CSC

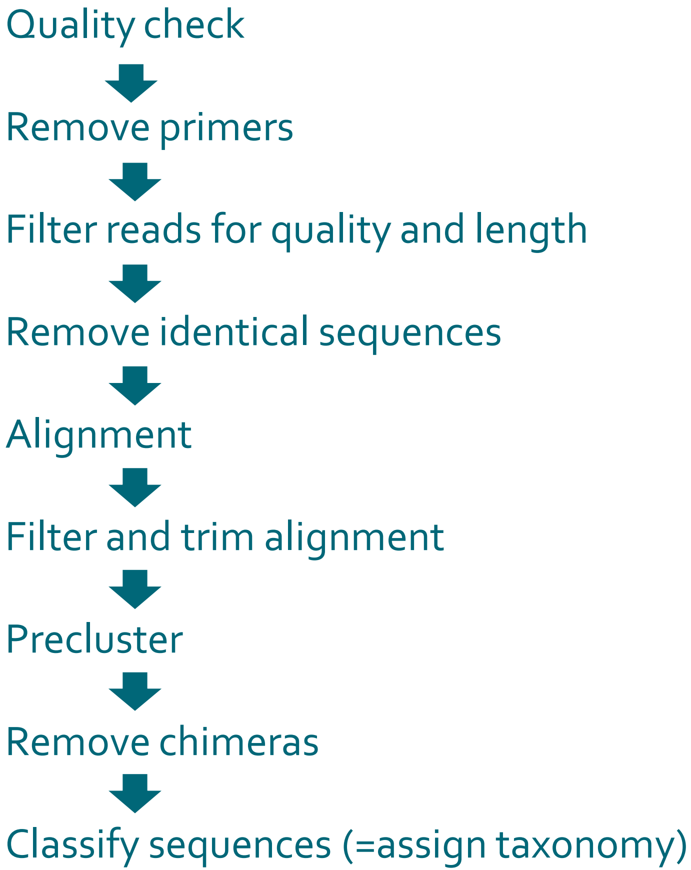
ICT Solutions for
Brilliant Minds

Bioinformatics for microbiome research
Day 2: microbial community analysis

Jyväskylä Summer School 2023

Heli Juottonen and Eija Korpelainen (slides modified from: Jesse Harrison)
CSC – IT Center for Science Ltd.
chipster@csc.fi

Outline of Day 1: Preprocessing reads



Quality check

Remove primers

Filter reads for quality and length

Remove identical sequences

Alignment

Filter and trim alignment

Precluster

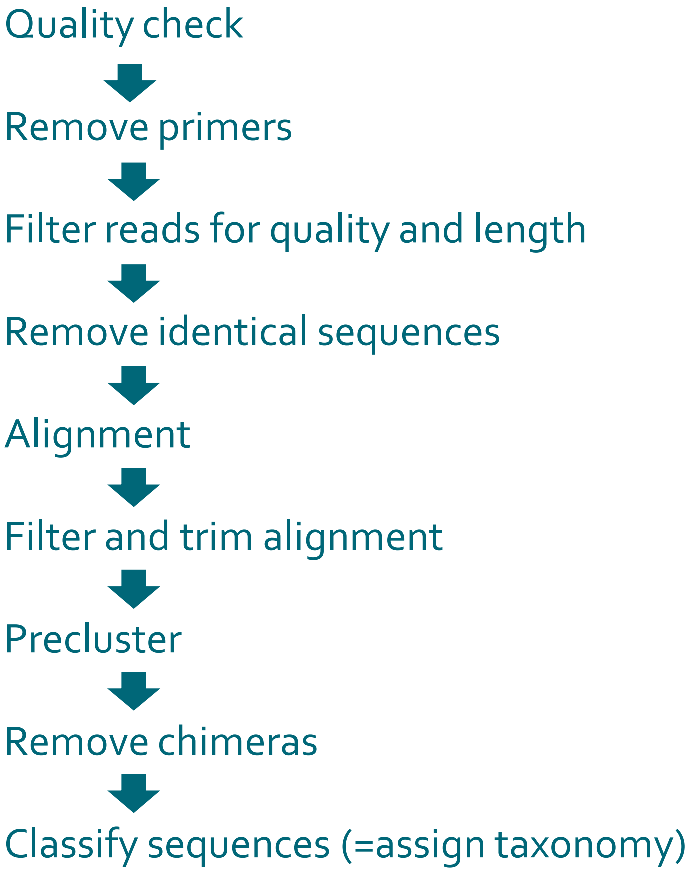
Remove chimeras

Classify sequences (=assign taxonomy)

Output so far:

1. FASTA file of processed reads
2. count file (which read in which sample)
3. taxonomy file (taxonomy of each read)

Outline of Day 1: Preprocessing reads



Quality check

Remove primers

Filter reads for quality and length

Remove identical sequences

Alignment

Filter and trim alignment

Precluster

Remove chimeras

Classify sequences (=assign taxonomy)

Output so far:

1. FASTA file of processed reads
2. count file (which read in which sample)
3. taxonomy file (taxonomy of each read)

Day 2: Community analysis

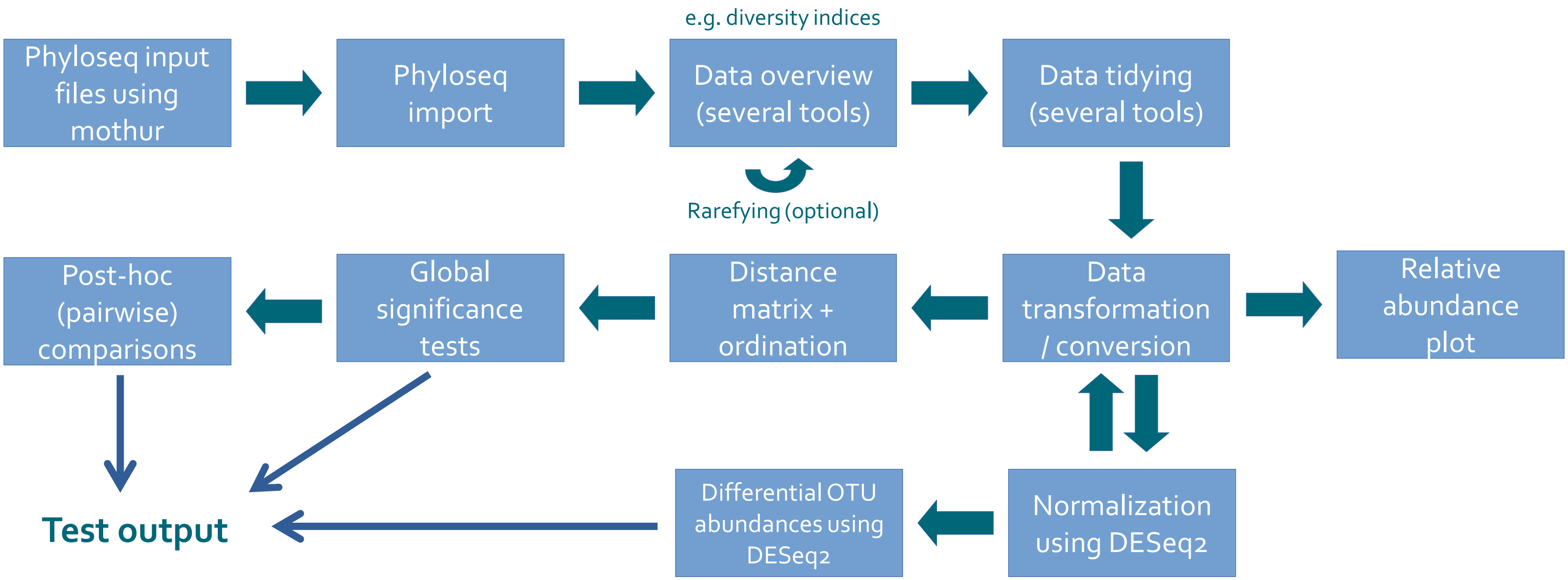
- Clustering into OTUs
- Import into **Phyloseq**
- Data tidying & transformations
- Taxonomy plots
- Alpha diversity
- Beta diversity: ordinations & statistics

CSC

ICT Solutions for
Brilliant Minds

Part 1: Tool overview and data importing

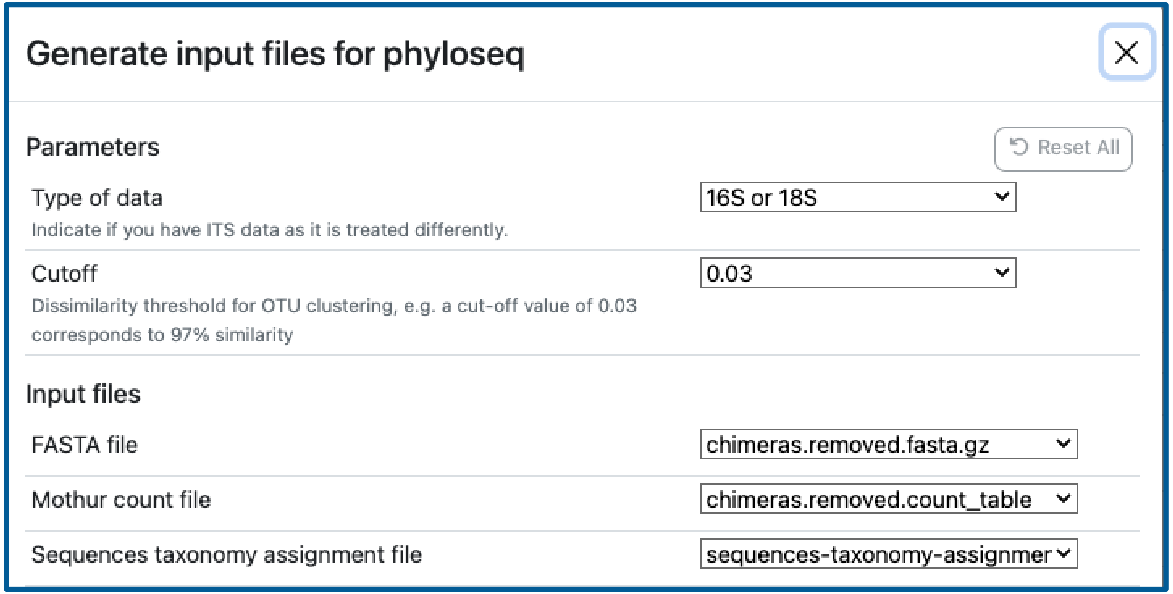
Workflow for microbial community analysis in Chipster


Phyloseq input files using mothur
Phyloseq import
e.g. diversity indices
Data overview (several tools)
Data tidying (several tools)
Rarefying (optional)
Post-hoc (pairwise) comparisons
Global significance tests
Distance matrix + ordination
Data transformation / conversion
Relative abundance plot
Test output
Differential OTU abundances using DESeq2
Normalization using DESeq2

Generating input files for phyloseq

Phyloseq is a multi-use R package for microbial community data processing and analysis

https://joey711.github.io/phyloseq/



Generate input files for phyloseq

Parameters

Type of data: 16S or 18S

Indicate if you have ITS data as it is treated differently.

Cutoff: 0.03

Dissimilarity threshold for OTU clustering, e.g. a cut-off value of 0.03 corresponds to 97% similarity

Input files

FASTA file: chimeras.removed.fasta.gz

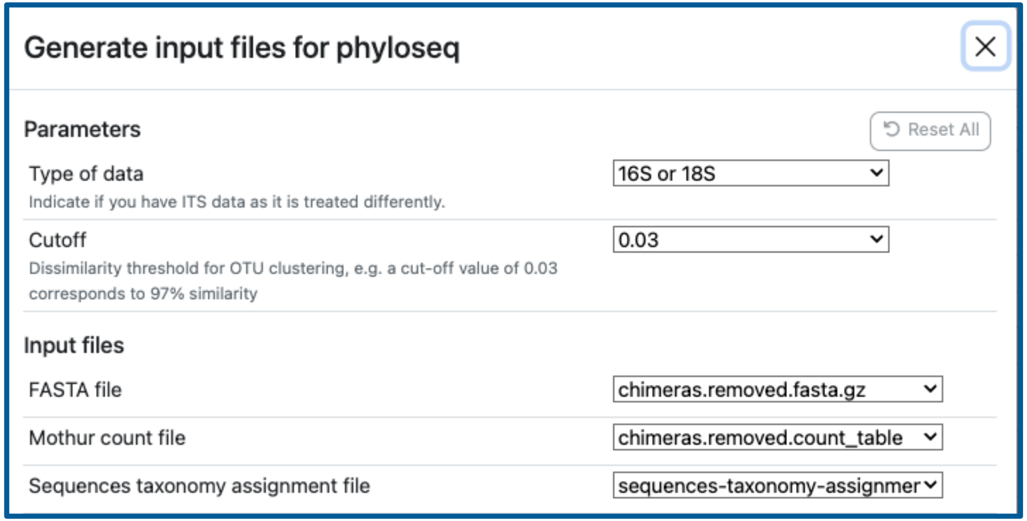
Mothur count file: chimeras.removed.count_table

Sequences taxonomy assignment file: sequences-taxonomy-assignmer

Generating input files for phyloseq

Specifications for creating phyloseq input files:

- type of data (16S/18S or ITS)
- % cut off for OTU clustering
- files produced by mothur
  - final FASTA
  - count file
  - taxonomy file = taxonomy assignment of each read



Generate input files for phyloseq

Parameters

Type of data: 16S or 18S

Indicate if you have ITS data as it is treated differently.

Cutoff: 0.03

Dissimilarity threshold for OTU clustering, e.g. a cut-off value of 0.03 corresponds to 97% similarity

Input files

FASTA file: chimeras.removed.fasta.gz

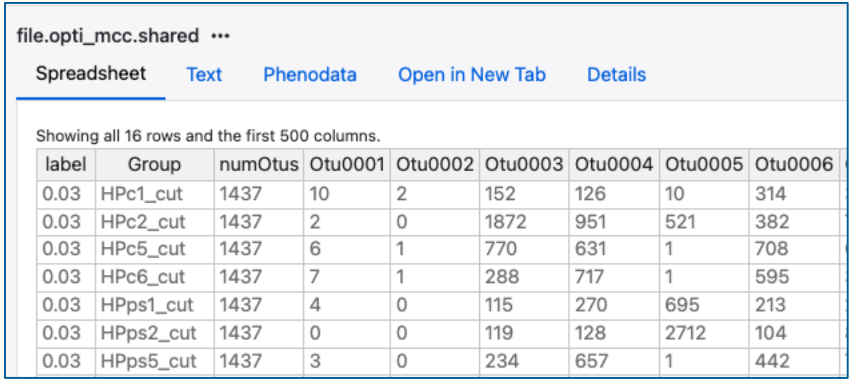
Mothur count file: chimeras.removed.count_table

Sequences taxonomy assignment file: sequences-taxonomy-assignmer

Generating input files for phyloseq

Generated input files:

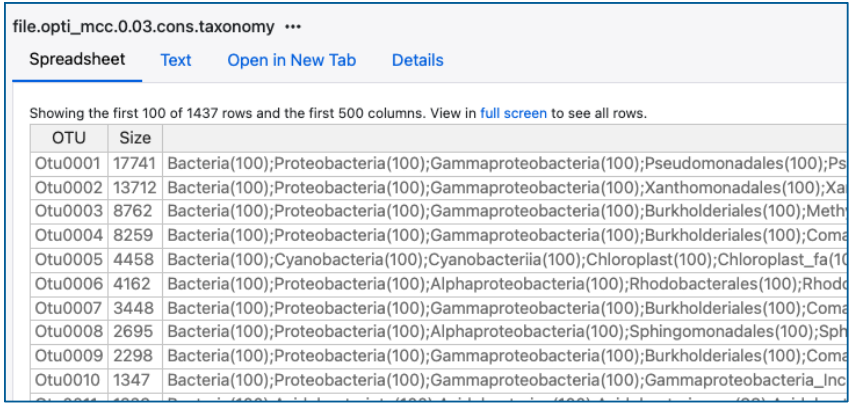
- .shared file (mothur file format)
  - samples in rows, OTUs in columns
  - how many reads of each OTU in each sample (OTU table)
- consensus taxonomy file
  - taxonomy assignments of OTUs
- phenodata file



file.opti_mcc.shared •••

Showing all 16 rows and the first 500 columns.

label	Group	numOtus	Otu0001	Otu0002	Otu0003	Otu0004	Otu0005	Otu0006
0.03	HPc1_cut	1437	10	2	152	126	10	314
0.03	HPc2_cut	1437	2	0	1872	951	521	382
0.03	HPc5_cut	1437	6	1	770	631	1	708
0.03	HPc6_cut	1437	7	1	288	717	1	595
0.03	HPps1_cut	1437	4	0	115	270	695	213
0.03	HPps2_cut	1437	0	0	119	128	2712	104
0.03	HPps5_cut	1437	3	0	234	657	1	442

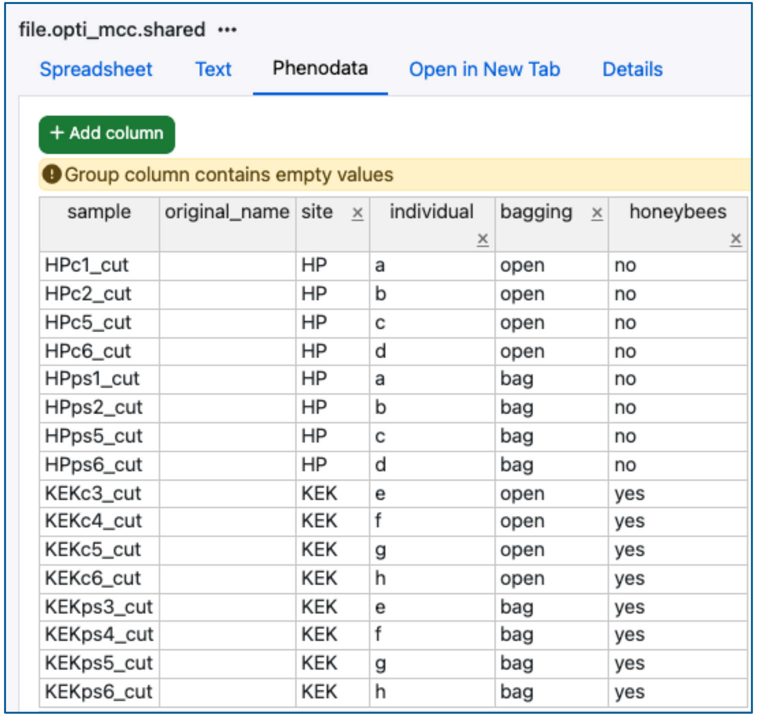


file.opti_mcc.0.03.cons.taxonomy •••

Showing the first 100 of 1437 rows and the first 500 columns. View in full screen to see all rows.

OTU	Size	
Otu0001	17741	Bacteria(100);Proteobacteria(100);Gammaproteobacteria(100);Pseudomonadales(100);Ps
Otu0002	13712	Bacteria(100);Proteobacteria(100);Gammaproteobacteria(100);Xanthomonadales(100);Xa
Otu0003	8762	Bacteria(100);Proteobacteria(100);Gammaproteobacteria(100);Burkholderiales(100);Meth
Otu0004	8259	Bacteria(100);Proteobacteria(100);Gammaproteobacteria(100);Burkholderiales(100);Coma
Otu0005	4458	Bacteria(100);Cyanobacteria(100);Cyanobacteriia(100);Chloroplast(100);Chloroplast_fa(10
Otu0006	4162	Bacteria(100);Proteobacteria(100);Alphaproteobacteria(100);Rhodobacterales(100);Rhodo
Otu0007	3448	Bacteria(100);Proteobacteria(100);Gammaproteobacteria(100);Burkholderiales(100);Coma
Otu0008	2695	Bacteria(100);Proteobacteria(100);Alphaproteobacteria(100);Sphingomonadales(100);Sph
Otu0009	2298	Bacteria(100);Proteobacteria(100);Gammaproteobacteria(100);Burkholderiales(100);Coma
Otu0010	1347	Bacteria(100);Proteobacteria(100);Gammaproteobacteria(100);Gammaproteobacteria_Inc

Phenodata file: fill in sample information



! Group column contains empty values

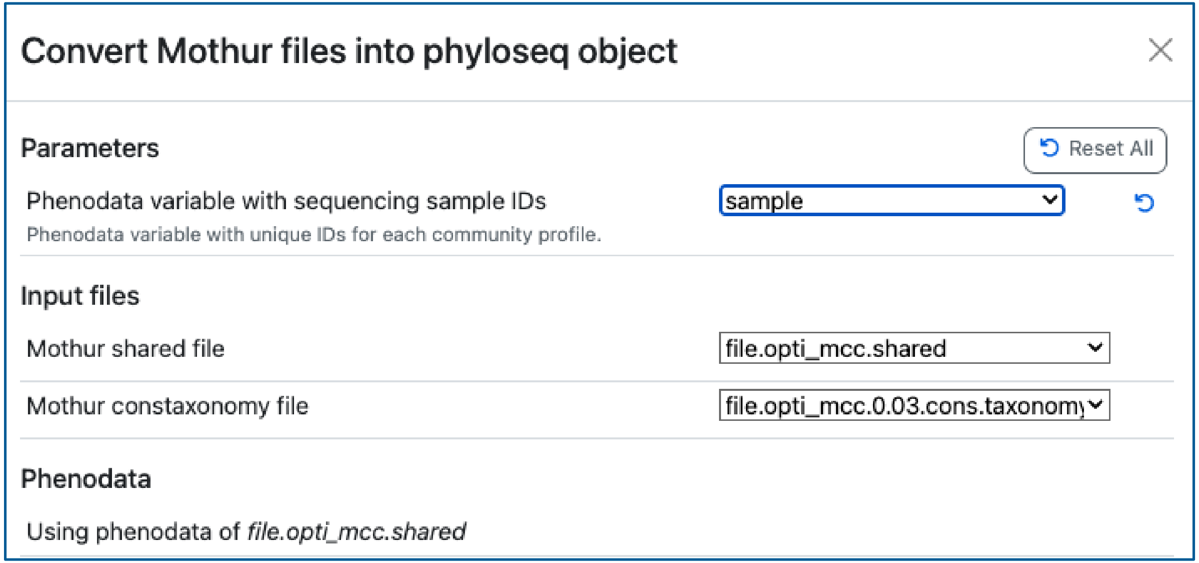
sample	original_name	site ×	individual ×	bagging ×	honeybees ×
HPc1_cut		HP	a	open	no
HPc2_cut		HP	b	open	no
HPc5_cut		HP	c	open	no
HPc6_cut		HP	d	open	no
HPps1_cut		HP	a	bag	no
HPps2_cut		HP	b	bag	no
HPps5_cut		HP	c	bag	no
HPps6_cut		HP	d	bag	no
KEKc3_cut		KEK	e	open	yes
KEKc4_cut		KEK	f	open	yes
KEKc5_cut		KEK	g	open	yes
KEKc6_cut		KEK	h	open	yes
KEKps3_cut		KEK	e	bag	yes
KEKps4_cut		KEK	f	bag	yes
KEKps5_cut		KEK	g	bag	yes
KEKps6_cut		KEK	h	bag	yes

The phenodata file is an editable table with

1) unique IDs for each sample

2) information on sample groupings

Converting input files into a phyloseq object



Parameters

Phenodata variable with sequencing sample IDs: sample

Phenodata variable with unique IDs for each community profile.

Input files

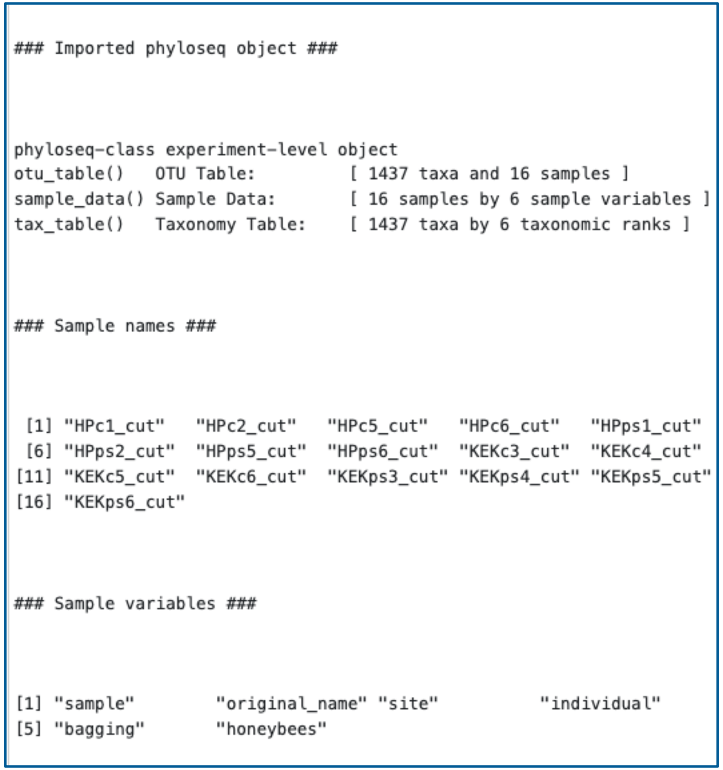
Mothur shared file: file.opti_mcc.shared

Mothur constaxonomy file: file.opti_mcc.0.03.cons.taxonom

Phenodata

Using phenodata of *file.opti_mcc.shared*

Converting input files into a phyloseq object



```
### Imported phyloseq object ###


phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 1437 taxa and 16 samples ]
sample_data() Sample Data:       [ 16 samples by 6 sample variables ]
tax_table()   Taxonomy Table:    [ 1437 taxa by 6 taxonomic ranks ]


### Sample names ###


 [1] "HPc1_cut"   "HPc2_cut"   "HPc5_cut"   "HPc6_cut"   "HPps1_cut"
 [6] "HPps2_cut"  "HPps5_cut"  "HPps6_cut"  "KEKc3_cut"  "KEKc4_cut"
[11] "KEKc5_cut"  "KEKc6_cut"  "KEKps3_cut" "KEKps4_cut" "KEKps5_cut"
[16] "KEKps6_cut"


### Sample variables ###


[1] "sample"        "original_name" "site"          "individual"
[5] "bagging"       "honeybees"
```

Produces **a phyloseq object (.Rda)** and a text summary

The Rda file is used as **the input for downstream analyses**

OTU table, taxonomy table and sample data can be exported for use outside Chipster (Microbial amplicon data analyses / **Extract information from the Phyloseq object**)

CSC

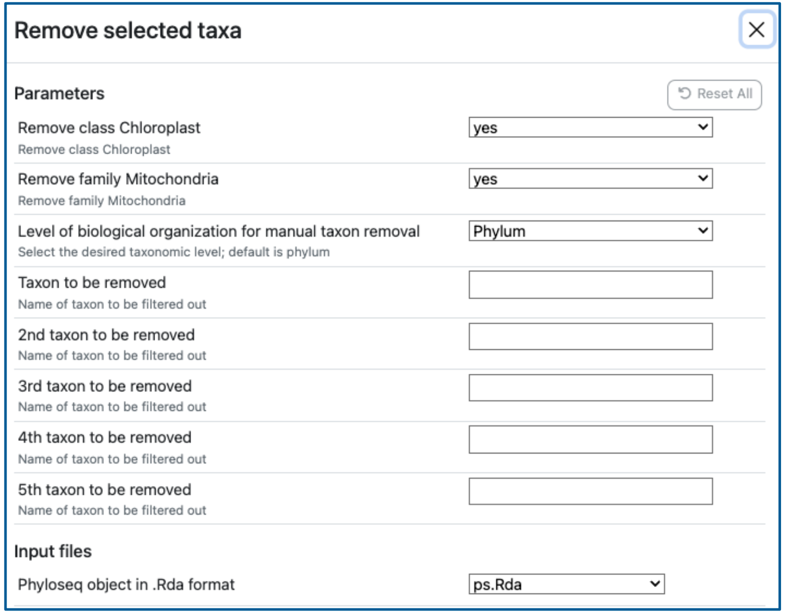
ICT Solutions for
Brilliant Minds

Part 2: Data tidying and alpha diversity

Taxon-level clean-up tools

Under Microbial amplicon data analyses:

- Filter by taxonomic group
  - Remove non-specific sequences (keep e.g. Bacteria or Archaea only)
- Remove selected taxa
  - Remove chloroplast and/or mitochondrial sequences
  - (Manually remove specific taxa)
- Overview of taxon composition
  - user-specified level



Remove selected taxa

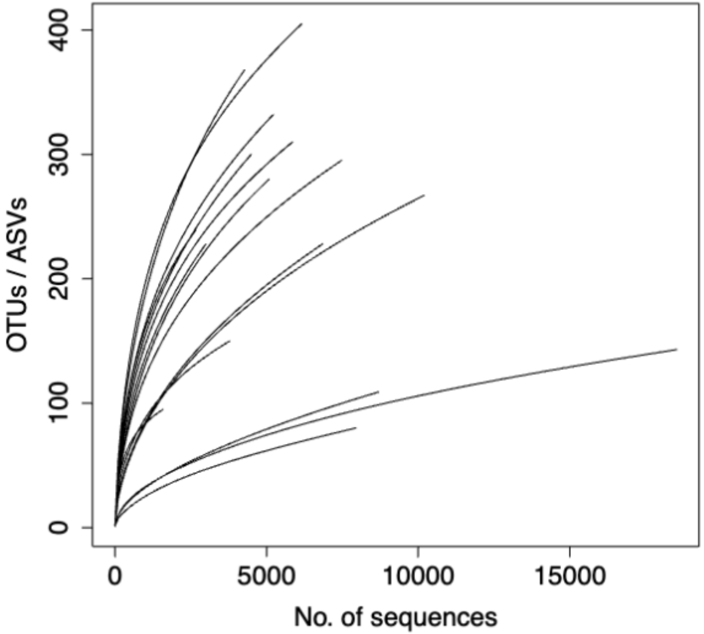
Parameters

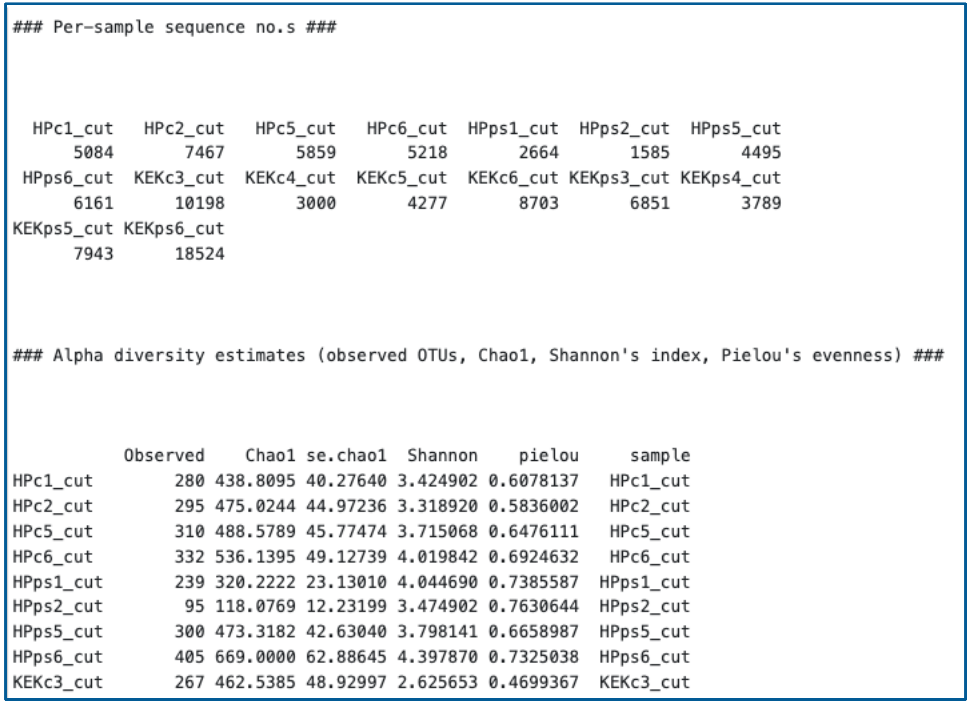
Remove class Chloroplast Remove class Chloroplast	yes
Remove family Mitochondria Remove family Mitochondria	yes
Level of biological organization for manual taxon removal Select the desired taxonomic level; default is phylum	Phylum
Taxon to be removed Name of taxon to be filtered out	
2nd taxon to be removed Name of taxon to be filtered out	
3rd taxon to be removed Name of taxon to be filtered out	
4th taxon to be removed Name of taxon to be filtered out	
5th taxon to be removed Name of taxon to be filtered out	

Input files

Phyloseq object in .Rda format	ps.Rda

Data inspection: Sequence numbers, rarefaction curve and alpha diversity estimates





```
### Per-sample sequence no.s ###



  HPc1_cut    HPc2_cut    HPc5_cut    HPc6_cut   HPps1_cut   HPps2_cut   HPps5_cut
      5084        7467        5859        5218        2664        1585        4495
 HPps6_cut  KEKc3_cut  KEKc4_cut  KEKc5_cut  KEKc6_cut KEKps3_cut KEKps4_cut
      6161       10198        3000        4277        8703        6851        3789
KEKps5_cut KEKps6_cut
      7943       18524



### Alpha diversity estimates (observed OTUs, Chao1, Shannon's index, Pielou's evenness) ###



           Observed    Chao1 se.chao1  Shannon    pielou     sample
HPc1_cut        280 438.8095 40.27640 3.424902 0.6078137   HPc1_cut
HPc2_cut        295 475.0244 44.97236 3.318920 0.5836002   HPc2_cut
HPc5_cut        310 488.5789 45.77474 3.715068 0.6476111   HPc5_cut
HPc6_cut        332 536.1395 49.12739 4.019842 0.6924632   HPc6_cut
HPps1_cut       239 320.2222 23.13010 4.044690 0.7385587  HPps1_cut
HPps2_cut        95 118.0769 12.23199 3.474902 0.7630644  HPps2_cut
HPps5_cut       300 473.3182 42.63040 3.798141 0.6658987  HPps5_cut
HPps6_cut       405 669.0000 62.88645 4.397870 0.7325038  HPps6_cut
KEKc3_cut       267 462.5385 48.92997 2.625653 0.4699367  KEKc3_cut
```

Filtering low-abundance OTUs

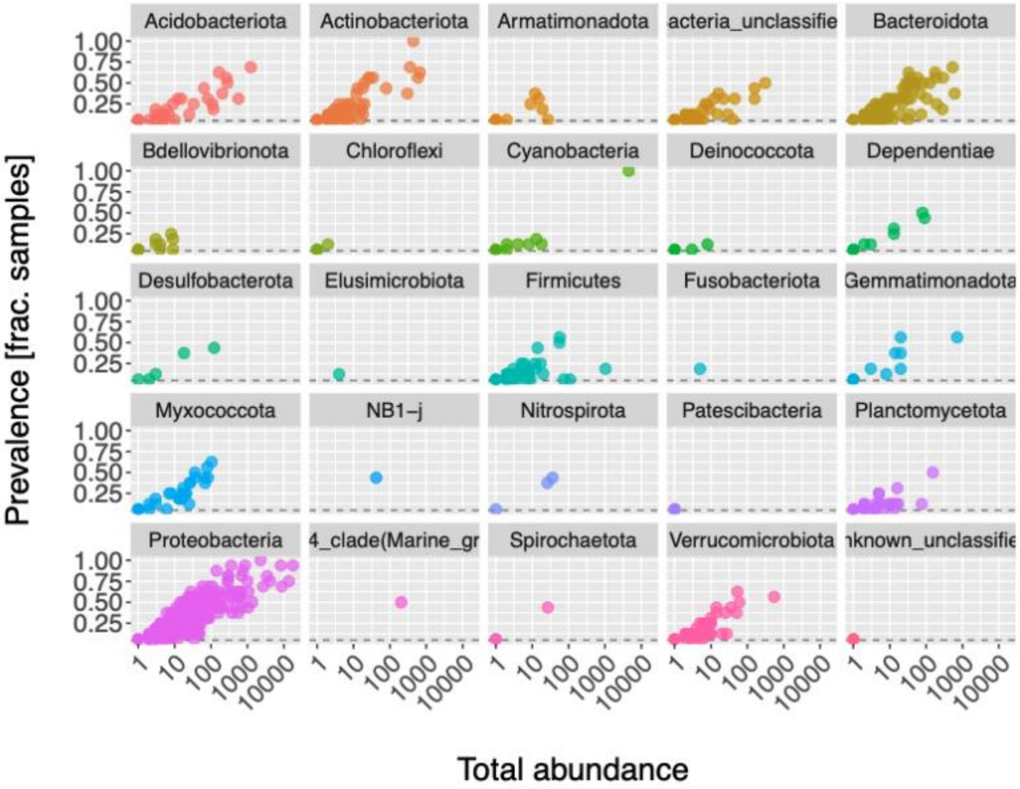
- Filter out OTUs that occur in less than x % of samples
  - Proportional prevalence filtering

- Remove singletons and doubletons
  - Remove OTUs with 0-2 occurrences

Visualizing low-abundance OTUs

Additional prevalence summaries

- Visualization of OTU prevalence at phylum level
- Text summary on low prevalence OTUs

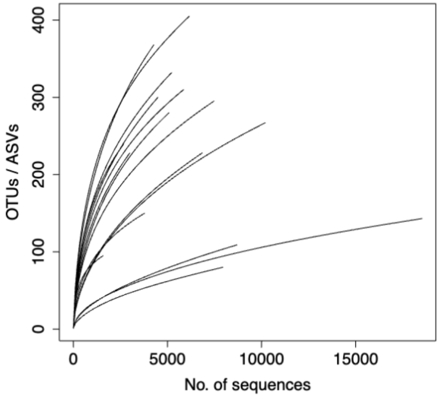


Alpha diversity

Diversity within a habitat: how many species are present, at which relative amounts?

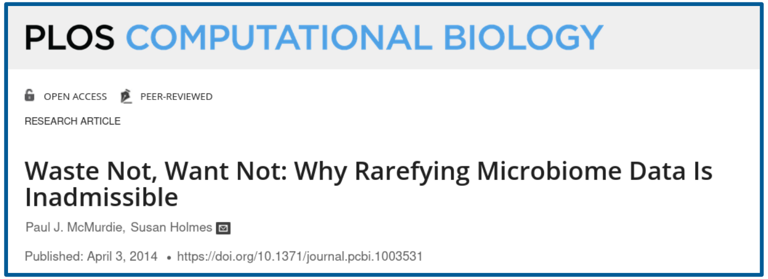
- Species richness - how many species?
  - observed number of OTUs
  - richness estimators (Chao1)
- Evenness – abundance distribution of species?
  - all OTUs equally abundant vs. a few dominant and a lot of rare OTUs
  - Pielou's evenness
- Many diversity indices **combine richness and evenness** (e.g. Shannon index)

Rarefaction?



Uneven sequence numbers among samples can bias comparisons, especially with alpha diversity. Solutions:

- rarefying: equal number of reads picked from all samples (Rarefy OTU data to even depth)
- data transformation (Transform OTU counts)



PLOS COMPUTATIONAL BIOLOGY

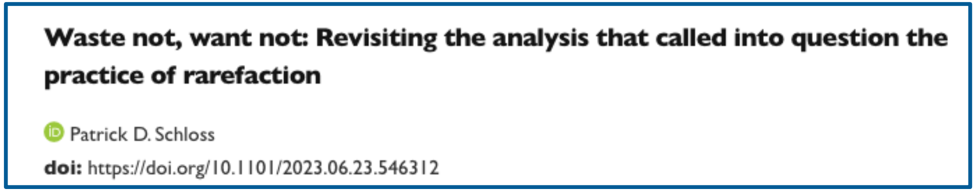
OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible

Paul J. McMurdie, Susan Holmes

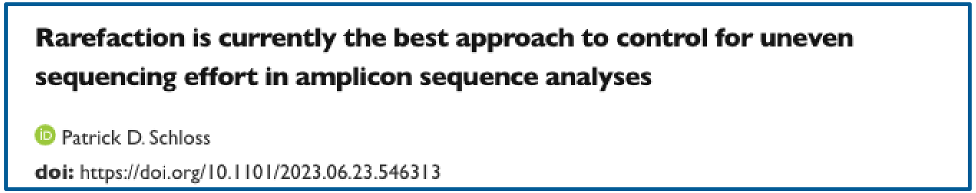
Published: April 3, 2014 • https://doi.org/10.1371/journal.pcbi.1003531



Waste not, want not: Revisiting the analysis that called into question the practice of rarefaction

Patrick D. Schloss

doi: https://doi.org/10.1101/2023.06.23.546312



Rarefaction is currently the best approach to control for uneven sequencing effort in amplicon sequence analyses

Patrick D. Schloss

doi: https://doi.org/10.1101/2023.06.23.546313

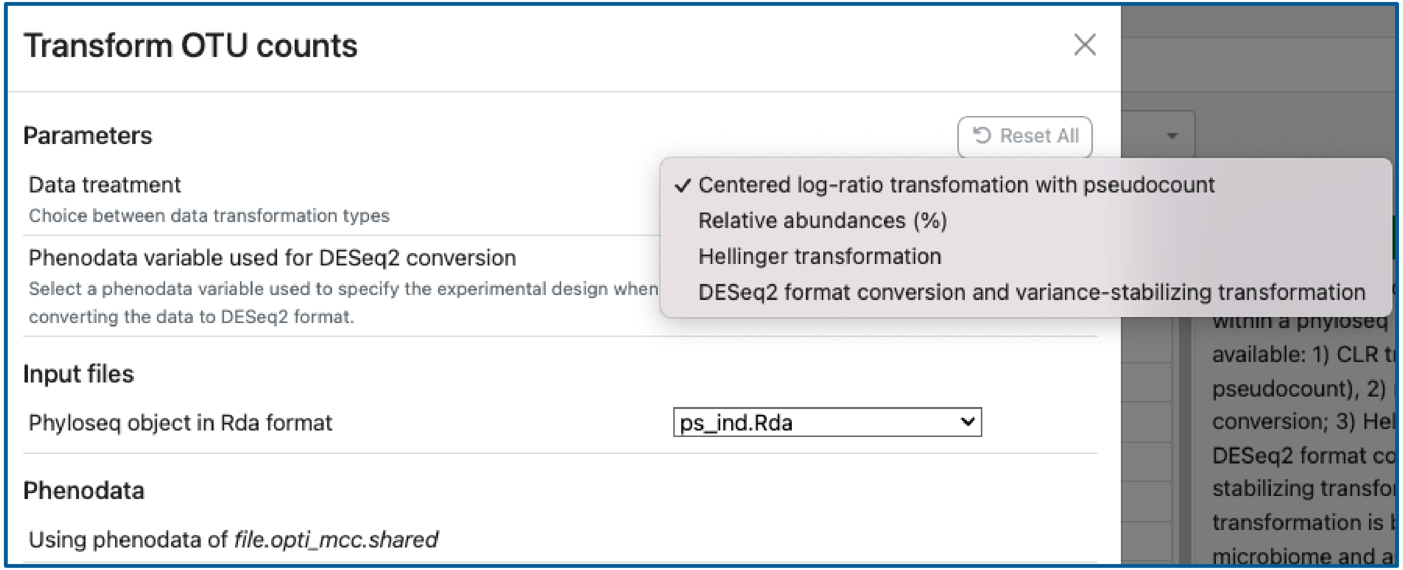
CSC

ICT Solutions for Brilliant Minds

Part 3: Transformations and ordinations

Transformation of OTU data

Four options (August 2023)



Transform OTU counts

Parameters

Data treatment
Choice between data transformation types

✓ Centered log-ratio transfomation with pseudocount
Relative abundances (%)
Hellinger transformation
DESeq2 format conversion and variance-stabilizing transformation

Phenodata variable used for DESeq2 conversion
Select a phenodata variable used to specify the experimental design when converting the data to DESeq2 format.

Input files

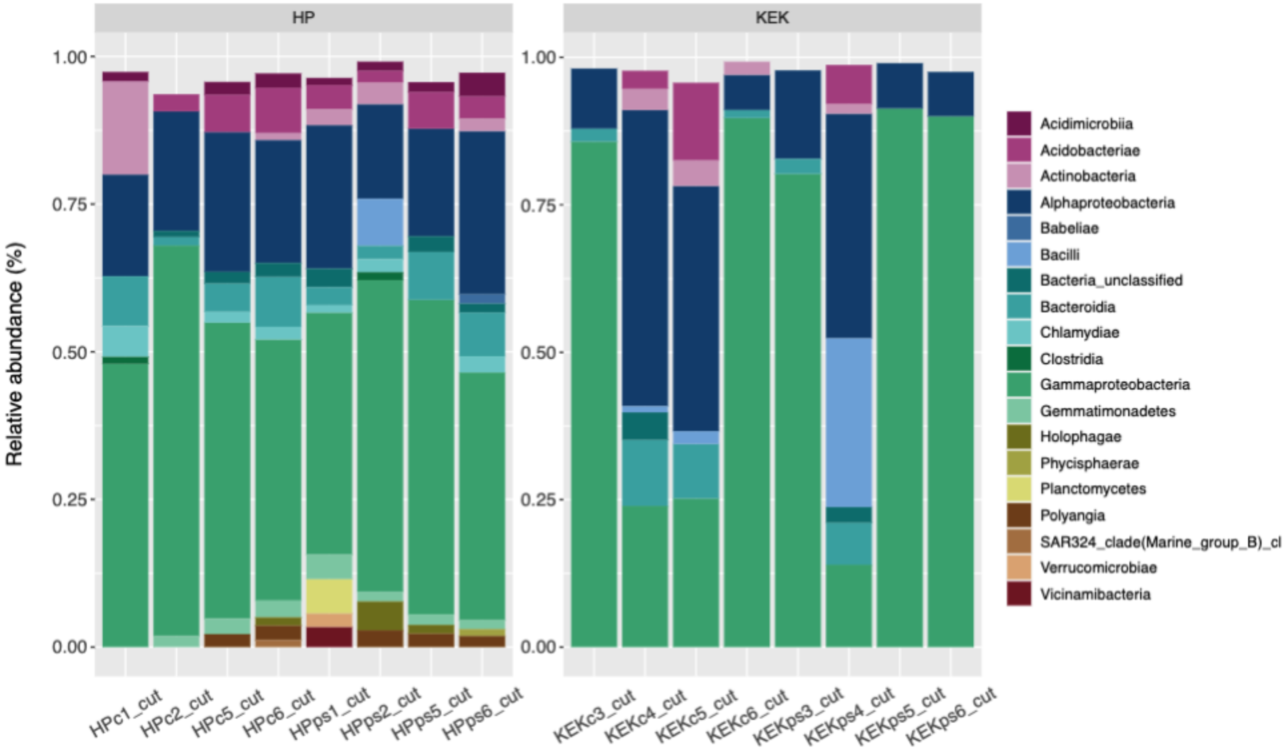
Phyloseq object in Rda format: ps_ind.Rda

Phenodata

Using phenodata of *file.opti_mcc.shared*

within a phyloseq
available: 1) CLR t
pseudocount), 2)
conversion; 3) Hel
DESeq2 format co
stabilizing transfo
transformation is
microbiome and a

Relative abundance (%) bar plots



Beta diversity

Change in community composition among habitats

- Does the microbial community composition in treatment A differ from treatment B?

- Unlike in alpha diversity, OTU identity matters
- Both identity and relative abundance of OTUs usually included
- Quantified with **distance or dissimilarity measures**

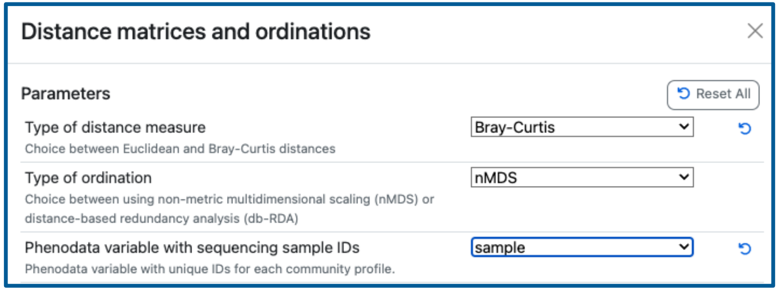
Distance matrices and ordinations

Distance measures available: **Bray-Curtis** or Euclidean

- centered log ratio (CLR) transformation + Euclidean distance = **Aitchinson distance**

Ordinations: visualizing beta diversity

- nMDS (non-metric multidimensional scaling)
  - overall variation among samples displayed
- db-RDA (distance-based redundancy analysis)
  - focus on the variation explained by phenodata variable(s)



Distance matrices and ordinations

Parameters

Type of distance measure: Bray-Curtis
Choice between Euclidean and Bray-Curtis distances

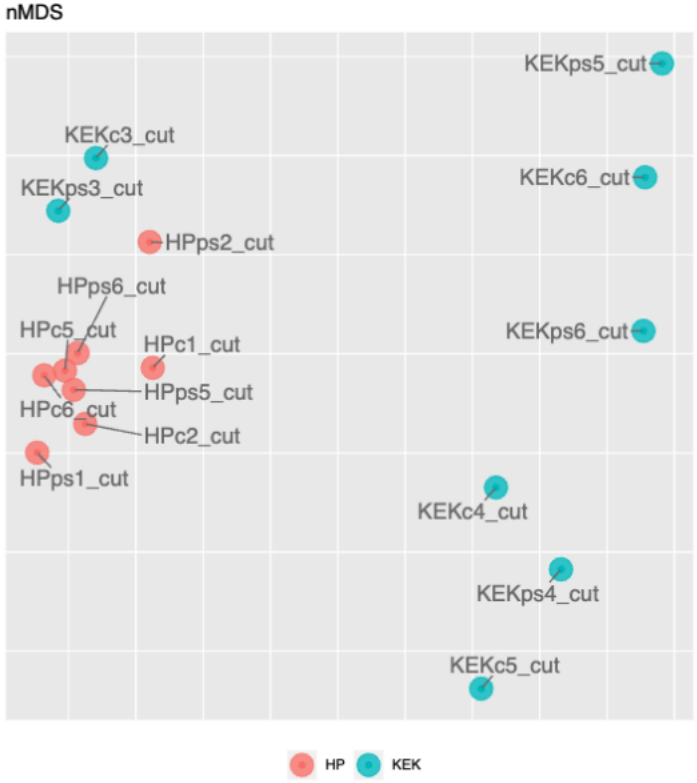
Type of ordination: nMDS
Choice between using non-metric multidimensional scaling (nMDS) or distance-based redundancy analysis (db-RDA)

Phenodata variable with sequencing sample IDs: sample
Phenodata variable with unique IDs for each community profile.

Recommended for more information:

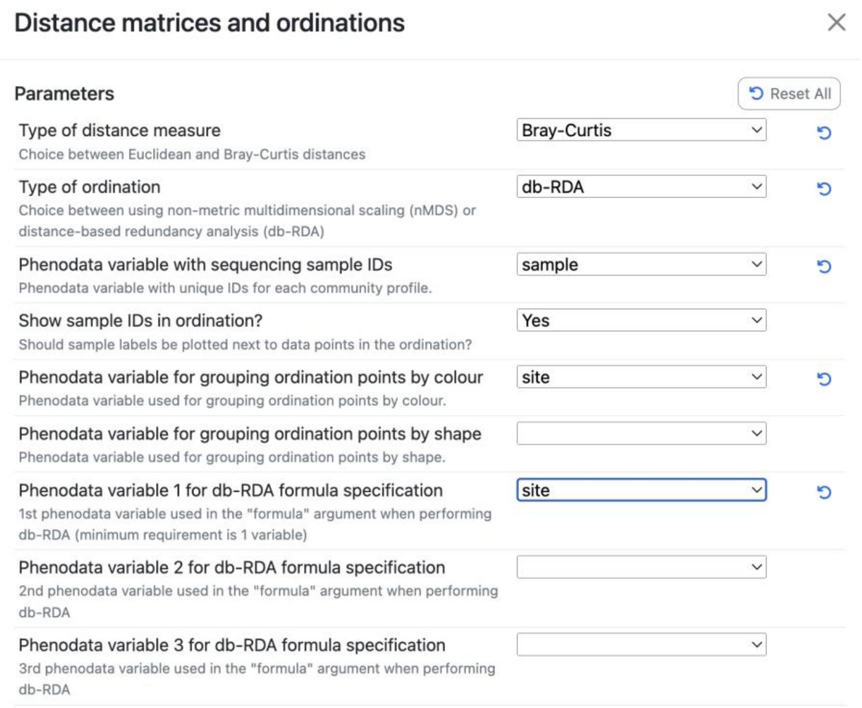
Guide to Statistical Analysis in Microbial Ecology: https://sites.google.com/site/mb3gustame/

Non-metric multidimensional scaling (nMDS)



Distance-based redundancy analysis (db-RDA)

Requires specifying one or more phenodata variables



Distance matrices and ordinations

Parameters

Type of distance measure: Bray-Curtis
Choice between Euclidean and Bray-Curtis distances

Type of ordination: db-RDA
Choice between using non-metric multidimensional scaling (nMDS) or distance-based redundancy analysis (db-RDA)

Phenodata variable with sequencing sample IDs: sample
Phenodata variable with unique IDs for each community profile.

Show sample IDs in ordination?: Yes
Should sample labels be plotted next to data points in the ordination?

Phenodata variable for grouping ordination points by colour: site
Phenodata variable used for grouping ordination points by colour.

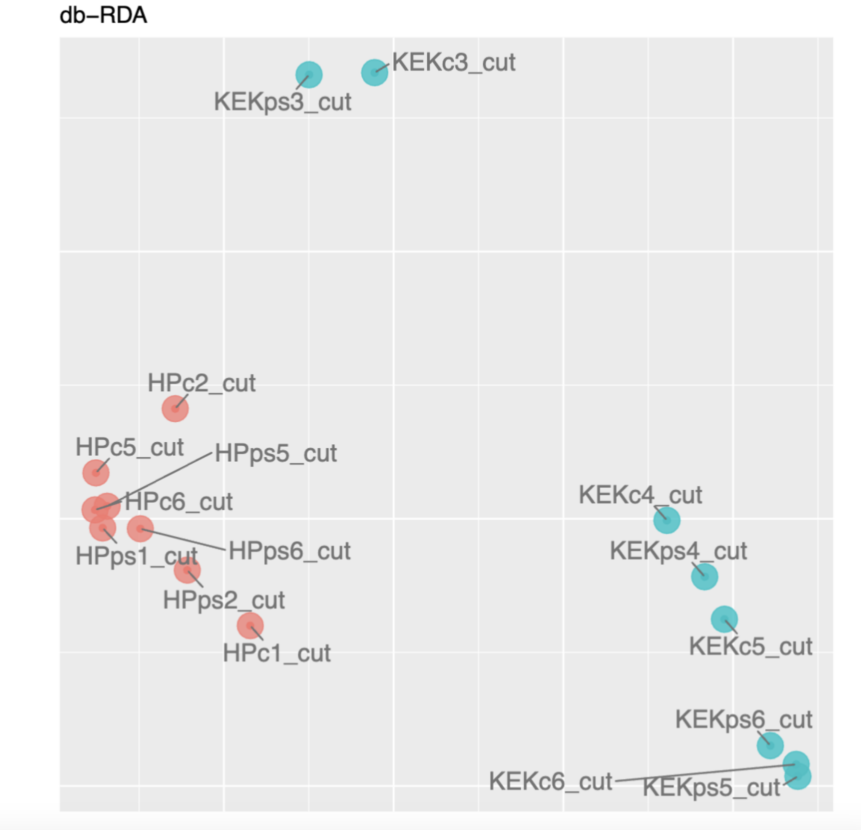
Phenodata variable for grouping ordination points by shape:
Phenodata variable used for grouping ordination points by shape.

Phenodata variable 1 for db-RDA formula specification: site
1st phenodata variable used in the "formula" argument when performing db-RDA (minimum requirement is 1 variable)

Phenodata variable 2 for db-RDA formula specification:
2nd phenodata variable used in the "formula" argument when performing db-RDA

Phenodata variable 3 for db-RDA formula specification:
3rd phenodata variable used in the "formula" argument when performing db-RDA

Distance-based redundancy analysis (db-RDA)



Constrained ordination = focus on the community variation explained by the phenodata variable(s)

Recommended for more information:
Guide to Statistical Analysis in Microbial Ecology: https://sites.google.com/site/mb3gustame/

CSC

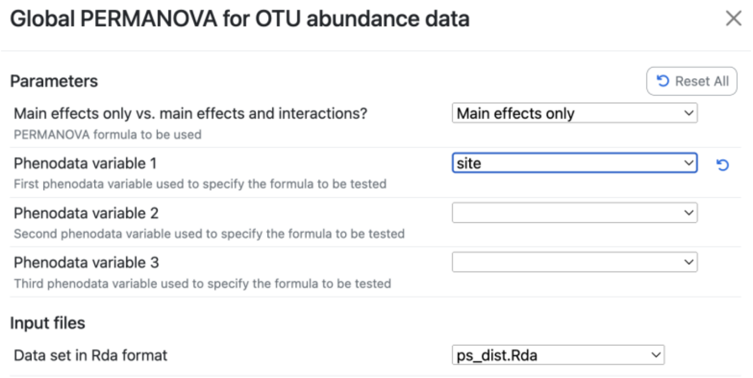
ICT Solutions for Brilliant Minds

Part 4: Statistics

PERMANOVA (permutational multivariate analysis of variance)

Input: distance matrix (ps_dist.Rda)

- Global test: ‘**Does community structure differ between sample groups**?’
    - Pairwise test: ‘Which groups differ from one another?’
- Currently: several phenodata variables -> added sequentially -> order matters!
- Influenced by both **location** and **dispersion**



Global PERMANOVA for OTU abundance data

Parameters

Main effects only vs. main effects and interactions?: Main effects only
PERMANOVA formula to be used

Phenodata variable 1: site
First phenodata variable used to specify the formula to be tested

Phenodata variable 2:
Second phenodata variable used to specify the formula to be tested

Phenodata variable 3:
Third phenodata variable used to specify the formula to be tested

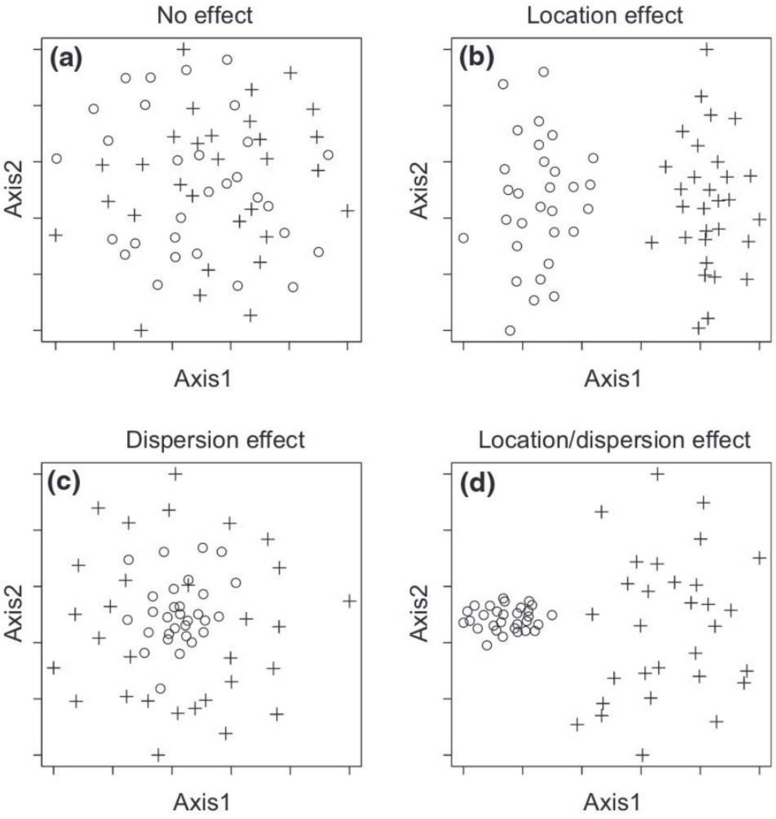
Input files

Data set in Rda format: ps_dist.Rda

Location vs. dispersion

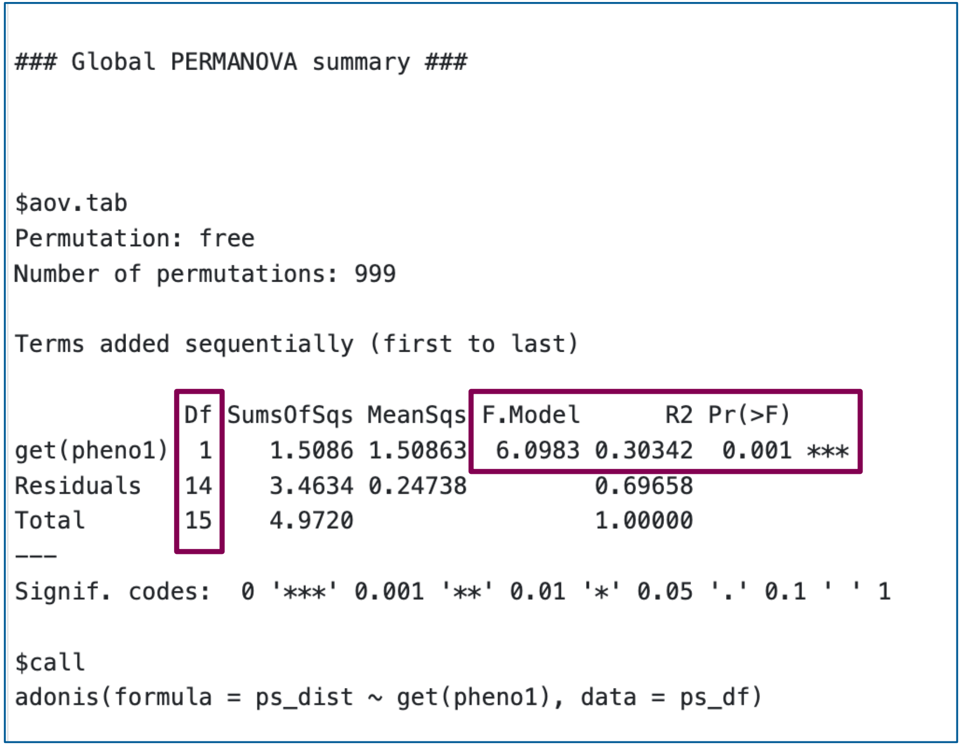
A significant PERMANOVA result can be due to:

- Location effect
- Dispersion effect
- Combination of both



Source: https://doi.org/10.1111/j.2041-210X.2011.00127.x

PERMANOVA output



```
### Global PERMANOVA summary ###



$aov.tab
Permutation: free
Number of permutations: 999

Terms added sequentially (first to last)

            Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
get(pheno1)  1    1.5086 1.50863  6.0983 0.30342  0.001 ***
Residuals   14    3.4634 0.24738         0.69658
Total       15    4.9720                 1.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$call
adonis(formula = ps_dist ~ get(pheno1), data = ps_df)
```

Df: Degrees of freedom

F.Model: Test statistic (pseudo-*F*)

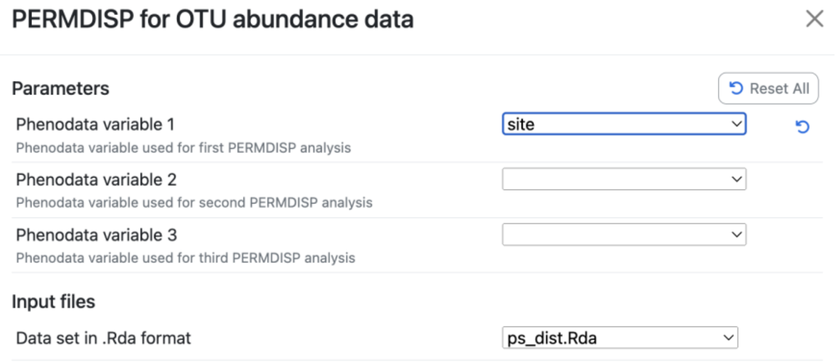
R2: Variation explained by the model

Pr(>F): Statistical significance (*p* value)

PERMDISP: test for the homogeneity of multivariate dispersion

Input: distance matrix

- Test if a significant PERMANOVA result is due to dispersion, not (only) location
- Significant result -> be careful with PERMANOVA interpretation



PERMDISP for OTU abundance data

Parameters

Phenodata variable 1: site
Phenodata variable used for first PERMDISP analysis

Phenodata variable 2
Phenodata variable used for second PERMDISP analysis

Phenodata variable 3
Phenodata variable used for third PERMDISP analysis

Input files

Data set in .Rda format: ps_dist.Rda

Post-hoc comparisons

Sample group comparisons **following significant global PERMANOVA:**

- Pairwise PERMANOVA (similar as global test but for sample pairs)

Dispersion comparisons **following significant PERMDISP:**

- Tukey's Honestly Significant Difference (HSD) test

- Both methods use a correction for multiple testing (Benjamin-Hochberg correction)

DESeq2

- Originates from the RNAseq field
- Addresses the question: '**Which taxa are differentially abundant between sample groups?'**
- Enables inferences such as: 'Illness x is associated with a reduction in the abundance of beneficial gut microbes y and z'
- Input: untransformed data -> convert to DESeq2 format with Transform OTU counts (corrects for differences in sequencing depth)
- Results given as **log fold changes**

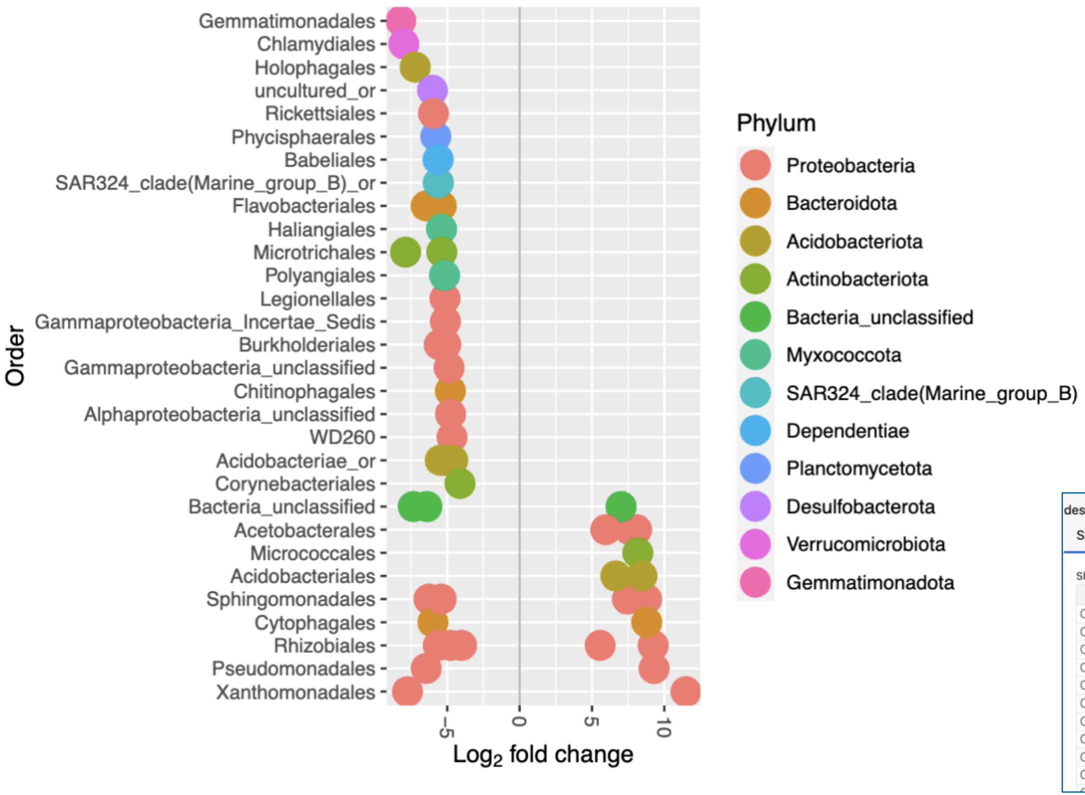
More information: https://joey711.github.io/phyloseq-extensions/DESeq2.html

DESeq2

Current tool configuration (August 2023):

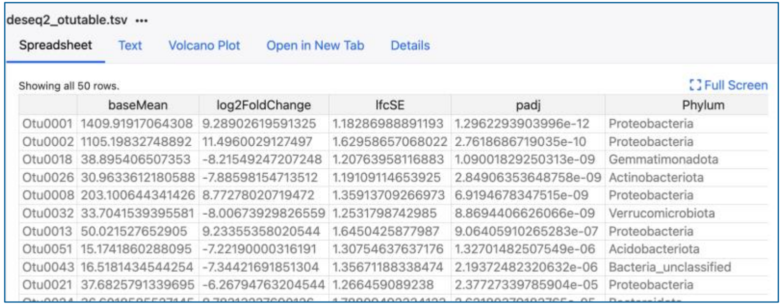
- **Focused on comparison of two groups at a time**
  - If selected phenodata column has >2 groups, specify a pair (Group 1 and Group 2)
- **Reference level selection**:
  - Phenodata column with two groups -> first in alphabet is the reference level (e.g. ’b vs. a’ or ‘sick vs. healthy’=
  - Phenodata column with >2 groups -> ‘Group 2’ is the reference level

DESeq2



8-fold increase compared to reference level = log2 fold change 3 (because $2^3=8$)

each dot = OTU with adjusted p value < 0.01



Showing all 50 rows.

	baseMean	log2FoldChange	lfcSE	padj	Phylum
Otu0001	1409.91917064308	9.28902619591325	1.18286988891193	1.2962293903996e-12	Proteobacteria
Otu0002	1105.19832748892	11.4960029127497	1.62958657068022	2.7618686719035e-10	Proteobacteria
Otu0018	38.895406507353	-8.21549247207248	1.20763958116883	1.09001829250313e-09	Gemmatimonadota
Otu0026	30.9633612180588	-7.88598154713512	1.19109114653925	2.84906353648758e-09	Actinobacteriota
Otu0008	203.100644341426	8.77278020719472	1.35913709266973	6.9194678347515e-09	Proteobacteria
Otu0032	33.7041539395581	-8.00673929826559	1.2531798742985	8.8694406626066e-09	Verrucomicrobiota
Otu0013	50.021527652905	9.23355358020544	1.6450425877987	9.06405910265283e-07	Proteobacteria
Otu0051	15.1741860288095	-7.22190000316191	1.30754637637176	1.32701482507549e-06	Acidobacteriota
Otu0043	16.5181434544254	-7.34421691851304	1.35671188338474	2.19372482320632e-06	Bacteria_unclassified
Otu0021	37.6825791339695	-6.26794763204544	1.266459089238	2.37727339785904e-05	Proteobacteria