CSC

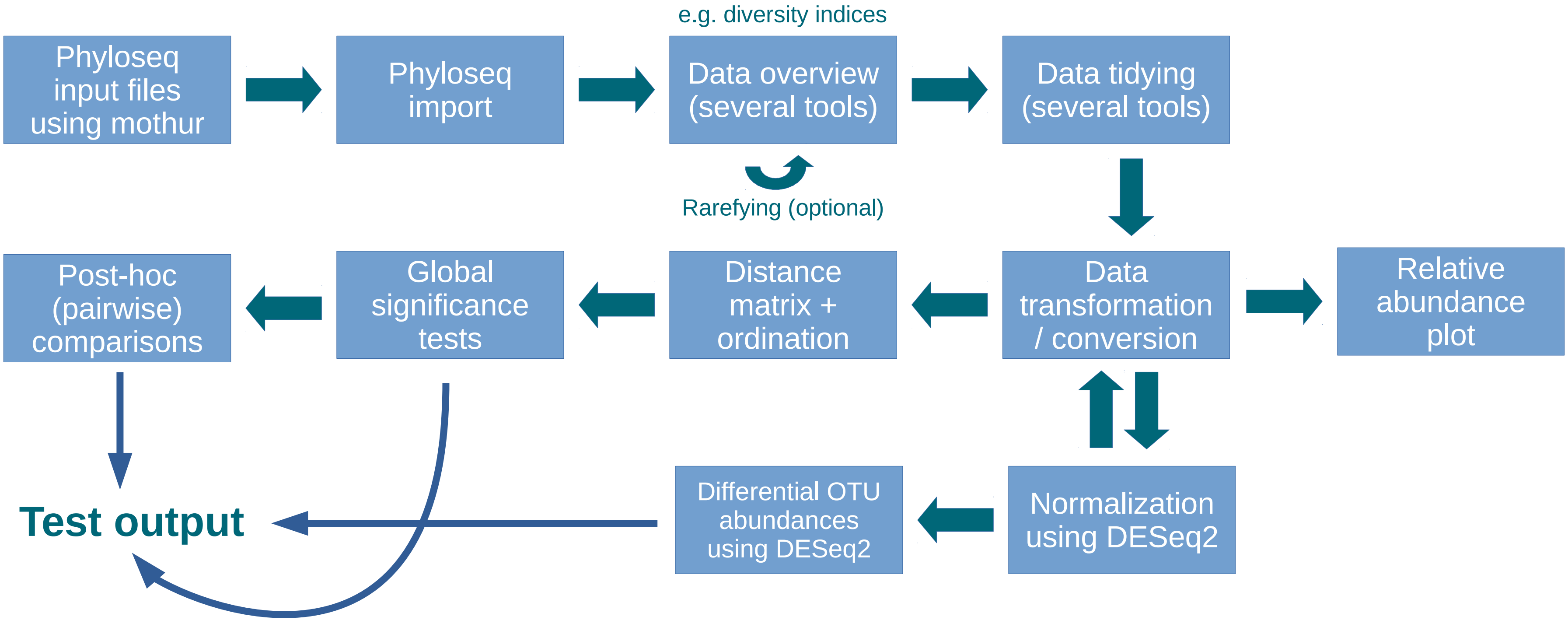
ICT Solutions for Brilliant Minds

Microbial community analysis using Chipster: data tidying, visualization and statistics

Part 1: Tool overview and data importing

Jesse Harrison and Eija Korpelainen
CSC – IT Center for Science Ltd.
chipster@csc.fi

Workflow for data tidying, analysis and statistics



Phyloseq input files using mothur

Phyloseq import

e.g. diversity indices

Data overview (several tools)

Rarefying (optional)

Data tidying (several tools)

Data transformation / conversion

Relative abundance plot

Distance matrix + ordination

Global significance tests

Post-hoc (pairwise) comparisons

Normalization using DESeq2

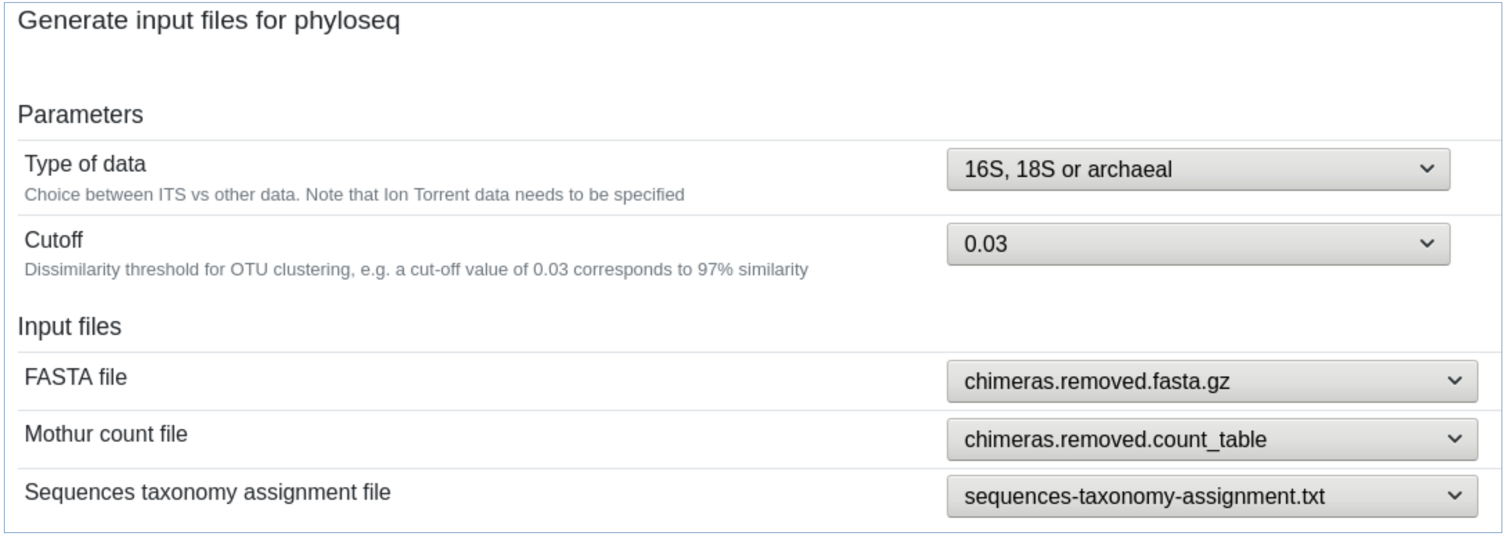
Differential OTU abundances using DESeq2

Test output

Generating phyloseq input files

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, 18S or archaeal

Choice between ITS vs other data. Note that Ion Torrent data needs to be specified

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

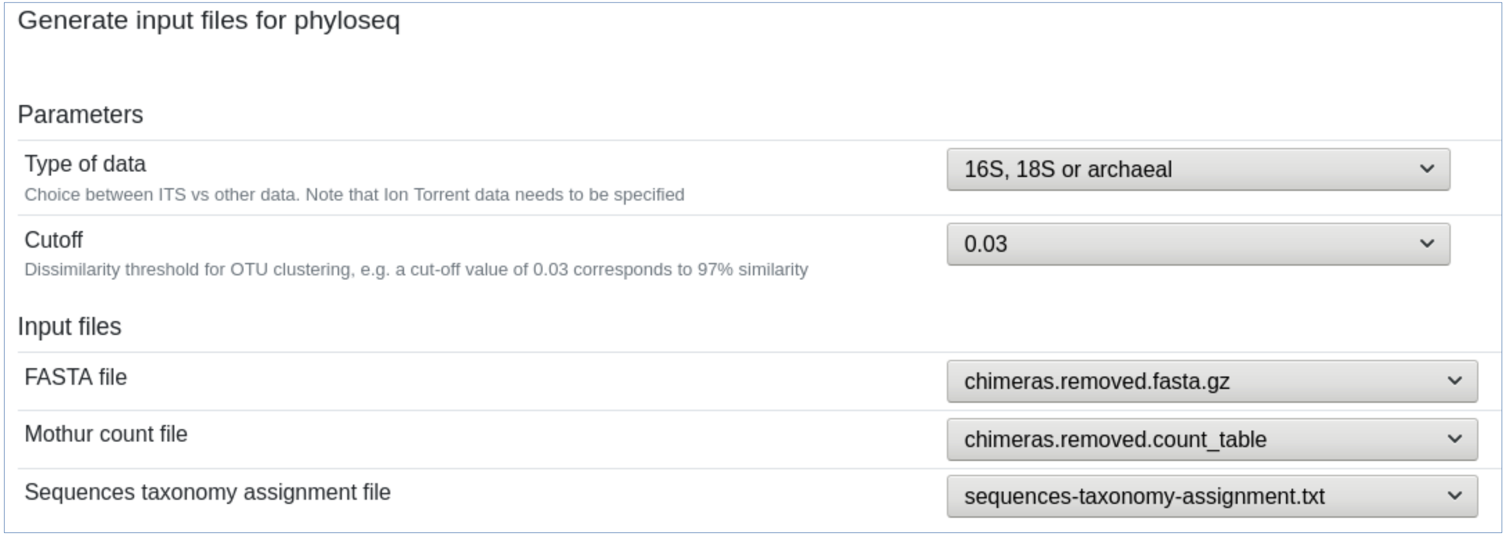
Mothur count file: chimeras.removed.count_table

Sequences taxonomy assignment file: sequences-taxonomy-assignment.txt

Generating phyloseq input files

Specifications for creating phyloseq input files:

Type of data, % cutoff and files produced by mothur (FASTA, count file, taxonomy file)



Generate input files for phyloseq

Parameters

Type of data: 16S, 18S or archaeal

Choice between ITS vs other data. Note that Ion Torrent data needs to be specified

Cutoff: 0.03

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

Input files

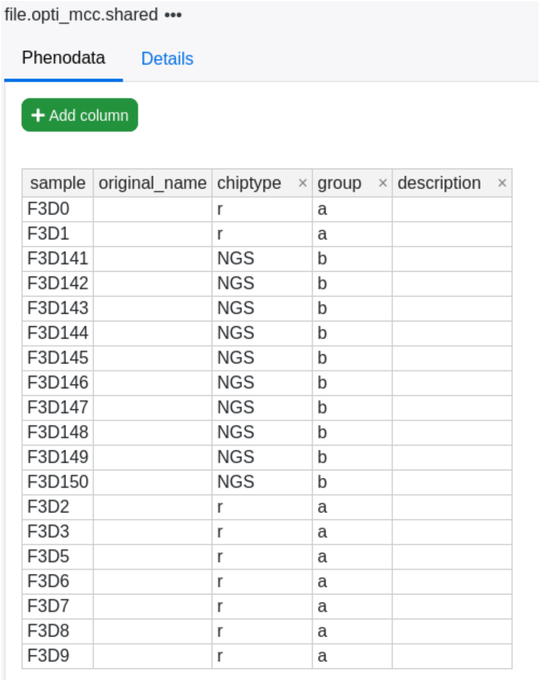
FASTA file: chimeras.removed.fasta.gz

Mothur count file: chimeras.removed.count_table

Sequences taxonomy assignment file: sequences-taxonomy-assignment.txt

The phenodata file

Generated input files: .shared + phenodata file, consensus taxonomy file

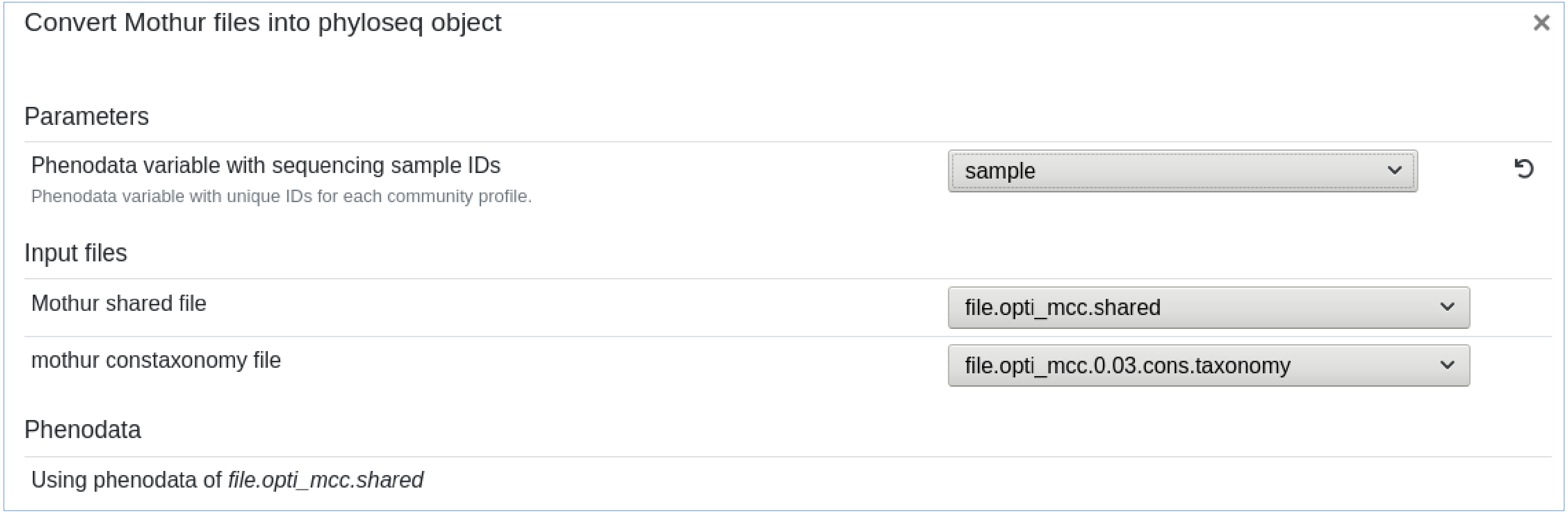


file.opti_mcc.shared •••

sample	original_name	chiptype ×	group ×	description ×
F3D0		r	a	
F3D1		r	a	
F3D141		NGS	b	
F3D142		NGS	b	
F3D143		NGS	b	
F3D144		NGS	b	
F3D145		NGS	b	
F3D146		NGS	b	
F3D147		NGS	b	
F3D148		NGS	b	
F3D149		NGS	b	
F3D150		NGS	b	
F3D2		r	a	
F3D3		r	a	
F3D5		r	a	
F3D6		r	a	
F3D7		r	a	
F3D8		r	a	
F3D9		r	a	

The phenodata file is an editable table with unique IDs for each sample and sample groupings

Converting input files into a phyloseq object



Convert Mothur files into phyloseq object

Parameters

Phenodata variable with sequencing sample IDs: sample

Phenodata variable with unique IDs for each community profile.

Input files

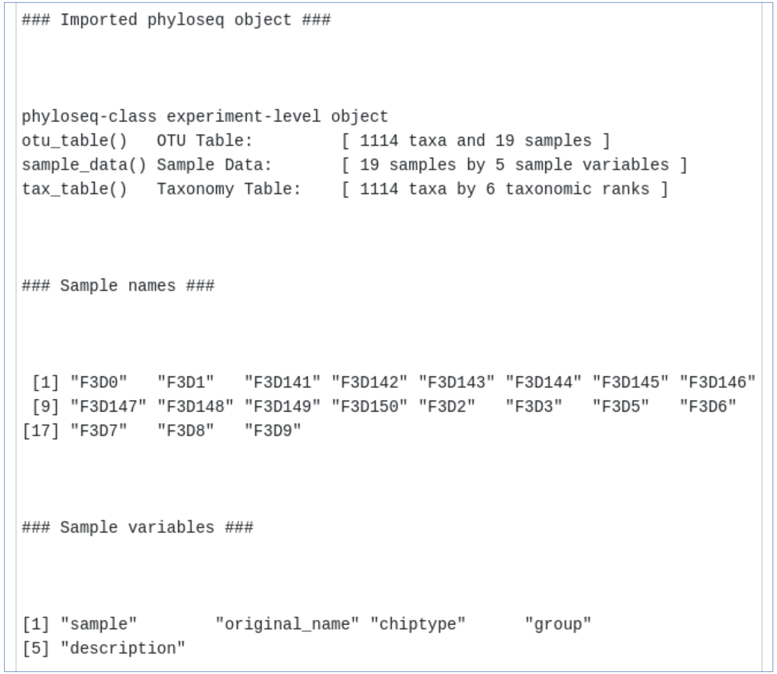
Mothur shared file: file.opti_mcc.shared

mothur constaxonomy file: file.opti_mcc.0.03.cons.taxonomy

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:         [ 1114 taxa and 19 samples ]
sample_data() Sample Data:       [ 19 samples by 5 sample variables ]
tax_table()   Taxonomy Table:    [ 1114 taxa by 6 taxonomic ranks ]



### Sample names ###



 [1] "F3D0"   "F3D1"   "F3D141" "F3D142" "F3D143" "F3D144" "F3D145" "F3D146"
 [9] "F3D147" "F3D148" "F3D149" "F3D150" "F3D2"   "F3D3"   "F3D5"   "F3D6"
[17] "F3D7"   "F3D8"   "F3D9"



### Sample variables ###



[1] "sample"        "original_name" "chiptype"      "group"
[5] "description"
```

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

The Rda file is used as the input for downstream analyses

CSC

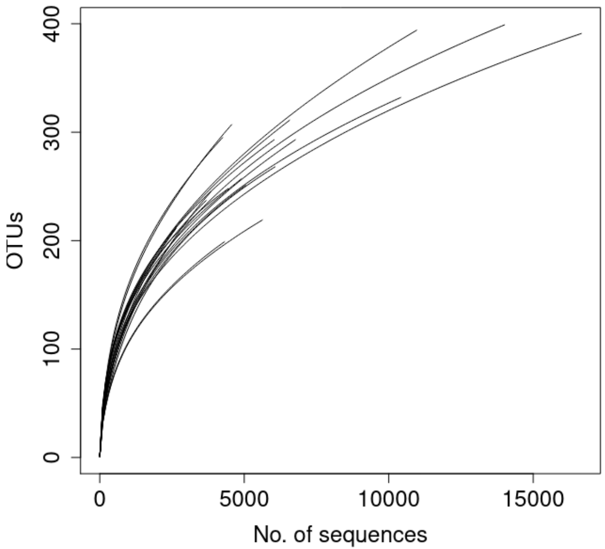
ICT Solutions for Brilliant Minds

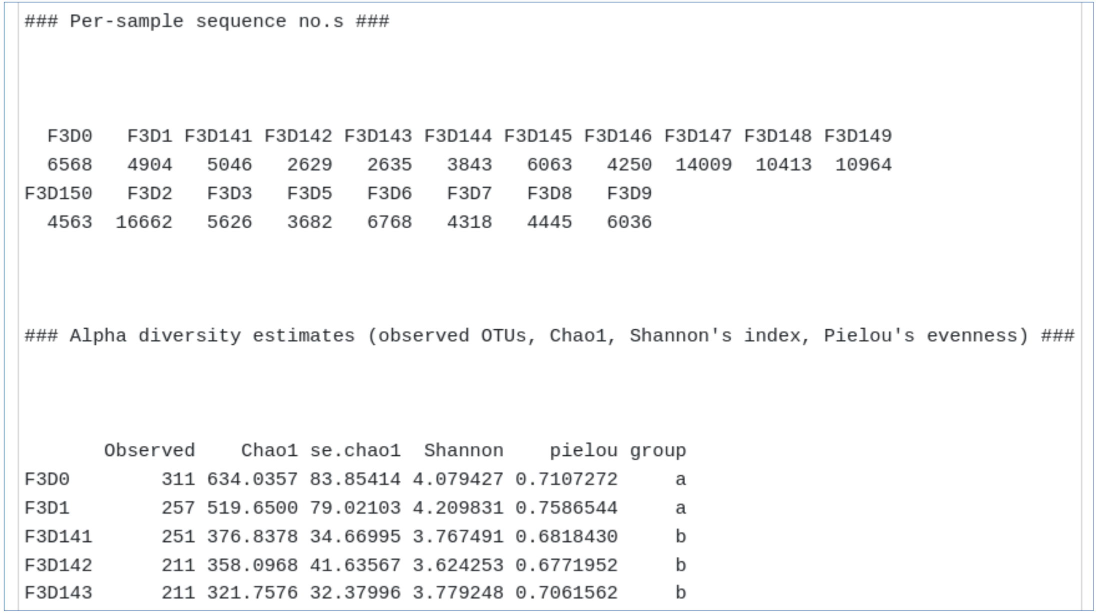
Microbial community analysis using Chipster: data tidying, visualization and statistics

Part 2: Data inspection and tidying

Jesse Harrison and Eija Korpelainen
CSC – IT Center for Science Ltd.
chipster@csc.fi

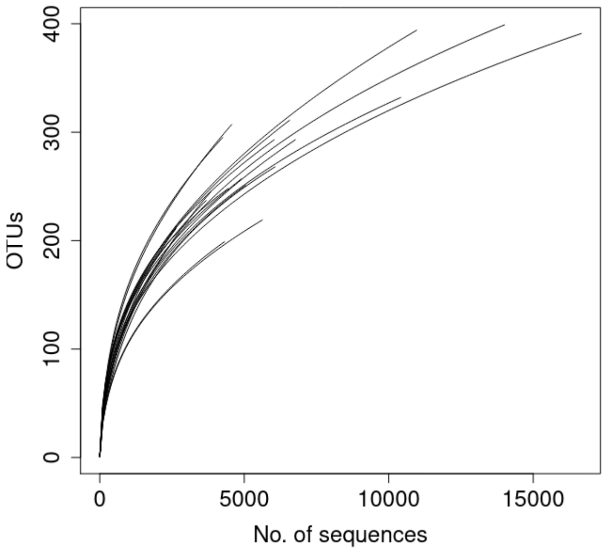
Sequence no.s, rarefaction curves and alpha diversity





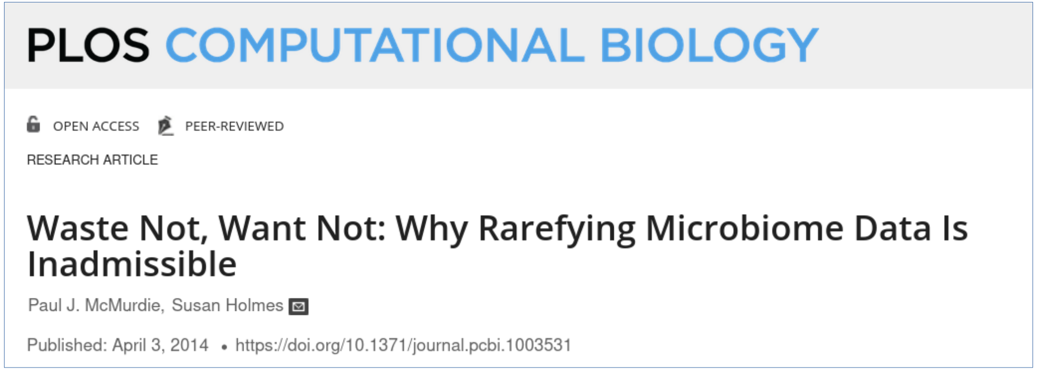
```
### Per-sample sequence no.s ###



  F3D0   F3D1 F3D141 F3D142 F3D143 F3D144 F3D145 F3D146 F3D147 F3D148 F3D149
  6568   4904   5046   2629   2635   3843   6063   4250  14009  10413  10964
F3D150   F3D2   F3D3   F3D5   F3D6   F3D7   F3D8   F3D9
  4563  16662   5626   3682   6768   4318   4445   6036



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



       Observed    Chao1 se.chao1  Shannon    pielou group
F3D0        311 634.0357 83.85414 4.079427 0.7107272     a
F3D1        257 519.6500 79.02103 4.209831 0.7586544     a
F3D141      251 376.8378 34.66995 3.767491 0.6818430     b
F3D142      211 358.0968 41.63567 3.624253 0.6771952     b
F3D143      211 321.7576 32.37996 3.779248 0.7061562     b
```

Sequence no.s, rarefaction curves and alpha diversity



Previously often used for library size normalization, but increasing evidence for drawbacks

Alternatives: data transformation / corrections for unequal library size



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

Taxon-level clean-up tools

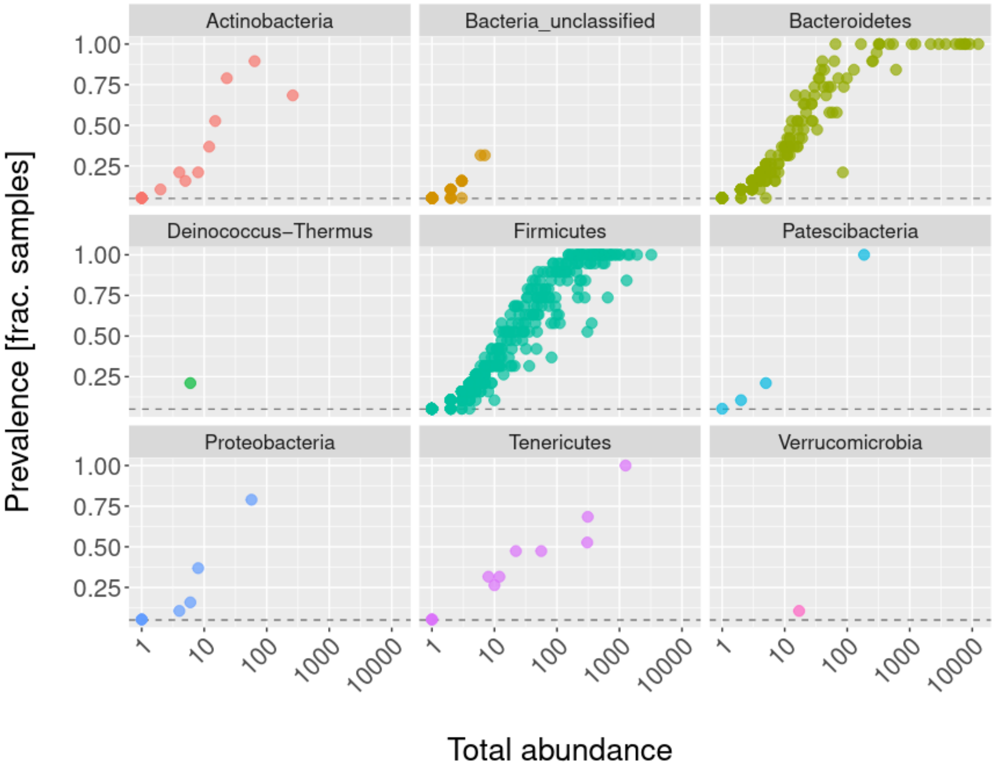
- Taxon composition overview (user-specified level)
- Removing non-specific sequences (keep e.g. Bacteria or Archaea only)
- Remove chloroplast and/or mitochondrial sequences
- Manually remove specific taxa

Split into a total of three different tools

Visualizing and filtering low-abundance features

- **Prevalence**
  - Definition: no. of samples in which a taxon appears at least once
  - Visualization and filtering tool
- **Singletons and doubletons**
  - Text summary and filtering tool

Visualizing and filtering low-abundance features



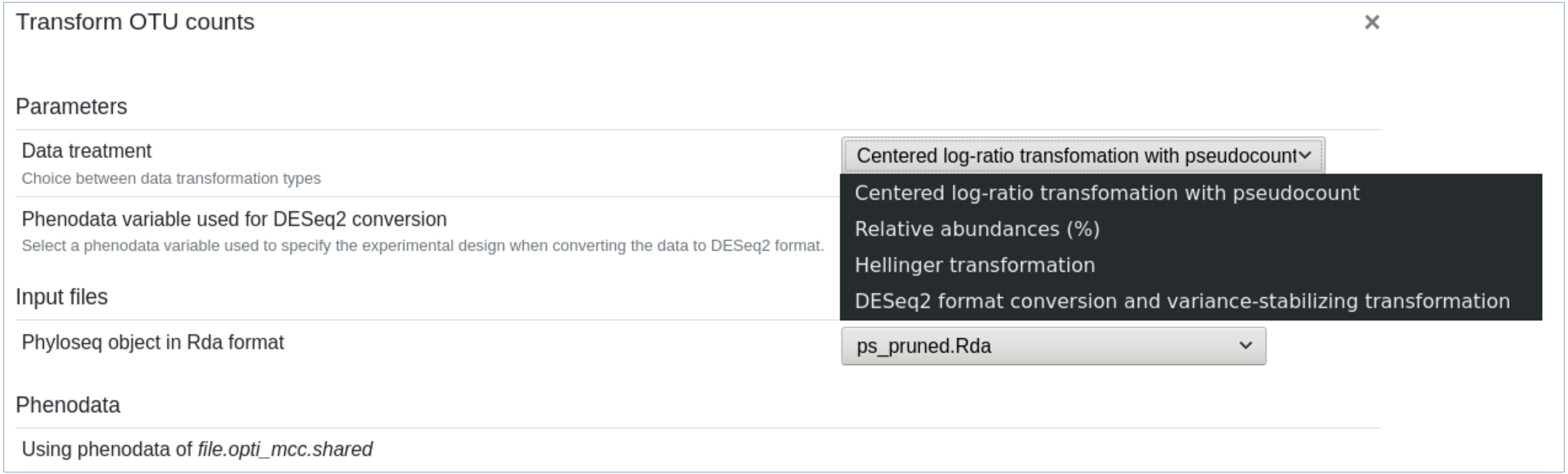
Microbial community analysis using Chipster: data tidying, visualization and statistics

Part 3: Transformations and ordinations

Jesse Harrison and Eija Korpelainen
CSC – IT Center for Science Ltd.
chipster@csc.fi

Data transformation

Four options (April 2021)



Transform OTU counts

Parameters

Data treatment

Choice between data transformation types

Centered log-ratio transfomation with pseudocount

- 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

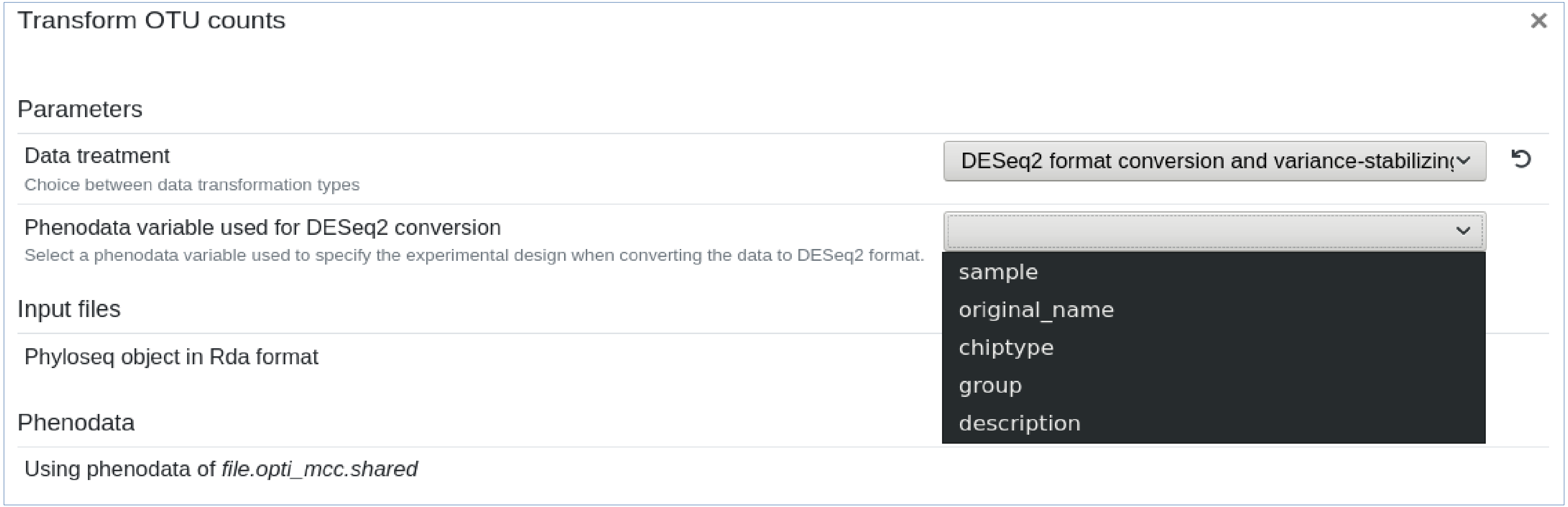
ps_pruned.Rda

Phenodata

Using phenodata of *file.opti_mcc.shared*

Data transformation

Four options (April 2021)



Transform OTU counts

Parameters

Data treatment

Choice between data transformation types

DESeq2 format conversion and variance-stabilizin

Phenodata variable used for DESeq2 conversion

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

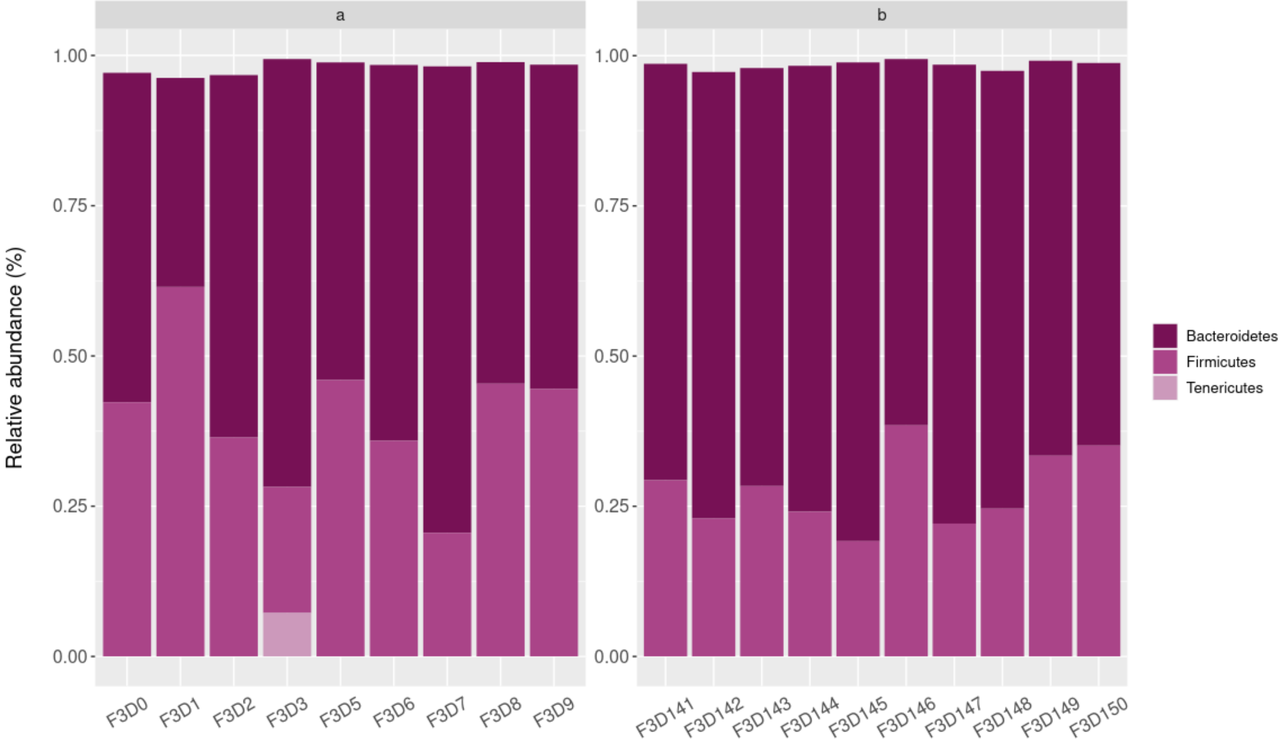
Input files

Phyloseq object in Rda format

Phenodata

Using phenodata of *file.opti_mcc.shared*

Relative abundance (%) bar plots

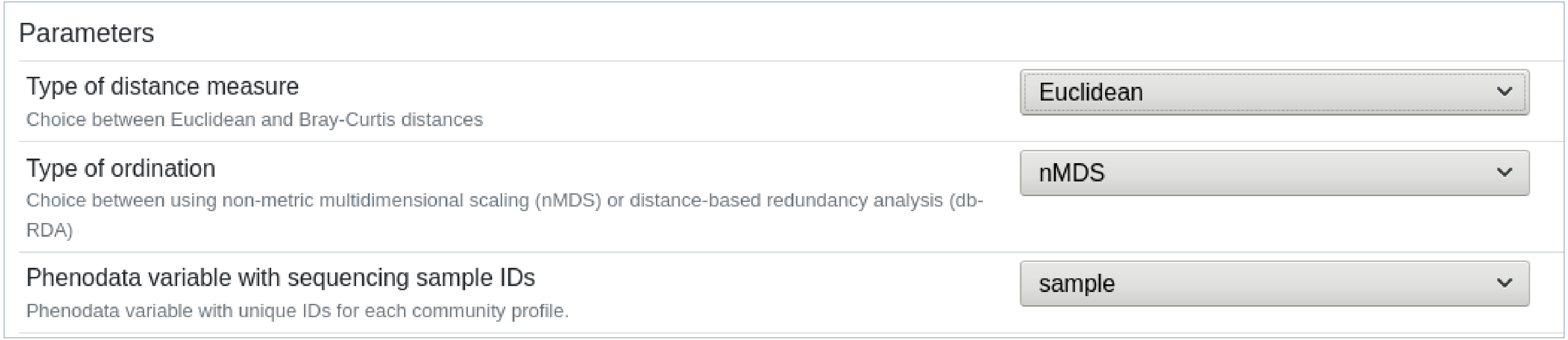


Distance matrices and ordinations

Distance measures: Euclidean or Bray-Curtis

Types of ordination: nMDS or db-RDA

(Aitchison distance = CLR + Euclidean)



Parameters

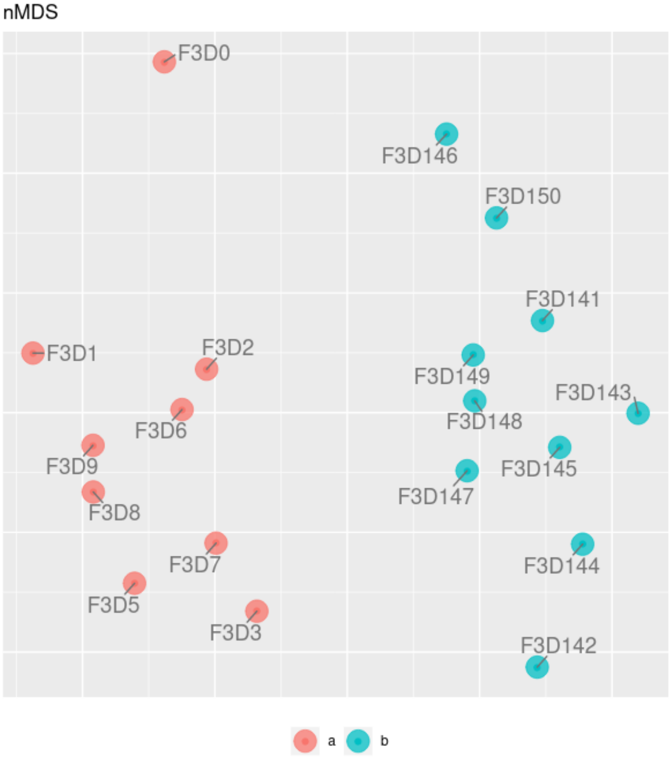
Type of distance measure
Choice between Euclidean and Bray-Curtis distances
Euclidean

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

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

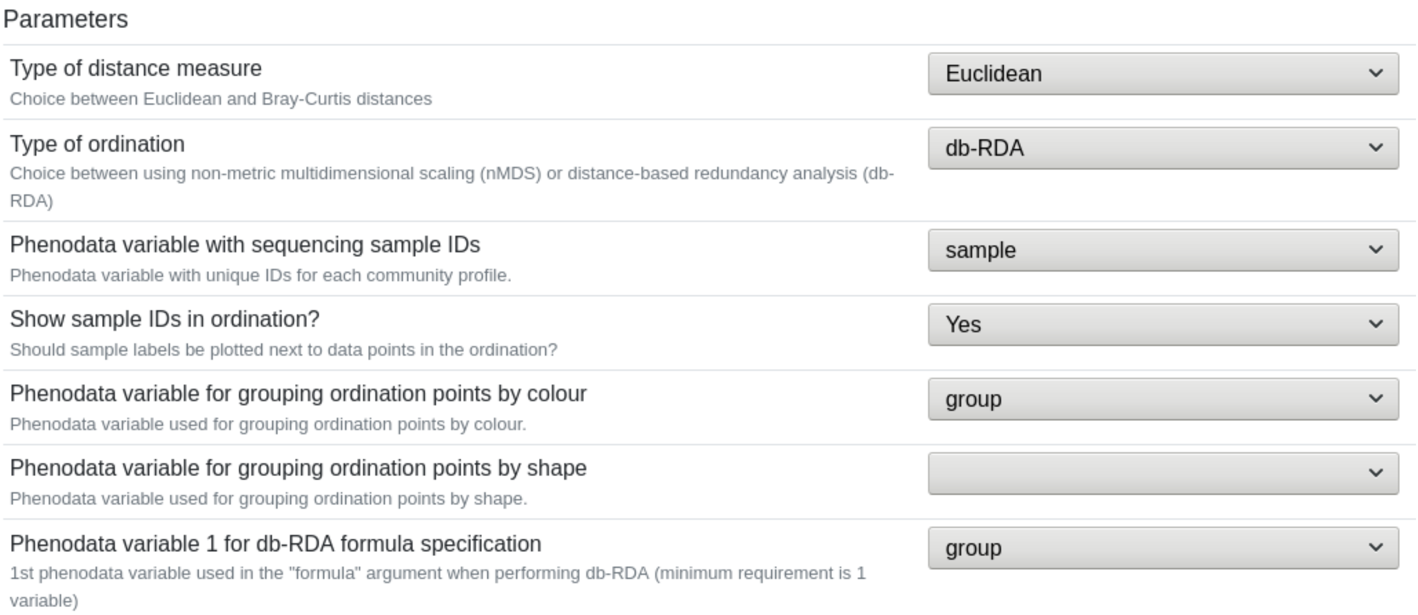
Guide to Statistical Analysis in Microbial Ecology:
https://sites.google.com/site/mb3gustame/

Non-metric multidimensional scaling (nMDS)



Distance-based redundancy analysis (db-RDA)

Requires specification of a phenodata variable



Parameters

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

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

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

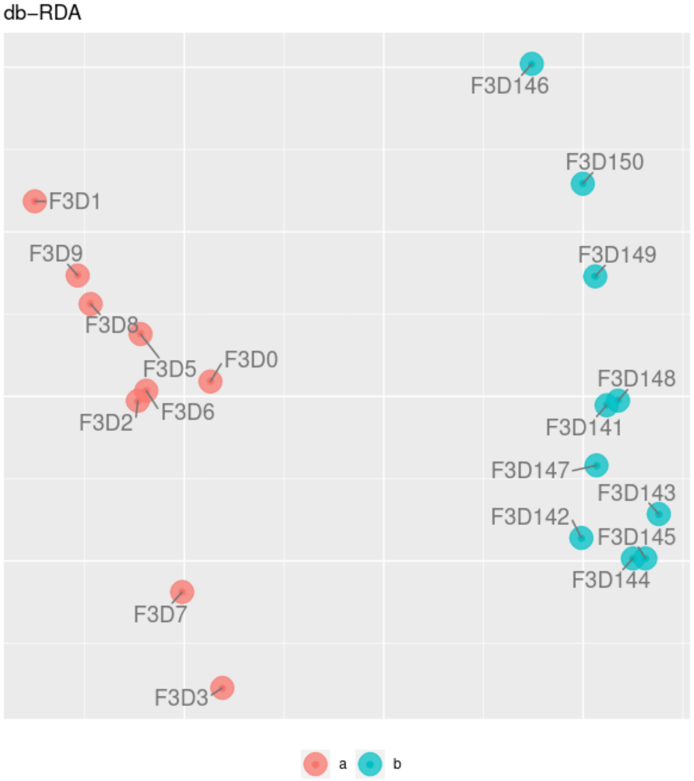
Show sample IDs in ordination?
Should sample labels be plotted next to data points in the ordination?
Yes

Phenodata variable for grouping ordination points by colour
Phenodata variable used for grouping ordination points by colour.
group

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
1st phenodata variable used in the "formula" argument when performing db-RDA (minimum requirement is 1 variable)
group

Distance-based redundancy analysis (db-RDA)



CSC

ICT Solutions for
Brilliant Minds

Microbial community analysis using Chipster: data tidying, visualization and statistics

Part 4: Statistics

Jesse Harrison and Eija Korpelainen
CSC – IT Center for Science Ltd.
chipster@csc.fi

PERMANOVA + PERMDISP

Require a distance matrix as the input

PERMANOVA (permutational multivariate analysis of variance)

- Global test: “Does community structure differ between sample groups?”
- Pairwise test: “Which particular groups differ from one another?”
- Influenced by both *location* and *dispersion* (more on these terms on the next slide)

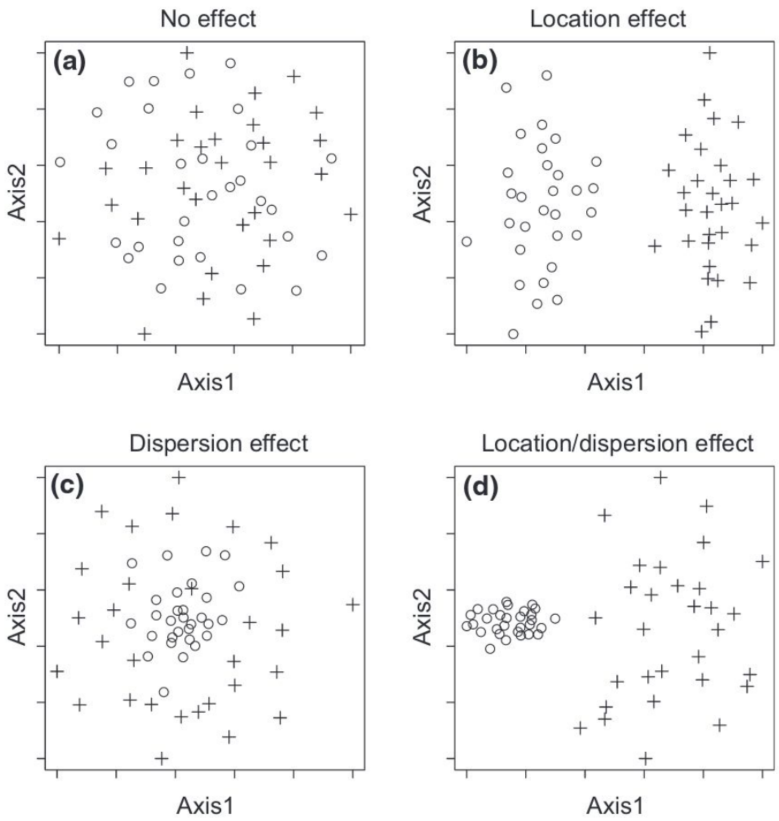
PERMDISP (test for the homogeneity of multivariate dispersions)

- Run only if get a significant (p < 0.05) PERMANOVA result
- Can help tell why the PERMANOVA result is significant

Location vs dispersion

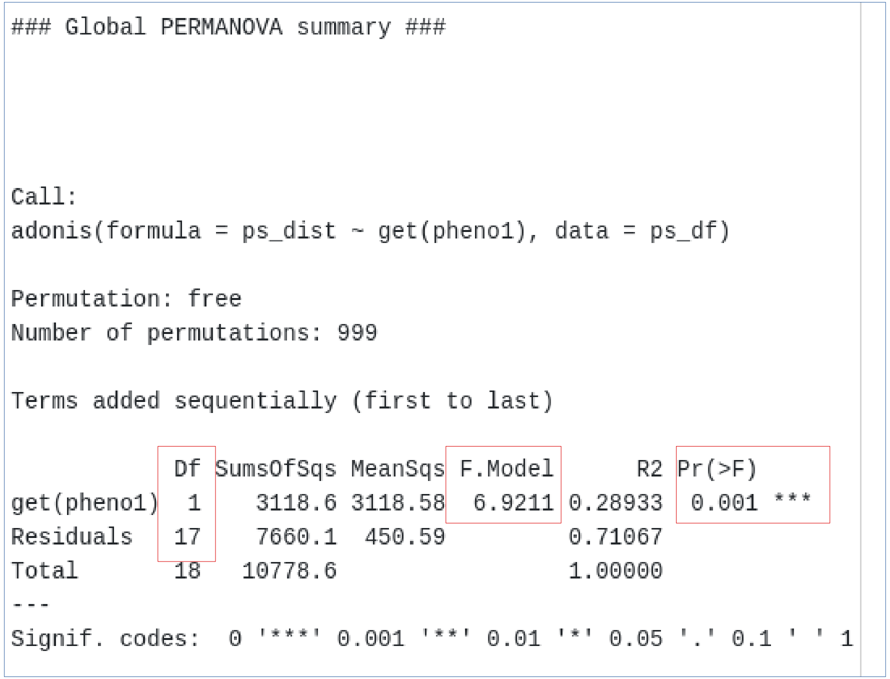
A significant PERMANOVA result can be due to:

- A location effect
- A dispersion effect
- A combination of both



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

PERMANOVA output



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




Call:
adonis(formula = ps_dist ~ get(pheno1), data = ps_df)

Permutation: free
Number of permutations: 999

Terms added sequentially (first to last)

            Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)    
get(pheno1)  1    3118.6 3118.58  6.9211 0.28933  0.001 ***
Residuals   17    7660.1  450.59         0.71067           
Total       18   10778.6                 1.00000           
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Df: Degrees of freedom

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

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

Post-hoc comparisons

Following significant global PERMANOVA:

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

Following significant PERMDISP:

- Tukey’s Honestly Significant Difference (HSD) test

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

DESeq2

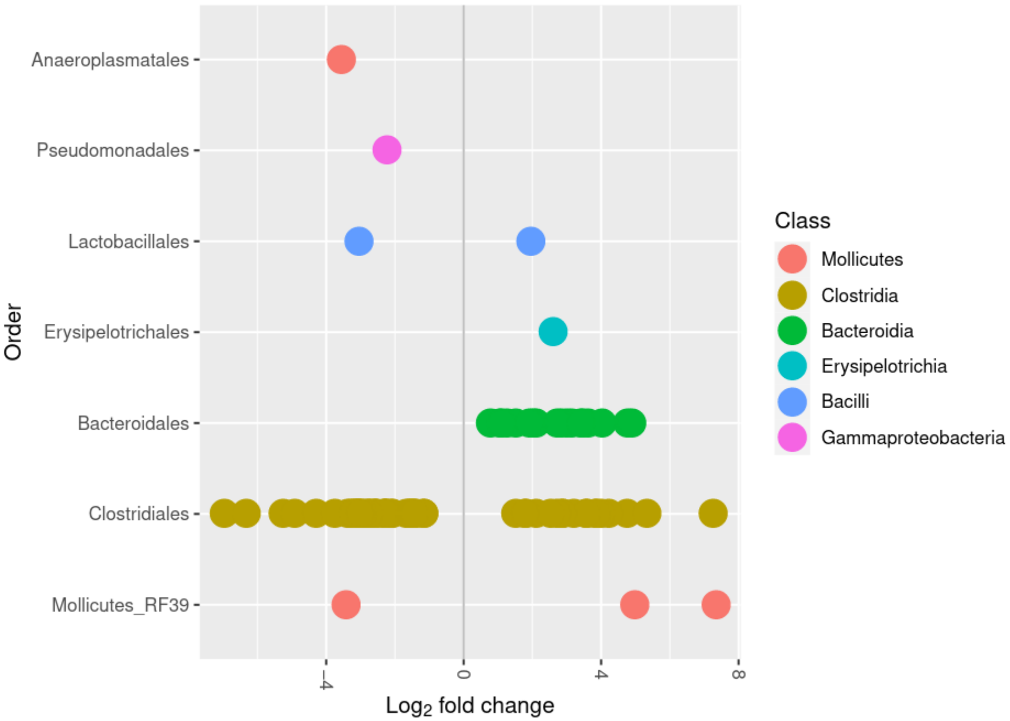
- Originates from the RNAseq field
- Used to address the question: “Which taxa are differentially abundant between sample groups?”
- Enables inferences such as: “Illness *x* is associated with a reduction in the abundance of beneficial gut microbes *y* and *z*”
- Untransformed data used as input (internally corrects for differences in library size)
- Results given as log fold changes
- Link to more info online:
  - https://joey711.github.io/phyloseq-extensions/DESeq2.html

DESeq2

Current tool configuration (April 2021)

- Focuses on comparisons of two groups at a time
  - If selected phenodata column has >2 groups, can specify a pair (Group 1 and Group 2)
- Phenodata column with two groups:
  - Reference level selected alphabetically (e.g. ‘b vs a’ or ‘sick vs healthy’)
- Phenodata column with >2 groups:
  - Reference level corresponds to ‘Group 2’

DESeq2


Anaeroplasmatales
Pseudomonadales
Lactobacillales
Erysipelotrichales
Bacteroidales
Clostridiales
Mollicutes_RF39
Order
-4
0
4
8
Log2 fold change
Class
Mollicutes
Clostridia
Bacteroidia
Erysipelotrichia
Bacilli
Gammaproteobacteria