
Qiita analysis tutorial

Release 0.2

Jan 12, 2020

1	Qiita tutorials:	3
1.1	Getting example data	3
1.2	Setting up Qiita	4
1.3	Studies in Qiita	5
1.4	Creating an example study	5
1.5	Adding sample information	7
1.6	Adding a preparation template and linking it to raw data	10
1.7	Exploring the raw data	14
1.8	Processing 16S data	15
1.9	The closed-reference workflow	19
1.10	The deblur workflow	19
1.11	Running the workflow	21
1.12	Analysis of Closed Reference Process	23
1.13	Rarefying Data	26
1.14	Taxa Bar Plots	30
1.15	Alpha Diversity Analysis	30
1.16	Beta Diversity Analysis	34
1.17	Filtering Data	48
1.18	Filtered Unweighted UniFrac Analysis	48
1.19	Altering Workflow Analysis Names	48

Materials below are intended for Jagiellonian University Bioinformatics 2 course. They include all information covered during the lab session.

For more information on Qiita, including Qiita philosophy and documentation, please visit [Qiita website](#).

A description of many of the terms used in this tutorial can be found in this [glossary](#).

This tutorial is adapted from the [University of California San Diego Center for Microbiome Innovation \(CMI\) Qiita/GNPS workshop](#). You can find more information on the CMI [here](#).

For more comprehensive tutorial on Qiita, please visit the CMI-workshop website. More advanced tutorials in QIIME 2 are available on the [QIIME 2 website](#).

If you have questions about this material, please [contact Tomasz Kosciolk](#).

CHAPTER 1

Qiita tutorials:

This tutorial will walk you through creation of your account and a test study in Qiita.

1.1 Getting example data

There are two separate example datasets made available to you - a *processing dataset* containing raw sequencing files which we will process to generate information about the identity and relative amounts of microbes in our samples (n=14), and an *analysis dataset* which contains a unique set of pre-processed samples (n=30) which we will use for statistical and community analyses.

NOTE

During this lab we are only going to perform the analysis step. Processing information is included only for background and context.

1.1.1 Processing dataset

You can download the [processing dataset](#) directly from GitHub. These files contain 16S rRNA microbiome data for 14 human skin samples. It is a subset of data that we will use later for analysis. Real sequencing data can be tens of gigabytes in size!

The files are:

- CMI_workshop_lane1_S1_L001_R1_001.fastq.gz # 16S sequences - forward reads
- CMI_workshop_lane1_S1_L001_R2_001.fastq.gz # 16S sequences - reverse reads
- CMI_workshop_lane1_S1_L001_I1_001.fastq.gz # 16S sequences - barcodes
- sample_info.txt # The sample information file
- prep_info_16S.txt # The preparation information file

1.1.2 Analysis dataset

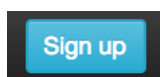
Example data that you can use for analysis are available to you directly on Qiita. You don't need to download anything to your hard drive. Instructions how to access these data are provided in the [analysis tutorial](#).

1.2 Setting up Qiita

1.2.1 Signing up for a Qiita account

Open your browser (it must be Chrome or Firefox) and go to [Qiita \(https://qiita.ucsd.edu\)](https://qiita.ucsd.edu).

Click on “Sign Up” on the upper-right-hand corner.



The “New User” link brings you to a page on which you can create a new account. Optional fields are indicated explicitly, while all other fields are required.

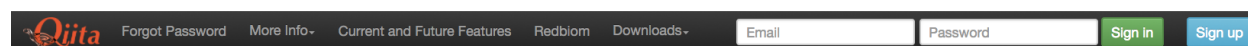
Enter User Information

Email	<input type="text" value="Email"/>	
Password	<input type="password" value="password"/>	
Confirm password	<input type="password" value="password"/>	
Name (Optional)	<input type="text" value="name"/>	
Affiliation (Optional)	<input type="text" value="affiliation"/>	
Address (Optional)	<input type="text" value="address"/>	
Phone # (Optional)	<input type="text" value="phone"/>	

Once the form is submitted, an email will be sent to you containing instructions on how to verify your email address.

1.2.2 Logging into your account and resetting a forgotten password

Once you have created your account, you can log into the system by entering your email and password.



If you forget your password, you will need to reset it. Click on “Forgot Password”.

This will take you to a page on which to enter your email address; once you click the “Reset Password” button, the system will send you further instructions on how to reset your lost password.

1.2.3 Updating your settings and changing your password

If you need to reset your password or change any general information in your account, click on your email at the top right corner of the menu bar to access the page on which you can perform these tasks.

User Information

Name

Affiliation

Address

Phone

Save Edits

Change Password

Old Password

New Password

Repeat New Password

Change Password

User Information

Name

Affiliation

Address

Phone

Save Edits

Change Password

Old Password

New Password

Repeat New Password

Change Password

1.3 Studies in Qiita

Studies are the source of data for Qiita. Studies can contain only one set of samples but can contain multiple sets of raw data, each of which can have a different preparation – for example, 16S, shotgun metagenomics, and metabolomics, or even multiple preparations of the same type (e.g., a plate rerun, biological and technical replicates, etc).

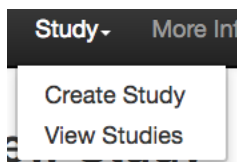
In the *analysis tutorial*, our study contains 30 samples, each with two types of data: 16S and metabolomic. To represent this project in Qiita, we created a single study with a single sample information file that contains all 30 samples. Then, we linked separate preparation files for each data type.

NOTE

You may skip the remainder of this section and proceed to *Analysis of Closed Reference Process*

1.4 Creating an example study

To create a study, click on the “Study” menu and then on “Create Study”. This will take you to a new page that will gather some basic information to create your study.



The “Study Title” has to be unique system-wide. Qiita will check this when you try to create the study, and may ask you to alter the study name if the one you provide is already in use.

A principal investigator is required, and a list of known PIs is provided. If you cannot find the name you are looking for in this list, you can choose to add a new one.

Create a new Study

* = Required Field

Study Title	<input type="text" value="[user's_name]"/>	*
Study Alias	<input type="text" value="[user's_name]"/>	*
DOI	<input type="text"/>	
Just values, no links, comma separated values		
PUBMED ID	<input type="text"/>	
Just values, no links, comma separated values		
Study Abstract	<input type="text" value="[user's abstract]"/>	*
Study Description	<input type="text" value="[description of user's study]"/>	*
Principal Investigator	<input type="text" value="Select an Option"/>	*
Lab Person	<input type="text" value="Select an Option"/>	
	Can't find the person you're looking for? Add a person	
Environmental Packages	<div> <div> <div>air</div> <div>built environment</div> <div>host-associated</div> <div>human-amniotic-fluid</div> <div>human-associated</div> <div>human-blood</div> <div>human-gut</div> <div>human-oral</div> <div>human-skin</div> <div>human-urine</div> <div>human-vaginal</div> <div>microbial mat/biofilm</div> <div>miscellaneous natural or artificial environment</div> <div>plant-associated</div> <div>sediment</div> <div>soil</div> <div>wastewater/sludge</div> <div>water</div> </div> </div>	
You can select multiple entries by control-clicking (mac: command-clicking)		*
Event-Based Data	<input type="text" value="No timeseries"/>	

Select the environmental package appropriate to your study. Different packages will request different specific information about your samples. For more details, see the [publication](#). For this test study for the *processing tutorial*, choose **human-skin**.

There is also an option to specify time series type (“Event-Based Data”) if you have such data. In our case, the samples come from a time series study design, so you should select “multiple intervention, real”. For more information on time series types, you can check out the [in-depth tutorial](#) on the Qiita website.

Once your study has been created, you will be informed by a green message; click on the study name to begin adding your data.

Study [user's_name] successfully created

1.5 Adding sample information

Sample information is the set of metadata that pertains to your biological samples: these are the measured variables that are motivating you to look for response variables in the microbiome. **IMPORTANT:** your metadata are your study; it is imperative that those data are consistent, correct, and sufficiently detailed. (To learn more, including how to format your own sample info file, check out the [in-depth documentation](#) on the Qiita website.)

The first point of entrance to a study is the study description page. Here you will be able to edit the study info, upload files, and manage all other aspects of your study.

The screenshot shows the Qiita web interface. At the top is a navigation bar with links: Analysis, Study, More Info, Current and Future Features, Redbiom, Downloads, and a welcome message. The main content area is titled "[user's_name] - ID 11568". On the left, there's a sidebar with buttons: "Study Information", "Sample Information", and "Upload Files". Below these is a message: "No preparation information has been added yet". The main content area displays study details: "Abstract" (with a placeholder "[user's abstract]"), "Study ID: 11568", "Publications:", "PI: Rob Knight (UCSD)", "Lab Contact: None", "Shared With:", "Samples: 0", and "EBI: not submitted". There are buttons for "Share", "Edit", and "Delete". To the right, there's a "Study Tags" section with a list of tags: "Previously admin", "Previously assigned", and "New". Below this is a text input field "Add more tags" and a "Save tags" button. At the bottom, there's a footer with a thank you message and contact information.

Since we are using a practice set of data, under “Study Tags” write “Tutorial” and select “Save Tags”. As part of our routine clean up efforts, this tag will allow us to find and remove studies and analyses generated using the template data and information.

The first step after study creation is uploading files. Click on the “Upload Files” button: as shown in the figure below, you can now drag-and-drop files into the grey area or simply click on “select from your computer” to select the fastq, fastq.gz or txt files you want to upload.

Note: Per our Terms of Condition for use, by uploading files to Qiita you are certifying that they do not contain: 1) Protected health information within the meaning of 45 Code of Federal Regulations part 160 and part 164, subparts A and E; [see checklist](#) 2) Whole genome sequencing data for any human subject; [HMP human sequence removal protocol](#) 3) Any data that is copyrighted, protected by trade secret, or otherwise subject to third party proprietary rights, including privacy and publicity rights, unless you are the owner of such rights or have permission from the

Study Tags

Previously admin, Previously assigned, New

Add more tags

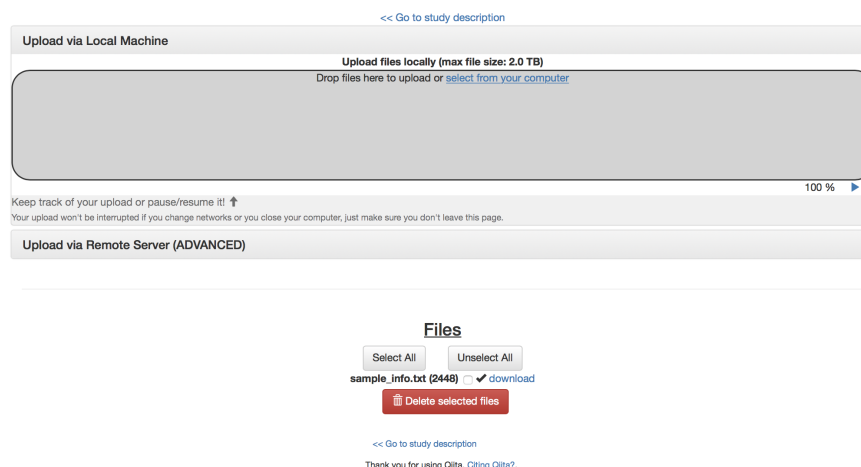
New tags are linked to the user that created them. Report abuse.

Save tags

rightful owner(s) to transfer the data and grant it to Qiita, on behalf of the Regents of the University of California, all of the license rights granted in our [Terms](#).

Uploads can be paused at any time and restarted again, as long as you do not refresh, navigate away from the page, or log out of the system from another browser window.

To proceed, drag the file named “sample_info.txt” into the upload box. It should upload quickly and appear below “Files” with a checkbox next to it below.



Once your file has uploaded, click on “Go to study description” and, once there, click on the “Sample Information” tab. Select your sample information from the dropdown menu next to “Upload information” and click “Create”.

If something is wrong with the sample information file, Qiita will let you know with a red banner at the top of the screen.

If the file processes successfully, you should be able to click on the “Sample Information” tab and see a list of the imported metadata fields.

To check out the different metadata values select the “Sample-Prep Summary” tab. On this page, select a metadata column to visualize in the “Add sample column information to table” dropdown menu and click “Add column.”

Next, we’ll add 16S raw data and process it.

Next: *Adding a preparation template and linking it to raw data*

[user's_name] - ID 11568

[user's_name]

Sample Information

Select sample information file:

sample_info.txt

If uploading a QIIME mapping file, select the data type of the prep information:

Choose a data type...

Create

❗ The 'sample_name' column is missing from your template, this file cannot be parsed.
Need help? Send us an [email](#).

[user's_name] - ID 11568

[user's_name]

Sample Information

📄 Sample Info

🗑 Delete

Sample Information

Sample-Prep Summary

Number of samples: 14

Number of columns: 21

Update sample information:

Choose file...

Information summary



anonymized_name: All the values in this category are different

[user's_name] - ID 11568

[user's_name]

Sample Information

Download Sample Info

Delete

Sample Information

Sample-Prep Summary

Sample Summary

Add sample column information to table

Add column

	Sample	
	11568.8D10	
	11568.4F2	

NOTE

Do not follow this section during the Bioinformatics 2 lab. Go directly to [Analysis of Closed Reference Process](#). This information is included here for context and everyone is encouraged to familiarize themselves with it **after class**.

Now, we'll upload some actual microbiome data to explore. To do this, we need to add the data themselves, along with some information telling Qiita about how those data were generated.

1.6 Adding a preparation template and linking it to raw data

Where the *sample info file* has the biological metadata associated with your samples, the *preparation info file* contains information about the specific technical steps taken to go from sample to data. Just as you might use multiple data-generation methods to get data from a single sample – for example, target gene sequencing and shotgun metagenomics – you can have multiple prep info files in a single study, associating your samples with each of these data types. You can learn more about prep info files at the [Qiita documentation](#).

Go back to the “Upload Files” interface. In the [example data](#), find and upload the 3 “.fastq.gz files” and the “[prep_info_16S.txt](#)” file.

These files will appear under “Files” when they finish uploading.

Then, go to the study description. Now you can click the “Add New Preparation” button. This will bring up the following dialogue:

Select “[prep_info_16S.txt](#)” from the “Select file” dropdown, and “16S” as the data type. Optionally, you can also select one of a number of investigation types that can be used to associate your data with other like studies in the database. Click “Create New Preparation”.

You should now be brought to a “Processing” tab of your preparation info:

By clicking on the “Summary” tab on this page you can see the preparation info that you uploaded.

In addition, you should see a “16S” button appear under “Data Types” on the menu to left:

You can click this to reveal the individual prep info files of that data type that have been associated with this study:

[<< Go to study description](#)

Upload via Local Machine

Upload files locally (max file size: 2.0 TB)

Drop files here to upload or [select from your computer](#)

Keep track of your upload or pause/resume it!

Your upload won't be interrupted if you change networks or you close your computer, just make sure you don't leave this page.

10 %

Upload via Remote Server (ADVANCED)

Files

Select All

Unselect All

CMI_workshop_lane1_S1_L001_R1_001.fastq.gz (4568505)

CMI_workshop_lane1_S1_L001_R1_001.fastq.gz (18733712)

CMI_workshop_lane1_S1_L001_R2_001.fastq.gz (17568260)

prep_info_16S.txt (2250)

Delete selected files

[<< Go to study description](#)

Thank you for using Qiita. [Citing Qiita?](#)
 Questions? qiita.help@gmail.com; don't forget to add your study or analysis id.
 Read our [terms and conditions](#).

Add a new preparation file: (* Required fields)

Name:

CMI tutorial

Select file: *

prep_info_16S.txt

Select data type: *

16S

Select Investigation Type:

Not sure what to select? [Check](#)

Unsure? [Check](#)

Create New Preparation

Thank you for using Qiita. [Citing Qiita?](#)
 Questions? qiita.help@gmail.com; don't forget to add your study or analysis id.
 Read our [terms and conditions](#).

[user's_name] - ID 11568

[user's_name]

[user's_prep_info] - ID 4220 (16S) Edit name Prep info QIIME map Delete

Summary

Processing

No files attached to this preparation


Select type: Choose a type...

Add a name for the file:

[user's_name] - ID 11568

[user's_name]

[user's_prep_info] - ID 4220 (16S)

 Edit name

 Prep info

 QIIME map

 Delete

Summary

Processing

Number of samples: 14

Number of columns: 14

Update prep information:

Select Investigation Type:

Information summary



barcode: All the values in this category are different.

Data Types (click on the tabs)

 16S

Data Types (click on the tabs)

 16S

[user's_prep_info] - ID 4220 - sandbox

None - ID None

None

If you have multiple 16S preparations (for example, if you sequenced using several different primer sets), these would each show up as a separate entry here.

Now, you can associate the sequence data from your study with this preparation.

[user's_name] - ID 11568

[user's_name]

[user's_prep_info] - ID 4220 (16S)

Edit name

Prep info

QIIME map

Delete

Summary

Processing

No files attached to this preparation

Select type:

Choose a type...

Add a name for the file:

In the prep info dialogue, there is a dropdown menu below the words *No files attached to this preparation*, labeled “Select type”. Click “Choose a type” to see a list of available file types. In our case, we’ve uploaded FASTQ-formatted file for all samples in our study, so we will choose “FASTQ - None”. In some cases outside of this tutorial, you may have per sample FASTQ files, so take care in considering which data type you are handling.

Magically, this will prompt Qiita to associate your uploaded files with the corresponding samples in your preparation info. (Our prep info file has a column named *run_prefix*, which associated the *sample_name* with the file name prefix for that particular sample).

You should see this as filenames showing up in the green: *raw barcodes* (file with *I1* in its name), *raw forward seqs* (*R1* in name) and *raw reverse seqs* (*R2* in name) columns below the import dropdown. You’ll want to give the set of these FASTQ files a name (*Add a name for the file* field below *Select type: FASTQ - None*), and then click “Add files” below.

[user's_name] - ID 11568

[user's_name]

[user's_prep_info] - ID 4220 (16S)

Edit name

Prep info

QIIME map

Delete

Summary

Processing

No files attached to this preparation

Select type:

FASTQ - None

Add a name for the file:

[user's_name]

Now, you can import files from other studies

Choose an artifact to import...

or click and drag your uploaded files to the correct file type

Please make sure that the correct files are in the correct column.

Note: the system will try to auto select the files based on run_prefix, if that doesn't work, either the type you selected doesn't support the use of run_prefix or the run_prefix is wrong

Available Files	raw barcodes	raw forward seqs	raw reverse seqs
	CMI_workshop_lane1_S1_L001_I1_001.fastq	CMI_workshop_lane1_S1_L001_R1_001.fas	CMI_workshop_lane1_S1_L001_R2_001.fas

Add files

That’s it! Your data are ready for processing.

1.7 Exploring the raw data

Click on the 16S menu on the left. Now that you’ve associated sequence files with this prep, you’ll have a “Processing network” displayed:

CMI tutorial - ID 6631 (16S) [Edit name](#) [Prep info](#) [QIIME map](#)

[Summary](#) [Processing](#) [Hide](#)

Processing network

Click on the graph to navigate through it. Click circles for more information. This graph will refresh in 2 seconds or reload [now](#)

Job status (circles): success running error in_construction queued waiting deleting

Artifact status (triangles): artifact type outdated deprecated

Thank you for using Qiita. [Citing Qiita?](#)
 Questions? qiita.help@gmail.com; don't forget to add your study or analysis id.
[Read our terms and conditions.](#)

If you see this message:

! This prep template is currently being updated

It means that your files need time to load. Refresh your screen after about 1 minute.

Your collection of FASTQ files for this prep are all represented by a single object in this network, currently called “CMI tutorial - 14 skin samples”. Click on the object.

Now, you’ll have a series of choices for interacting with this object. You can click “Edit” to rename the object, “Process” to perform analyses, or “Delete” to delete it. In addition, you’ll see a list of the actual files associated with this object.

[user's_name] (ID: 43489) Visibility: sandbox [Edit name](#) [Process](#) [Delete](#) [Request approval](#)

Available files:

[CMI_workshop_lane1_S1_L001_R2_001.fastq.gz \(raw reverse seqs\)](#)

[CMI_workshop_lane1_S1_L001_I1_001.fastq.gz \(raw barcodes\)](#)

[CMI_workshop_lane1_S1_L001_R1_001.fastq.gz \(raw forward seqs\)](#)

Currently, no summary exists.

[Generate summary](#)

Scroll to the bottom, and you’ll also see an option to generate a summary of the object.

Currently, no summary exists.

Generate summary

If you click this button, it will be replaced with a notification that the summary generation has been added to the processing queue.

To check on the status of the processing job, you can click the rightmost icon at the top of the screen:



This will open a dialogue that gives you information about currently running jobs, as well as jobs that failed with some sort of error. *Please note*, this dialogue keeps the entire history of errors that Qiita encountered for your jobs, so take notice of dates and times in the *Heartbeat* column.

Active Jobs

×

successful jobs are not shown

Search:

Heartbeat	Name	Status	Step
2017-12-05 15:51:40	release_validators	running	
	Generate HTML summary	queued	

Close

The summary generation shouldn't take too long. You may need to refresh your screen. When it completes, you can click back on the FASTQ object and scroll to the bottom of the page to see a short peek at the data in each of the FASTQ files in the object. These summaries can be useful for troubleshooting.

Now, we'll process the raw data into something more interesting.

1.8 Processing 16S data

Scroll back up and click on the "CMI tutorial - 14 skin samples(FASTQ)" artifact, and select "Process". Below the files network, you will now see a "Choose command" dropdown menu. Based on the type of object, this dropdown menu will give you a list of available processing steps.

For 16S "FASTQ" objects, the only available command is "Split libraries FASTQ". This converts the raw FASTQ data into the file format used by Qiita for further analysis (you can read more extensively about this file type [here](#)).

[Open summary in a new window](#)

CMI_workshop_lane1_S1_L001_I1_001.fastq.gz (raw_barcodes)

MD5:: c37e6591036167ca2fa064e6913ae9a0

```
@D00611:254:HKV3NBCXX:1:2202:3627:54691 1:N:0:1
TTCACACAGTGG
+
<<<<. <. <<<<. <
@D00611:254:HKV3NBCXX:1:1111:19044:55327 1:N:0:1
TCTTAAGATTTG
+
<<<<@D00611:254:HKV3NBCXX:1:2205:20391:30288 1:N:0:1
ATGTGCTGCTCG
```

CMI_workshop_lane1_S1_L001_R1_001.fastq.gz (raw_forward_seqs)

MD5:: cd5c636f6df04a12653769e132601edb

[illegible]

CMI_workshop_lane1_S1_L001_R2_001.fastq.gz (raw_reverse_seqs)

MD5:: 8a1226efccd6d860516bb363220cecad

[illegible]

Select the “Split libraries FASTQ” step. Now, you will be able to select the specific combination of parameters to use for this step in the “Choose parameter set” dropdown menu.

For our files, choose “Multiplexed FASTQ; Golay 12 base pair reverse complement mapping file barcodes with reverse complement barcodes”. The specific parameter values used will be displayed below. **For most raw data coming out of the Knight Lab you will use the same setting.**

Click “Add Command”.

You'll see the files network update. In addition to the original white object, you should now see the processing command (represented in yellow) and the object that will be produced from that command (represented in grey).

You can click on the command to see the parameters used, or on an object to perform additional steps.

Next we want to trim to a particular length, to ensure our samples will be comparable to other samples already in the database. Click back on the “demultiplexed (Demultiplexed)”. This time, select the Trimming operation. Currently, there are seven trimming length options. Let’s choose “100 basepairs”, which trims to the first 100bp, for this run, and click “Add Command”.

Click “Add Command”, and you will see the network update:

Note that the commands haven't actually been run yet! (We'll still need to click "Run" at the top.) This allows us to add multiple processing steps to our study and then run them all together.

We're going to process our sequences files using two different workflows. In the first, we'll use a conventional reference-based OTU picking strategy to cluster our 16S sequences into OTUs. This approach matches each sequence to a reference database, ignoring sequences that don't match the reference. In the second, we will use [deblur](#), which uses an algorithm to remove sequence error, allowing us to work with unique sequences instead of clustering into OTUs. Both of these approaches work great with Qiita, because we can compare the observations between studies without having to do any sort of re-clustering!

CMI tutorial - ID 6631 (16S)

Edit name

Prep info

QIIME map

Summary

Processing

Processing network

Hide

Run

Click on the graph to navigate through it. Click circles for more information. This graph will refresh in 6 seconds or reload now

Job status (circles):

success

running

error

in_construction

queued

waiting

deleting

Artifact status (triangles):

artifact

type

outdated

deprecated

```

graph LR
    A[CMI tutorial - 14 skin samples (FASTQ)] --> B((Split libraries FASTQ))
    B --> C[demultiplexed (Demultiplexed)]
    
```

The graph shows a linear workflow. It starts with a triangle artifact node labeled 'CMI tutorial - 14 skin samples (FASTQ)'. An arrow points to a circle job node labeled 'Split libraries FASTQ'. Another arrow points to a triangle artifact node labeled 'demultiplexed (Demultiplexed)'. The interface includes a top bar with the tutorial name and action buttons, a processing network header, a 'Run' button, a detailed instruction, job status legends, artifact status legends, and navigation controls.

Thank you for using QIITA. [Citing QIITA?](#)

Questions? qiita.help@gmail.com; don't forget to add your study or analysis id.

[Read our terms and conditions.](#)

CMI tutorial - ID 6631 (16S)

Edit name

Prep info

QiIME map

Summary

Processing

Processing network

Hide

Run

Start workflow:

Click on the graph to navigate through it. Click circles for more information. This graph will refresh in 5 seconds or reload now

Job status (circles):

success

running

error

in_construction

queued

waiting

deleting

Artifact status (triangles): artifact type

outdated

deprecated

⌂

⌂

⌂

⌂

⌂

⌂

⌂

⌂

CMI tutorial - 14 skin samples (FASTQ)

Split libraries FASTQ

demultiplexed (Demultiplexed)

Choose command:

Trimming

Required parameters:

input data:

demultiplexed (Demultiplexed)

Optional parameters:

Parameter set:

100 base pairs

Note: changing default parameter values not allowed

length:

100

Add Command

Thank you for using Qiita. Citing Qiita?

Questions? qiita.help@gmail.com; don't forget to add your study or analysis id.

Read our [terms and conditions](#).

CMI tutorial - Justin update

Do you want to submit to EBI-ENA? Review the [submission checklist](#)

CMI tutorial - ID 6631 (16S)

Edit name

Prep info

QiIME map

Summary

Processing

Processing network

Hide

Run

Start workflow:

Click on the graph to navigate through it. Click circles for more information. This graph will refresh in 10 seconds or reload now

Job status (circles):

success

running

error

in_construction

queued

waiting

deleting

Artifact status (triangles): artifact type

outdated

deprecated

⌂

⌂

⌂

⌂

⌂

⌂

⌂

⌂

CMI tutorial - 14 skin samples (FASTQ)

Split libraries FASTQ

demultiplexed (Demultiplexed)

Trimming

Trimmed Demultiplexed (Demultiplexed)

Thank you for using Qiita. Citing Qiita?

Questions? qiita.help@gmail.com; don't forget to add your study or analysis id.

Read our [terms and conditions](#).

18

Chapter 1. Qiita tutorials:

1.9 The closed-reference workflow

To do closed reference OTU picking, click on the “Trimmed Demultiplexed 100 (Demultiplexed)” object and select the “Pick closed-reference OTUs” command. We will use the “Defaults” parameter set for our data, which are relatively small. For a larger data set, we might want to use the “Defaults - parallel” implementation.

CMI tutorial - ID 6631 (16S) [Edit name](#) [Prep info](#) [QIIME map](#)

[Summary](#) [Processing](#)

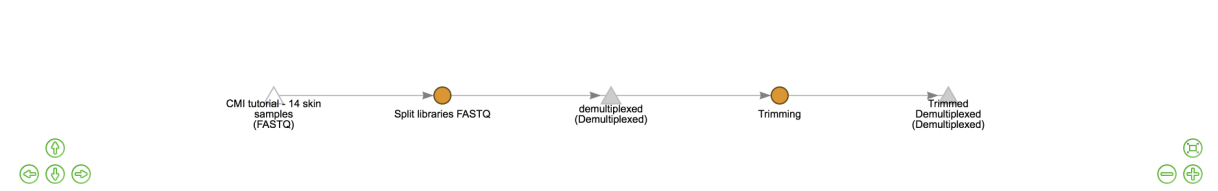
Processing network [Hide](#)

Start workflow: [Run](#)

Click on the graph to navigate through it. Click circles for more information. This graph will refresh in 4 seconds or reload [now](#)

Job status (circles): success running error in_construction queued waiting deleting

Artifact status (triangles): artifact type outdated deprecated



Choose command: [Pick closed-reference OTUs](#)

Required parameters:

input data: [Trimmed Demultiplexed \(Demultiplexed\)](#)

Optional parameters:

Parameter set: [Defaults](#)

Note: changing default parameter values not allowed

reference-seq: [/databases/gg/13_8/rep_set/97_otus.fasta](#)

reference-tax: [/databases/gg/13_8/taxonomy/97_otu_taxonomy.txt](#)

similarity: [0.97](#)

sortmerna coverage: [0.97](#)

sortmerna e_value: [1](#)

sortmerna max_pos: [10000](#)

threads: [1](#)

[Add Command](#)

Thank you for using Qiita. [Citing Qiita?](#)
 Questions? qiita.help@gmail.com; don't forget to add your study or analysis id.
[Read our terms and conditions.](#)

By default, Qiita uses the GreenGenes 16S reference database. You can also choose to use the Silva 119 18S database, or the UNITE 7 fungal ITS database.

Click “Add Command”, and you will see the network update:

Here you can see the blue “Pick closed-reference OTUs” command added, and that the product of the command is a BIOM-formatted OTU table.

That’s it!

1.10 The deblur workflow

The deblur workflow is only marginally more complex. Although you can deblur the demultiplexed sequences directly, “deblur” works best when all the sequences are the same length. By trimming to a particular length, we can also ensure our samples will be comparable to other samples already in the database.

Click back on the “Trimmed Demultiplexed 100 (Demultiplexed)” object. This time, select the *Deblur* operation. Choose “Deblur” from the “Choose command” dropdown, and “Defaults” for the parameter set.

Add this command to create this workflow:

CMI tutorial - Justin update

Do you want to submit to EBI-ENA? Review the [submission checklist](#)

CMI tutorial - ID 6631 (16S)

Edit name

Prep info

QIIME map

Summary

Processing

Processing network

Hide

Run

Start workflow:

Click on the graph to navigate through it. Click circles for more information. This graph will refresh in 10 seconds or reload now

Job status (circles):

success

running

error

in_construction

queued

waiting

deleting

Artifact status (triangles): artifact type

outdated

deprecated

CMI tutorial - 14 skin samples (FASTQ)

Split libraries FASTQ

demultiplexed (Demultiplexed)

Trimming

Trimmed Demultiplexed (Demultiplexed)

Pick closed-reference OTUs

OTU table (BIOM)

Thank you for using Qiita. [Citing Qiita?](#)

Questions? qiita.help@gmail.com; don't forget to add your study or analysis id.

[Read our terms and conditions.](#)

Click on the graph to navigate through it. Click circles for more information. This graph will refresh in 5 seconds or reload now

Job status (circles):

success

running

error

in_construction

queued

waiting

deleting

Artifact status (triangles): artifact type

outdated

deprecated

CMI tutorial - 14 skin samples (FASTQ)

Split libraries FASTQ

demultiplexed (Demultiplexed)

Trimming

Trimmed Demultiplexed (Demultiplexed)

Pick closed-reference OTUs

OTU table (BIOM)

Choose command:

Deblur

Required parameters:

Demultiplexed sequences:

Trimmed Demultiplexed (Demultiplexed)

Optional parameters:

Parameter set:

Defaults

Note: changing default parameter values not allowed

Error probabilities for each Hamming distance:

1, 0.06, 0.02, 0.02, 0.01, 0.005, 0.005, 0.005, 0.001, 1

Indexed negative filtering database:

default

Indexed positive filtering database:

default

Insertion/deletion (indel) probability:

0.01

Jobs to start:

5

Maximum number of insertion/deletion (indel):

3

Mean per nucleotide error rate:

0.005

Minimum dataset-wide read threshold:

0

Minimum per-sample read threshold:

2

Negative filtering database:

default

Positive filtering database:

default

Reference phylogeny for SEPP:

Greengenes_13.8

Sequence trim length (-1 for no trimming):

-1

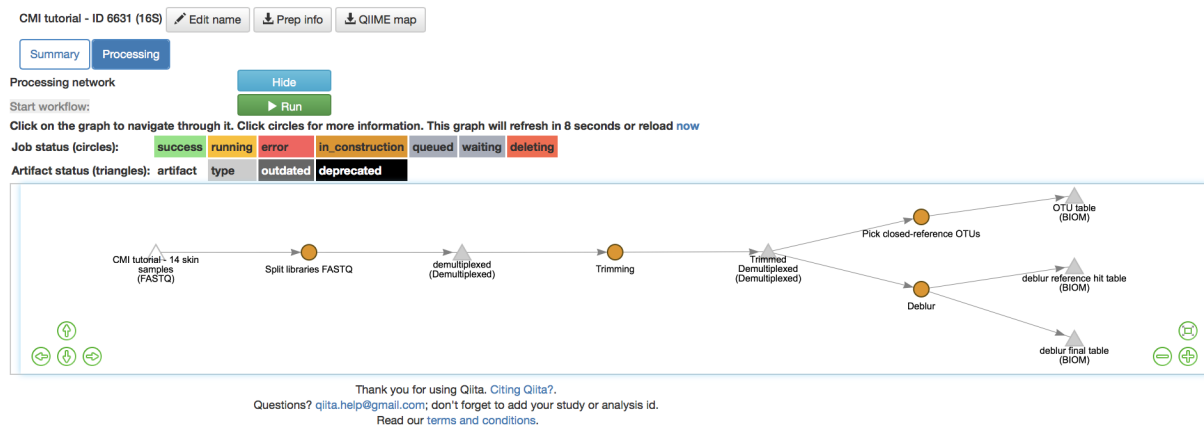
Threads per sample:

1

Add Command

20

Chapter 1. Qiita tutorials:

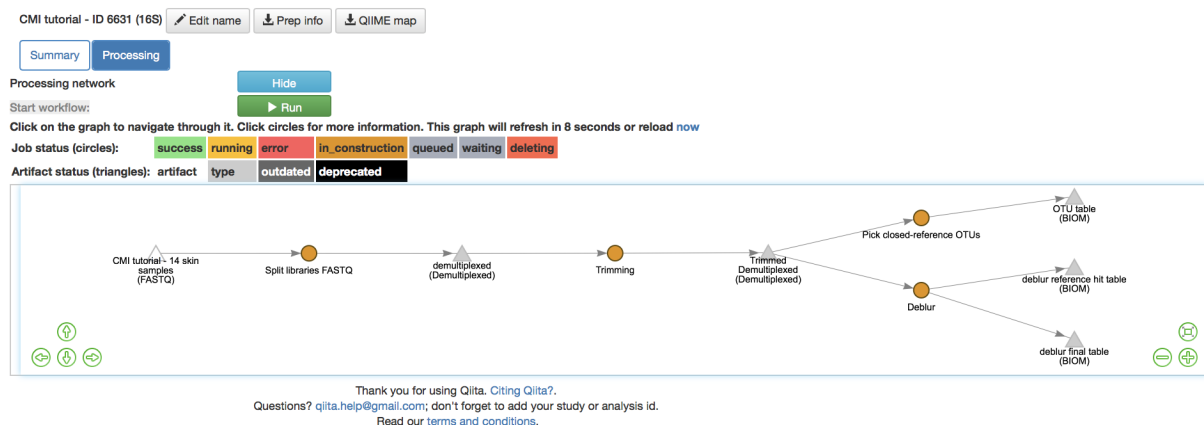


Now you can see that we have the same “Trimmed Demultiplexed (Demultiplexed)” object being used for two separate processing steps – closed-reference OTU picking, and deblur.

As you can see, “deblur” produces two BIOM-formatted OTU tables as output. The “deblur reference hit table (BIOM)” contains deblurred sequences that have been filtered to try and exclude things like organellar mitochondrial reads, while “deblur final table (BIOM)” has all the sequences.

1.11 Running the workflow

Now, we can see the whole set of commands and their output files:

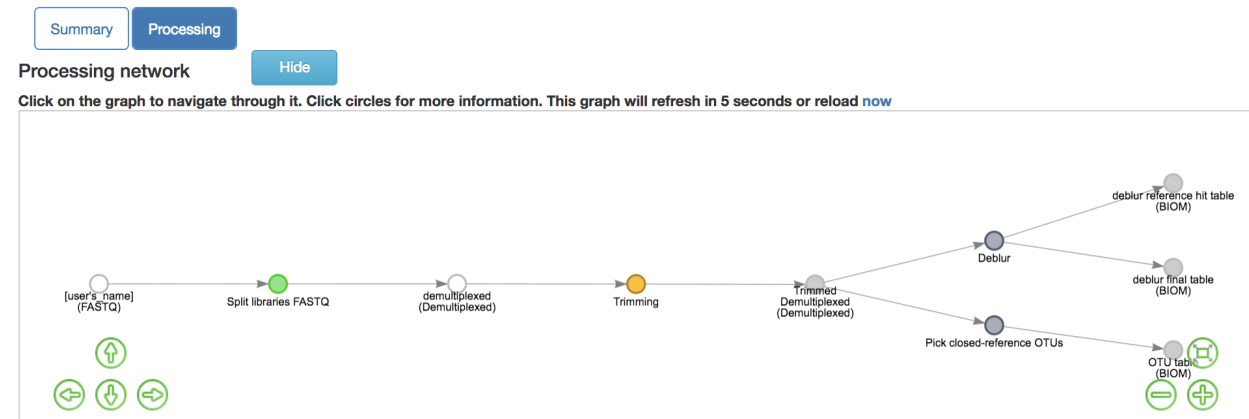


Click “Run” at the top of the screen, and Qiita will start executing all of these jobs. You’ll see a “Workflow submitted” banner at the top of your window.

The full workflow can take time to load depending on the amount of samples and Qiita workload. You can keep track of what is running by looking at the colors of the command artifacts. If yellow, the commands are being run now. If green, the commands have successfully been run. If red, the commands have failed.

Once objects have been generated, you can generate summaries for them just as you did for the original “FASTQ” object.

The summary for the “demultiplexed (Demultiplexed)” object gives you information about the length of sequences in the object:



Open summary in a new window

Features

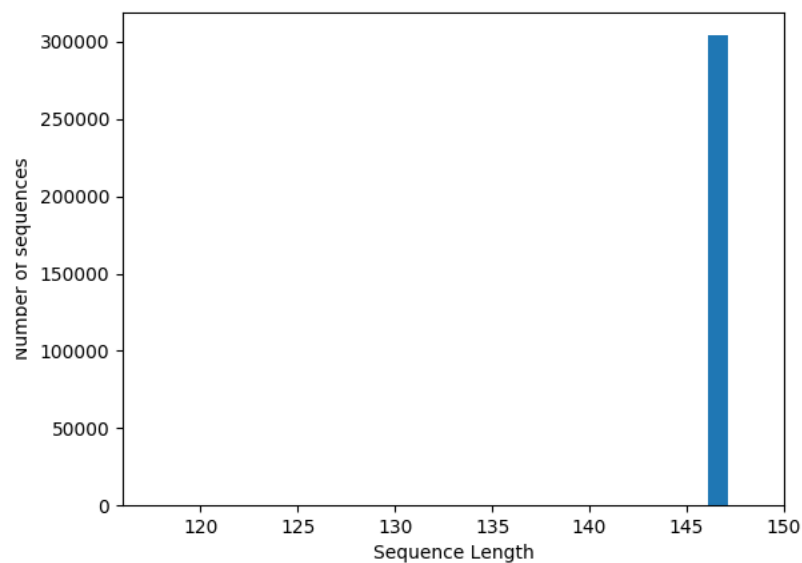
Total: 304000

Max: 150

Mean: 149

Standard deviation: 150

Median: 0



The summary for a BIOM-format OTU table gives you a table summary, details regarding the frequency per sample, and a histogram of the number of features per sample:

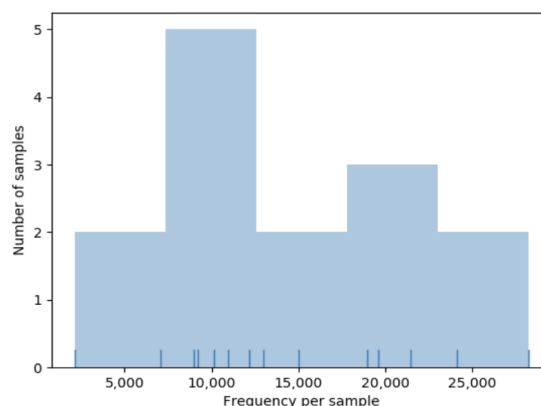
Table summary

Metric	Sample
Number of samples	14
Number of features	1,049
Total frequency	201,137

Frequency per sample

	Frequency
Minimum frequency	2,099.0
1st quartile	9,447.75
Median frequency	12,588.0
3rd quartile	19,446.75
Maximum frequency	28,285.0
Mean frequency	14,366.92857142857

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



Next: *Analysis of Closed Reference Process*

1.12 Analysis of Closed Reference Process

To create an analysis, select “Create new analysis” from the top menu.

This will take you to a list of studies with samples available to you for analysis, divided between your studies and publicly available studies (“Public Studies”).

Filter by column data (Title, abstract, PI, etc):

Filter studies by tags: (Admin, User)

Select tags for filtering

Your Studies

Show 5 entries

Expand for analysis	Title	Study ID	Samples	Shared With These Users	Principal Investigator	Publications	Status	Qiita EBI submission
1	CMI workshop analysis CMIWorkshop	11269	30	Modify Owner: Tomasz Austin Swafford, Alison Vrbanac, CMI, fernando.vargas0341@gmail.com, Justine Debellus, miv023@ucsd.edu, Robert Quinn, Yoshiki	Tomasz Kosciolk		public	not submitted
1	[user's_name]	11574	14	Modify Owner: CMI	Rob Knight		sandbox	not submitted
No BIOMs	Microbiome and metabolome in opiod and methamphetamine addicted patients	11480	124	Modify Owner: CMI Austin Swafford, Bryn Taylor, Carolina S. Carpenter, fernando.vargas0341@gmail.com, Greg Humphrey, sorchanian@eng.ucsd.edu	Karsten Zengler		sandbox	not submitted

Public Studies

Expand for analysis	Title	Study ID	Samples	Principal Investigator	Publications	Qiita EBI submission
1	Seasonal restructuring of the ground squirrel gut microbiota over the annual hibernation cycle	926	46	Hannah Carey	23152108, 10.1152/ajpregu.00387.2012	not submitted

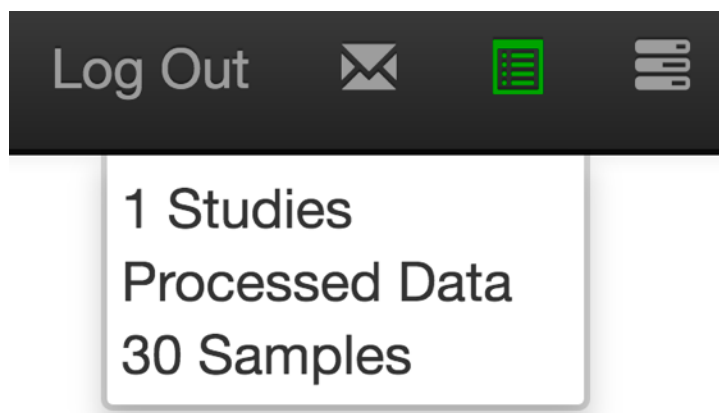
Find the “CMI workshop analysis” study in Public Studies. You can use the search window at the top right, or filter by tags (“CMIWorkshop” tag). Click the green plus sign at the left of the row. This will expand the study to expose all the objects from that study that are available to you for analysis.

Expand for analysis	Title	Study ID	Samples	Shared With These Users	Principal Investigator	Publications	Status	Qiita EBI submission
1	CMI workshop analysis CMIWorkshop	11269	30	Modify Owner: Tomasz Austin Swafford, Alison Vrbancac, CMI, fernando.vargas0341@gmail.com, Justine Debelius, miv023@ucsd.edu, Robert Quinn, Yoshiki	Tomasz Kosciolok		public	not submitted
<div> <div>Artifacts</div> <div>Processing method</div> <div>Data type</div> </div>								
Add all	Per Artifact (1)							165 ()

To look more closely at the details of the artifact, select “Per Artifact (1).” Here you can add each of these objects to the analysis by selecting the “Add” button. We will just add the Closed Reference OTU table object by clicking “Add” in that row.

Artifacts	Processing method	Data type									
Add all	Per Artifact (1)	165 ()									
	<table> <thead> <tr> <th>Name</th><th>Samples in Prep Info</th><th>Files</th></tr> </thead> <tbody> <tr> <td>Add</td><td>raw_biom_closed-reference (33736 - 2017-08-17 13:40:15)</td><td>30</td></tr> <tr> <td></td><td></td><td>test.biom</td></tr> </tbody> </table>	Name	Samples in Prep Info	Files	Add	raw_biom_closed-reference (33736 - 2017-08-17 13:40:15)	30			test.biom	
Name	Samples in Prep Info	Files									
Add	raw_biom_closed-reference (33736 - 2017-08-17 13:40:15)	30									
		test.biom									

Now, the second-right-most icon at the top bar should turn green, indicating that there are samples selected for analysis.



Clicking on the icon will take you to a page where you can refine the samples you want to include in your analysis. Here, all 30 of our samples are currently included:

You could optionally exclude particular samples from this set by clicking on “Show/Hide samples”, which will show each individual sample name along with a “remove” option. (Removing them here will mask them from the analysis, but will not affect the underlying files in any way.)

This should be good for now. Click the “Create Analysis” button, enter a name and description, then click “Create Analysis”.

This brings you to the processing network page. Here, pulling down the “Processing Network” tab. This may take 2 to 5 minutes to load. You can analyze data that has been run.

Before we process the data, let’s have a look at the summary of the contents of the biom file. Select the “dflt_name (BIOM)” artifact to see a summary of this file displaying a table summary, details regarding the frequency per sample, histogram of the number of features per sample:

As you can see, this file contains 30 samples with roughly 36,000 features. The features in our case are OTUs (Operational Taxonomic Units), because the features were generated using the closed-reference OTU picking.

Question

Selected Samples

[Create Analysis](#) [Clear Selected](#)

CMI workshop analysis

Processed Data

Id	Datatype	Processed Date	Algorithm	Parameters	Samples selected from Prep Info	
33738 	16S	2017-08-17 13:40:15.365324	None		30	Show/Hide samples Remove

Create new analysis ×

Analysis name

Description

Create analysis

Processing network

[Hide](#)

Click on the graph to navigate through it. Click circles for more information. This graph will refresh in 5 seconds or reload [now](#)

Job status (circles): success running error in_construction queued waiting deleting

Artifact status (triangles): artifact type outdated deprecated

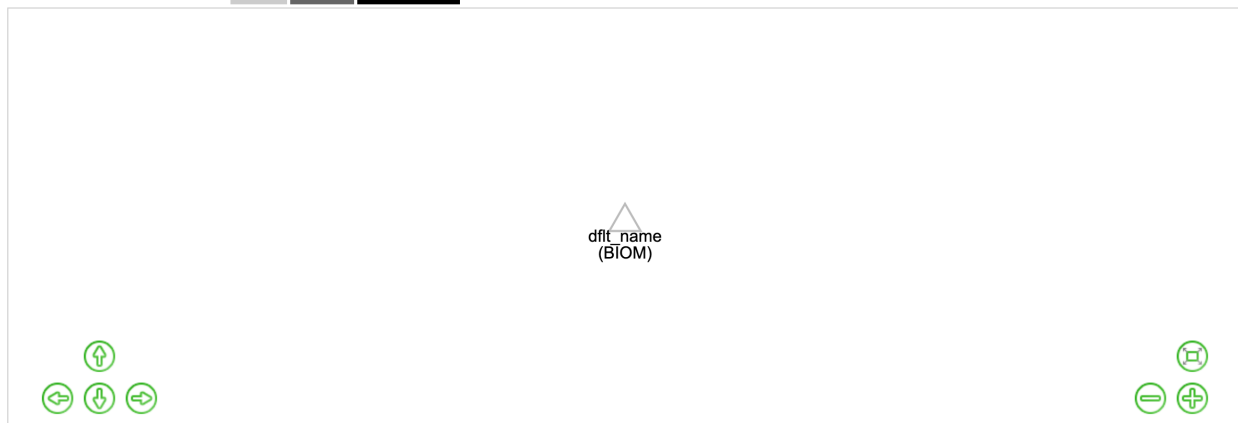


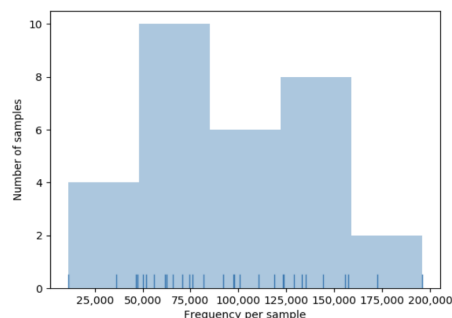
Table summary

Metric	Sample
Number of samples	30
Number of features	35,723
Total frequency	2,879,112

Frequency per sample

	Frequency
Minimum frequency	11,030.0
1st quartile	61,778.25
Median frequency	94,809.0
3rd quartile	127,517.25
Maximum frequency	196,075.0
Mean frequency	95,970.4

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



Are OTUs equivalent to bacterial species? Please, provide a justification to your answer. You may find the [Qiita glossary](#) useful.

Now we can begin analyzing these samples. Let's go ahead and select "dft_name (BIOM)" then select "Process". This will take us to the commands selection page. Once there, the commands pull down tab can be accessed which will display twenty-five actions.

The text in brackets is the actual underlying commands from QIIME2. We will now go through the use of some of the most-used commands which will enable you to generate summaries, plot your data, and calculate statistics to help you get the most out of your data.

1.13 Rarefying Data

For certain analyses such as those we are about to conduct, the data should be *rarefied*. This means that all the samples in the analysis will have their features, in this case OTUs, randomly subsampled to the same, desired number, reducing potential alpha and beta diversity biases. Samples with fewer than this number of features will be excluded, which can also be useful for excluding things like blanks. To choose a good cutoff for your data, view the histogram that was made when we generated the summary of the data.

An appropriate cutoff would exclude clear outliers, but retain most of the samples. Here we have already removed blanks from our data and eliminated the outliers prior to analysis so we will just use the minimum number of features observed in our samples (11030) as the cutoff.

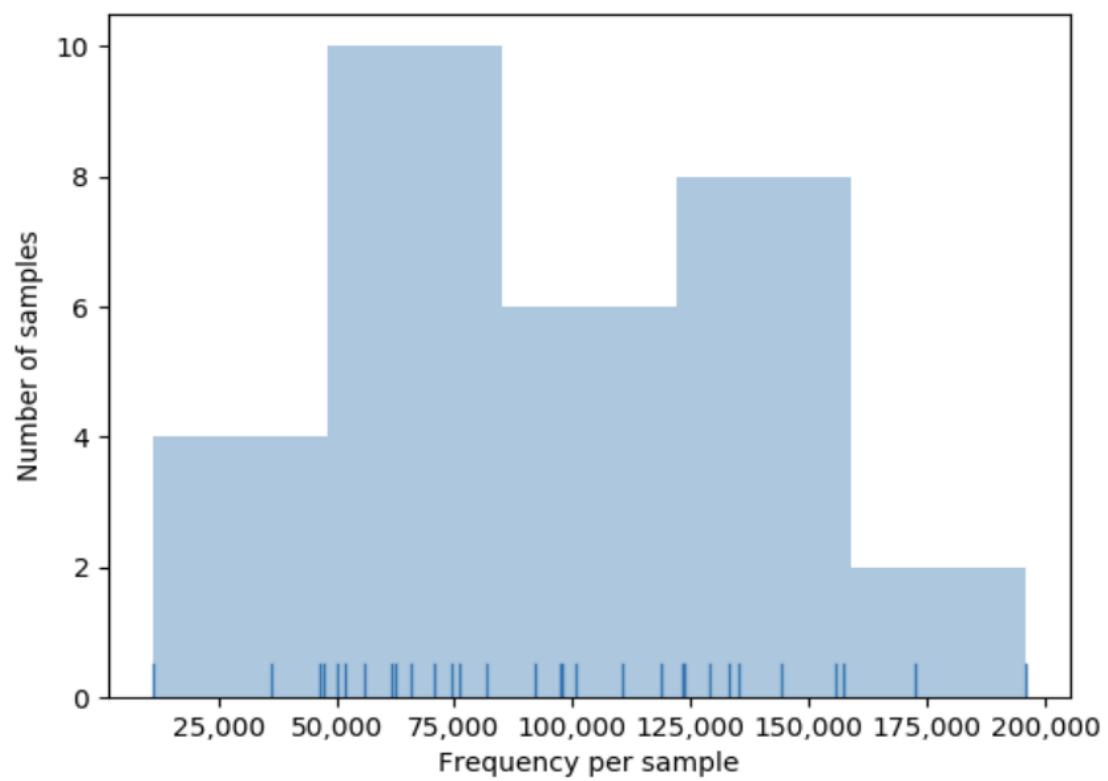
To rarefy the data, select "Rarefy table" from the drop-down menu. The parameters will appear below the workflow diagram:

Several parameters will have only one option which will be automatically selected for you. In the field, "The total frequency that each sample should be rarefied to... (sampling depth)", we will specify the number of features to rarefy our samples to. Enter "11030" in this box, and click "Add Command".

Click the "Run" button above the workflow network to start the process of rarefaction. Then, click on the "dft_name (BIOM)" artifact to see blue "Jobs using this data" button. Once you click on it, you can see the current status of your job. You can also view it clicking on the server button in the top-right corner of the screen:

✓ Choose command...

Add pseudocount to table [add_pseudocount]
 Alpha diversity (phylogenetic) [alpha_phylogenetic]
 Alpha diversity [alpha]
 Alpha rarefaction curves [alpha_rarefaction]
 Beta diversity (phylogenetic) [beta_phylogenetic]
 Beta diversity [beta]
 Beta diversity rarefaction [beta_rarefaction]
 Collapse features by their taxonomy at the specified level [collapse]
 Compute first differences or difference from baseline between sequential states [first_differences]
 Convert (and merge) positive numeric metadata (in)to feature table. [metatable]
 Convert to presence/absence [presence_absence]
 Convert to relative frequencies [relative_frequency]
 Core diversity metrics (non-phylogenetic) [core_metrics]
 Core diversity metrics (phylogenetic and non-phylogenetic) [core_metrics_phylogenetic]
 Filter features from table [filter_features]
 Filter samples from table [filter_samples]
 Generate a heatmap representation of a feature table [heatmap]
 Generate heatmap of important features. [heatmap]
 Generate interactive volatility plot [volatility]
 Group samples or features by a metadata column [group]
 Identify core features in table [core_features]
 Linear mixed effects modeling [linear_mixed_effects]
 Nested cross-validated supervised learning classifier. [classify_samples_ncv]
 Nested cross-validated supervised learning regressor. [regress_samples_ncv]
 Nonparametric microbial interdependence test [nmit]
 Paired difference testing and boxplots [pairwise_differences]
 Plot longitudinal feature volatility and importances [plot_feature_volatility]
 Pre-fitted sklearn-based taxonomy classifier [classify_sklearn]
 Rarefy table [rarefy]
 Remove features from table if they're not present in tree. [filter_table]
 Split a feature table into training and testing sets. [split_table]
 Subsample table [subsample]
 Summarize table [summarize]
 Taxonomy-based feature table filter. [filter_table]
 Transpose a feature table. [transpose]
 Visualize taxonomy with an interactive bar plot [barplot]



Choose command:

Rarefy table [rarefy]

Required parameters:

The feature table to be rarefied.:

dflt_name (BIOM)

Optional parameters:

Parameter set:

Default

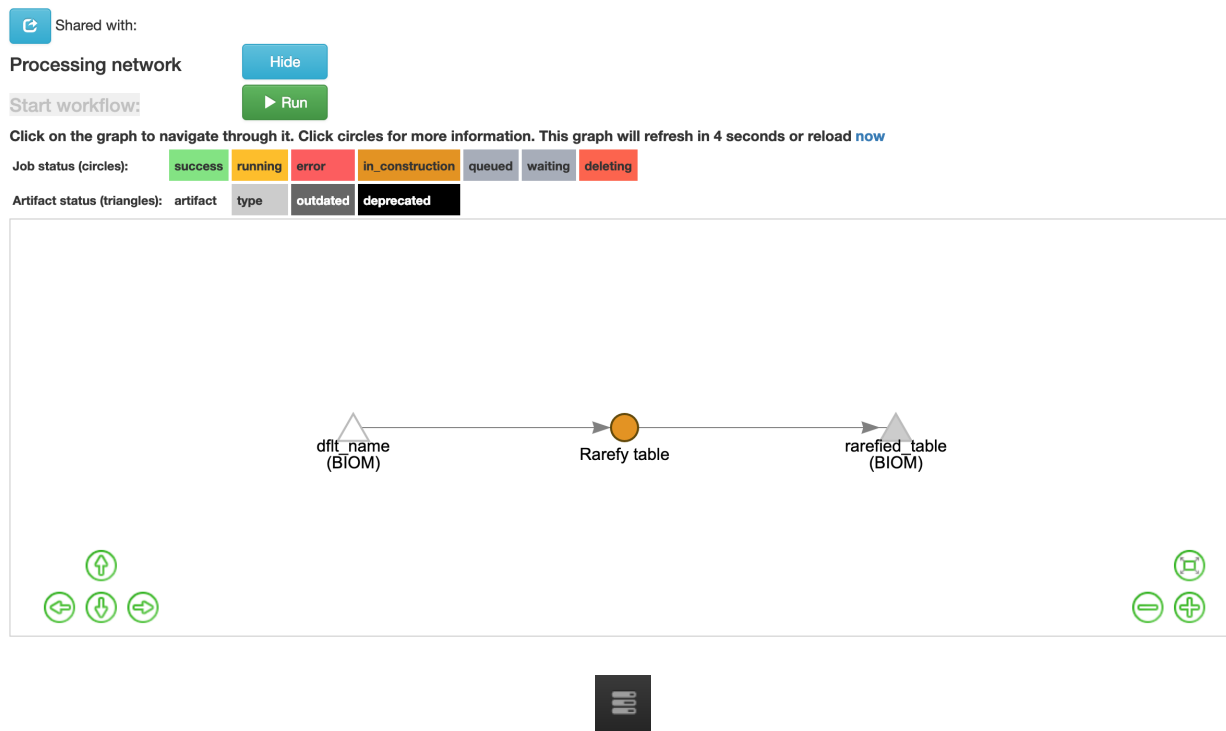
Rarefy with replacement by sampling from the multinomial distribution instead of rarefying without replacement. (with replacement):

☐

The total frequency that each sample should be rarefied to. Samples where the sum of frequencies is less than the sampling depth will be not be included in the resulting table unless subsampling is performed with replacement. (sampling depth):

11030

Add Command



The view will return to the original screen, while the rarefied feature-table generation job runs. Your browser will automatically refresh every 15 seconds until the “rarefied table (BIOM)” artifact appears:

Select the newly generated “rarefied_table (BIOM)” artifact. This time instead of seeing a histogram of the rarefied samples, you instead see a brief summary confirming that your samples have all be rarefied to the same depth. Now that the data are rarefied, we can begin the analysis.

1.14 Taxa Bar Plots

NOTE

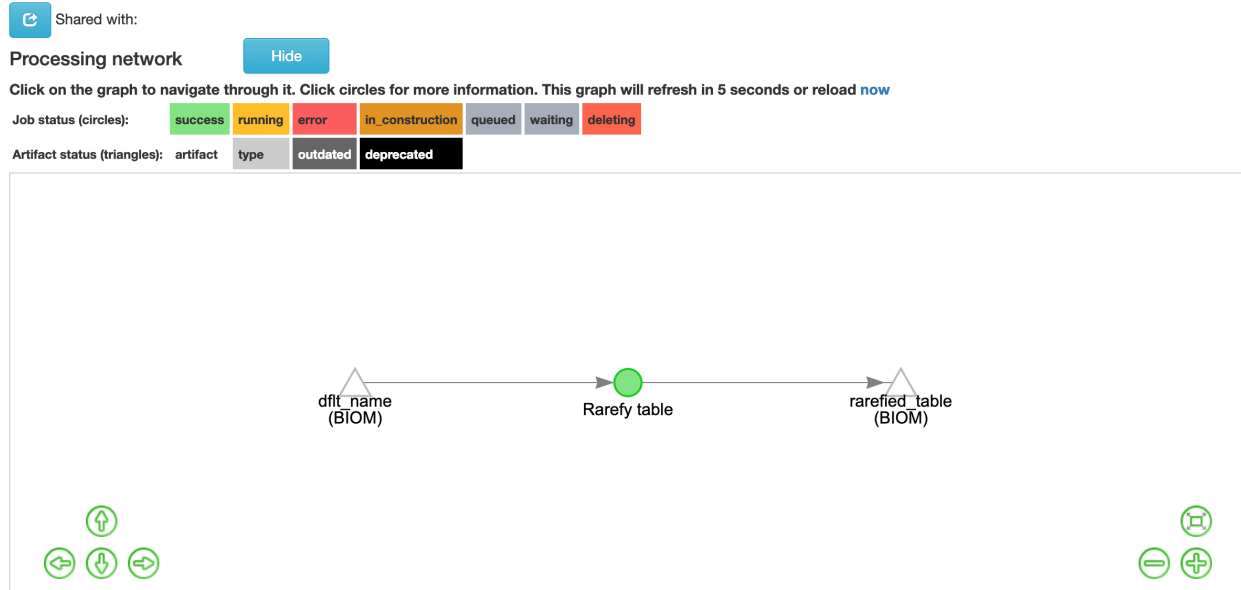
Taxonomy is outside the scope of this lab session. However, if you are interested in this topic, you are encouraged to follow the [CMI Qiita/GNPS tutorial afterwards](#).

1.15 Alpha Diversity Analysis

Now, let’s analyze the alpha diversity of your samples. Alpha diversity metrics describe the diversity of features within a sample or a group of samples. This is used to analyze the diversity within rather than between samples or a group of samples.

1.15.1 Observed Operational Taxonomic Units

One type of analysis for alpha diversity, and the simplest, is looking at the number of observed, unique features, or OTUs in this example, also known as feature richness. This type of analysis will provide the number of unique OTUs



found in a sample or group of samples.

To perform an alpha diversity analysis of feature richness, select the rarefied “rarefied table (BIOM)” artifact in the processing network and select “Process”. Select “Alpha diversity” from the drop-down menu. The parameters will appear below the workflow diagram:

Several parameters have been automatically selected for you since these options cannot be changed. In the field, “The alpha diversity metric... (metric)”, we will specify the alpha diversity metric to run in our analysis. Select “Number of distinct features” from the drop-down menu in this box, and click “Add Command”.

Once the command is added the workflow should appear as follows:

Click the run button to start the process of the alpha diversity analysis. The view will return to the original screen, while the alpha diversity analysis job runs.

1.15.2 Faith’s Phylogenetic Diversity Index

Another alpha diversity analysis in this tutorial uses Faith’s phylogenetic diversity index. This index also measured abundance and diversity but considers the phylogenetic distance spanning all features in a sample. The results can also be displayed as a phylogeny, rather than as a plot.

To perform an alpha diversity analysis using Faith’s phylogenetic diversity index, select the “rarefied table (BIOM)” artifact in the processing network and select “Process”. Select “Alpha diversity (phylogenetic)” from the drop-down menu. The parameters will appear below the workflow diagram:

Several parameters have been automatically selected for you. For example, in the field, “The alpha diversity metric... (metric)”, “Faith’s Phylogenetic Diversity” has already been chosen from the drop-down menu in this box. In the “Phylogenetic tree” field select “/databases/gg/13_8/trees/97_otus_no_none.tree” then click “Add Command”.

Once the command is added the workflow should appear as follows:

Click the run button to start the process of the alpha diversity analysis. The view will return to the original screen, while the alpha diversity analysis job runs.

Shared with: [Hide](#)

Processing network

Click on the graph to navigate through it. Click circles for more information. This graph will refresh in 13 seconds or reload [now](#)

Job status (circles): success running error in_construction queued waiting deleting

Artifact status (triangles): artifact type outdated deprecated

Choose command:

Required parameters:

The feature table containing the samples for which alpha diversity should be computed.:

Optional parameters:

Parameter set:

The alpha diversity metric to be computed. (metric):

[Add Command](#)

Shared with: [Hide](#)

Processing network

Start workflow: [Run](#)

Click on the graph to navigate through it. Click circles for more information. This graph will refresh in 5 seconds or reload [now](#)

Job status (circles): success running error in_construction queued waiting deleting

Artifact status (triangles): artifact type outdated deprecated

Processing network Hide

Click on the graph to navigate through it. Click circles for more information. This graph will refresh in 7 seconds or reload [now](#)

Job status (circles): success running error in_construction queued waiting deleting

Artifact status (triangles): artifact type outdated deprecated

Choose command:

Required parameters:

The feature table containing the samples for which alpha diversity should be computed.:

Optional parameters:

Parameter set:

Phylogenetic tree:

The alpha diversity metric to be computed. (metric):

Add Command

Shared with:

Processing network Hide

Run

Click on the graph to navigate through it. Click circles for more information. This graph will refresh in 1 seconds or reload [now](#)

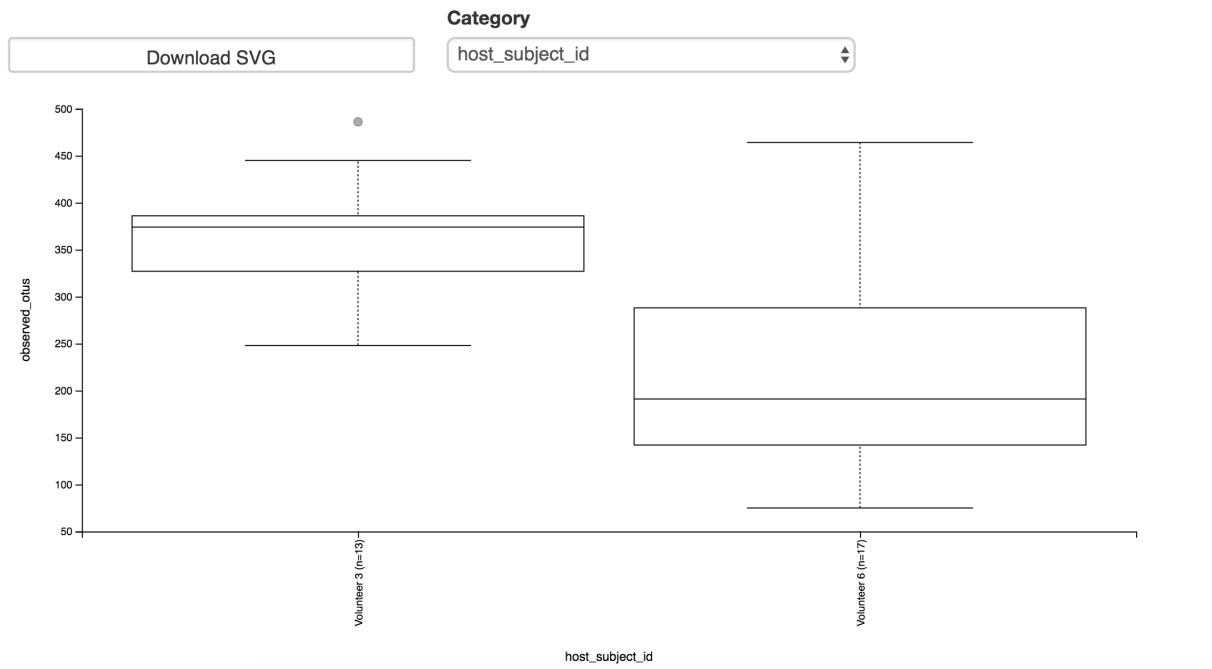
Job status (circles): success running error in_construction queued waiting deleting

Artifact status (triangles): artifact type outdated deprecated

1.15.3 Alpha Diversity Outputs

Each alpha diversity analysis will output an interactive boxplot that shows how that alpha diversity metric correlates with different metadata categories:

Alpha Diversity Boxplots



To change the category, choose the “Category” pull-down menu and choose the metadata category you would like to analyze:

You will also be given the outcomes to Kruskal-Wallis tests:

Question

Which alpha diversity metric produces a higher between-subject effect size?

1.16 Beta Diversity Analysis

One can also measure beta diversity in Qiita. Beta diversity measures feature turnover among samples (i.e., the diversity between samples rather than within each sample). This is used to compare samples to one another.

1.16.1 Bray-Curtis Dissimilarity

One commonly used beta diversity metric is Bray-Curtis dissimilarity. This metric quantifies how dissimilar samples are to one another.

To perform an analysis of beta diversity using the Bray-Curtis dissimilarity metric, select the “rarefied table (BIOM)” artifact in the processing network and select “Process”. Then select “Beta diversity” from the drop-down menu. The parameters will appear below the workflow diagram:

Category

✓ host_subject_id
 sex
 subject
 extraction_robot
 side
 product_metabolites
 extractionkit_lot
 collection_date
 timepoint
 well_id
 lane
 processing_robot
 common_body_product_chemicals_present
 LinkerPrimerSequence
 BarcodeSequence
 phase
 sample_plate
 tm50_8_tool
 tm300_8_tool

Kruskal-Wallis (all groups)

	Result
H	9.723809725184108
p-value	0.0018189808066191715

Kruskal-Wallis (pairwise)

[Download CSV](#)

Group 1	Group 2	H	p-value	q-value
Volunteer 3 (n=13)	Volunteer 6 (n=17)	9.72381	0.001819	0.001819

Choose command:

Beta diversity [beta]

Required parameters:

The feature table containing the samples over which beta diversity should be computed.:

Rarefied 11030 (BIOM)

Optional parameters:

Parameter set:

Default

A pseudocount to handle zeros for compositional metrics. This is ignored for other metrics. (pseudocount):

1

The beta diversity metric to be computed. (metric):

Aitchison distance

The number of jobs to use for the computation. This works by breaking down the pairwise matrix into n jobs even slices and computing them in parallel. If -1 all CPUs are used. If 1 is given, no parallel computing code is used at all, which is useful for debugging. For n_jobs below -1, $(n_cpus + 1 + n_jobs)$ are used. Thus for $n_jobs = -2$, all CPUs but one are used. (Description from `sklearn.metrics.pairwise_distances`) (n_jobs):

1

Add Command

Several parameters have been automatically selected for you. In the field, “The beta diversity metric... (metric), we will specify the beta diversity analysis to run. Select “Bray-Curtis dissimilarity” from the drop-down menu in this box, and click “Add Command”.

To create a principal coordinates plot of the Bray-Curtis dissimilarity distance matrix, select the “distance matrix (distance matrix)” artifact and select “Process”. Select “Perform Principal Coordinate Analysis (PCoA)” from the drop-down menu. The parameters will appear below the workflow diagram:

All of the parameter have automatically selected for you just click “Add Command”.

Once the command is added the workflow should appear as follows:

Click the run button to start the process of the beta diversity analysis. The view will return to the original screen, while the beta diversity analysis job runs.

1.16.2 Unweighted UniFrac Analysis

Another commonly used distance metric for measuring beta diversity is unweighted UniFrac distance. *Unweighted* refers to that the metric considers only feature richness and not abundance, when comparing samples to one another. This differs from the weighted UniFrac distance metric, which takes into account both feature richness and abundance, for each sample.

To perform unweighted UniFrac analysis, select the “rarefied table (BIOM)” artifact in the processing network and select “Process”. Then select “Beta diversity (phylogenetic)” from the drop-down menu. The parameters will appear below the workflow diagram:

All of the parameters have been automatically selected for you, just click “Add Command”.

To create a principal coordinates plot of the unweighted Unifrac distance matrix, select the “distance_matrix (distance_matrix)” artifact that will be generated using Unweighted UniFrac distance. Note that, unless you rename each distance matrix (see below: Altering Workflow Analysis Names), they will appear identical until you select them to view their provenance information. Once you have selected the distance matrix artifact, select “Perform Principal Coordinate Analysis (PCoA)” from the drop-down menu. The parameters will appear below the workflow diagram:

All of the parameters have been automatically selected for you just click “Add Command”. Once the command is added the workflow should appear as follows:

Click the run button to start the process of the beta diversity analysis. The view will return to the original screen, while the beta diversity analysis job runs.

Question

Is there a scenario in which unweighted UniFrac value can be < 0 ? Which of the two distance metrics used produces more homogenous results (eg. smaller variance)?

1.16.3 Principal Coordinate Analysis

Clicking on the “pcoa (ordination_results)” (Principal Coordinate Analysis) artifact will open an interactive visualization of the similarity among your samples. Generally speaking, the more similar the samples with respect to their features, the closer they are likely to be in the PCoA ordination plot. The Emperor visualization program offers a very useful way to explore how patterns of similarity in your data associate with different metadata categories.

Once the Emperor visualization program loads, the PCoA result will look like:

You will see tabs including “Color”, “Visibility”, “Opacity”, “Scale”, “Shape”, “Axes”, and “Animations”.

Under “Color” you will notice two pull-down menus:

Choose command:

Principal Coordinate Analysis [pcoa]

Required parameters:

The distance matrix on which
PCoA should be computed.:

distance_matrix (distance_matrix)

Optional parameters:

Parameter set:

Default

Dimensions to reduce the distance matrix to. This number determines how many eigenvectors and eigenvalues are returned, and influences the choice of algorithm used to compute them. By default, uses the default eigendecomposition method, SciPy's `eigh`, which computes all eigenvectors and eigenvalues in an exact manner. For very large matrices, this is expected to be slow. If a value is specified for this parameter, then the fast, heuristic eigendecomposition algorithm `fsvd` is used, which only computes and returns the number of dimensions specified, but suffers some degree of accuracy loss, the magnitude of which varies across different datasets. (number of_dimensions):

Add Command



Under “Select a Color Category” you can select how the samples will be grouped. Under “Classic QIIME Colors”, you can select how each group will be colored.

Under the “Visibility” tab you will notice 1 pull-down menu:

Under “Select a Visibility Category” you can select which group will be displayed on the PCoA plot.

Under the “Opacity” tab you will notice 1 pull-down menu:

Under “Select an Opacity Category” you can select the categories in which the opacity will change on the PCoA plot. Once chosen, these groups will be displayed under “Global Scaling” and, when selected, you can change the opacity of each group separately. Under “Global Scaling” you can change the opacity of all of the samples.

Under the “Scale” tab you will notice 1 pull-down menu:

Under “Select a Scale Category” you can choose the grouping of your samples. Under “Global Scaling” you can change the point size for each group on the PCoA plot.

Under the “Shape” tab you will notice 1 pull-down menu:

Under “Select a Shape Category” you can alter the shape of each group on the PCoA plot to the following:

Under the “Axis” tab you will notice 5 pull-down menus:

The first 3 pull-down menus located under “Visible” allow you to change the axis that are being displayed. The “Axis and Labels Color” menu allow you to change the color of your axis and label of the PCoA. The “Background Color” menu allows you to change the color of the background of the PCoA. The % Variation Expanded graph displays how different the most dissimilar samples are by percentage for each axis that can be used.

Under the “Animations” tab you will notice 2 pull-down menus:

Under “Category to sort samples” you can choose the category that you will be sorting the samples by. Under “Category to group sample” you can choose the category that you will be grouping the samples by.

Let’s take a few minutes now to explore the various features of Emperor. Open a new browser window with the [Emperor tutorial](#) and follow along with your test data.

Question

From the unweighted UniFrac PCoA plot, what is the main driver of bacterial community separation, subject (*host_subject_id*), body side (*side*), or phase of the experiment (*phase_discreet*)? Is the same true for Bray-Curtis results?

Required parameters:

The feature table containing the samples over which beta diversity should be computed.:

Rarefied 11030 (BIOM)

Optional parameters:

Parameter set:

Default

In a bifurcating tree, the tips make up about 50% of the nodes in a tree. By ignoring them, specificity can be traded for reduced compute time. This has the effect of collapsing the phylogeny, and is analogous (in concept) to moving from 99% to 97% OTUs (bypass tips):

☐

Perform variance adjustment based on Chang et al. BMC Bioinformatics 2011. Weights distances based on the proportion of the relative abundance represented between the samples at a given node under evaluation. (variance adjusted):

☐

Phylogenetic tree:

/databases/gg/13_8/trees/97_otus_no_none.tree

The beta diversity metric to be computed. (metric):

Unweighted UniFrac

The number of workers to use. (n jobs):

1

This parameter is only used when the choice of metric is generalized unifrac. The value of alpha controls importance of sample proportions. 1.0 is weighted normalized UniFrac.

40.0 is close to unweighted UniFrac, but only if the sample proportions are dichotomized. (alpha):

Shared with: Hide Run

Processing network

Start workflow: Run

Click on the graph to navigate through it. Click circles for more information. This graph will refresh in 6 seconds or reload [now](#)

Circle status: success running error in_construction queued waiting deleting

Circle types: artifact type deprecated

Choose command:

Required parameters:

Dimensions to reduce the distance matrix to. This number determines how many eigenvectors and eigenvalues are returned and influences the choice of algorithm used to compute them. By default, uses the default eigendecomposition method, SciPy's eigh, which computes all eigenvectors and eigenvalues in an exact manner. For very large matrices, this is expected to be slow. If a value is specified for this parameter, then the fast, heuristic eigendecomposition algorithm fvdb is used, which only computes and returns the number of dimensions specified, but suffers some degree of accuracy loss, the magnitude of which varies across different datasets. (number of dimensions):

The distance matrix on which PCoA should be computed:

Optional parameters:

Parameter set:

Add Command

Shared with: Hide Run

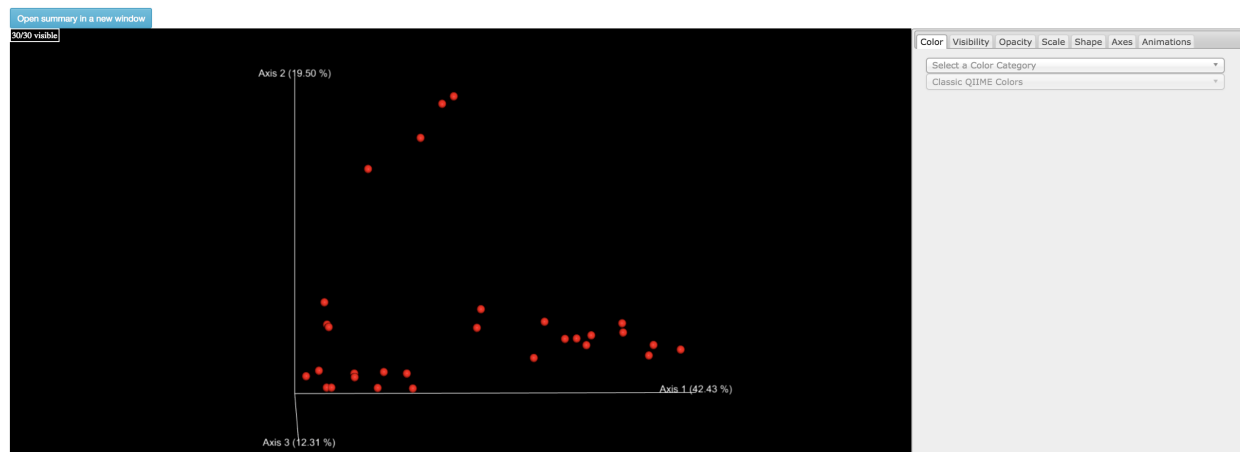
Processing network

Start workflow: Run

Click on the graph to navigate through it. Click circles for more information. This graph will refresh in 0 seconds or reload [now](#)

Circle status: success running error in_construction queued waiting deleting

Circle types: artifact type deprecated



Color | Visibility | Opacity | Scale | Shape | Axes

Animations

Select a Color Category

Classic QIIME Colors

Color | Visibility | Opacity | Scale | Shape | Axes

Animations

Select a Visibility Category

Color | Visibility | Opacity | Scale | Shape | Axes

Animations

Select a Opacity Category

☐ Change opacity by values

Global Scaling

1.0

Color
Visibility
Opacity
Scale
Shape
Axes

Animations

Select a Scale Category ▼

☐ Change scale by values

Global Scaling

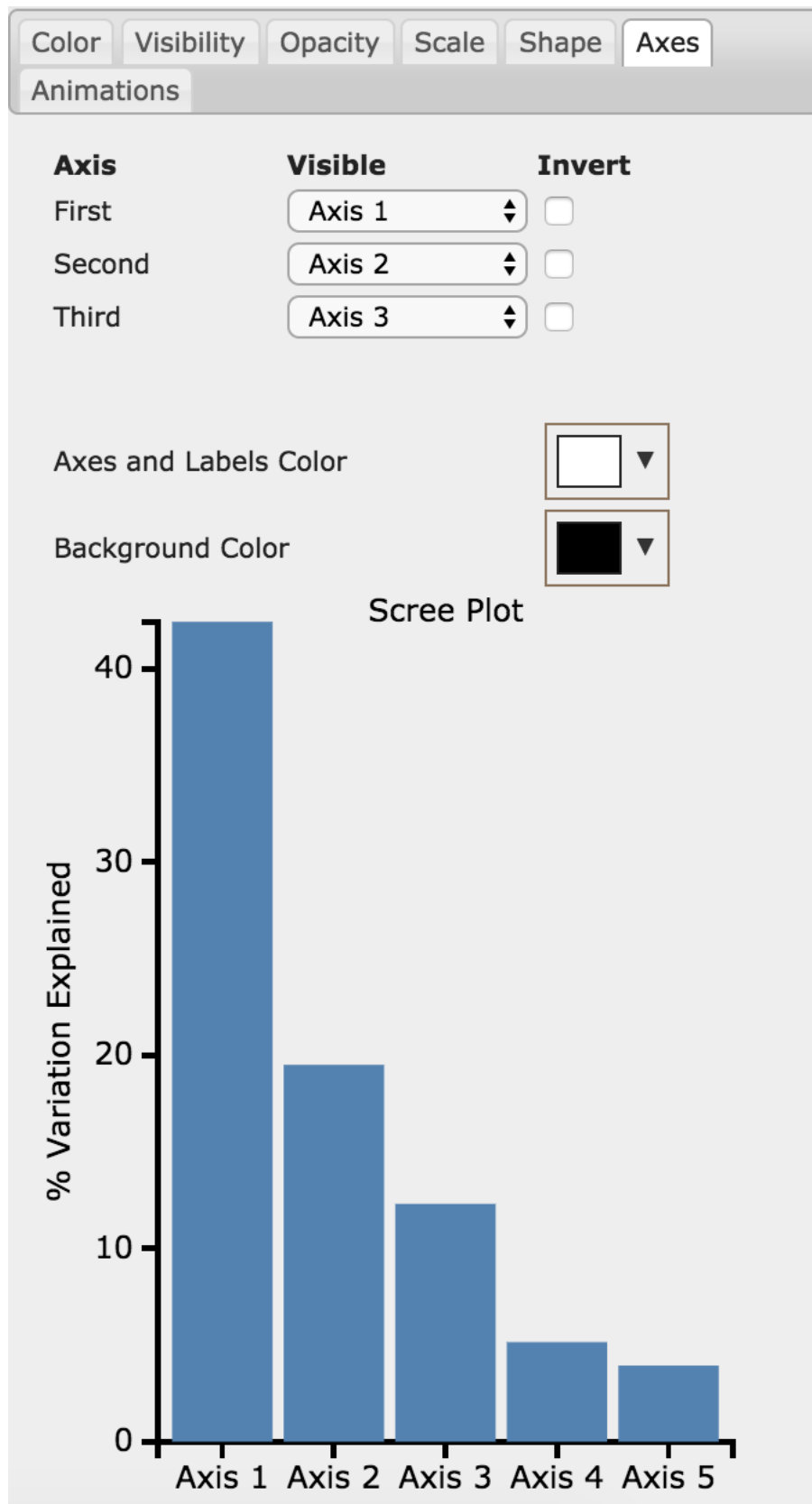
1.0

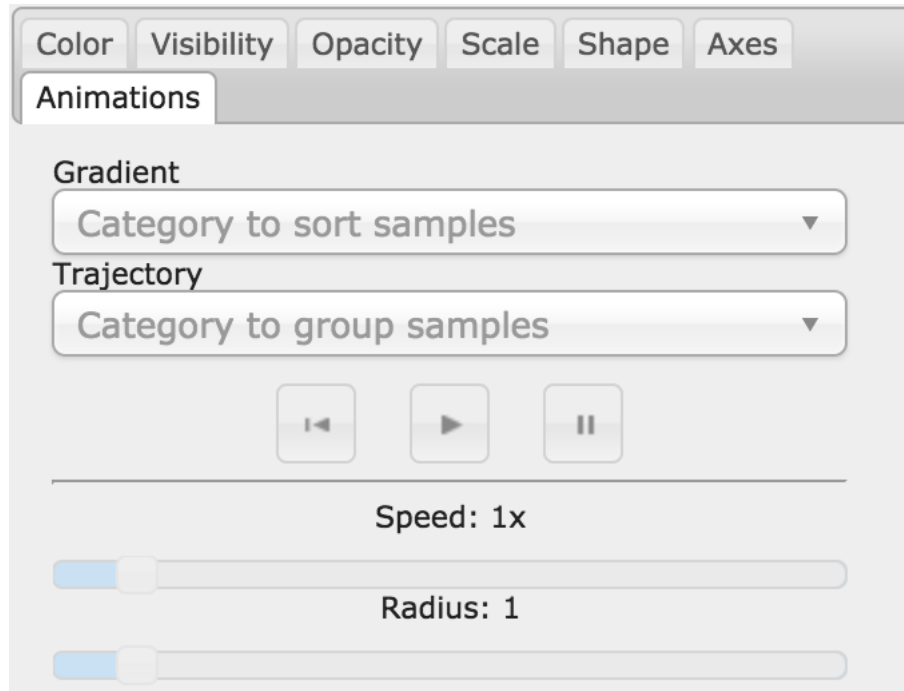
Color
Visibility
Opacity
Scale
Shape
Axes

Animations

Select a Shape Category ▼

- ✓ Sphere
- Cube
- Cone
- Icosahedron
- Cylinder





1.16.4 Beta Diversity Group Significance

Another way to study the beta diversity is by measuring the beta diversity group significance. Beta diversity group significance measures whether groups of samples are significantly different from one another using a permutation-based statistical test. Sample groups are designated by metadata variables.

If you have completed the tutorial up to this point, you can begin analysis of beta diversity group significance from one of your beta diversity distance matrices (jump down two paragraphs). Here we begin with the rarefied feature-table. To perform a beta group significance analysis, select the “rarefied table (BIOM)” artifact in the processing network and select “Process”. Select “Beta diversity” from the drop-down menu. The parameters will appear below the workflow diagram:

Shared with: Hide

Processing network

Click on the graph to navigate through it. Click circles for more information. This graph will refresh in 0 seconds or reload now

Circle status: success running error in construction queued waiting deleting

Circle types: artifact type deprecated

Choose command: Beta diversity

Required parameters:

The feature table containing the samples over which beta diversity should be computed:

rarefied_table (BIOM)

Optional parameters:

Parameter set: Default

A pseudocount to handle zeros for compositional metrics. This is ignored for other metrics. (pseudocount): 1

The beta diversity metric to be computed. (metric): Sokal-Michener coefficient

The number of jobs to use for the computation. This works by breaking down the pairwise matrix into n jobs even slices and computing them in parallel. If -1 all CPUs are used. If 1 is given, no parallel computing code is used at all, which is useful for debugging. For n jobs below -1, (n_cpus + 1 + n_jobs) are used. Thus for n_jobs = -2, all CPUs but one are used. (Description from sklearn.metrics.pairwise_distances) (n_jobs): 1

Add Command

Several parameters have been automatically selected for you. In the field, “The beta diversity metric... (metric)”, we will specify the beta diversity distance metric to use in our analysis. Note that if you attempt to create a distance matrix that already exists in the Processing network, you will get an error stating such. For example, if you have already created a beta diversity distance matrix using the Bray-Curtis dissimilarity metric, you will have to select a unique metric here (e.g., “Aitchison distance”). In the “Phylogenetic tree” field enter “/databases/gg/13_8/trees/97_otus.tree”, and click “Add Command”.

To create the beta group significance analysis, select the “distance_matrix (distance_matrix)” artifact of interest in the Processing network, and select “Beta diversity group significance” from the drop-down menu. The parameters will appear below the workflow diagram:

The screenshot shows the Qiita Processing network interface. At the top, there's a 'Shared with:' section and a 'Processing network' title. Below it, a 'Start workflow' button and a 'Run' button are visible. A status bar indicates 'Click on the graph to navigate through it. Click circles for more information. This graph will refresh in 1 seconds or reload now'. The status bar also shows 'Circle status: success running error in construction queued waiting starting' and 'Circle types: artifact type deprecated'. The main area displays a workflow diagram with nodes like 'data diversity', 'beta diversity', 'beta diversity (phylogenetic)', 'alpha diversity', 'alpha diversity (phylogenetic)', 'beta diversity (distance_matrix)', 'alpha diversity (distance_matrix)', 'beta diversity (distance_matrix)', 'alpha diversity (distance_matrix)', 'Principal Coordinates Analysis', and 'Visualize beta diversity with an'. Below the diagram, the 'Choose command:' dropdown is set to 'Beta diversity group significance'. The 'Required parameters:' section shows 'Matrix of distances between pairs of samples:' set to 'distance_matrix (distance_matrix)'. The 'Optional parameters:' section shows 'Parameter set:' set to 'Default'. The 'Metadata column to use:' field is empty. The 'Perform pairwise tests between all pairs of groups in addition to the test across all groups. This can be very slow if there are a lot of groups in the metadata column. (pairwise):' checkbox is unchecked. The 'The group significance test to be applied. (method):' dropdown is set to 'PERMANOVA'. The 'The number of permutations to be run when computing p-values. (permutations):' field is set to '999'. An 'Add Command' button is at the bottom left.

Several parameters have been automatically selected for you. In the “Metadata column to use” field we will specify the category from the metadata file to be used for determining significance between groups (e.g., subject). Using the “Perform pairwise tests...” checkbox we can indicate if we would like the group significance to be run “Pairwise”, otherwise the analysis will be done across all groups (i.e., Non-pairwise). Note that for metadata variables for which there are only two groups, this distinction makes no difference. In the field, “The group significance test... (method)”, we will specify the correlation test that will be applied (e.g., [PERMANOVA \[Permutational multivariate analysis of variance\]](#)). Then click “Add Command”. Once the command is added the workflow should appear as follows:

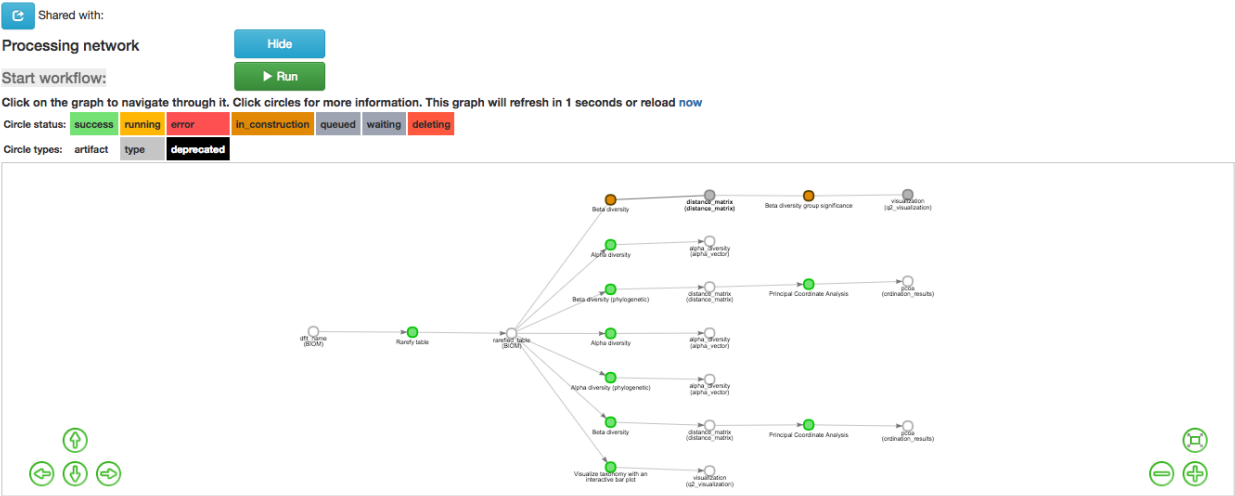
Click the run button to start the process of the beta diversity group significance analysis. The view will return to the original screen, while the beta diversity group significance analysis job runs.

Beta Group Significance Output Analysis

Once the beta group significance “visualization (q2_visualization)” artifact is chosen in the network, the beta diversity group significance Overview, which in our case shows results from the PERMANOVA (i.e., across all groups) and Group significance plots will appear:

The results from pairwise PERMANOVA tests will also be displayed if included in the analysis:

The command ‘Beta diversity group significance’ provides PERMANOVA that can be run on a single categorical metadata variable. If you instead would like to provide multiple terms in the form of an equation, you can use the



command ‘adonis PERMANOVA test for beta group significance’. This latter command implements the ‘adonis’ function from the R package, vegan.

NOTE

The sections below are **optional**. You can do them only if you have the time.

1.17 Filtering Data

Using QIITA you can also filter your data. This allows you to filter out samples.

To filter the data, select the “rarefied table (BIOM)” artifact in the processing network and select “Process”. Then select “Filter samples from table” from the drop-down menu. The parameters will appear below the workflow diagram:

Several parameters have been automatically selected for you. In the “SQLite WHERE-clause” field we are filtering out all samples except for certain samples. In this case we wanted to filter out all samples except those in which `subject = 'Volunteer 3'`, and click “Add Command”. If instead you want to filter out all of Volunteer 3’s samples, either use the SQLite WHERE-clause above while also checking the box “If true, the samples selected... will be excluded”, or alternatively use the SQLite WHERE-clause `subject != 'Volunteer 3'`, and click “Add Command”. If you want to filter for samples containing an apostrophe, write it out in the following format: `subject = \"Volunteer 3's samples\"`. **Keep in mind that all fields are case sensitive.**

Click “Run” to execute the filtering process.

An example of how you can use filtering in your analysis is explained in the following “Filtered Unweighted UniFrac Analysis” section.

1.18 Filtered Unweighted UniFrac Analysis

By filtering, you can perform unweighted UniFrac analysis but this time without certain sample.

After filtering your data (shown in the previous “Filtering Data” section), you can perform a beta diversity analysis by selecting the “filtered_table (BIOM)” in the Processing network and clicking “Process”. Select “Beta diversity (phylogenetic)” from the drop-down menu. The parameters will appear below the workflow diagram:

All of the parameters have been automatically selected for you, just click “Add Command”.

To create a principal coordinates plot of the unweighted Unifrac distance matrix, select the “distance_matrix (distance_matrix)” artifact that you set up above, and select “Perform Principal Coordinate Analysis (PCoA)” from the drop-down menu. The parameters will appear below the workflow diagram:

All of the parameters have been automatically selected for you just click “Add Command”. Once the command is added the workflow should appear as follows:

Click the run button to start the process of the beta diversity analysis. The view will return to the original screen, while the beta diversity analysis job runs.

1.19 Altering Workflow Analysis Names

To alter the name of a result, click the artifact then use the edit button on the processing network page.

This will cause a window to pop-up where you can input the name you’d like to replace it with.

Choose command:

Filter samples from table [filter_samples]

Required parameters:

The feature table from which samples should be filtered.:

Rarefied 11030 (BIOM)

Optional parameters:

Parameter set:

Default

If true, the samples selected by `metadata` or `where` parameters will be excluded from the filtered table instead of being retained. (exclude ids):

☐

SQLite WHERE clause specifying sample metadata criteria that must be met to be included in the filtered feature table. If not provided, all samples in `metadata` that are also in the feature table will be retained. (where):

The maximum number of features that a sample can have to be retained. If no value is provided this will default to infinity (i.e., no maximum feature filter will be applied). (max features):

The maximum total frequency that a sample can have to be retained. If no value is provided this will default to infinity (i.e., no maximum frequency filter will be applied). (max frequency):

The minimum number of features that a sample must have to be retained. (min features):

The minimum total frequency that a sample must have to be retained. (min frequency):

Add Command

Required parameters:

The feature table containing the samples over which beta diversity should be computed.:

Rarefied 11030 (BIOM)

Optional parameters:

Parameter set:

Default

In a bifurcating tree, the tips make up about 50% of the nodes in a tree. By ignoring them, specificity can be traded for reduced compute time. This has the effect of collapsing the phylogeny, and is analogous (in concept) to moving from 99% to 97% OTUs (bypass tips):

☐

Perform variance adjustment based on Chang et al. BMC Bioinformatics 2011. Weights distances based on the proportion of the relative abundance represented between the samples at a given node under evaluation. (variance adjusted):

☐

Phylogenetic tree:

/databases/gg/13_8/trees/97_otus_no_none.tree

The beta diversity metric to be computed. (metric):

Unweighted UniFrac

The number of workers to use. (n jobs):

1

This parameter is only used when the choice of metric is generalized unifrac. The value of alpha controls importance of sample proportions. 1.0 is weighted normalized UniFrac.

50.0 is close to unweighted UniFrac, but only if the sample proportions are dichotomized. (alpha):

[illegible]

Updating artifact 33455 name

Introduce the new name: