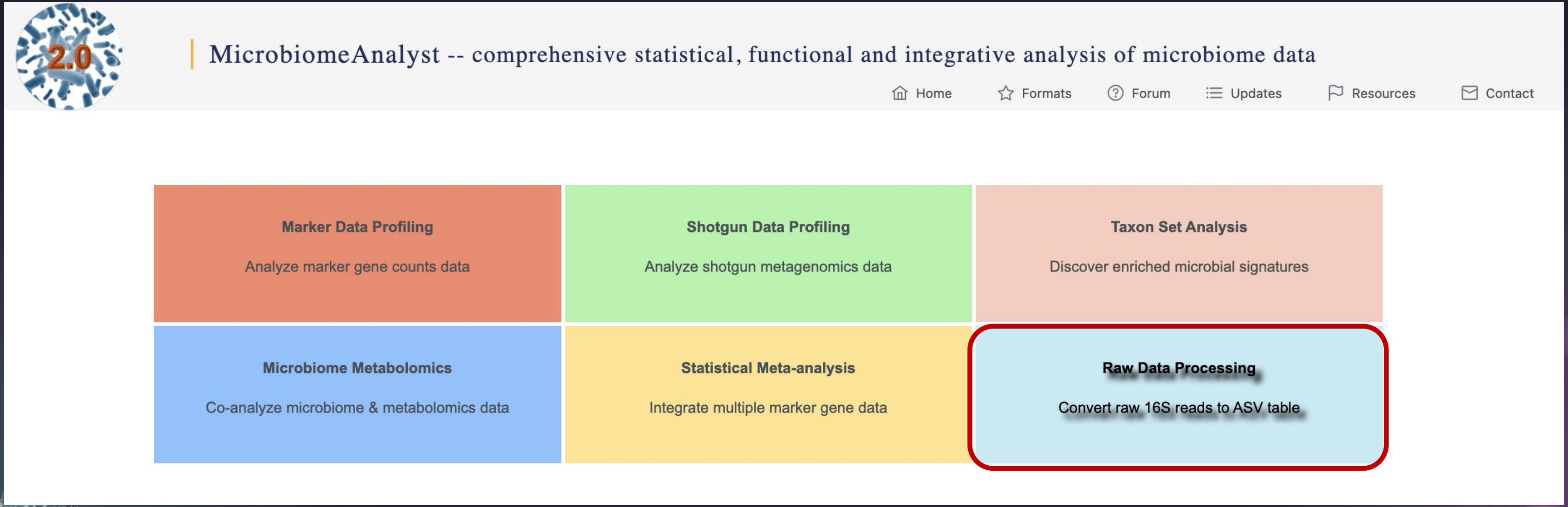
MicrobiomeAnalyst 2.0

Comprehensive statistical, functional and integrative analysis of microbiome data

xialab@mcgill 2023-Mar-02

Tutorial for Raw Data Processing



MicrobiomeAnalyst -- comprehensive statistical, functional and integrative analysis of microbiome data

Marker Data Profiling

Analyze marker gene counts data

Shotgun Data Profiling

Analyze shotgun metagenomics data

Taxon Set Analysis

Discover enriched microbial signatures

Microbiome Metabolomics

Co-analyze microbiome & metabolomics data

Statistical Meta-analysis

Integrate multiple marker gene data

Raw Data Processing

Convert raw 16S reads to ASV table

Background

- Amplicon sequencing has enabled comprehensive profiling of microbial communities, bypassing traditional wet lab culturing methods.

- Traditional Operational Taxonomic Units (OTU) picking methods work by clustering sequences based on a similarity threshold (usually around 97%). However, this method tends to introduce sequencing level errors into the reads due to the arbitrary clustering threshold.

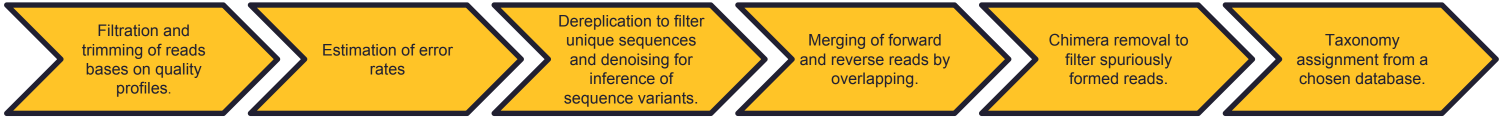
- The Divisive Amplicon Denoising Algorithm (DADA) was introduced to improve the accuracy of amplicon sequence variant (ASV) inference from high-throughput sequencing data.

- DADA2 uses a statistical model-based approach that corrects these incorporated errors and infers higher quality and more accurate ASVs, which help improve our understanding of complex and previously understudied microbial ecosystems.

Overview

- **Goal**: To provide a user-friendly web-based platform for the raw data processing of marker gene sequencing data of microbial communities.

- **Workflow**:

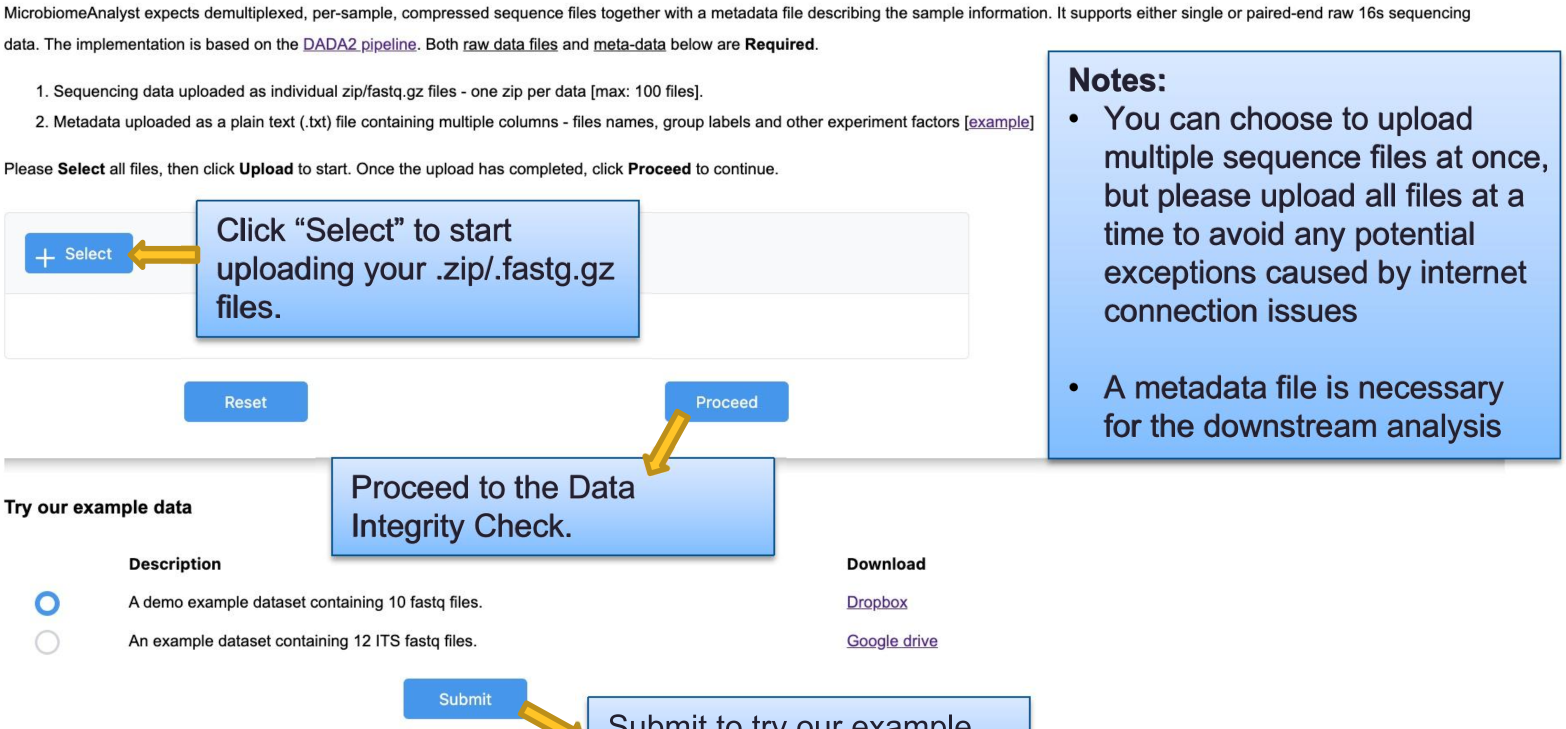


Filtration and trimming of reads bases on quality profiles.
Estimation of error rates
Dereplication to filter unique sequences and denoising for inference of sequence variants.
Merging of forward and reverse reads by overlapping.
Chimera removal to filter spuriously formed reads.
Taxonomy assignment from a chosen database.

- **Data requirements**:
  - Demultiplexed individual fastq files with no primers or any other non-biological nucleotides.
  - For paired-end data, the forward and reverse fastq files should have matching ordered names with "_R1" for forward reads and "_R2" for reverse reads, as shown in the example data.
  - Additionally, a metadata file indicating the groups is required to facilitate a streamlined input into the other MicrobiomeAnalyst modules.

- **Other considerations for paired-end data**:
  - What is the length of the forward and reverse reads? For e.g., 2x200bp
  - What was the target region of the 16S rRNA gene that was sequenced and what were your primer lengths? For e.g., V4, V3-V4, etc.

Data Upload:



MicrobiomeAnalyst expects demultiplexed, per-sample, compressed sequence files together with a metadata file describing the sample information. It supports either single or paired-end raw 16s sequencing data. The implementation is based on the DADA2 pipeline. Both raw data files and meta-data below are **Required**.

1. Sequencing data uploaded as individual zip/fastq.gz files - one zip per data [max: 100 files].
2. Metadata uploaded as a plain text (.txt) file containing multiple columns - files names, group labels and other experiment factors [example]

Please **Select** all files, then click **Upload** to start. Once the upload has completed, click **Proceed** to continue.

Try our example data

	Description	Download
●	A demo example dataset containing 10 fastq files.	Dropbox
○	An example dataset containing 12 ITS fastq files.	Google drive

Click “Select” to start uploading your .zip/.fastg.gz files.

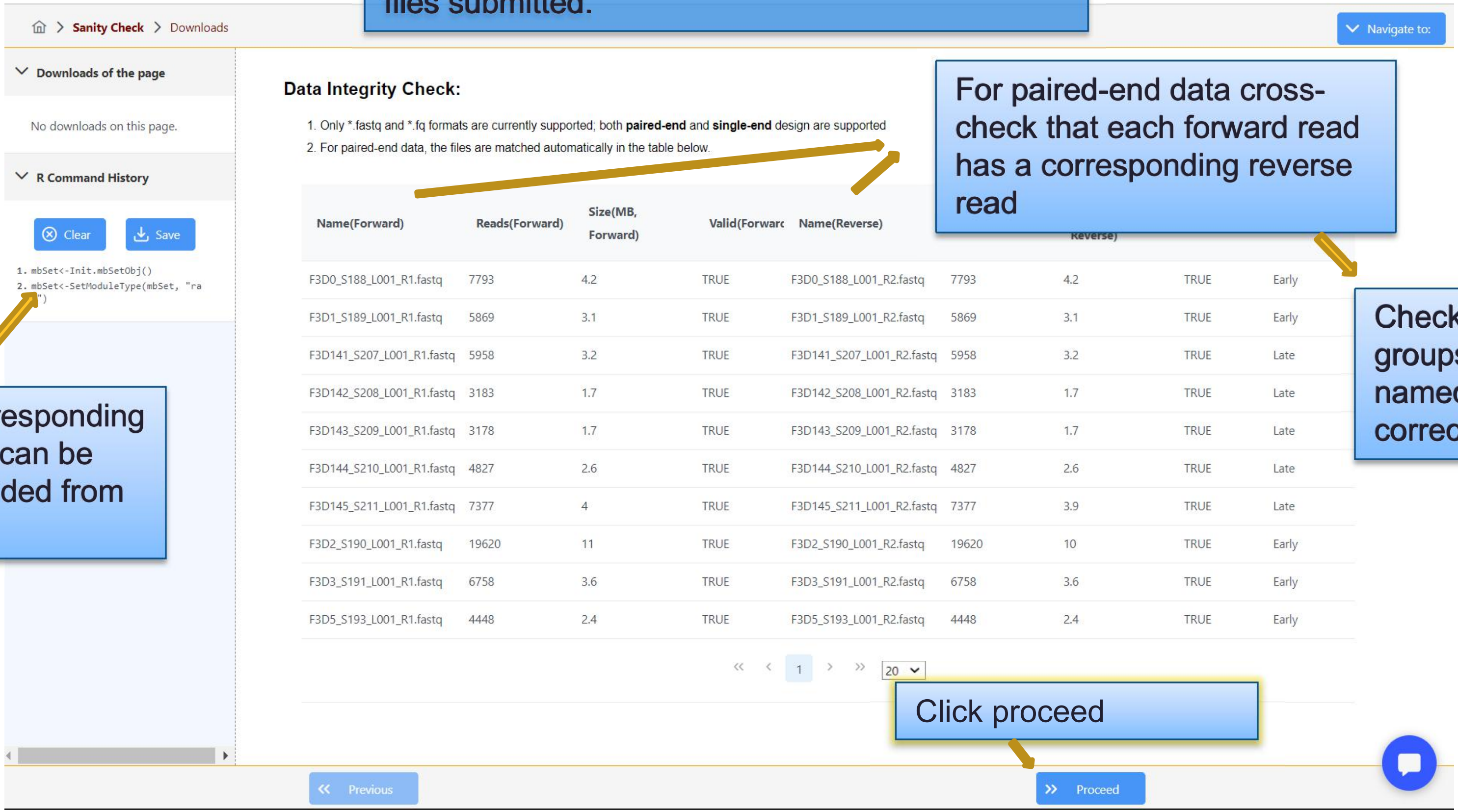
Proceed to the Data Integrity Check.

Submit to try our example here.

Notes:

- You can choose to upload multiple sequence files at once, but please upload all files at a time to avoid any potential exceptions caused by internet connection issues
- A metadata file is necessary for the downstream analysis

Data Integrity Check:



```
1. mbSet<-Init.mbSetObj()
2. mbSet<-SetModuleType(mbSet, "ra
   ")
```

Data Integrity Check:

1. Only *.fastq and *.fq formats are currently supported; both **paired-end** and **single-end** design are supported
2. For paired-end data, the files are matched automatically in the table below.

Name(Forward)	Reads(Forward)	Size(MB, Forward)	Valid(Forwarc	Name(Reverse)		Reverse)		
F3D0_S188_L001_R1.fastq	7793	4.2	TRUE	F3D0_S188_L001_R2.fastq	7793	4.2	TRUE	Early
F3D1_S189_L001_R1.fastq	5869	3.1	TRUE	F3D1_S189_L001_R2.fastq	5869	3.1	TRUE	Early
F3D141_S207_L001_R1.fastq	5958	3.2	TRUE	F3D141_S207_L001_R2.fastq	5958	3.2	TRUE	Late
F3D142_S208_L001_R1.fastq	3183	1.7	TRUE	F3D142_S208_L001_R2.fastq	3183	1.7	TRUE	Late
F3D143_S209_L001_R1.fastq	3178	1.7	TRUE	F3D143_S209_L001_R2.fastq	3178	1.7	TRUE	Late
F3D144_S210_L001_R1.fastq	4827	2.6	TRUE	F3D144_S210_L001_R2.fastq	4827	2.6	TRUE	Late
F3D145_S211_L001_R1.fastq	7377	4	TRUE	F3D145_S211_L001_R2.fastq	7377	3.9	TRUE	Late
F3D2_S190_L001_R1.fastq	19620	11	TRUE	F3D2_S190_L001_R2.fastq	19620	10	TRUE	Early
F3D3_S191_L001_R1.fastq	6758	3.6	TRUE	F3D3_S191_L001_R2.fastq	6758	3.6	TRUE	Early
F3D5_S193_L001_R1.fastq	4448	2.4	TRUE	F3D5_S193_L001_R2.fastq	4448	2.4	TRUE	Early

Each column gives information about the fastq files submitted.

For paired-end data cross-check that each forward read has a corresponding reverse read

Check if the groups are named correctly.

The corresponding R script can be downloaded from here

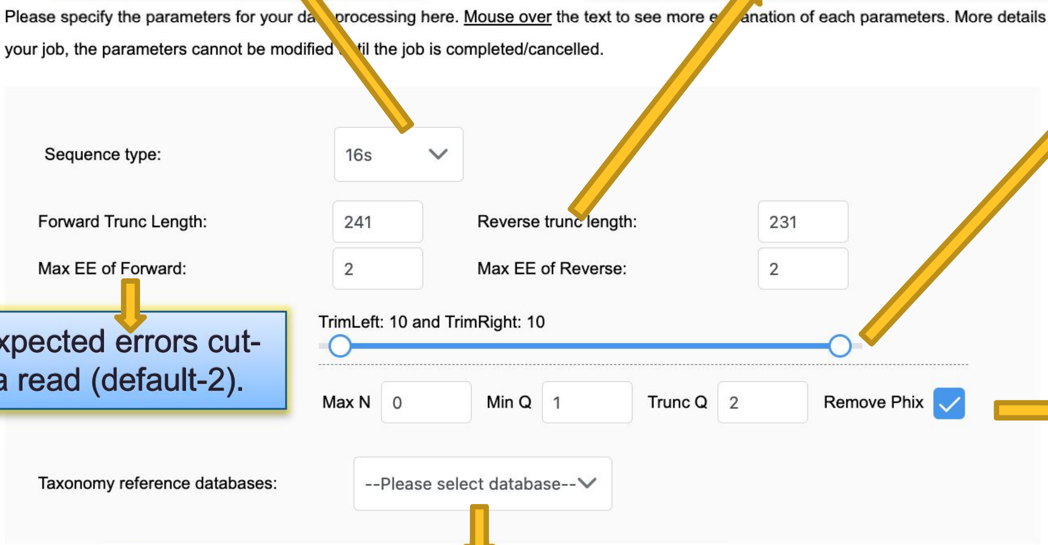
Click proceed

Parameter Settings:

This is the most critical step of the entire pipeline where the read quality profiles need to be examined to determine the filtering and trimming parameters.

Select the type of marker gene used: 16S for bacteria, 18S for eukaryotes, and ITS for fungi.

Choose the cut-off length for the forward and reverse reads based on the quality profile (see below). This will truncate the reads to a maximum length, maintaining reads of uniform length which is important during taxonomy assignment.



Please specify the parameters for your da processing here. Mouse over the text to see more e anation of each parameters. More details on thes

your job, the parameters cannot be modified til the job is completed/cancelled.

Sequence type: 16s

Forward Trunc Length: 241

Reverse trunc length: 231

Max EE of Forward: 2

Max EE of Reverse: 2

TrimLeft: 10 and TrimRight: 10

Max N 0

Min Q 1

Trunc Q 2

Remove Phix ☑

Taxonomy reference databases: --Please select database--

This is used to trim low quality bases on the 5’ end (TrimLeft) and 3’ end (TrimRight).

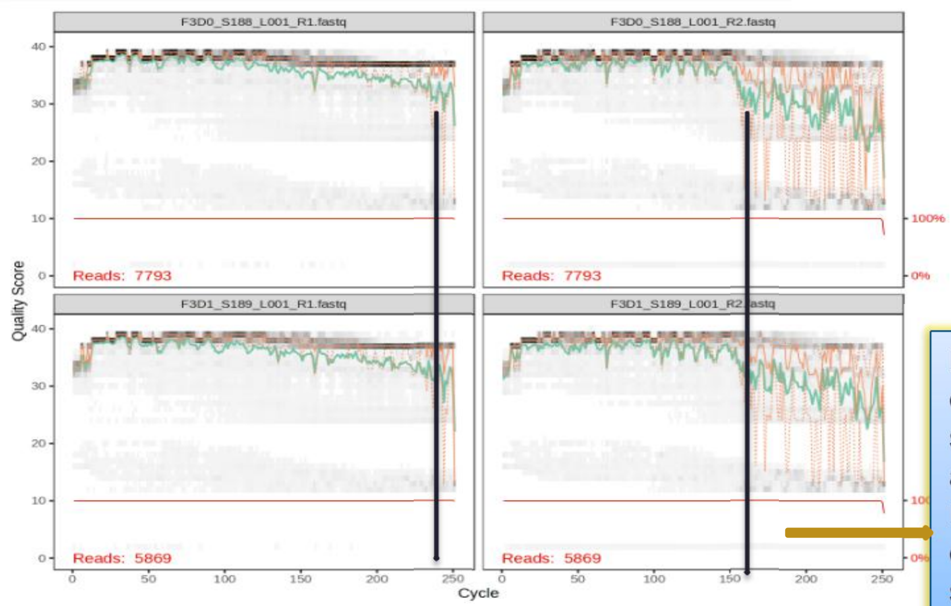
The expected errors cut-off in a read (default-2).

MaxN determines the number of ambiguous bases allowed. Typically, this is by default=0 which means no ambiguous bases would be allowed to pass through. **MinQ** and **TruncQ** are used to respectively filter out bases below a min. quality score and to truncate reads at the first instance of quality drop, below the specified score in the read. **RemPhix** removes reads that match an Illumina control genome called Phix. This ensures that only reads originating from the sample pass through.

Select the database of choice for taxonomy assignment.

Quality control:

The quality score of the raw sequences can be viewed on the Parameter Settings page to help adjust the parameters.



Typically any reads dropping below a quality score of 30 are considered to be low quality and are trimmed.

Forward reads tend to have better quality profiles than reverse reads.

For the forward reads (left panel) the quality drops off slightly at the end and so we will set the **forward trunc length** as 240.
For the reverse reads (right panel) the quality drops off around 170 cycles and so the **reverse trunc length** should be set as 170.

Note: In order to ensure overlap of forward and reverse reads, the trunc length parameters depend on the type of primer used. Refer to the “other considerations section on slide 4.

Parameter optimisation:

- **Do your results have very few reads passing through?** Consider changing the following parameters**:**
  - For **multi-V-regions such as V3-V4**, the overlap of merged reads is determined as follows:
  - For 2x250bp, 16S-341F and 16S-805R primers of the V3-V4 region,

    ```
    (forward read) + (reverse read) - (length of amplicon)  = overlap
    250              + 250            -  (805-341)              = 36
    ```
  - If the forward read is truncated at 240 and reverse read is truncated at 150,

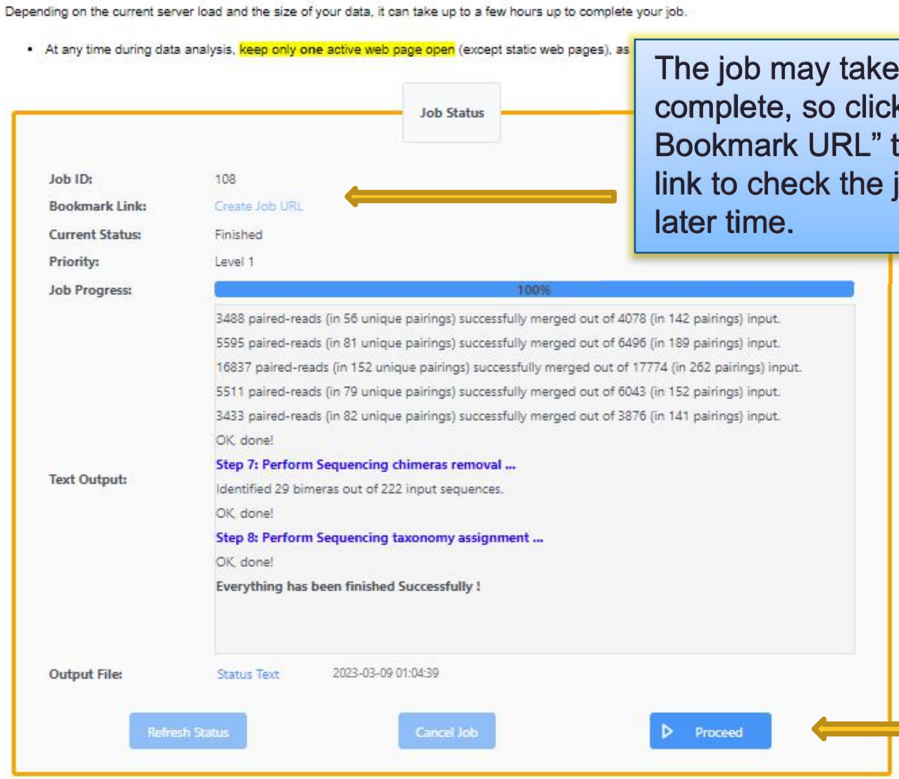
    ```
    240              + 150            -  464                    = -74 (No overlap!!!!)
    ```
  - Thus the parameters should be adjusted accordingly to ensure an overlap of >20nt.
  - For the **V4 region**, there is usually less variability and the parameters can be directly based off the quality profiles.
  - For more information visit- https://forum.qiime2.org/t/merging-quality-control-and overlapping/12618/2

- Do you still find very few reads passing through? Consider increasing the **Max EE parameter** which would allow less stringent filtering, especially for reverse reads. E.g.: Max EE of reverse= 5

- Is the percentage of **chimera removal >25%**? Check if all non-biological nts such as adapters and primers were removed properly. Consider trimming your sequences more using the Trim parameters. If the chimera removal is still high but the number of reads passing through are sufficient, you could consider moving ahead with the results. More information - https://forum.qiime2.org/t/loss-of-reads-after-dada2-as-chimeras/9503/2

Job Status Tracking:

Track the processing status here. The job status will update here in real-time.



Depending on the current server load and the size of your data, it can take up to a few hours up to complete your job.

- At any time during data analysis, keep only one active web page open (except static web pages), as

Job Status

Job ID: 108

Bookmark Link: Create Job URL

Current Status: Finished

Priority: Level 1

Job Progress: 100%

Text Output:

```
3488 paired-reads (in 56 unique pairings) successfully merged out of 4078 (in 142 pairings) input.
5595 paired-reads (in 81 unique pairings) successfully merged out of 6496 (in 189 pairings) input.
16837 paired-reads (in 152 unique pairings) successfully merged out of 17774 (in 262 pairings) input.
5511 paired-reads (in 79 unique pairings) successfully merged out of 6043 (in 152 pairings) input.
3433 paired-reads (in 82 unique pairings) successfully merged out of 3876 (in 141 pairings) input.
OK, done!
Step 7: Perform Sequencing chimeras removal ...
Identified 29 bimeras out of 222 input sequences.
OK, done!
Step 8: Perform Sequencing taxonomy assignment ...
OK, done!
Everything has been finished Successfully !
```

Output File: Status Text 2023-03-09 01:04:39

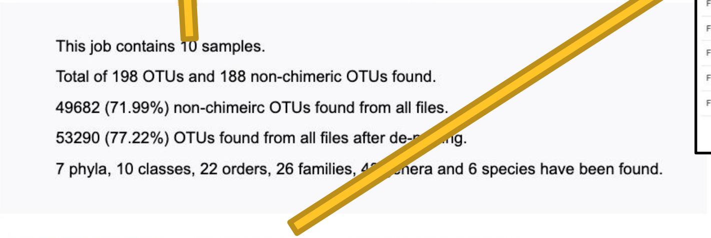
The job may take some time to complete, so click "Create Bookmark URL" to save the job link to check the job status at a later time.

Once the job is completed, click proceed.

Note: Keep only one active web page open. Multiple tabs/windows will interfere with each other, leading to unpredictable results

Result:

Summary of denoising and chimera removal results.



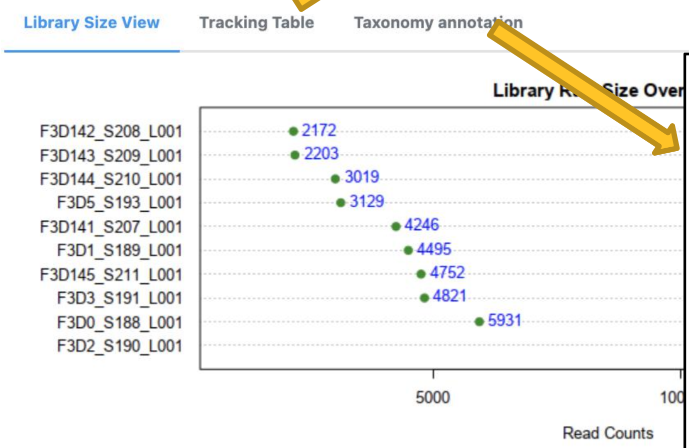
This job contains 10 samples.

Total of 198 OTUs and 188 non-chimeric OTUs found.

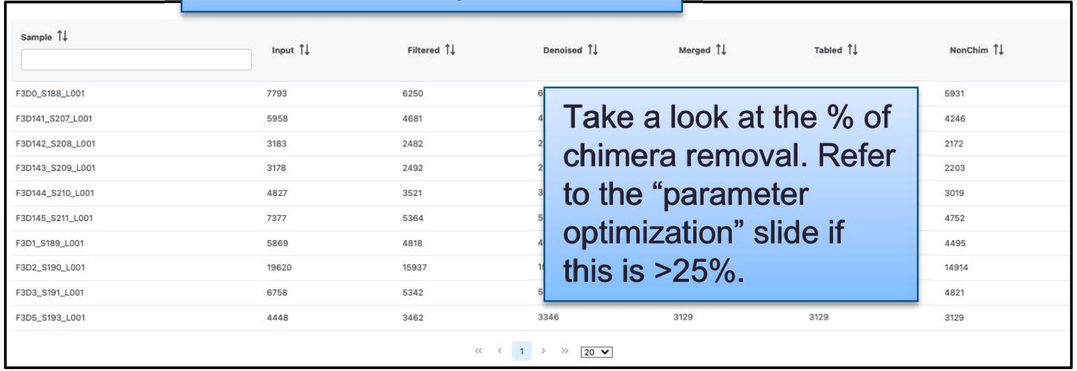
49682 (71.99%) non-chimeirc OTUs found from all files.

53290 (77.22%) OTUs found from all files after de-n[illegible]ing.

7 phyla, 10 classes, 22 orders, 26 families, 4[illegible]enera and 6 species have been found.



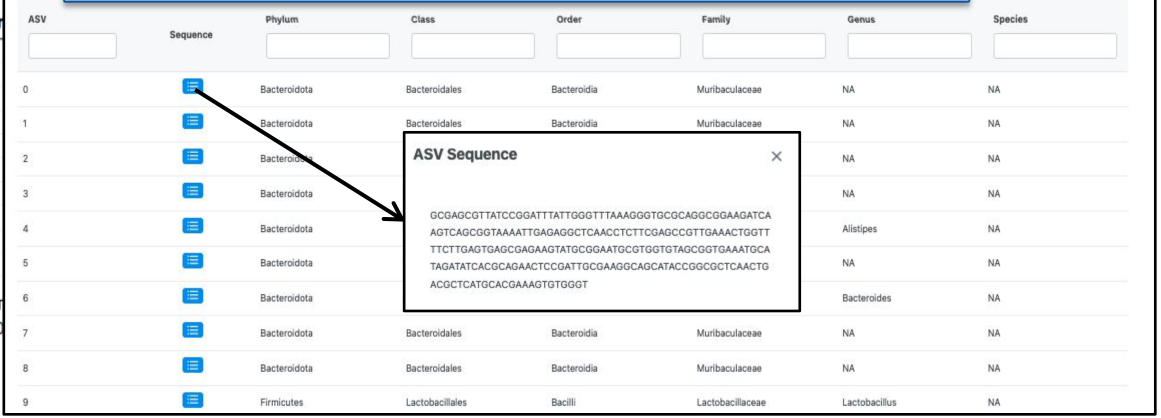
Track reads through the pipeline



Sample ⇅	Input ⇅	Filtered ⇅	Denoised ⇅	Merged ⇅	Tabled ⇅	NonChim ⇅
F3D0_S188_L001	7793	6250	6			5931
F3D141_S207_L001	5958	4681	4			4246
F3D142_S208_L001	3183	2482	2			2172
F3D143_S209_L001	3178	2492	2			2203
F3D144_S210_L001	4827	3521	3			3019
F3D145_S211_L001	7377	5364	5			4752
F3D1_S189_L001	5869	4818	4			4495
F3D2_S190_L001	19620	15937	1			14914
F3D3_S191_L001	6758	5342	5			4821
F3D5_S193_L001	4448	3462	3346	3129	3129	3129

Take a look at the % of chimera removal. Refer to the "parameter optimization" slide if this is >25%.

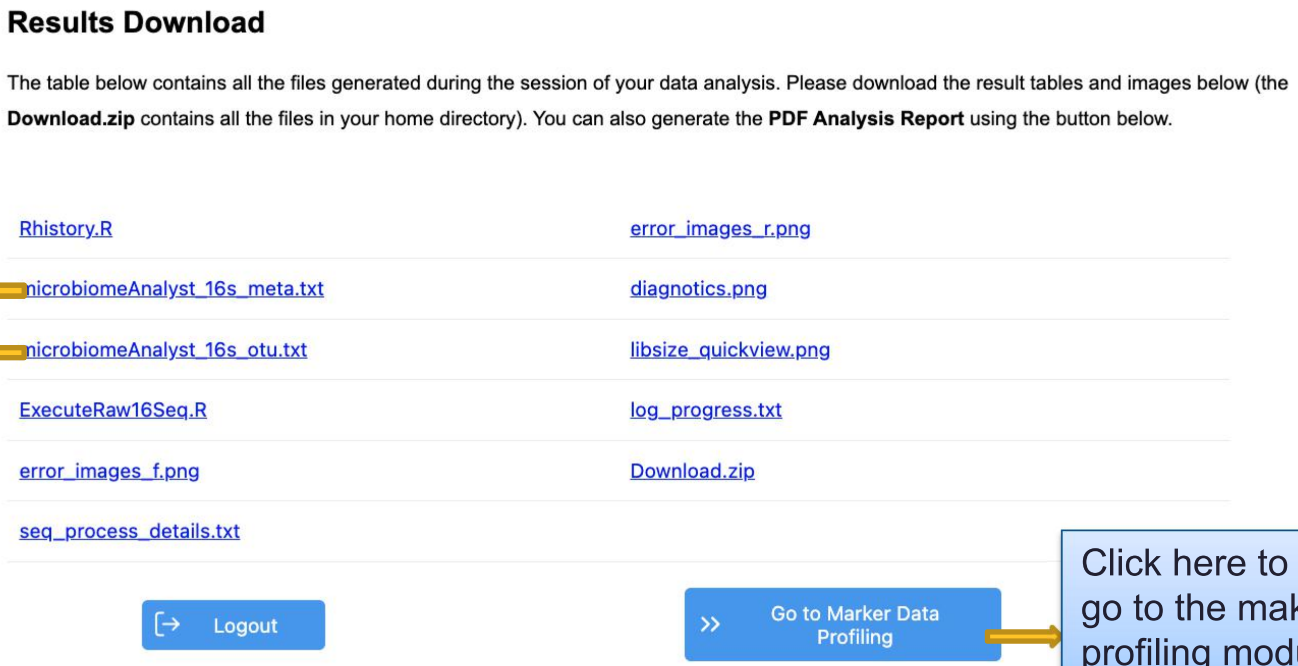
Check taxonomy annotation here. It is common to have lesser assignment at the Species level with 16S sequencing.



ASV	Sequence	Phylum	Class	Order	Family	Genus	Species
0	☰	Bacteroidota	Bacteroidales	Bacteroidia	Muribaculaceae	NA	NA
1	☰	Bacteroidota	Bacteroidales	Bacteroidia	Muribaculaceae	NA	NA
2	☰	Bacteroidota				NA	NA
3	☰	Bacteroidota				NA	NA
4	☰	Bacteroidota				Alistipes	NA
5	☰	Bacteroidota				NA	NA
6	☰	Bacteroidota				Bacteroides	NA
7	☰	Bacteroidota	Bacteroidales	Bacteroidia	Muribaculaceae	NA	NA
8	☰	Bacteroidota	Bacteroidales	Bacteroidia	Muribaculaceae	NA	NA
9	☰	Firmicutes	Lactobacillales	Bacilli	Lactobacillaceae	Lactobacillus	NA

ASV Sequence ×

```
GCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGCAGGCGGAAGATCA
AGTCAGCGGTAAAATTGAGAGGCTCAACCTCTTCGAGCCGTTGAAACTGGTT
TTCTTGAGTGAGCGAGAAGTATGCGGAATGCGTGGTGTAGCGGTGAAATGCA
TAGATATCACGCAGAACTCCGATTGCGAAGGCAGCATACCGGCGCTCAACTG
ACGCTCATGCACGAAAGTGTGGGT
```



Results Download

The table below contains all the files generated during the session of your data analysis. Please download the result tables and images below (the **Download.zip** contains all the files in your home directory). You can also generate the **PDF Analysis Report** using the button below.

- Rhistory.R
- microbiomeAnalyst_16s_meta.txt
- microbiomeAnalyst_16s_otu.txt
- ExecuteRaw16Seq.R
- error_images_f.png
- seq_process_details.txt
- error_images_r.png
- diagnotics.png
- libsize_quickview.png
- log_progress.txt
- Download.zip

Input files for MDP module of MicrobiomeAnalyst.

Click here to directly go to the maker data profiling module for downstream analysis

The End

For more information, visit Tutorials, Resources
and Contact pages on www.microbiomeanalyst.ca
Also visit our forum for FAQs on www.omicsforum.ca