minimum rarefaction depth, and the number of rarefied tables to compute at each depth (10 by default). See the help documentation for more information.

```
qiime diversity alpha-rarefaction --help
```

We can use this plot in combination with our feature table summary to decide at which depth we would like to rarefy. We need to choose a sequencing depth at which the diversity in most of the samples has been captured and most of the samples have been retained.



At a sampling depth of 10,000, we lose ~50 samples and retain 87% of samples, while at 15,000, we only retain ~80% of samples.

We can also check out the stability of our beta diversity metrics at a given sequencing depth using `qiime diversity beta-rarefaction`, which produces an "Emperor jackknifed PCoA plot, samples clustered by UPGMA or neighbor joining with support calculation, and a

heatmap showing the correlation between rarefaction trials of that beta diversity metric" (<https://docs.qiime2.org/2022.8/plugins/available/diversity/beta-rarefaction/>).

*Rarefying is generally applied only to diversity analyses, and many of the methods in QIIME 2 will use plugin specific normalization methods (e.g., q2-breakaway (<https://github.com/statdivlab/q2-breakaway>), ANCOM (<https://docs.qiime2.org/2022.8/plugins/available/composition/ancom/>), ALDEX2 (<https://library.qiime2.org/plugins/q2-aldex2/24/>)).*

## Core metrics phylogenetic

We will produce a number of core diversity metrics (alpha and beta) using a QIIME 2 pipeline, `qiime diversity core-metrics-phylogenetic`.

The parameters we need to know include the path to our rooted tree (`--i-phylogeny`), the path to our feature table (`--i-table`), the sampling depth at which we would like to rarefy (`--p-sampling-depth`), the path to the sample information (`--m-metadata-file`), and the name of the directory we would like to save our results to (`--output-dir`). If you do not have a tree, or you are not interested in phylogenetic diversity metrics, you can also use `qiime diversity core-metrics`. We can speed up this command by including the `--p-n-jobs-or-threads` parameter.

```
qiime diversity core-metrics-phylogenetic \
--i-phylogeny /data/cancer_data_catchup/phylogeny-align-to-tree-main.qza \
--i-table filtered-table-3.qza \
--p-sampling-depth 10000 \
--p-n-jobs-or-threads 3 \
--m-metadata-file /data/sample-metadata.tsv \
--output-dir diversity-core-metrics-phylogenetic
```

Let's take a look at the output.

```
ls -l diversity-core-metrics-phylogenetic
```

You should see something like this:

```
emmonsal@NCI-02243046-ML ~/Documents/CourseResources/QIIME2/amplicon_course/downstream/diversity-core-metrics-phylogenetic
$ ls -lth
total 51432
-rw-r--r-- 1 emmonsal NIH\Domain Users  1.9M Oct 28 14:46 bray_curtis_emperor.qzv
-rw-r--r-- 1 emmonsal NIH\Domain Users  1.9M Oct 28 14:46 jaccard_emperor.qzv
-rw-r--r-- 1 emmonsal NIH\Domain Users  1.9M Oct 28 14:46 weighted_unifrac_emperor.qzv
-rw-r--r-- 1 emmonsal NIH\Domain Users  1.9M Oct 28 14:46 unweighted_unifrac_emperor.qzv
-rw-r--r-- 1 emmonsal NIH\Domain Users  1.5M Oct 28 14:46 bray_curtis_pcoa_results.qza
-rw-r--r-- 1 emmonsal NIH\Domain Users  2.0M Oct 28 14:46 jaccard_pcoa_results.qza
-rw-r--r-- 1 emmonsal NIH\Domain Users  1.4M Oct 28 14:46 weighted_unifrac_pcoa_results.qza
-rw-r--r-- 1 emmonsal NIH\Domain Users  1.5M Oct 28 14:46 unweighted_unifrac_pcoa_results.qza
-rw-r--r-- 1 emmonsal NIH\Domain Users  1.1M Oct 28 14:46 bray_curtis_distance_matrix.qza
-rw-r--r-- 1 emmonsal NIH\Domain Users  1.4M Oct 28 14:46 jaccard_distance_matrix.qza
-rw-r--r-- 1 emmonsal NIH\Domain Users  1.9M Oct 28 14:46 weighted_unifrac_distance_matrix.qza
-rw-r--r-- 1 emmonsal NIH\Domain Users  1.9M Oct 28 14:46 unweighted_unifrac_distance_matrix.qza
-rw-r--r-- 1 emmonsal NIH\Domain Users  924K Oct 28 14:46 evenness_vector.qza
-rw-r--r-- 1 emmonsal NIH\Domain Users  924K Oct 28 14:46 shannon_vector.qza
-rw-r--r-- 1 emmonsal NIH\Domain Users  921K Oct 28 14:46 observed_features_vector.qza
-rw-r--r-- 1 emmonsal NIH\Domain Users  924K Oct 28 14:46 faith_pd_vector.qza
-rw-r--r-- 1 emmonsal NIH\Domain Users  1.1M Oct 28 14:46 rarefied_table.qza
(giime2-2022_8)
```

For alpha diversity, core-metrics-phylogenetic returns an evenness\_vector.qza, shannon\_vector.qza, observed\_features\_vector.qza, and faith\_pd\_vector.qza.

## Alpha diversity comparison

Let's remember back to the design of the study we are examining (*Reconstitution of the gut microbiota of antibiotic-treated patients by autologous fecal microbiota transplant* (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6468978/>)).

This study included a randomized controlled longitudinal trial involving 25 patients (14 received auto-FMT following allogeneic stem cell transplantation and 11 did not). Antibiotics were given prior to the transplant to avoid serious infection, but a loss in microbial gut diversity can consequently leads to adverse health outcomes following transplantation. This study showed that auto-FMT could restore gut microbial diversity and composition to a pre-transplantation state.

To assess differences in alpha diversity by metadata category, we can use `qiime diversity alpha-group-significance`.

```
qiime diversity alpha-group-significance \
--i-alpha-diversity diversity-core-metrics-phylogenetic/observed_f
--m-metadata-file /data/sample-metadata.tsv \
--o-visualization alpha-group-sig-obs-feats.qzv
```

This reports kruskal-wallis results and kruskal-wallis pairwise results with Benjamini and Hochberg FDR corrected p-values (q-value in this case). If we are interested in assessing the relationship between an alpha diversity metric and a continuous variable, we could use `qiime diversity alpha-correlation`, which applies a Spearman correlation.

We notice there is a lot of variation within an individual, but because we are examining longitudinal assumptions, this type of test is inappropriate because it violates the assumption that data are independent.

## Q2-longitudinal

Luckily, there is a [q2-longitudinal plugin](https://docs.qiime2.org/2022.8/plugins/available/longitudinal/) (<https://docs.qiime2.org/2022.8/plugins/available/longitudinal/>) to handle dependent longitudinal data.

The q2-longitudinal plugin includes:

- interactive plotting (e.g., 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 selection

For example, we can use a linear mixed effects model to better explore changes in richness related to the timing of the bone marrow transplant.

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.

The linear-mixed-effects action in q2-longitudinal uses statsmodels' "mixedlm" function to compute LME models. --- [Bokulich et al. 2018](https://journals.asm.org/doi/10.1128/mSystems.00219-18) (<https://journals.asm.org/doi/10.1128/mSystems.00219-18>)

LMEs take into account both fixed and random effects, where fixed effects are similar to factor levels and random effects represent sources of unknown variation. For example, here, we can set `PatientID` as a random effect; each patient has a different starting microbiome. Note: "a random intercept for each individual is set by default" ([plugin description](https://docs.qiime2.org/2022.8/plugins/available/longitudinal/linear-mixed-effects/) (<https://docs.qiime2.org/2022.8/plugins/available/longitudinal/linear-mixed-effects/>)).

Check out available parameters:

```
qiime longitudinal linear-mixed-effects
```

Note: we can specify two input metadata files to merge metadata information.

Let's see how this works before we run the model.

```
qiime metadata tabulate \
--m-input-file /data/sample-metadata.tsv diversity-core-metrics-ph \
--o-visualization merged_meta_alpha_summ.qzv
```

Let's run the LME.

```
qiime longitudinal linear-mixed-effects \
--m-metadata-file /data/sample-metadata.tsv diversity-core-metrics \
--p-state-column DayRelativeToNearestHCT \
--p-individual-id-column PatientID \
--p-metric observed_features \
--o-visualization lme-obs-features-HCT.qzv
```

And, we can look at the impact of the auto-FMT.

```
qiime longitudinal linear-mixed-effects \
--m-metadata-file /data/sample-metadata.tsv diversity-core-metrics \
--p-state-column day-relative-to-fmt \
--p-individual-id-column PatientID \
--p-metric observed_features \
--o-visualization lme-obs-features-FMT.qzv
```

Let's check out the qzv files produced by these commands. As always, visualization files should be moved to the `~/public` directory. See the [Cancer Microbiome Intervention Tutorial \(\*https://docs.qiime2.org/jupyterbooks/cancer-microbiome-intervention-tutorial/030-tutorial-downstream/060-alpha-diversity.html\*\)](https://docs.qiime2.org/jupyterbooks/cancer-microbiome-intervention-tutorial/030-tutorial-downstream/060-alpha-diversity.html) for a description of the results.

## Optional filtering of samples

If we want to retain only the samples included in our diversity analyses, we can use `qiime feature-table filter-samples` to drop samples with read depths less than 10,000.

```
qiime feature-table filter-samples \
--i-table filtered-table-3.qza \
--p-min-frequency 10000 \
--o-filtered-table filtered-table-4.qza
```

Note: I generally use qiime2 for upstream processing (denoising, classification, tree building) and then use R for the remainder of my analyses. R is a language and statistical computing environment, and there are many different packages that are specific to microbiome analysis (e.g., phyloseq, microbiomeSeq).

# Lesson 6 .

## Learning Objectives

1. Introduce several beta diversity metrics
2. Discover different ordination methods
3. Learn about statistical methods that are applicable

## Beta diversity

Beta diversity is between sample diversity. This is useful for answering the question, how different are these microbial communities?

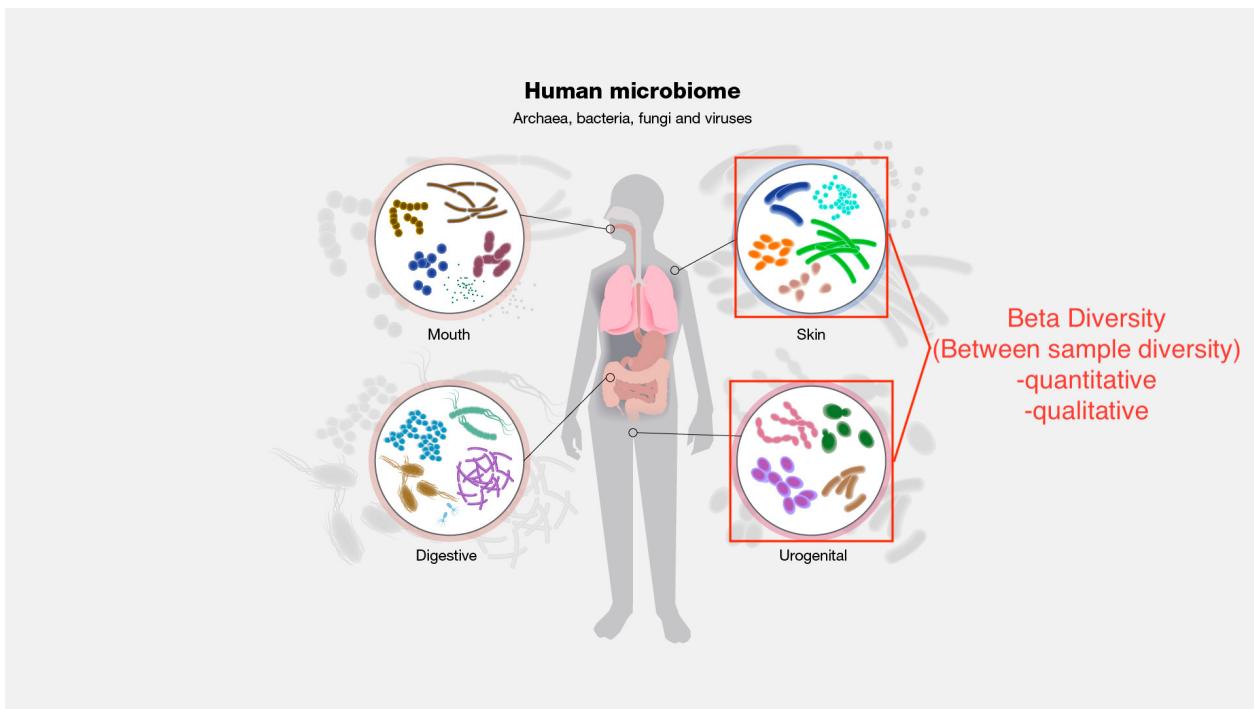


Image modified from <https://www.genome.gov/genetics-glossary/Microbiome> (<https://www.genome.gov/genetics-glossary/Microbiome>)

Beta diversity is measured using distance and dissimilarity metrics. The core-metrics-phylogenetic pipeline automatically produces Bray-Curtis, Jaccard, weighted UniFrac, and unweighted UniFrac. More on these below.

## Distance and dissimilarity metrics

### Bray-Curtis dissimilarity

- quantitative
- Takes into consideration abundance and presence absence

### Jaccard

- qualitative - presence / absence - percentage of taxa not found in both samples (<https://forum.qiime2.org/t/pca-vs-pcoa-which-is-the-appropriate-one-for-microbiome-data/5974/6>)

### Weighted UniFrac

- quantitative
- similar to Bray-Curtis but takes into consideration phylogenetic relationships

### Unweighted UniFrac

- qualitative
- like Jaccard focuses on presence / absence of taxa but also includes phylogenetic relationships
- percentage of phylogenetic branch length not found in both samples (<https://forum.qiime2.org/t/pca-vs-pcoa-which-is-the-appropriate-one-for-microbiome-data/5974/6>)

### Aitchison

- an answer to the compositional nature of the data
- "euclidean distances between clr-transformed compositions" (Quinn et al. 2018) (<https://academic.oup.com/bioinformatics/article/34/16/2870/4956011>).
- a clr transformation sets the features in a data set relative to the geometric mean of the composition

{}{Sdet}{}{Esum}{{}}

What is compositional data?{}{Esum}{{}}

Compositional data have two unique properties. First, the total sum of all component values (i.e. the library size) is an artifact of the sampling procedure (van den Boogaart and Tolosana-Delgado, 2008). Second, the difference between component values is only meaningful proportionally [e.g. the difference between 100 and 200 counts carries the same information as the difference between 1000 and 2000 counts (van den Boogaart and Tolosana-Delgado, 2008)].--- Quinn et al. 2018 (<https://academic.oup.com/bioinformatics/article/34/16/2870/4956011>)

See this paper (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5695134/>) for more information on compositional data.

{}{Edet}{}{Esum}{{}}

Some other notable metrics are described [here](https://docs.onecodex.com/en/articles/4150649-beta-diversity) (<https://docs.onecodex.com/en/articles/4150649-beta-diversity>).

*These methods result in large distance / dissimilarity matrices. In all methods, a value closer to zero indicates similarity between microbial communities, while a value closer to one indicates dissimilarity.*

## Beta rarefaction

Again, rarefaction is used to eliminate issues due to differences in library size prior to beta diversity. This method is built-in to QIIME 2 core metrics pipelines. We can examine the stability of a beta diversity metric using `qiime diversity beta-rarefaction`.

```
qiime diversity beta-rarefaction \
--i-table filtered-table-3.qza \
--p-metric braycurtis \
--p-clustering-method nj \
--p-sampling-depth 10000 \
--m-metadata-file /data/sample-metadata.tsv \
--o-visualization braycurtis-rarefaction-plot.qzv
```

This will rarefy your feature table multiple times at a given depth. The output provides a jackknifed emperor plot, with variability around a community represented by the ellipsoids around a point. A correlation heatmap and a UPGMA/NJ sample-clustering tree is also output.

## Ordination methods

Methods to reduce dimensionality in the data and visualize trends in the data. The following list includes commonly used methods and is not exhaustive.

### PCoA

- most common
- similar to PCA but works on distance metrics beyond euclidean
- maximizes linear correlation
- prone to the [horseshoe effect](https://journals.asm.org/doi/10.1128/mSystems.00166-16) (<https://journals.asm.org/doi/10.1128/mSystems.00166-16>)  
(also observed in PCA)

### UMAP (Uniform Manifold Approximation and Projection)

- non-linear
- can be used on multiple distance / dissimilarity metrics
- improved resolution in clusters
- More information [here](https://journals.asm.org/doi/10.1128/mSystems.00691-21) (<https://journals.asm.org/doi/10.1128/mSystems.00691-21>).

## NMDS (Not available in QIIME 2)

- better for rank ordered data (e.g., Bray-Curtis)
- dimensions are specified
- stress indicates how well the ordination represents the data (stress < 0.1 ~ good)
- no single solution

See this resource (<http://ordination.okstate.edu/overview.htm>) for more information on ordination metrics.

## Generating a PCoA and UMAP in QIIME2

### PCoA

PCoA was included by default in our `core-metrics-phylogenetic` pipeline. Because these are longitudinal data, we will customize the axis to include the variable, `week-relative-to-hct`.

```
qiime emperor plot \
--i-pcoa diversity-core-metrics-phylogenetic/unweighted_unifrac_pcoa.qza \
--m-metadata-file /data/sample-metadata.tsv diversity-core-metrics-phylogenetic \
--p-custom-axes week-relative-to-hct \
--o-visualization uu-pcoa-emperor-w-time.qzv

qiime emperor plot \
--i-pcoa diversity-core-metrics-phylogenetic/weighted_unifrac_pcoa.qza \
--m-metadata-file /data/sample-metadata.tsv diversity-core-metrics-phylogenetic \
--p-custom-axes week-relative-to-hct \
--o-visualization wu-pcoa-emperor-w-time.qzv
```

### UMAP

First, we perform the ordination.

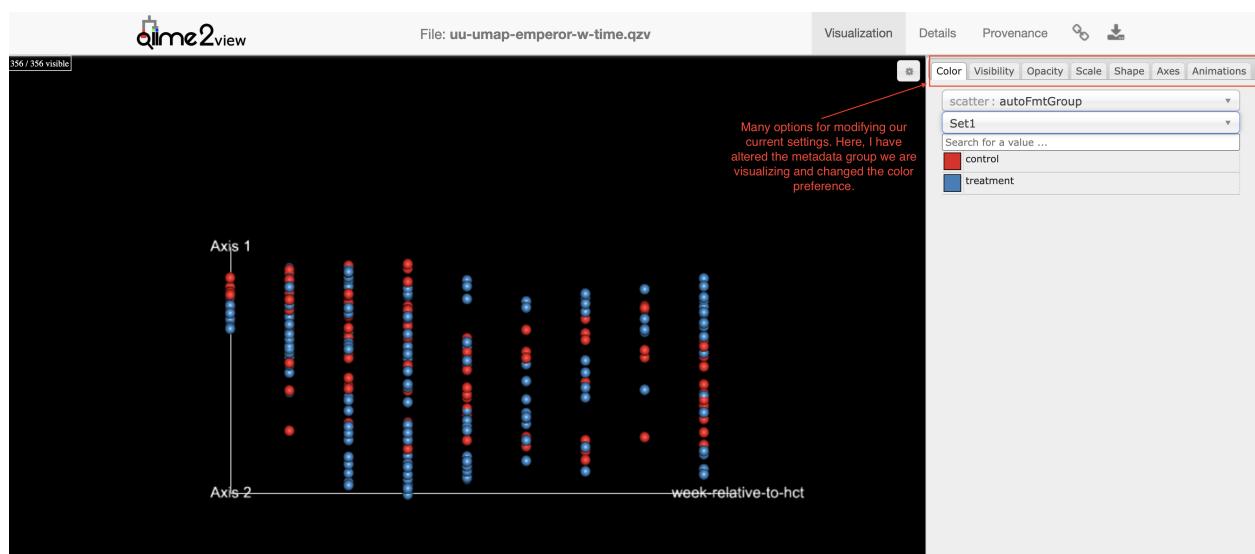
```
qiime diversity umap \
--i-distance-matrix diversity-core-metrics-phylogenetic/unweighted_unifrac.qza \
--o-umap uu-umap.qza

qiime diversity umap \
--i-distance-matrix diversity-core-metrics-phylogenetic/weighted_unifrac.qza \
--o-umap wu-umap.qza
```

Then we use emperor to plot. Though the input parameter is `--i-pcoa`, we can also input umap results.

```
qiime emperor plot \
--i-pcoa uu-umap.qza \
--m-metadata-file /data/sample-metadata.tsv diversity-core-metrics \
--p-custom-axes week-relative-to-hct \
--o-visualization uu-umap-emperor-w-time.qzv
qiime emperor plot \
--i-pcoa wu-umap.qza \
--m-metadata-file /data/sample-metadata.tsv diversity-core-metrics \
--p-custom-axes week-relative-to-hct \
--o-visualization wu-umap-emperor-w-time.qzv
```

When we view these files, there are many options for customizing our plot and toggling our view. Let's look at these in more detail.



UMAP (unweighted UniFrac): A 2-D representation of axis-1 vs week-relative-to-hct. We can toggle our view to see this in 3D.

Longitudinal trends are difficult to view here because the data are dependent, but these can be teased apart in greater detail using the q2-longitudinal plugin.

Let's take a look at the moving pictures data set ([https://view.qiime2.org/?src=https%3A%2F%2Fdocs.qiime2.org%2F2022.8%2Fdata%2Ftutorials%2Fmoving-pictures%2Fcore-metrics-results%2Funweighted\\_unifrac\\_emperor.qzv](https://view.qiime2.org/?src=https%3A%2F%2Fdocs.qiime2.org%2F2022.8%2Fdata%2Ftutorials%2Fmoving-pictures%2Fcore-metrics-results%2Funweighted_unifrac_emperor.qzv)) for a clearer example.

## Statistics

Some typical statistical tests applied to beta diversity metrics include the following:

Adonis (PERMANOVA)

- Similar to a MANOVA, but is permutational and non-parametric.
- Sensitive to group dispersion, so it is worth running alongside a beta-dispersion method.

- generates a pseudo-F ratio; larger pseudo-F suggests larger group separation.
- requires data independence

### ANOSIM (Analysis of Similarity)

- uses a ranked approach (complementary to NMDS)
- The ANOSIM statistic compares the mean of ranked dissimilarities between groups to the mean of ranked dissimilarities within groups. An R value close to "1.0" suggests dissimilarity between groups while an R value close to "0" suggests an even distribution of high and low ranks within and between groups. R values below "0" suggest that dissimilarities are greater within groups than between groups. --- [gustame documentation \(\*https://sites.google.com/site/mb3gustame/hypothesis-tests/anosim\*\)](https://sites.google.com/site/mb3gustame/hypothesis-tests/anosim)
- Also sensitive to differences in group dispersion.

These methods, including a permutational dispersion test, can be run using `qiime diversity beta-group-significance`. The PERMANOVA implementation here is one-way. To include more than one variable with potential interactions, use `qiime diversity adonis`.

Again, because we are looking at longitudinal data, these are not as relevant in this specific case.

# Lesson 7: Course Wrap-Up

## Learning Objectives

1. Introduce the QIIME2 microbiome workflow for Biowulf
2. Review key concepts
3. Showcase additional plugins

## QIIME 2 on Biowulf

As mentioned previously, QIIME 2 is installed on Biowulf.

To see available versions use

```
module avail qiime
```

Also, check out the [QIIME2 Biowulf help page \(<https://hpc.nih.gov/apps/QIIME.html#:~:text=QIIME2%20on%20Biowulf&text=QIIME%20is%20a%20powerful,quality%20figures>\)](https://hpc.nih.gov/apps/QIIME.html#:~:text=QIIME2%20on%20Biowulf&text=QIIME%20is%20a%20powerful,quality%20figures)

The default version on Biowulf is qiime2-2021.4, and the latest installed version is qiime2-2022.2.

If you are interested in a reproducible workflow to use on Biowulf, Samantha Chill, a bioinformatician with CCBR, created a workflow that is readily available from [github \(\[https://github.com/CCBR/BETP\\\_microbiome\\\_2022\]\(https://github.com/CCBR/BETP\_microbiome\_2022\)\)](https://github.com/CCBR/BETP_microbiome_2022).

# Review

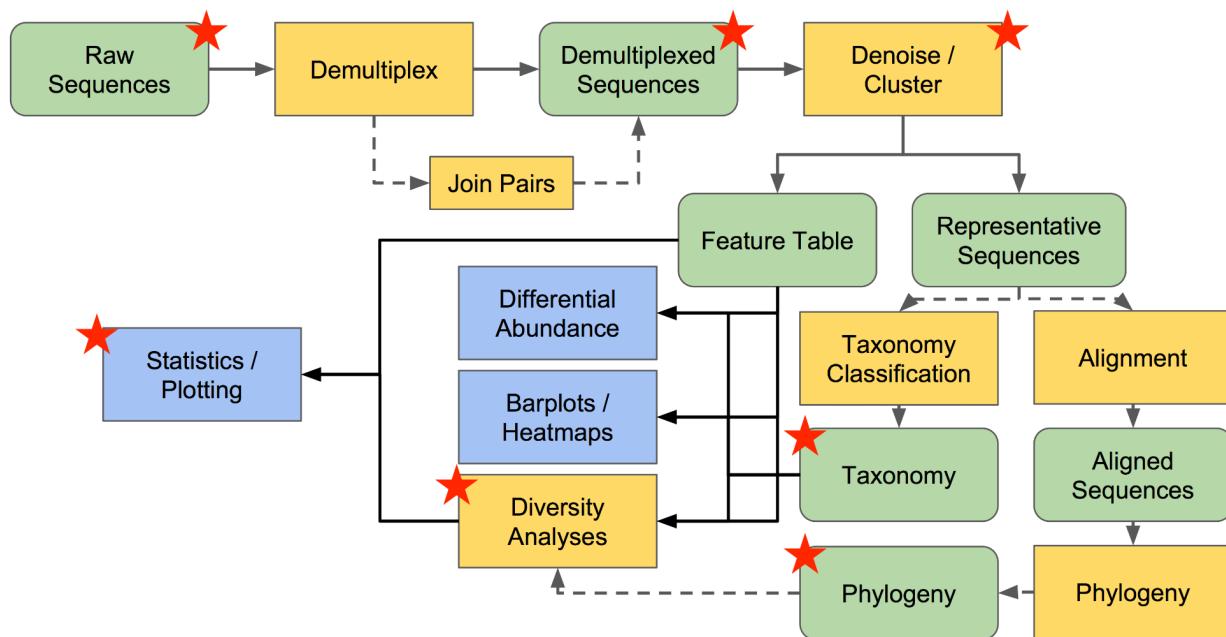


Image adapted from docs.qiime2.org (Conceptual Overview of QIIME 2).

## What have we done?

Using a small subset of data:

1. Imported raw fastq files using `qiime tools import`. Data was paired-end CASAVA format.
2. Checked for primers using `qiime cutadapt trim-paired`.
3. Denoised with `qiime dada2 denoise-paired` and generated summaries of our feature table and representative sequences.

Using the larger data set:

1. Filtered samples and features based on metadata categories and other thresholds.
2. Classified our sequences using a Greengenes trained (V4-V5) classifier and `qiime feature-classifier classify-sklearn`
3. Applied taxonomic filtering
4. Generated a de novo phylogenetic tree using `qiime phylogeny align-to-tree-mafft-fasttree`
5. Chose a rarefaction depth using the qiime 2 feature table summary and rarefaction curve
6. Generated several core alpha and beta diversity metrics and visualizations

These are our core steps, but let's also take a look at some of the other analysis plugins and methods available in QIIME 2.

# Other plugins of interest

## Differential abundance testing

Differential abundance testing examines which taxa are significantly different in abundance between conditions. However, challenges such as sparsity, compositionality, and library size differences make this challenging to determine.

### Methods in QIIME 2

**ANCOM** (Analysis of Composition of Microbiomes) (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4450248/>)

- additive log ratio approach
- assumes that less than 25 % of features change between groups
- q2-composition plugin (<https://docs.qiime2.org/2022.8/plugins/available/composition/>)
- Need to filter rare taxa
- w-statistic - the number of null hypotheses rejected

See the **Moving Pictures** tutorial (<https://docs.qiime2.org/2022.8/tutorials/moving-pictures/>).

**gneiss** (<https://journals.asm.org/doi/10.1128/mSystems.00162-16>)

- uses balance trees (isometric log-ratio transformation) (<https://docs.qiime2.org/2022.8/tutorials/gneiss/>)
- Need to filter rare taxa
- Check out this explanation (<https://www.youtube.com/watch?v=HAULM1WQkew>)

**ALDEx2** (<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0067019>)

- center log ratio approach (<https://library.qiime2.org/plugins/q2-aldex2/24/>)
- used for multiple types of NGS data (e.g., RNA-Seq)
- tested for QIIME 2 version 2019.7

Note: Different methods produce different results, and methods are sensitive to upstream processing decisions. **ALDEx2** and **ANCOM-II** were found to be more conservative but less sensitive than other methods, which means they are less prone to false positives (<https://www.nature.com/articles/s41467-022-28034-z>).

## ANCOM example

Let's use our practice data set to run ANCOM.

Step 1: Filter out low abundance / low prevalent ASVs. Note: this will shift the composition of the samples, and thus could bias results.

```
mkdir ancom

qiime feature-table filter-features \
--i-table /data/practice/04_filter/filtered-table3.qza \
--p-min-frequency 50 \
--p-min-samples 2 \
--o-filtered-table ancom/ancomfilt.qza
```

Step 2: Add pseudo-counts - This method does not tolerate zeros.

```
qiime composition add-pseudocount \
--i-table ancom/ancomfilt.qza \
--o-composition-table ancom/comp-table.qza
```

The only metadata category of interest in the data set is `DataType`, old vs young. Now, let's run `ancom`.

```
qiime composition ancom \
--i-table ancom/comp-table.qza \
--m-metadata-file /data/practice/metadata.txt \
--m-metadata-column DataType \
--o-visualization ancom/ancom-OY.qzv
```

## Core microbiome

If interested in highly prevalent taxa, you could use `qiime feature-table core-features`, which identifies "features observed in a user-defined fraction of the samples." By default, this will return features observed in at least 50% of samples.

## Random forest regression and classification

Can we use microbial community composition to predict a condition? For example, maybe we are interested in whether microbial community composition can predict a cancer state from a non-cancer state.

In QIIME 2, we could use the [q2-sample-classifier](https://docs.qiime2.org/2022.8/tutorials/sample-classifier/) (<https://docs.qiime2.org/2022.8/tutorials/sample-classifier/>), which uses supervised learning (default = Random Forest classification).

Supervised learning classifiers predict the categorical metadata classes of unlabeled samples by learning the composition of labeled training samples. --- <https://docs.qiime2.org/2022.8/tutorials/sample-classifier/> (<https://docs.qiime2.org/2022.8/tutorials/sample-classifier/>).

## Random Forest example

Let's use our practice data set again, and see if we can predict group membership (old vs young) by microbial composition. We will use the `sample-classifier` pipeline. This pipeline splits our data into training and testing sets, trains the model using the `--p-estimator` of choice, performs k-fold cross-validation (5 by default), tests the model on the test set, and calculates model accuracy by comparing true values versus predicted values of the test set.

```
qiime sample-classifier classify-samples \
--i-table /data/practice/04_filter/filtered-table3.qza \
--m-metadata-file /data/practice/metadata.txt \
--m-metadata-column DataType \
--p-optimize-feature-selection \
--p-parameter-tuning \
--p-estimator RandomForestClassifier \
--p-random-state 123 \
--output-dir rforest
```

We can move our visualizations to `~/public` to view some of these outputs.

## Other notable plugins

[DEICODE](https://library.qiime2.org/plugins/deicode/19/) (<https://library.qiime2.org/plugins/deicode/19/>)

- compositional beta diversity with biplots
- performs a Robust Aitchison PCA

[q2-clawback](https://library.qiime2.org/plugins/q2-clawback/7/) (<https://library.qiime2.org/plugins/q2-clawback/7/>)

- can improve taxonomic classifications
- uses taxonomic weights based on environment

[q2-picrust2](https://library.qiime2.org/plugins/q2-picrust2/13/) (<https://library.qiime2.org/plugins/q2-picrust2/13/>)

- functional prediction from 16S rRNA data

[q2-sidle](https://library.qiime2.org/plugins/q2-sidle/35/) (<https://library.qiime2.org/plugins/q2-sidle/35/>)

- a new implementation of the Short MUltiple Regions Framework (SMURF) (<https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-017-0396-x>)

[provenance\\_lib](https://library.qiime2.org/plugins/provenance_lib/43/) ([https://library.qiime2.org/plugins/provenance\\_lib/43/](https://library.qiime2.org/plugins/provenance_lib/43/))

- provenance replay (<https://forum.qiime2.org/t/provenance-replay-alpha-release-and-tutorial/23279>)
- generate reproducible code based on your QIIME 2 inputs and outputs

## Exporting results

Many of the QIIME 2 visualizations allow you to directly download results in a tab-delimited format. You can also unzip any QIIME 2 visualization (.qzv) or QIIME 2 artifact (.qza) and access data in the data directory.

There is also an [export method](https://docs.qiime2.org/2022.8/tutorials/exporting/) (<https://docs.qiime2.org/2022.8/tutorials/exporting/>) in QIIME 2 (`qiime tools export`). Let's export a feature table.

First, let's simply unzip a feature table artifact.

```
unzip -d filtered-table3 filtered-table-3.qza
```

Now, let's use `qiime tools export`.

```
qiime tools export \
--input-path filtered-table-3.qza \
--output-path exported-feature-table-3
```

## Working in R

There are many packages available to work with microbiome data in R. While there is an R API in the works for QIIME 2, for now, users can use the R package, [qiime2R](https://github.com/jbisanz/qiime2R) (<https://github.com/jbisanz/qiime2R>), to easily import QIIME 2 files.

[Phyloseq](http://joey711.github.io/phyloseq/) (<http://joey711.github.io/phyloseq/>) is a fantastic Bioconductor package for microbiome analysis with R, and qiime2R can import QIIME 2 files as phyloseq objects.

## Struggling with command line?

1. Try QIIME 2 Galaxy implementation
2. [Nephele](https://nephele.niaid.nih.gov/) (<https://nephele.niaid.nih.gov/>)
3. [MicrobiomeAnalyst](https://www.microbiomeanalyst.ca/) (<https://www.microbiomeanalyst.ca/>)

If you have any questions about your microbiome analysis, do not hesitate to email us at [ncibtep@nih.gov](mailto:ncibtep@nih.gov).

**Practice**

## Practice Lesson 2

For the help sessions, we will work on processing sequences generated in Zhang Z, Feng Q, Li M, Li Z, Xu Q, Pan X, Chen W. Age-Related Cancer-Associated Microbiota Potentially Promotes Oral Squamous Cell Cancer Tumorigenesis by Distinct Mechanisms. *Front Microbiol.* 2022 Apr 15;13:852566. doi: 10.3389/fmicb.2022.852566. PMID: 35495663; PMCID: PMC9051480. (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9051480/>)



[Front Microbiol.](#) 2022; 13: 852566.

PMCID: PMC9051480

Published online 2022 Apr 15. doi: [10.3389/fmicb.2022.852566](https://doi.org/10.3389/fmicb.2022.852566)

PMID: [35495663](https://pubmed.ncbi.nlm.nih.gov/35495663/)

### Age-Related Cancer-Associated Microbiota Potentially Promotes Oral Squamous Cell Cancer Tumorigenesis by Distinct Mechanisms

Zhen Zhang,<sup>1, 2, 3, 4, 5</sup> Qiang Feng,<sup>6, 7</sup> Meihui Li,<sup>6, 7</sup> Zhihui Li,<sup>1, 2, 3, 4, 5</sup> Qin Xu,<sup>1, 2, 3, 4, 5</sup> Xinhua Pan,<sup>1, 2, 3, 4, 5</sup> and Wantao Chen<sup>1, 2, 3, 4, 5, \*</sup>

This study examined differences in the oral microbiome of patients with oral squamous cell cancer. The goal was to determine whether the oral tumor microbiome in young patients was related to disease progression. While this study lacks controls that would make the authors' arguments stronger - for example, it would be nice to see tumor and non-tumor samples as well as healthy controls - the small sample size (20 young and 20 old) makes it fairly easy to reproduce. However, you should not consider this an example of a model experimental design.

Note: the authors did not make the sample information available beyond the sample ids. The metadata provided here resulted from inferring young vs old samples by sample name, as provided in the SRA, alone.

## Download the sequences and import for further processing with the QIIME2 platform.

The data is available in the Sequence Read Archive (BioProject PRJNA803155), so the first step is to grab the data from the SRA. For your convenience, we have also created a compressed archive of the sequence files (/data/practice/PRJNA803155.tar.gz).

Make a directory called Practice and unpack this file. You will also need to grab the accession list from the SRA in Step 1. You can skip step 2.

{{Sdet}}

Solution{{Esum}}

```
mkdir Practice
cd Practice
tar -xvf /data/practice/PRJNA803155.tar.gz
```

{{Edet}}

## Step 1: Get the run info from the SRA

According to the data availability statement, the data can be found in PRJNA803155.

Change to the Practice directory created above or make it now. Then make a new directory named raw\_data.

{{Sdet}}

Solution{{Esum}}

```
cd Practice
mkdir raw_data
```

{{Edet}}

Get the SRA Accession IDs using e-utilities or from NCBI's [Run Selector](https://0-www.ncbi-nlm-nih-gov.braininfo.ac.uk/Traces/study/) (<https://0-www.ncbi-nlm-nih-gov.braininfo.ac.uk/Traces/study/>). Save the file containing the accession IDs to `Practice/sra\_id.txt'.

{{Sdet}}

Solution{{Esum}}

```
esearch -db sra -query PRJNA803155 | efetch -format runinfo | cut -f
```

{{Edet}}

## Step 2: Download the data

Download the data using prefetch and fasterq-dump.

{{Sdet}}

Solution{{Esum}}

```
cd raw_data  
cat ./sra_id.txt | while read sra_id; do prefetch $sra_id; fasterq-dump -S $sra_id --gzip > ${sra_id}.fastq.gz; done
```

{{Edet}}

What format are the sequences in? How can you import them? See [this forum post](https://forum.qiime2.org/t/importing-and-demultiplexing-sequence-data-quick-reference/14002) (<https://forum.qiime2.org/t/importing-and-demultiplexing-sequence-data-quick-reference/14002>) for guidance.

### Step 3: Create the manifest

We will need to use a manifest file to import. See the Import tutorial (<https://docs.qiime2.org/2022.8/tutorials/importing/>). Note: The manifest file can be comma separated depending on the format that you use at import (<https://forum.qiime2.org/t/manifest-file-tsv-or-csv/23523>), despite what is written in the import tutorial. We will create the manifest file in our Practice directory.

{{Sdet}}

Solution{{Esum}}

```
cd ~/Practice  
  
echo "sample-id,absolute-filepath,direction" > q2_manifest.csv  
  
cat sra_id.txt | while read sra_id; do echo "$sra_id,$PWD/raw_data/$sra_id.fastq.gz,forward" >> q2_manifest.csv; done
```

{{Edet}}

### Step 4: Import

To import we will need to keep in mind that our samples are paired-end with quality information and that we are using a manifest format. Note: Phred 64 quality scores are associated with older data, so most data will have quality scores that are Phred 33. For more information on quality scores, see this [technical note](https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf) ([https://www.illumina.com/documents/products/technotes/technote\\_Q-Scores.pdf](https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf)) from Illumina.

We will save our imported data to a new directory named `01_import`.

{{Sdet}}

Solution{{Esum}}

```
mkdir 01_import  
qiime tools import \
```

```
--type 'SampleData[PairedEndSequencesWithQuality]' \
--input-path q2_manifest.csv \
--output-path 01_import/import.qza \
--input-format PairedEndFastqManifestPhred33
```

{{Edet}}

## Summarize import

{{Sdet}}

Solution{{Esum}}

```
qiime demux summarize \
--i-data 01_import/import.qza \
--o-visualization 01_import/import.qzv
```

{{Edet}}

Again, to view this file, you will need to move it to public.

Note: It is easier to create the comma separated manifest. However, as you have seen the recommended format is tab separated with the header.

```
sample-id    forward-absolute-filename    reverse-absolute-filename
```

To get this to work, you could use the following:

{{Sdet}}

Solution{{Esum}}

```
#Create tab delimited manifest

echo sample-id$'\t'forward-absolute-filename$'\t'reverse-absolute-filename$'\t'> manifest2.tsv

cat sra_id.txt | while read sra_id; do echo $sra_id$'\t'$PWD/raw_data/$sra_id$'\t'>> manifest2.tsv; done

#Import
qiime tools import \
--type 'SampleData[PairedEndSequencesWithQuality]' \
--input-path q2_manifest2.tsv \
--output-path 01_import/import2.qza \
--input-format PairedEndFastqManifestPhred33V2
```

{}{{E}det{}}

# Practice Lesson 3

This practice lesson is associated with Lesson 3 of the Microbiome Analysis with QIIME 2. In this practice lesson, we will work on generating a feature table and representative sequences. We will continue working with the data from [Zhang et al. 2022](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9051480/) (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9051480/>).

1. Change directories to Practice (cd Practice).
2. Check and trim primers and non-biological sequences. We will trim primers targeting V3-V4 (F: CCTACGGGNGGCWGCAG, R: GACTACHVGGGTATCTAATCC). You should output trimmed sequences to a directory called 02\_trim.

Note: Zhang et al. 2022 stated,

The universal primers 515F 5'-GTGCCAGCMGCCGCGG-3' and 907R 5'-CCGTCAATTCTTTRAGTTT-3' were applied to capture the V4–V5 region of 16S rDNA.

However, this was clearly not the case. If you use q2-cutadapt in combination with these primers, you will notice that the forward primer is found toward the center / ends of the reads and the reverse primer cannot be located.

{}{Sdet}}

Solution{}{Esum}}

```
mkdir 02_trim
qiime cutadapt trim-paired \
--i-demultiplexed-sequences 01_import/import.qza \
--p-front-f CCTACGGGNGGCWGCAG \
--p-front-r GACTACHVGGGTATCTAATCC \
--p-overlap 6 \
--p-discard-untrimmed \
--o-trimmed-sequences 02_trim/demux-trimmed.qza \
--verbose | tee 02_trim/cutadaptresults.log
```

{}{Edet}}

3. Create a new summary table.

{}{Sdet}}

Solution{}{Esum}}

```
qiime demux summarize \  
--i-data 02_trim/demux-trimmed.qza \  
--o-visualization 02_trim/demux-trimmed-summary.qza
```

{{Edet}}

4. Generate a feature table using DADA2 and save to **03\_denoise**.

{{Sdet}}

Solution{{Esum}}

```
mkdir 03_denoise  
# takes 30 minutes without multi-threading  
qiime dada2 denoise-paired \  
--i-demultiplexed-seqs 02_trim/demux-trimmed.qza \  
--p-trunc-len-f 0 \  
--p-trunc-len-r 0 \  
--p-n-threads 2 \  
--o-representative-sequences 03_denoise/asv-sequences.qza \  
--o-table 03_denoise/feature-table.qza \  
--o-denoising-stats 03_denoise/dada2-stats.qza
```

{{Edet}}

5. Summarize DADA2 stats using `qiime metadata tabulate`.

{{Sdet}}

Solution{{Esum}}

```
qiime metadata tabulate \  
--m-input-file 03_denoise/dada2-stats.qza \  
--o-visualization 03_denoise/dada2-stats-summ.qzv
```

{{Edet}}

What range of percentages of sequences were retained following quality filtering, denoising, merging, and removal of chimeric sequences? (Check out the last column in your stats summary).

{{Sdet}}

Solution{{Esum}}

66.58-87.64%

{{Edet}}

- Summarize your feature table and representative sequences. The path to the sample information (sample metadata) is /data/practice/metadata.txt.
6. {{Sdet}}

Solution{{Esum}}

```
qiime feature-table summarize \
--i-table 03_denoise/feature-table.qza \
--m-sample-metadata-file /data/practice/metadata.txt \
--o-visualization 03_denoise/feature-table-summ.qzv
qiime feature-table tabulate-seqs \
--i-data 03_denoise/asv-sequences.qza \
--o-visualization 03_denoise/asv-sequences-summ.qzv
```

{{Edet}}

Follow-up questions:

How many unique ASVs resulted?

{{Sdet}}

Solution{{Esum}}

6,881

{{Edet}}

Did any interesting patterns emerge in read frequencies by sample?

{{Sdet}}

Solution{{Esum}}

There is an interesting bifurcation in the read frequencies between old versus young samples. This could have either an underlying biological or technical explanation.

{{Edet}}

How can we grab a quick summary of our sample metadata?

{{Sdet}}

Hint{{Esum}}

Check out `qiime tools --help`

{{Edet}} {{Sdet}}

Solution{{Esum}}

```
qiime tools inspect-metadata /data/practice/metadata.txt
```

{{Edet}}

# Practice Lesson 4

This practice lesson is associated with Lesson 4 of the Microbiome Analysis with QIIME 2. In this practice lesson, we will work on filtering our feature table and representative sequences, classify our features, and generate a phylogenetic tree. We will continue working with the data from [Zhang et al. 2022](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9051480/) (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9051480/>).

1. Change directories to Practice (cd Practice).

2. Let's do some initial filtering.

1. Create a directory named 04\_filter to save filtered tables.

{Sdet}

Solution{Esum}

```
mkdir 04_filter
```

{Edet}

2. Filter any samples that do not have a total read frequency greater than 1,000.

{Sdet}

Solution{Esum}

```
qiime feature-table filter-samples \
--i-table 03_denoise/feature-table.qza \
--p-min-frequency 1000 \
--o-filtered-table 04_filter/filtered-table1.qza
```

{Edet}

3. Filter any features with a total abundance across all samples less than 10.

{Sdet}

Solution{Esum}

```
qiime feature-table filter-features \
--i-table 04_filter/filtered-table1.qza \
```

```
--p-min-frequency 10 \
--o-filtered-table 04_filter/filtered-table2.qza
```

}}Edet}}

3. Classify the features using a trained classifier (/data/practice/gg-13-8-99-V3V4-nb-classifier.qza) and generate a summary of the results. These should be saved to a new directory (**05\_taxonomy**). Feel free to try an alternative classification method.

}}Sdet}}

Solution}}Esum}}

```
mkdir 05_taxonomy
qiime feature-classifier classify-sklearn \
--i-classifier /data/practice/gg-13-8-99-V3V4-nb-classifier.qza
--i-reads 03_denoise/asv-sequences.qza \
--o-classification 05_taxonomy/taxonomy.qza

qiime metadata tabulate \
--m-input-file 05_taxonomy/taxonomy.qza \
--o-visualization 05_taxonomy/taxonomy.qzv
```

}}Sdet}}

Solution}}Esum}}

4. Filter out mitochondria, chloroplasts, and assignments only at the kingdom / domain level. Generate a bar plot.

}}Sdet}}

Solution}}Esum}}

```
qiime taxa filter-table \
--i-table 04_filter/filtered-table2.qza \
--i-taxonomy 05_taxonomy/taxonomy.qza \
--p-mode contains \
--p-include p__ \
--p-exclude 'p__,Chloroplast,Mitochondria' \
--o-filtered-table 04_filter/filtered-table3.qza

qiime taxa barplot \
--i-table 04_filter/filtered-table3.qza \
--i-taxonomy 05_taxonomy/taxonomy.qza \
--m-metadata-file /data/practice/metadata.txt \
--o-visualization 05_taxonomy/taxa-barplot.qzv
```

{{Edet}}

5. Generate a phylogenetic tree using SEPP fragment insertion. You can find a Greengenes SEPP reference file [here](https://docs.qiime2.org/2022.8/data-resources/#) (<https://docs.qiime2.org/2022.8/data-resources/#>). Also, see the Parkinson's Mouse Tutorial (<https://docs.qiime2.org/2022.8/tutorials/pd-mice/#generating-a-phylogenetic-tree-for-diversity-analysis>) for help.

{{Sdet}}

Solution{{Esum}}

```
mkdir 06_tree

wget \
-O "06_tree/sepp-refs-gg-13-8.qza" \
"https://data.qiime2.org/2022.8/common/sepp-refs-gg-13-8.qza"

qiime fragment-insertion sepp \
--i-representative-sequences 03_denoise/asv-sequences.qza \
--i-reference-database 06_tree/sepp-refs-gg-13-8.qza \
--o-tree 06_tree/tree.qza \
--o-placements 06_tree/tree_placements.qza \
--p-threads 10
```

{{Edet}}

This will take a long time to complete. Feel free to **ctrl + C** to terminate. There will be a de novo phylogenetic tree available for the next exercise.



# Lesson 5 Practice

This practice lesson is associated with Lesson 5 of the Microbiome Analysis with QIIME 2. In this practice lesson, we will work on choosing a sampling depth to rarefy, running core-metrics, and comparing alpha diversity between our two metadata groups (Old vs young). We will continue working with the data from [Zhang et al. 2022](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9051480/) (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9051480/>).

1. Navigate to your Practice directory.

{{Sdet}}

Solution{{Esum}}

```
cd Practice
```

{{Edet}}

2. Generate a summary of your recently filtered feature table (i.e., `04_filter/filtered-table3.qza`).

{{Sdet}}

Solution{{Esum}}

```
qiime feature-table summarize \
--i-table 04_filter/filtered-table3.qza \
--m-sample-metadata-file /data/practice/metadata.txt \
--o-visualization 04_filter/filtered-table3-summ.qzv
```

{{Edet}}

3. Generate an alpha rarefaction plot and save the plot to a new directory named `07_analysis`. There is a tree available at `/data/practice/phylogeny-align-to-tree-mafft-fasttree/rooted_tree.qza`.

{{Sdet}}

Solution{{Esum}}

```
mkdir 07_analysis
qiime diversity alpha-rarefaction \
```

```
--i-table 04_filter/filtered-table3.qza \
--i-phylogeny /data/practice/06_tree/phylogeny-align-to-tree-maf
--m-metadata-file /data/practice/metadata.txt \
--p-max-depth 40000 \
--o-visualization 07_analysis/alpha-rarefaction-plot.qzv
```

{{Edet}}

4. Choose a sampling depth to rarefy using the feature table summary (`filtered-table3-summ.qzv`) and the alpha-rarefaction plot (``alpha-rarefaction-plot.qzv``).

{{Sdet}}

Solution{{Esum}}

Diversity plateaus around 20k reads, so it is safe to rarefy at the smallest library depth (69,355).

{{Edet}}

5. Generate core metrics using the `core-metrics-phylogenetic` pipeline.

{{Sdet}}

Solution{{Esum}}

```
qiime diversity core-metrics-phylogenetic \
--i-phylogeny /data/practice/06_tree/phylogeny-align-to-tree-maf
--i-table 04_filter/filtered-table3.qza \
--p-sampling-depth 69355 \
--p-n-jobs-or-threads 8 \
--m-metadata-file /data/practice/metadata.txt \
--output-dir 07_analysis/diversity-core-metrics-phylogenetic
```

{{Edet}}

6. Compare old vs young using a kruskal-wallis test (`alpha-group-significance`).

{{Sdet}}

Solution{{Esum}}

```
qiime diversity alpha-group-significance \
--i-alpha-diversity 07_analysis/diversity-core-metrics-phylogene
--m-metadata-file /data/practice/metadata.txt \
--o-visualization 07_analysis/alpha-group-sig-obs-feats.qzv
```

{}{Edet}{}

How does this compare to the paper?

# Practice Lesson 6

This practice lesson is associated with Lesson 6 of the Microbiome Analysis with QIIME 2. In this practice lesson, we will view beta diversity results and determine whether our two conditions (old vs young) demonstrate significant differences in microbial community structure. We will continue working with the data from [Zhang et al. 2022](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9051480/) (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9051480/>).

1. Examine the weighted and unweighted unifrac PCoAs. Do samples demonstrate group clustering?

{}{Sdet}{}{Esum}{{}}

Solution{}{Esum}{{}}

Yes, more so with unweighted UniFrac distances.

{}{Edet}{}{Esum}{{}}

2. Is there a significant difference between microbial communities from old vs young individuals?

{}{Sdet}{}{Esum}{{}}

Solution{}{Esum}{{}}

```
qiime diversity beta-group-significance \
--i-distance-matrix 07_analysis/diversity-core-metrics-phylogene \
--m-metadata-file /data/practice/metadata.txt \
--m-metadata-column 'DataType' \
--p-method 'permanova' \
--o-visualization 07_analysis/permanova_results.qzv
```

{}{Edet}{}{Esum}{{}}

3. Is there a significant difference in group dispersion?

{}{Sdet}{}{Esum}{{}}

Solution{}{Esum}{{}}

```
qiime diversity beta-group-significance \
--i-distance-matrix 07_analysis/diversity-core-metrics-phylogene \
--m-metadata-file /data/practice/metadata.txt \
```

```
--m-metadata-column 'DataType' \
--p-method 'permdisp' \
--o-visualization 07_analysis/permdisp_results.qzv
```

{{Eedit}}

# Getting the Data

The data used in this course are freely available from the Sequence Read Archive (SRA-NCBI) and [qiime2.org](https://qiime2.org/) (<https://qiime2.org/>). For your convenience, we are also including a compressed archive containing the data [here](#).

## References

# References

This course series primarily used information from QIIME2.org (<https://qiime2.org/>) and the QIIME2 forum (<https://forum.qiime2.org/>). Specifically, this course series focused on data and code from the QIIME2 Cancer Microbiome Intervention Tutorial (<https://docs.qiime2.org/jupyterbooks/cancer-microbiome-intervention-tutorial/040-appendices/citations.html#id2>).

Special thanks goes to the QIIME 2 development team for making these resources available!

## References for the main content

1. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvall C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu YX, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, and Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* 37: 852–857. <https://doi.org/10.1038/s41587-019-0209-9>
2. Chen Liao, Bradford P Taylor, Camilla Ceccarani, Emily Fontana, Luigi A Amoretti, Roberta J Wright, Antonio L C Gomes, Jonathan U Peled, Miguel-Angel Perales, Marcel R M van den Brink, Eric Littmann, Eric G Pamer, Jonas Schluter, and Joao B Xavier. Compilation of longitudinal microbiota data and hospitalome from hematopoietic cell transplantation patients. *Sci Data*, 8(1):71, March 2021.
3. Ying Taur, Katharine Coyte, Jonas Schluter, Elizabeth Robilotti, Cesar Figueroa, Mergim Gjonbalaj, Eric R Littmann, Lilan Ling, Liza Miller, Yangtsho Gyaltshen, Emily Fontana, Sejal Morjaria, Boglarka Gyurkocza, Miguel-Angel Perales, Hugo Castro-Malaspina, Roni Tamari, Doris Ponce, Guenther Koehne, Juliet Barker, Ann Jakubowski, Esperanza Papadopoulos, Parastoo Dahi, Craig Sauter, Brian Shaffer, James W Young, Jonathan Peled, Richard C Meagher, Robert R Jenq, Marcel R M van den Brink, Sergio A Giralt, Eric G Pamer, and Joao B Xavier. Reconstitution of the gut microbiota of antibiotic-treated patients by autologous fecal microbiota transplant. *Sci. Transl. Med.*, September 2018.

## **Additional Resources**

# Additional Resources

## The QIIME 2 docs and forum

For general support throughout your microbiome data analysis, see the QIIME 2 documentation (<https://docs.qiime2.org/2022.8/>) and the QIIME 2 forum (<https://forum.qiime2.org/>).

## Related readings

1. q2book (<https://gregcaporaso.github.io/q2book/front-matter/reading.html>)
2. Some useful slides from UCSD ([http://comppbio.ucsd.edu/wp-content/uploads/2018/07/20180621\\_oslo\\_university\\_microbiome\\_analysis\\_with\\_qiime2\\_tutorial.pdf](http://comppbio.ucsd.edu/wp-content/uploads/2018/07/20180621_oslo_university_microbiome_analysis_with_qiime2_tutorial.pdf))

## Linux help

1. Introduction to Unix from the *Bioinformatics Workbook* (<https://bioinformaticsworkbook.org/Appendix/Unix/unix-basics-1.html#gsc.tab=0>)
2. Unix from Happy Belly Bioinformatics (<https://astrobiomike.github.io/unix/>)
3. Linux Command List (from fosswire.com) (<https://files.fosswire.com/2007/08/fwunixref.pdf>)
4. Bioinformatics for Beginners: Module 1 - Linux and Biowulf ([https://btep.ccr.cancer.gov/docs/b4b/Module1\\_Unix\\_Biowulf/Lesson1/](https://btep.ccr.cancer.gov/docs/b4b/Module1_Unix_Biowulf/Lesson1/))

## Other microbiome analysis platforms / tools

- Nephele (<https://nephele.niaid.nih.gov/>)
- Bacterial and Viral Bioinformatics resource tools (<https://www.bv-brc.org/>)
- Kbase (<https://www.kbase.us/>)
- Microbiome Analyst (<https://www.microbiomeanalyst.ca/>)
- Mothur (<https://mothur.org/wiki/>)
- drive5 Bioinformatics software and services (<https://drive5.com/>)
- Phyloseq (<https://joey711.github.io/phyloseq/>)