Qiita-GNPS-workshop

Release 0.01

May 08, 2017

Creating a study in Qiita

1 Qiita tutorials: 3

1.1 Getting CMI Workshop example data . . . 3
1.2 Setting up Qiita . . . 3
1.3 Creating a test study . . . 5
1.4 Creating an example study . . . 5
1.5 Adding sample information . . . 8
1.6 16S Data Processing in Qiita . . . 11
1.7 16S Microbiome Analysis in Qiita . . . 23
1.8 Notes on metabolomics . . . 33
1.9 Feature finding with MZmine2 . . . 35
1.10 Metabolomics demo data in Qiita . . . 46
1.11 GNPS tutorial for MS/MS data annotation . . . 46

Materials below are intended for CMI Qiita/GNPS workshop participants. They include all information covered during days 1 and 2 of the workshop.

For more information on Qiita, including Qiita philosophy and documentation, please visit Qiita website.

For general information about workshops, please contact Tomasz Kosciolek directly.

CHAPTER 1

Qiita tutorials:

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

Getting CMI Workshop example data

First, we'll download some example data. These files contain both 16S and shotgun metagenomics data for 12 samples from the American Gut Project.

For this tutorial, the relevant files are:

```
qiita-files/16S/*.fastq.gz        # The actual 16S sequences, one per sample
qiita-files/sample_information.txt   # The sample information file
qiita-files/prep_template_16S.txt # The prep information file
```

Next, we'll sign up for Qiita and create a study for these data.

Setting up Qiita

Signing up for a Qiita account

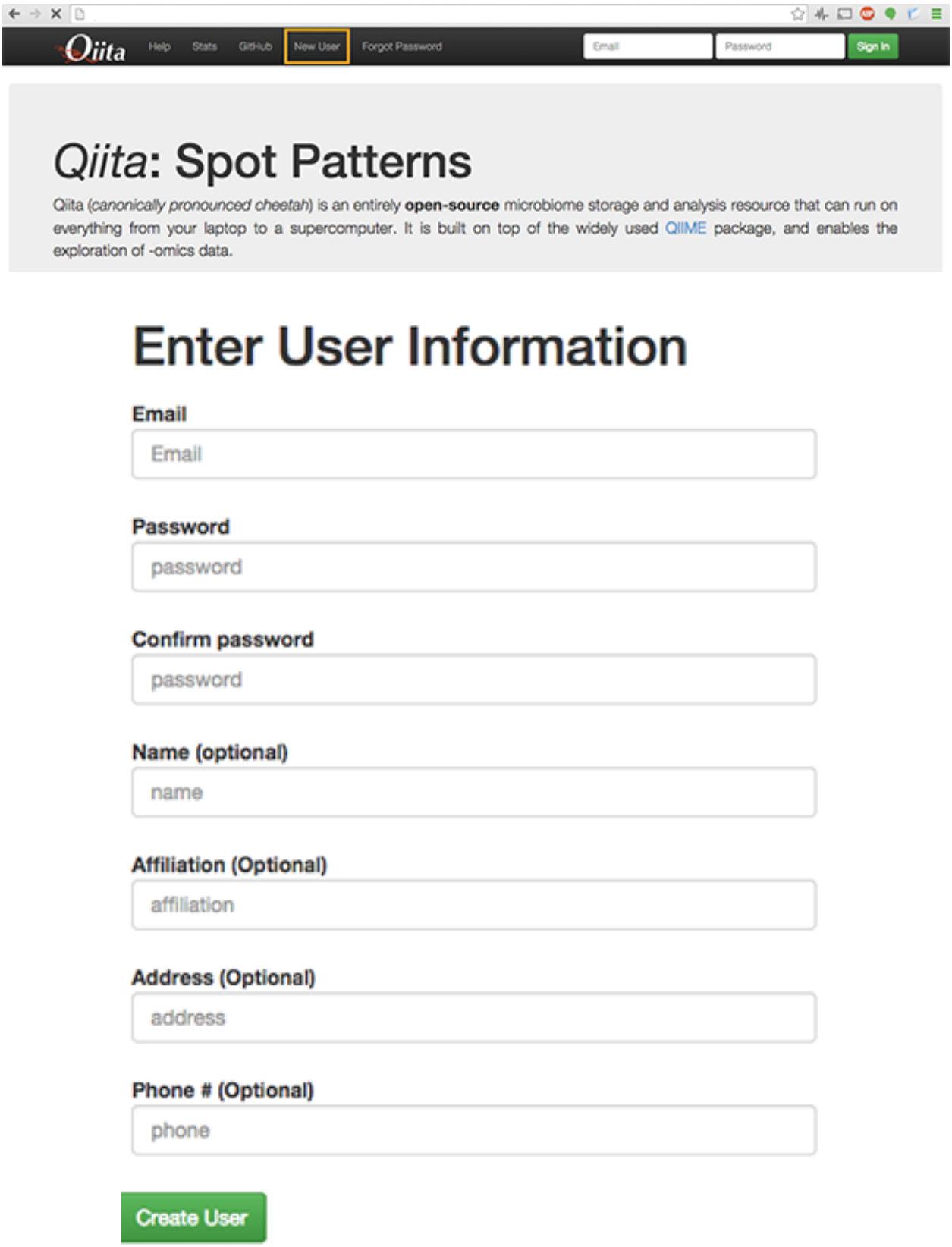
Open your browser (it must be Chrome or Firefox) and go to Qiita (https://qiita.ucsd.edu).

Click on "New User".

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

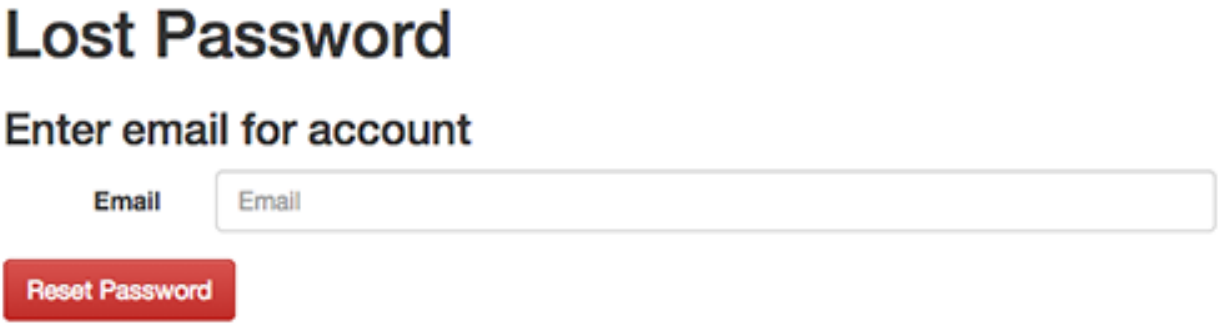
Logging into your account and resetting a forgotten password

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


Qiita
Help
Stats
GitHub
New User
Forgot Password
Email
Password
Sign in
Qiita: Spot Patterns
Qiita (canonically pronounced cheetah) is an entirely open-source microbiome storage and analysis resource that can run on everything from your laptop to a supercomputer. It is built on top of the widely used QIIME package, and enables the exploration of -omics data.
Enter User Information
Email
Email
Password
password
Confirm password
password
Name (optional)
name
Affiliation (Optional)
affiliation
Address (Optional)
address
Phone # (Optional)
phone
Create User

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.



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.

Creating a test study

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 this tutorial, our study contains 12 samples, each with two types of data: 16S and shotgun metagenomics. To represent this project in Qiita, you will need to create a single study with a single sample information file that contains all 12 samples. Then, you will link separate preparation files for each data type.

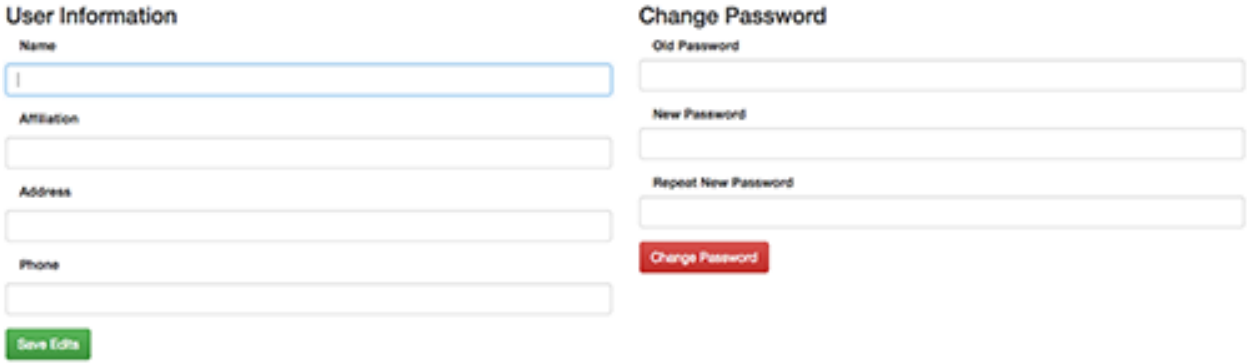
Creating an example study

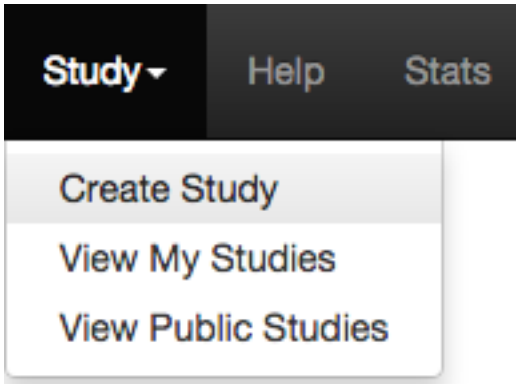
To create a study, click on the "Study" menu and then on "Create Study". This will take you to a new page that will gather some basic information to create your study.

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.


Welcome antgonza@gmail.com
Log Out


User Information
Name
Affiliation
Address
Phone
Save Edits
Change Password
Old Password
New Password
Repeat New Password
Change Password


Study
Help
Stats
Create Study
View My Studies
View Public Studies

Create a new Study

* = Required Field

Study Title JGS April 2017 CMI workshop *

Study Alias April 2017 CMI *

DOI
Just values, no links, comma separated values

PUBMED ID
Just values, no links, comma separated values

Study Abstract An example study *

Study Description Just a test *

Principal Investigator Rob Knight, UCSD *

Lab Person Jon Sanders, UCSD

Can't find the person you're looking for? Add a person

Environmental Packages *
You can select multiple entries by control-clicking (mac: command-clicking)

air
built environment
host-associated
human-amniotic-fluid
human-associated
human-blood
human-gut
human-oral
human-skin
human-urine
human-vaginal
microbial mat/biofilm
miscellaneous natural or artificial environment
plant-associated
sediment
soil
wastewater/sludge
water

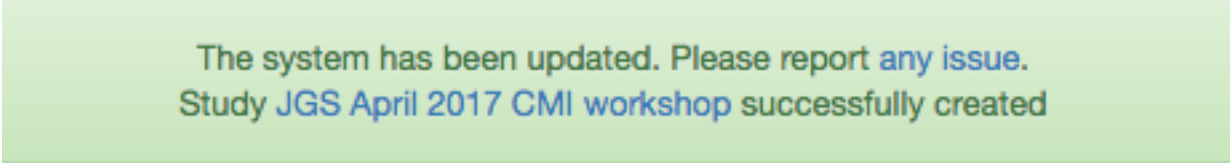
Event-Based Data No timeseries

Create Study

Select the environmental package appropriate to your study. Different packages will request different specific information about your samples. This information is optional; for more details, see the metadata section.

There is also an option to specify time series type ("Event-Based Data") if you have time series data. In our case, the samples come from a cross-sectional study design, so you should select "No time series." For more information on time series types, you can check out the in-depth tutorial on the Qiita website.

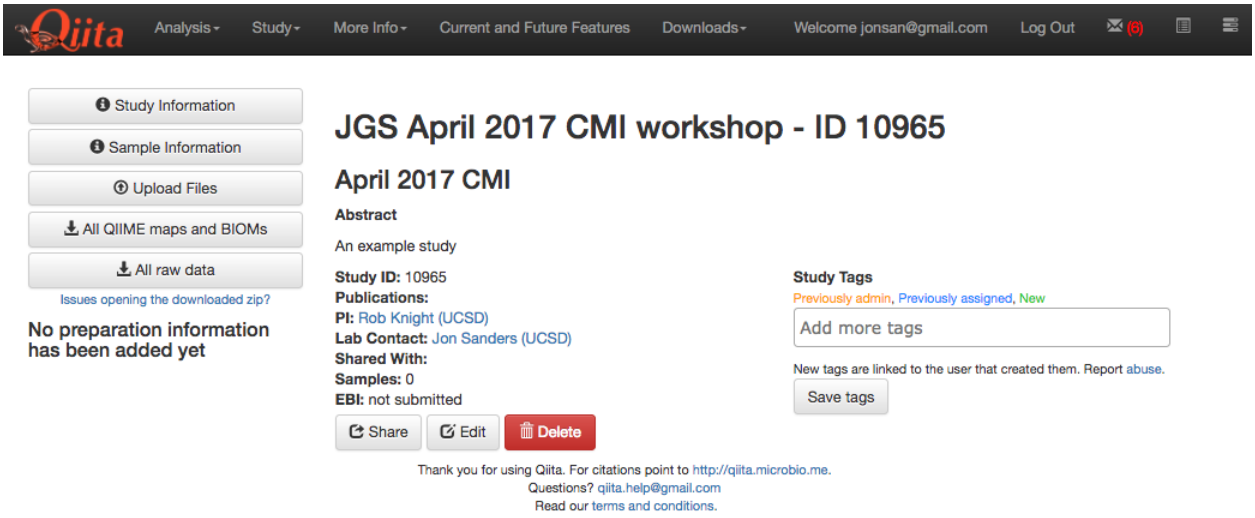
Once your study has been created, you will be informed by a green message; click on the study name to begin adding your data.



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.)

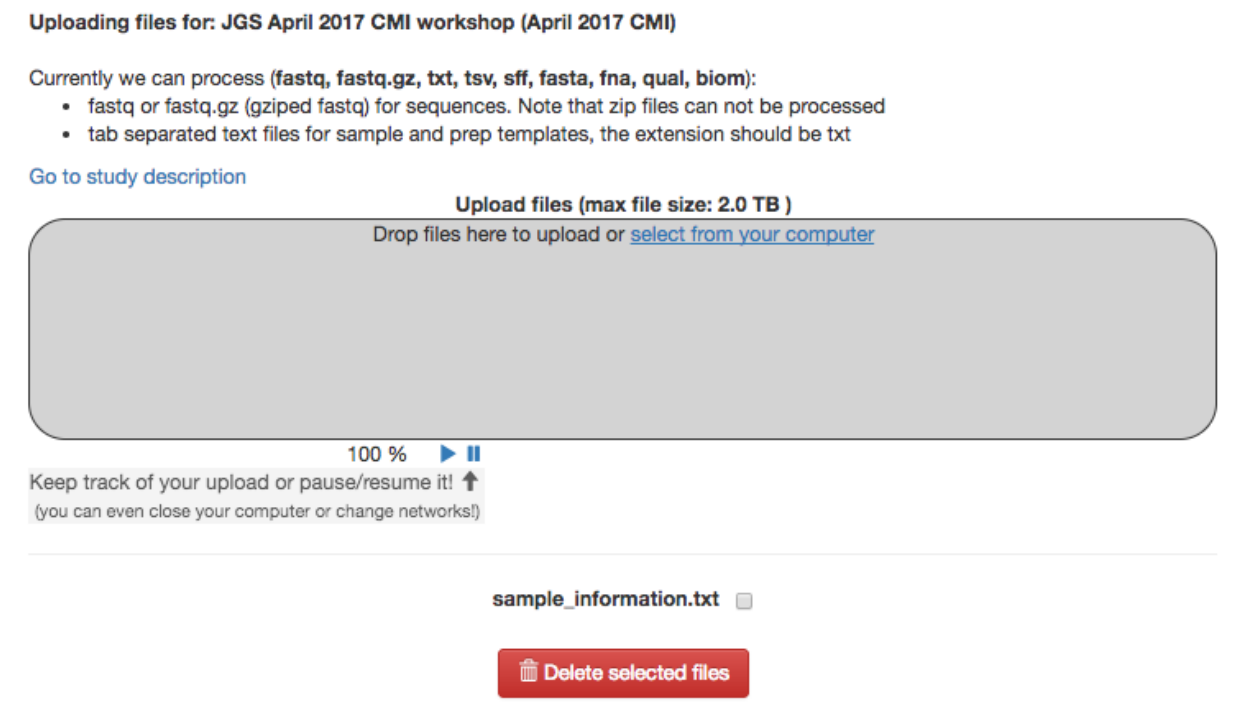
The first point of entrance to a study is the study description page. Here you will be able to edit the study info, upload files, and manage all other aspects of your study.



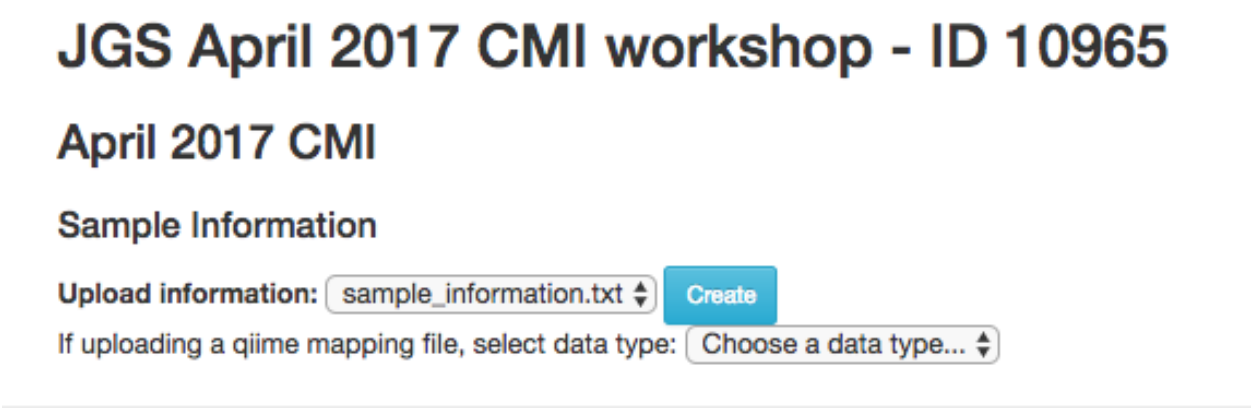
The first step after study creation is uploading files. Click on the "Upload Files" button: as shown in the figure below, you can now drag-and-drop files into the grey area or simply click on "select from your computer" to select the fastq, fastq.gz or txt files you want to upload.

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

Drag the file named "sample_information.txt" into the upload box. It should upload quickly and appear 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".



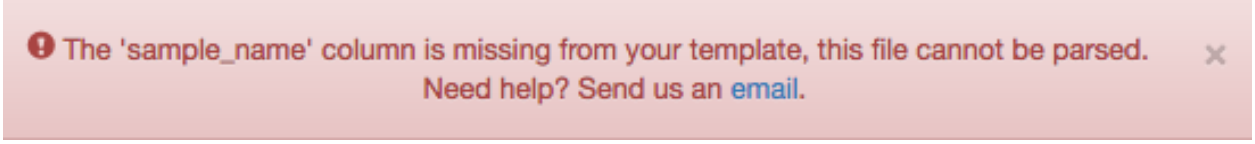
If something is wrong with the sample information file, Qiita will let you know with a red banner a the top of the screen.

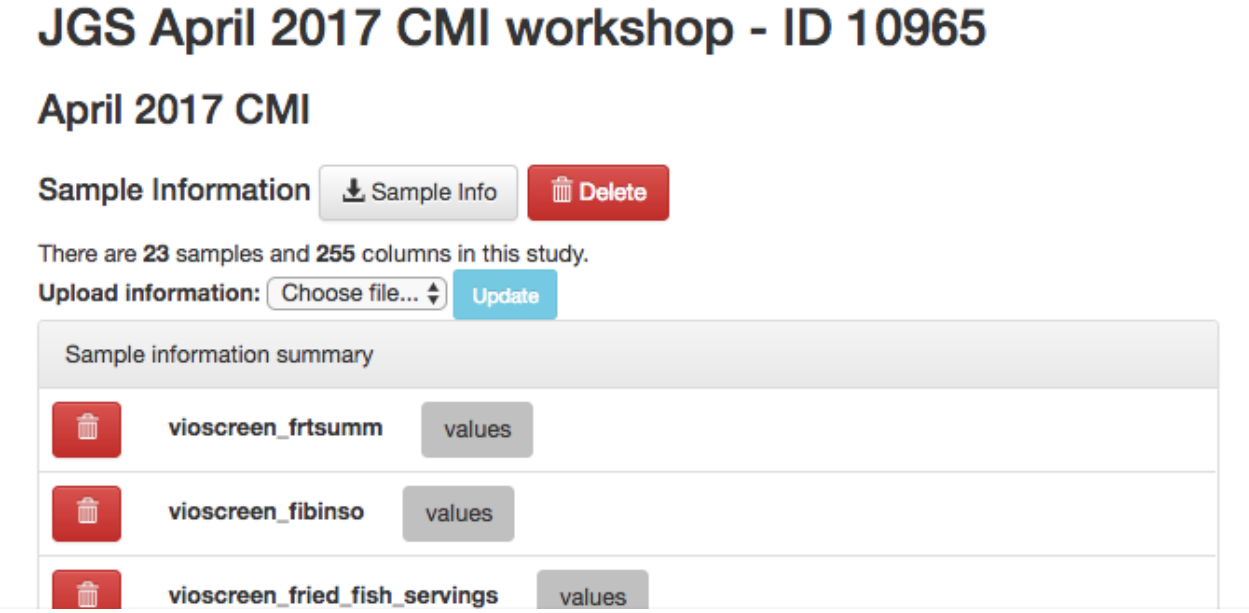
If the file processes successfully, you should be able to click on the "Sample Information" tab and see a list of the imported metadata fields.

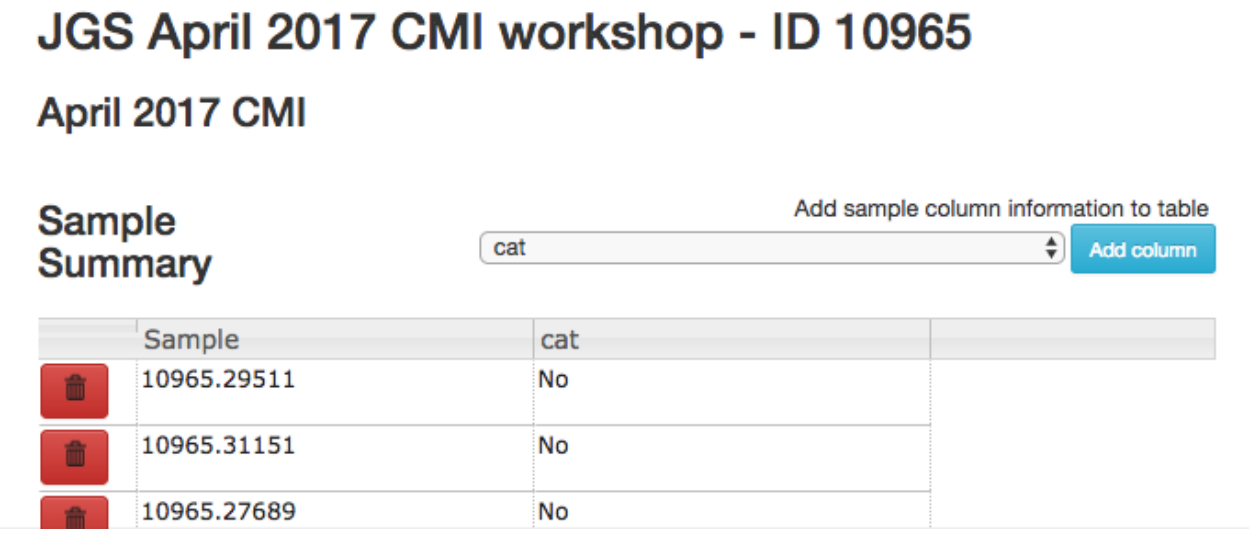
You can also click on "Sample Summary" to check out the different metadata values. Select a metadata column to visualize in the dropdown menu and click "Add column."

In this cohort, only three people were sensible enough to own a cat.

Next, we'll add 16S data and do a preliminary analysis.


The 'sample_name' column is missing from your template, this file cannot be parsed.
Need help? Send us an email.


JGS April 2017 CMI workshop - ID 10965
April 2017 CMI
Sample Information
Sample Info
Delete
There are 23 samples and 255 columns in this study.
Upload information:
Choose file...
Update
Sample information summary
vioscreen_frtsumm
values
vioscreen_fibinso
values
vioscreen_fried_fish_servings
values


JGS April 2017 CMI workshop - ID 10965
April 2017 CMI
Sample Summary
Add sample column information to table
cat
Add column
Sample
cat
10965.29511
No
10965.31151
No
10965.27689
No

Next: *16S Data Processing in Qiita*

16S Data Processing in Qiita

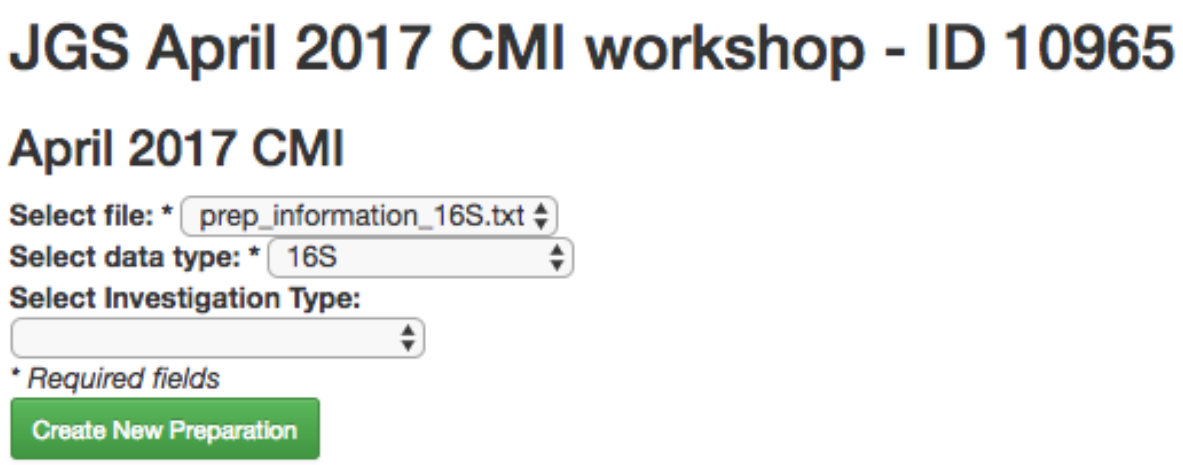
Now, we'll upload some actual microbiome data to explore. To do this, we need to add the data themselves, along with some information telling Qiita about how those data were generated.

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 files in the *16S* folder and the file called *prep_information_16S.txt*.

Now you can click the "Add New Preparation" button. This will bring up the following dialogue:



Select *prep_information_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".

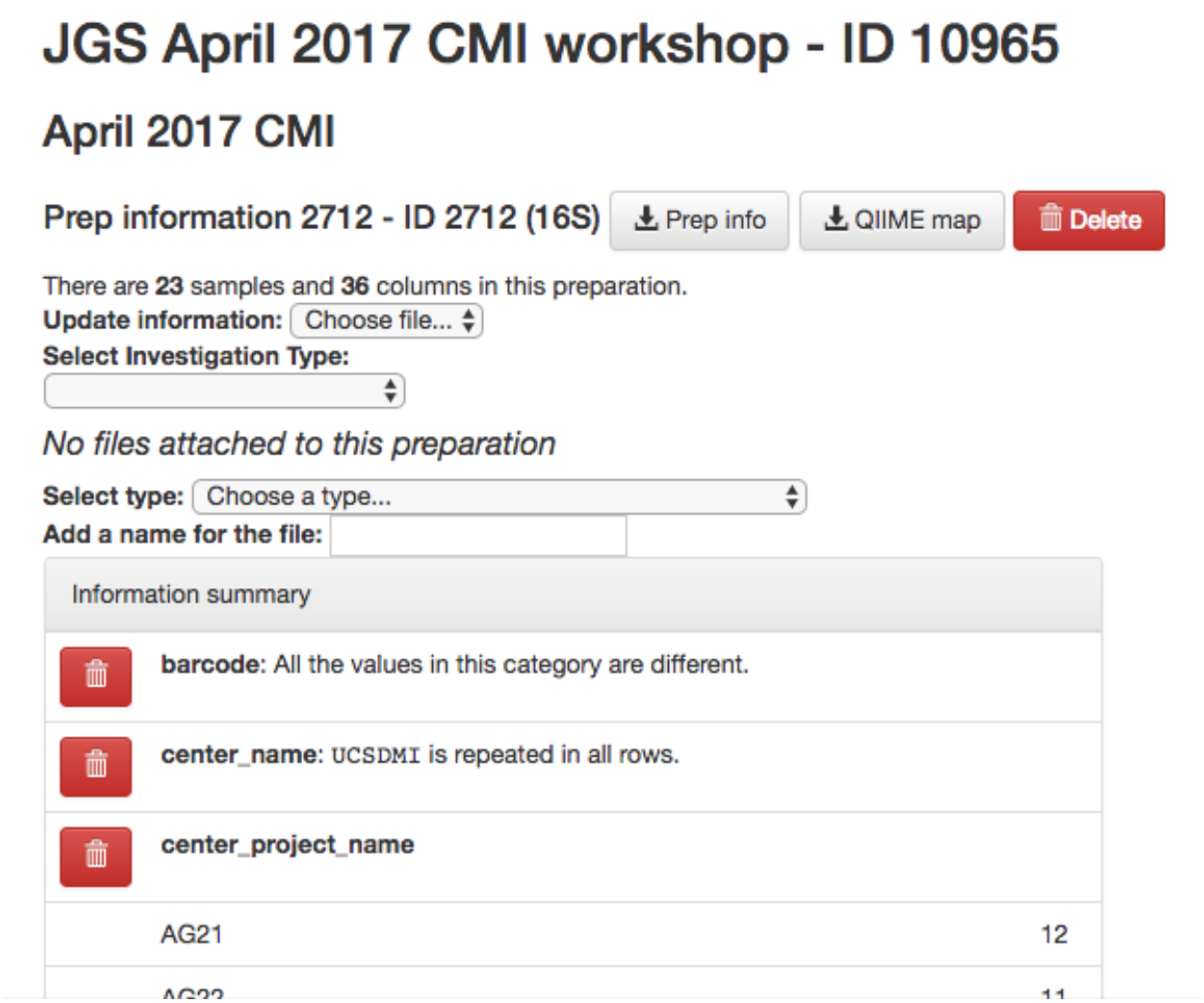
You should now see a summary of your preparation info, similar to the summary we saw of the sample info:

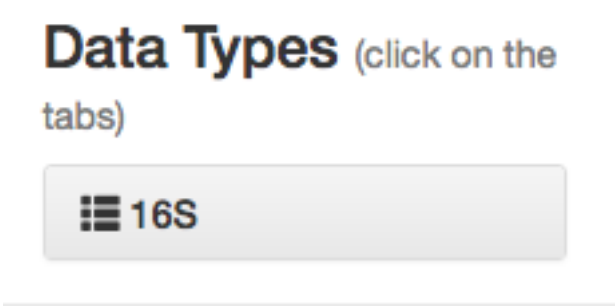
In addition, you should see a "16S" button appear under "Data Types" on the menu to left:

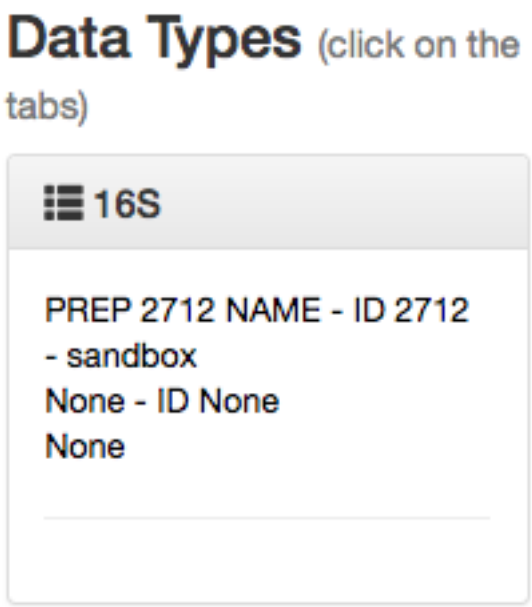
You can click this to reveal the individual prep info files of that data type that have been associated with this study:

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.


JGS April 2017 CMI workshop - ID 10965
April 2017 CMI
Prep information 2712 - ID 2712 (16S)
Prep info
QIIME map
Delete
There are 23 samples and 36 columns in this preparation.
Update information:
Choose file...
Select Investigation Type:
No files attached to this preparation
Select type:
Choose a type...
Add a name for the file:
Information summary
barcode: All the values in this category are different.
center_name: UCSDMI is repeated in all rows.
center_project_name
AG21
12


Data Types (click on the tabs)
16S



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 files per each sample in our study, so we will choose *per_sample_FASTQ*.

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 a list of filenames showing up in the green *raw forward seqs* column below the import dropdown. You'll want to give the set of these per-sample-FASTQ files a name (**Add a name for the file**), and then click "Add files" below.

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

Exploring the raw data

Click back through on your 16S preparation. Now that you've associated sequence files with this prep, you'll have a *Files network* displayed:

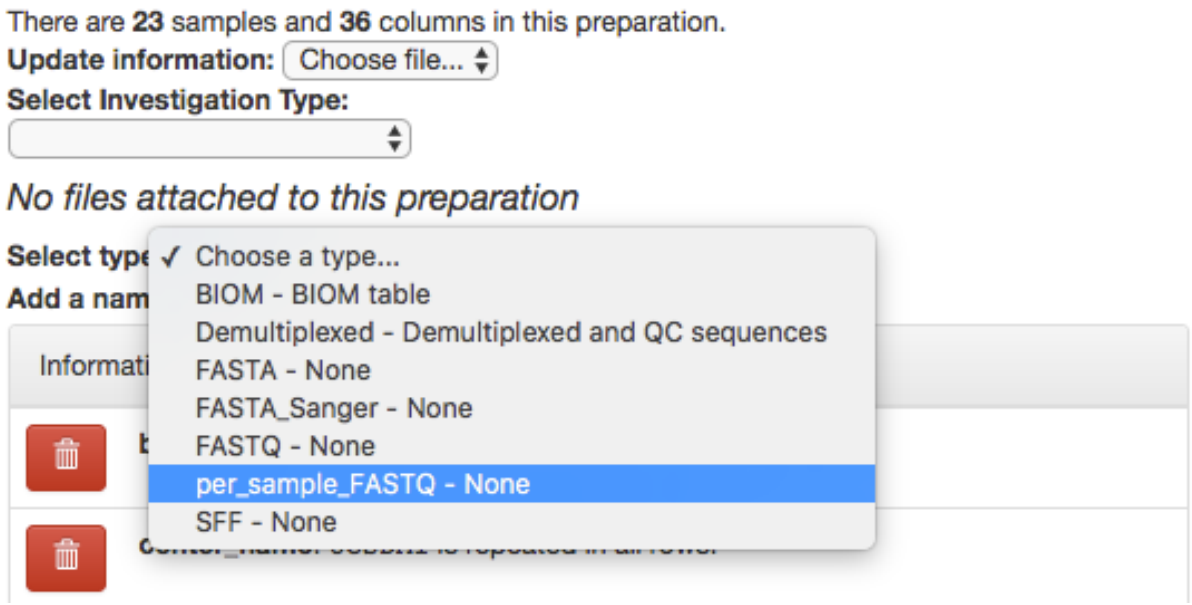
Your collection of FASTQ files for this prep are all represented by a single object in this network, currently called *dflt_name*. Click on the object.

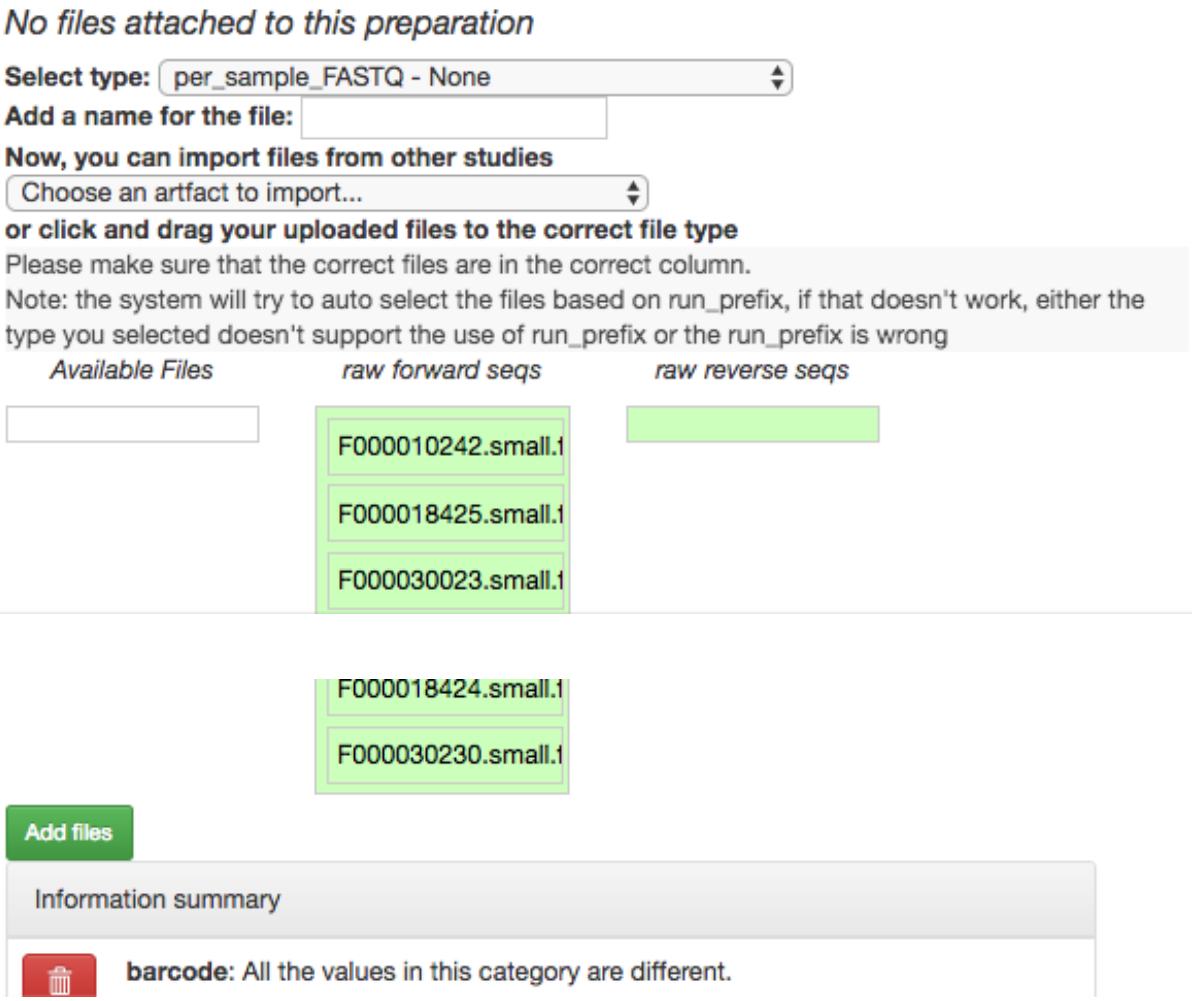
Now, you'll have a series of choices for interacting with this object. You can click "Edit" to rename the object, "Process" to perform analyses, or "Delete" to delete it. In addition, you'll see a list of the actual files associated with this object.

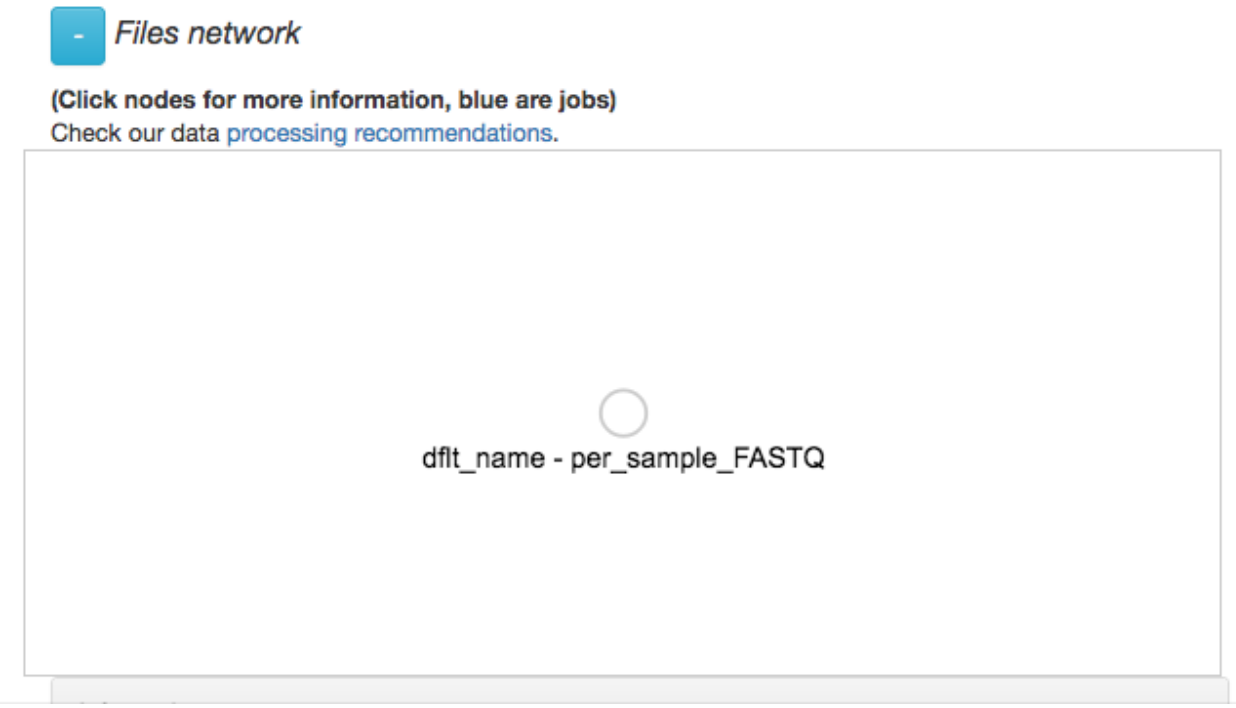
Scroll to the bottom, and you'll also see an option to generate a summary of the object.

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:


There are 23 samples and 36 columns in this preparation.
Update information: Choose file...
Select Investigation Type:
No files attached to this preparation
Select type
Add a nam
Informati
Choose a type...
BIOM - BIOM table
Demultiplexed - Demultiplexed and QC sequences
FASTA - None
FASTA_Sanger - None
FASTQ - None
per_sample_FASTQ - None
SFF - None


No files attached to this preparation
Select type: per_sample_FASTQ - None
Add a name for the file:
Now, you can import files from other studies
Choose an artfact 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 forward seqs
raw reverse seqs
F000010242.small.1
F000018425.small.1
F000030023.small.1
F000018424.small.1
F000030230.small.1
Add files
Information summary
barcode: All the values in this category are different.



This will open a dialogue that gives you information about currently running jobs, as well as jobs that failed with some sort of error.

The summary generation shouldn't take too long. When it completes, you can click back on the per_sample_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.

Processing 16S data

Scroll back up and click on the *per_sample_FASTQ* object, and select "Process". This will bring you to another network visualization interface. Here, you can add processing steps to your objects.

Click again on the *per_sample_FASTQ* object. Below the files network, you will see an option to *Choose command*. Based on the type of object, this dropdown menu will give a you a list of available processing steps.

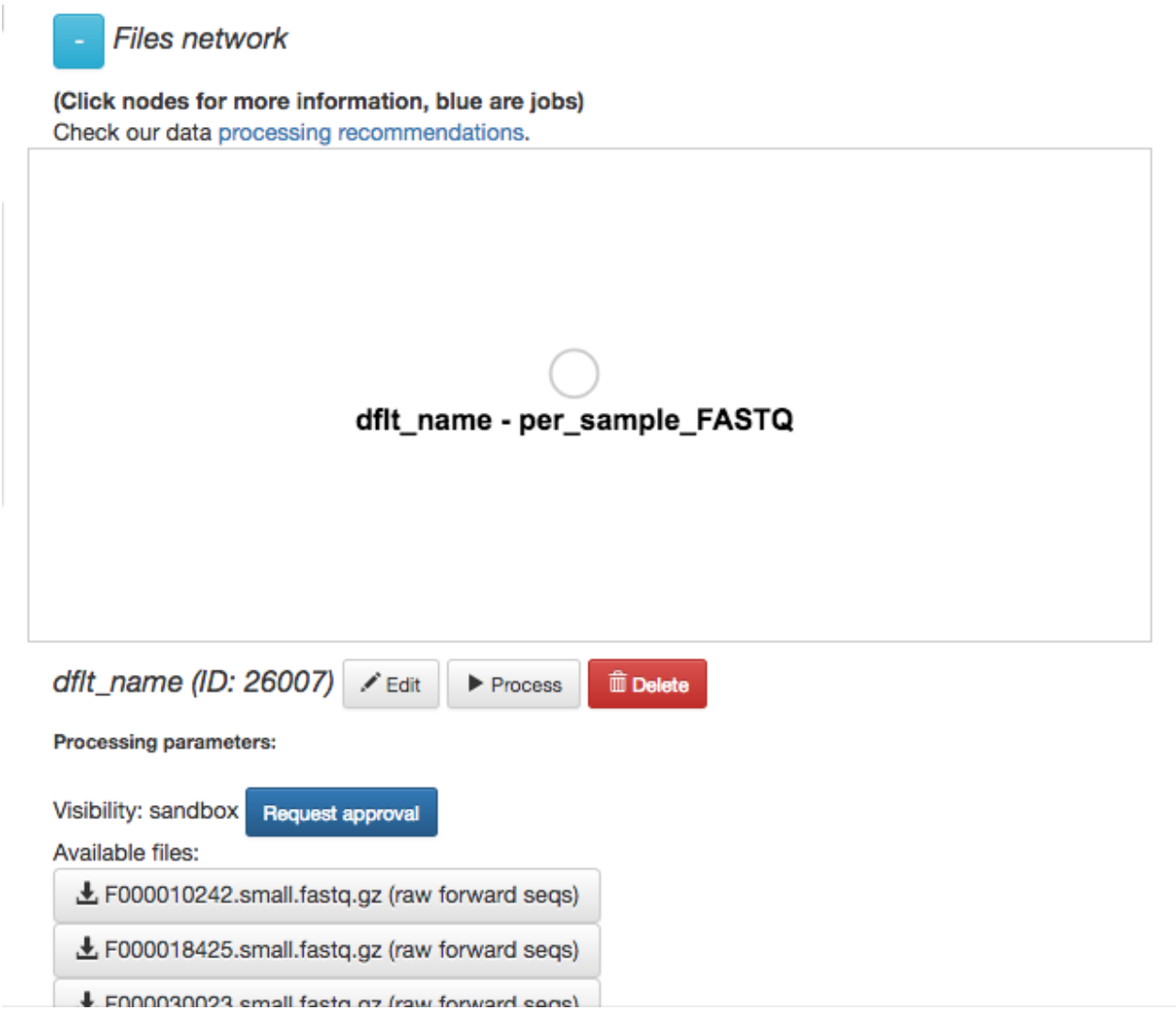
For 16S *per_sample_FASTQ* objects, the only available command is *Split libraries FASTQ*. The converts the raw FASTQ data into the file format used by Qiita for further analysis (you can read more extensively about this file type here).

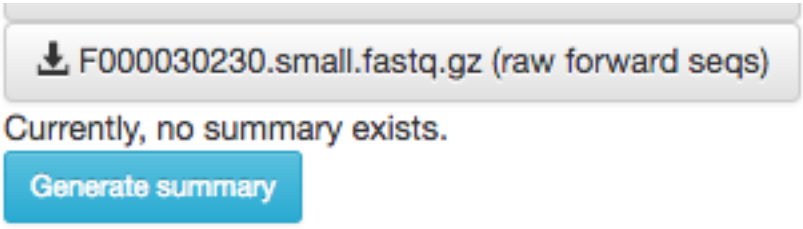
Select the *Split libraries FASTQ* step. Now, you will be able to select the specific combination of parameters to use for this step in the *Choose parameter set* dropdown menu.

For our files, choose *per sample FASTQ defaults, phred_offset 33*. The specific parameter values used will be displayed below. (The other commonly used choice for data generated at the CMI is *golay_12, reverse complement mapping file barcodes, reverse complement barcodes*, which is what you will select if you have one set of non-demultiplexed FASTQ files (forward, reverse, and barcode) containing all of your samples.)

Click "Add Command".

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


Files network
(Click nodes for more information, blue are jobs)
Check our data processing recommendations.
dflt_name - per_sample_FASTQ
dflt_name (ID: 26007)
Edit
Process
Delete
Processing parameters:
Visibility: sandbox
Request approval
Available files:
F000010242.small.fastq.gz (raw forward seqs)
F000018425.small.fastq.gz (raw forward seqs)


F000030230.small.fastq.gz (raw forward seqs)
Currently, no summary exists.
Generate summary


Log Out
(6)

Processing Jobs

(skipping successful jobs)

Search:

Heartbeat	Name	Status	Step
2017-04-18 20:17:12	Generate HTML summary	running	
2017-02-14 07:54:13	Validate	error	Step 2: Validatin 'per_sample_FA files

F000030230.small.fastq.gz (raw forward seqs)

artifact_26010.html (html summary)

F000010242.small.fastq.gz (raw_forward_seqs)

MD5:: 6370d7049892c401037c72161c52b0e8

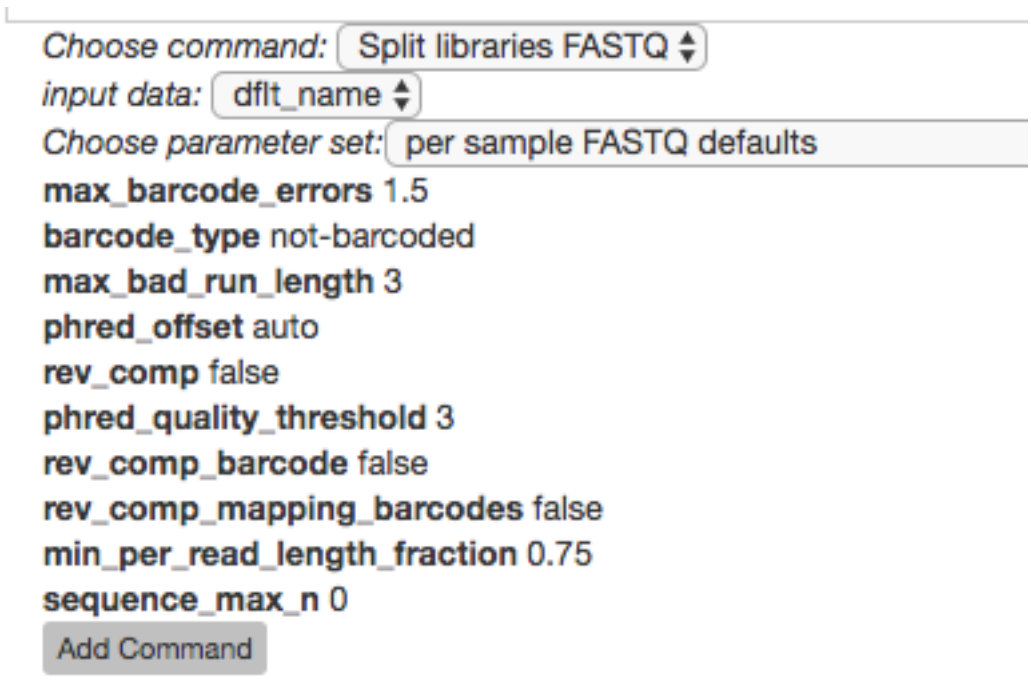
```
@10317.000010242_58737 orig_bc=GCGTCTGTAGCA new_bc=ACGTCTGTAGCA
bc_diffs=1
TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGAAAT
+
BBBBBB@10317.000010242_70616 orig_bc=TCGTCTGTAGCA new_bc=ACGTCTGTAGCA
bc_diffs=1
TACGTAGGGGGCGAGCGTTATCCGGATTCATTGGGCGTAAAGCGCGCGTAGGCGGCCCGGCAGGCCGGGGGTCGAAG
+
BBBB@10317.000010242_2480 orig_bc=ACGTCTGTAGCA new_bc=ACGTCTGTAGCA
bc_diffs=0
TACAGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCGCGTAGGTGGTTTGTTAAGTTGAATGTGAAAT
```

F000018425.small.fastq.gz (raw_forward_seqs)

MD5:: a7e9451e53b580b17eb9204c0731acd7

```
@10317.000018425_61392 orig_bc=TTATGGGTAAGA new_bc=TTATGGGTAAGA
bc_diffs=0
```


Processing dflt_name (ID: 26010)
Processing workflow
Run
Don't forget to hit "Run" once you are done with your workflow!
Wondering what to select? Check our data processing recommendations.
dflt_name - (per_sample_FASTQ)
Choose command:
Choose command...


Choose command:
Split libraries FASTQ
input data:
dflt_name
Choose parameter set:
per sample FASTQ defaults
max_barcode_errors 1.5
barcode_type not-barcoded
max_bad_run_length 3
phred_offset auto
rev_comp false
phred_quality_threshold 3
rev_comp_barcode false
rev_comp_mapping_barcodes false
min_per_read_length_fraction 0.75
sequence_max_n 0
Add Command

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

Note that the command hasn'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.

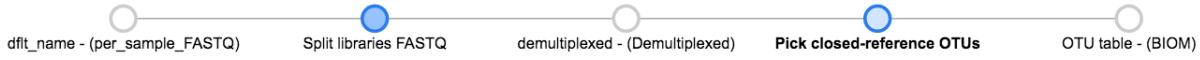
We're going to process our sequences files using two different workflows. In the first, we'll use a conventional reference-based OTU picking strategy to cluster our 16S sequences into OTUs. This approach matches each sequence to a reference database, ignoring sequences that don't match the reference. 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!

The closed reference workflow

To do closed reference OTU picking, click on the *demultiplexed* object and select the *Pick closed-reference OTUs* command. We will use the *default - serial* parameter set for our data, which are relatively small. For a larger data set, we might want to use the parallel implementation.

By default, Qiita uses the GreenGenes 16S reference database. You can also choose to use Silva, or the Unite 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.

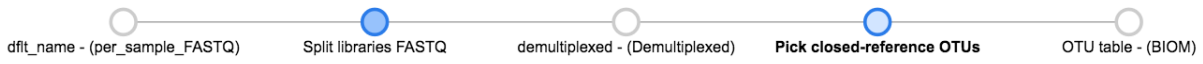
That's it!

The deblur workflow

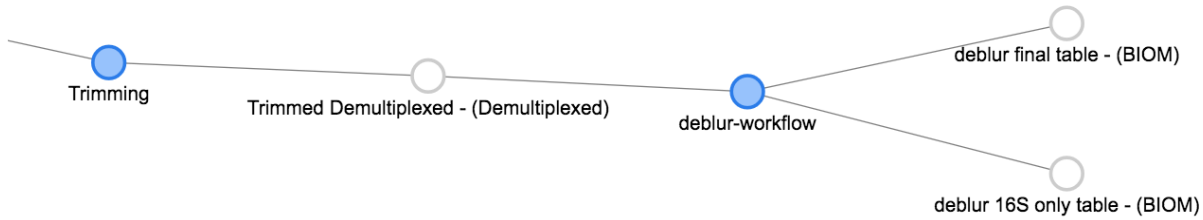
The deblur workflow is only marginally more complex. Although you can deblur the demultiplexed sequences directly, *deblur* works best when all the sequences are the same length. By trimming to a particular length, we can also ensure our samples will be comparable to other samples already in the database.

Click back on the *demultiplexed* object. this time, select the *Trimming* operation. Currently, there are three trimming length options. Let's choose *Trimming 100*, which trims to the first 100bp, for this run, and click "Add Command".

Now you can see that we have the same *demultiplexed* object being used for two separate processing steps – closed-reference OTU picking, and trimming.



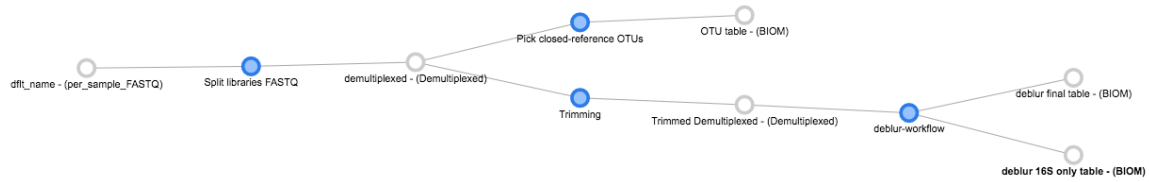
Now we can click the *Trimmed Demultiplexed* object and add a deblur step. Choose *deblur-workflow* from the *Choose command* dropdown, and *Defaults* for the parameter set. Add this command.



As you can see, *deblur* produces two BIOM-formatted OTU tables as output. The *deblur 16S only table* contains deblurred sequences that have been filtered to try and exclude things like organellar mitochondrial reads, while *deblur final table* has all the sequences.

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.

As noted above, you can follow the process of your commands in the dialogue at the top right of the window.

You can also click on the objects in the prep info file network, and see status updates from the commands running on that object at the bottom of the page:

Once objects have been generated, you can generate summaries for them just as you did for the original *per_sample_FASTQ* object.

The summary for the *demultiplexed* object gives you information about the length of sequences in the object:

The summary for a BIOM-format OTU table gives you a histogram of the the number of sequences per sample:

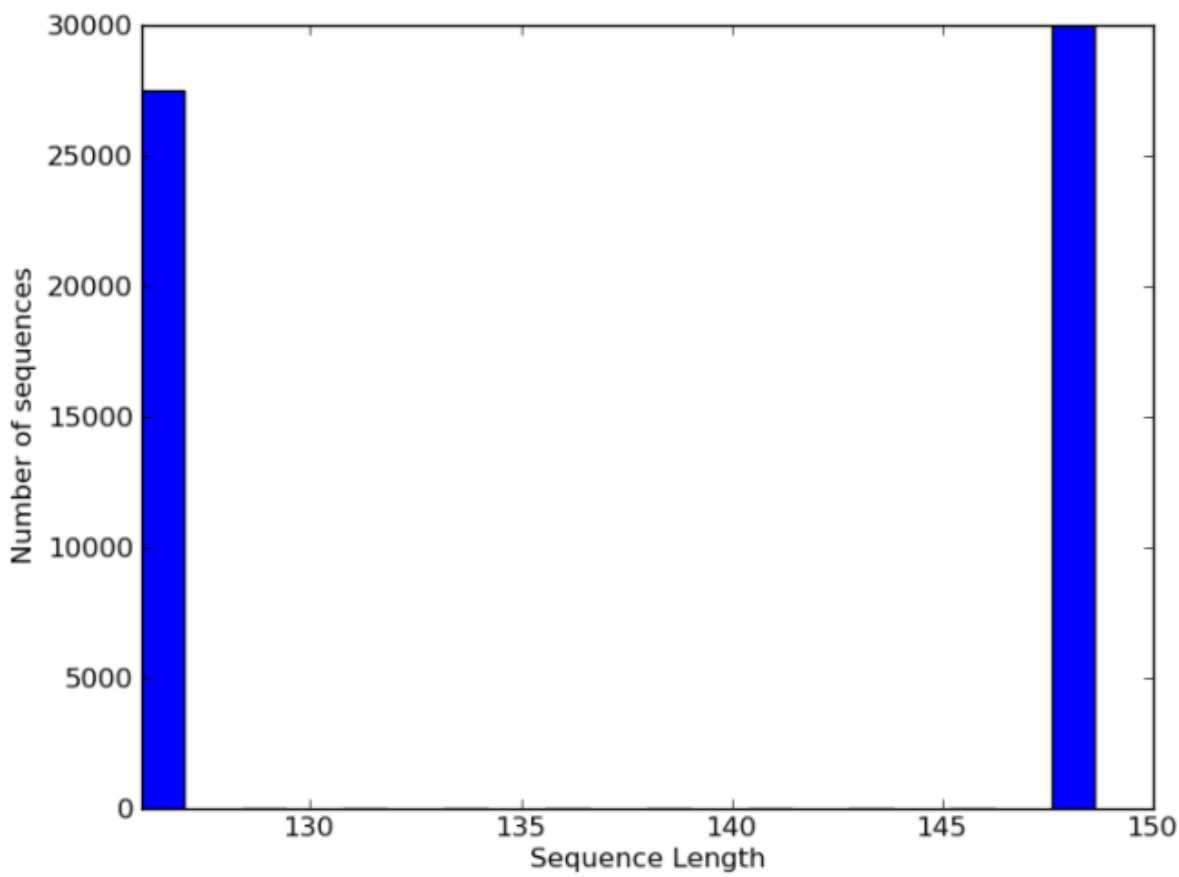
Next: *16S Microbiome Analysis in Qiita*

Jobs using this set of files:

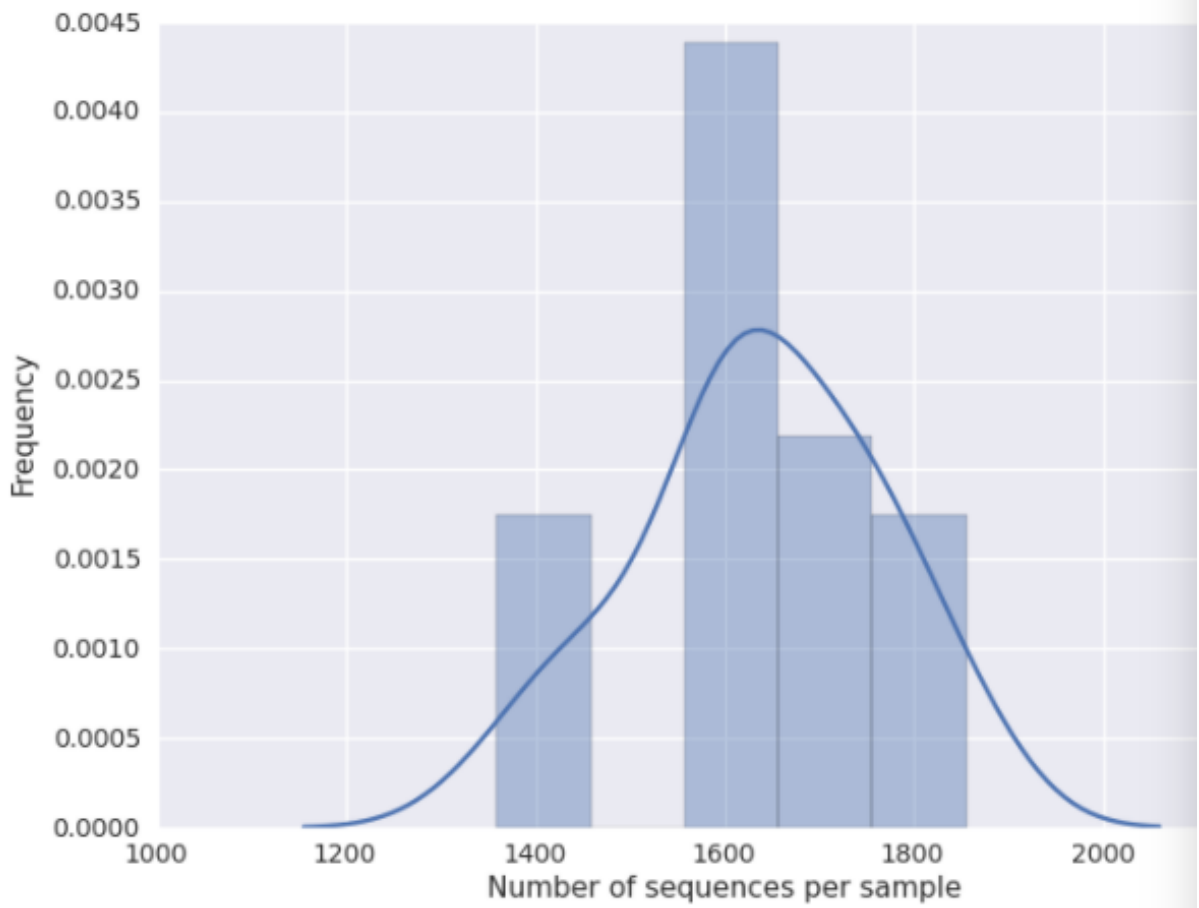
Job **2915b8ad-b5a3-4e55-8179-d07f0b61595f** (Pick closed-reference OTUs).
Status: *running*. Step: *Step 3 of 4: Executing OTU picking*
Currently, no summary exists.

Features

Total: 57500
Max: 150
Mean: 138
Standard deviation: 150
Median: 11



Number of samples: 23
Number of features: 207
Minimum count: 1359
Maximum count: 1855
Median count: 1631
Mean count: 1635

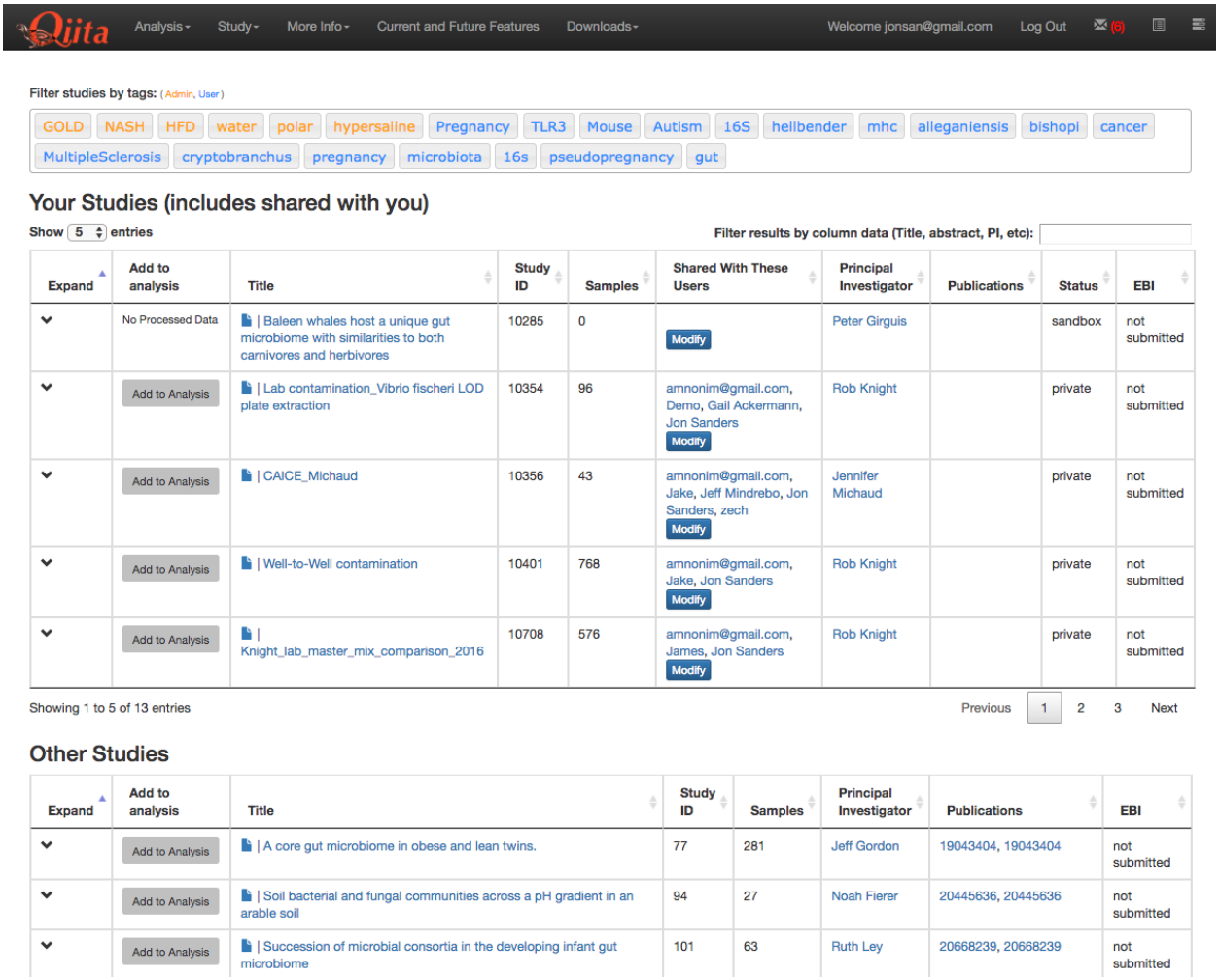


16S Microbiome Analysis in Qiita

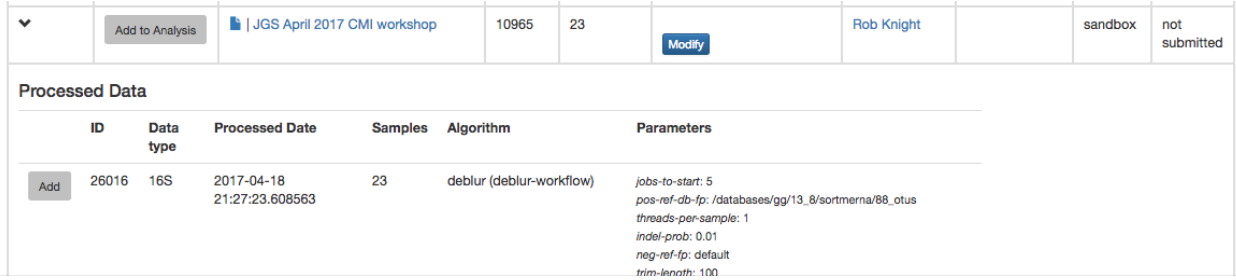
Analysis of Closed Reference processing

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 publically available ('Other') studies.



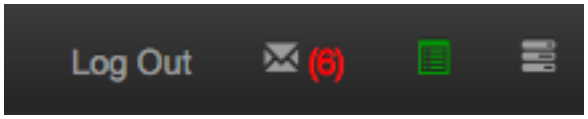
Find the study you created for this tutorial under "Your Studies". Click the down arrow at the left of the row. This will expand the study to expose all the objects from that study that are available to you for analysis.



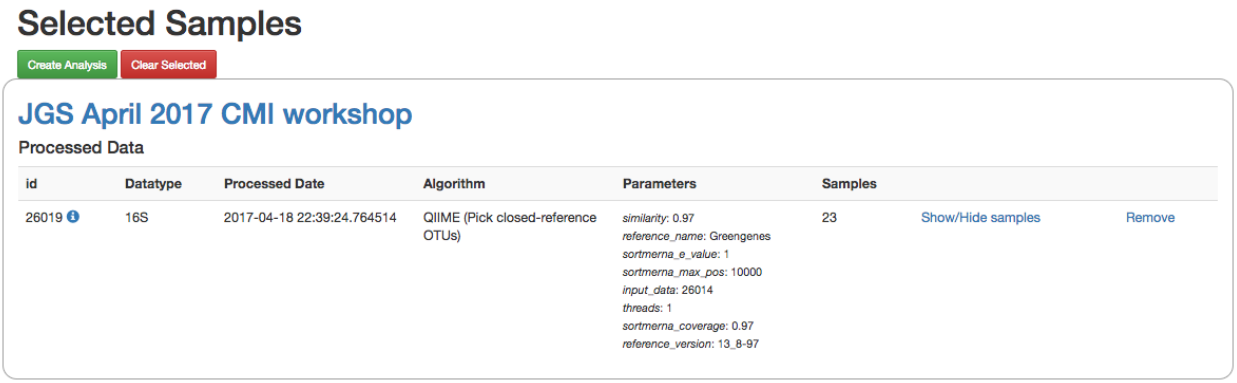
You could add all of these objects to the analysis by selecting the *Add to Analysis* button. We will just add the Closed Reference OTU table object by clicking *Add* in that row.

Add	26019	16S	2017-04-18 22:39:24.764514	23	QIIME (Pick closed-reference OTUs)	*similarity*: 0.97 *reference_name*: Greengenes *sortmerna_e_value*: 1 *sortmerna_max_pos*: 10000 *threads*: 1 *sortmerna_coverage*: 0.97 *reference_version*: 13_8-97

Now, the second-right-most icon at the top bar should be 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 23 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 an any way.)

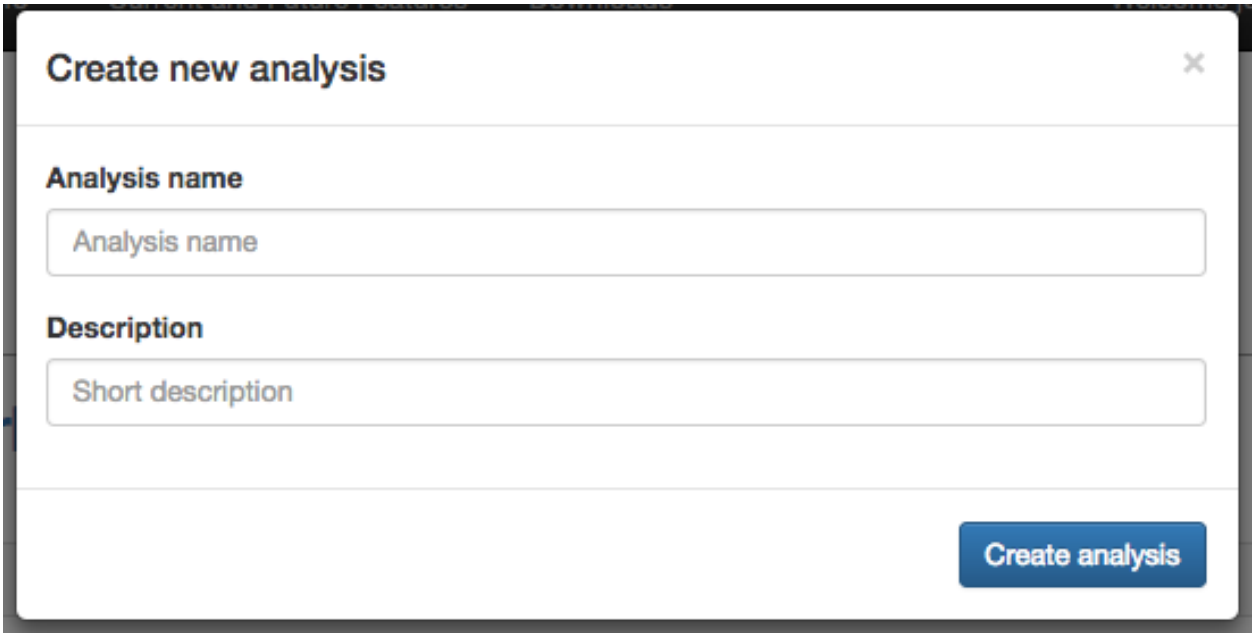
This should be good for now. Click the "Create Analysis" button, enter a name and description, then click "Create analysis".

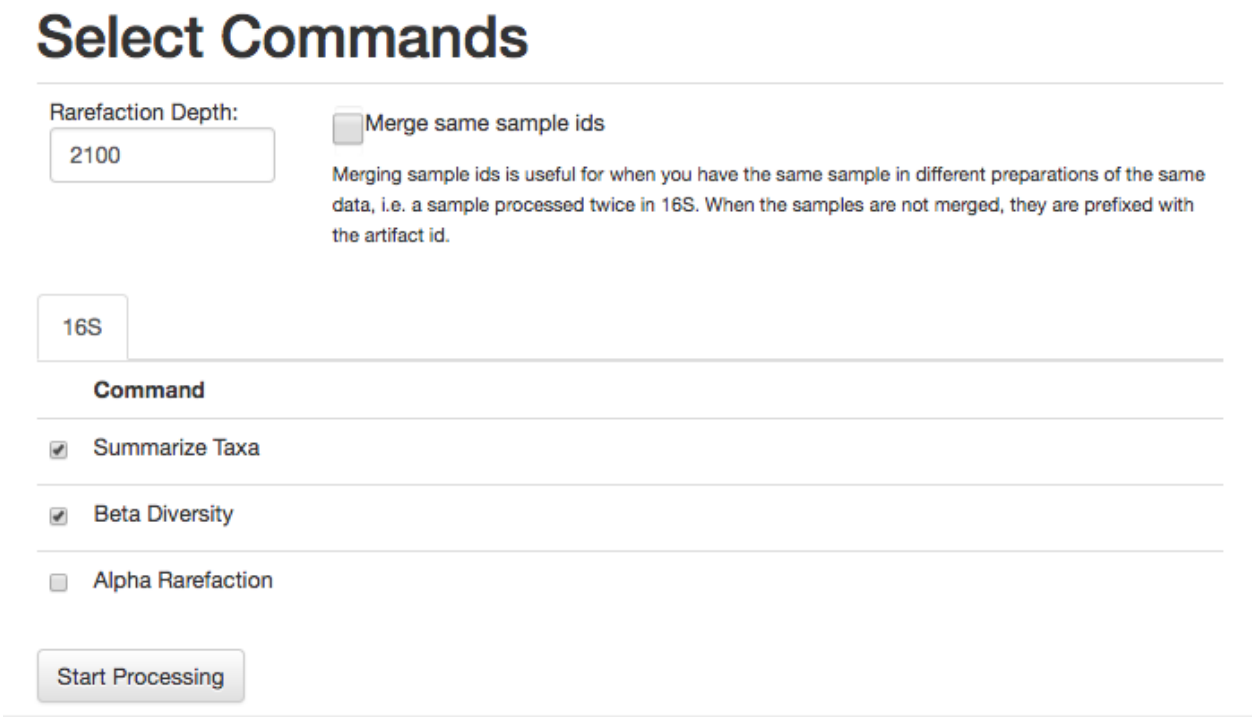
This brings you to the analysis commands selection page, where you can specify the steps in your analysis.

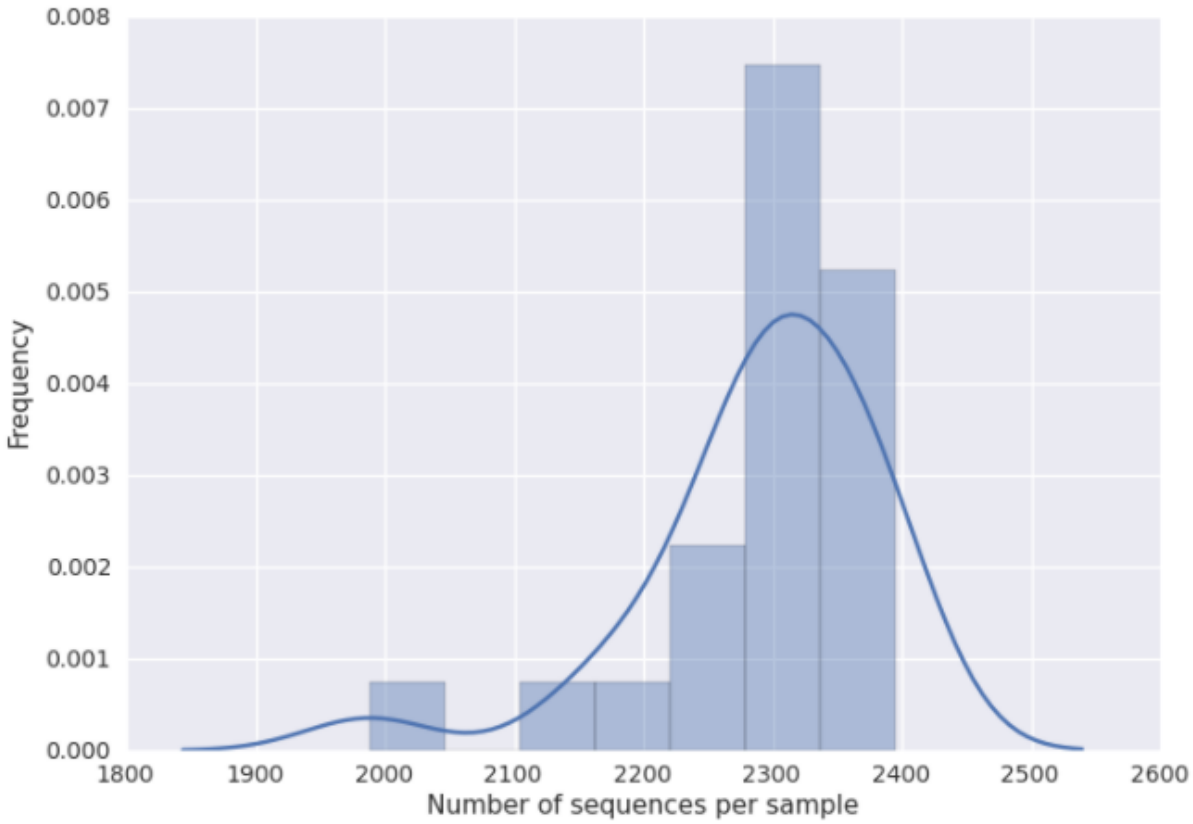
For this analysis, let's go ahead and select the commands Summarize Taxa and Beta Diversity (Alpha Rarefaction can take some time to run).

We will also need to specify an even sampling or rarefaction depth. All the samples in the analysis will be randomly subsampled to this number of sequences, reducing potential biases. Samples with fewer than this number of sequences will be excluded, which can also be useful for excluding things like blanks.

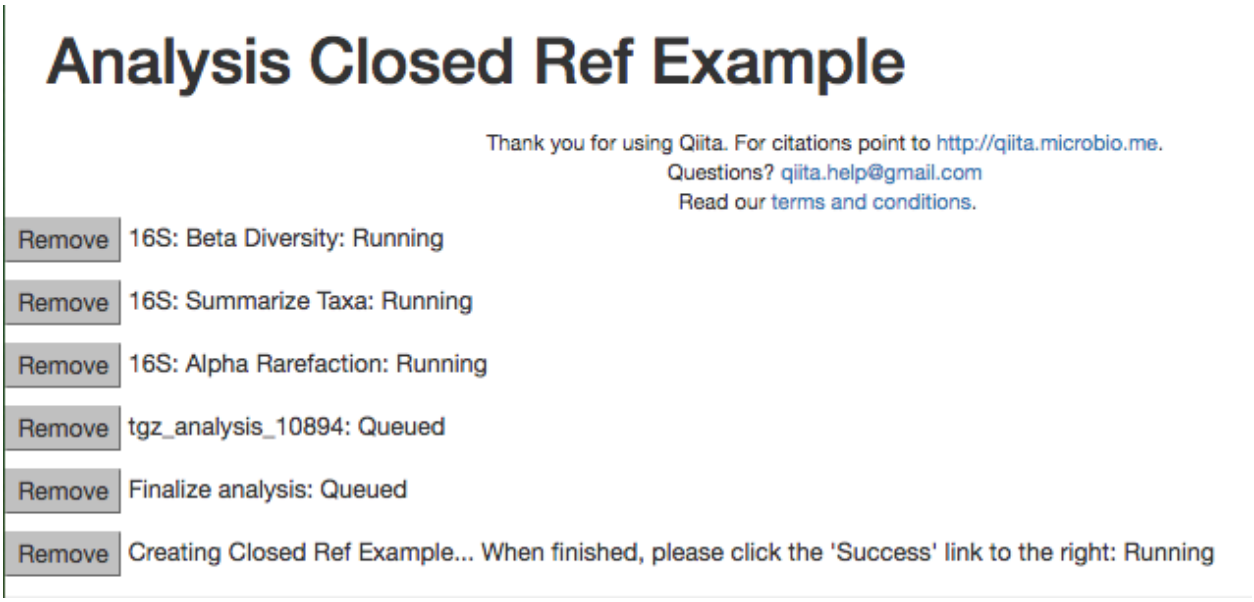
You can get a good idea of where to set this threshold by looking at the histogram generated by summarizing the input closed-reference OTU table, as discussed in *16S Microbiome Analysis in Qiita*. Here, it looks like 2100 would be an appropriate cutoff: it excludes one clear outlier, but retains most of the samples.


Create new analysis
Analysis name
Analysis name
Description
Short description
Create analysis


Select Commands
Rarefaction Depth:
2100
Merge same sample ids
Merging sample ids is useful for when you have the same sample in different preparations of the same data, i.e. a sample processed twice in 16S. When the samples are not merged, they are prefixed with the artifact id.
16S
Command
Summarize Taxa
Beta Diversity
Alpha Rarefaction
Start Processing


0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
Frequency
1800
1900
2000
2100
2200
2300
2400
2500
2600
Number of sequences per sample

Enter 2100 in the rarefaction depth field, select the check boxes for Summarize Taxa and Beta Diversity, and click "Start Processing". You will see a list each step in the analysis, followed by its status:



When the analysis is finished, click the 'Success' link to see the results.

The results page will have sections indication which samples were dropped due to insufficient numbers of reads, as well as sections for each data type.

Here, we have taxonomy summaries and beta diversity PCoA plots available.

Clicking on *bar_charts.html* under "Summarize Taxa" will take you to a visualization of the taxa that were found in your sample:

Under "Beta Diversity", you will have a selection of Principle Coordinates Analyses of different measures of beta diversity, or the similarity between samples.

Clicking on one (say, *unweighted unifrac emperor pcoa plot*) will open an interactive visualization of the similarity among your samples. Generally speaking, the more similar the samples, the closer the are likely to be in the PCoA ordination. The Emperor visualization program offers a very useful way to explore how patterns of similarity in your data associate with different metadata categories. Here, I've colored the points in our test data by cat ownership.

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.

Finally, if you ran Alpha Rarefaction, you will also have a link to interactive plots that can be used to show how different measures of alpha diversity correlate with different metadata categories:

Analysis of deblur processing

Creating an analysis of your deblurred data is virtually the same as the process for the Closed Reference data, but there are a few quirks.

First, because the deblur process creates two separate BIOM tables, you'll want to make a note of the specific object ID number for the artifact you want to use. In my case, that's ID 26017, the deblurred table with 'only-16s' reads.

The specific ID for your table will be unique, so make a note of it, and you can use it to select the correct table for analysis.

Analysis: CMI Workshop Test

Shared with:

Dropped Samples

16S

JGS April 2017 CMI workshop:
Total dropped: 1
10965.000030083

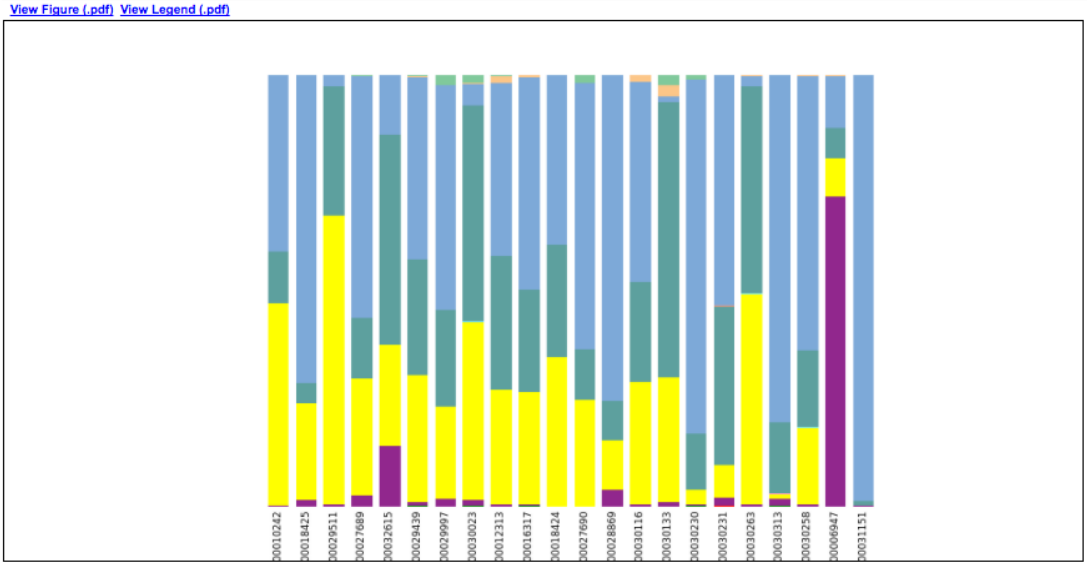
16S

Summarize Taxa
area_charts.html
bar_charts.html

Beta Diversity
unweighted unifrac emperor pcoa plot
weighted unifrac emperor pcoa plot

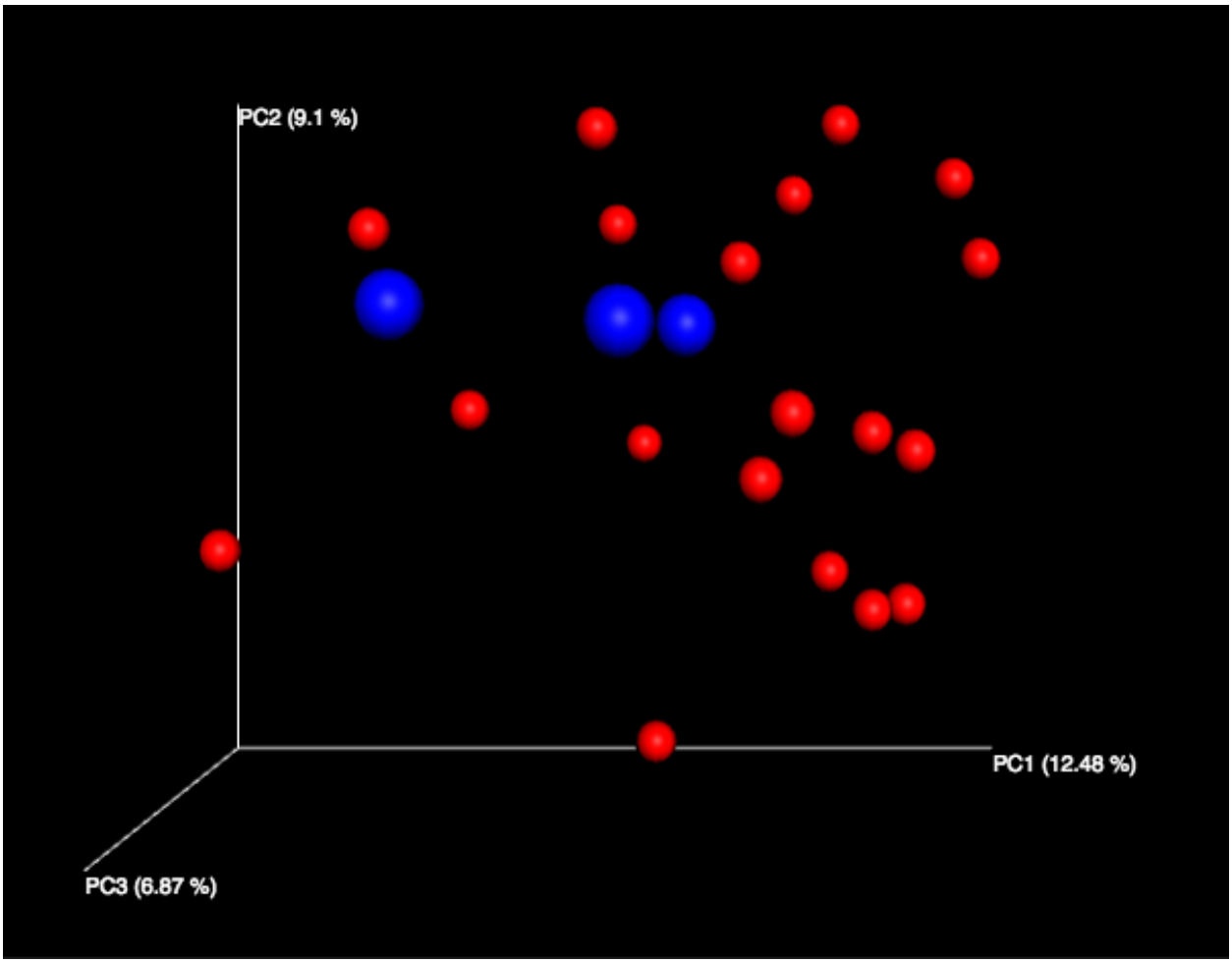
Taxonomy Summary. Current Level:

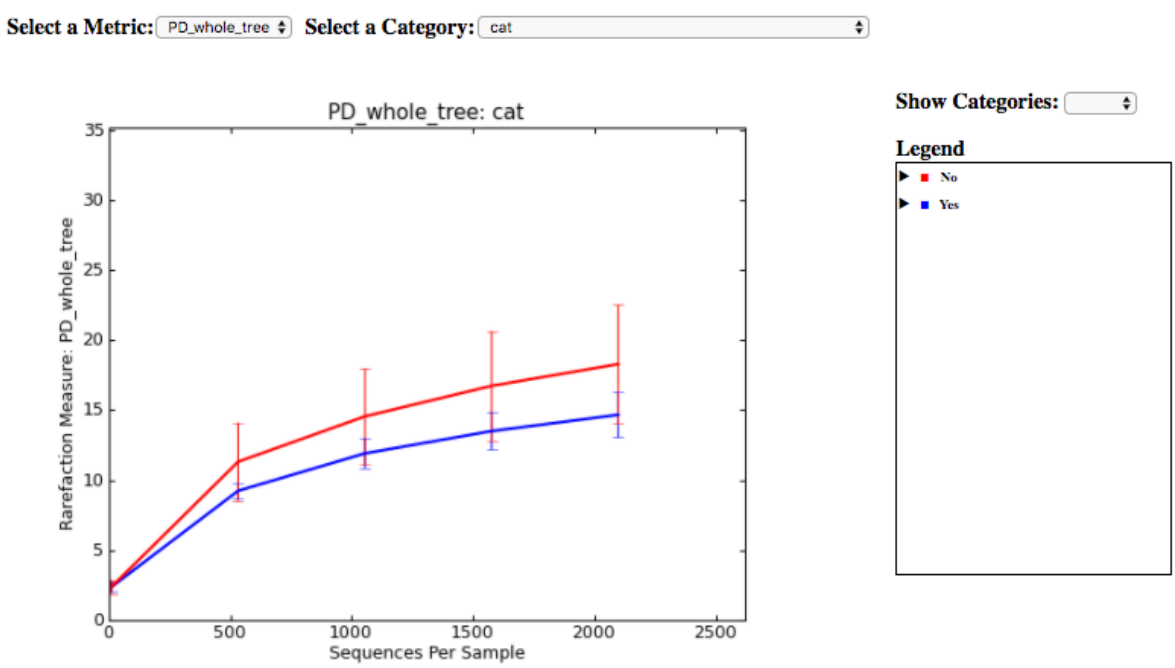
View Figure (.pdf) View Legend (.pdf)

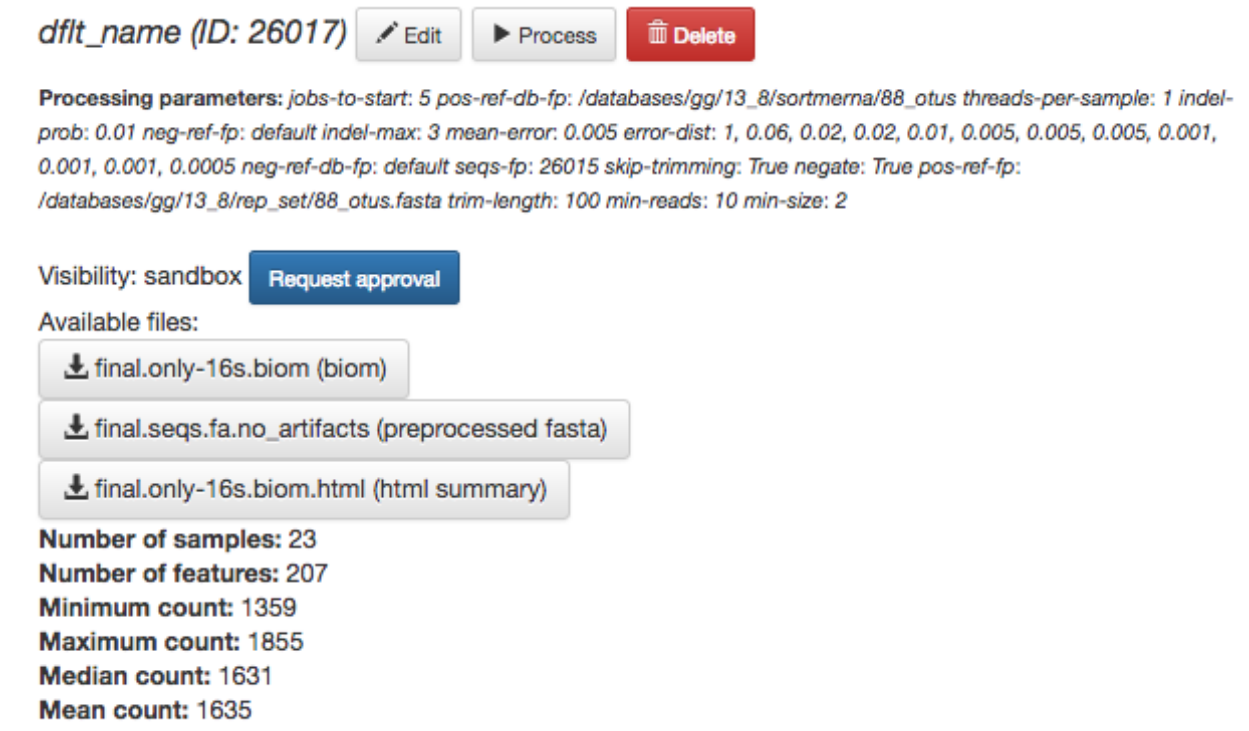


View Table (.txt)

Legend	Taxonomy	Total	10965.000010242	10965.000018425	10965.000029511	10965.000027689	10965.000032615	10965.000029439	10965.000029997	10965.000030023	10965.000012313	10965.000016317	1
		%	%	%	%	%	%	%	%	%	%	%	
	k__Archaea;p__Crenarchaeota	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	
	k__Archaea;p__Euryarchaeota	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	
	k__Bacteria;p__AD3	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	
	k__Bacteria;p__Acidobacteria	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.1%	
	k__Bacteria;p__Actinobacteria	4.8%	0.1%	1.5%	0.6%	2.6%	13.9%	1.0%	1.7%	1.4%	0.4%	0.3%	
	k__Bacteria;p__Bacteroidetes	24.8%	47.0%	22.3%	66.7%	27.0%	23.5%	29.3%	21.3%	41.1%	26.5%	26.1%	
	k__Bacteria;p__Cyanobacteria	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.3%	0.0%	0.0%	
	k__Bacteria;p__FBP	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	
	k__Bacteria;p__Firmicutes	24.4%	11.9%	4.9%	30.1%	14.2%	48.7%	26.9%	22.5%	49.9%	31.1%	23.8%	
	k__Bacteria;p__Fusobacteria	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	
	k__Bacteria;p__Gemmatimonadetes	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	
	k__Bacteria;p__Lentisphaerae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	
	k__Bacteria;p__Planctomycetes	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	
	k__Bacteria;p__Proteobacteria	45.0%	41.0%	71.2%	2.5%	56.0%	13.8%	42.1%	52.1%	5.1%	40.1%	49.1%	
	k__Bacteria;p__Tenericutes	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	1.5%	0.5%	
	k__Bacteria;p__Verrucomicrobia	0.5%	0.0%	0.0%	0.0%	0.2%	0.0%	0.3%	2.3%	2.0%	0.2%	0.0%	


PC2 (9.1 %)
PC1 (12.48 %)
PC3 (6.87 %)


Select a Metric: PD_whole_tree
Select a Category: cat
PD_whole_tree: cat
Rarefaction Measure: PD_whole_tree
Sequences Per Sample
0
5
10
15
20
25
30
35
0
500
1000
1500
2000
2500
Show Categories:
Legend
No
Yes



Second, currently only the Beta Diversity analysis command option is working with deblurred data.

Creating a meta-analysis

One of the most powerful aspects of Qiita is the ability to compare your data with hundreds of thousands of samples from across the planet. Right now, there are almost 130,000 samples publicly available for you to explore:

(You can get up-to-date statistics by clicking "Stats" under the "More Info" option on the top bar.)

Creating a meta-analysis is just like creating an analysis, except you choose data objects from multiple studies. Let's start creating a meta-anlysis by adding our Closed Reference OTU table to a new analysis.

Next, we'll look for some additional data to compare against.

You noticed the 'Other Studies' table below 'Your Studies' when adding data to the analysis. (Sometimes this takes a while to load - give it a few minutes.) These are publicly available data for you to explore, and each should have processed data suitable for comparison to your own.

There are a couple tools provided to help you find useful public studies.

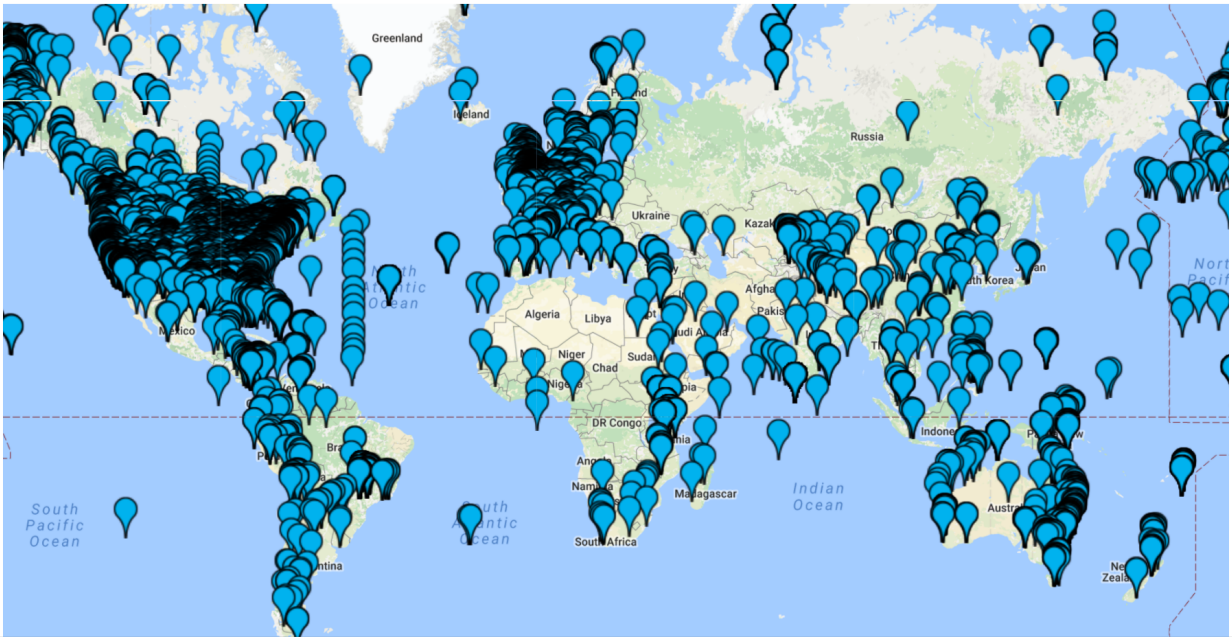
First, there are a series of "tags" listed at the top of the window:

There are two types of tags: admin-assigned (yellow), and user-assigned (blue). You can tag your own study with any tag you'd like, to help other users find your data. For some studies, Qiita administrators will apply specific reserved tags to help identify particularly relevant data. The "GOLD" tag, for example, identifies a small set of highly-curated, very well-explored studies. If you click on one of these tags, all studies not associated with that tag will disappear from the tables.

Second, there is a search field that allows you to filter studies in real time. Try typing in the name of a known PI, or a particular study organism – the thousands of publicly available studies will be filtered down to something that is easier to look through.

Generated on: 04-19-17 03:03:51

Studies	Samples	Users
sandbox: 793 *public*: 310 *private*: 311 *submitted to EBI*: 184	*sandbox*: 130,583 *public*: 129,997 *private*: 115,951 *submitted to EBI*: 107,888 *submitted to EBI (prep)*: 84,039	2,978



Filter studies by tags: (Admin, User)

GOLD NASH HFD water polar hypersaline Pregnancy TLR3 Mouse Autism 16S hellbender mhc alleganiensis bishopi cancer MultipleSclerosis cryptobranchus pregnancy microbiota 16s pseudopregnancy gut

Your Studies (includes shared with you)

Show 5 entries

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

Expand	Add to analysis	Title	Study ID	Samples	Shared With These Users	Principal Investigator	Publications	Status	EBI

Let's try comparing our data to the "Global Gut" dataset of human microbiomes from the US, Africa, and South America from the study "Human gut microbiome viewed across age and geography" by Yatsunenko et al. We can search for this dataset using the DOI from the paper: *10.1038/nature11053*.

Other Studies

Expand	Add to analysis	Title	Study ID	Samples	Principal Investigator	Publications	EBI
	Add to Analysis	| Human gut microbiome differentiation viewed across cultures, ages and families illumina	850	528	Jeff Gordon	22699611, 10.1038/nature11053	not submitted

Processed Data

	ID	Data type	Processed Date	Samples	Algorithm	Parameters
Add	2458	16S	2015-10-15 15:14:56.647454	528	QIIME (Pick closed-reference OTUs)	*similarity*: 0.97 *reference_name*: Greengenes *sortmerna_e_value*: 1 *sortmerna_max_pos*: 10000 *threads*: 5 *sortmerna_coverage*: 0.97 *reference_version*: 13_8-97

Add the closed reference OTU table from this study to your analysis. You should now be able to click the green analysis icon in the upper right and see both your own OTU table and the public study OTU table in your analysis staging area:

You can now click "Create Analysis" just as before to begin specifying analysis steps. This time, let's just do the beta diversity step. Select the *Beta Diversity* command, enter a rarefaction depth of 2100, and click "Start Processing".

Because you've now expanded the number of samples in your analysis by more than an order of magnitude, this step will take a little longer to complete. But when it does, you will be able to use Emperor to explore the samples in your test dataset to samples from around the world!

Notes on metabolomics

Edited for the Dorrestein Lab by Louis-Felix Nothias, Daniel Petras and Ricardo Silva on December 2016. Last edit on April 2017.

About the metabolomics workshop

In the following documentation, we are providing step-by-step tutorials to perform basic analysis of liquid chromatography coupled to tandem mass spectrometry data (LC-MS/MS). These tutorials can be employed to process untargeted metabolomics data, such as those generated for seed funded project.

- The GNPS web-platform will be used to generate a qualitative analysis of your sample LC-MS/MS data. Such as the annotation of known compounds (by MS/MS spectral matching with public library), along as annotating unknown compounds by molecular networking (by spectral similarity).
- And we will used MZmine2 to process LC-MS/MS data in order to generate a feature table. This feature table contains the list of detected compounds and their relative distribution accross samples. This feature table will be used to generate statistical analysis in Qiita.

Selected Samples

Create Analysis | Clear Selected

JGS April 2017 CMI workshop

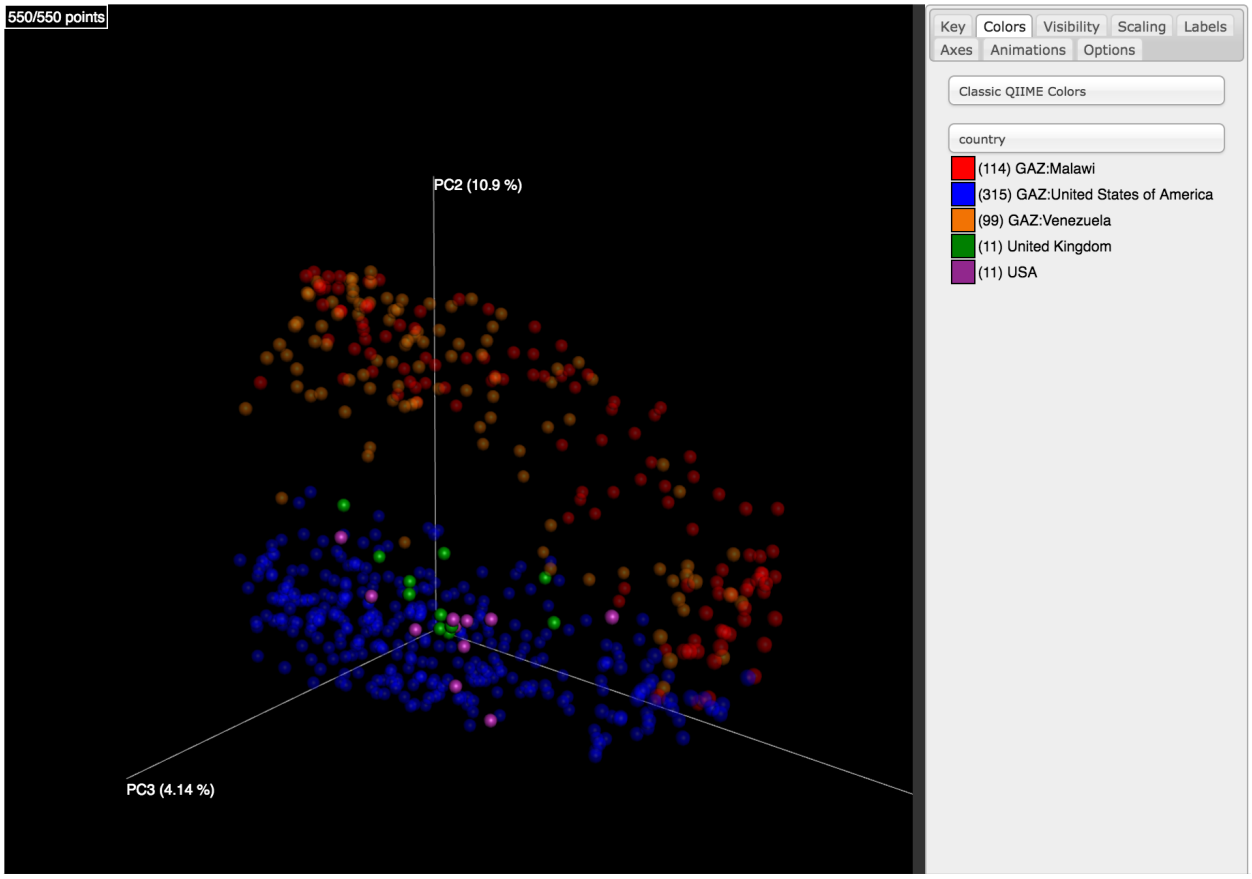
Processed Data

id	Datatype	Processed Date	Algorithm	Parameters	Samples		
26019	16S	2017-04-18 22:39:24.764514	QIIME (Pick closed-reference OTUs)	*similarity*: 0.97 *reference_name*: Greengenes *sortmerna_e_value*: 1 *sortmerna_max_pos*: 10000 *input_data*: 26014 *threads*: 1 *sortmerna_coverage*: 0.97 *reference_version*: 13_8-97	23	Show/Hide samples	Remove

Human gut microbiome differentiation viewed across cultures, ages and families illumina

Processed Data

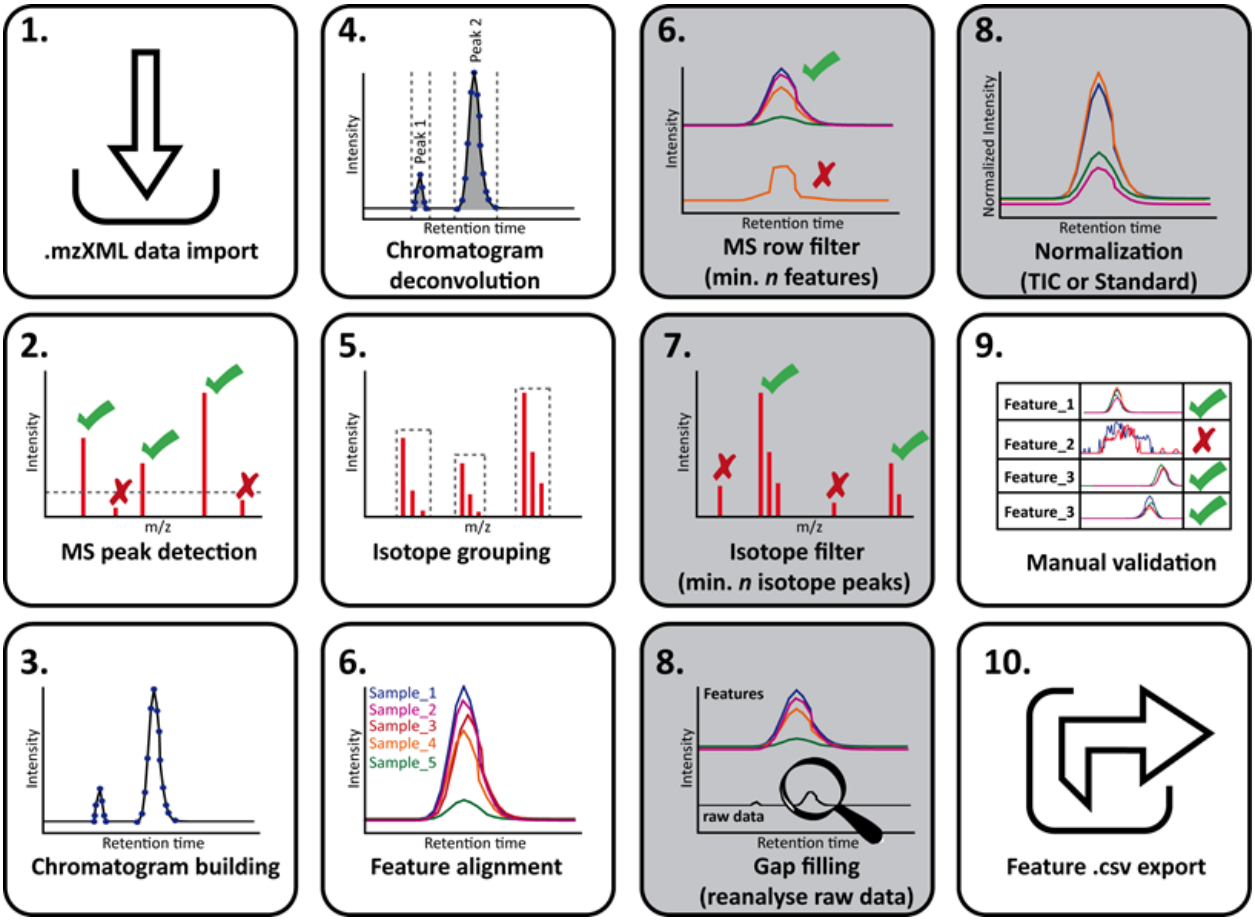
id	Datatype	Processed Date	Algorithm	Parameters	Samples		
2458	16S	2015-10-15 15:14:56.647454	QIIME (Pick closed-reference OTUs)	*similarity*: 0.97 *reference_name*: Greengenes *sortmerna_e_value*: 1 *sortmerna_max_pos*: 10000 *input_data*: 171 *threads*: 5 *sortmerna_coverage*: 0.97 *reference_version*: 13_8-97	528	Show/Hide samples	Remove



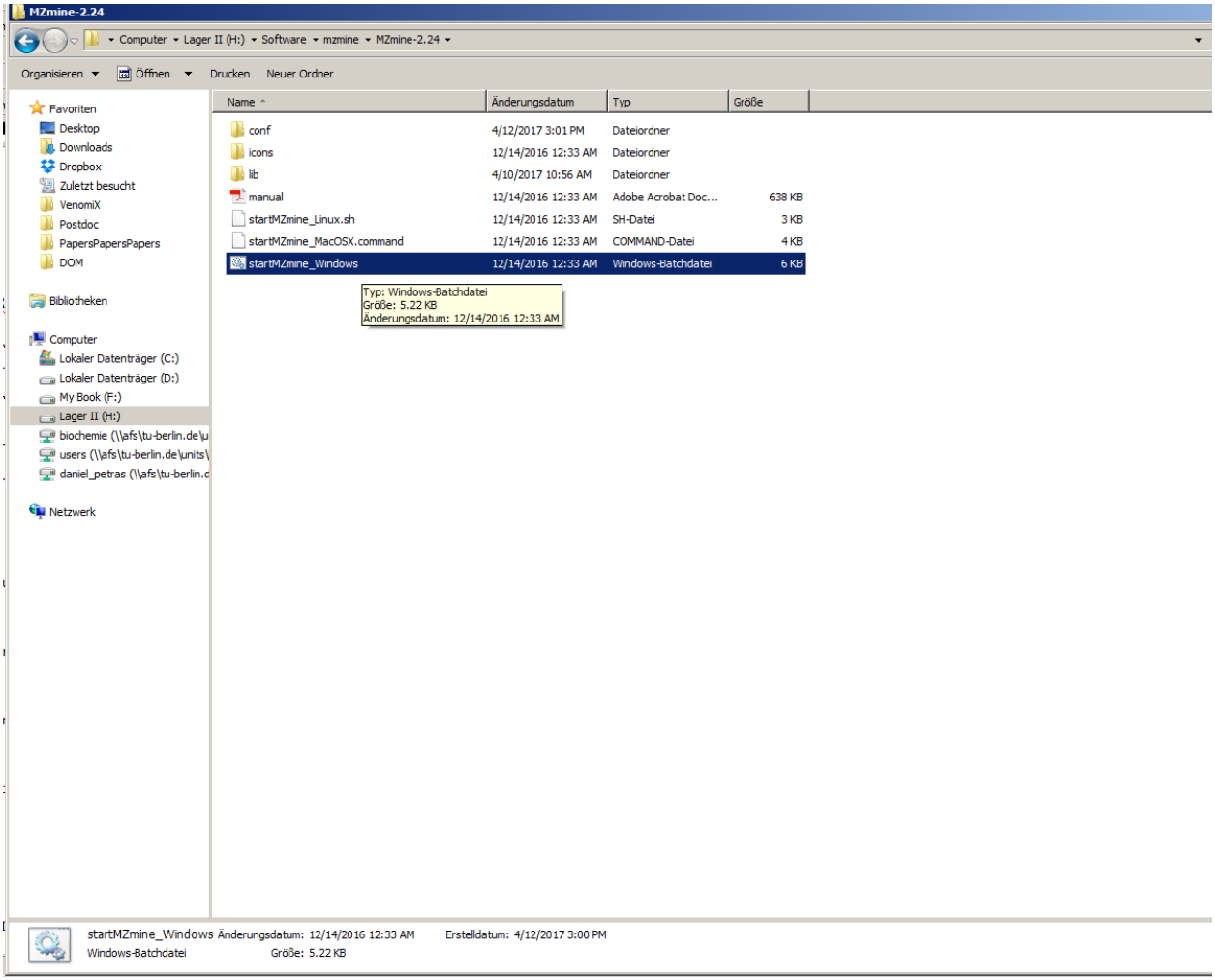
Feature finding with MZmine2

Please follow this (link) to install the software and dependencies.

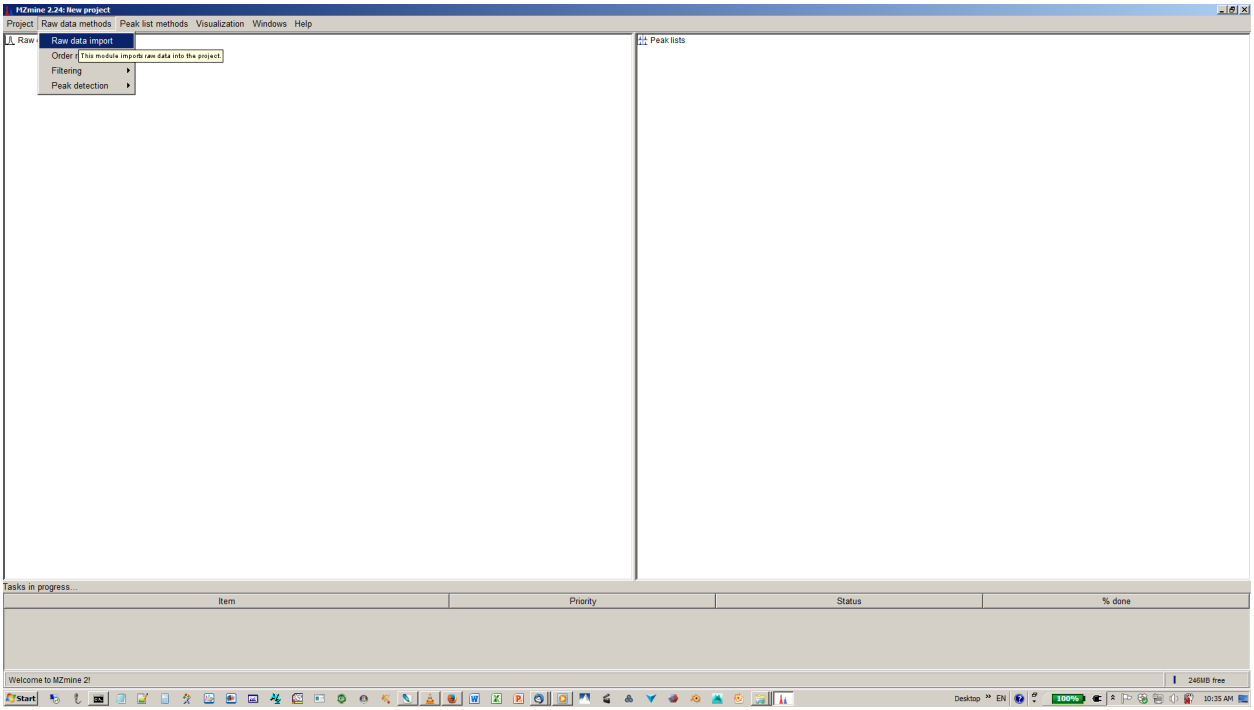
Complete workflow view



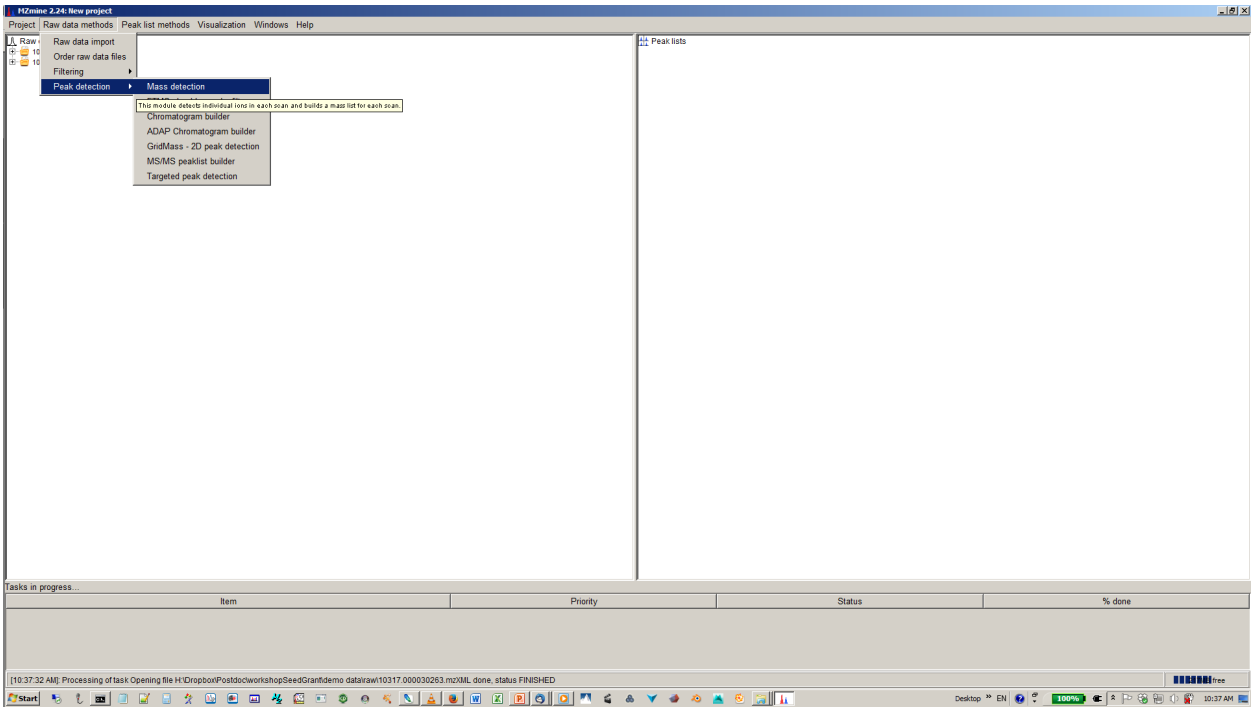
1. Start mzMine2



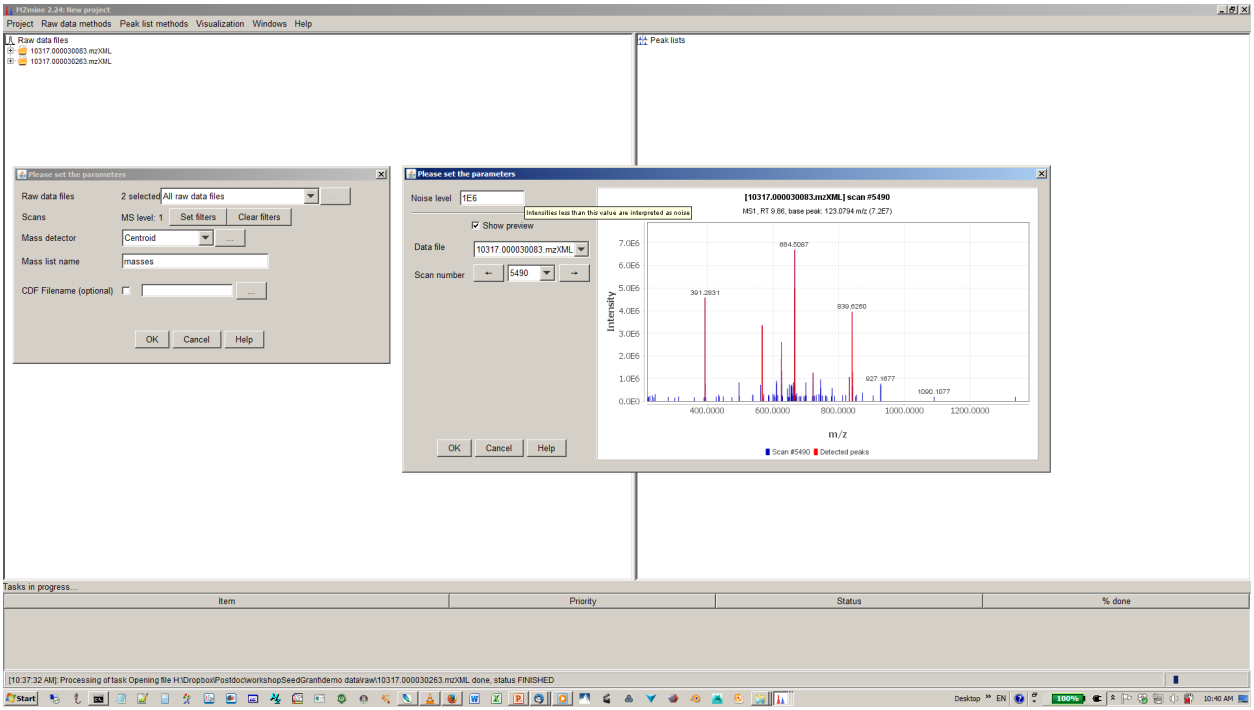
2. Click on raw data import in drop down menu and select .mzxml files



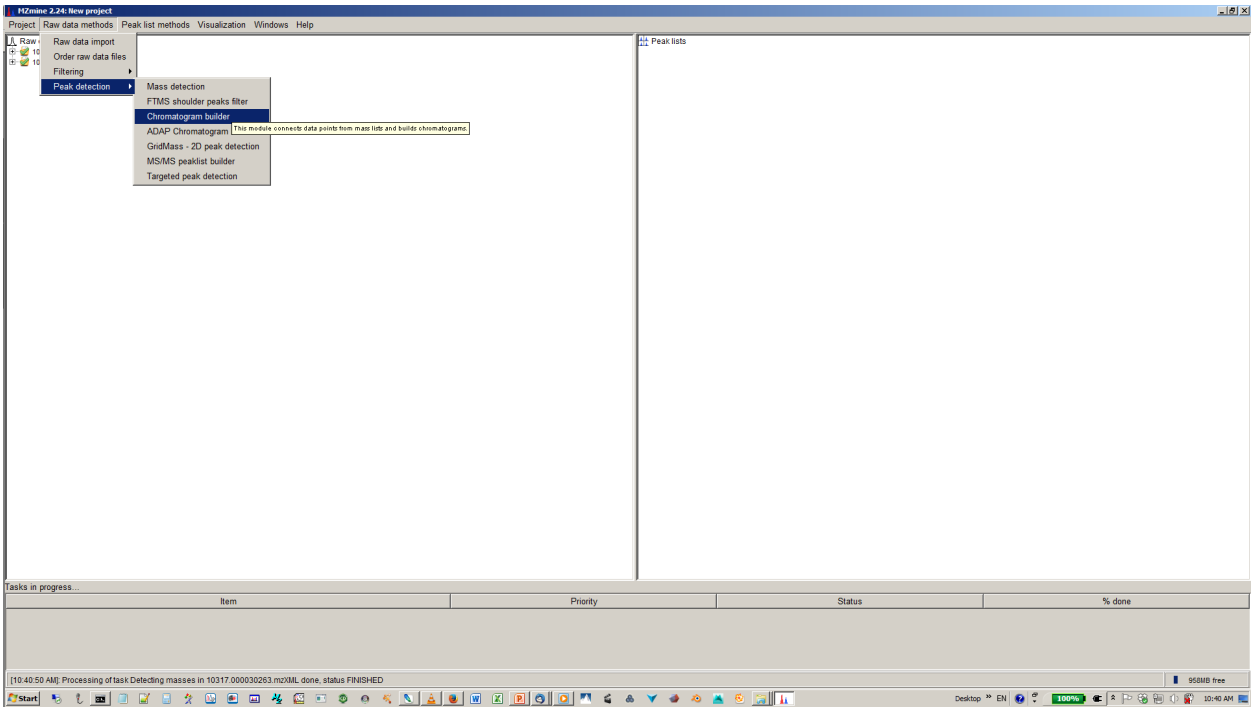
3. Click on mass detection in drop down menu



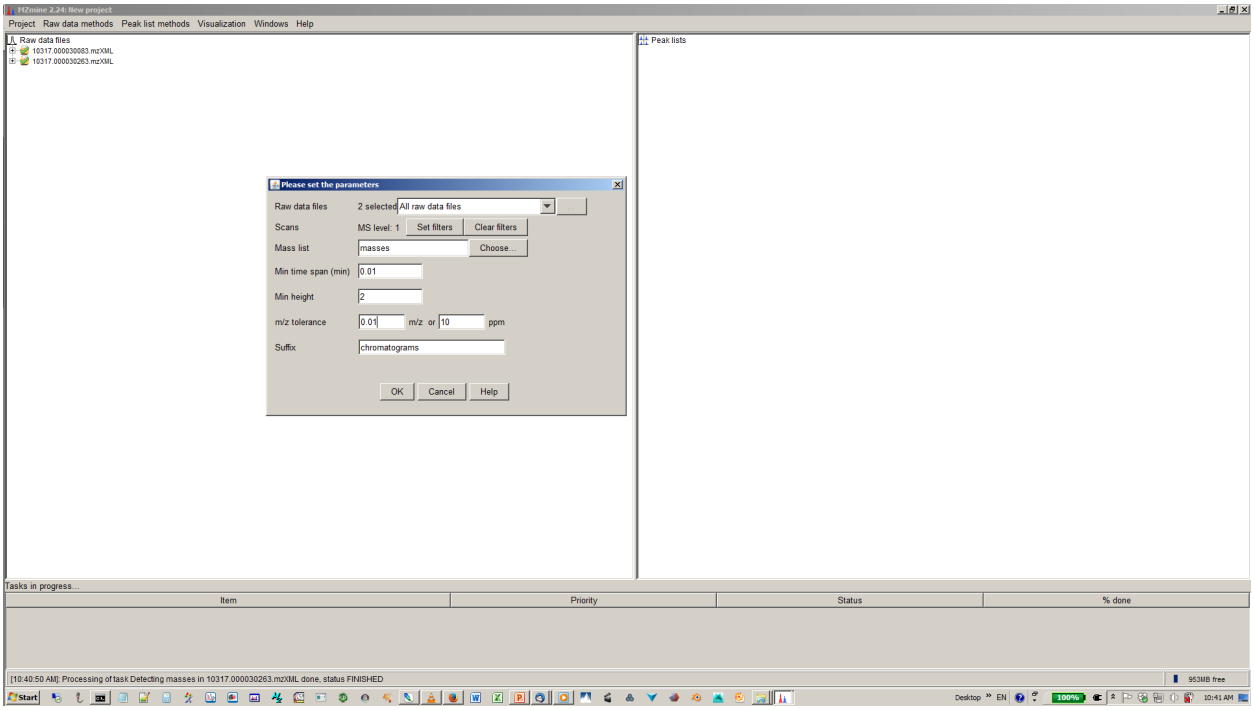
4. Specify intensity cut-off and mass list



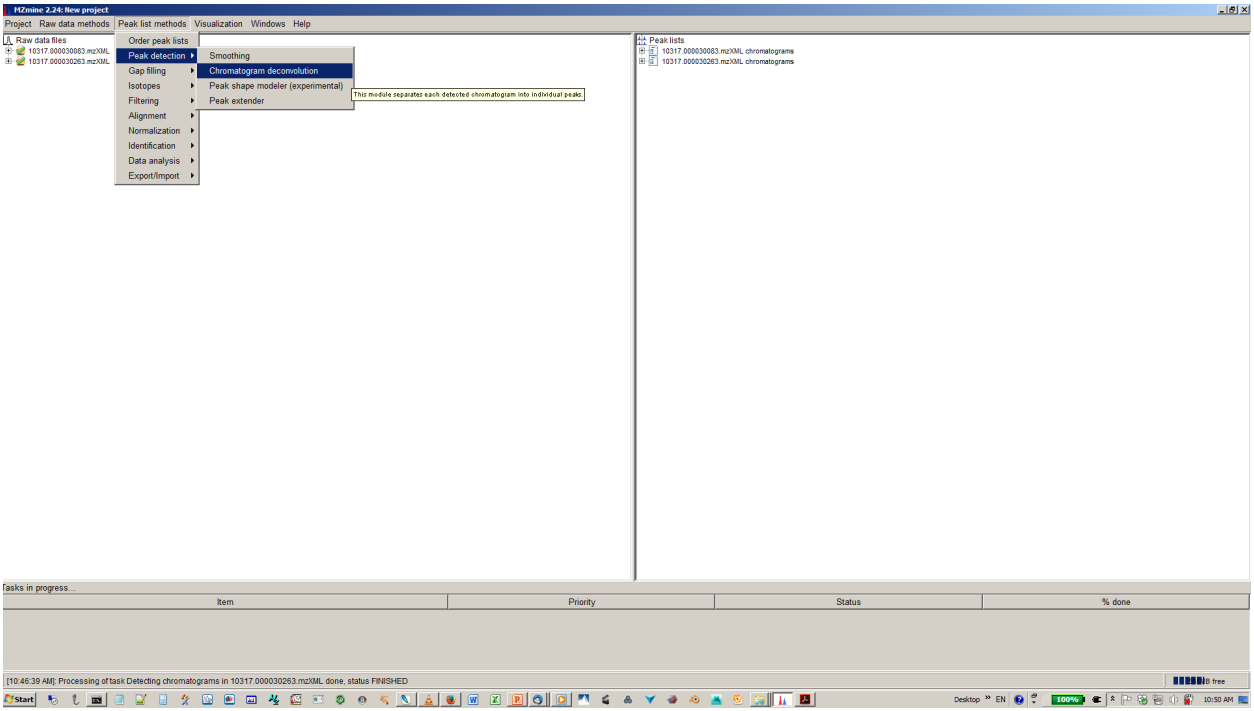
5. Build XICs with chromatogram builder



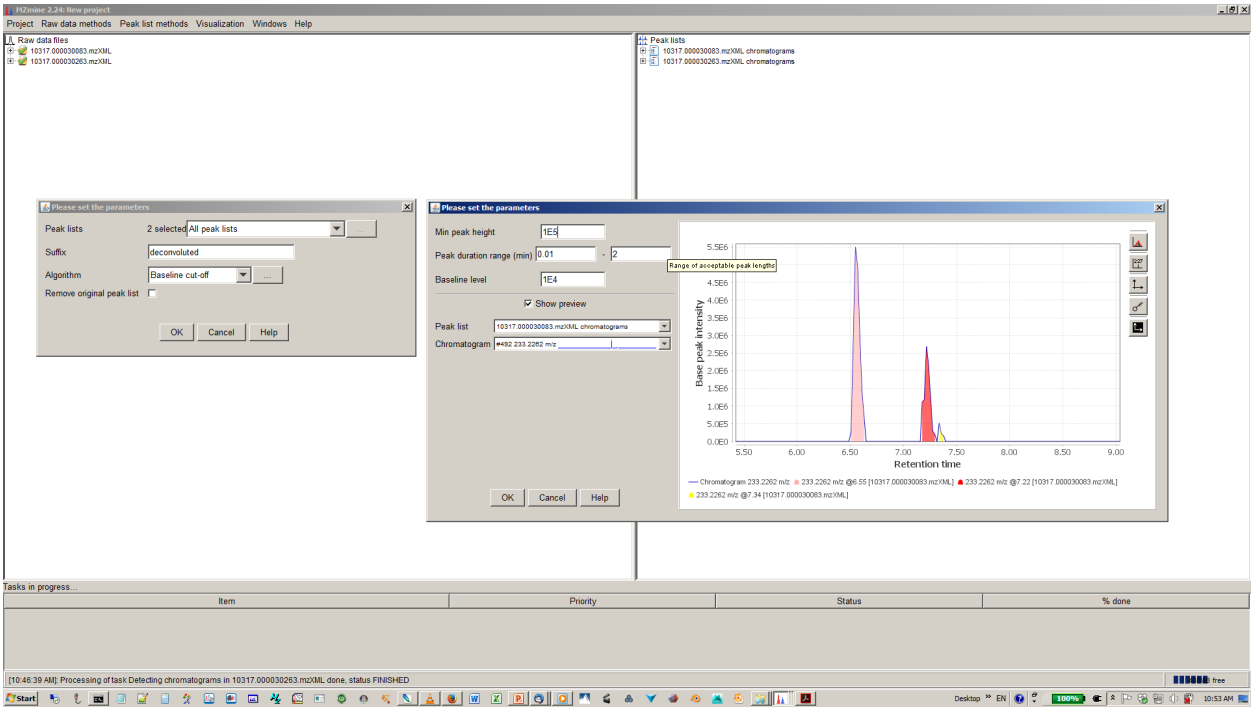
6. Specify mass list, mass tolerance min. time span and min. hight



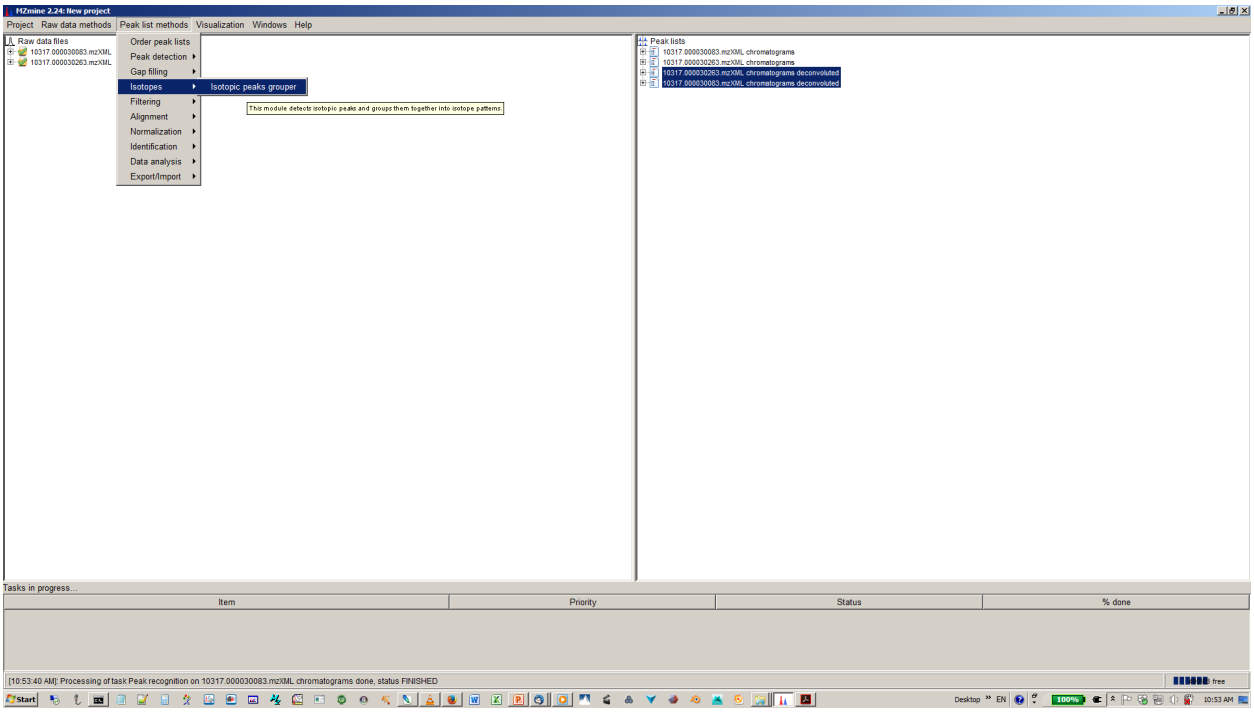
7. Deconvolute isobaric peaks with chromatogram deconvolution



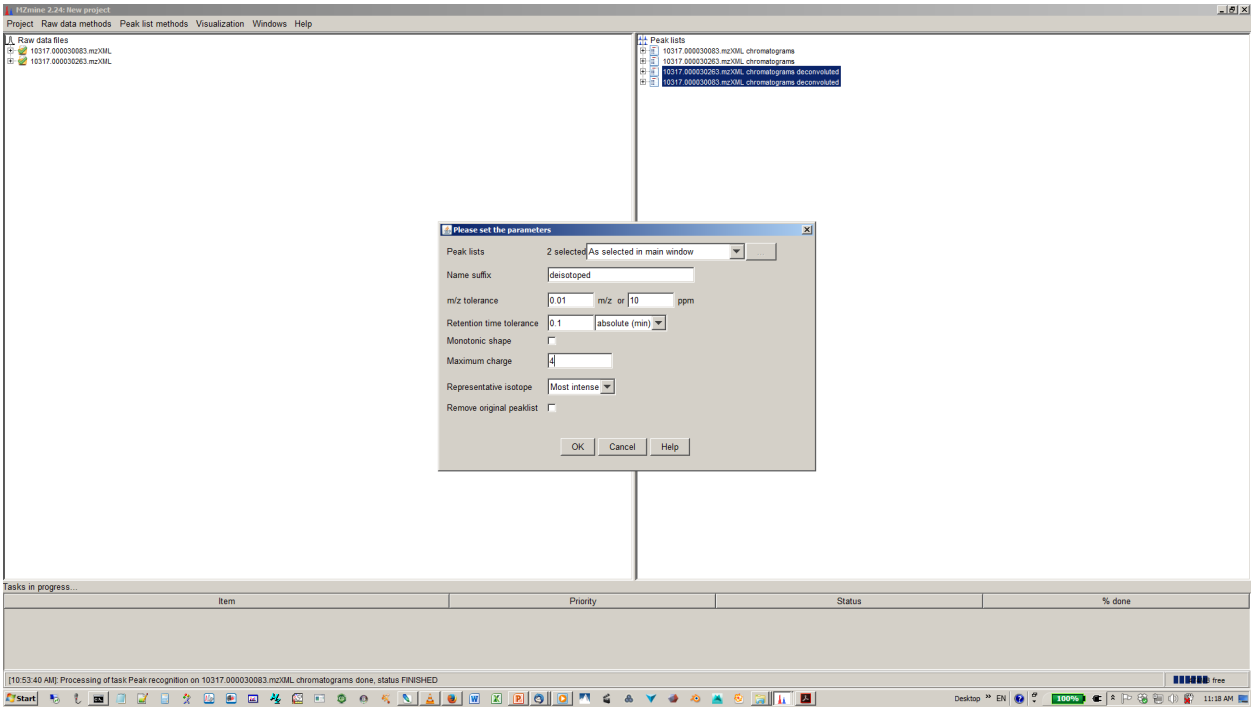
8. Specify algorithm (base line cut-off or local minimum search and parmaters



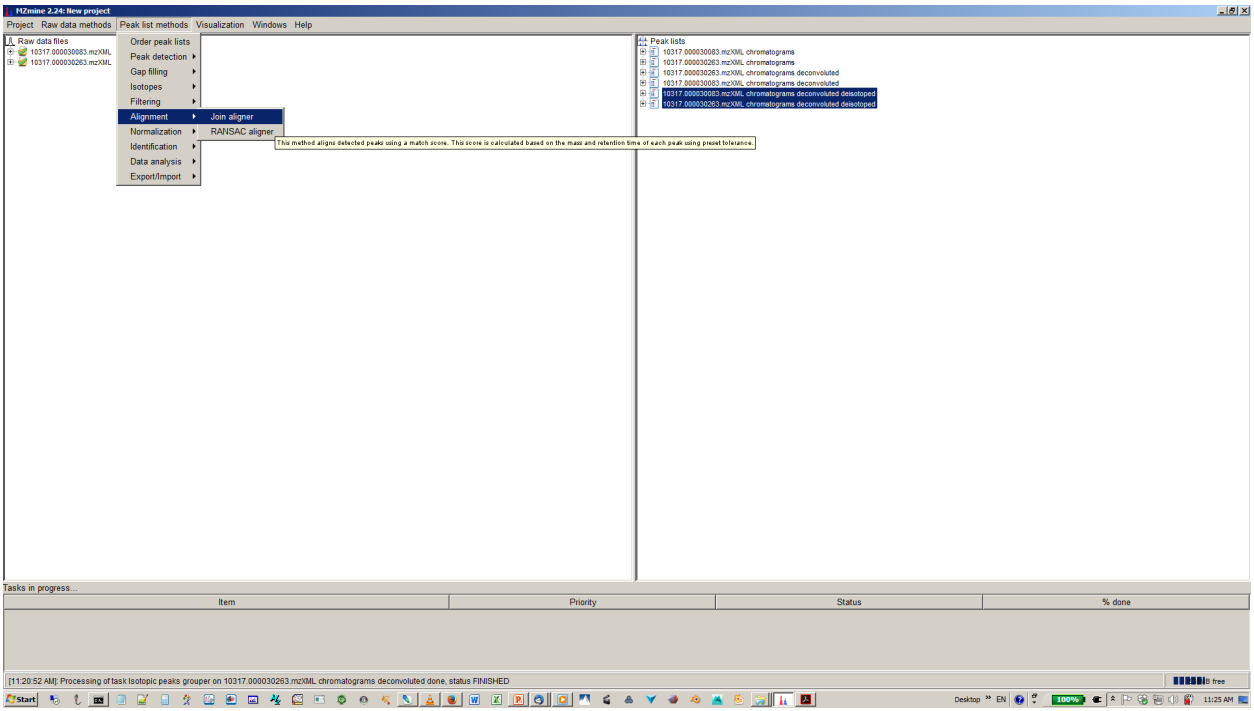
9. Perform deisotopization through isotope peak grouper



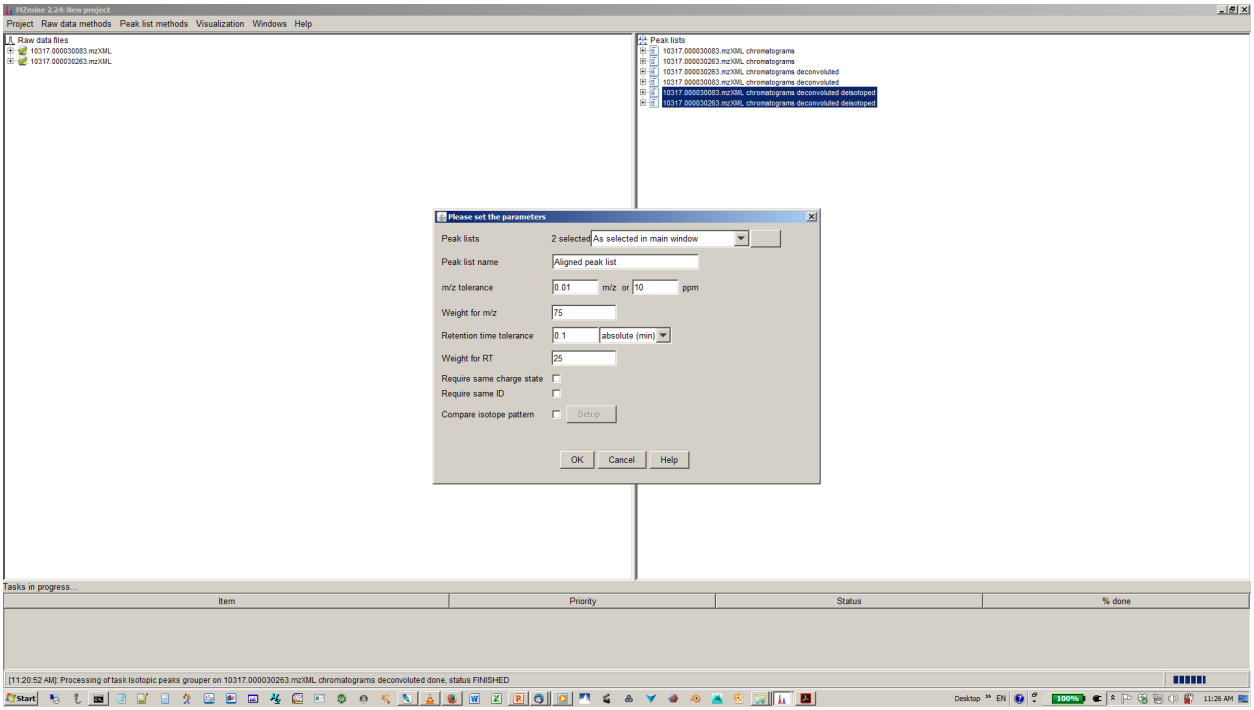
10. Specify parameters for isotope peak grouping



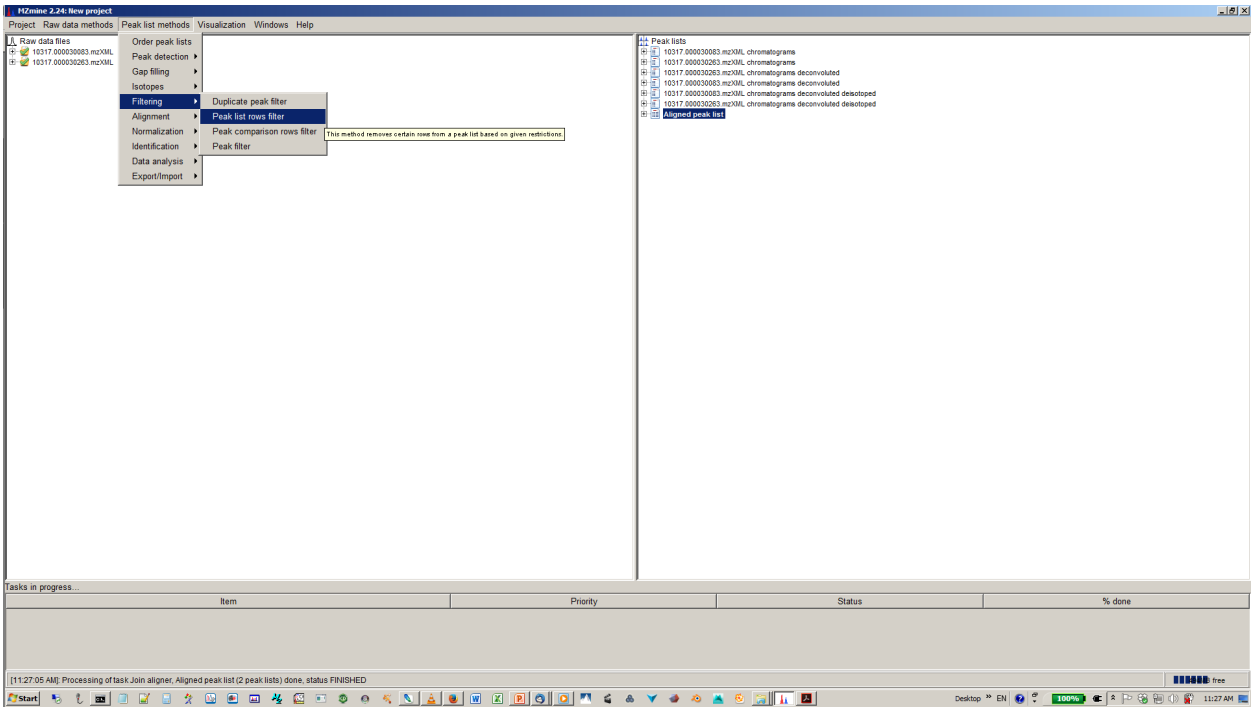
11. Align XICs from different sample to one matrix



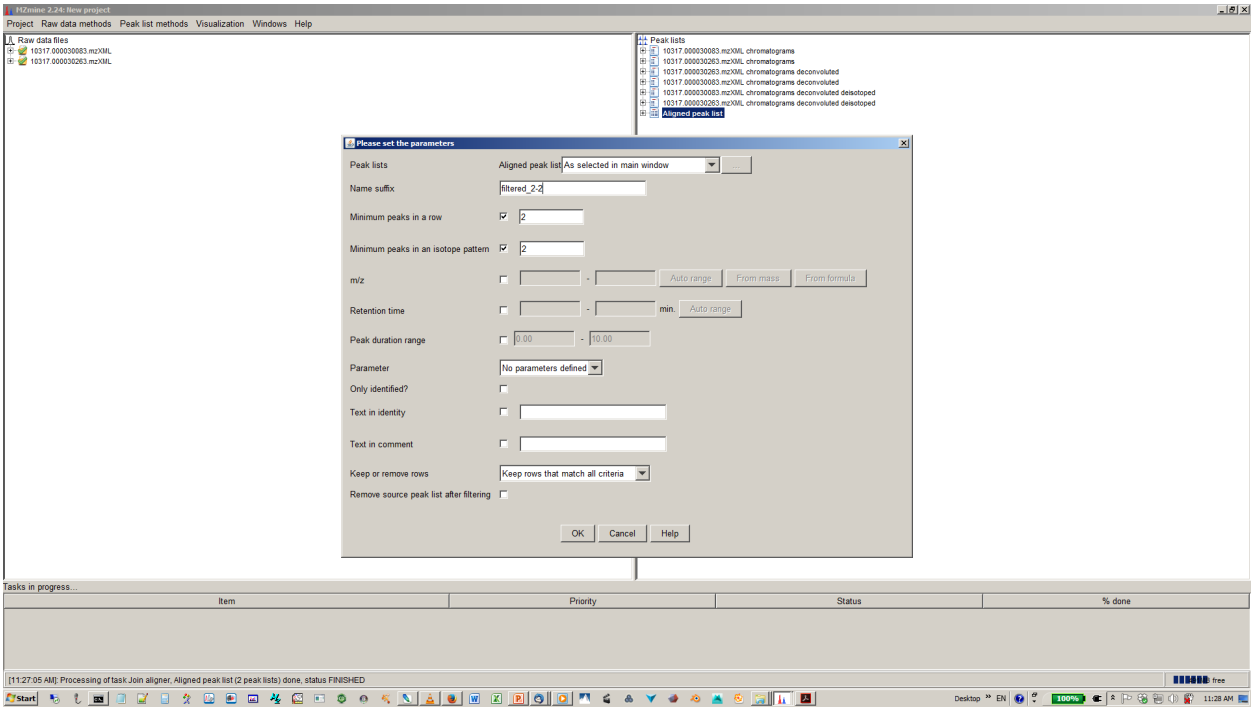
12. Specify join aligner parameters



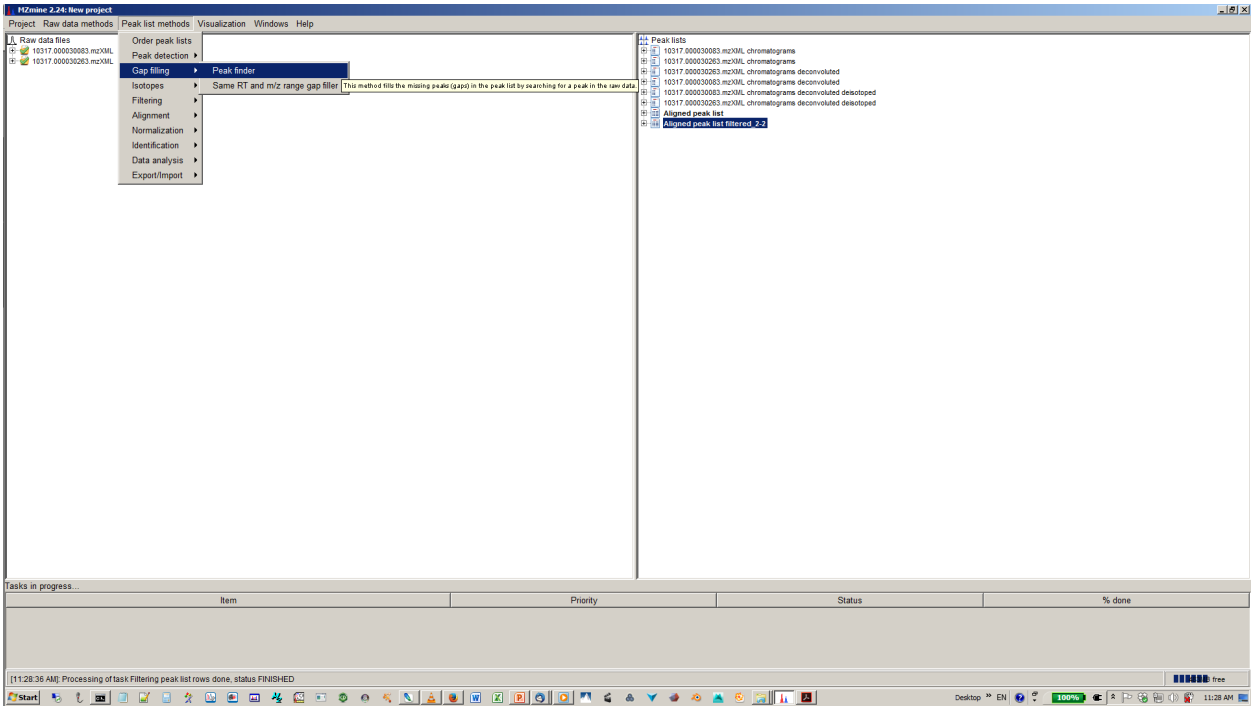
13. [optional] Filter aligned feature matrix with peak list row filter



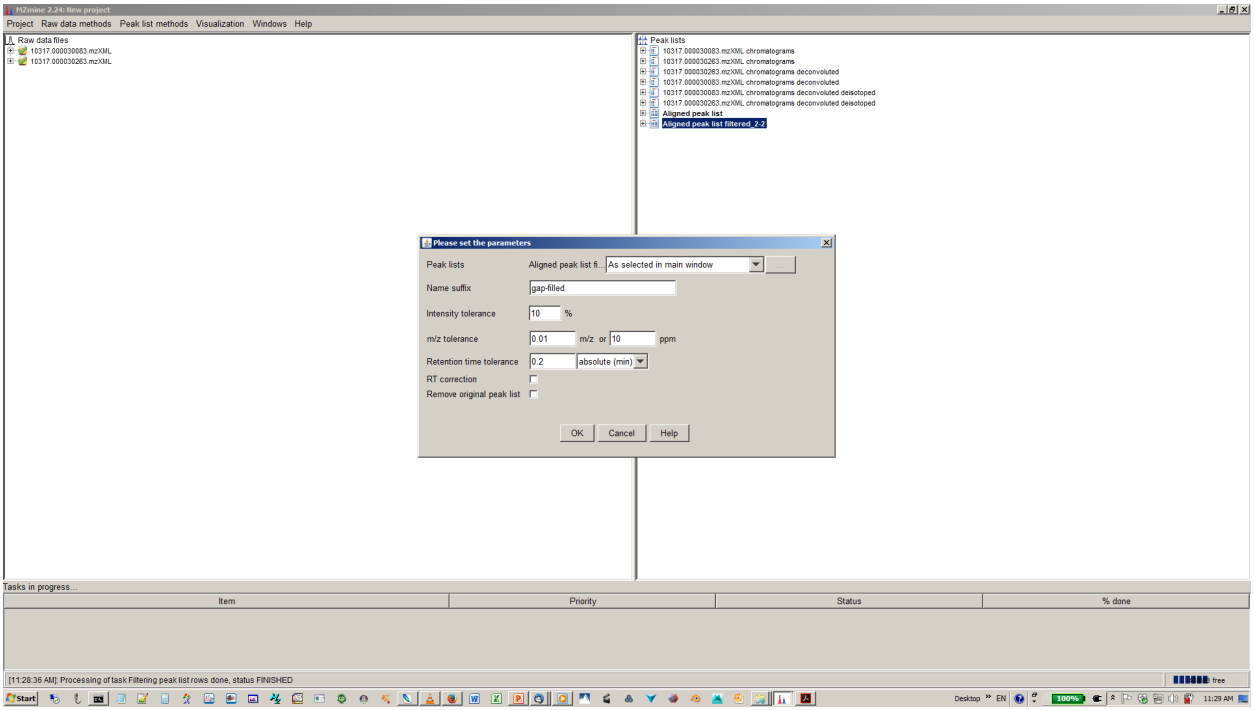
14. [optional] Depending of your experimental design use n minimum peaks in a row (n should be around the number of replicates or samples you expect to be similar) and 2-3 minimum peaks per isotope pattern



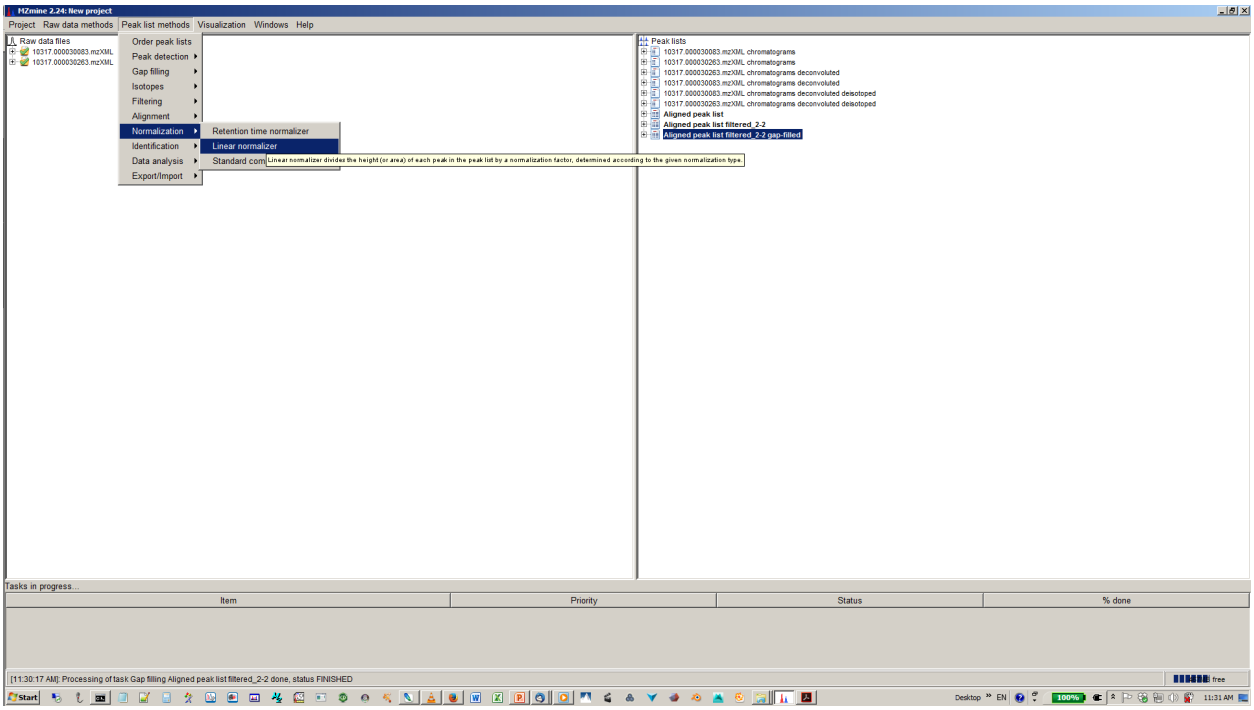
15. [optional] You gap filling the re-analyses missed peaks and fill gaps in the feature matrix



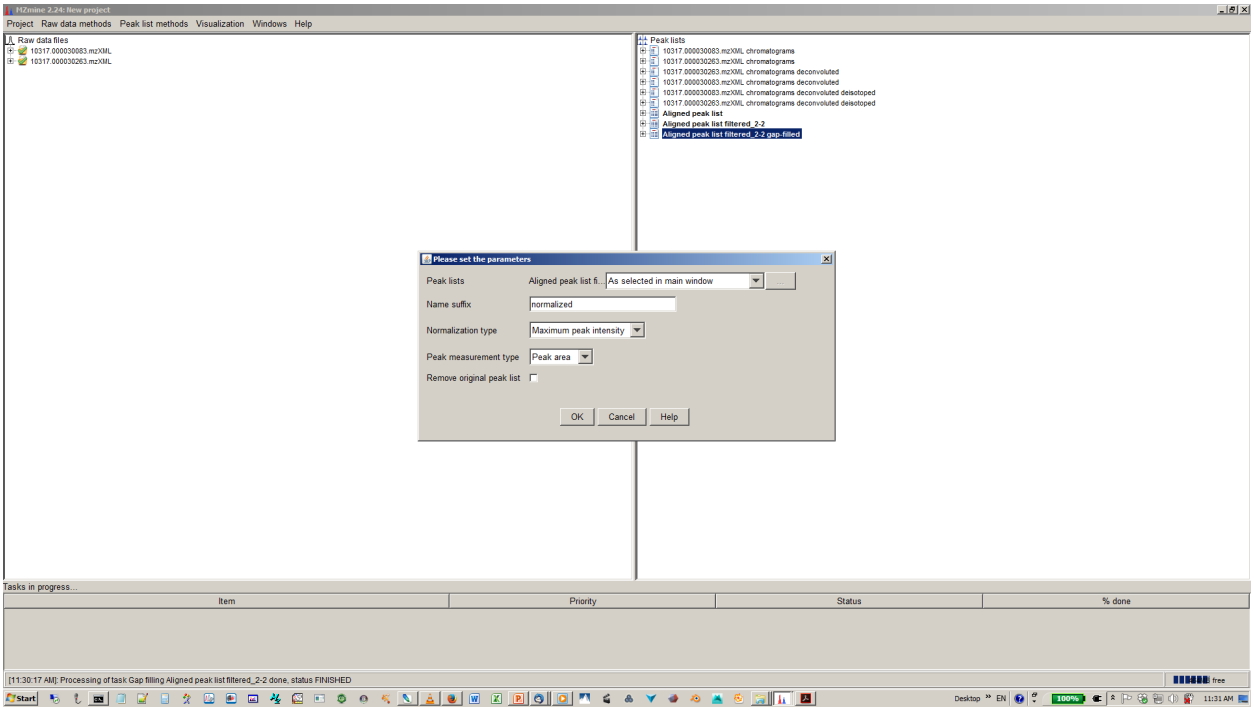
16. [optional] Depending on experimental design you can normalize your peak intensities to internal standards, TICs or total peak area.



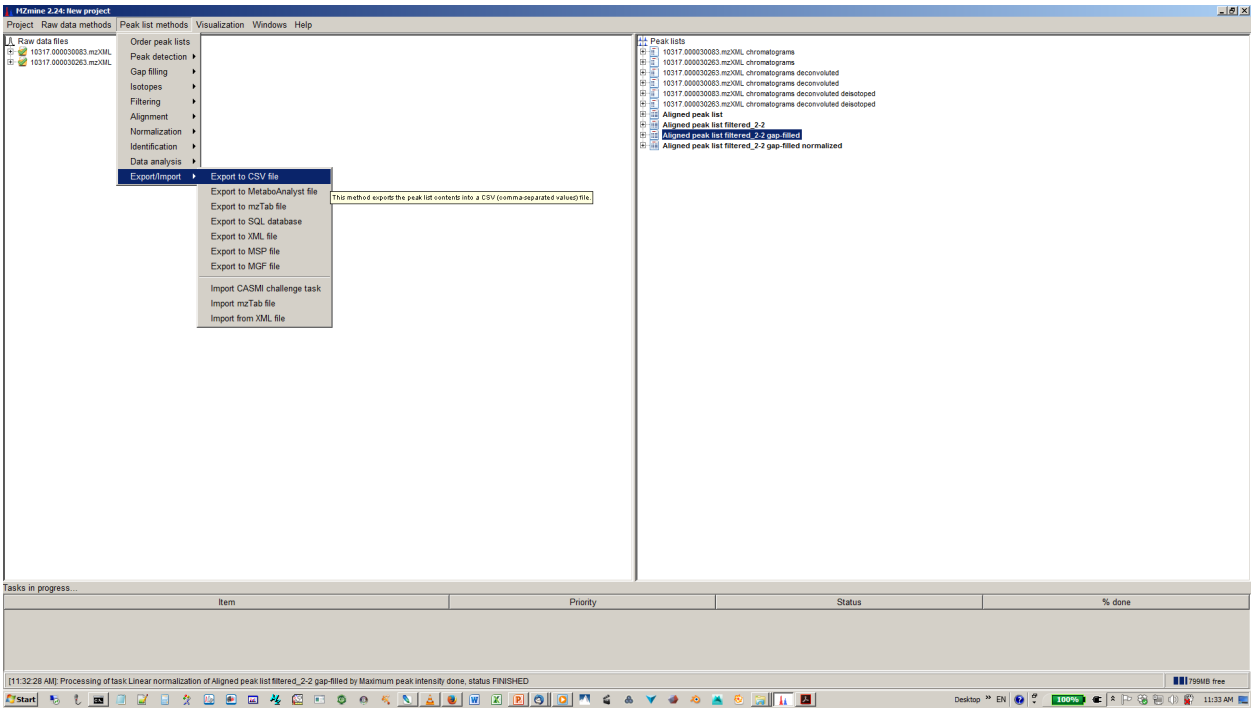
17. [optional] Specify normalization parameters

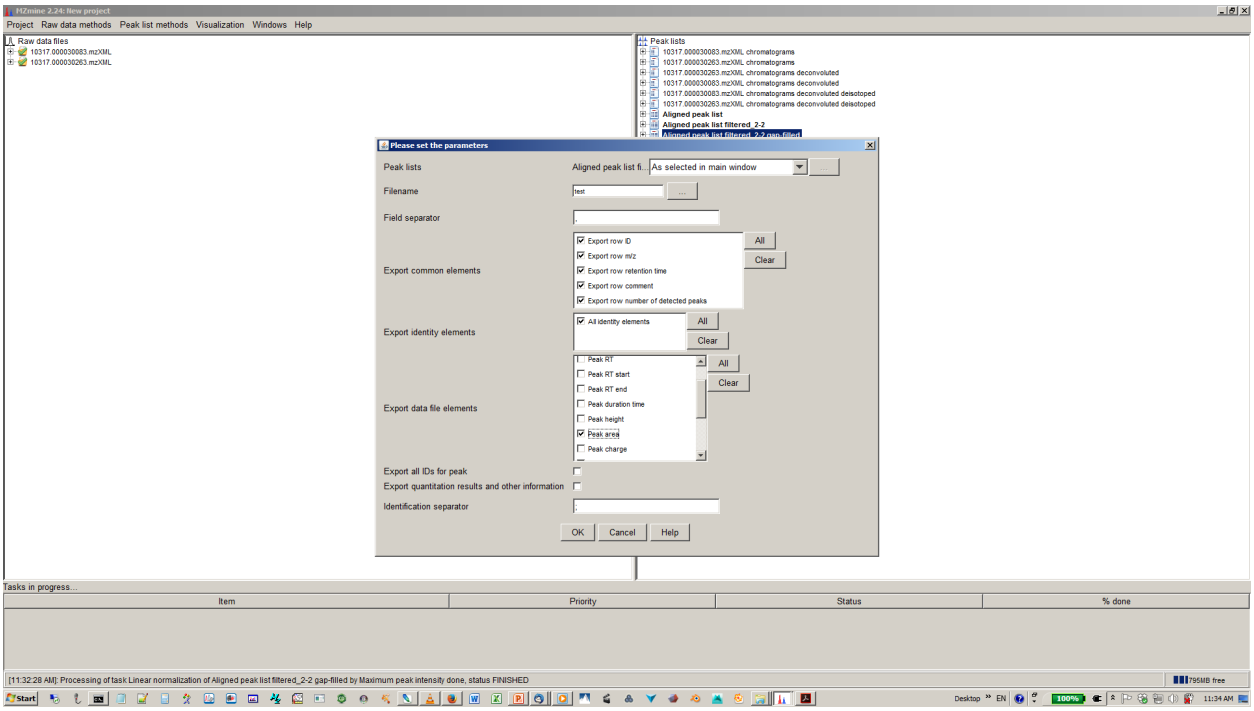


18. Export your matrix as .csv file for down stream data analysis



19. select file name and parameters you want to export





Here is also a video for MZmine 2 documentation:

Metabolomics demo data in Qiita

- Refer to the Qiita documentation about Principal Coordinates Analysis (PCoA) here

GNPS tutorial for MS/MS data annotation

Global Natural Products Social Molecular Networking GNPS web-platform provides public data set deposition and/or retrieval through the Mass Spectrometry Interactive Virtual Environment (MassIVE) data repository. The GNPS analysis infrastructure further enables online dereplication, automated molecular networking analysis, and crowdsourced MS/MS spectrum curation. Each data set added to the GNPS repository is automatically reanalyzed in the next monthly cycle of continuous identification. For more information, please check out the GNPS paper published in Nature Biotechnology by Ming et al 2016 here as well as the video and the ressource on Youtube, and well as on the online documentation

Tutorial: Generation of Molecular Networks in 15 minutes: Exploring MS/MS data with the GNPS Data Analysis workflow

Step 1- Go to GNPS and create an account

Go to GNPS main page in an other window http://gnps.ucsd.edu and create your own account first (important!)

Please Login to Use Workflows

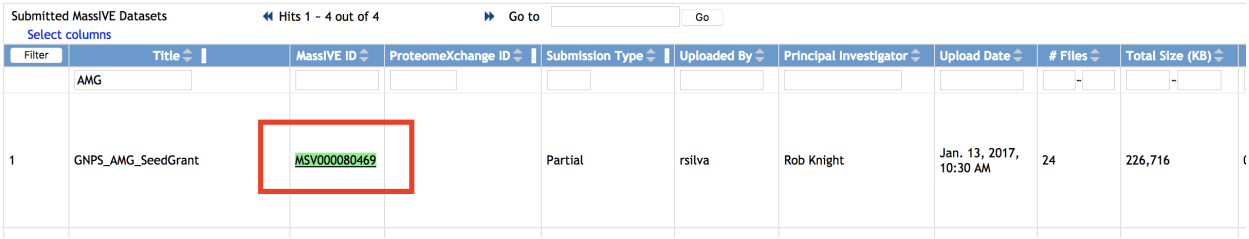
The Future of Natural Products Research and Mass Spectrometry

Step 2- Find a MS/MS dataset on MassIVE (Mass spectrometry Interactive Virtual Environment)

The Future of Natural Products Research and

A) Go to GNPS and access the MassIVE datasets repository.

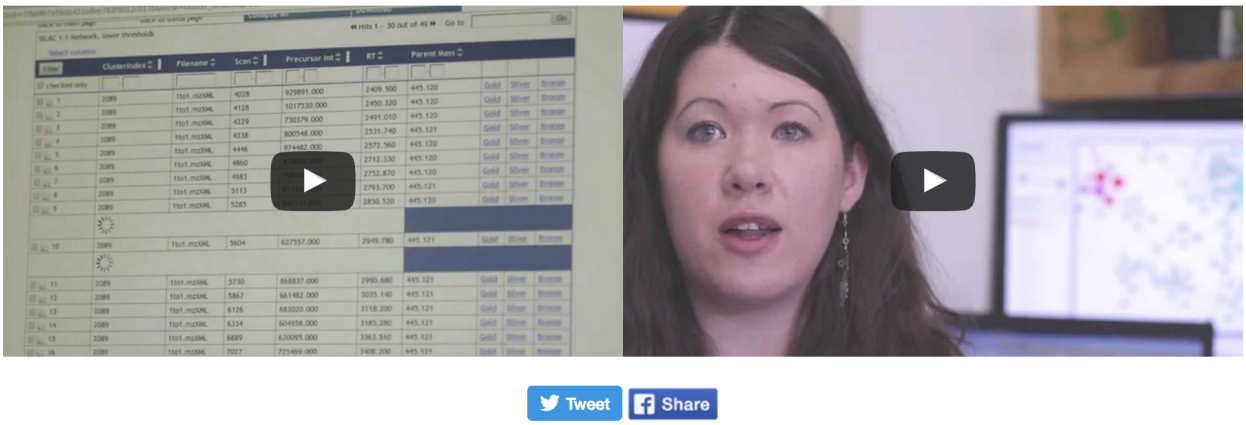
B) Search for the MassIVE datasets named "GNPS Workshop" (or "GNPS_AMG_SeedGrant" for a larger example with American Gut Projects samples).Explore its content, and copy the MassIVE ID number (MSV)



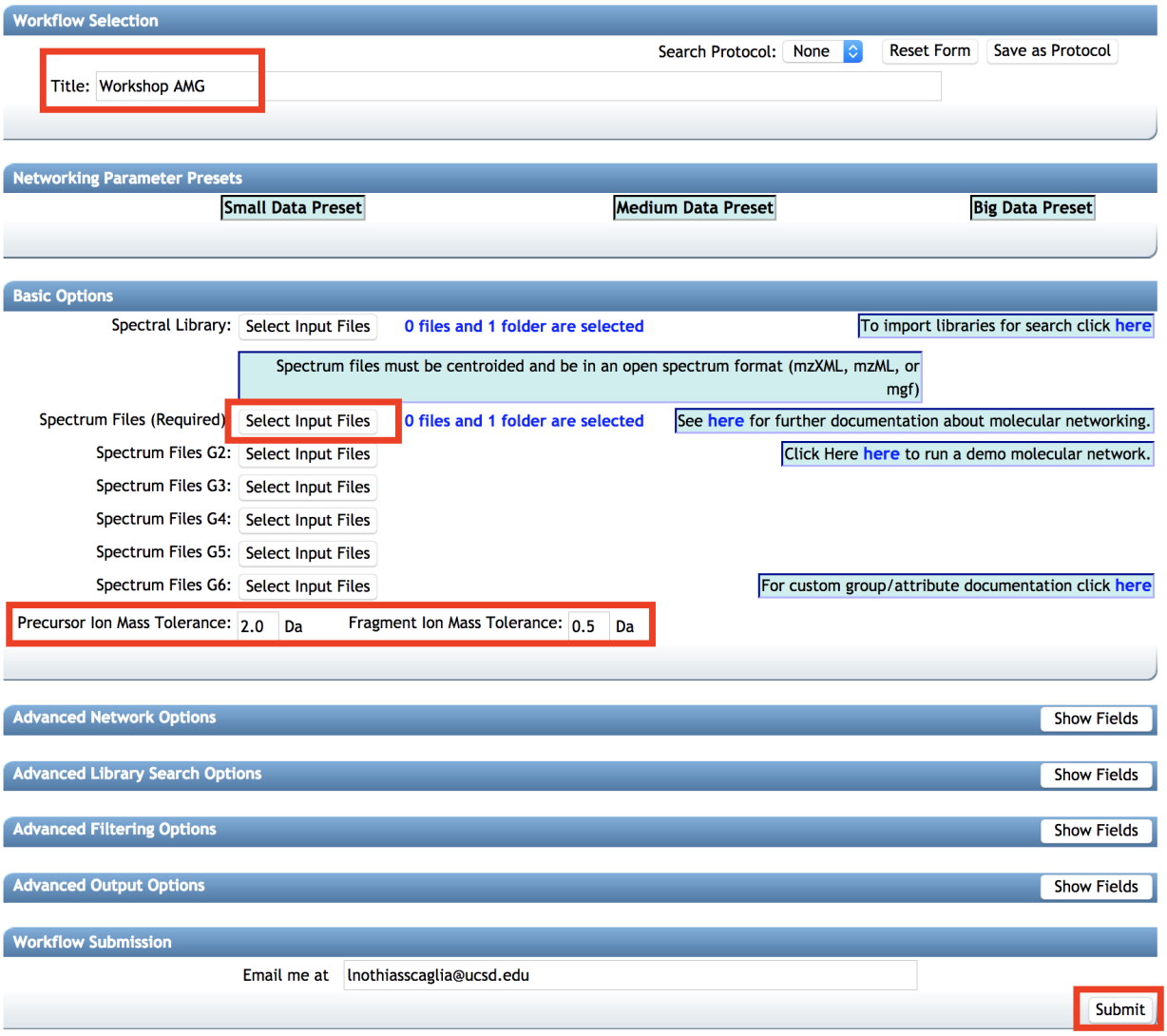
Note: If you want to upload your own data, follow the DorresteinLab youtube chanel, here is the video:

Step 3 - Access to the Data Analysis workflow

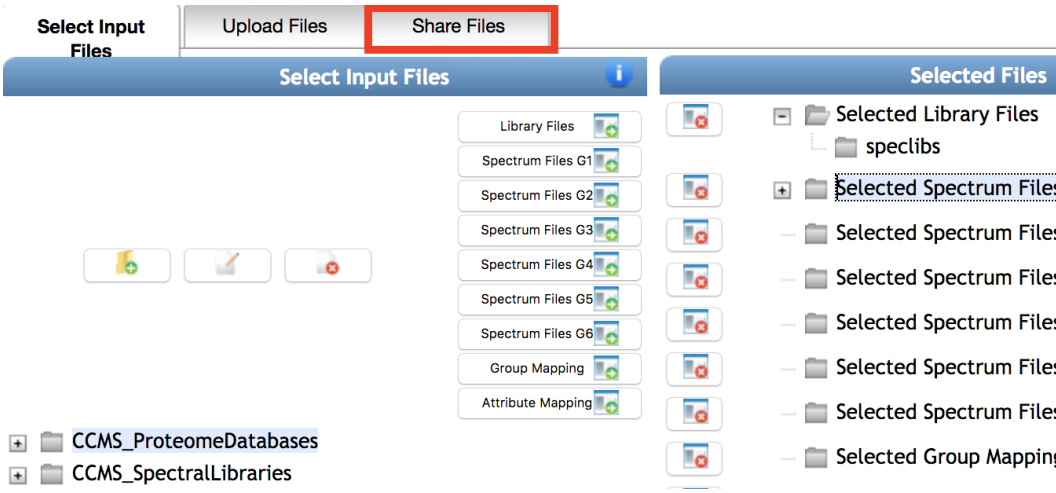
Go back GNPS main page and open the Data Analysis workflow.



Step 4 - Configure and launch the Data Analysis workflow



A) Indicate a Title.



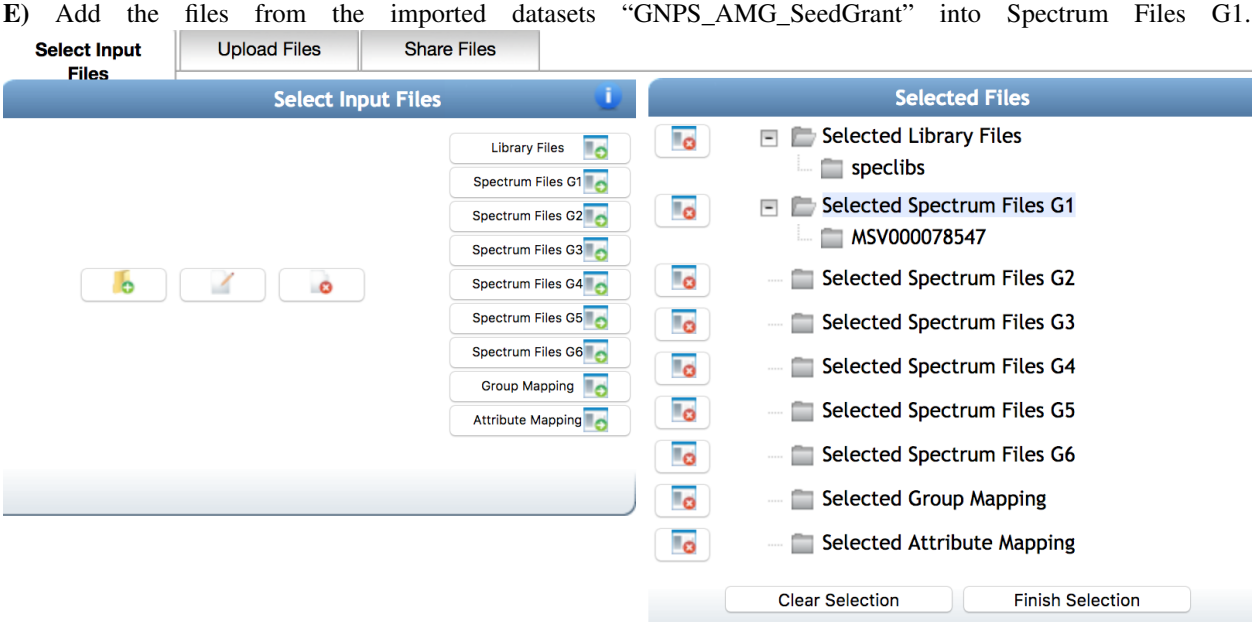
B) Clic on Spectrum Files (required)

C) Go to the Share Files spreadsheet and import the Massive dataset files for the "GNPS workshop" or "GNPS_AMG_SeedGrant" with the Import Data Share (use the MassIVE ID).

figs/GNPS_import.png

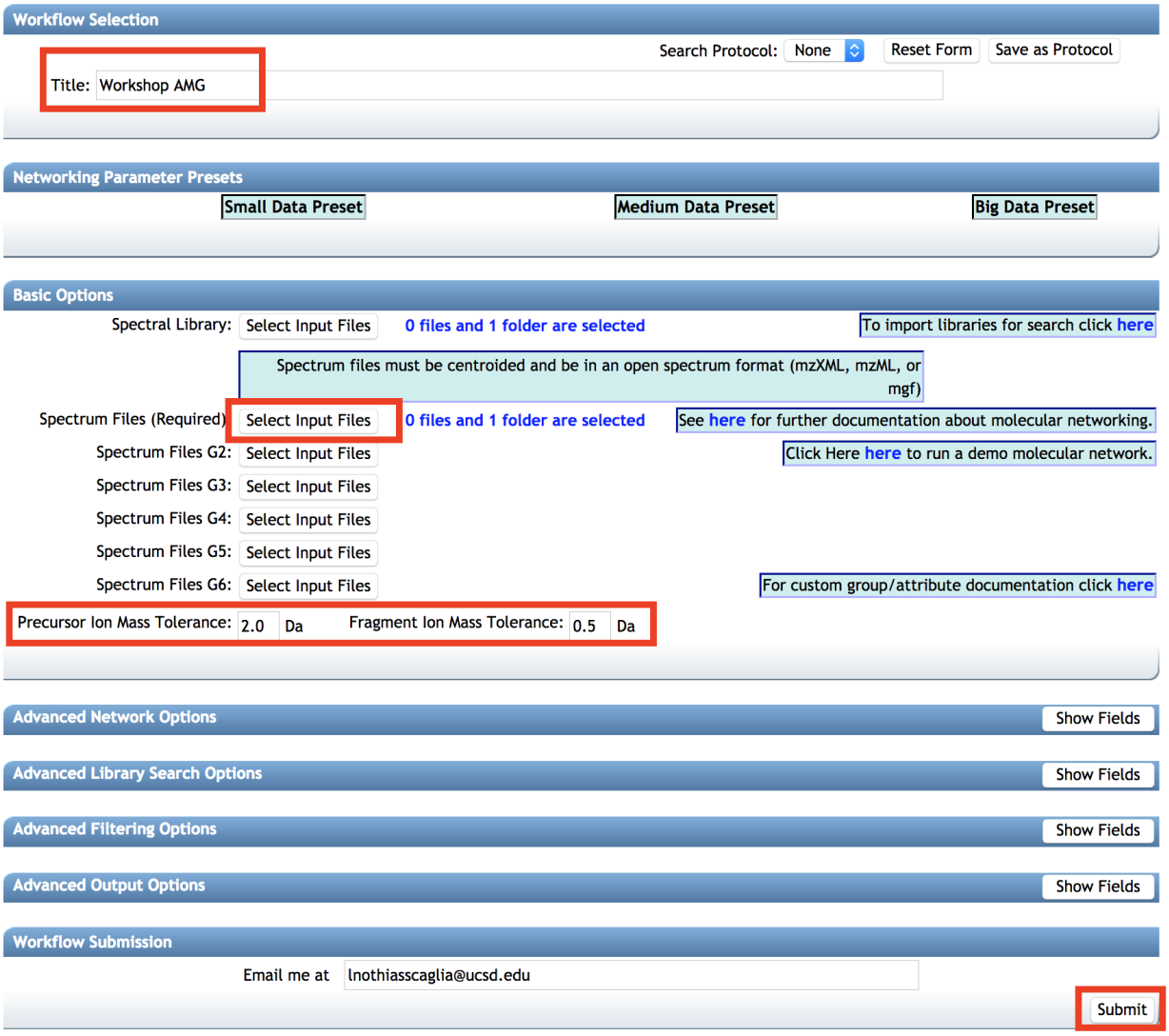
D) Go back to the Select Input Files spreadsheet.

E) Add the files from the imported datasets "GNPS_AMG_SeedGrant" into Spectrum Files G1.



F) Validate the selection with Finish Selection button.

G) Modifiy parameters to meet high-resolution mass spectrometry: Precursor Ion Mass Tolerance (0.02), Fragment Ion Mass Tolerance (0.02), Min Pairs Cos (0.6), Minimum Matched Fragment Ions (2), Minimum cluster size (use 1)

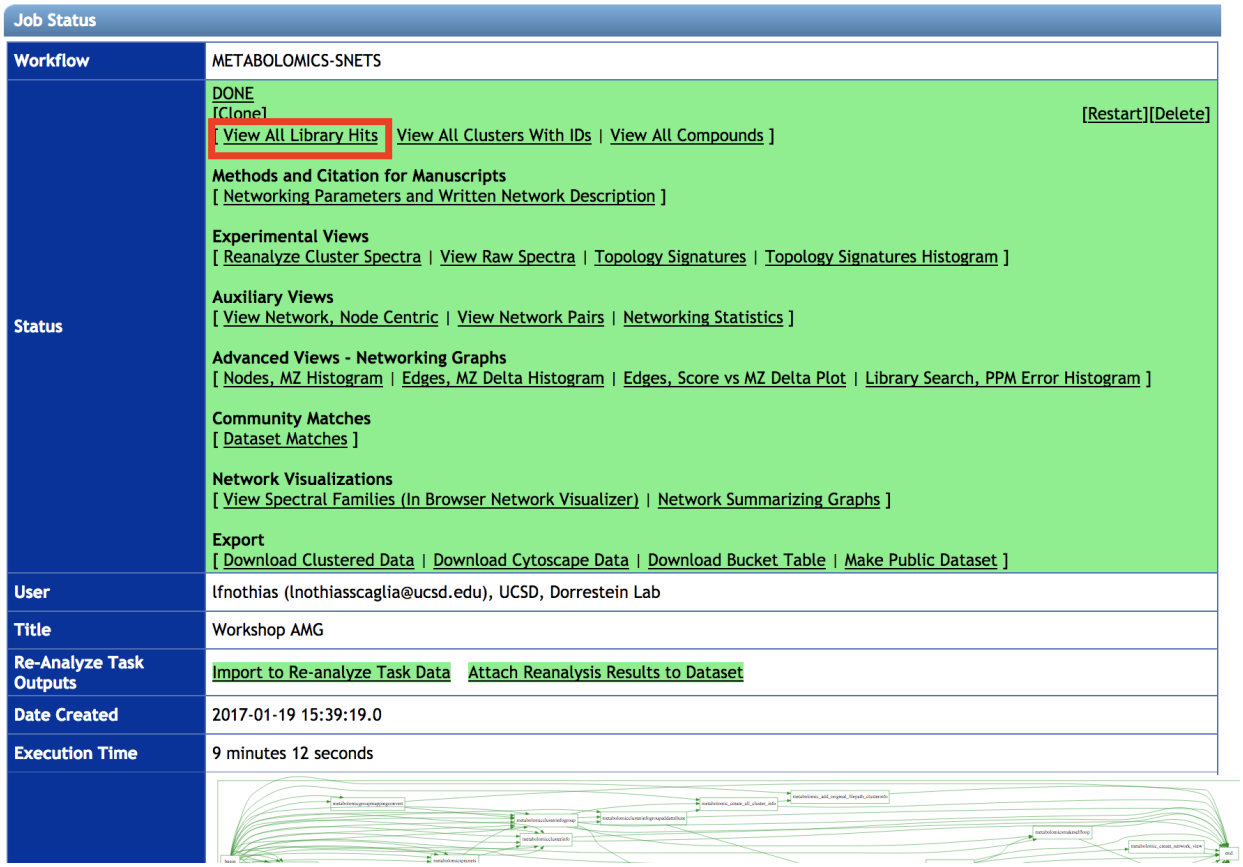


H) Launch the Data Analysis workflow using the Submit button.

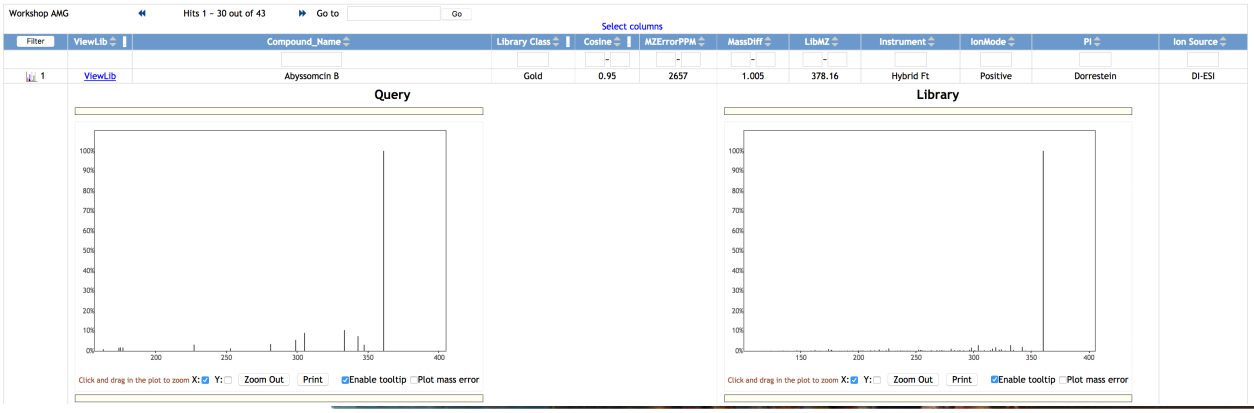
Step 5 - Visualize the Data Analysis workflow output

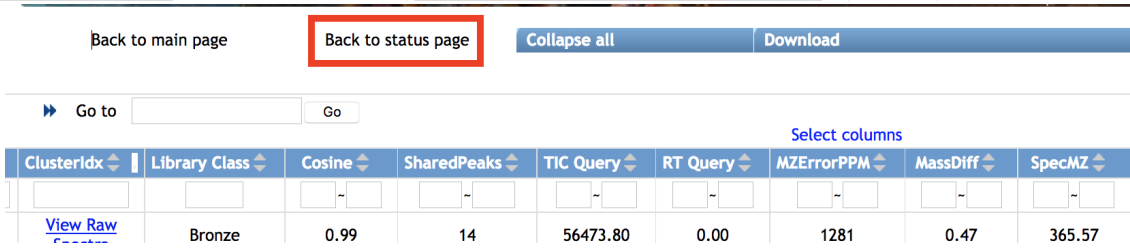
A) Return to GNPS main page and go to the Jobs page. Please find here an example of GNPS data analysis output with American Gut Project.

The Future of Natural Products Research and Mass Spectrometry



B) Explore the molecule annotated using public spectral library available on GNPS. Click on View All Library Hits.



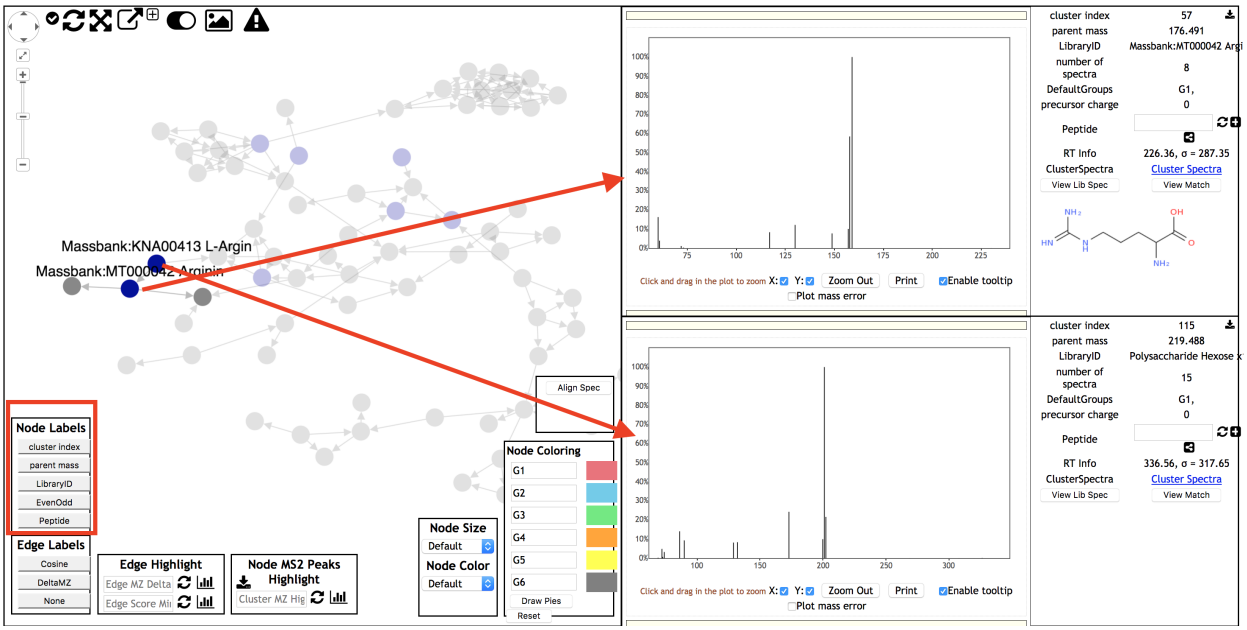


C) Go back to the Status Page

D) Clic on the View Spectral families and visualize the molecular network 1

Status	**Auxiliary Views** [View Network, Node Centric | View Network Pairs | Networking Statistics] **Advanced Views - Networking Graphs** [Nodes, MZ Histogram | Edges, MZ Delta Histogram | Edges, Score vs MZ Delta Plot | Library Search, PPM Error Histogram] **Community Matches** [Dataset Matches] **Network Visualizations** [View Spectral Families (In Browser Network Visualizer) | Network Summarizing Graphs] **Export** [Download Clustered Data | Download Cytoscape Data | Download Bucket Table | Make Public Dataset]
User	lfnothias (lnothiasscaglia@ucsd.edu), UCSD, Dorrestein Lab
Title	Workshop AMG

E) In Node Labels (bottom left), map the parent mass, or the LibraryID, in the molecular network.



F) Visualize a first MS/MS spectrum by left-clicking on one node. Visualize a second MS/MS spectrum by right-clicking on a second node.

More on navigating into the results with the following video: