

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

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

Updated

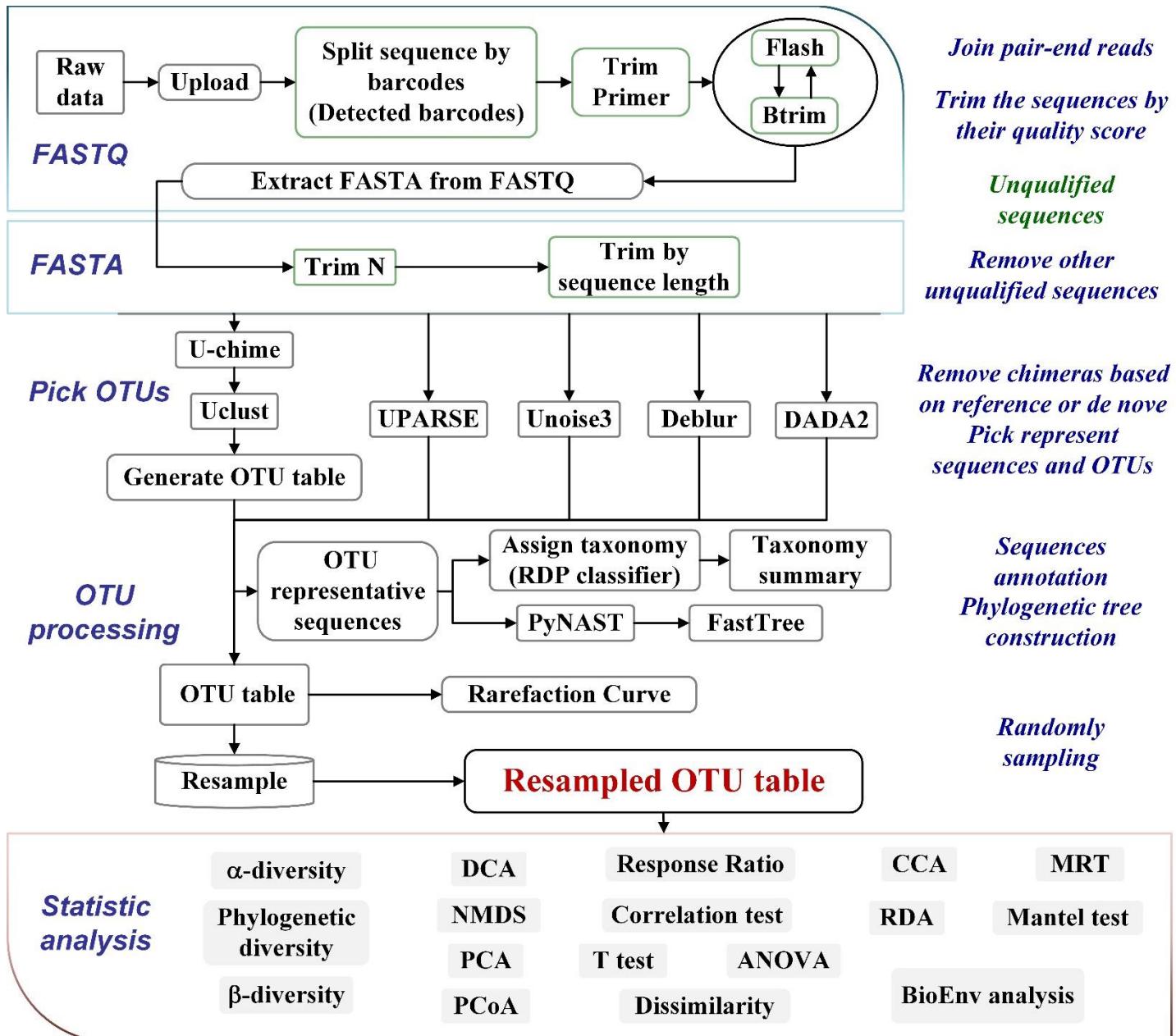
Nov. 2018

@ 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

A. Data Treatment and quantity control.....	1
1. Upload File	1
2. Detect barcodes(FASTQ).....	2
3. Trim Primer(FASTQ).....	2
4. Remove end base (FASTQ) (Optional).....	3
5. Flash (Combine R1 and R2) (FASTQ)	3
6. Btrim (FASTQ).....	4
7. Extract FASTA from FASTQ	5
8. Trim N (FASTA) (Trimming sequences containing “N”)	5
9. Trim by Sequence Length	6
10. Framebot (optional; only necessary for functional genes)	7
11. Generate OTU table	7
UPARSE (Recommend; UPARSE for FASTA).....	7
UNOISE (Unoise for FASTA to generate ZOTUs).....	8
UNOISE (Unoise for FASTA using Vsearch)	8
Deblur	9
Uclust	10
12. ITSx Extractor (Optional for ITS)	11
13. Compare otu table to sequence file (optional)	13
14. Rarefaction Curve.....	13
15. RDP Classifier	14
16. Resample OTU table	15
B. Statistics analysis	17
1. Diversity methods.....	17
1.1 α -diversity (Calculate taxonomy alpha diversity and evenness)	17
1.2 Hill number	17
1.3 β -diversity	18
2. Community structure.....	18
2.1 PCA	18
2.2 DCA.....	19
2.3 NMDS	19
2.4 PD&PCoA	20
2.5 Relative abundance.....	21

3. Comparison analysis	21
3.1 Response ratio calculation	21
3.2 Paired and unpaired t test.....	22
3.3 Dissimilarity (MRPP, adonis, anosim).....	23
4. Environmental associations	25
4.1 Correlation test	25
4.2 Multivariate Regression Tree (MRT)	26
4.3 BioEnv Analysis	26
4.4 CCA.....	27
4.5 Mantel Test.....	27
5. Plotting figures.....	28
5.1 Venn Diagrams	28
5.2 Heatmap	29
5.3 Hierarchical cluster	29
C. Ecological process analysis	30
1. Null model test	30
2. Null model test on Permdisp	30
3. Beta NTI calculation	31
4. RC distance	31
5. Summary ecological process.....	32
D. Functional profile prediction approaches	33
1. PICRUSt.....	33
1.1 Pick up ref OTU.....	33
1.2 Normalize by Copy Number	33
1.3 Predict Metagenome	33
1.4 Categorize by Function	34
Convert Biom to Tabular	34
2. Tax4Fun	35
2.1 Preparation for Tax4Fun	35
2.2 Tax4Fun.....	35
3. FAPROTAX.....	36
4. BugBase	37
4.1 Pick up ref OTU.....	37
4.2 BugBase Analysis.....	37
5. FunGuild	38

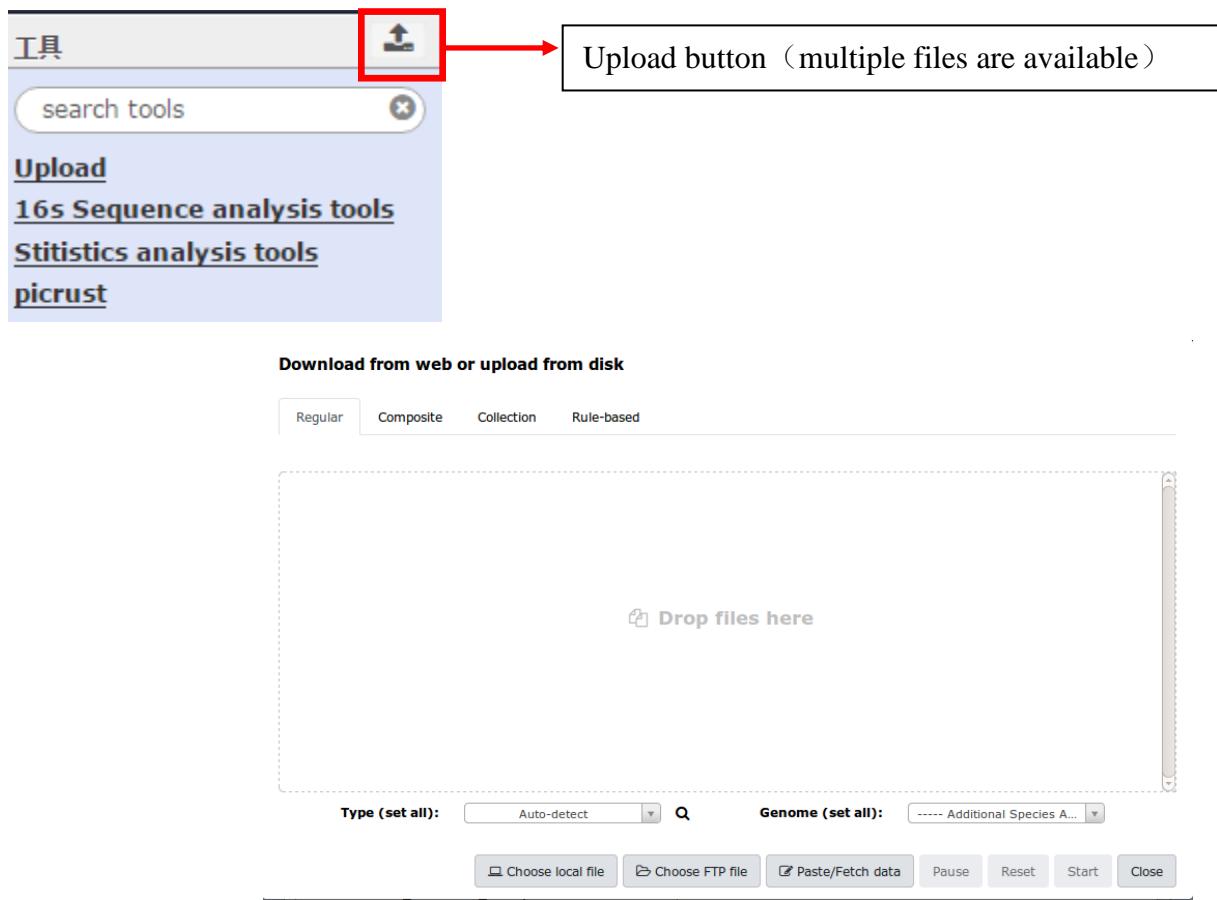
E. Other analysis tools in microbial ecology	39
1. LEfSe	39
1.1 Input prepare for LefSe analysis.....	39
1.2 A) Format Data for LefSe.....	39
1.3 B) LDA Effect Size (LEfSe)	40
1.4 C) Plot LEfSe Results	40
1.5 D) Plot Cladogram.....	40
1.6 E) Plot One Feature	41
1.7 F) Plot Differential Features.....	41
2. Source Tracker	41
F. Auxilliary tools in miscellaneous section	43
1. FastQC.....	43
2. Split files into separated samples	43
3. FASTQ format check	44
4. Length Statistics	44
5. Sequence number for each tag	45
6. Merge and add tags for each sequence.....	46
7. Merge files.....	46
8. Data location	47
G. Operation tricks and common problem solutions.....	48
1. Basic operations in Galaxy	48
2. Shared test datasets.....	49
3. Shared libraries	49
4. Dataset deletions	50
5. Share histories to other users	51

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 quantity control

1. Upload File

Upload the three sequencing data individually by Galaxy, barcodes file (txt) of your samples also need to be uploaded before downstream 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_barcode	forward_primer_515F	reverse_primer_806R	Primers
A1	AGCCAGTCATAC	GTTGGTTGGCAT	GTGCCAGCMGCCGCGTAA	GGACTACHVGGGTWTCTAAT	3
A2	AGCGAACCTGTT	TTCCACACGTGG	GTGCCAGCMGCCGCGTAA	GGACTACHVGGGTWTCTAAT	4
A3	GTTTGCTCGAGA	AACCCAGATGAT	GTGCCAGCMGCCGCGTAA	GGACTACHVGGGTWTCTAAT	5
A4	CAAACGCACTAA	GTAGTGTCAACA	GTGCCAGCMGCCGCGTAA	GGACTACHVGGGTWTCTAAT	6
A5	GAACAAAGAGCG	TGGAGAGGAGAT	GTGCCAGCMGCCGCGTAA	GGACTACHVGGGTWTCTAAT	7
A6	GCTAAGTGATGT	CGTATAATGCG	GTGCCAGCMGCCGCGTAA	GGACTACHVGGGTWTCTAAT	8
B1	AAGGGACAAGTG	AATACAGACCTG	GTGCCAGCMGCCGCGTAA	GGACTACHVGGGTWTCTAAT	9
B2	AGTGTGATTG	GACTCAACCAGT	GTGCCAGCMGCCGCGTAA	GGACTACHVGGGTWTCTAAT	10
B3	CTATTAAGCGGC	GGAAAGAAGTAGC	GTGCCAGCMGCCGCGTAA	GGACTACHVGGGTWTCTAAT	11
B4	GAGTCCGTTGCT	ACACCGCACAAAT	GTGCCAGCMGCCGCGTAA	GGACTACHVGGGTWTCTAAT	13
B5	GATAACTGTACG	GTCTCCTCCCTT	GTGCCAGCMGCCGCGTAA	GGACTACHVGGGTWTCTAAT	14

2. Detect barcodes(FASTQ)

Inputs:

Detect barcodes (FASTQ) This script performs demultiplexing of Fastq sequence data where forward and reverse barcodes are contained in two separate fastq files respectively (Deng lab improved) (Galaxy Version 1.0.0)

Sequence file 1 (FASTQ)
 2: R1.fastq

Sequence file 2 (FASTQ)
 3: R2.fastq

Sample list
 1: barcode.txt

maximum number of errors in barcode

Options

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

Outputs:

6: tagged1_R1.fastq

 7: tagged2_R2.fastq

 8: barcode_summary.txt

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

Trim Primer Remove primers at the beginning of sequences (

Sequence file
 6: tagged1_R1.fastq

Primer
 GTGCCAGCMGCCGCGTAA

Allowed mismatch

Maximum starting position

Reverse

Trim Primer Remove primers at the beginning of sequences

Sequence file
 7: tagged2_R2.fastq

Primer
 GGACTACHVGGGTWTCTAAT

Allowed mismatch

Maximum starting position

Parameter options:

Primer used:

515F GTGCCAGCMGCCGCGTAA (R1)

806R GGACTACHVGGGTWTCTAAT (R2)

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

515MF GTGYCAGCMGCCGCGTAA (R1)

806MR GGACTACNVGGGTWTCTAAT (R2)

gITS7F GTGARTCATCGARTCTTG (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

[10: TrimPrimer_summary.txt](#)

[9: TrimPrimer_tagged_1_R1.fastq.fastq](#)

Reverse

[12: TrimPrimer_summary.txt](#)

[11: TrimPrimer_tagged_2_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:

Remove end base (FASTQ) Remove some low quality bases at end side for fastq file (Galaxy Version 1.0.0)

fastq file required to remove base

Remove length

Execute

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

Flash Pair-end joining program for FASTQ (Galaxy Version 1.0.0)

Sequence file 1 (FASTQ)
 9: TrimPrimer_tagged1_R1.fastq.fastq

Sequence file 2 (FASTQ)
 11: TrimPrimer_tagged2_R2.fastq.fastq

The minimum required overlap length (bp)

The maximum overlap length expected in approximately 90% of read pairs(bp)

The maximum allowed ratio of the number of mismatches and the overlap length

phredOffset

phredOffset is the smallest ASCII value of the characters used to represent quality values of bases in fastq files. It should be set to either 33, which corresponds to the later Illumina platforms and Sanger platforms, or 64, which corresponds to the earlier Illumina platforms.
Default: 33.

Average reads length

avg_frag_length

standard deviation of fragment lengths

If you do not know standard deviation of the fragment library, you can probably assume that the standard deviation is 10% of the average fragment length.

Execute

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 approximately 90% of read pairs (bp): 250 ($250 = 220 \times 2 - 253 + 25 \times 2.5$)

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: $25 = \text{avg_frag_length} * 10\%$

Outputs:

<u>20: FlashHist.txt</u>
<u>19: notCombined_TrimPrimer_tagged2_R2.fastq.fastq</u>
<u>18: notCombined_TrimPrimer_tagged1_R1.fastq.fastq</u>
<u>17: Combined.fastq</u>

6. Btrim (FASTQ)

Inputs:

Btrim Trimming tool for FASTQ (Galaxy Version 1.0.0)

Sequence file(FASTQ)
 17: Combined.fastq

Format
Sanger

Average Quality Score
20

Minimum Length
140

Window Size
5

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 window 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 discarded.)

Outputs:

[22: length_distribution.txt](#)
[21: Trimmed_Combined.fastq](#)

7. Extract FASTA from FASTQ

Transforming FASTQ file (Trimmed_combined.fastq) to FASTA file

Inputs:

Extract FASTA from FASTQ Extract sequences in FASTA format from FASTQ file without quality scores.
(Galaxy Version 1.0.0)

FASTQ file
 21: Trimmed_Combined.fastq

Execute

Outputs:

[23: {Trimmed_Combined.fasta}.fasta](#)

8. Trim N (FASTA) (Trimming sequences containing “N”)

Trimming sequences containing “N”

Inputs:

Trim N The program delete the sequences contains N or trim them (remove the bases after N) (Galaxy Version 1.0.0)

Fasta file



23: {Trimmed_Combined.fastq}.fasta

If the sequence contains N

Delete

'Delete' means to remove the entire sequence; 'Trim' means to remove the bases after N

Trim by length (bp)

200

Execute

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:

[25: Remove_N.sum](#)

[mary.txt](#)

[24: Remove_N.fasta](#)

9. Trim by Sequence Length

Inputs:

Trim by Sequence Length The program trim the sequences based on the length. Only sequences longer than the minimum length and shorter (or equal) than the maximum length are kept (Galaxy Version 1.0.0)

File format

FASTA

FASTA(4 54)



24: Remove_N.fasta

Minimum length (bp)

245

Trim_or_delete

Trim sequence base over the maximum length

Maximum length (bp)

260

Execute

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:

[27: Trim_length_summary.txt](#)

[26: Trim_length](#)

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

Inputs:

FrameBot parellel (version 1.0.0)

Sequences: 1: Trim_length_sul1_1

References: 14: sul1DB201606-Framebot.fasta

Execute

Parameter options:

References: unaligned_protein.fasta (own database without alignment)

Outputs:

[36: FrameBot_log.txt](#)

[35: frameBot_corr_prot.fasta](#)

[34: frameBot_corr_nucl.fasta](#)

11. Generate OTU table

UPARSE (Recommend; UPARSE for FASTA)

Inputs:

UPARSE for FASTA Clustering method to generate OTU for FASTA format, without quality trimming (Galaxy Version usearch v7.0.1001_i86linux64)

FASTA file with sample ID
 26: Trim_length

Reference sequence for chimera checking
 29: Galaxy68-[core_set_16s_unaligned.fasta].fasta

Trim length

 Sequences will all be trimmed by this length. Fill 'N/A' if you don't want to trim the sequences.

Clustering threshold

Majority rule, cut=?

 Example: 1 means remove all the singlet

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):

[32: UPARSE otu sequence names.txt](#)

[31: UPARSE rep seq.fasta](#)

[30: UPARSE otu table.txt](#)

UNOISE (Unoise for FASTA to generate ZOTUs)

Inputs:

Unoise for FASTA to generate ZOTUs Unoise give all the correct biological sequences in the reads (ZOTUs) (Galaxy Version usearch10) ▼ Options

FASTA file with sample ID
 26: Trim_length

The minimum abundance
 8

Input sequences with lower abundances are discarded. For example, 1 means keep all sequences, 2 means remove all the singleton. Default is 8.

Execute

Parameter options:

The minimum abundance: 8 (default value)

Outputs:

[41: Unoise otu seq.fasta](#)

[40: Unoise3 otu table.txt](#)

[39: Unoise3 seqs summary.fasta](#)

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)

Unoise for FASTA using Vsearch Based on Vsearch (Galaxy Version 2.7.2) ▼ Options

FASTA file with sample ID

 26: Trim_length

The minimum abundance
 8
 Input sequences with lower abundances are discarded. For example, 1 means keep all sequences, 2 means remove all the singleton. Default value of Unoise3 algorithm is 8.

Alpha parameter
 2
 Default is 2. Generally, this is not recommended to change.

Threshold for mapping
 0.97
 Denoised OTUs also use a 0.97 identity threshold by default to allow for sequencing and PCR error. This value varied from 0 to 1.

Deblur

Inputs:

Deblur A novel sub-operational-taxonomic-unit (sOTU) approach (Galaxy Version 1.0.4) ▼ Options

FASTA files with sample ID

 26: Trim_length

Trim length
 240
 Sequence trim length. All reads shorter than this value will be discarded. A value of -1 can be specified to skip trimming, but this assumes all sequences have an identical length.

Minimum reads
 10
 Keep only the sequences which appear at least min-reads study wide. Zero is to ignore this parameter and minimum value can be set as 1.

Minimum size
 2
 Keep only sequences which appear at least min-size times per-sample.

Positive reference filtering database
 Use default database (Greengene 13_8)
 Select your own database

Parameter options:

Trim length: 240 (change based on the instructions)

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:

U-Chime Detect chimeras (Galaxy Version USEARCH 5.2.32) Options

Sequence file
 26: Trim_length
 The input sequences should include identical sequences

select
 Reference database

Reference sequence file
 29: Galaxy68-[core_set_16s_unaligned.fasta].fasta
 The reference database should contain trusted sequences that are chimera-free (nucleotide sequences)

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:

uclust Clustering method to generate OTU, fast (Galaxy Version usearch 5.2.32) Options

Sort by length?
 Yes

Sequence file
 42: Uchime.fasta
 FASTA file Dataset collection

Clustering threshold
 0.97

Parameter options:

Clustering threshold: 0.97

Outputs:

[45: Uclust clustering.txt](#)

[44: Uclust seeds.fasta](#)

11.5-3 Generate OTU Table

Inputs:

Generate OTU table Generate OTU tables from CD-HIT and output clustering file(s). [Options](#)

Combine the identical sequences removed before if applicable. (Galaxy Version 1.0.0)

Sequence File (FASTA)
 26: Trim_length
 used to identify all samples and pick representative sequences

Clustering method
 Uclust
 CD-HIT/Uclust

Keep forward and reverse tags (454)
 No
 Yes - two columns for two tags; No - one column

Did you combine forward and reversed sequences (454)
 No
 If you did, you need to provide three cluster files to generate the table

Clustering file from CD-HIT/UCLUST
 45: Uclust_clustering.txt

Redundancy map
 43: redundancy_map.txt

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:

ITSx Extractor ITSx -- Identifies ITS sequences (Galaxy Version 1.1b)

Input Fasta
 616: Galaxy8-[ITS_UPARSE_rep_seq.fasta].fasta

Domain E-value Cutoff
 1e-05
 Domain E-value cutoff a sequence must obtain in the HMMER-based step to be included in the output.

Domain Score Cutoff
 0
 Domain score cutoff that a sequence must obtain in the HMMER-based step to be included in the output.

Minimum Number of Domains
 2
 The minimum number of domains (different HMM gene profiles) that must match a sequence or it to be included in the output (detected as an ITS sequence). Setting the value lower than two will increase the number of false positives, while increasing two will decrease ITSx detection abilities on fragmentary data.

HMMER Search Type
 Search E-value
Search E-value
 0.01
 The actual E-value cutoff used in the HMMER search. High numbers may slow down the process. Should never be set to a than the Domain E-value Cutoff option. Cannot be used in combination with Search Score option.

Re-creates the HMM-database before ITSx is run
 Yes No

Allow profiles not to be in the expected order on the extracted sequences
 Yes No

Check both DNA strands against the database
 Yes No

Use HMMER's heuristic filtering
 Yes No

Preserve sequence headers instead of printing out ITSx headers
 Yes No

Remove ends of ITS sequences if they are outside of the ITS region
 Yes No

Options:

Usually not required 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
  Stramenopiles: 1
  Tracheophyta: 79
-----
ITSx run finished at Sat Apr  7 12:00:56 2018.
```

➤ Identified Fungi ITS fasta File

```
>OTU_1
AACGCACATGCGOCCTGGTATTCAGGggCATGCGCTTTCAGGTCATTTCCTCAACATCTCTT
TGTtttttttCAGAGagAGTTCTCGCTCTGGAGTATAATGCAAGTAGGTCGTTTAAAGTT
agCGTCTAGGCGAACATGTTAAAGTTGACCTCAAAATCAGGTAGGAGTACCGCTGAACTTAA
>OTU_2
AACGCACCTGCGCTctctGGTATTCGGAGAGCATGCGCTTTCAGTACGTTATCATGAAATCTCAACCATAGG
GCGTGTTAAGTTGCTATCTGGGTTAAAGTTGCGCTGGTACGACTTGAGAAGTGGCTTCTAA
A
>OTU_3
AACGCACCTGCGCTctctGGTATTCGGAGAGCATGCGCTTTCAGTACGTTATCATGAAATCTCAACCATAGG
AGCGTCTTAACATATGCTATCTGGGTTAAAGTTGCGCTGGTACGACTTGAGAAGTGGCTTCTAA
A
A
```

➤ ITSx results associated with other eukaryotes

ITSx result summary

The file of [ITS_identify.ITS1.fasta](#) is available to download.

The file of [ITS_identify.ITS2.fasta](#) is available to download.

The file of [ITS_identify.extraction.results](#) is available to download.

The file of [ITS_identify.extraction_Alveolates.fasta](#) is available to download.

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:

Compare otu table to sequence file Quickly compare the otu table file to the sequence file or reversely. (Galaxy Version 1.0.0) ▼ Options

fasta file
 31: UPARSE_rep_seq.fasta

OTU table
 32: UPARSE_otu_sequence_names.txt

Outputs:

- [52: normalized_summary.txt](#)
- [51: normalized_UPARSE_otu_sequence_names.txt](#)
- [50: normalized_UPARSE_rep_seq.fasta](#)

14. Rarefaction Curve

① Richness rarefaction

Inputs:

Rarefaction Curve Generate rarefaction curve from OTU table. (Galaxy Version 1.0.0) ▼ Options

OTU table
 30: UPARSE_otu_table.txt

Sample list
 Nothing selected

Richness Shannon,simpson
Richness rarefaction

Steps in rarefaction calculations. Suggested:1000 as default value
1000

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:

Rarefaction Curve Generate rarefaction curve from OTU table. (Galaxy Version 1.0.0) ▼ Options

OTU table
 30: UPARSE_otu_table.txt

Sample list
 Nothing selected

Richness Shannon,simpson
Shannon and Simpson rarefaction

Maximum reads in these samples, please fill hundreds or thousands according to your steps,e.g.
41000 correponding to 1000 steps
40000

Steps in rarefaction calculations. Suggested:1000
100 (slower)

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:

RDP Classifier Assign 16S rRNA or Fungal LSU sequences to the bacterial and fungal taxonomy (Galaxy Version 1.0.0) ▼ Options

Sequences to classify: (FASTA)
 31: UPARSE_rep_seq.fasta

gene
16S rRNA (RDP training set RDP release 11.5)

conf
0.5

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:

gene

Select your own provided taxonomy database

Bacteria/Archaea or Fungi

Eukaryote (fungal, protist, etc)

FASTA file containing a unique name for each sequence

1645: sh_refs_qiime_ver7_97_s_01.12.2017.fasta

Text file containing assigned taxonomy for each sequence name

1646: sh_taxonomy_qiime_ver7_97_s_01.12.2017.txt

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
 - CCGAACTGTCGACAGAGTTGGCTGCCCTCAACAGGGGGCATGTGCACA
 - >SH492954.07FU_DQ656654_reps_singleton
 - CATGAGCCTTGATCTCGCCGTTAACAGAGGCCCTACGGTCCGGGGTAATCT
 - >SH628622.07FU_LC131409_reps
 - TCGAAGAACACACTTCTCCAACCCCTGTGAACCGTCGTCGAGCATGATGCTCGGACGCTCCATACACT
 - ...
- 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:

[60: ClassifierSummary.txt](#)

[59: Classifier of 16srrna.txt](#)

[16. Resample OTU table](#)

Inputs:

Resample OTU table Randomly resample for each tag/sample from OTU table (Galaxy Version 1.0.0)

OTU table to be resampled

30: UPARSE_otu_table.txt

Resample size

33007

Execute

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:

Calculate taxonomy alpha diversity and evenness Calculate taxonomy alpha diversity (Shannon, Simpson) and evenness (Pielou evenness, Simpson evenness) (Galaxy Version 1.0.0)

File1(tabular file)
61: resample_UPARSE_otu_table.txt

whether you want ggplot2 plots for alpha-diversity index
No

Parameter options:

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

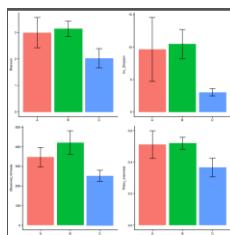
A1 A
A2 A
A3 A
A4 A
A5 A
A6 A
B1 B
B2 B
B3 B
B4 B
B5 B
B6 B

Outputs:

63: alpha diversity report

19 lines
格式: tabular, 数据库: ?

1	2	3	4	5	6
""	"Shannon"	"Inv_Simpson"	"Observed_richness"	"Pielou_evenness"	"Simpson_evenness"
"C3"	2.41778708952995	3.98882178969163	281	0.428810749114447	0.0141950953369809
"A4"	3.54611218566935	13.3270376921694	393	0.593609842975853	0.033911037384655
"C5"	1.6863675501679	2.72543151106386	213	0.314544982449162	0.0127954531035862
"C1"	2.29284476382483	3.33508036713978	286	0.405383324537111	0.0116611201648244



1.2 Hill number

或者根据 Chao 的最新文章计算 Hill number:

The screenshot shows a software interface titled "Hill number estimation (version 1.0.0)". At the top, it says "OTU table for Hill number calculation: 432: FunGuild result for resample_otu_table_soil_ITS_6_group.txt". Below this is a "Execute" button. The main area contains the following text:
What it does
The program calculates the Hill number based on abundance data of OTU table.
Hill numbers include the three widely used species diversity measures as special cases: Species richness(q=0), Shannon diversity (q=1), and Simpson diversity (q=2).

Hill numbers

q=0: Hill number is simply species richness, which counts species equally without regard to their relative abundances.

q=1: q tends to 1 is the exponential of the Shannon index, referred to as Shannon diversity.

q=2: Simpson diversity is the inverse of Simpson concentration index.

1.3 β-diversity

Inputs:

The screenshot shows a software interface titled "Calculate distance indexes (Jaccard, Bray, Horn, Euclidean) (Galaxy Version 1.0.0)". It has a "File1(tabular file)" section with a file input field containing "61: resample_UPARSE_otu_table.txt". Below it is a "data type" section with a dropdown menu set to "quantitative data". A note says "Recommend:quantitative data". At the bottom is a "Execute" button.

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:

The screenshot shows a list of output distance metrics:
69: Euclidean
68: Horn-Morisita
67: Bray-Curtis
66: Jaccard

2. Community structure

2.1 PCA

Inputs:

Principal Component Analysis Principal Component Analysis (Galaxy Version 1.0.0)

File_in(tabular file)



61: resample_UPARSE_otu_table.txt

whether you have different groups of sample (for ggplot2)

No

Outputs:

[71: pca plot](#)

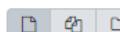
[70: pca result](#)

2.2 DCA

Inputs:

Detrended Correspondence Analysis Detrended Correspondence Analysis (Galaxy Version 1.0.0)

File_in(tabular file)



61: resample_UPARSE_otu_table.txt

whether you have different groups of sample (for ggplot2)

No

Outputs:

[73: dca plot](#)

[72: dca result](#)

2.3 NMDS

Inputs:

Non-metric multidimensional scaling. (NMDS) Non-metric multidimensional scaling (Galaxy Version 1.0.0)

File_in(tabular file)



61: resample_UPARSE_otu_table.txt

distance_type

Jaccard distance

dimension_type

Two dimensions

whether you have different groups of sample(for ggplot2)

No

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:

PyNAST Alignment Use PyNAST to align sequences. Pre-aligned reference sequences are required
1.0.0)

Sequence to be aligned (FASTA)
 31: UPARSE_rep_seq.fasta

Aligned reference sequences
16S - GreenGene

Minimum length to include in the alignment
200

Parameter options:

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

Outputs:

76: PyNAST_aligned

2) FastTree

Inputs:

FastTree Use FastTree to construct phylogenetic tree based on aligned sequences

Sequence to be aligned (FASTA)
 76: PyNAST_aligned

Execute

Outputs:

3) UniFrac

Inputs:

Unifrac A tool for comparing microbial community diversity in a phylogenetic context (Galaxy Version 1.0.0) ▼ Options

OTU table
 61: resample_UPARSE_otu_table.txt

Tree file
 77: FastTree.nwk

Outputs:

Weighted PcoA and unweighted PcoA are the results of PCOA.

2.5 Relative abundance

Inputs:

Taxonomy summary and relative abundance plot Taxonomy summary for each sample at selected similarity cutoff (Galaxy Version 1.0.0) ▾ Options

Resample OTU table
61: resample_UPARSE_otu_table.txt

Sample list
Nothing selected

Sample list for all samples grouping

OTU classification result from rdp classifier
59: Classifier of 16srna.txt

Count species richness or count species abundance
 Species richness
 Species abundance

Summary result type for each sample
 Numbers
 Percentage

Select which level to calculate result
#4:Class

Do not select first column, #1:ID.

No. of species showing in the plot
0

0 means all species would be shown in the relative abundance plot

Execute

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 be 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:

Response Ratio Calculation (Galaxy Version 1.0.0)

OTU table file
 61: resample_UPARSE_otu_table.txt

Treatments:
 Select/Unselect all
 #2:C3 #4:C5 #5:C1 #14:C6 #16:C2
 Do not select #1:ID.

Controls:
 Select/Unselect all
 #3:A4 #7:A5 #9:A3 #12:A2 #13:A1 #19:A6 |
 Do not select #1:ID

Blanks treatment:
 considered as missing values (excluded from the analysis)
 considered as 0 (included)
 fill 0 if paired with a valid value, then calculate avg and SD on whole set
 fill 0 if paired with a valid value, get average on each sample and then calculate avg and SD of all samples
 fill 0 if paired with a valid value, get sum on each sample and then calculate avg and SD of all samples

Confidence interval:
 90
 95
 99

Draw significant genes in the plot only
 Yes No

Execute

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:

91: Response ratio result

90: Response ratio plot

3.2 Paired and unpaired t test

Inputs:

Paired and unpaired t test Paired or unpaired t test (Galaxy Version 1.0.0)

OTU/Gene table(tabular file)/Data for t test with replicates
 61: resample_UPARSE_otu_table.txt

Group file to separate the samples into two groups
 86: Treatment file for t test.txt
 Please follow the example below for paired and unpaired t test

Paired or Unpaired
 Unpaired t test
 Paired t test

A character string specifying the alternative hypothesis
 Two sided (default)
 Greater
 less

Taxonomic file/Category file or not
 No
 Yes
 Please pay attention to the file format.

Execute

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:

Calculate dissimilarity Calculate dissimilarity using MRPP,ANOSIM,PERMANOVA (Galaxy Version 1.0.0)

File in(resampled OTU table)
 61: resample_UPARSE_otu_table.txt

Sample list(tabular file)
 62: Treatment file for dissimilarity.txt

Distance method
 Bray-Curtis distance
 Jaccard distance

How many groups types have you uploaded
1

Calculate the dissimilarity of different taxonomy group
No

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 taxonomy 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
Camylobacteriales	2.8721	0.002	0.1754

4. Environmental associations

4.1 Correlation test

Inputs:

Correlation test Correlations between community data and environmental variables (Galaxy Version 1.0.0) ▼ Options

Community data/Data matrix (tabular separated file)
 61: resample_UPARSE_otu_table.txt

Comparison way:
 Genes vs Environment data
 Among genes/factors
 Among samples

The option of among_genes is not recommended for large genes/OTUs.

Environmental variables(tabular file)
 93: Env file for test samples.txt

Correlation method:
 Pearson Correlation
 Spearman's ranked correlation
 Kendall's ranked correlation

Standardization method:
 standardize environmental data only (scale each factor to zero mean and unit variance)
 standardize genes and environmental data (scale each factor to zero mean and unit variance)
 divide by maximum (both genes and env)
 divide by maximum and multiply by the number of non-zero items (both genes and env)
 standardize values into range 0...1 (both genes and env)

If you only have one data file, please neglect the clues for environmental data.

Missing values in Genes/OTUs
 fill with 0 (before standardization)
 fill with 0 (after standardization)

This option is only available for data matrix file, not for environmental data.

Adjust P-values for multiple comparisons
 None
 Bonferroni correction (P-values are multiplied by the number of comparisons)
 Holm (1979) A simple sequentially rejective multiple test procedure
 Hochberg (1988) A sharper Bonferroni procedure for multiple tests of significance
 Hommel (1988) A stagewise rejective multiple test procedure based on a modified Bonferroni test
 False discovery rate (1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing)
 Benjamini and Yekutieli (2001) The control of the false discovery rate in multiple testing under dependency

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	Sample	factor Name1	factor Name2	factor Name3	...	factorN
factor1	data11	data12	...	data1N	Sample1	data11	data12	data13	...	data1N
factor2	data21	data22	...	data2N	Sample2	data21	data22	data23	...	data2N
factor3	data31	data32	...	data3N	Sample3	data31	data32	data33	...	data3N

Correlation method: Pearson / Spearman / Kendall

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:

Multivariate Regression Tree MRT analysis (Galaxy Version 1.0.0) ▼ Options

OTU table(tabular file)/Community data
 61: resample_UPARSE_otu_table.txt

Environmental variables(tabular file)
 93: Env file for test samples.txt

Selected factors in MRT analysis:
 Select/Unselect all
 #2:pH #3:H2 #4:Ce #5:H2Rec #6:EneRec #7:SubRec |
 Do not select first column #1:

Splitting times:
 Determined by the best tree from cross-validation
 Give the best tree within one SE of the overall best (high confidence)
 Split to specific groups

Dissimilarity measures before splitting:
 Do not calculate distance before splitting
 Euclidean distance
 Bray-Cutis distance
 manhattan distance

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:

BioEnv Analysis Best Subset of Environmental Variables with Maximum (Rank) Correlation with Community Dissimilarity (Galaxy Version 1.0.0) ▼ Options

OTU table(tabular file)/Community data
 61: resample_UPARSE_otu_table.txt

Environmental variables(tabular file)
 93: Env file for test samples.txt

Dissimilarity index:
 euclidean bray

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

Parameter options:

Dissimilarity index: euclidean / bray

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

Outputs:

	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:

Canonical Correspondence Analysis Canonical Correspondence Analysis (Galaxy Version 1.0.0) ▼ Options

File_in(tabular file)
61: resample_UPARSE_otu_table.txt

Env_file(tabular file)
93: Env file for test samples.txt

Execute

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:

Mantel and partial mantel test Default or user defined calculation (Galaxy Version 1.0.0) ▼ Options

File_in (tabular file)
61: resample_UPARSE_otu_table.txt
Resample OTU table or functional gene table

Env_file (environmental factors in tabular format)
93: Env file for test samples.txt

Include geographic information with latitude and longitude
Not contain latitude and longitude

calculate the partial mantel test according to your selection (User defined)
Default Calculation

Execute

For user-defined partial mantel test:

Mantel and partial mantel test Default or user defined calculation (Galaxy Version 1.0.0)

File_in (tabular file)
 61: resample_UPARSE_otu_table.txt
 Resample OTU table or functional gene table

Env_file (environmental factors in tabular format)
 93: Env file for test samples.txt

Include geographic information with latitude and longitude
 Not contain latitude and longitude

calculate the partial mantel test according to your selection (User defined)
 User defined for partial mantel test

Included environmental factors
 Select/Unselect all
 #2:pH
 Please use CTRL to select multiple factors. Do not select #1:ID.

Excluded environmental factors
 Select/Unselect all
 #3:H2 #4:Ce #5:H2Rec #6:EneRec #7:SubRec #8:RT
 Please use CTRL to select multiple factors. Do not select #1:ID

Output:

102: mantel test and partial mantel test report

5. Plotting figures

5.1 Venn Diagrams

Venn Diagrams Draw venn diagrams based on OTU table (Galaxy Version 1.0.0)

OTU table file
 61: resample_UPARSE_otu_table.txt

Sample list (Optional)
 Nothing selected

2nd sequence file
 4 categories

A vector of numbers to indicate colors from 1 to 100
 20,1,50,70,90
 The numbers should be separated by comma and this value must correspond to selected categories.

Execute

Input Format

- OTU table:
 The OTU table should contain one head row starting 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

Plotting heatmap Generating heatmap image (Galaxy Version 1.0.0) ▼ Options

Data with headers and row names for heatmap plot:

61: resample_UPARSE_otu_table.txt

Standardization method before plotting heatmap:

Nothing to do with the data
 Standardize the data (scale each factor to zero mean and unit variance)
 divide by maximum
 divide by maximum and multiply by the number of non-zero items
 standardize values into range 0...1

The values should be centered and scaled in either the row direction or the column direction, or none.

Row direction
 Column direction
 None

If rows should be clustered or hclust object

TRUE
 FALSE

If columns should be clustered or hclust object

TRUE
 FALSE

Distance method if selected clusters

The complete linkage method finds similar clusters.
 Average (UPGMA)
 Mcquitty (WPGMA)
 Median (WPGMC)
 Centroid (UPGMC)

5.3 Hierarchical cluster

Hierarchical Cluster Hierarchical clustering analysis with heatmap (Galaxy Version 1.0.0) ▼ Options

Community data/Data matrix (tabular separated file)

61: resample_UPARSE_otu_table.txt

Data preparation (on each sample):

None
 Standardization (scale to zero mean and unit variance)
 Taking logarithm ($\log(x)+1$; log base is 2)
 Making sum of squares equal to one

Distance Method:

Pearson Correlation
 Bray-Curtis distance
 Euclidean distance
 Maximum

Clustering Algorithm:

Average
 Complete
 Median
 Centroid

Figure option:

Simple hierarchy tree on samples
 Heatmap with hierarchy tree on samples only
 Heatmap with hierarchy tree on both genes and samples (Not recommend for more than 500 genes or OTUs)

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:

Null model test Effect size et al. (Galaxy Version 1.0.0) ▼ Options

OTU table file
 ▼

Group file (treatment list)
 ▼

Distance method:
 Jaccard distance
 Bray-Curtis distance

Data transformation:
 No transformation
 Transfer to the presence/absence data
 Transfer to the round integers

Null model:
 Chase 2010 EcoSim null model (Randomize community data matrix with the independent swap algorithm (Gotelli 2000) maintaining both row and column sums constant)
 Chase 2011 Ecosphere null model (keep alpha and gamma diversity of the whole/group data constant)
 Randomize community data matrix abundances within species (only keep column sum constant)

Keep gamma diversity in:
 Total dataset
 Group by group

✓ Execute

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:

Null model test on Permdisp (Galaxy Version 1.0.0)

OTU table file
 1258: Uparse_OTU_table_resample.txt

Group file (treatment list)
 1259: Group_sample_null_model.txt

Distance method:
 Jaccard distance (presence/absence)
 Bray-Curtis distance (abundance)
 Sorenson distance (presence/absence)

Null model:
 Chase 2010 EcoSim null model (Randomize community data matrix with the independent swap algorithm (Gotelli 2000) maintaining both row and column sums constant)
 Chase 2011 Ecosphere null model (keep alpha and gamma diversity of the whole/group data constant)
 Randomize community data matrix abundances within species (only keep column sum constant)

Keep gamma diversity in:
 Total dataset
 Group by group

Execute

Output:

: Null model test on Permdisp

3. Beta NTI calculation

Input:

Beta NTI calculation bNTI (Galaxy Version 1.0.0)

OTU table file
 1285: Uparse_OTU_table_resample.txt

Phylogenetic tree file
 1286: Uparse_FastTree.nwk

Weighted:
 Weighted
 Unweighted

Randomization

Output:

bNTI result

4. RC distance

Input:

RC distance Raup-Crick based on taxonomic dissimilarity index (Galaxy Version 1.0.0) ▼ Options

OTU table file
 1285: Uparse_OTU_table_resample.txt

Distance method:
 Jaccard distance
 Bray-Curtis distance

Community matrix type
 Default is to use abundance data
 Transfer to the presence/absence data

Use Chases's method (Chase 2011)
 Abundance weighted
 Based on Jaccard dissimilarity (not abundance weighted)

Randomization
 1000

Execute

Output:

1295: Raup-Crick result

5. Summary ecological process

Input:

Summary ecological process Based on bNTI and RC distance (Galaxy Version 1.0.0) ▼ Options

bNTI result
 1291: bNTI result

Raup-Crick result
 1295: Raup-Crick result

Execute

Output:

1295: Ecological process summary

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:

A screenshot of a Galaxy tool interface titled "pick up ref OTU Generate OTU table based on GreenGene references (Galaxy Version 1.0.0)". The input field "Sequences to generate OTUs: (FASTA)" contains "108: Trim_length". A "Execute" button is visible at the bottom left.

Output:

OTUs_GG_ref.biom

1.2 Normalize by Copy Number

Input:

A screenshot of a Galaxy tool interface titled "Normalize by Copy Number (Galaxy Version 1.1.1)". The input file is "1693: OTUs_GG_ref.biom". The "GreenGenes Version (used to generate your OTU table)" dropdown is set to "GG 13.5". A "Execute" button is visible at the bottom left.

Output:

1694: Normalize by Copy Number on data 1693
1,129 lines
格式: biom, 数据库: ?

1.3 Predict Metagenome

Input:

Predict Metagenome (Galaxy Version 1.1.1)

Input file
 1694: Normalize by Copy Number on data 1693

GreenGenes Version (used to generate your OTU table)
 GG 13.5

Type of functional predictions
 KEGG Orthologs

Execute

Output:

1696: Predict Metagenome on data 1694

30 lines

格式: **txt**, 数据库: ?



```
#Sample Metric Value
A2      Weighted NSTI  0.04519505782896097
6       Weighted NSTI  0.1693524444199479
4       Weighted NSTI  0.16395798594477903
2       Weighted NSTI  0.1492311122685514
```

1695: Predict Metagenome on data 1694

3,481 lines

格式: **biom**, 数据库: ?

1.4 Categorize by Function

Input:

Categorize by Function (Galaxy Version 1.1.1)

Input file
 1695: Predict Metagenome on data 1694

Pathway Hierarchy Level
 3

Metadata category that describes hierarchy (NOTE: RFAM categories cannot be collapsed).
 KEGG Pathways

Execute

Output:

1697: Categorize by Function on data 1695

Convert Biom to Tabular

Input:

Convert Biom to Tabular Convert Biom file to Tabular file (Galaxy Version 1.0.0)

Biom file
 1697: Categorize by Function on data 1695

Execute

Output:

```
1698: Tabular file for Categorize by Function on d  
ata 1695  
328 lines  
格式: txt, 数据库: ?  
[Icon] [Icon] [Icon] [Icon] [Icon] ?  
# Constructed from biom file  
#OTU ID A2 6 4 2 3 5  
1,1,1-Trichloro-2,2-bis(4-chlorophenyl)ethane (DDT) c
```

2. Tax4Fun

2.1 Preparation for Tax4Fun

Input:

Preparation for Tax4Fun Generate taxonomy file for Tax4Fun (Galaxy Version 1.0.0) ▼ Options

Resample OTU table
[Icon] [Icon] [Icon] 61: resample_UPARSE_otu_table.txt

OTU classification result
[Icon] [Icon] [Icon] 103: Classifier of silva16s.txt
Recommend: The taxonomy result based on SILVA database in rdp classifier

Execute

Output:

- Import file for Tax4Fun

2.2 Tax4Fun

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

Input:

Tax4Fun Generate Tax4Fun results based on taxonomy information (Galaxy Version 1.0.0) ▼ Options

Tax4Fun input file containing OTU and taxonomy information
[Icon] [Icon] [Icon] 105: Import file for Tax4Fun

Functional profile approach
 Functional capabilities of microbial communities
 Metabolic capabilities according to MoP aproach

Method for pre-computing the functional reference profiles
 UProC
 PAUDA

Reads length for reference profiles
Based on 100 bp reads

Normalize by the 16S rRNA gene copy number
 Yes
 No

Execute

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:

Functional Annotation of Prokaryotic Taxa Functional annotation based on a database of prokaryotic clades (Galaxy)
Version 1.0.0 ▼ Options

Resample OTU table
 61: resample_UPARSE_otu_table.txt

OTU taxonomy information from rdp classifier
 59: Classifier of 16srna.txt

Normalization approaches
 Normalize before collapsing
 Normalize after collapsing
 Normalize before collapsing excluding unassigned groups
 none

How to normalize the table (before_collapsing normalizes the input table prior to processing, after_collapsing normalizes the output table).

Approach for group name definition

Omit unrepresented groups

partition each group by scores in report

The score of an entry is its total number of hits across all data columns, divided by the total number of hits across all entries and data columns. Please use comma(,) to separate the scores.

Output:

111: Sub-tables for each functional group.zip
110: Functional group overlaps (Jaccard)
109: Matched functional definition
108: Group-record associations
107: Report of FAPROTAX
106: FAPROTAX result for resample_UPARSE_otu_table.txt

4. BugBase

4.1 Pick up ref OTU

Input:

pick up ref OTU Generate OTU table based on GreenGene references (Galaxy Version 1.0.0) ▼ Options

Sequences to generate OTUs: (FASTA)

26: Trim_length

Execute

Output:

[112: OTUs_GG_ref.biom](#)

4.2 BugBase Analysis

Input:

BugBase Analysis Determine high-level phenotypes present in microbiome samples (Galaxy Version 0.1.0) ▼ Options

OTU table (Biom format)

112: OTUs_GG_ref.biom
Picked OTU table against GreenGene database

Use coefficient of variance instead of variance to determine thresholds

Yes No

Centered Log-Ratio Transformation Data

Yes No
Instead of converting to relative abundance, you can centered log-ratio transform the data. This helps prevent issues with the compositionality of sequencing data.

Specific Thresholds

NA
The threshold must be a float between 0 and 1. Default (left NA), BugBase will use the threshold with the highest variance in your data.

Taxa level to plot otu contributions

2
Default is 2 (phylum), others should be within the list, 1,2,3,4,5,6,7.

Use KEGG modules?

Not use KEGG modules

Proceed with mapping file or without mapping file

Without mapping file

Execute

Output:

[116: Threshold used in OTUs_GG_ref.biom.zip](#)

[115: Predicted phenotypes for OTUs_GG_ref.bio
m.zip](#)

[114: OTU contributions for OTUs_GG_ref.biom.zip](#)

[113: Normalized otu table for OTUs_GG_ref.biom](#)

If you have a mapping 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:

FunGuild Analysis Parsing fungal community datasets with ecological guild (Galaxy Version 1.0.0)

Resample OTU table
 429: resample_otu_table_soil_ITS.txt

OTU classification result from rdp classifier
 428: Galaxy84-[Classifier_Soil_ITS.txt].txt

Database type
 Fungi
 Nematode

Assign a specified database to the program

Execute

Output:

- Funguild result for “resample_otu_table_soil_ITS.txt”

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

E. Other analysis tools in microbial ecology

1. LEfSe

1.1 Input prepare for LefSe analysis

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

Input:

Input prepare for LEfSe analysis (Galaxy Version 1.0.0)

OTU table (tabular file)
61: resample_UPARSE_otu_table.txt

Classification summary (tabular file)
59: Classifier of 16srrna.txt
The output file from rdp classifier step

Execute

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

A) Format Data for LEfSe (Galaxy Version 1.0)

Upload a tabular file of relative abundances and class labels (possibly also subclass and subjects labels) for LEfSe - See samples below - Please use Galaxy Get-Data/Upload-File. Use File-Type = tabular
369: test files

Select whether the vectors (features and meta-data information) are listed in rows or columns
Rows

Select which row to use as class
#1:oxygen_availability

Select which row to use as subclass
#2:body_site

Select which row to use as subject
#3:subject_id

Per-sample normalization of the sum of the values to 1M (recommended when very low values are present)
Yes

Execute

1.3 B) LDA Effect Size (LEfSe)

B) LDA Effect Size (LEfSe) (Galaxy Version 1.0) ▼ Options

Select data
1142: A) Format Data for LEfSe on data 369

Alpha value for the factorial Kruskal-Wallis test among classes
0.05

Alpha value for the pairwise Wilcoxon test between subclasses
0.05

Threshold on the logarithmic LDA score for discriminative features
2.0

Do you want the pairwise comparisons among subclasses to be performed only among the subclasses with the same name?
No

Set the strategy for multi-class analysis
All-against-all (more strict)

Execute

1.4 C) Plot LEfSe Results

C) Plot LEfSe Results (Galaxy Version 1.0) ▼ Options

Select data
373: B) LDA Effect Size (LEfSe) on data 371

Set text and label options (font size, abbreviations, ...)
Default

Set some graphical options to personalize the output
Default

Output format
png

Set the dpi resolution of the output
150

Execute

1.5 D) Plot Cladogram

D) Plot Cladogram (Galaxy Version 1.0) ▼ Options

Select data
1144: B) LDA Effect Size (LEfSe) on data 1142

Set structural parameters of the cladogram
Default

Set text and label options (font size, abbreviations, ...)
Default

Set some graphical options to personalize the output
Default

Output format
png

Set the dpi resolution of the output
150

Execute

1.6 E) Plot One Feature

E) Plot One Feature (Galaxy Version 1.0)

The formatted datasets
1142: A) Format Data for LEfSe on data 369
The input is the result of A

The LEfSe output
1144: B) LDA Effect Size (LEfSe) on data 1142
The input is the result of B

Select the feature names among biomarkers or all features
Biomarkers only

Select the feature to plot
Bacteria.Actinobacteria

Set some graphical options to personalize the output
Default

Output format
png

Set the dpi resolution of the output
150

Execute

1.7 F) Plot Differential Features

F) Plot Differential Features (Galaxy Version 1.0)

The formatted datasets
1142: A) Format Data for LEfSe on data 369
The input is the result of A

The LEfSe output
1144: B) LDA Effect Size (LEfSe) on data 1142
The input is the result of B

Do you want to plot all features or only those detected as biomarkers?
Biomarkers only

Set some graphical options to personalize the output
Default

Output format
png

Set the dpi resolution of the output
150

Execute

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)

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:

Fastqc: Fastqc QC using FastQC from Babraham (Galaxy Version 0.3)

Short read data from your current history
9: TrimPrimer_tagged1_R1.fastq.fastq

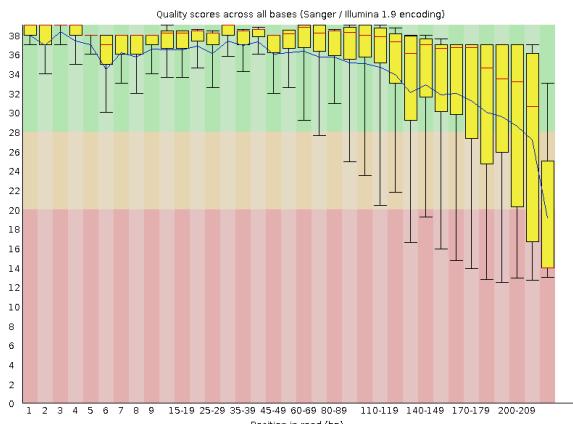
Title for the output file - to remind you what the job was for
FastQC

Contaminant list
Nothing selected

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Prir

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 Archive (SRA).

FASTA

Input:

split fasta file into samples(FASTA) To split one fasta file into different samples (Deng lab) (Galaxy Version 1.0.0) ▼ Options

Sequence file 1 (FASTA)
584: {Combined.fastq}.fasta

Execute

Output:

588: split_{Combined.fastq}.fasta.gz

325,937 lines

格式: zip, 数据库: ?

```
10.fasta  
11.fasta  
12.fasta  
13.fasta  
14.fasta  
15.fasta  
16.fasta  
17.fasta  
18.fasta  
19.fasta  
1.fasta  
20.fasta  
21.fasta  
22.fasta  
23.fasta  
24.fasta  
25.fasta  
26.fasta
```

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:

FASTQ format check Check the format of FASTQ files (Galaxy Version 1.0.0)

Sequence file

1190: Combined.fastq

Execute

Output:

- Checked Combined.fastq
- Fastq check summary

4. Length Statistics

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

Input:

Length statistics Quick information about the sequences length distribution (Galaxy Version 1.0.0)

Input file format

FASTA

fasta file

990: Remove_N.fasta

Execute

Options:

Input file format: FASTA / FASTQ

Output:

- length_stat.html

Summary:

The total sequence number: 38049.

The average length of the sequences: 293.36

Sequence length distribution:

The minimum length is: 201

The maximum length is: 427

Data

Length Seq number

201 117

202 165

203 132

204 89

205 100

... ...

5. Sequence number for each tag

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

FASTQ

Input:

Sequence number for each tag (FASTQ) Quick information about the sequences distribution among tags (Galaxy Version 1.0.0)

fastq file

21: Trimmed_Combined.fastq

Tags are linked by

--

eg. Sequence ID: >HG9RAK004JHMDS--T4F, the linker is '--'

Execute

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:

Merge and add tags for each sequence One file and multiple files (Galaxy Version 1.0) Options

Select input file format
Fasta format

Select multiple files
3: 3C.Tags.fasta
4: 3A.Tags.fasta
3: 2S.fasta
2: 2C.fasta
1: 2A.fasta

Please rename your selected file referring to specific format (showing below). Select multiple files using CTRL.

rename the merged file
Renamed file

Execute

Please rename your selected sequence file as follows, like A1.fasta, B1.fasta, C1.fasta or A1.fastq, B1.fastq, C1.fastq .

The extension names for the fasta or fastq file also include A1.fa, A1.fq.

A1 is the sample name. Please use simple names and not use special symbols in the name, like ".", "-", "/", "#".

Options:

Input file format: Fasta format / Fastq format

Output:

- Renamed file for fasta

```
>2A_Tag1--2A
TACGGAGGGTGCAAGCGTTGCTCGGAATTACTGGCGTA/
>2A_Tag2--2A
TACGGAGGTGCAAGCGTTATCCGATTCTGGGTTTA/
>2A_Tag3--2A
CACCGGGCGCTCGAGTGGTAACCGTTATTATTGGGTCTA/
>2A_Tag4--2A
CACCGGCAGCTCAAGTGGTGGCCATTTTATTGGGCCTA/
>2A_Tag5--2A
CACCGGGCGCTCGAGTGGTAACCGTTATTATTGGGTCTA/
>2A_Tag6--2A
CACCGGCAGCTCAAGTGGTGGCCATTTTATTGGGCCTA/
```

- 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:

Merge Files Merge two files (Galaxy Version 1.0)

File1
 97: DXAL

File2
 58: XXAL.fastq

Additional file for merging

1: Additional file for merging
 Additional file
 38: TTS.fastq

2: Additional file for merging
 Additional file
 29: XXBN_CBS.fastq

+ Insert Additional file for merging Insert more files if you have multiple files to merge rename the merged file

Succeision

Execute

Output:

- Succeision (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 administrators when you have problems.

Input:

Data Location Data location in the galaxy server (Galaxy Version 1.0.0)

Input file
 1690: R2.fastq

Execute

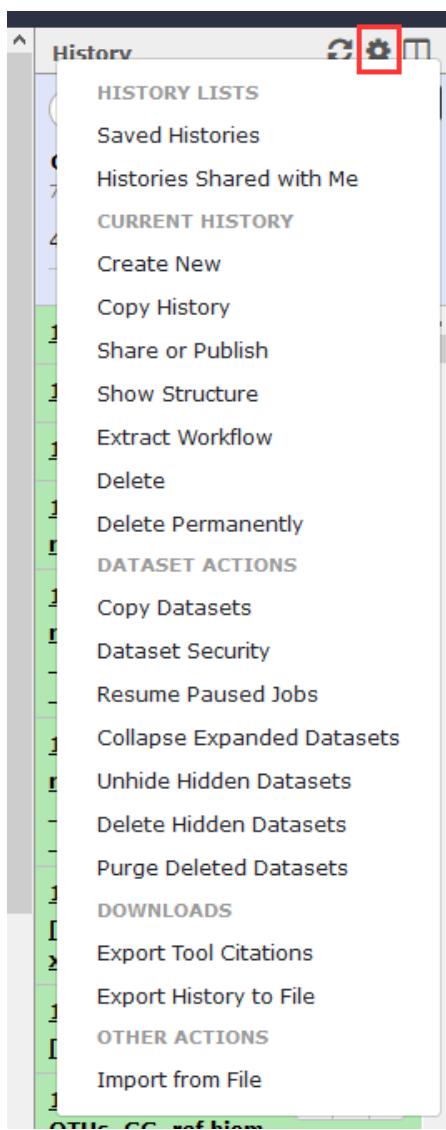
Output:

- `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:

The screenshot shows the 'Copy History Items' dialog. On the left, under 'Source History:', a dropdown menu shows '2: Galaxy_Text_compare'. Below it are buttons for 'All' and 'None', and a list of items: '1: Galaxy50-[tagged1_16S_2_wzy_R1.fastq]', '2: Galaxy51-[tagged2_16S_2_wzy_R2.fastq]', '3: Galaxy52-[barcode_summary_16S_2_wzy.txt.txt]', '7: 16S_2__barcode_wzy.txt', and '11: Galaxy4-'. Some items have checkboxes next to them. An arrow points to the right side of the dialog. On the right, under 'Destination History:', a dropdown menu shows '1: For test analysis'. Below it is a link 'Choose multiple histories'. At the bottom, there is a section for 'New history named:' with a text input field and a 'Copy History Items' button.

2. Shared test datasets

The screenshot shows the Galaxy DengLab interface. At the top, there is a navigation bar with 'Analyze Data', 'Workflow', 'Visualize', 'Shared Data', 'Admin', 'Help', and 'User'. A red arrow points from the 'Shared Data' button to a dropdown menu. The dropdown menu is titled 'Data Libraries' and includes options: Histories, Workflows, Visualizations, and Pages. Another red arrow points from the 'Data Libraries' option to its submenu. Below the navigation bar, there is a 'Welcome to Metagenomics for (DengLab)!' banner. On the left, there is a 'Tools' sidebar with various analysis tools listed. The main area displays a table of data libraries:

name	description	synopsis	actions
16S data library	for denglab		Edit Manage
18S ref	For classification and phylogenetic analysis. silva_132_97_18S.fna is from Qiime...	Silva and PR2 database uploaded by Lishuzhen.	Edit Manage
ITS refs			Edit Manage
mcrA	Functional gene		Edit Manage
Test data	For users to learn how to use this pipeline		Edit Manage

Below the table, there is a 'DATA LIBRARIES' toolbar with buttons for 'Create Folder', 'To History', 'Download', 'Delete', 'Details', and 'Help'. A red arrow points from the 'To History' button to its icon. The URL in the address bar is 'Libraries / Test data'. The page lists five items:

name	description	data type	size	time updated (UTC)	state
barcode.txt		tabular	1.4 KB	2018-10-30 04:58 AM	Edit Manage
R1.fastq		fastqsanger	538.2 MB	2018-10-30 04:58 AM	Edit Manage
R2.fastq		fastqsanger	539.4 MB	2018-10-30 04:58 AM	Edit Manage
Sample list for rarefaction curve.txt		txt	60 bytes	2018-10-30 04:58 AM	Edit Manage
Treatment file for dissimilarity.txt		tabular	90 bytes	2018-10-30 04:58 AM	Edit Manage

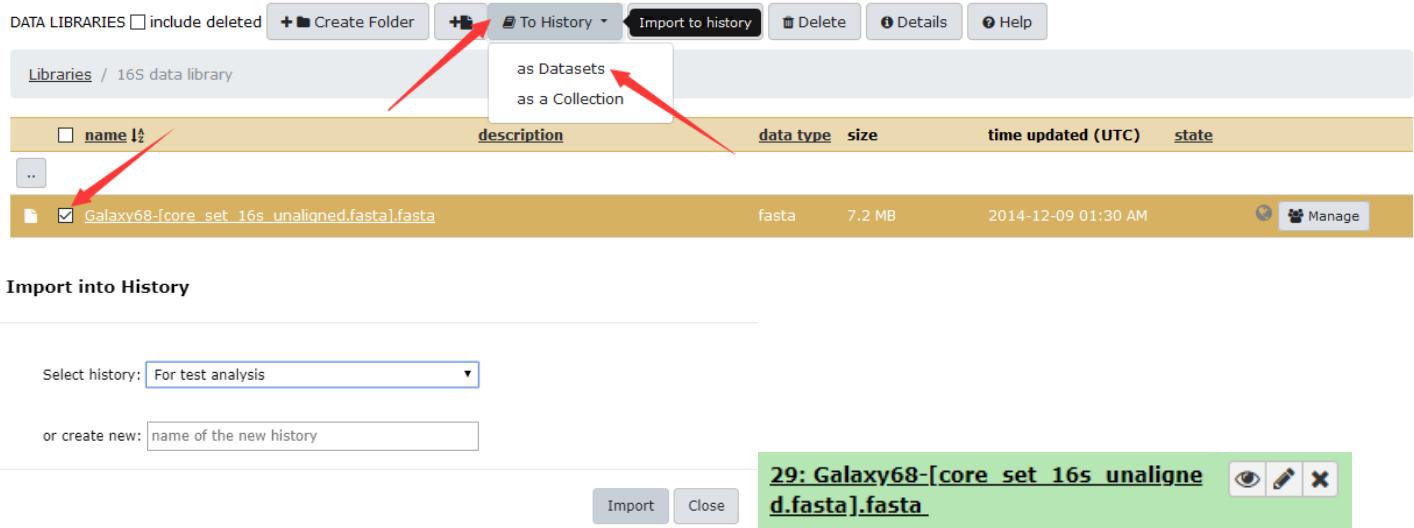
At the bottom, there is a pagination control with buttons for '<<', '0', '1', '2', and '>>'. The text '5 items shown (change) 5 total' is also present.

3. Shared libraries

This screenshot is identical to the one above, showing the Galaxy DengLab interface with the 'Shared Data' menu open. A red arrow points from the 'Shared Data' button to the 'Data Libraries' option in the dropdown menu. The main area displays the same 'Welcome to Metagenomics for (DengLab)!' banner and tools sidebar. The data library table and history list are also identical to the previous screenshot.

DATA LIBRARIES		« 0 1 2 »	5 libraries shown (change) 5 total	Great Library	<input type="checkbox"/> include deleted <input type="checkbox"/> exclude restricted	+ New Library	Help
name	description	synopsis					
16S data library	for denglab					 Edit	 Manage
18S ref	For classification and phylogenetic analysis. silva_132_97_18S.fna is from Qiime...	Silva and PR2 database uploaded by Lishuzhen.				 Edit	 Manage
ITS_refs						 Edit	 Manage
mcrA	Functional gene					 Edit	 Manage
Test data	For users to learn how to use this pipeline					 Edit	 Manage

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



The screenshot shows the Galaxy web interface for managing data libraries. At the top, there's a navigation bar with buttons for 'Create Folder', 'To History', 'Import to history', 'Delete', 'Details', and 'Help'. Below the navigation bar, a sub-menu for 'To History' is open, showing options 'as Datasets' and 'as a Collection'. A red arrow points to the 'Import to history' button. Another red arrow points to the 'as Datasets' option in the sub-menu. A third red arrow points to the checkbox next to the dataset file name 'Galaxy68-[core_set_16s_unaligned.fasta].fasta' in the list below. The list includes columns for name, description, data type, size, time updated (UTC), and state. The dataset 'Galaxy68-[core_set_16s_unaligned.fasta].fasta' is selected, indicated by a checked checkbox. At the bottom, there's an 'Import into History' section with a dropdown for 'Select history' set to 'For test analysis' and a text input for 'or create new' with placeholder 'name of the new history'. A green success message box at the bottom right says '29: Galaxy68-[core_set_16s_unaligned.fasta].fasta' with edit and delete icons.

4. Dataset deletions

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

Saved Histories

[Advanced Search](#)

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

Saved Histories

[Close Advanced Search](#)

name:	<input type="text"/>	
tags:	<input type="text"/>	
sharing:	accessible all private published shared	
status:	active all deleted	

5. Share historys to other users

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

Share or Publish History 'Galaxy_Text_compare'

Make History Accessible via Link and Publish It

This history is currently restricted so that only you and the users listed below can access it. You can:

[Make History Accessible via Link](#)

Generates a web link that you can share with other people so that they can view and import the history.

[Make History Accessible and Publish](#)

Makes the history accessible via link (see above) and publishes the history to Galaxy's [Published Histories](#) section, where it is publicly listed and searchable.

Share History with Individual Users

You have not shared this history with any users.

[Share with a user](#)

[Back to Histories List](#)

