User Manual of Amplicon Seuqnce Analysis Procedure for Galaxy-based pipeline in Denglab

http://mem.rcees.ac.cn:8080

Updated

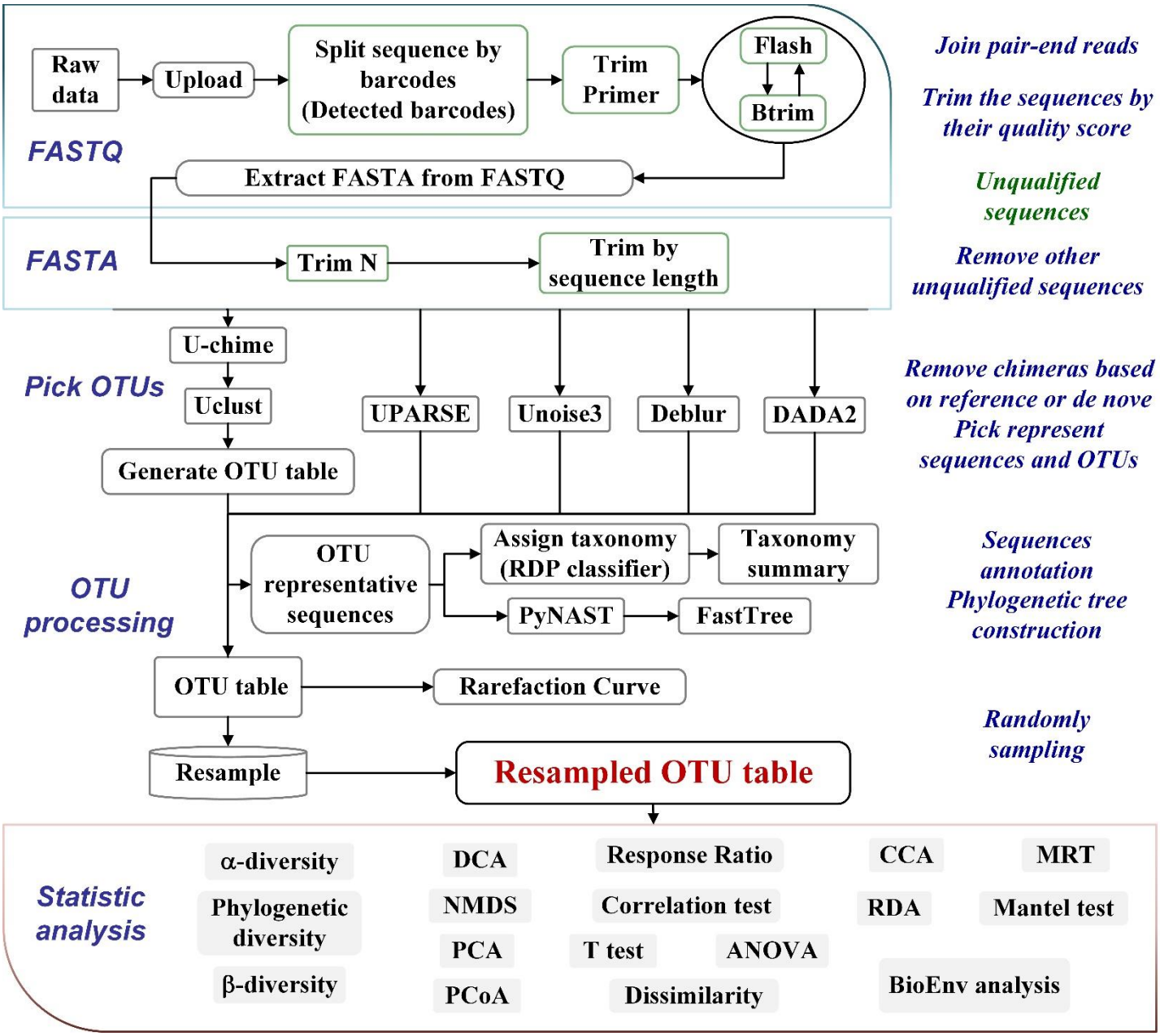
Nov. 2018

@ Denglab

Metagenomics for Environmental Microbiology (MEM)

Research Center for Eco-Environmental Sciences, CAS

Steps of amplicon sequencing data (16S/ITS/18S/Functional gene) preprocessing on Galaxy pipeline and basic statistical analysis procedures in Denglab



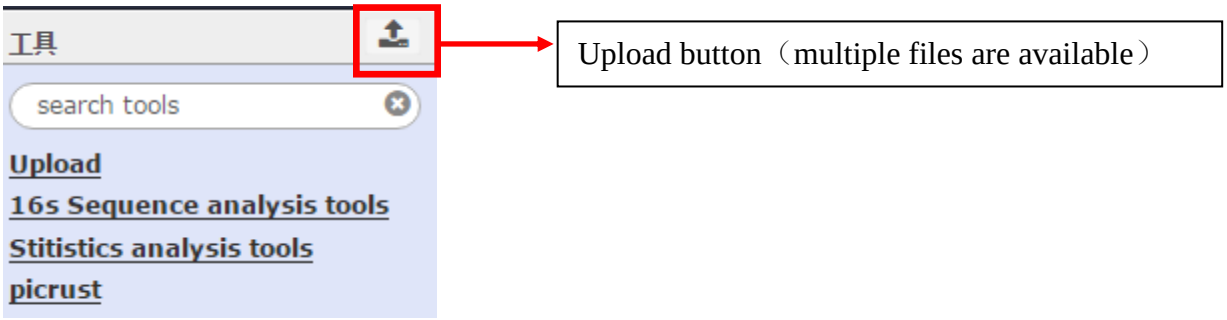
Content

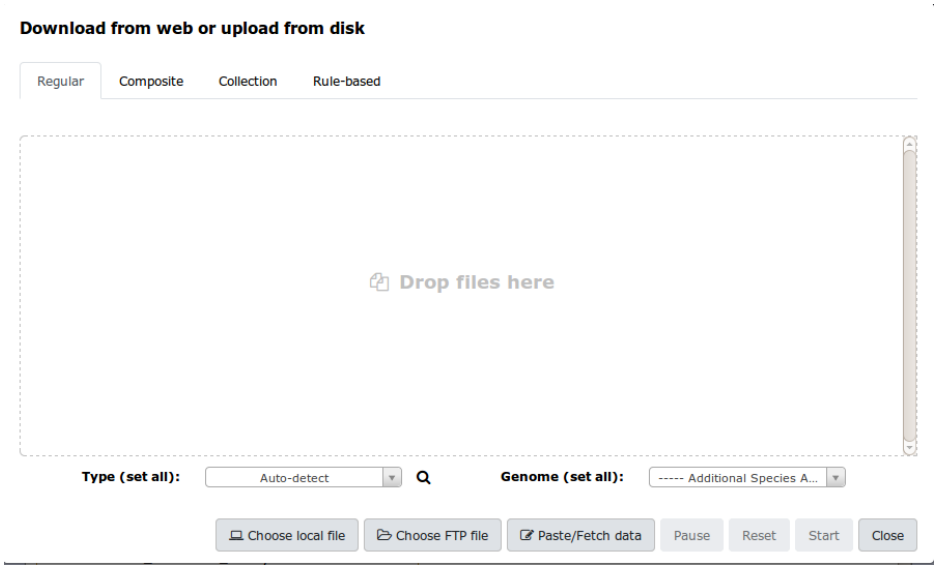
If you are interested in other popular analysis tools and want to make a contribution for our pipeline, please contact Prof. Ye Deng (yedeng@rcees.ac.cn).

A. Data Treatment and quanlity control

1. Upload File

Upload the three sequencing data idividually by Galaxy, barcodes file (txt) of your samples also need to be uploaded before downstreem analysis.





Required files:

You can find following test data from the "shared library/test data" directory and import these three files there.

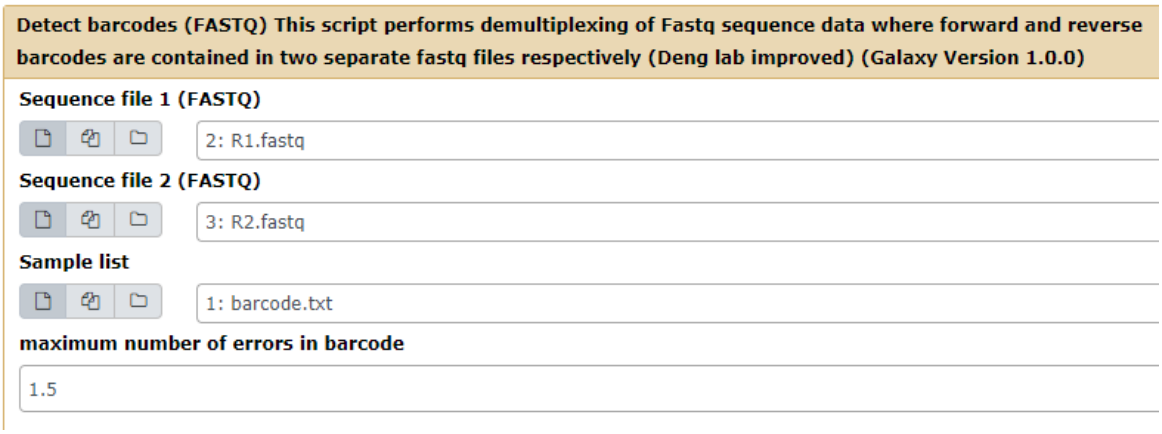
①sequencing data： R1.fastq R2.fastq Raw sequence data of R1 and R2.

②barcodes file (txt)：sample list barcode.txt （this file need to be finished by yourself）

Sample	forward_barcode	reverse_bacode	forward_primer_515F	reverse_primer_806R	Primers
A1	AGCCAGTCATAC	GTTGGTTGGCAT	GTGCCAGCMGCCGCGGTAA	GGACTACHVGGGTWTCTAAT	3
A2	AGCGAACCTGTT	TTCCACACGTGG	GTGCCAGCMGCCGCGGTAA	GGACTACHVGGGTWTCTAAT	4
A3	GTTTGCTCGAGA	AACCCAGATGAT	GTGCCAGCMGCCGCGGTAA	GGACTACHVGGGTWTCTAAT	5
A4	CAAACGCACTAA	GTAGTGTCAACA	GTGCCAGCMGCCGCGGTAA	GGACTACHVGGGTWTCTAAT	6
A5	GAACAAAGAGCG	TGGAGAGGAGAT	GTGCCAGCMGCCGCGGTAA	GGACTACHVGGGTWTCTAAT	7
A6	GCTAAGTGATGT	CGTATAAATGCG	GTGCCAGCMGCCGCGGTAA	GGACTACHVGGGTWTCTAAT	8
B1	AAGGGACAAGTG	AATACAGACCTG	GTGCCAGCMGCCGCGGTAA	GGACTACHVGGGTWTCTAAT	9
B2	AGTGTCGATTCG	GACTCAACCAGT	GTGCCAGCMGCCGCGGTAA	GGACTACHVGGGTWTCTAAT	10
B3	CTATTAAGCGGC	GGAAGAAGTAGC	GTGCCAGCMGCCGCGGTAA	GGACTACHVGGGTWTCTAAT	11
B4	GAGTCCGTTGCT	ACACCGCACAAT	GTGCCAGCMGCCGCGGTAA	GGACTACHVGGGTWTCTAAT	13
B5	GATAACTGTACG	GTCTCCTCCCTT	GTGCCAGCMGCCGCGGTAA	GGACTACHVGGGTWTCTAAT	14

2. Detect barcodes(FASTQ)

Inputs:



Options

Maximum number of errors in barcode: >= 0. 1.5 means allowing one mismatch.

Outputs:

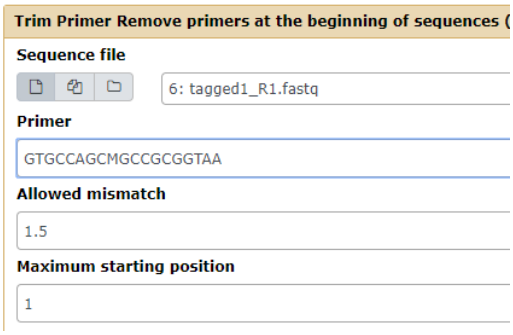
The barcode sequences were trimed in tagged1_xxx.fastq and tagged2_xxx.fastq files. "tagged1_xxx.fastq" is forward direction and "tagged2_xxx.fastq" is reverse direction.

3. Trim Primer(FASTQ)

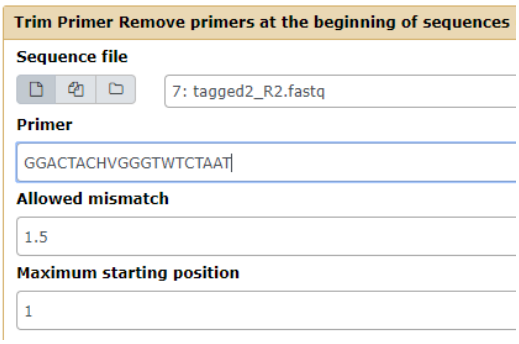
The forward and reverse primer sequences should be trimmed in the raw sequencing data separately.

Inputs:

Forward



Reverse



Parameter options:

Primer used:

515F GTGCCAGCMGCCGCGGTAA (R1)

806R GGACTACHVGGGTWTCTAAT (R2)

Modified 515F/806R (Walters W. et al. mSystem. 2015. 1(1):e00009-15)

515MF GTGYCAGCMGCCGCGGTAA (R1)

806MR GGACTACNVGGGTWTCTAAT (R2)

gITS7F GTGARTCATCGARTCTTTG (R1)

ITS4R：TCCTCCGCTTATTGATATGC (R2)

Allowed mismatch: 1.5

Maximum starting position: 1 (according to your primer position in the raw sequences, usually primers start at 1 position)

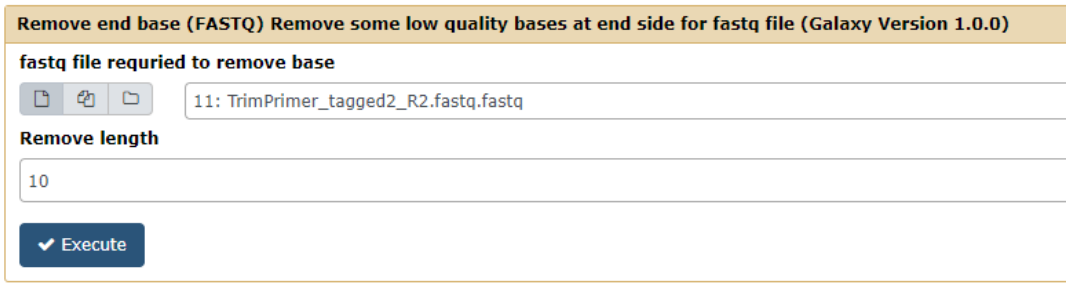
Outputs:

Forward	Reverse
10: TrimPrimer_summary.txt	12: TrimPrimer_summary.txt
9: TrimPrimer_tagged1_R1.fastq.fastq	11: TrimPrimer_tagged2_R2.fastq.fastq

4. Remove end base (FASTQ) (Optional)

Remove the certain bases for each sequences. It is useful to discard bases with low quality scores.

Inputs:



Outputs:

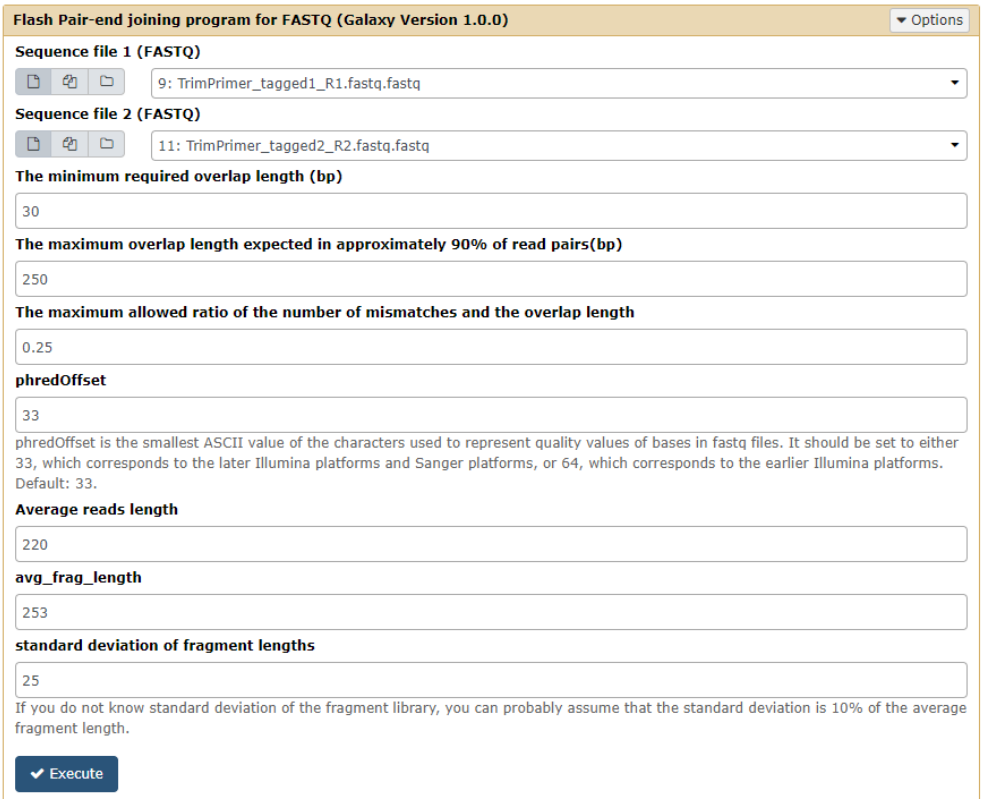
- Remove_TrimPrimer_tagged2_R2.fastq.fastq

5. Flash (Combine R1 and R2) (FASTQ)

Inputs:

Forward sequences for sequence file 1

Reverse sequences for sequence file 2



Parameter options:

- **The maximum overlap length is usually "250", and the other parameters can be changed when necessary.**

The minimum required overlap length: 30 (sequence length for enough overlap)

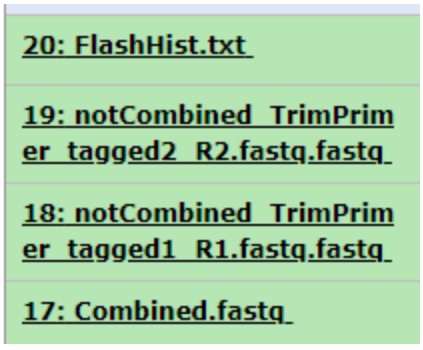
The maximum overlap length expected in approaximately 90% of read pairs (bp): 250 (250 = 220*2 – 253 + 25*2.5)

Average reads length: 220 (based on the length_statistics for sequences after trim primer)

Avg_frag_length: 253 (515F-806R); 265 (gITS7F-ITS4R);

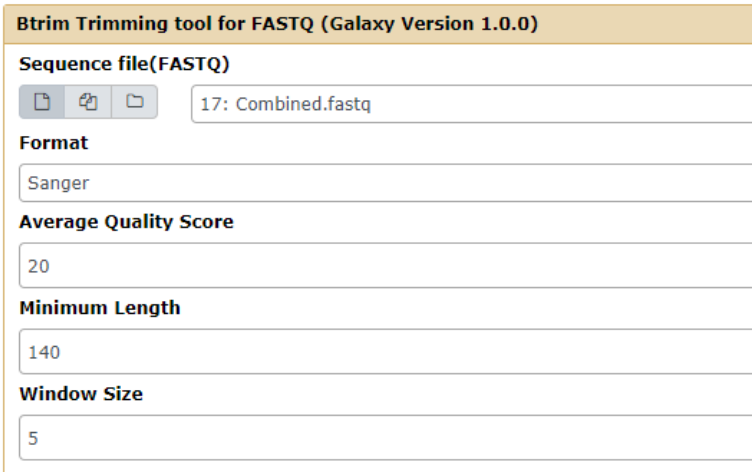
Standard deviation of fragment length: 25 = avg_frag_length * 10%

Outputs:



6. Btrim (FASTQ)

Inputs:



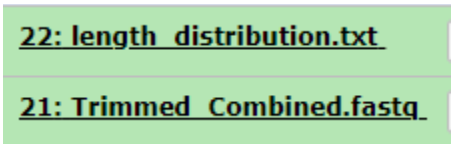
Parameter options:

Average Quality Score: 20

Minimum length: 140 (Determined by FastQC results to select maximum sequence length of Q30 or Q20)

Window Size: 5 (every 5 bases was treated as a windon size to check the quality. If the average score of these 5 bases is < 20, but the sequence length is > 140, then the bases after this 5 bases will be removed and left bases will remain. Otherwise, the whole sequence will be discard.)

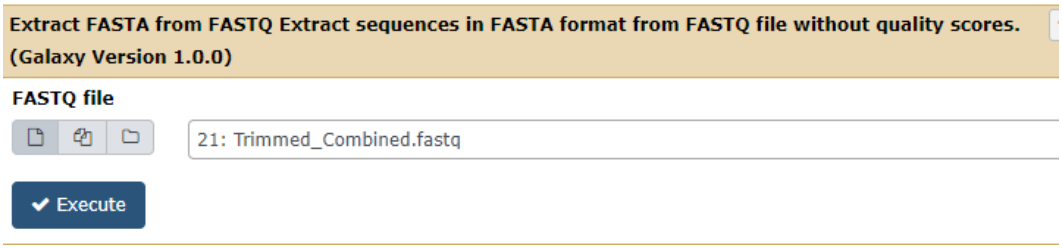
Outputs:



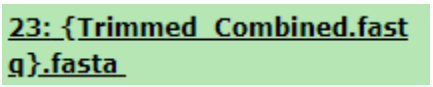
7. Extract FASTA from FASTQ

Transforming FASTQ file (Trimmed_combined.fastq) to FASTA file

Inputs:



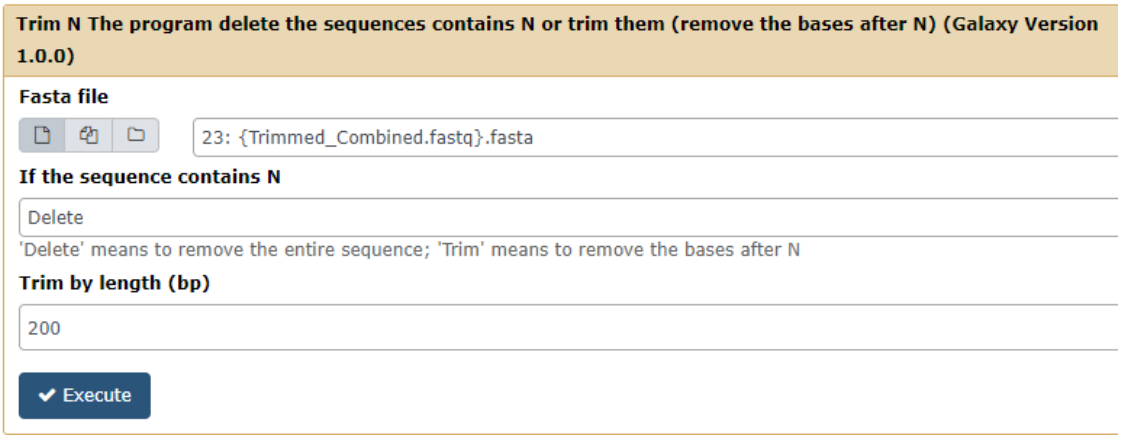
Outputs:



8. Trim N (FASTA) (Trimming sequences containing "N")

Trimming sequences containing "N"

Inputs:



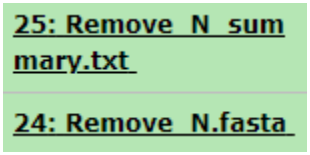
Parameter options:

Delete: remove the entire sequence once it contained 'N' (more strict)

Trim: remove the bases in a sequence after 'N'

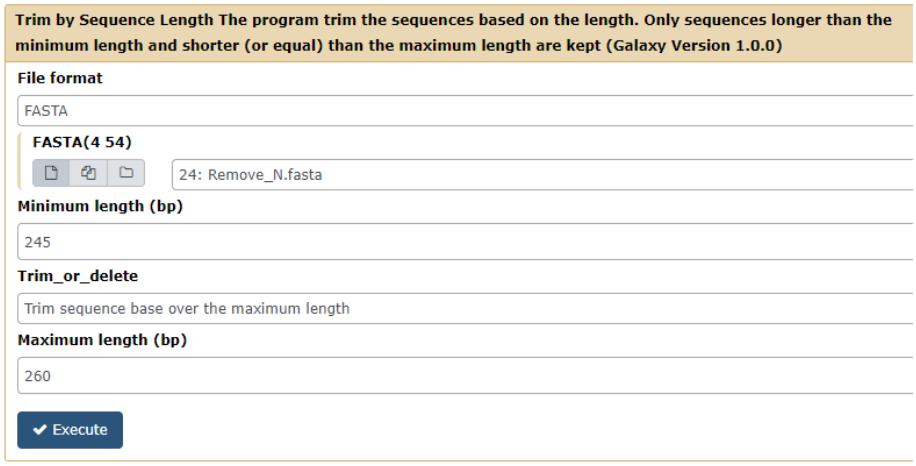
Trim by length (bp): 200 (default; you can change it according to your case)

Outputs:



9. Trim by Sequence Length

Inputs:



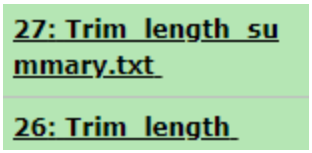
Parameter options:

File format: FASTA / FASTQ (select according to your case)

Trim or delete: Trim sequence base over the maximum length (strictly, you can choose delete)

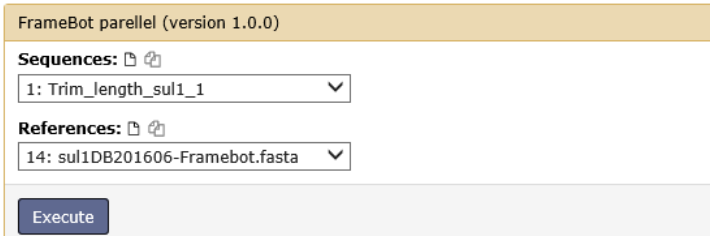
Length range: For 16S (515F/806R) primers, generally 245~260 bp; For ITS (gITS7F~ ITS4R), generally varies around 265 bp; the length range can be changed according to your specific case.

Outputs:



10. Framebot (optional; only necessary for functional genes)

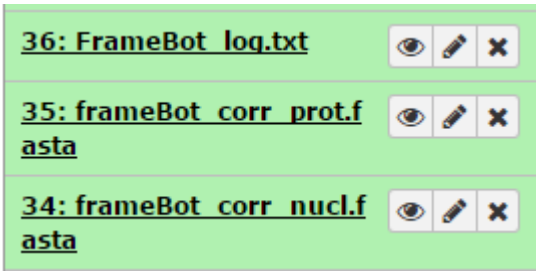
Inputs:



Parameter options:

References: unaligned_protein.fasta (own database without alignment)

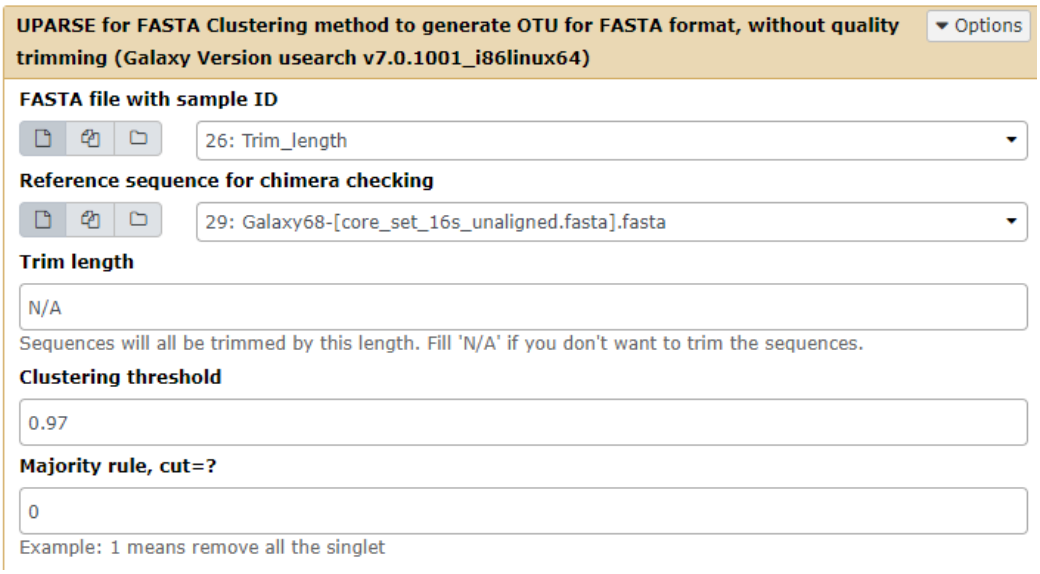
Outputs:



11. Generate OTU table

UPARSE (Recommend; UPARSE for FASTA)

Inputs:



Parameter options:

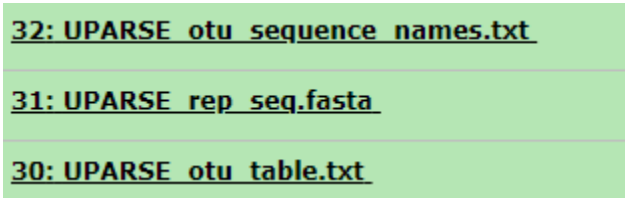
Reference sequence for chimera checking: find the database in shared library or upload your own database.

Trim length: N/A

Clustering threshold: 97% (similarity of OTU clustering; usually 0.97)

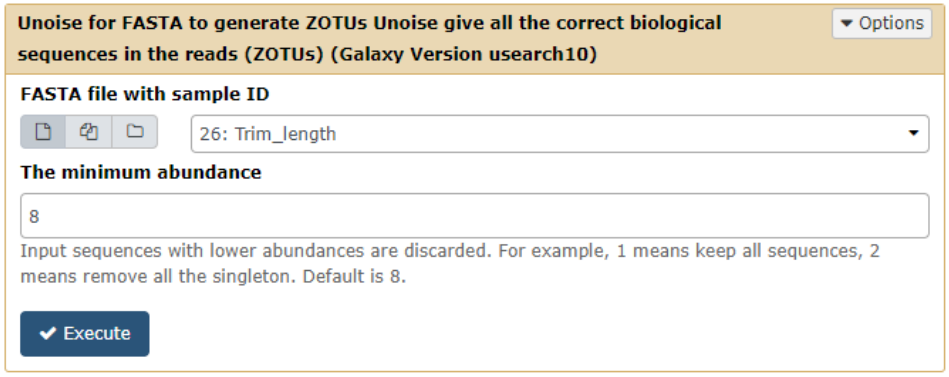
Majority rule, cut=?: 0 (keep all singletons)

Outputs (3 files):



UNOISE (Unoise for FASTA to generate ZOTUs)

Inputs:



Parameter options:

The minimum abundance: 8 (default value)

Outputs:

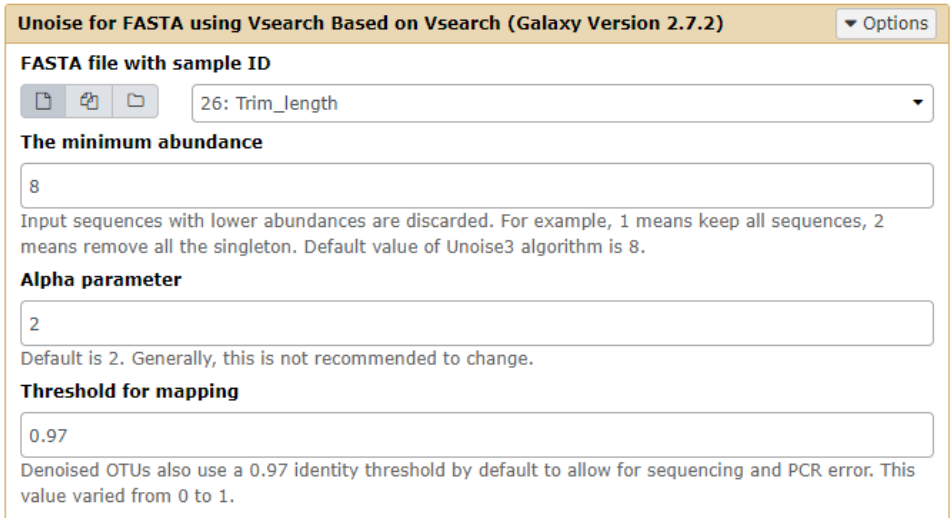
UNOISE (Unoise for FASTA using Vsearch)

This program is mainly for large dataset for OTU clustering, such as greater than 4GB. Similar options as above instructions.

The minimum abundance: 8

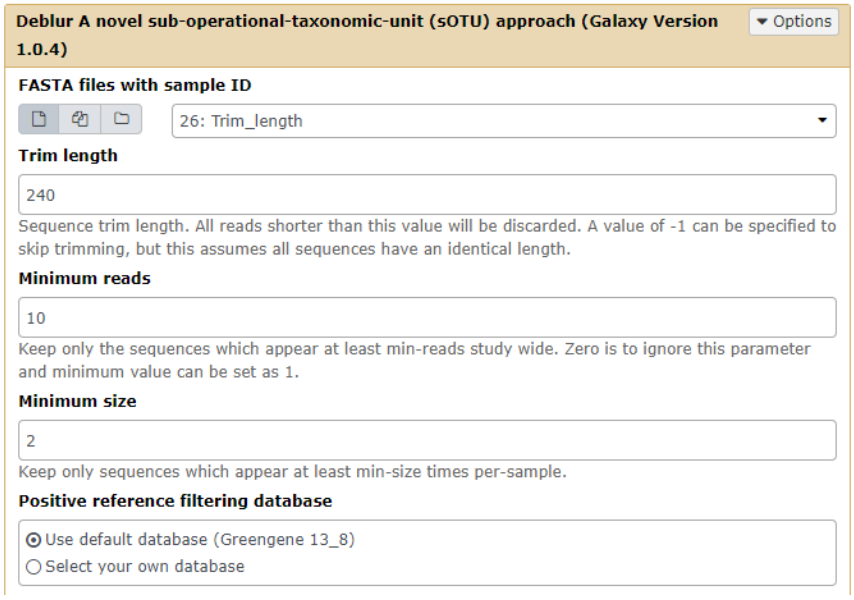
Alpha parameter: 2 (not recommended to change)

Threshold for mapping: 0.97 (default value)



Deblur

Inputs:



Parameter options:

Trim length: 240 (change based on the intructions)

Minimum reads: 10 (default; it depends)

Minimum size: 2 (default; it depends)

Positive reference filtering database: default Greengene 13.8 version

Outputs:

38: Deblur reference-non-hit rep seq

37: Deblur reference-non-hit

36: Deblur sOTU rep seq (reference-hit)

35: Deblur sOTU table (reference-hit)

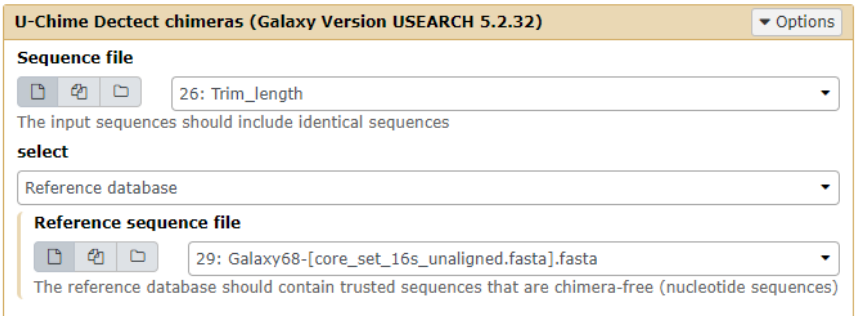
34: Deblur all seq (include hit and non-hit)

33: Deblur all table (include hit and non-hit)

Uclust

11.5-1 Uchime

Inputs:



Parameter options:

Select: de novo / Reference database (in shared library or upload your own database)

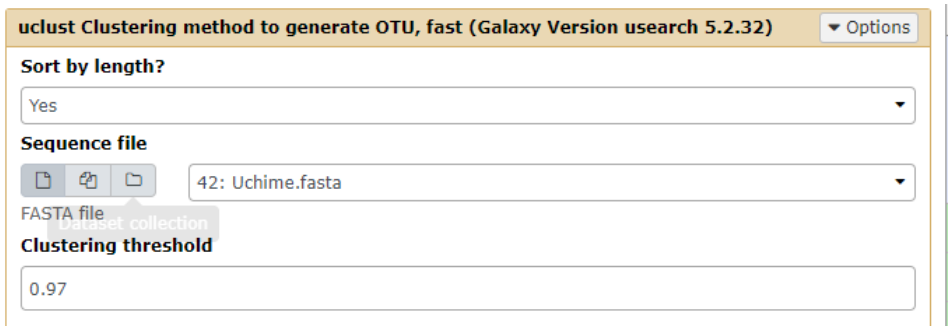
Outputs:

43: redundancy_map.txt

42: Uchime.fasta

11.5-2 Uclust

Inputs:



Parameter options:

Clustering threshold: 0.97

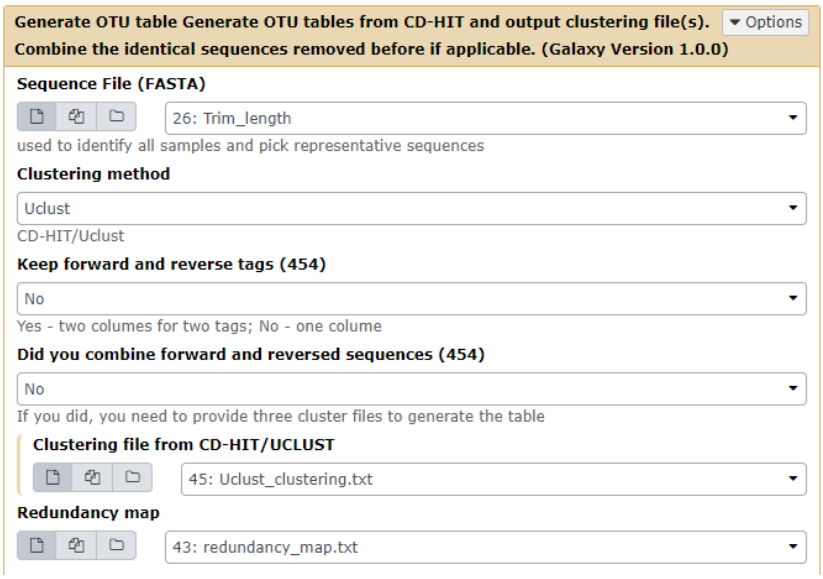
Outputs:

45: Uclust_clustering.txt

44: Uclust_seeds.fasta

11.5-3 Generate OTU Table

Inputs:



Parameter options:

Clustering method: Uclust

Other parameters keep default

Outputs:

49: rep_seq_no_singlet.fasta

48: rep_seq.fasta

47: OTU_table_without_singlet.txt

46: OTU_table.txt

6,127 lines

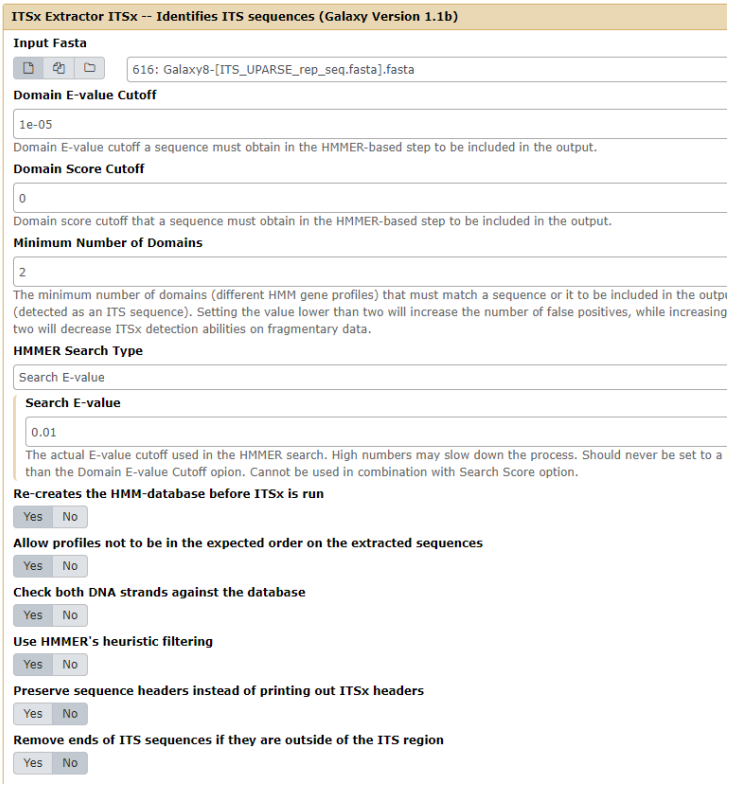
格式: **txt**, 数据库: ?

OTU	A1	A2	A3	A4	A5
OTU_0	3104	7489	75	15	3490
OTU_1	1	7	0	1	1
OTU_10	105	4	102	113	2
OTU_100	0	0	0	0	0

12. ITSx Extractor (Optional for ITS)

This tool is to identify ITS sequences and extracts the ITS regions..

Input:



Options:

Usually not requried to change.

Output:

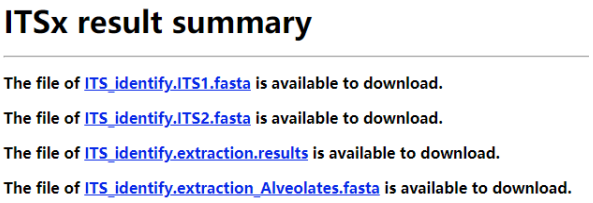
- ITSx Summary

```
ITSx run started at Sat Apr  7 11:58:12 2018.
--------------------------------------------------
Number of sequences in input file:            2661
Sequences detected as ITS by ITSx:      2320
  On main strand:               2320
  On complementary strand:      0
ITS sequences by preliminary origin:
  Alveolates:           60
  Amoebozoa:            50
  Bacillariophyta:      39
  Brown algae:          0
  Bryophytes:           0
  Euglenozoa:           0
  Eustigmatophytes:     0
  Fungi:                1562
  Green algae:          439
  Liverworts:           0
  Metazoa:              69
  Microsporidia:        0
  Oomycetes:            0
  Prymnesiophytes:      0
  Raphidophytes:        0
  Red algae:            4
  Rhizaria:             17
  Synurophyceae:        1
  Tracheophyta:         79
--------------------------------------------------
ITSx run finished at Sat Apr  7 12:00:56 2018.
```

- Identified Fungi ITS fasta File

```
>OTU_1
AACGCACATTGCGCCCCTTGGTATTCCAGGgggCATGCCTGTTTGAGCGTCATTTCCTTCTCAAACATTCT
TGTTtttttttCCAAAGagagGTTTCTCTGCGTGCTTGAGGTATAATGCAAGTACGGTCGTTTTAGGTTTT
agCGTCTAGGCGAACAATGTTCTTAAAGTTTGACCTCAAATCAGGTAGGAGTACCCGCTGAACTTAA
>OTU_2
AACGCACCTTGCGCTctctGGTATTCCGGAGAGCATGCCTGTTTGAGTATCATGAAATCTCAACCATTAGG
GCGTGTTTAACTTGTCATATCTGGCGTAATAAGTTTCGCTGGTGTACGACTTGAGAAGTGCGCTTCTAATC
A
>OTU_3
AACGCAACTTGCGCTctctGGTATTCCGGAGAGCATGCCTGTTTGAGTGTCATGAAATCTCAACCATTAGG
AGCGAGTTTAACTATTGCTATCTGGCGTAATAAGTTTCGCTGGAATGGTATTGTGAAGCgcgcTTCTAATC
```

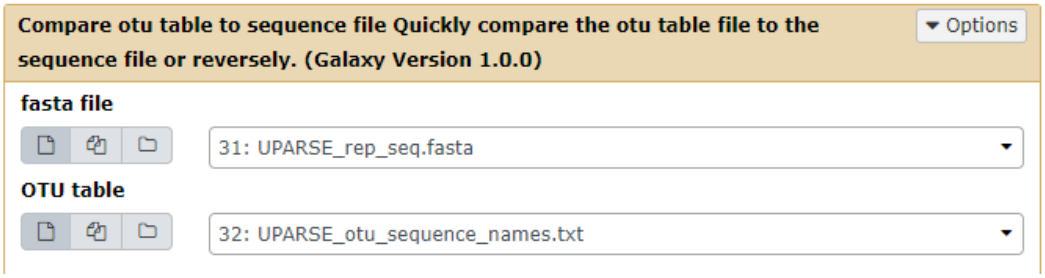
- ITSx results associated with other eukaryotes



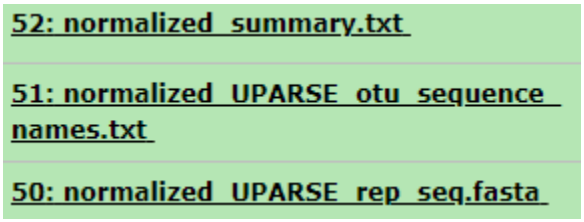
13. Compare otu table to sequence file (optional)

Useful when the OTU numbers for rep_seq and table file were inconsistent. Compare the otu table file to the sequence file. Only keep the matched sequence or otu name left.

Inputs:



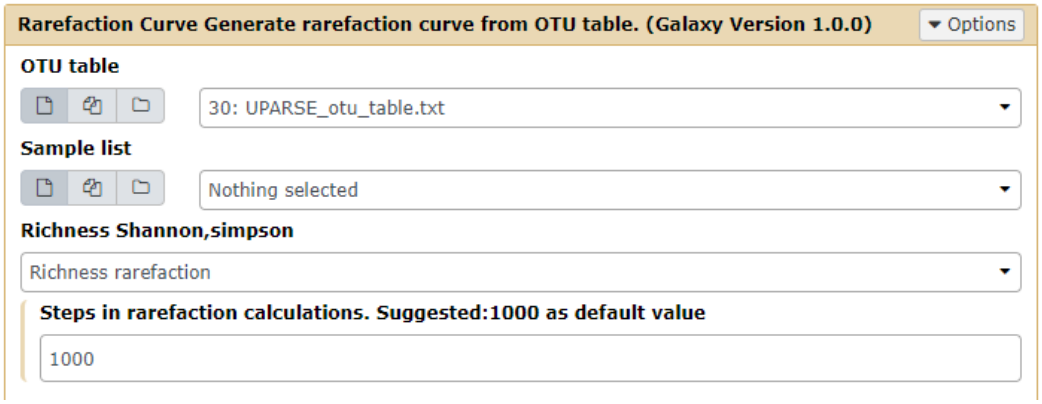
Outputs:



14. Rarefaction Curve

① Richness rarefaction

Inputs:



Parameter options:

Sample list (optional):

- Sample list (optional):

```
Sample1_name:tag1,tag2,tag3
Sample2_name:tag4,tag5,tag6
......
```

```
A:A1,A2,A3,A4,A5,A6
B:B1,B2,B3,B4,B5,B6
C:C1,C2,C3,C4,C5,C6
```

Richness rarefaction: 1000 steps

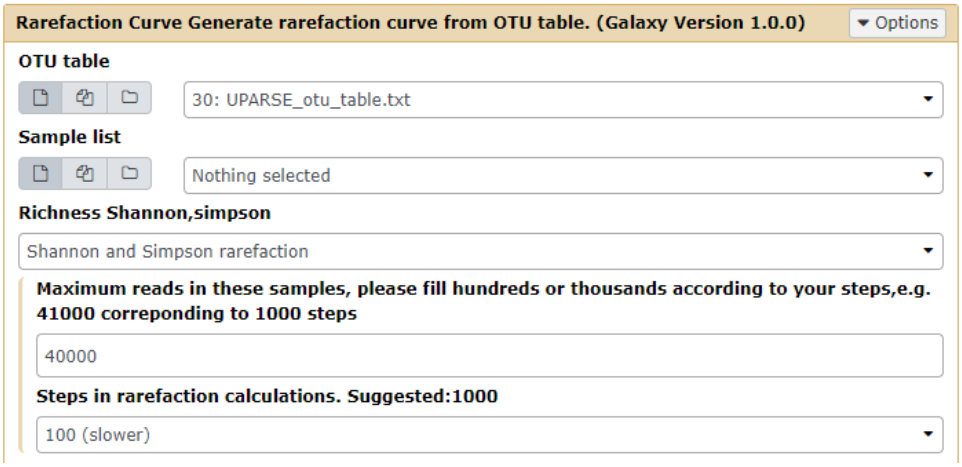
Outputs:

54: chao.txt

53: rarefaction.txt

② Shannon and Simpson Rarefaction

Inputs:



Parameter options:

Maximum reads: 40000 (depends on maximum reads of all samples in your OTU table; must be changed to multiples for selected steps)

Outputs:

58: simpson_diversity

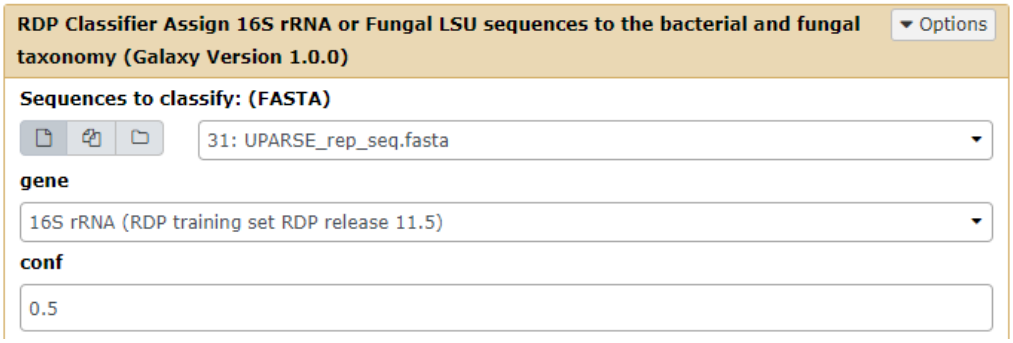
57: shannon_diversity

56: rarefaction_simpson.txt

55: rarefaction_shannon.txt

15. RDP Classifier

Inputs:



Parameter options:

Conf: 0.5 (recommended value)

Gene:

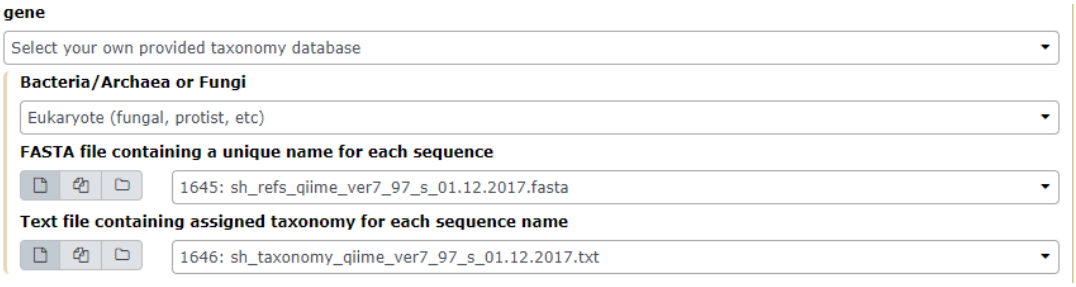
For 16S, RDP, Greengene, SILVA

For fungal ITS, warcup, Unite

For fungal LSU, RDP

For 18S, SILVA

If you want to use your own database to assign taxonomy, it is available now:



Notice when you choose own taxonomy database

You should provide two files, fasta file and text file, to assign sequences based on your taxonomy database using RDP classifier. If you want to add an option for your own trained database in this pipeline, please contact Prof. Ye Deng.

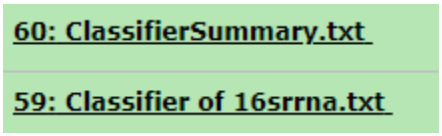
- The uploaded FASTA file must have a stable or unique name for each sequence

```
- >SH009881.07FU_EF634088_reps
- CCGAACTGTCGACACGAGTTGTTGCTGGCCTCTCAAACGGGGGGCATGTGCACA
- >SH492954.07FU_DQ656654_reps_singleton
- CATGAGCCTTGATCTCGCCGGTTAAGCAGAGGCCTCACGGTCCGCGGGTAATCT
- >SH628622.07FU_LC131409_reps
- TCGAAGAAGCACACTTCTTCTCCAACCCCTGTGAACCGTCGTCCGAGCATGATGCTCGGACGCTCCATACATACT
- ...
```

- The uploaded txt file must be tab-devided text file and its format should be as same as below, containing two columns, sequence name and taxonomy information. Please notice k is for kingdom, p is for phylum, c is for class, o is for order, f is for family, g is for genus and s is for species.

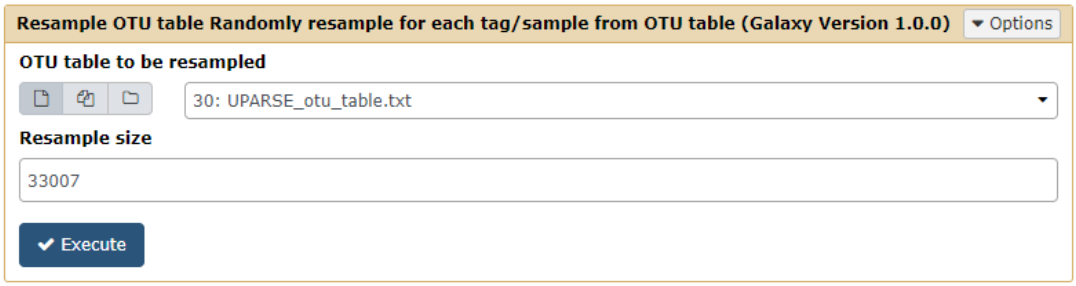
```
- SH009881.07FU_EF634088_reps        k__Fungi;p__Basidiomycota;c__Agaricomycetes;o__Thelephorales;f__Thelephoraceae;g__Tomentell
- SH492954.07FU_DQ656654_reps_singleton      k__Fungi;p__Ascomycota;c__Orbiliomycetes;o__Orbiliales;f__Orbiliaceae;g__Hyalorbili
- SH628622.07FU_LC131409_reps        k__Fungi;p__Basidiomycota;c__Dacrymycetes;o__Dacrymycetales;f__Dacrymycetaceae;g__Calocera;
```

Outputs:



16. Resample OTU table

Inputs:



Parameter options:

Resample size: change according to your own OTU table. Usually it is the minimum value of reads numbers for all samples in the OTU table with singletons. For test data, the resample size is 33007.

```
   C3    A4    C5    C1    B2    A5    B3    A3    B5    B1    A2    A1    C6    B4    C2
52348 44169 45372 56772 33065 40556 35288 64117 33007 51750 38349 41994 56878 40327 44666
   B6    C4    A6
44579 39361 44393
```

For soil samples, 16S analysis > 30000; ITS analysis > 10000

Outputs:

61: resample_UPARSE_otu_table.txt

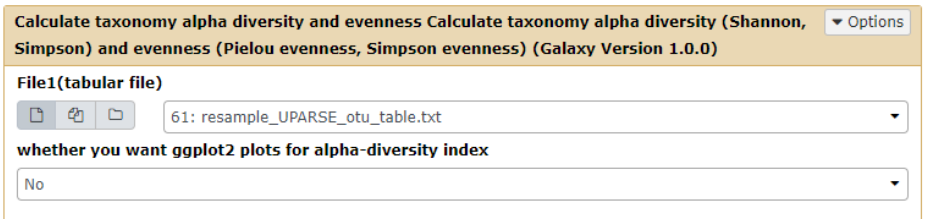
B. Statistics analysis

All statistics analysis methods were based on OTU table after resample process (resampled OTU table).

1. Diversity methods

1.1 α-diversity (Calculate taxonomy alpha diversity and evenness)

Inputs:

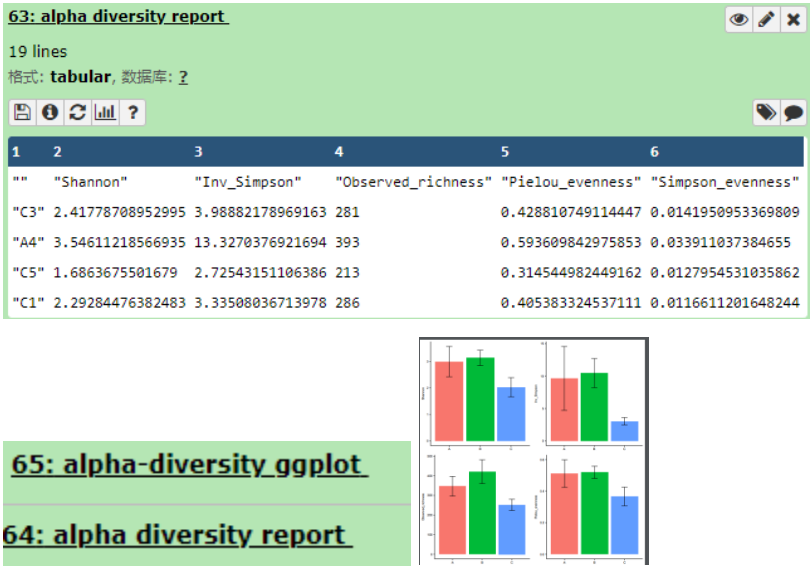


Parameter options:

Plot figure using ggplot2: No / Yes (provide a group file, see the example below)

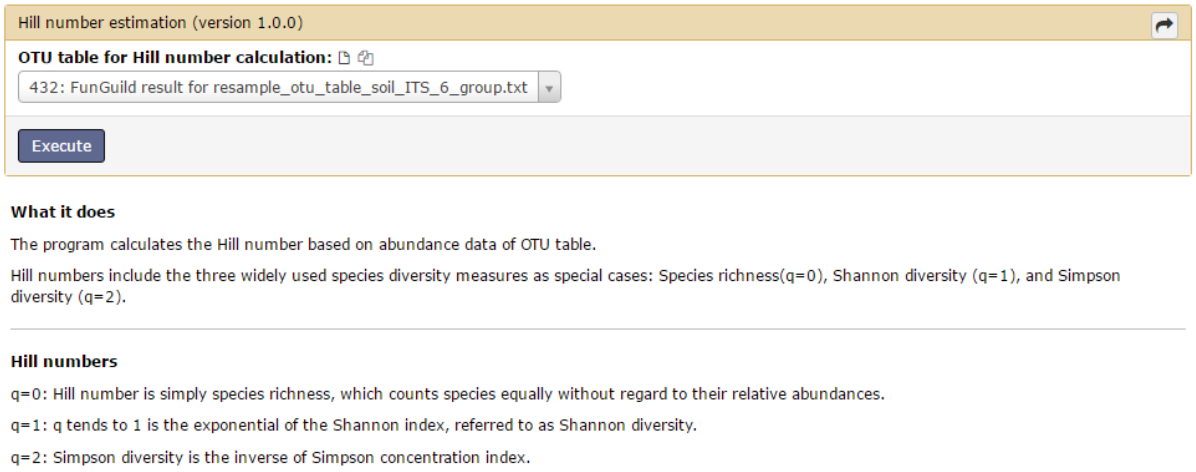
```
A1	A
A2	A
A3	A
A4	A
A5	A
A6	A
B1	B
B2	B
B3	B
B4	B
B5	B
B6	B
```

Outputs:



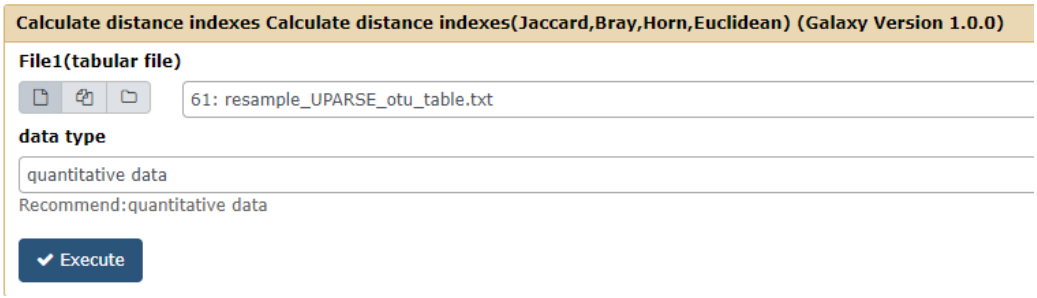
1.2 Hill number

或者根据 Chao 的最新文章计算 Hill number:



1.3 β-diversity

Inputs:

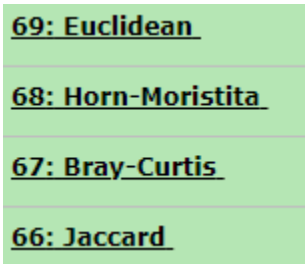


Parameter options:

Quantitative type is corresponding to the abundance type of the input file. The distance calculation is based on abundance data. (Recommend)

Absence/Presence type is corresponding to the 1/0 data type. The input file will be standardized to 0/1 scale.

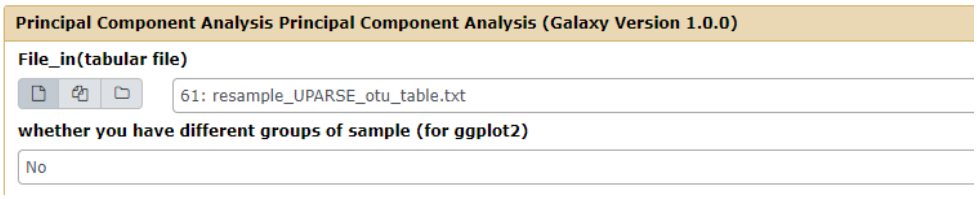
Outputs:



2. Community structure

2.1 PCA

Inputs:



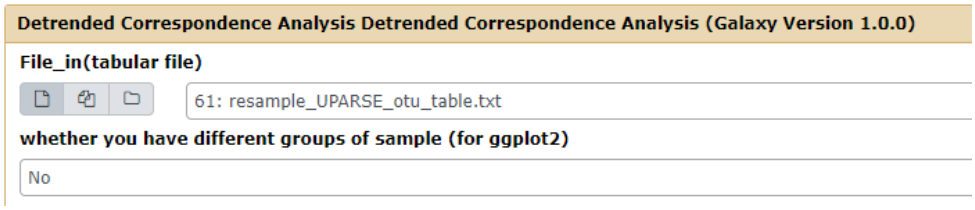
Outputs:

71: pca plot

70: pca result

2.2 DCA

Inputs:



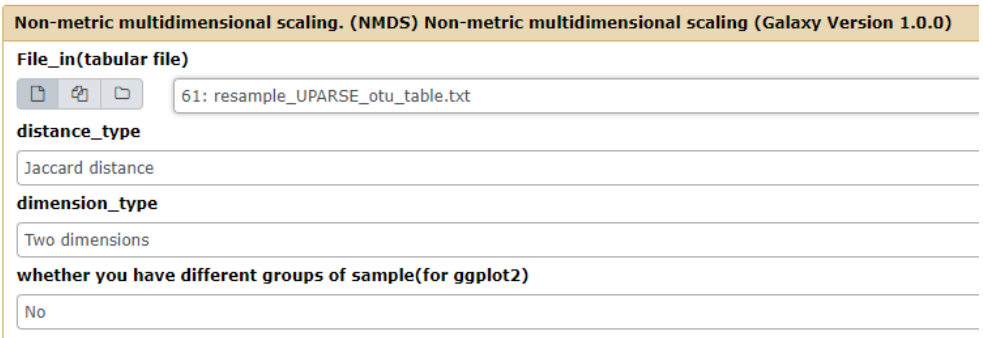
Outputs:

73: dca plot

72: dca result

2.3 NMDS

Inputs:



Parameter options:

Distance type: jaccard / bray

Dimentison type: 2D / 3D

Plot using ggplot2: provide a group file if you have

Outputs:

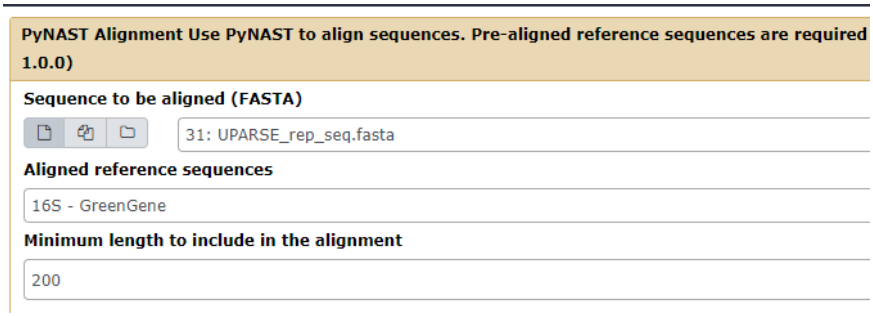
75: NMDS plot

74: NMDS result

2.4 PD&PCoA

1) PyNAST alignment

Inputs:



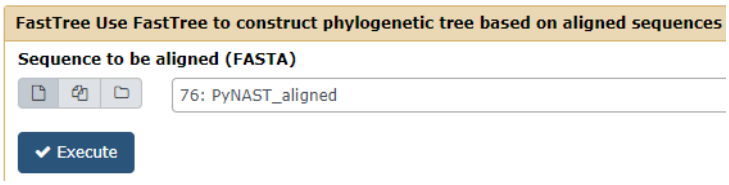
Parameter options:

Aligned reference sequences: 16S-GreenGene / own aligned database (from shared library or upload)

Outputs:

2) FastTree

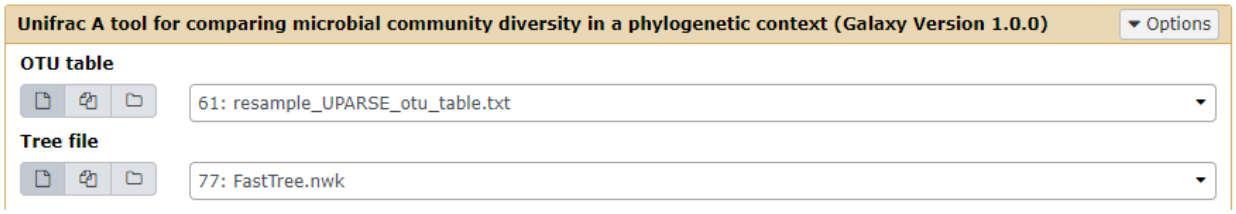
Inputs:



Outputs:

3) UniFrac

Inputs:

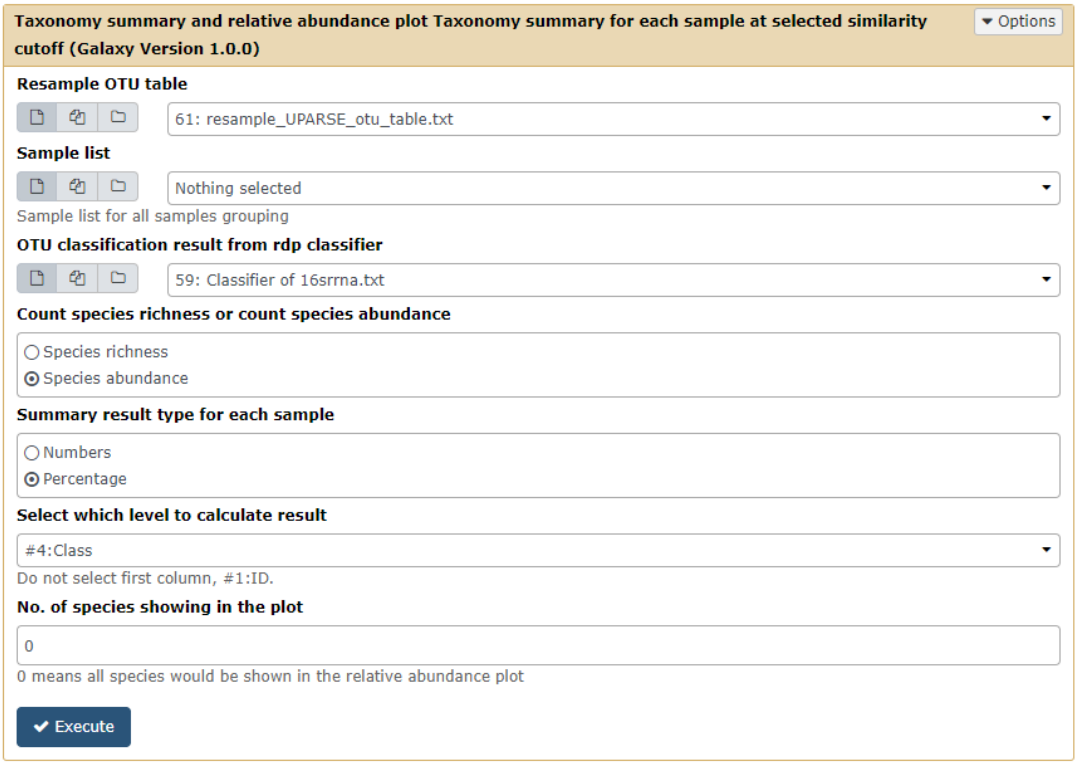


Outputs:

Weighted PcoA and unweighted PcoA are the results of PCOA.

2.5 Relative abundance

Inputs:



Parameter options:

Count species richness or count species abundance: richness / abundance

Result type for each sample: numbers / percentage

Taxonomy level：Domain, Phylum, Class, Order, Family, Genus, Species (this will shown if you put OTU classification file in)

No. of species showing in the plot: 0 is to show all species.

Outputs:

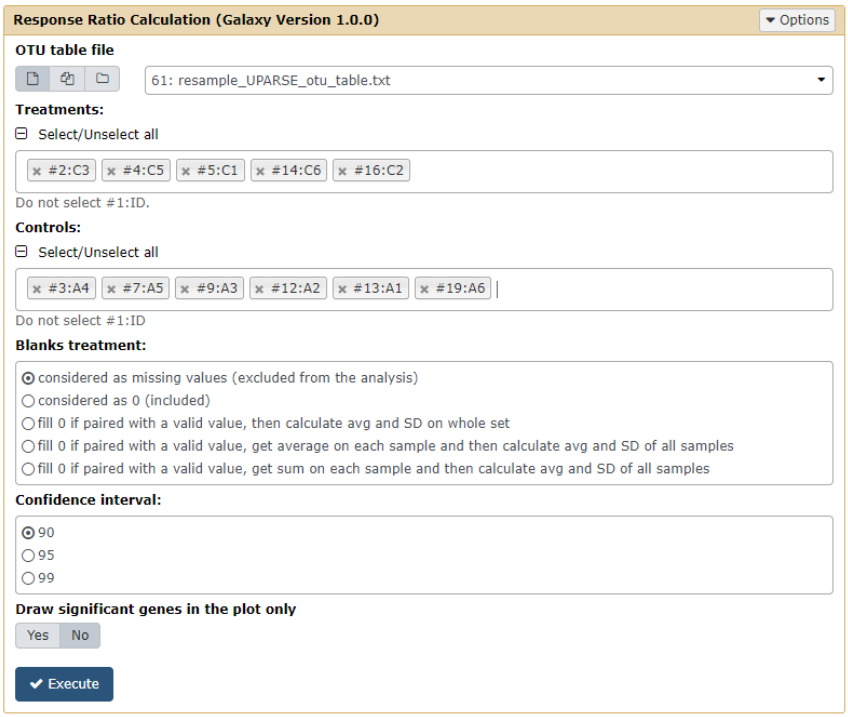
83: Relative abundance at 3 level for abundance count

82: Taxonomy summary at 3 level for abundance count

3. Comparison analysis

3.1 Response ratio calculation

Inputs:



Parameter options:

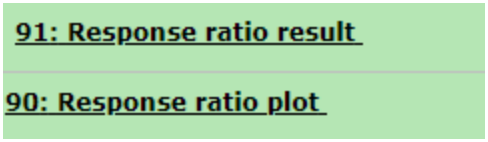
Treatments / Controls: select specific column in the OTU table file

Confidence interval: 90 / 95 / 99

Blanks treatment: considered as missing values

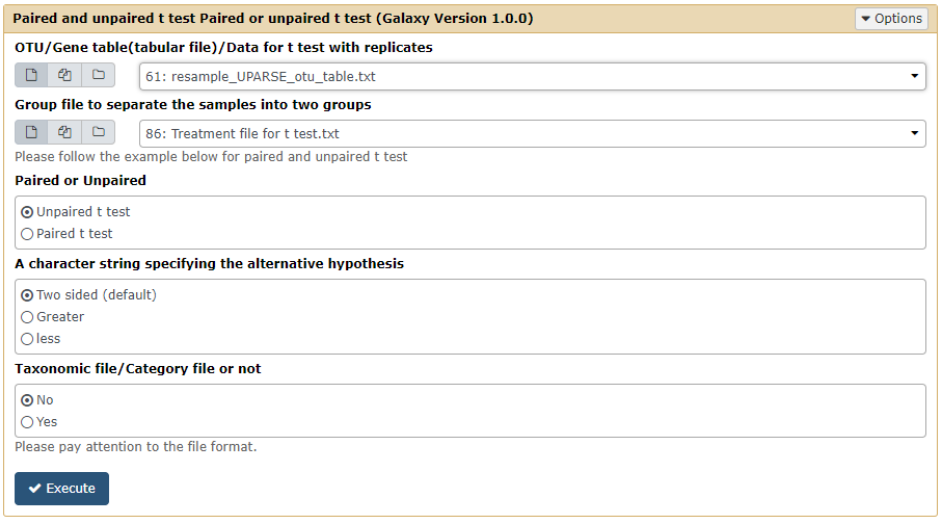
Draw significant genes in the plot only: No (depends)

Outputs:



3.2 Paired and unpaired t test

Inputs:



Parameter options:

Group files for unpaired t test and paired t test:

Samples	Group	Pairs	Group1	Group2
Sample1	group1	Pair1	sample1	sample9
Sample2	group1	Pair2	sample2	sample10
Sample3	group1	Pair3	sample3	sample11
Sample4	group1	Pair4	sample4	sample12
Sample5	group2	Pair5	sample5	sample13
Sample6	group2	Pair6	sample6	sample14
Sample7	group2	Pair7	sample7	sample15
Sample8	group2	Pair8	sample8	sample16

Paired or unpaired: paired t test / unpaired t test

Taxonomic file/category file: No (default) / Yes (provide another taxonomy file and select certain column)

Outputs:

87: Paired or Unpaired t test result

Paired or unpaired t test report

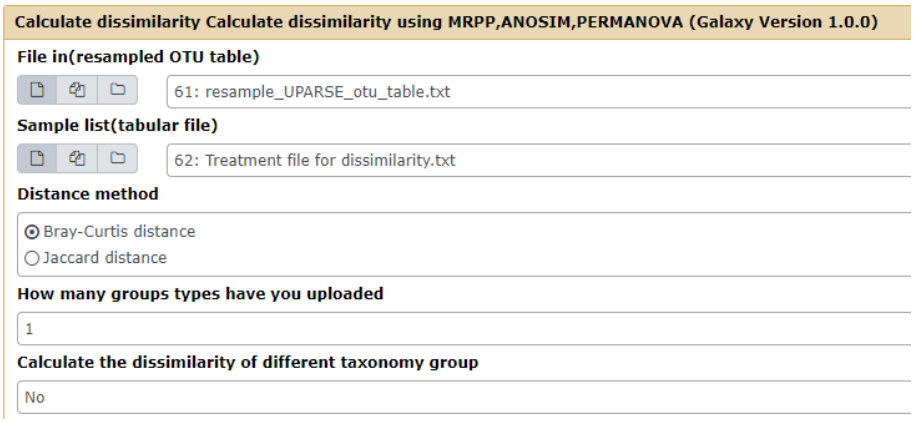
T test for each data

Origin	df	t	signif(p)
OTU_11	4.9398	0.9047	0.4075
OTU_5	4.8699	0.9726	0.3765
OTU_1	4.6596	1.3029	0.2532
OTU_1041	6.5780	-0.0119	0.9908

3.3 Dissimilarity (MRPP, adonis, anosim)

Using resample_otu_table and group_file to calculate the dissimilarity based on three different methods, MRPP, ANOSIM and PERMANOVA.

Inputs:



Parameter options:

Distance method: Bray-Curtis distance / Jaccard distance

Groups: 1 / 2 (Please select according to your sample list)

Calculate the dissimilarity of different taxonomy group: No (default) / Yes (Provide another taxonmy tabular file for each OTU)

Sample list for 1 group and 2 group:

A1	A
A2	A
A3	A
A4	A
A5	A
A6	A
B1	B
B2	B
B3	B
B4	B
B5	B
B6	B
C1	C
C2	C
C3	C
C4	C
C5	C
C6	C

A2R2	A	Recovery
A4R1	A	Recovery
A4R2	A	Recovery
A6R1	A	Recovery
A6R2	A	Recovery
B2A1	B	Post
B2A2	B	Post
B3A1	B	Post
B3A2	B	Post
B6A1	B	Post
B6A2	B	Post
B2B1	B	PRecovery
B2B2	B	PRecovery
B3B1	B	PRecovery

You can upload either 1 or 2 types of grouping approaches in the grouping file and choose the corresponding number "1" or "2" to define the calculation result.

Outputs:

92: dissimilarity result of bray_

Methods	Whole dataset
MRPP.delta	0.3457
MRPP.P	0.001
ANOSIM.r	0.8069
ANOSIM.P	0.001
PERMANOVA.F	11.1642
PERMANOVA.P	0.001

MRPP	A	B	C
A	0	0.006	0.005
B	0.4243	0	0.001
C	0.3123	0.3006	0

ANOSIM	A	B	C
A	0	0.004	0.004
B	0.5055	0	0.003
C	0.8240	0.9814	0

PERMANOVA	A	B	C
A	0	0.005	0.001
B	4.2306	0	0.004
C	15.7546	18.2536	0

Attention: The values of upper triangular matrices are the significance value (p-value). The values of lower triangular matrices for MRPP, ANOSIM and PERMANOVA are delta, R value and F-value, respectively.

The dissimilarity test for different taxonomy profiles are also available in this analysis tool. The annotation file is the classifier file related to the resample_otu_table you have used above. The result of this annotation part might look like this:

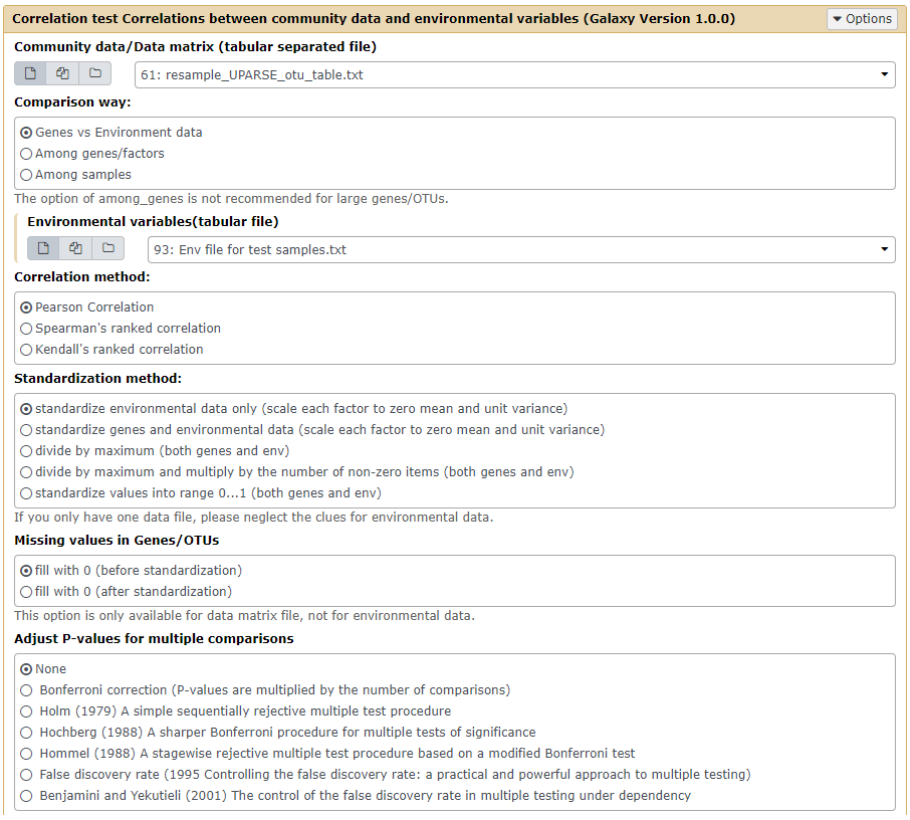
For each taxonomy category

	F.model	P-value	R2
Acidimicrobiales	5.7781	0.001	0.2007
Actinomycetales	5.2298	0.001	0.1701
Alteromonadales	6.9174	0.001	0.2201
Anaerolineales	5.0798	0.001	0.1661
Bacillales	2.1603	0.03	0.0841
Bacteroidales	2.1191	0.018	0.0767
Bdellovibrionales	3.0417	0.007	0.1104
Burkholderiales	5.4732	0.001	0.1767
Caldilineales	11.6734	0.001	0.3140
Campylobacterales	2.8721	0.002	0.1754

4. Environmental associations

4.1 Correlation test

Inputs:



Parameter options:

Comparison way: Genes vs Environment data (provide a environmental data) / Among genes (factors) / Among samples

Please attention the file format:

Factors	Sample1	Sample2	...	SampleN
factor1	data11	data12	...	data1N
factor2	data21	data22	...	data2N
factor3	data31	data32	...	data3N

Sample	factor Name1	factor Name2	factor Name3	...	factorN
Sample1	data11	data12	data13	...	data1N
Sample2	data21	data22	data23	...	data2N
Sample3	data31	data32	data33	...	data3N

Correlation method: Pearson / Spearman / Kendell

Standardization method: default selection (change if necessary)

Missing value: fill with zero (before; change if necessary)

Ajust P-value for multiple comparisons: None (change if necessary)

Outputs:

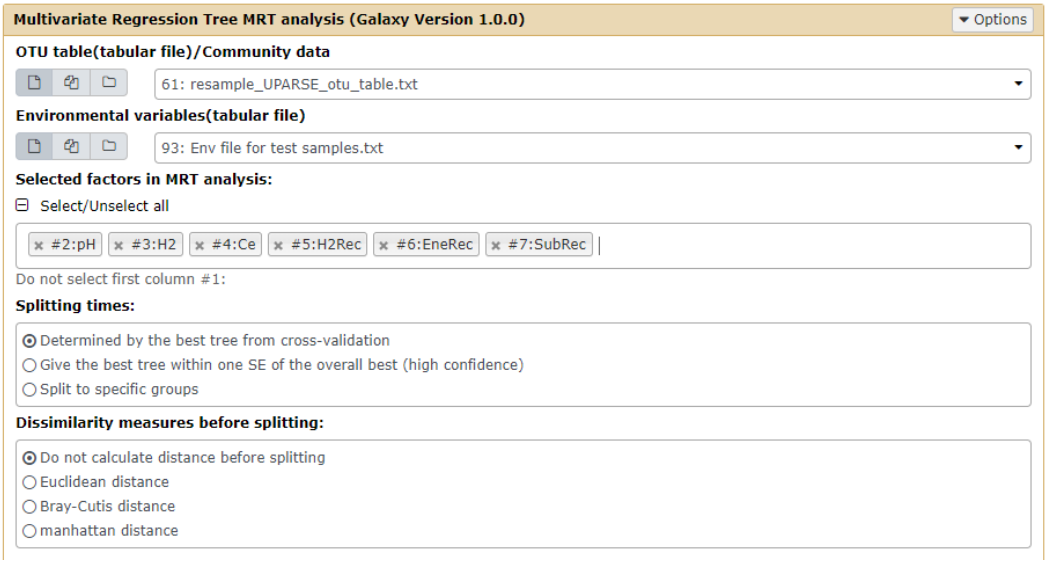
94: pearson correlation coefficient and significance result

```
C3      A4      C5      C1      B2      A5      B3
C3      1(0)    0.291(1.41e-27) 0.991(0)        0.
A4      0.291(1.41e-27) 1(0)    0.327(1.06e-34) 0.
99(2.87e-131)
```

correlation coefficient (*P* value)

4.2 Multivariate Regression Tree (MRT)

Inputs:



Parameter options:

Select factors in MRT analysis: select what you need for the model

Splitting times: Determined by the best tree from cross-validation (recommend)

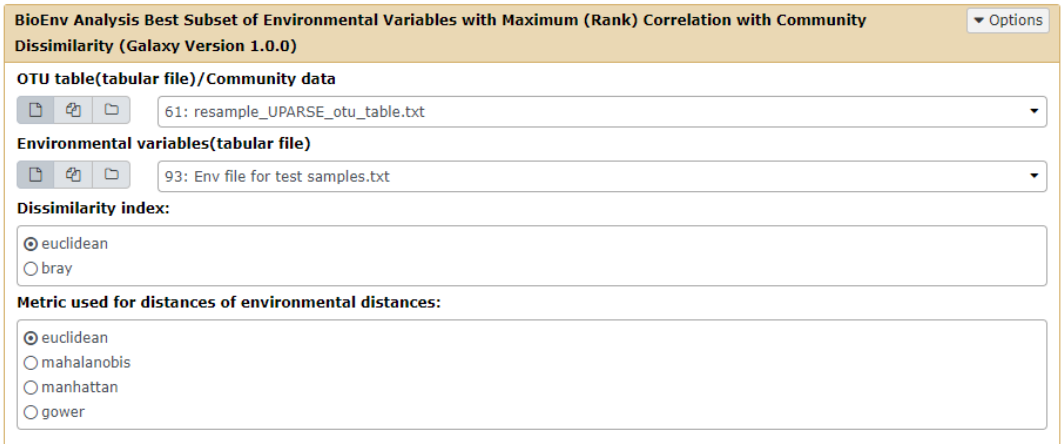
Dissimilarity measures before splitting: Do not calculate distance (change it if errors happened)

Outputs:

95: MRT result

4.3 BioEnv Analysis

Inputs:



Parameter options:

Dissimilarity index: euclidean / bray

Metric used for distance of environmental distances: euclidean / mahalanobis / manhattan / gower

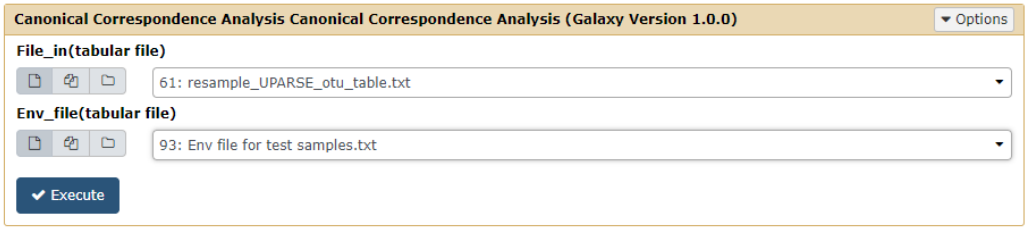
Outputs:

97: BioEnv result

	size	correlation
Ce	1	0.0568
Ce RT	2	0.0492
Ce EneRec RT	3	0.0196
H2 Ce EneRec RT	4	0.0376

4.4 CCA

Inputs:



Outputs:

101: cca plot

100: individual anova test F and p values

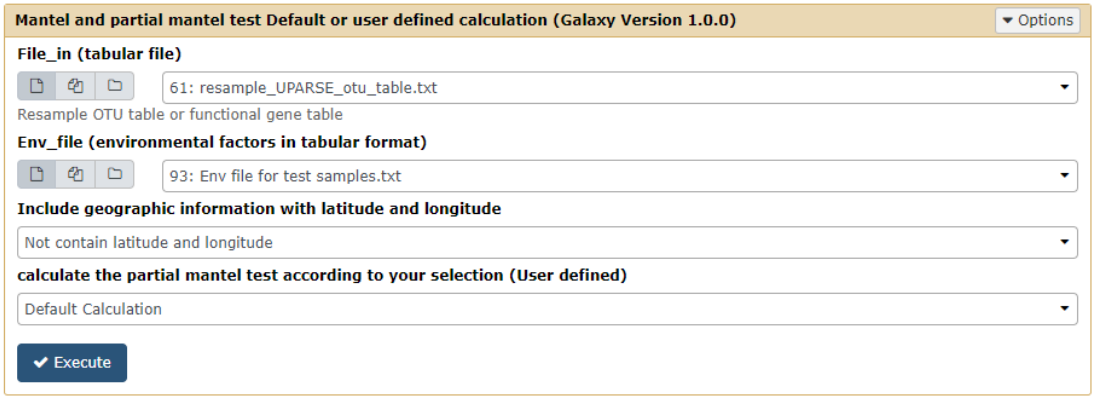
99: inflation factors

98: cca result

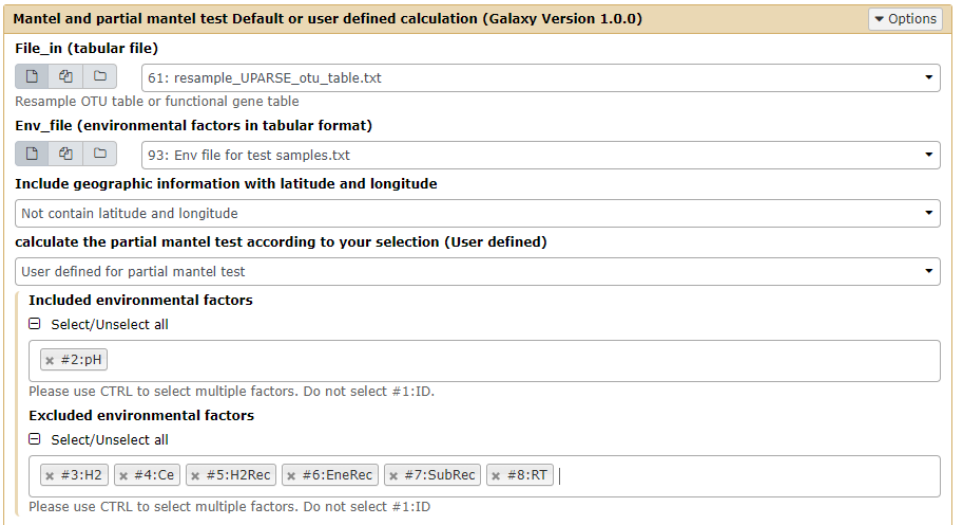
4.5 Mantel Test

Input:

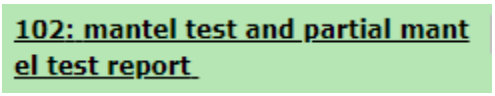
For mantel and partial mantel test:



For user-defined partial mantel test:

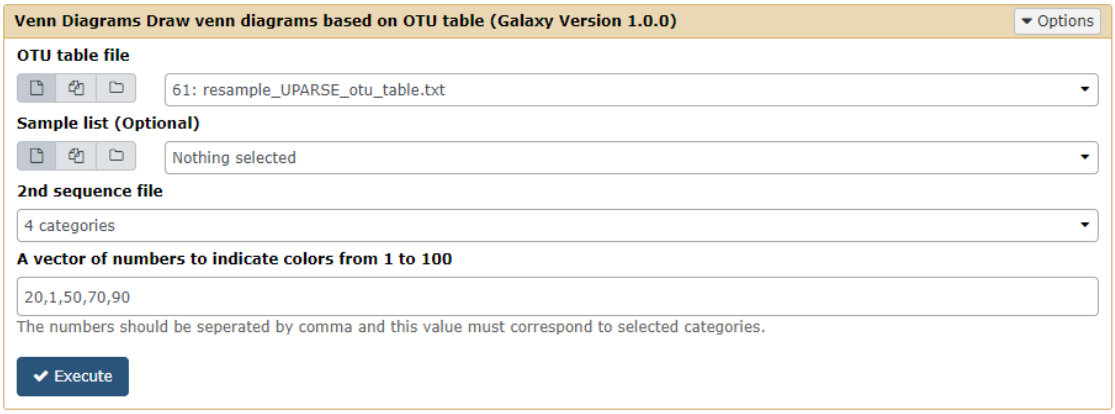


Output:



5. Plotting figures

5.1 Venn Diagrams



Input Format

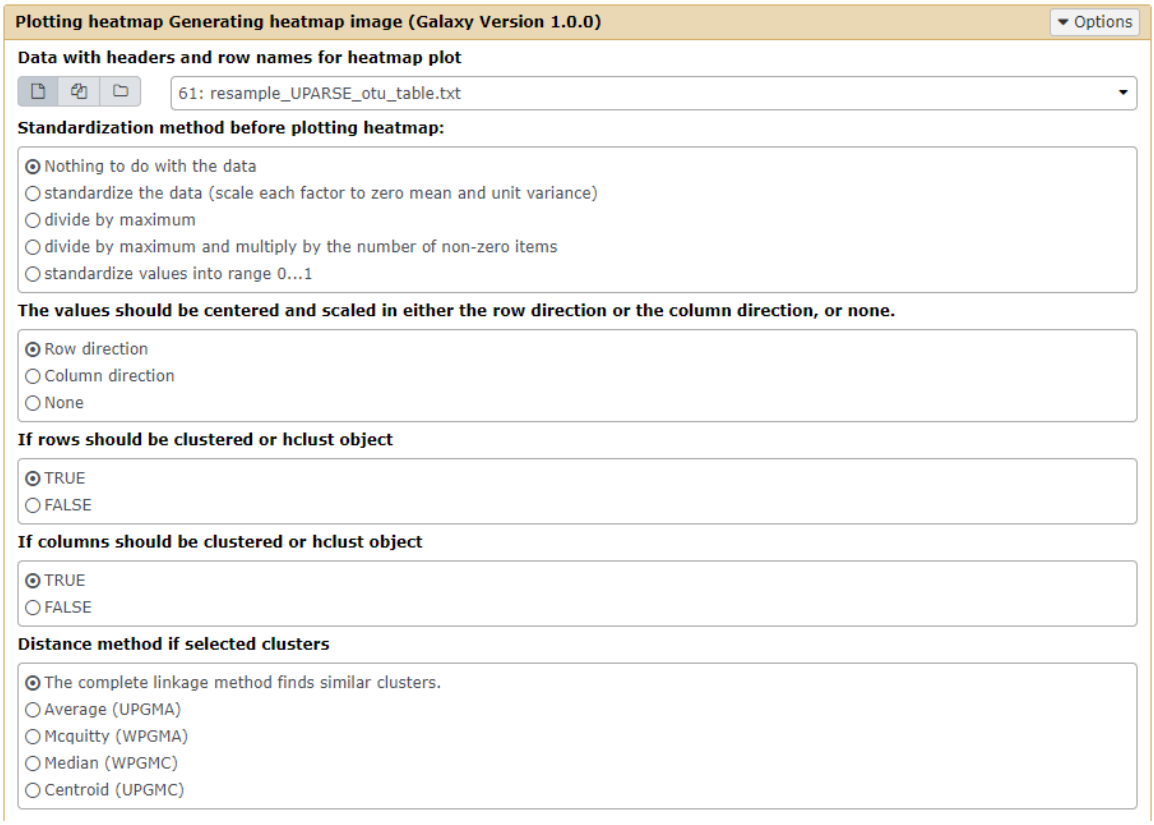
- OTU table:

  The OTU table should contain one head row starts with "OTU" and then the tag/sample/treatment list.

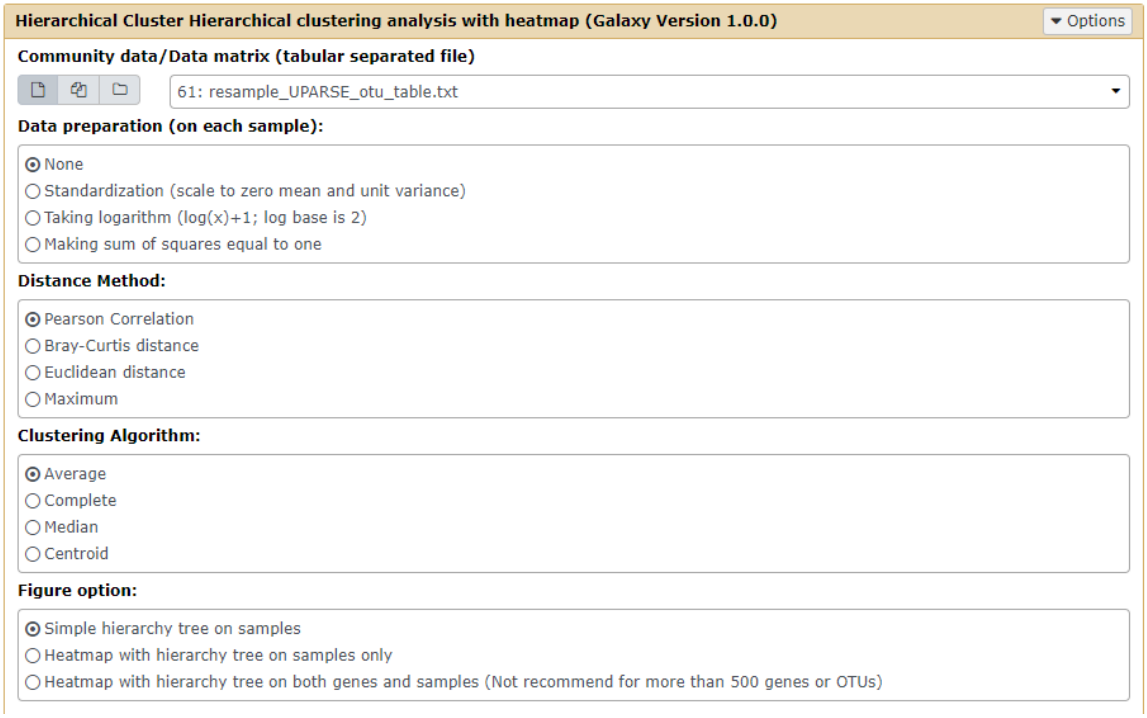
- Sample list (optional):

```
    Sample1_name:tag1,tag2,tag3
    Sample2_name:tag4,tag5,tag6
    Sample3_name:tag7,tag8,tag9
......
```

5.2 Heatmap



5.3 Hierarchical cluster

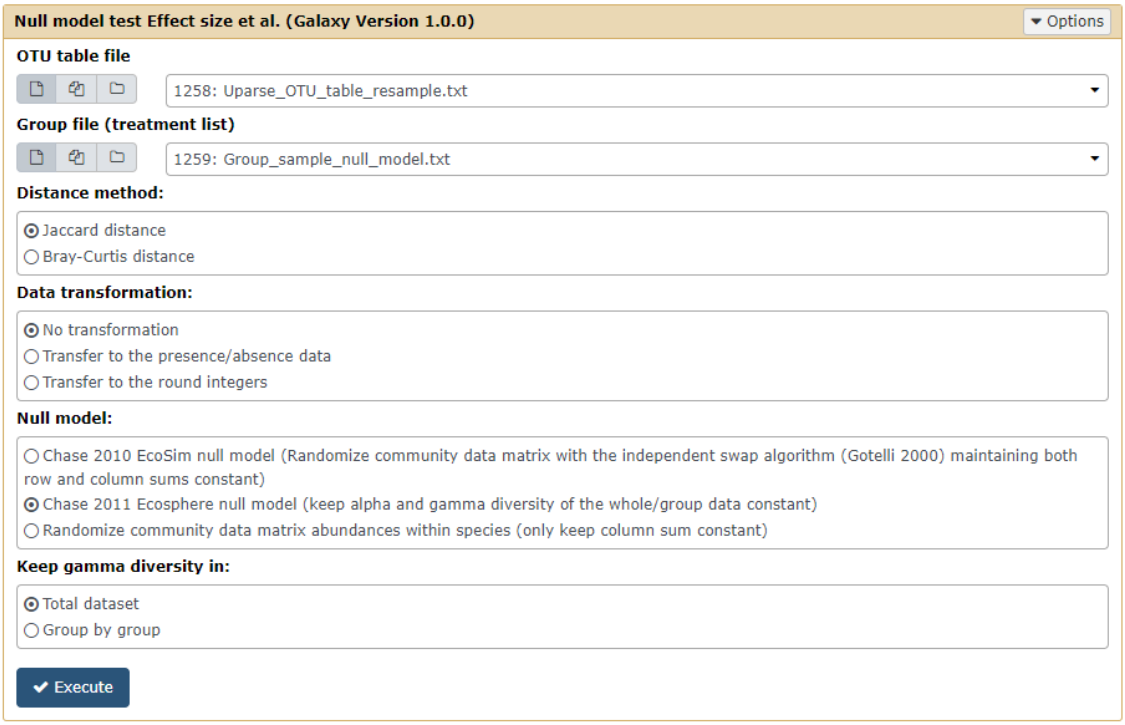


C. Ecological process analysis

1. Null model test

If you are interested in this analysis, please further read: Zhou JZ, Deng Y, Zhang P, Xue K, Liang YT, Van Nostrand JD, et al. Stochasticity, succession, and environmental perturbations in a fluidic ecosystem. Proc Natl Acad Sci. 2014;111:E836-E45.

Input:



Output:

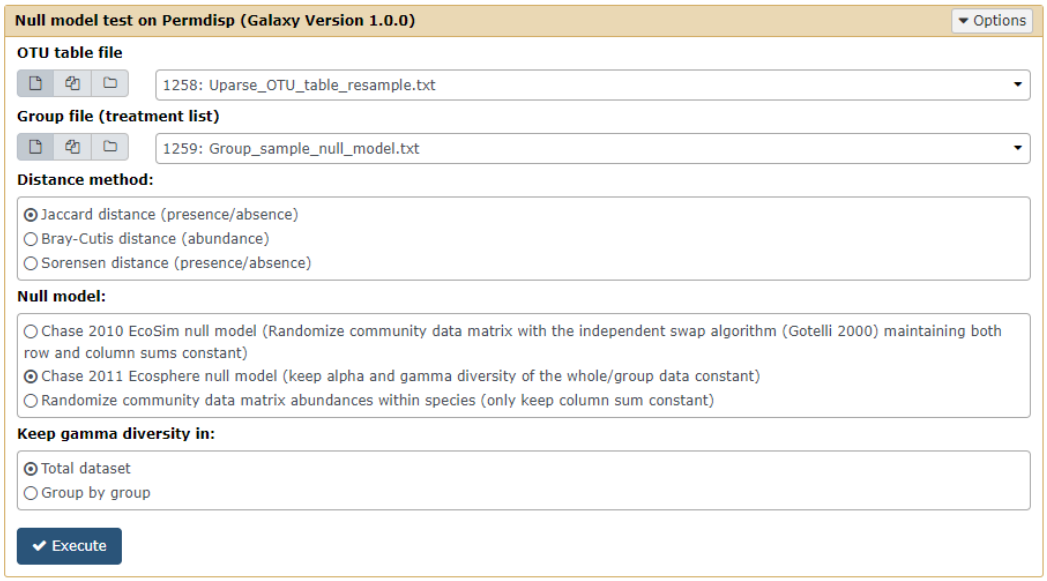
ANOVA test for null model

Null model result

2. Null model test on Permdisp

If you are interested in this analysis, please further read: Zhou JZ, Deng Y, Zhang P, Xue K, Liang YT, Van Nostrand JD, et al. Stochasticity, succession, and environmental perturbations in a fluidic ecosystem. Proc Natl Acad Sci. 2014;111:E836-E45.

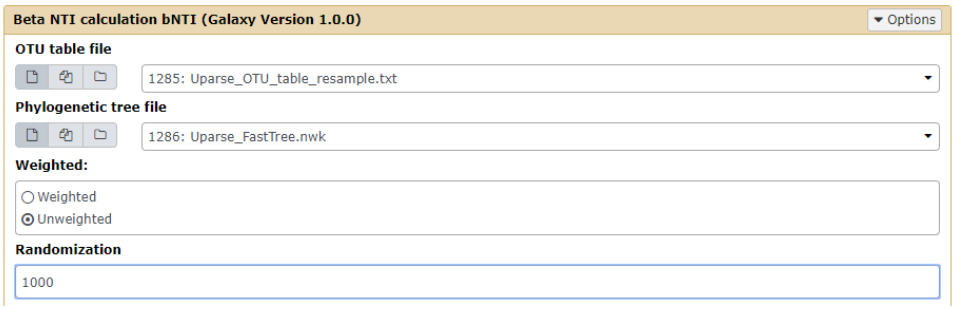
Input:



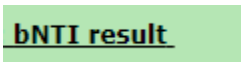
Output:

3. Beta NTI calculation

Input:

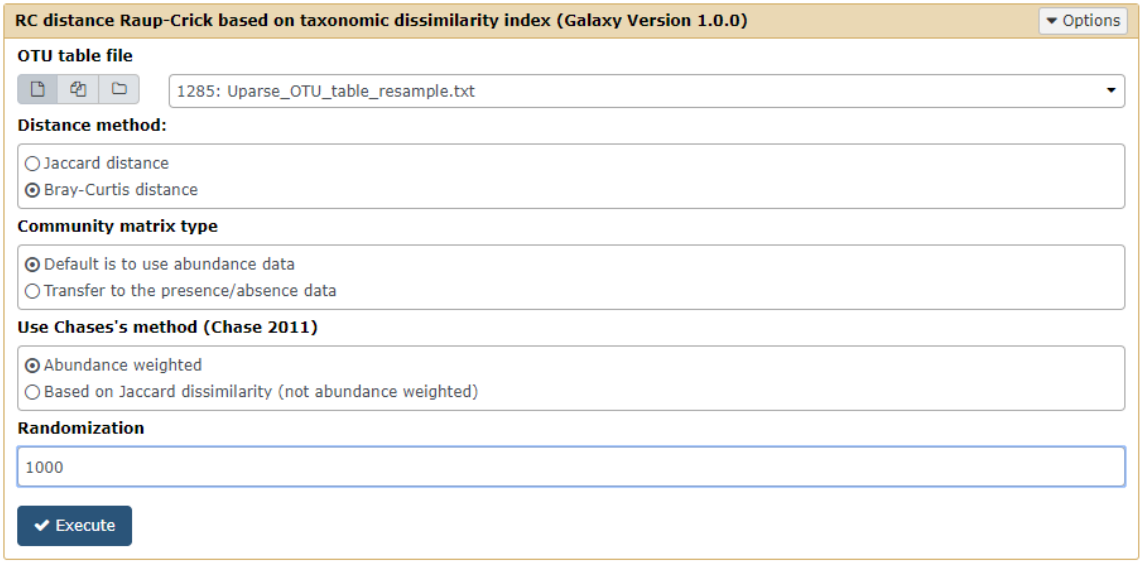


Output:

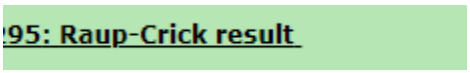


4. RC distance

Input:

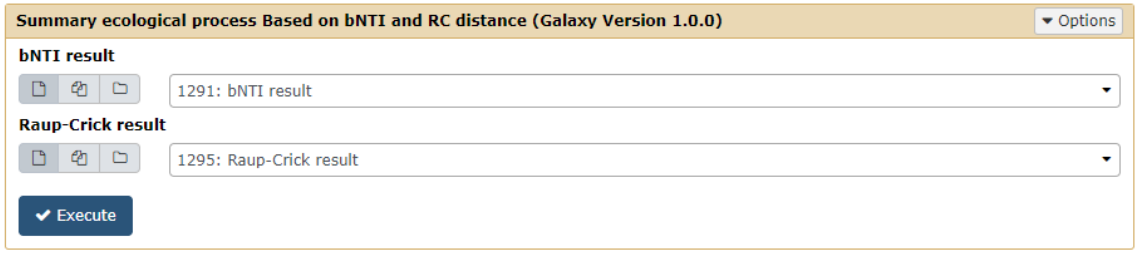


Output:

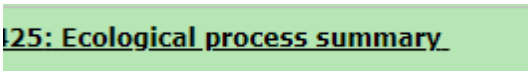


5. Summary ecological process

Input:



Output:



D. Functional profile prediction approaches

The introductions of this section will only cover some fundamental operations in our analysis pipeline. The results including the inference and plotting figures should be referred to the original literatures for each method.

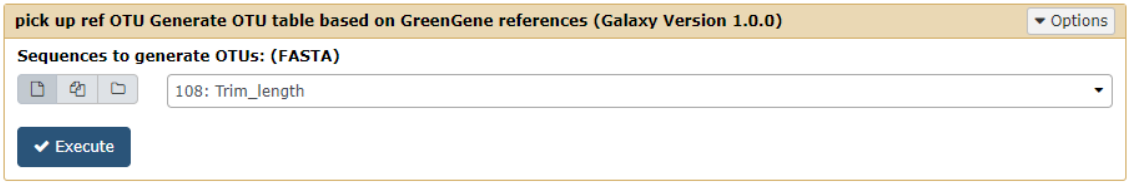
The functional profile predictions are mainly divided into two parts according to their amplicon sequences: 16S-based and ITS-based analysis. The functional profiles prediction tools for 16S-based sequences mainly included PICRUSt, Tax4Fun, FAPROTAX and BugBase. For ITS-based sequences, there was only FunGuild method in this analysis pipeline.

1. PICRUSt

Please see this paper when you have some questions: Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Langille, M. G.I.; Zaneveld, J.; Caporaso, J. G.; McDonald, D.; Knights, D.; a Reyes, J.; Clemente, J. C.; Burkepile, D. E.; Vega Thurber, R. L.; Knight, R.; Beiko, R. G.; and Huttenhower, C. Nature Biotechnology, 1-10. 8 2013.

1.1 Pick up ref OTU

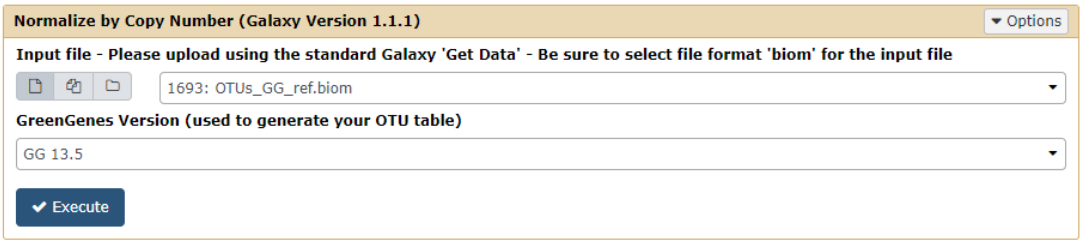
Input:



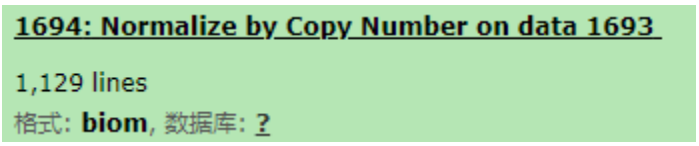
Output:

1.2 Normalize by Copy Number

Input:

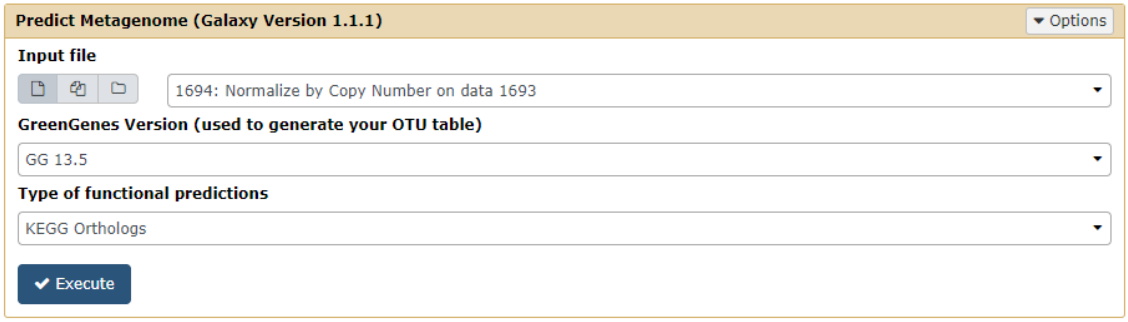


Output:

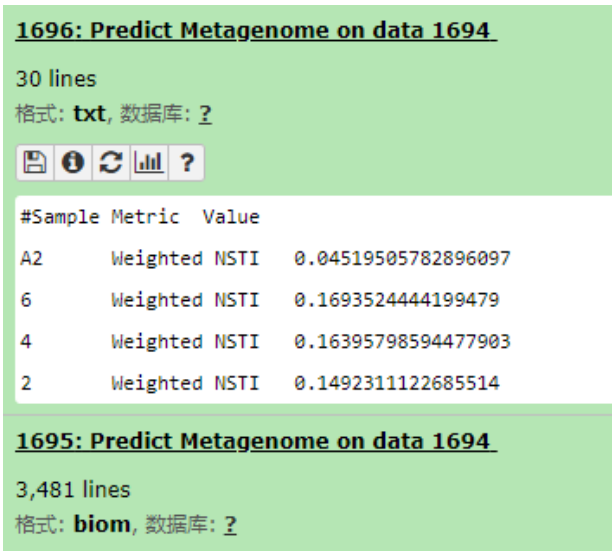


1.3 Predict Metagenome

Input:

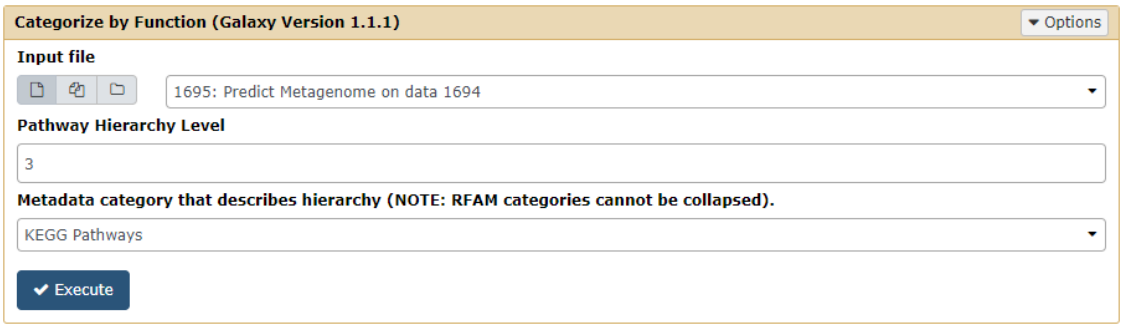


Output:

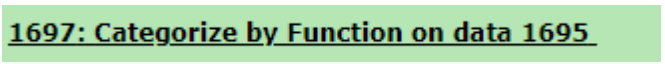


1.4 Categorize by Function

Input:

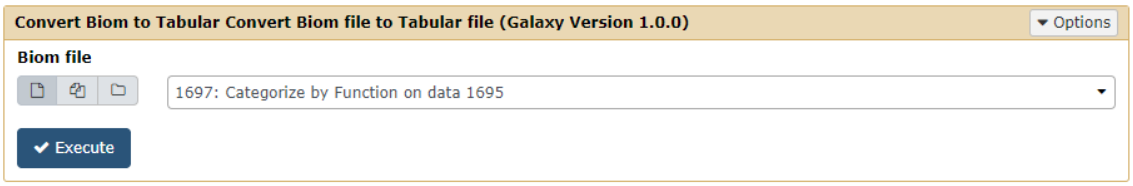


Output:

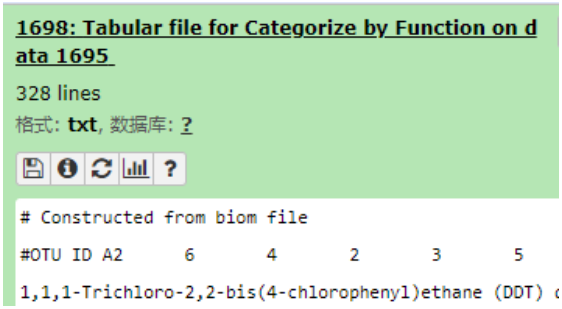


Convert Biom to Tabular

Input:



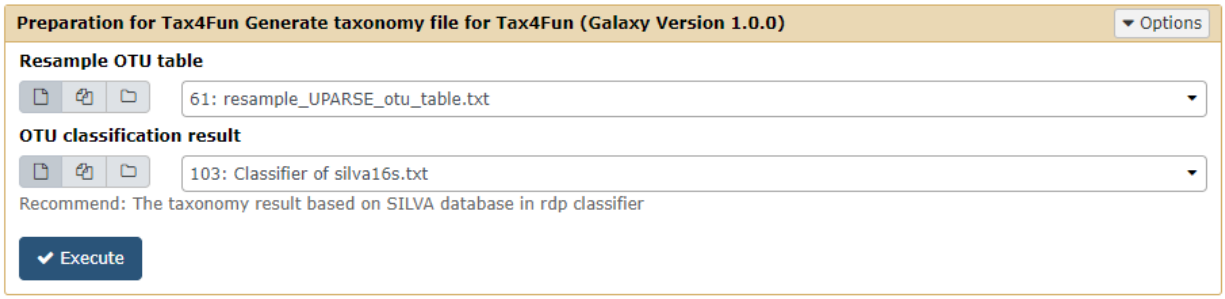
Output:



2. Tax4Fun

2.1 Preparation for Tax4Fun

Input:



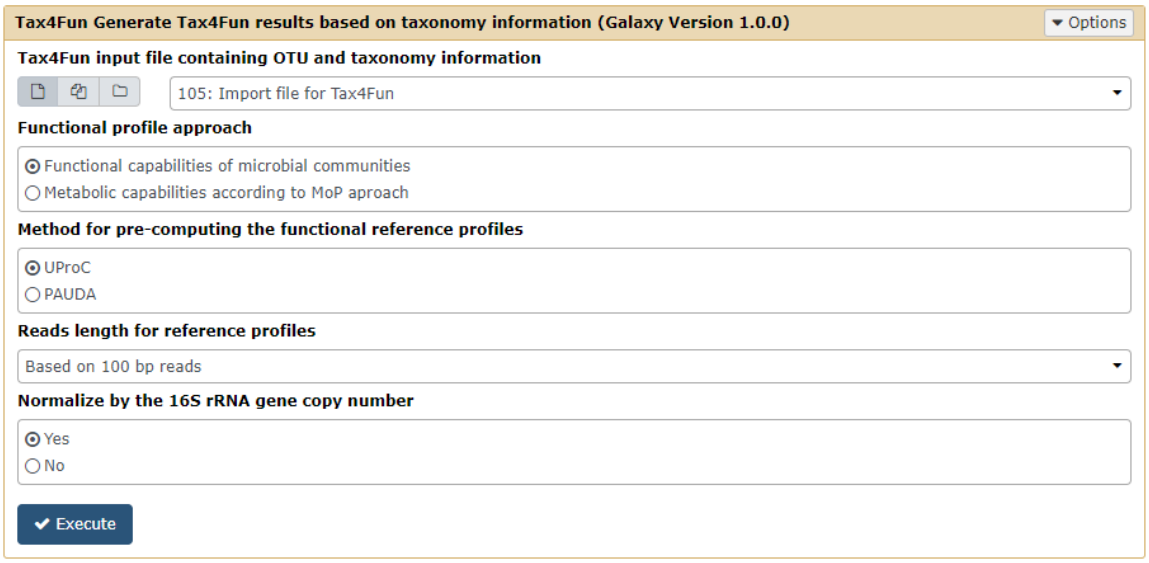
Output:

- Import file for Tax4Fun

2.2 Tax4Fun

This tool is to predict functional profiles from metagenomic 16S rRNA data using Tax4Fun.

Input:



Output:

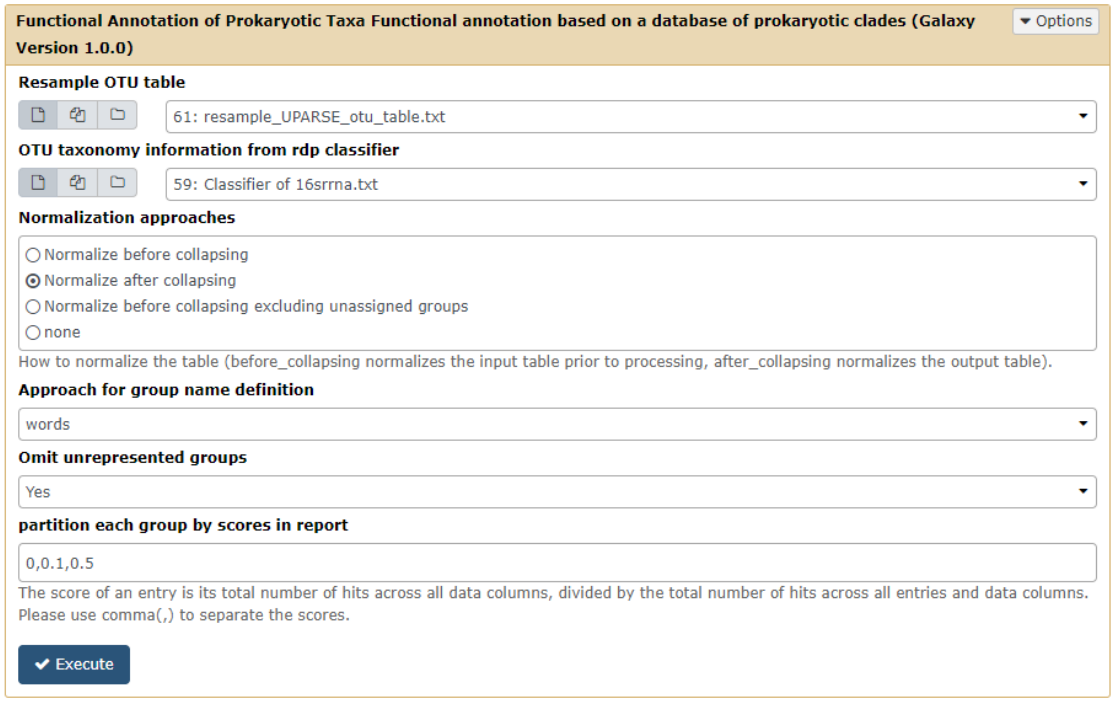
- Tax4Fun results

- FTU (fraction of taxonomic units unexplained)

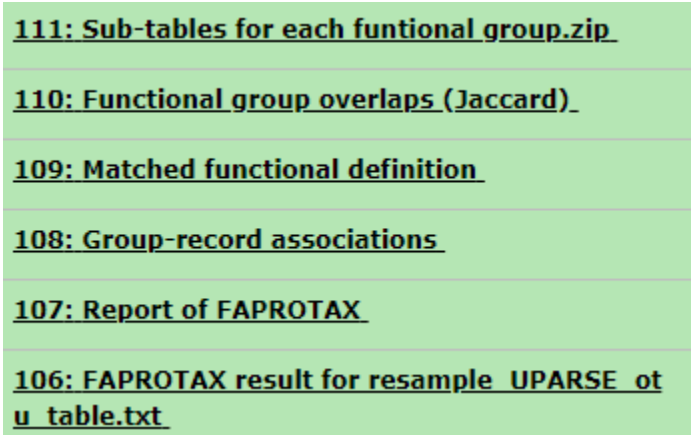
3. FAPROTAX

This is FAPROTAX (Functional Annotation of Prokaryotic Taxa), a database that maps prokaryotic clades (e.g. genera, species or subspecies) to established metabolic or other ecologically relevant functions based on the current literature. FAPROTAX includes software for converting taxonomic microbial community profiles (e.g. in the form of an OTU table) into putative functional profiles, based on taxa identified in a sample. The web site is http://www.zoology.ubc.ca/louca/FAPROTAX/. Please cite: Louca, S. and Parfrey, L. W. and Doebeli, M. (2016). Decoupling function and taxonomy in the global ocean microbiome. In Science, 353 (6305), pp. 1272-1277. [doi:10.1126/science.aaf4507]

Input:



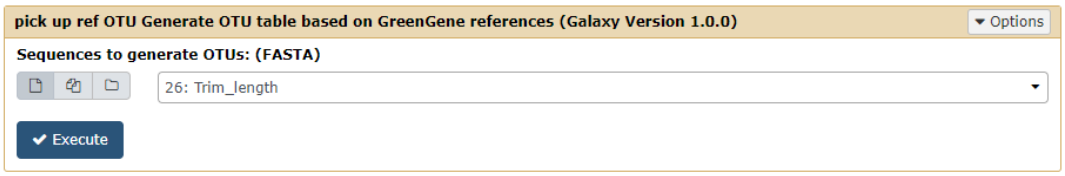
Output:



4. BugBase

4.1 Pick up ref OTU

Input:

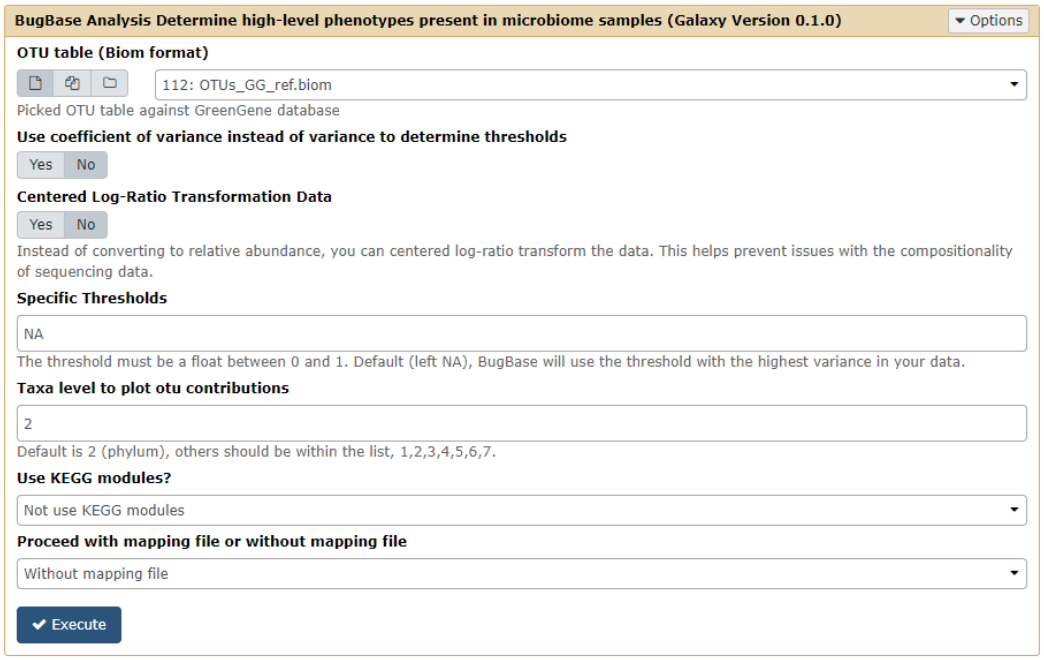


Output:

112: OTUs_GG_ref.biom

4.2 BugBase Analysis

Input:



Output:

116: Threshold used in OTUs_GG_ref.biom.zip

115: Predicted phenotypes for OTUs_GG_ref.biom.zip

114: OTU contributions for OTUs_GG_ref.biom.zip

113: Normalized otu table for OTUs_GG_ref.biom

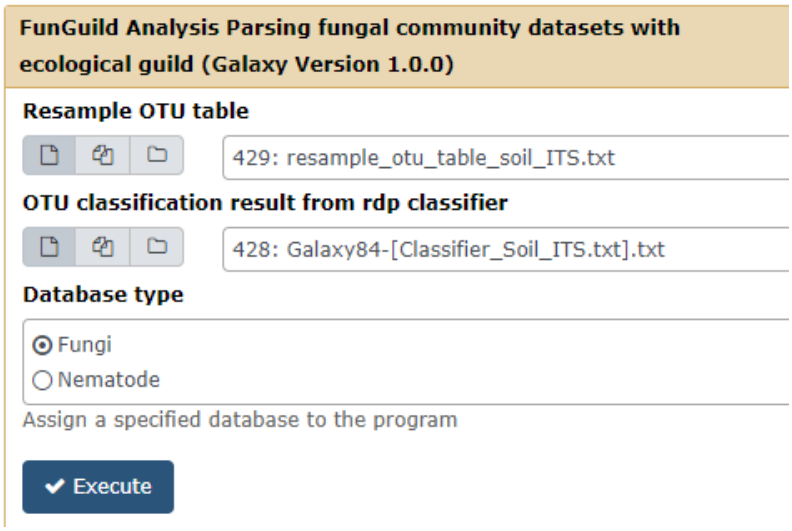
If you have a mappling file, you can proceed with mapping file:

- **Mapping file:**
  - Be a tab-delimited text file
  - Have sample IDs in the first column
  - Have column headers in the first row
  - Have #SampleID as the first header
  - Contain only letters, numbers, underscores and hyphens
  - Not contain spaces, commas or quotes
  - Never contain confidential information

```
-  #SampleID   Group   Location        Details
-  Sample1     A       Lab 1           PCR_water_sample_1
-  Sample2     A       Lab 2           PCR_water_sample_2
-  Sample3     B       Lab 1           PCR_soil_sample_1
-  Sample4     B       Lab 2           PCR_soil_sample_2
```

5. FunGuild

Input:



Output:

➢ Funguild result for "resample_otu_table_soil_ITS.txt"

OTUID	HGT	HZT	QGT	QZT	ZGT	ZZT	taxonomy	Taxon	Taxon Leve	Trophic Mo	Guild	Growth Mo	Trait	Confidence	Notes	Citation/Source	
OTU_1	8114	997	1377	8273	3336	16274	Fungi;Asco	Aleuria	13	Saprotropl	Undefined	NULL	NULL	Probable	Probable	NULL	Tedersoo L, et al. 2014. S
OTU_2	5262	3749	6208	8525	9986	3052	Fungi;Asco	Didymella	13	Pathotropl	Plant Path	NULL	NULL	Probable	Probable	NULL	Tedersoo L, et al. 2014. S
OTU_5	4674	1909	5033	7469	9330	6523	Fungi;Zygc	Mortierella	13	Saprotropl	Undefined	NULL	NULL	Probable	Probable	NULL	Tedersoo L, et al. 2014. S
OTU_4	2844	11521	79	3510	451	543	Fungi;Zygc	Mortierella	13	Saprotropl	Undefined	NULL	NULL	Probable	Probable	NULL	Tedersoo L, et al. 2014. S
OTU_3	3150	3950	184	6140	227	886	Fungi;Zygc	Mucor	13	Saprotropl	Undefined	NULL	NULL	Probable	Probable	NULL	Tedersoo L, et al. 2014. S
OTU_8	1800	682	2287	594	3682	1014	Fungi;Asco	-	-	-	-	-	-	-	Unassigne	-	
OTU_6	39	6	118	77	5702	4076	Fungi;Asco	Saccharom	7	Saprotropl	Undefined	Yeast	NULL	Possible	Possible	NULL	Sterkenburg E, et al. 2015
OTU_7	699	4754	42	3744	228	168	Fungi;Asco	Debaryom	20	Saprotropl	Undefined	Yeast	NULL	Highly Prol	Highly Prol	NULL	Kurtzman CP, et al. (eds.)
OTU_9	633	6620	154	1192	556	230	Fungi;Asco	Alternaria	20	Pathotropl	Endophyte	NULL	NULL	Possible	Possible	Host - Soi	Costa IPMW, et al. 2012. I
OTU_13	2724	729	1934	11	786	1232	Fungi;Chyt	Monoblep	13	Saprotropl	Undefined	NULL	NULL	Probable	Probable	NULL	Tedersoo L, et al. 2014. S
OTU_11	888	1220	549	687	2486	1056	Fungi;Basic	Rhodotoru	13	Pathotropl	Animal Syr	NULL	NULL	Probable	Probable	Opportuni	http://www.cdc.gov
OTU_10	367	174	4862	554	638	31	Fungi;Asco	Saccharom	7	Saprotropl	Undefined	Yeast	NULL	Possible	Possible	NULL	Sterkenburg E, et al. 2015
OTU_3343	114	196	1634	4462	40	10	Fungi;Zygc	Mortierella	13	Saprotropl	Undefined	NULL	NULL	Probable	Probable	NULL	Tedersoo L, et al. 2014. S
OTU_20	929	123	1866	373	1844	1046	Fungi	-	-	-	-	-	-	-	Unassigne	-	

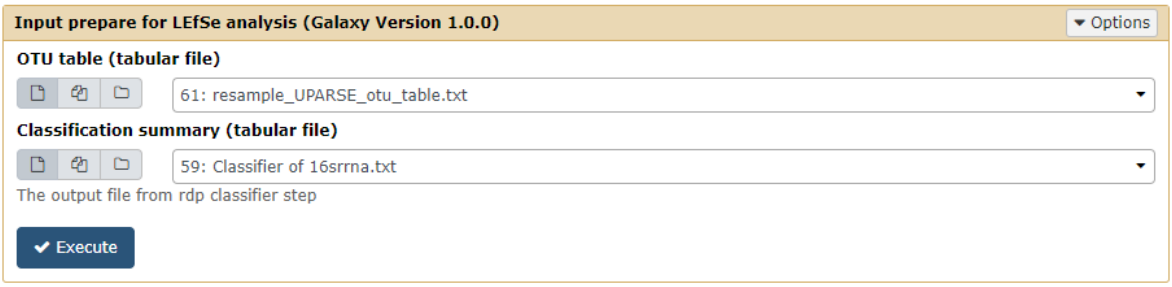
E. Other analysis tools in microbial ecology

1. LEfSe

1.1 Input prepare for LefSe anlaysis

This tool is mainly used to merge multiple files into one file.

Input:



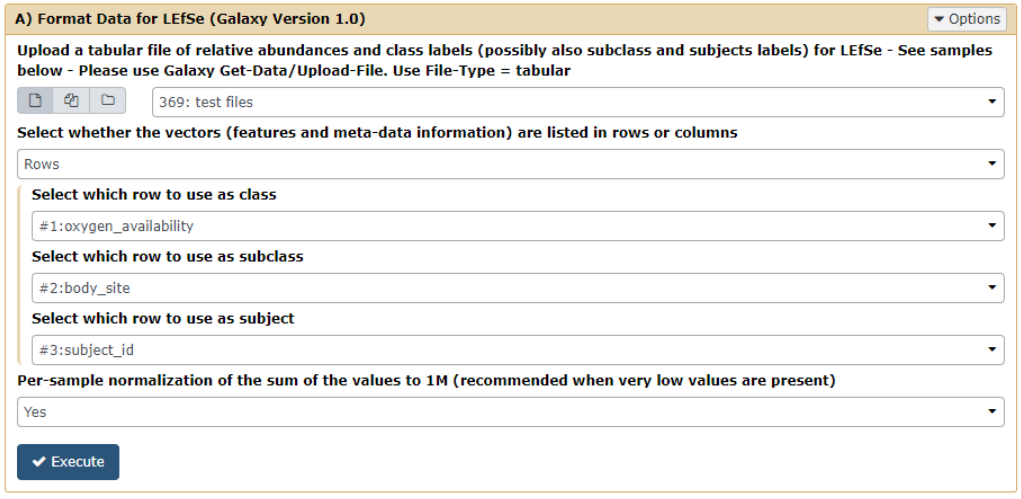
Output:

117: Input for LEfSe analysis

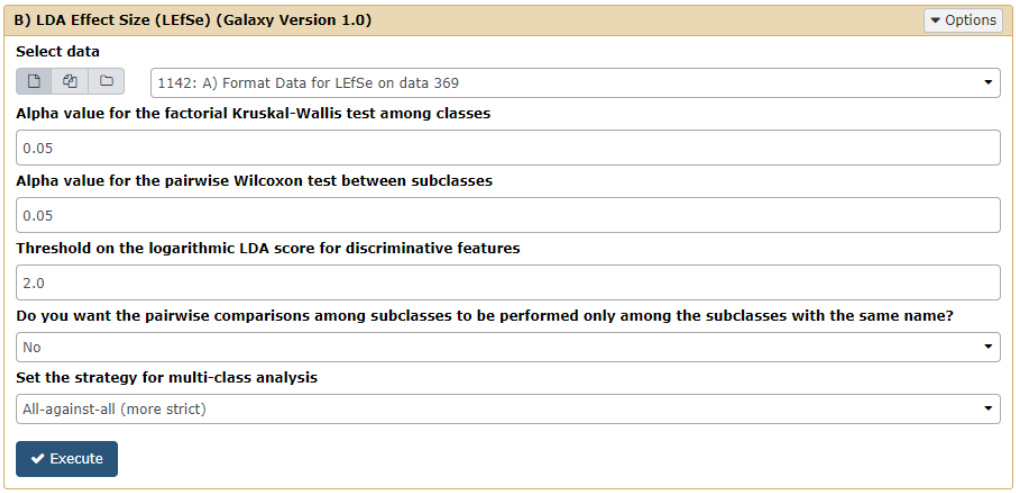
""	"C3"	"A4"	"C5"
"Archaea"	"0.0269639773381404"	"0.117429636137789"	"0.0140273275365831"
"Bacteria"	"0.97303602266186"	"0.882570363862211"	"0.985972672463417"
"Archaea|Crenarchaeota"	"3.02966037507195e-05"	"0"	"0"
"Archaea|Euryarchaeota"	"0.0269336807343897"	"0.117429636137789"	"0.0140273275365831"
"Archaea|Thaumarchaeota"	"0"	"0"	"0"
"Archaea|Unclassified"	"0"	"0"	"0"
"Bacteria|Acidobacteria"	"0.00405974490259642"	"0.00487775320386585"	"0.00130275396128094"
"Bacteria|Actinobacteria"	"0.00490804980761657"	"0.0122398279152907"	"0.00281758414881692"
"Bacteria|Aquificae"	"0"	"6.05932075014391e-05"	"0"
"Bacteria|Armatimonadetes"	"0.000181779622504317"	"0.000363559245008635"	"0.000242372830005756"
"Bacteria|Bacteroidetes"	"0.0969188353985518"	"0.102978156148696"	"0.0265398248856303"

This program is used for summarizing the relative abundance for samples at multiple taxonomic levels, which is required for further LEfSe analysis. The output format should be very similar to the following shape. Later you could modify this table according to your demand, like adding different separation standards.

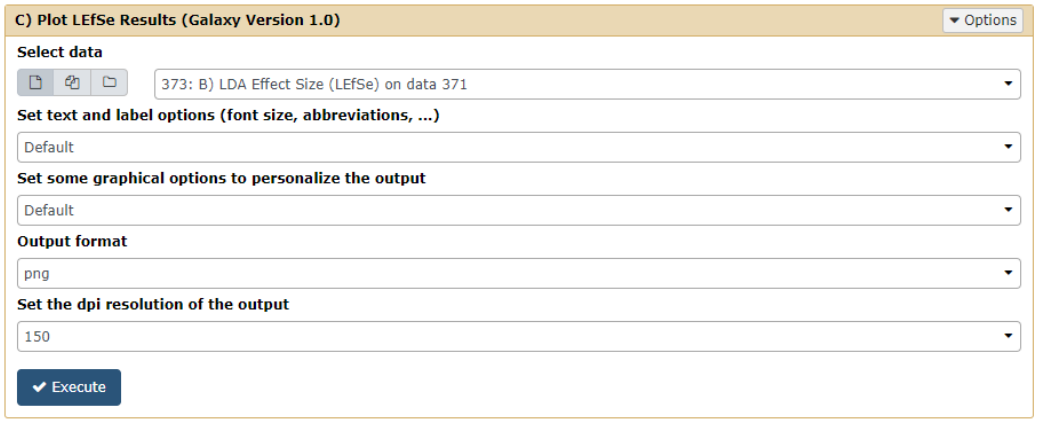
1.2 A) Format Data for LefSe



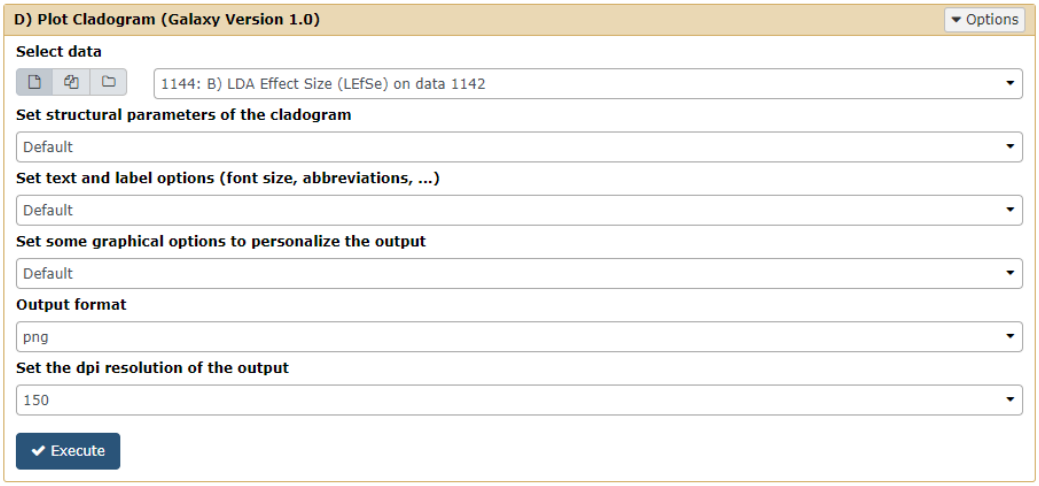
1.3 B) LDA Effect Size (LEfSe)



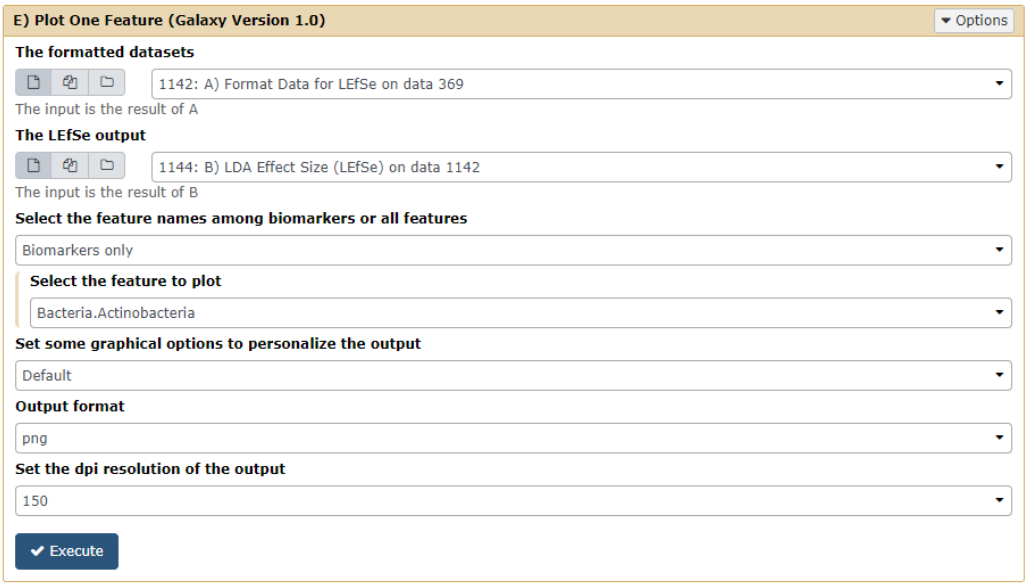
1.4 C) Plot LEfSe Results



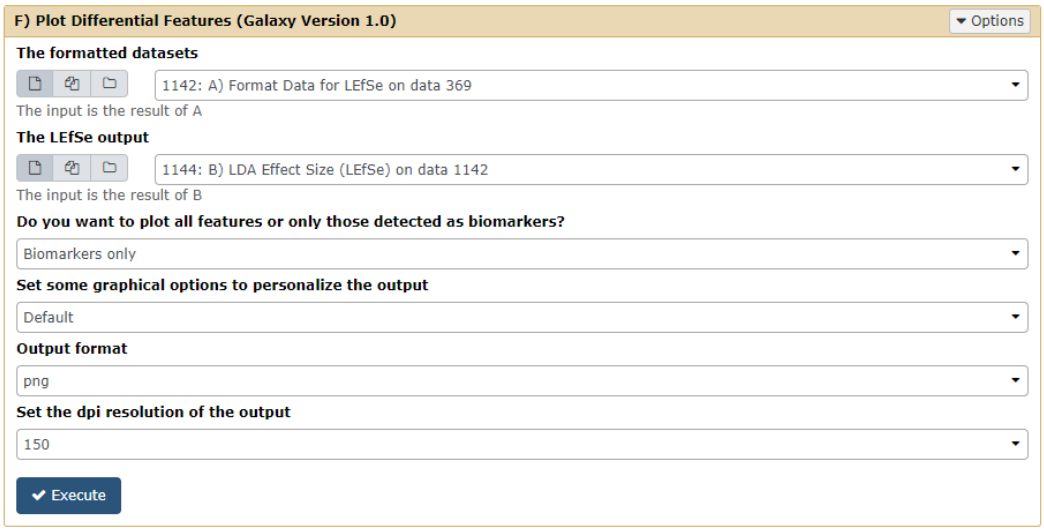
1.5 D) Plot Cladogram



1.6 E) Plot One Feature



1.7 F) Plot Differential Features

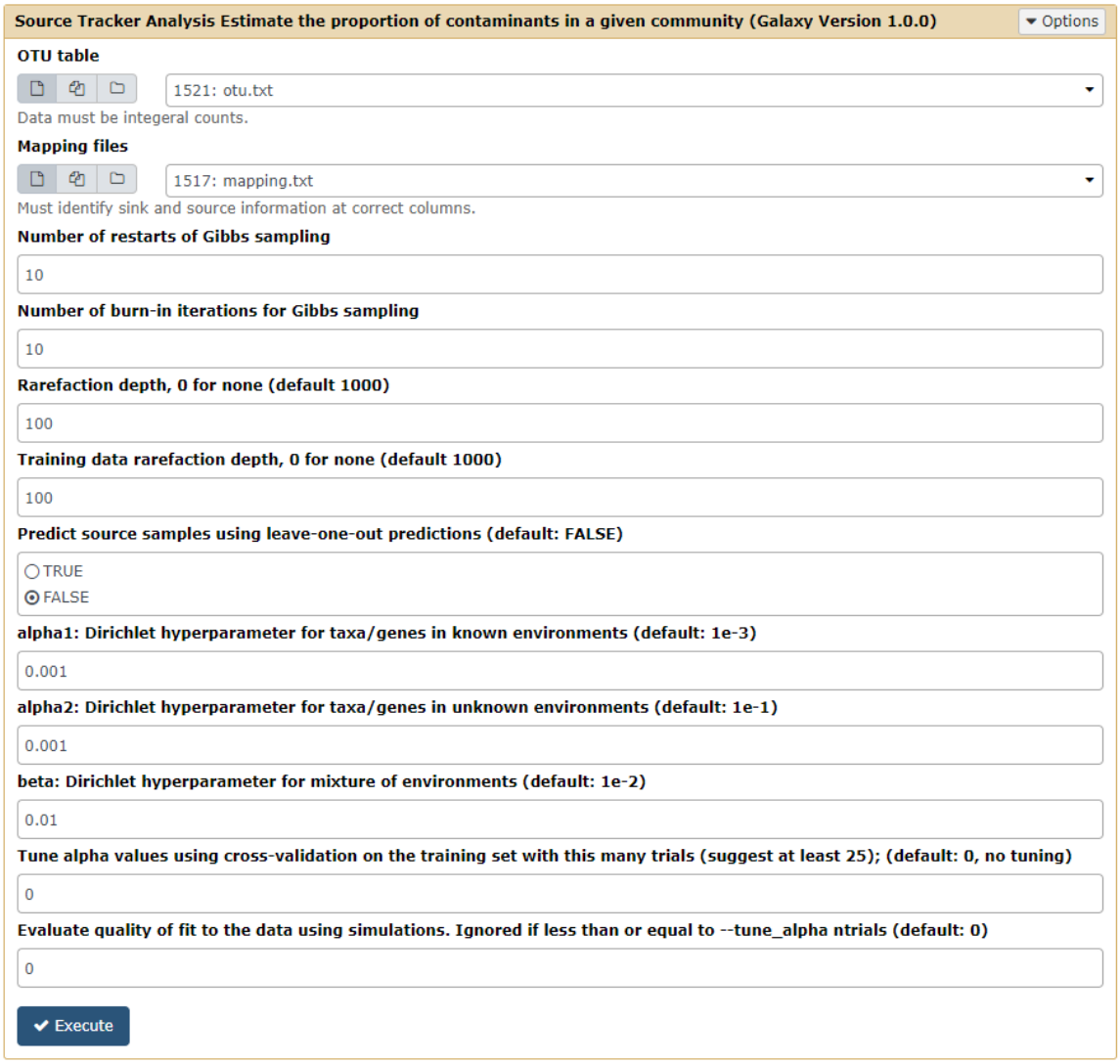


2. Source Tracker

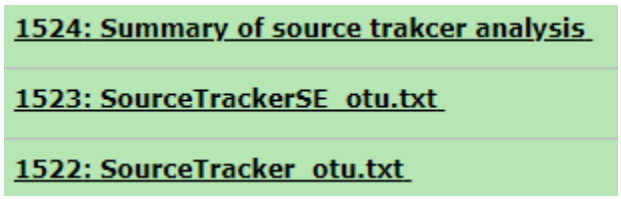
Please make a mapping file (tabular-separated txt) by yourself as the following format:

SampleID	Description	Env	SourceSink	Study	Details
Sample1	PCR water 1	A1	sink	Lab 1	PCR_water_sample_1
Sample2	PCR water 2	A2	sink	Lab 2	PCR_water_sample_2
Sample3	PCR soil 1	B1	source	Lab 1	PCR_soil_sample_1
Sample4	PCR soil 2	B2	source	Lab 2	PCR_soil_sample_2

Input:


Source Tracker Analysis Estimate the proportion of contaminants in a given community (Galaxy Version 1.0.0)
Options
OTU table
1521: otu.txt
Data must be integeral counts.
Mapping files
1517: mapping.txt
Must identify sink and source information at correct columns.
Number of restarts of Gibbs sampling
10
Number of burn-in iterations for Gibbs sampling
10
Rarefaction depth, 0 for none (default 1000)
100
Training data rarefaction depth, 0 for none (default 1000)
100
Predict source samples using leave-one-out predictions (default: FALSE)
TRUE
FALSE
alpha1: Dirichlet hyperparameter for taxa/genes in known environments (default: 1e-3)
0.001
alpha2: Dirichlet hyperparameter for taxa/genes in unknown environments (default: 1e-1)
0.001
beta: Dirichlet hyperparameter for mixture of environments (default: 1e-2)
0.01
Tune alpha values using cross-validation on the training set with this many trials (suggest at least 25); (default: 0, no tuning)
0
Evaluate quality of fit to the data using simulations. Ignored if less than or equal to --tune_alpha ntrials (default: 0)
0
Execute

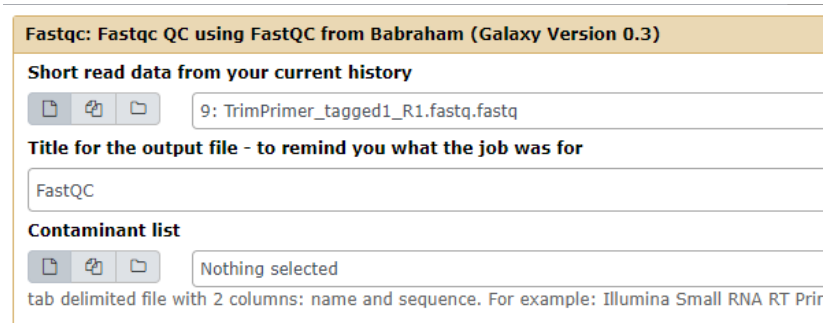
Output:


1524: Summary of source trakcer analysis
1523: SourceTrackerSE_otu.txt
1522: SourceTracker_otu.txt

F. Auxilliary tools in miscellaneous section

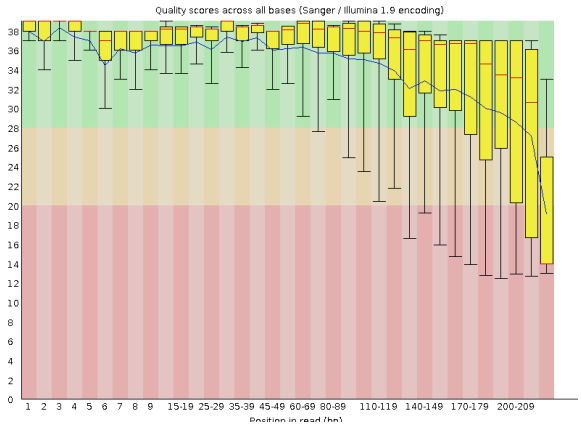
1. FastQC

Input:



Output:

- FastQC.html

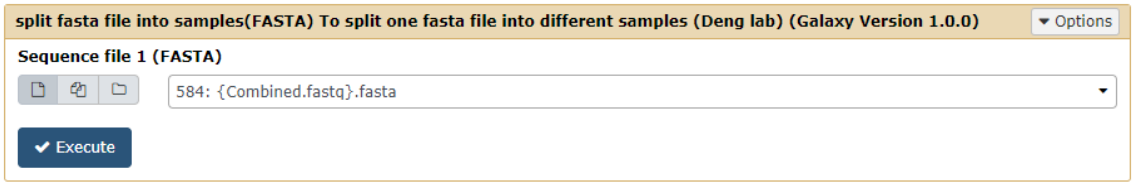


2. Split files into separated samples

This tool is mainly used to separate sequences within one file to multiple independent files. Then you could easily upload these sequence files to NCBI Sequence Read Archieve (SRA).

FASTA

Input:



Output:

After download it to local directory, you need to unzip this file twice. For the first step of unzipping process, you can easily unzip it. For the second step of unzipping process, you need to rename the extension file type to ".zip" or ".gz" and thereafter you could to unzip this file. After the two steps of unzipping, you can see the separated fasta or fastq files.

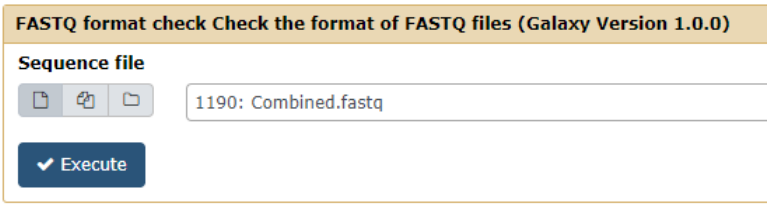
FASTQ

Similar options as above FASTQ program.

3. FASTQ format check

This tool is mainly used to check the file format of a fastq file, four lines for each sequence.

Input:



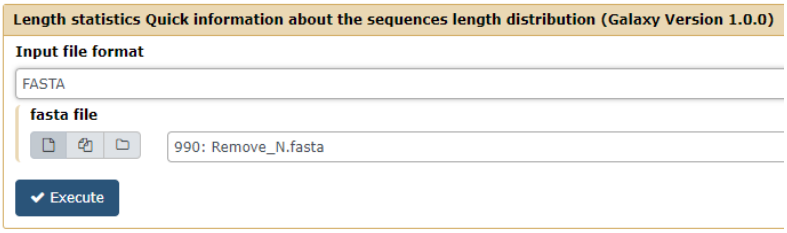
Output:

- Checked Combined.fastq
- Fastq check summary

4. Length Statistics

This tool is to summary the length distribution for fasta or fastq file.

Input:

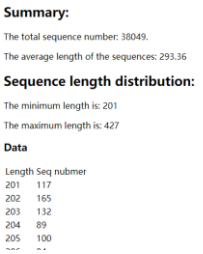


Options:

Input file format: FASTA / FASTQ

Output:

- length_stat.html

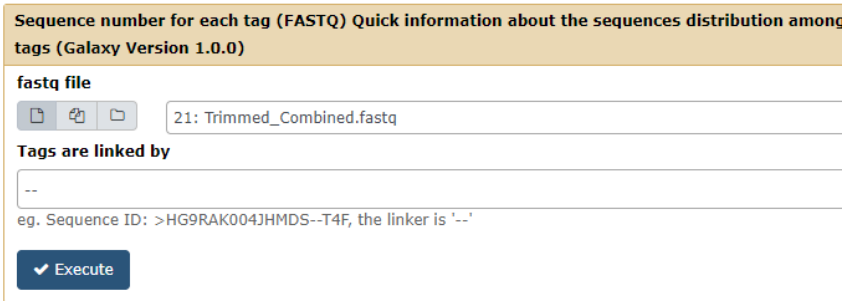


5. Sequence number for each tag

This tool is mainly used to make a summary sequence numbers for each tag.

FASTQ

Input:



Output:

- **Tag_stat.txt**

# Total sequence number 909188	
A1	48461
A2	44966
A3	72585
A4	51772
A5	44331

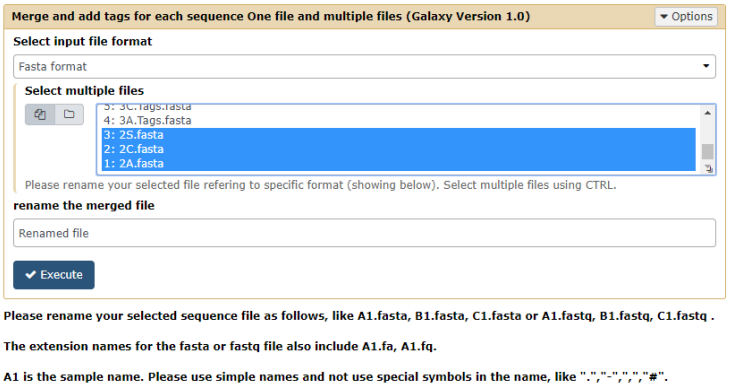
FASTA

Similar options as above FASTQ program.

6. Merge and add tags for each sequence

This tool is mainly used to merge multiple files and add a tag "--tag" for each sequence, which is required in this analysis pipeline. This is very helpful if you have multiple sample files and want to use this pipeline to conduct sequencing and statistical analysis. Please rename the file names like **A1.fasta, B1.fasta, C1.fasta or A1.fastq, B1.fastq, C1.fastq.**

Input:



Options:

Input file format: Fasta format / Fastq format

Output:

- Renamed file for fasta

```
>2A_Tag1--2A
TACGGAGGGTGCAAGCGTTGCTCGGAATTACTGGGCGTA
>2A_Tag2--2A
TACGGAGGATGCAAGCGTTATCCGGATTCATTGGGTTTA
>2A_Tag3--2A
CACCGGCGGCTCGAGTGGTAACCGTTATTATTGGGTCTA
>2A_Tag4--2A
CACCGGCAGCTCAAGTGGTGGCCATTTTTATTGGGCCTA
>2A_Tag5--2A
CACCGGCGGCTCGAGTGGTAACCGTTATTATTGGGTCTA
>2A_Tag6--2A
CACCGGCAGCTCAAGTGGTGGCCATTTTTATTGGGCCTA
```

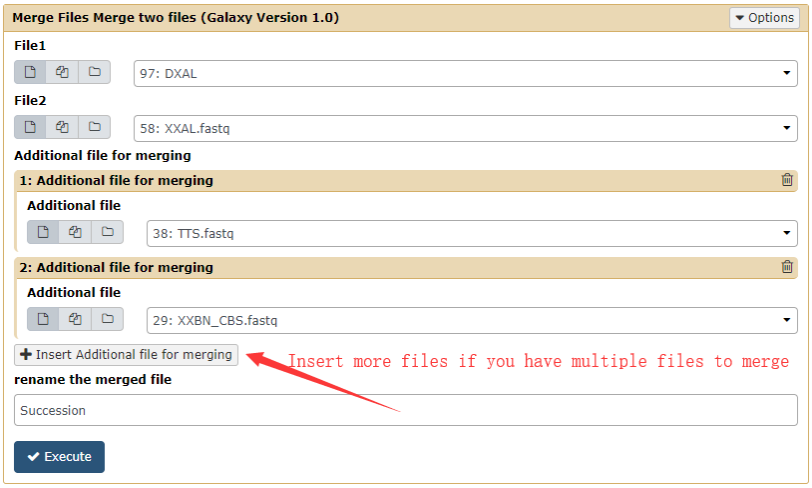
- Sequences numbers for each tag

2A	34925
2C	35278
2S	35599

7. Merge files

This tool is mainly used to merge multiple files into one file.

Input:



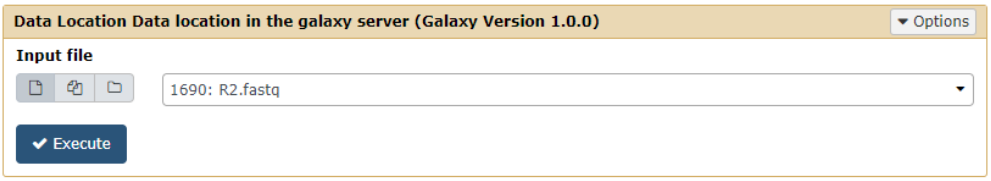
Output:

- Succession (If you put another name in the “rename the merged file”, it will show what you have fill in.)

8. Data location

This tool is mainly used to find the data location for certain dataset in the server. The data location is helpful to find the dataset for Galaxy adminstrators when you have problems.

Input:



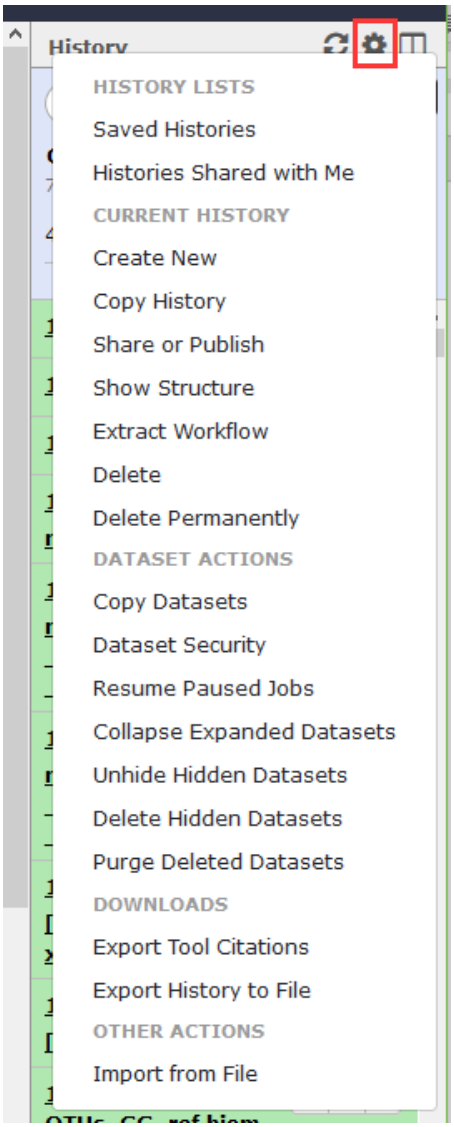
Output:

- **file_location.txt**

```
/newdatabase1/galaxy/user_data/datasets/000/105/dataset_105483.dat
```

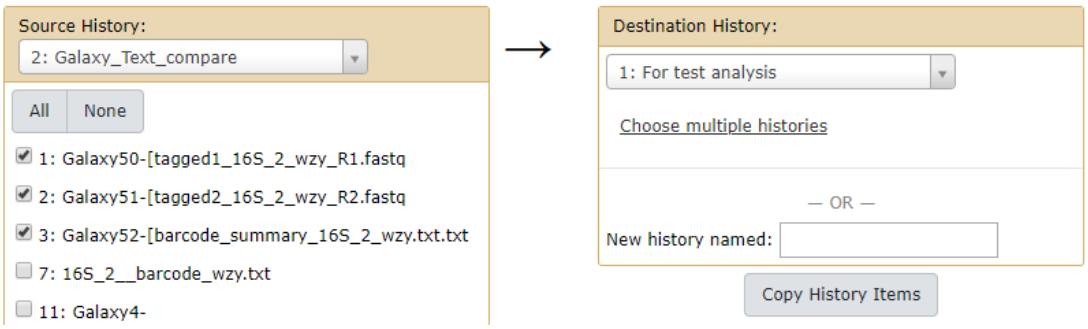
G. Operation tricks and common problem solutions

1. Basic operations in Galaxy

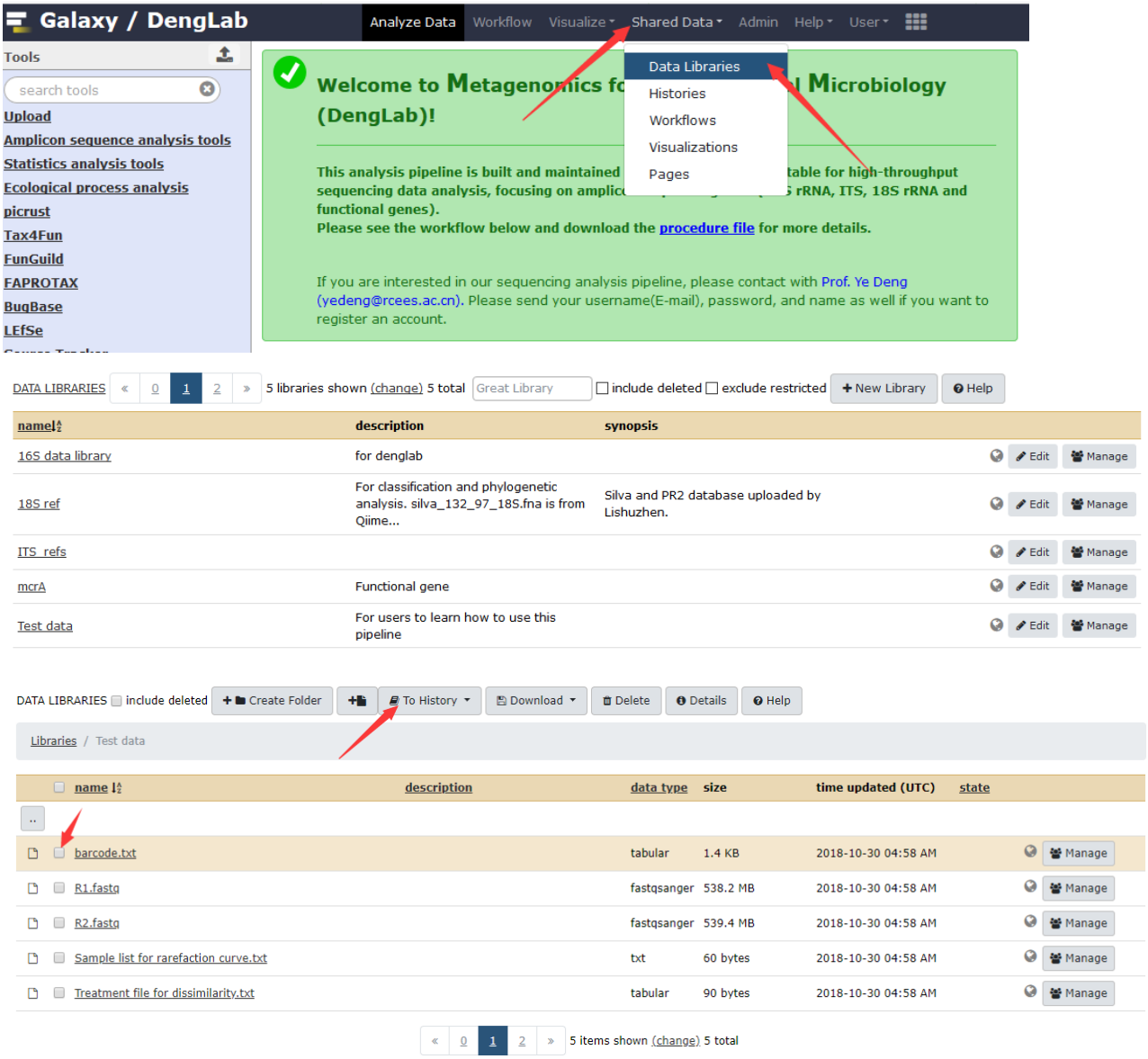


Please remember to choose “choose permanently” if you want to erase your history permanently, otherwise it will store into a temporary place and your quota will not decrease. See the below introduction for how to find temporarily deleted history.

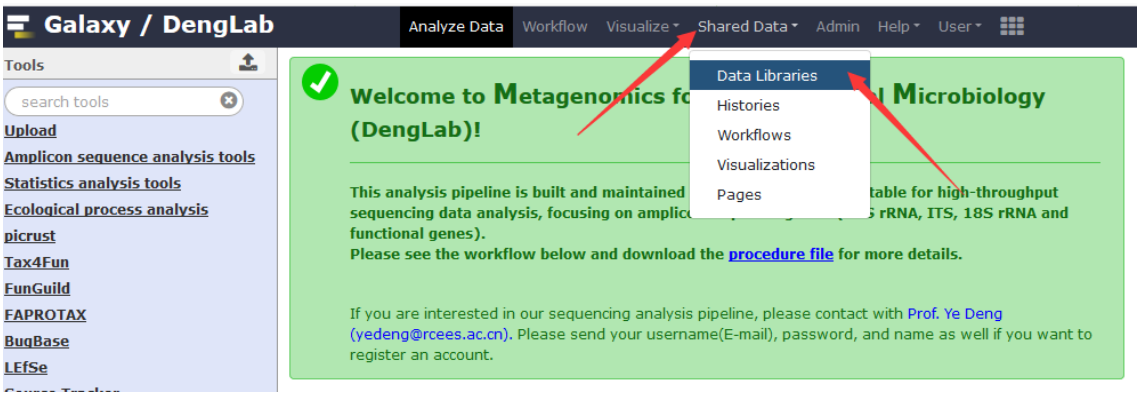
Copy datasets:

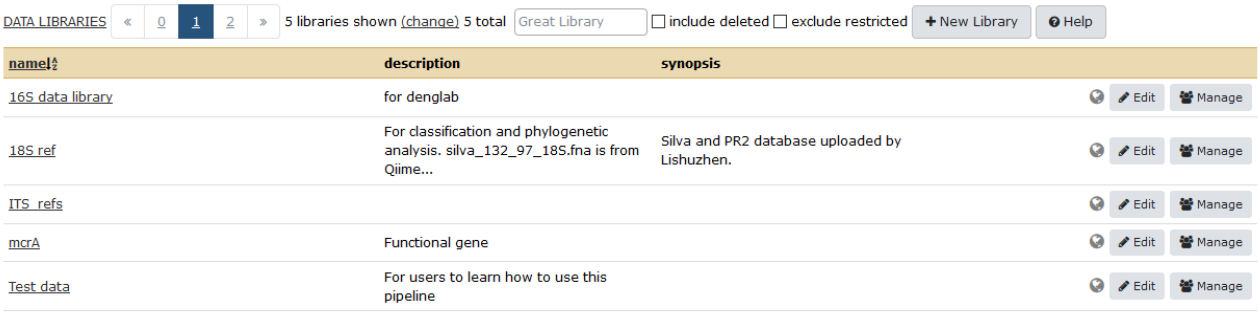


2. Shared test datasets

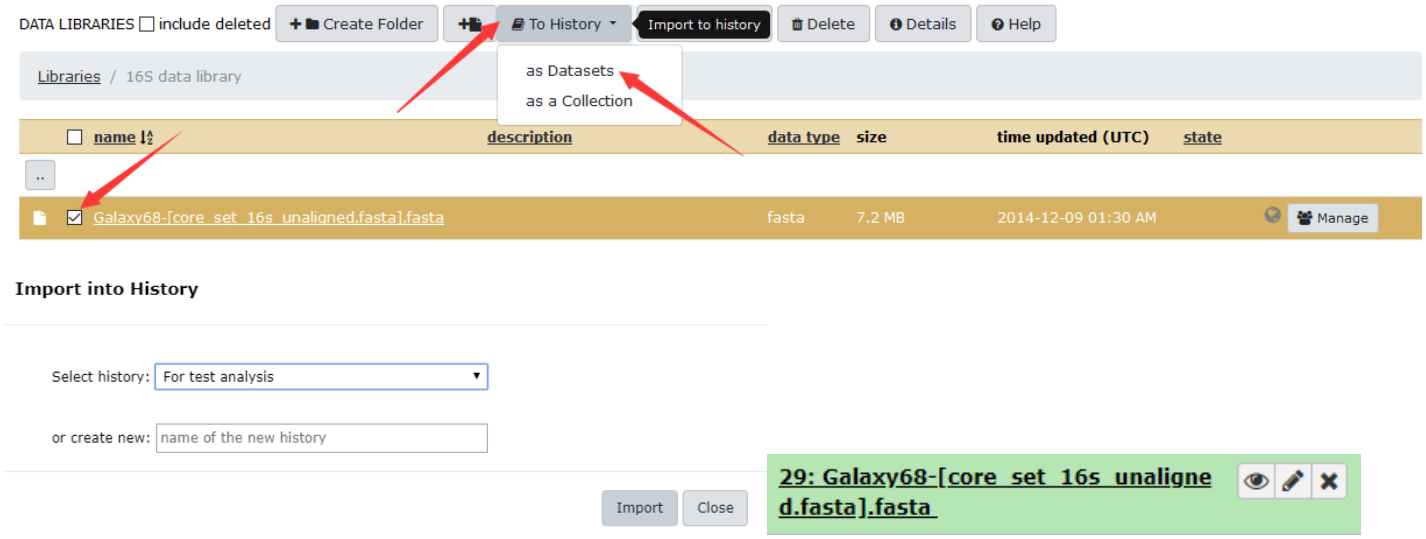


3. Shared libraries



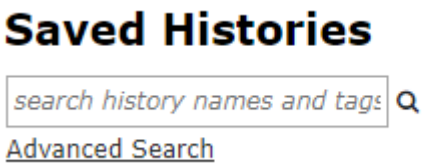


Currently, we provided three shared libraries for 16S, 18S, ITS and mcrA. Please find relevant datasets to import into your history.

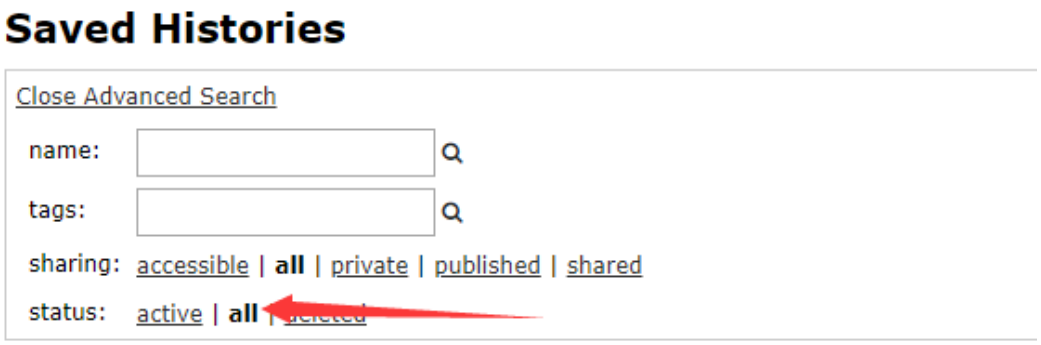


4. Dataset deletions

Select “saved history” and further choose “Advanced Search” button:



Choose “all” button to show all history that you have created in your account. And select the deleted history to further erase or retrieve.



5. Share historys to other users

Select “share or publish” of a certain history, then fill in the individual users: