Metagenomics using Chipster

Statistical analysis of marker gene data
Comparing diversity and abundance between groups
Visualization

Jarno Tuimala PhD, adjunct professor

https://www.csc.fi/web/training/-/metagenomics

Topics

- Visualizations
  - Rarefaction curves (vegan::specaccum)
  - Rank-abundance curves (BiodiversityR::rankabundance)
- Statistical analysis for marker gene studies
  - Contribution diversity approach (vegan::contribdiv)
  - Permutational Multivariate Analysis of Variance Using Distance Matrices (vegan::adonis)
  - Analysis of Molecular Variance (pegas::amova)
  - Multivariate homogeneity of groups dispersions (variances) (vegan::betadisper)
  - Dufrene-Legendre Indicator Species Analysis (labdsv::indval)
  - Indicator Species Analysis Minimizing Intermediate Occurrences (labdsv::isamic)
- "Visual data analysis"
  - RDA plot (vegan::rda)
  - Heatmap (pheatmap::pheatmap)

Demo data

- Costello et al., stool analysis
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3602444/
- https://www.mothur.org/wiki/Costello_stool_analysis

Chipster

- Metagenomics / Statistical analysis for marker gene studies

- Visualisation / Heatmap

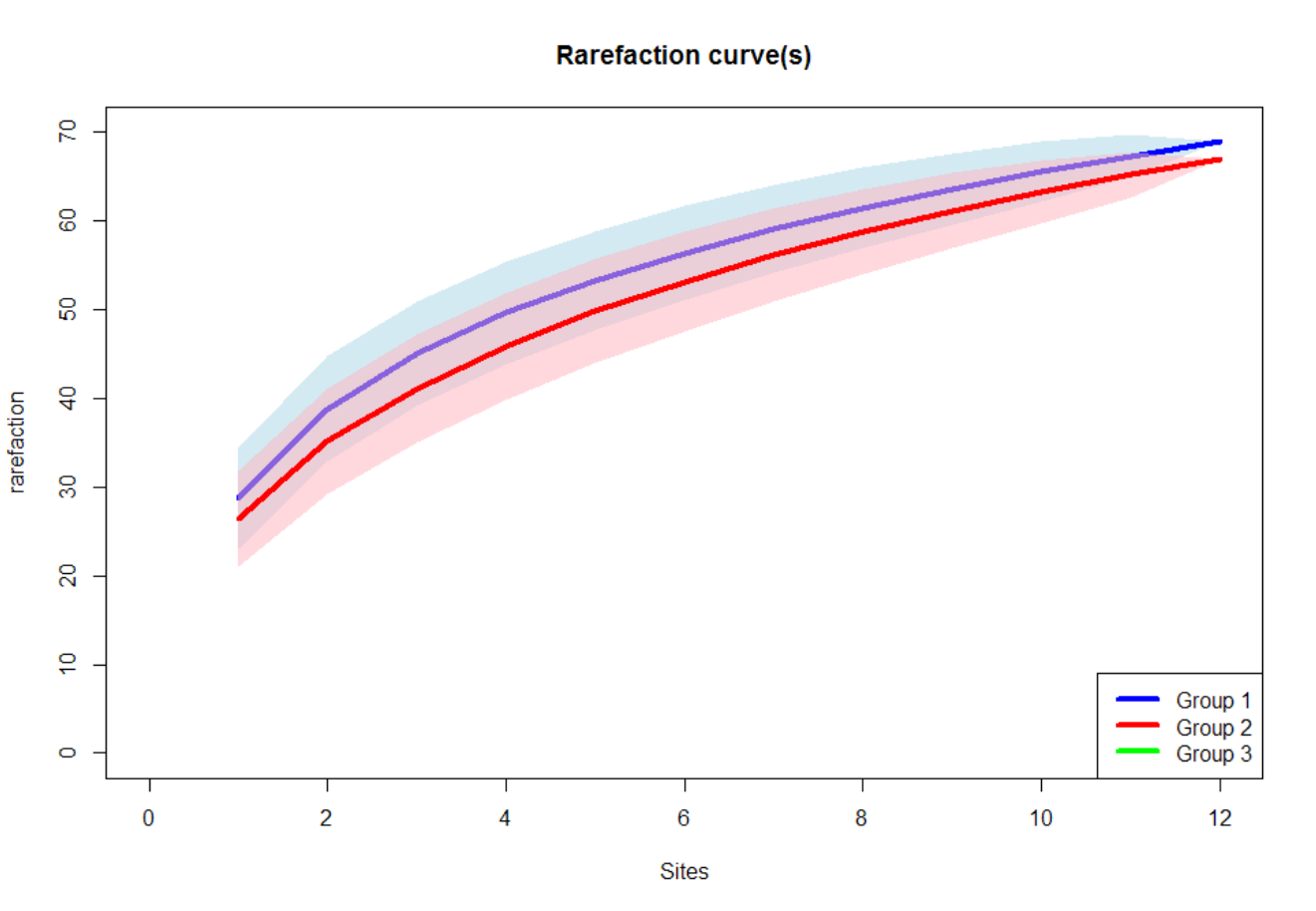
Rarefaction curve

https://en.wikipedia.org/wiki/Rarefaction_(ecology)

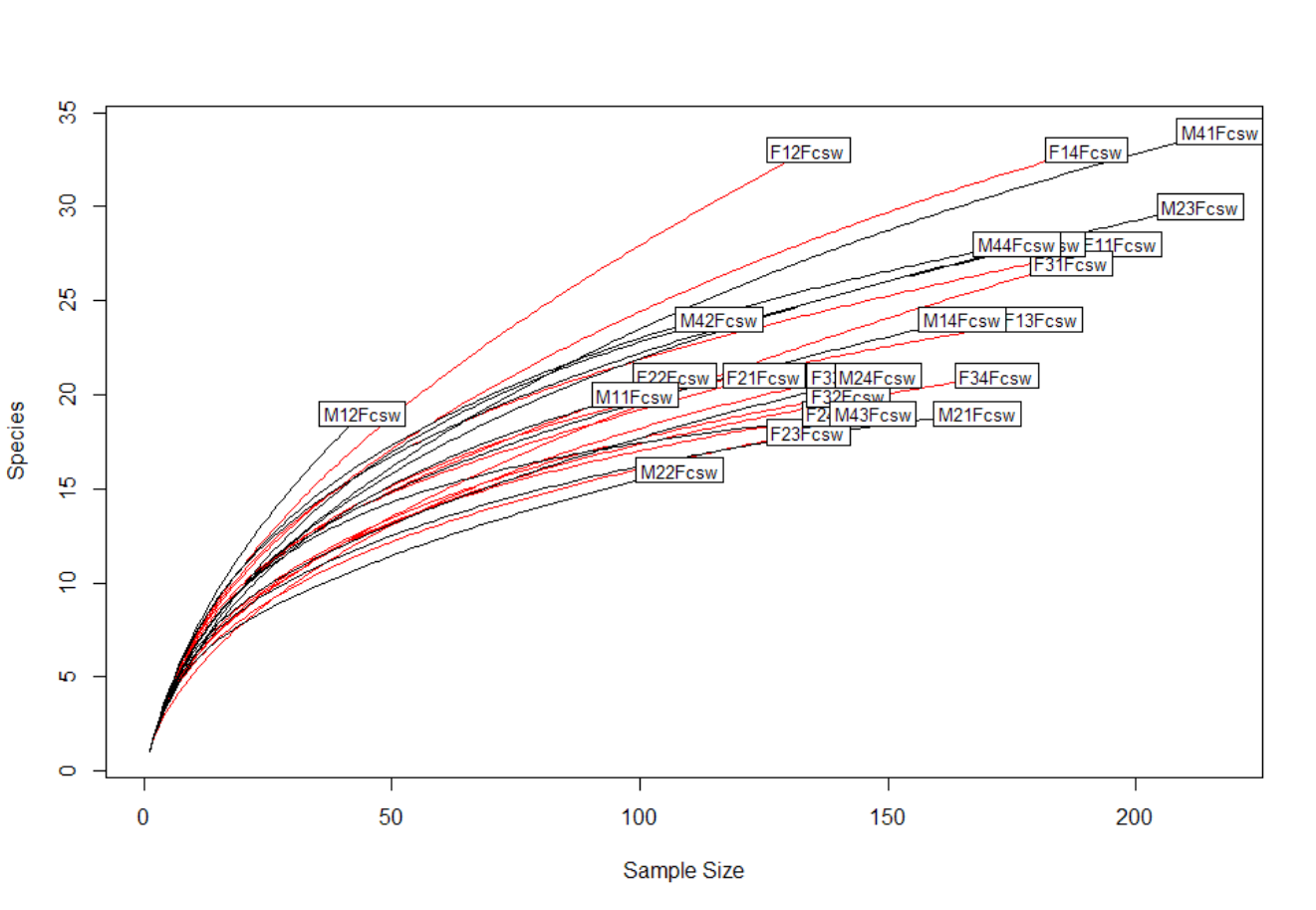
Rarefaction curve

- Used for estimating species richness
- Can be plotted as a function of the number of samples taken or the number of species
- Chipster plots the curves as a function of samples (see the next slide)
- Can also be used for visualizing the cumulative species richness per sample

Rarefaction curve



Rarefaction



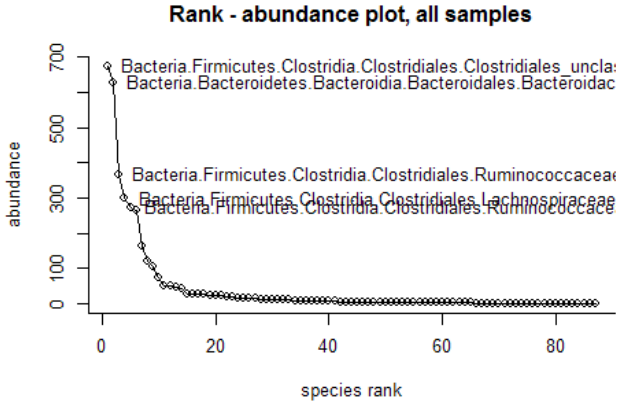
Rank abundance curve

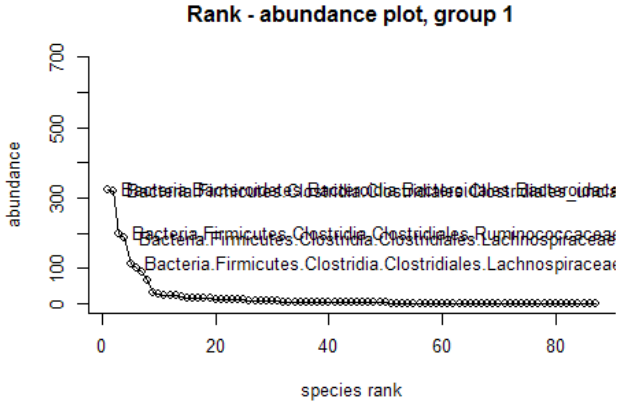
https://en.wikipedia.org/wiki/Rank_abundance_curve

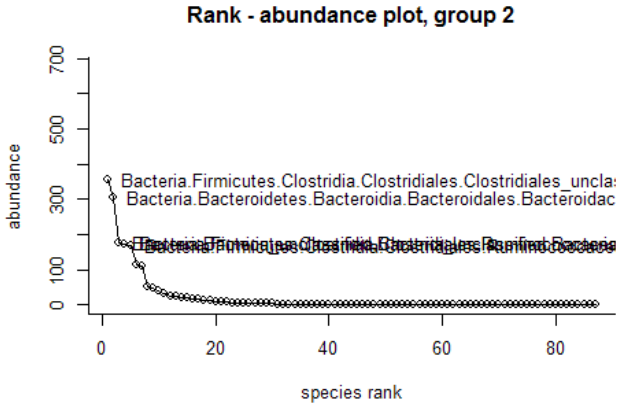
Rank abundance curves

- Displays relative species richness
- Relative abundance on the y-axis, and abundance rank on the x-axis
- Species evenness is depicted by the shape of the curve

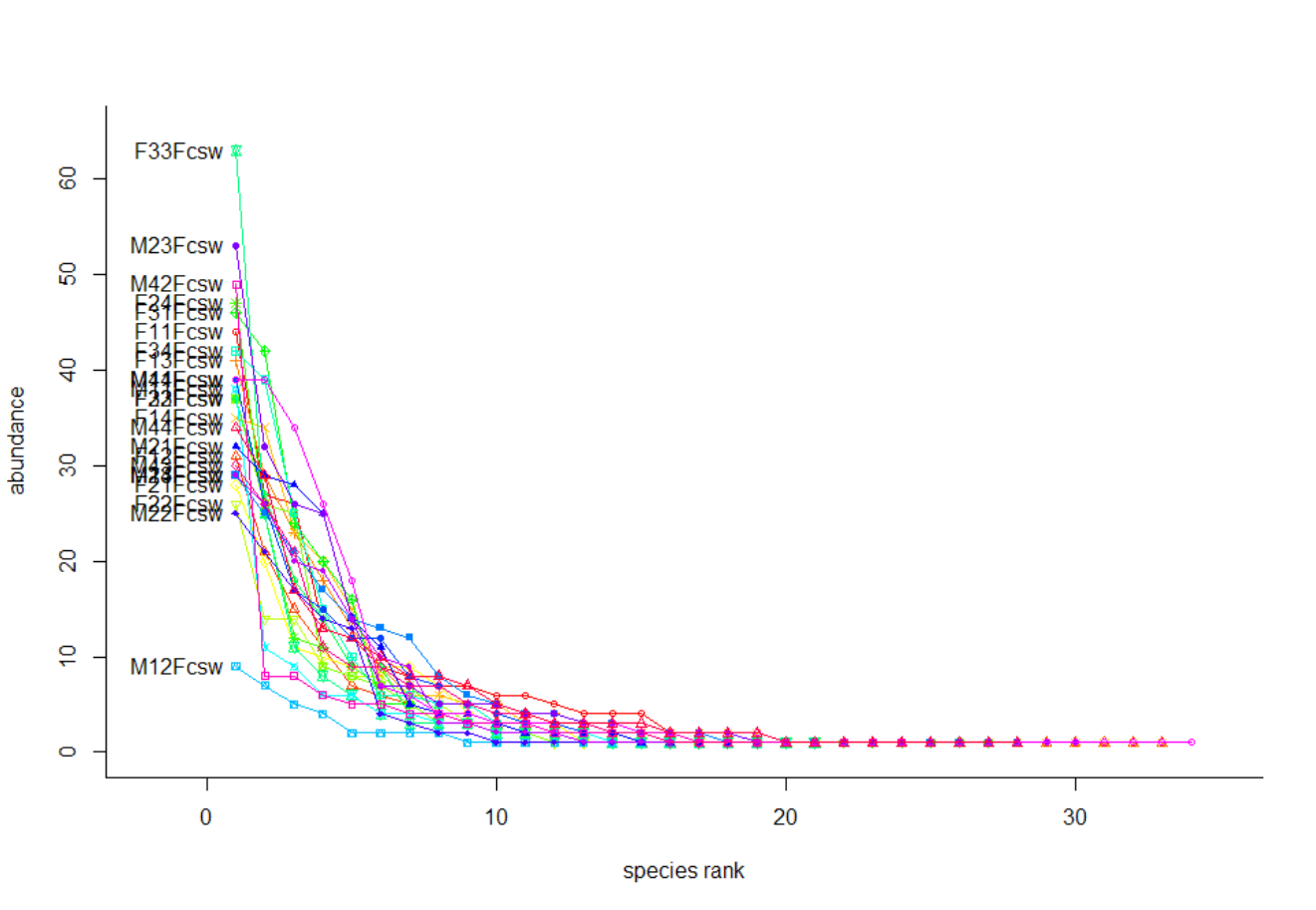
Rank abundance curves





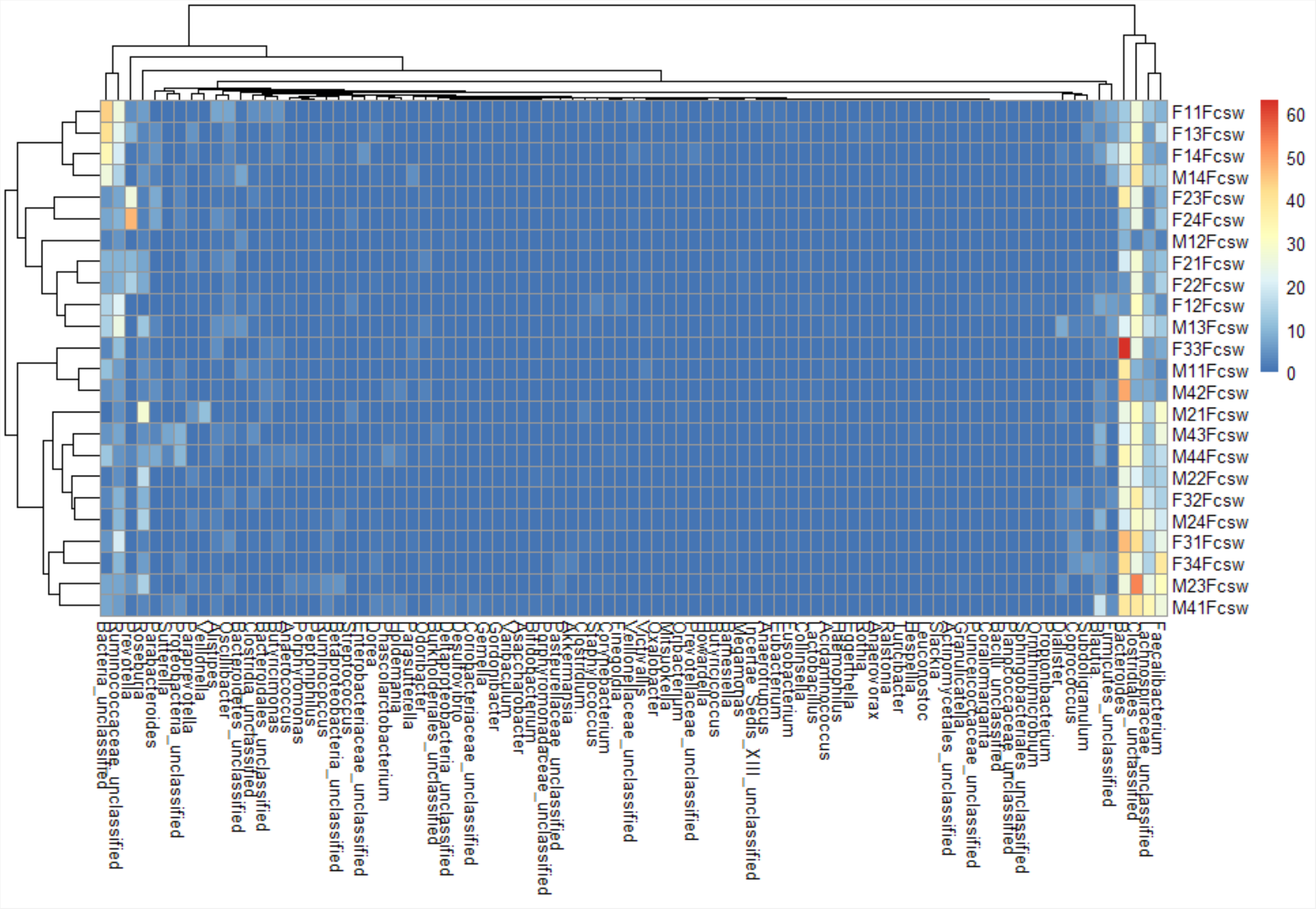


Rank abundance curves

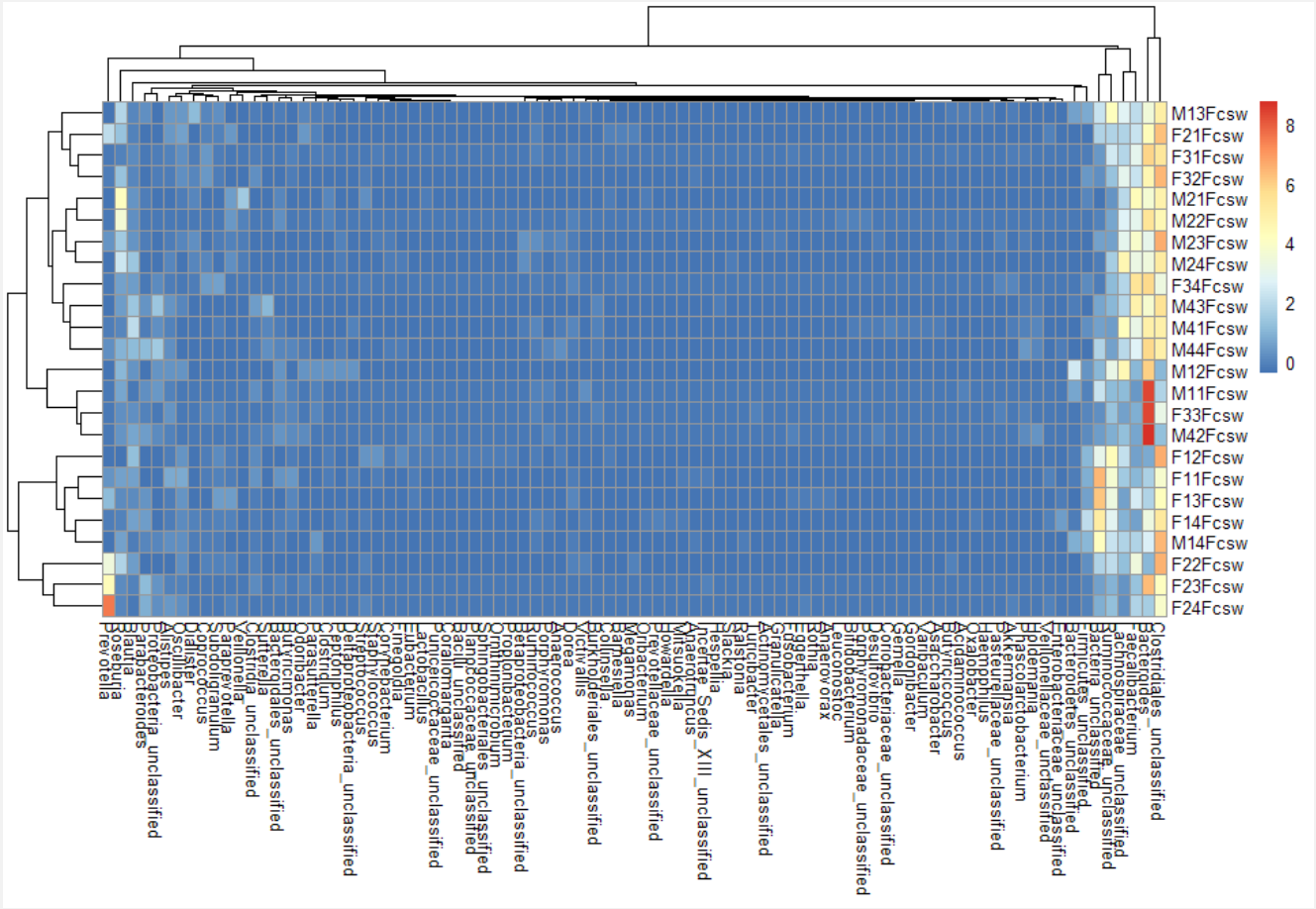


Heatmap

Heatmap


F11Fcsw
F13Fcsw
F14Fcsw
M14Fcsw
F23Fcsw
F24Fcsw
M12Fcsw
F21Fcsw
F22Fcsw
F12Fcsw
M13Fcsw
F33Fcsw
M11Fcsw
M42Fcsw
M21Fcsw
M43Fcsw
M44Fcsw
M22Fcsw
F32Fcsw
M24Fcsw
F31Fcsw
F34Fcsw
M23Fcsw
M41Fcsw
60
50
40
30
20
10
0

Heatmap (data scaled first)



Ordination analysis

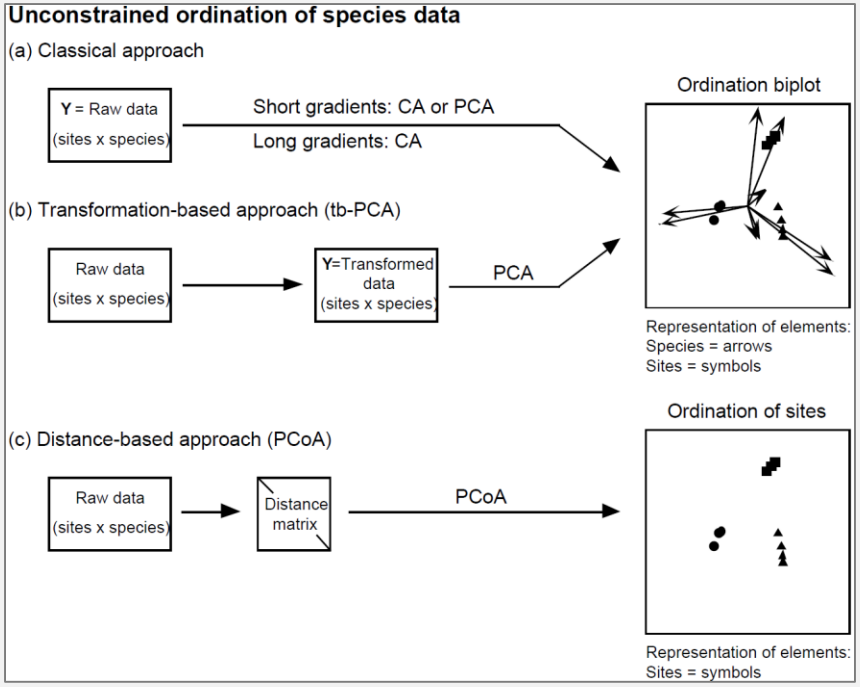
Ordination analysis

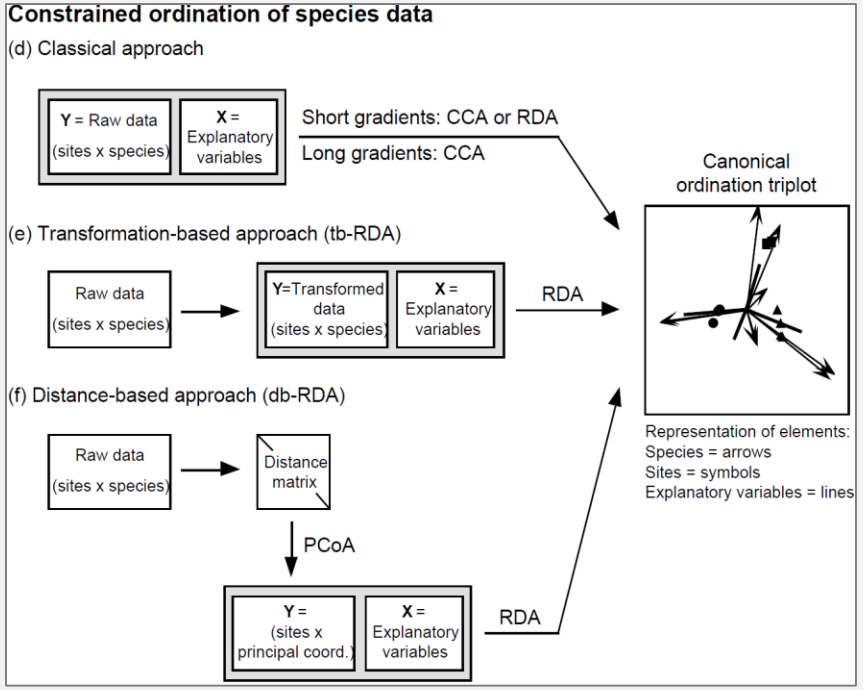
- In regression we had one response variable, and several explanatory variables.
- What if we have several response variables?
  - Multivariate analysis of variance (MANOVA)?
  - Analysis of varience using distance matrices (ADONIS)?
  - Ordination?

Ordination analysis

- Allows displaying a dataset with several dimensions with a smaller number of dimensions
- Takes a species count table (rows=taxa, columns=samples, cells=frequency)
- Additionally takes a comparable matrix of environmental measurements (phenodata)
- Creates an image, and allows testing for the significance of the environmental factors to the species occurrance or frequency

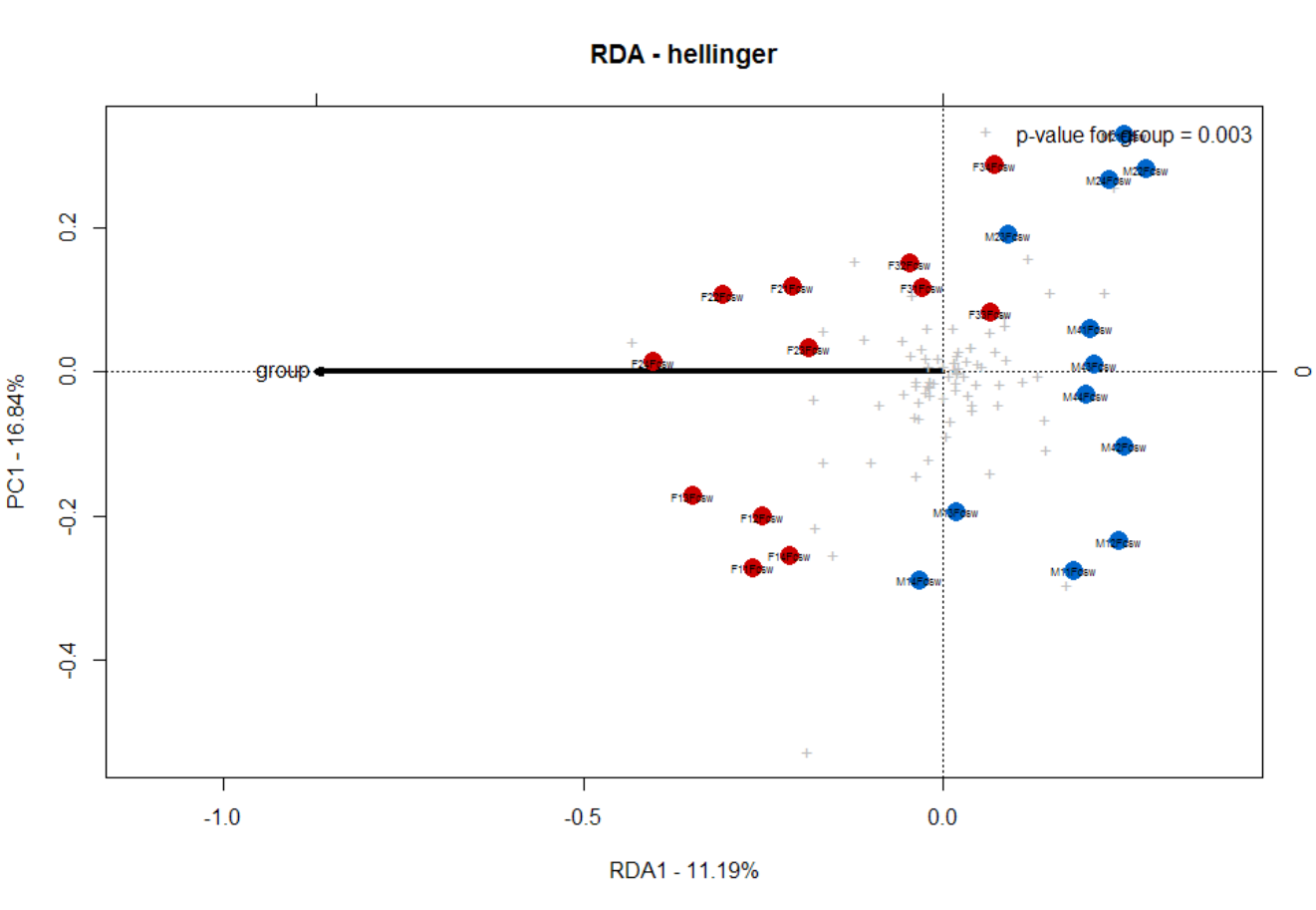
Ordination approaches



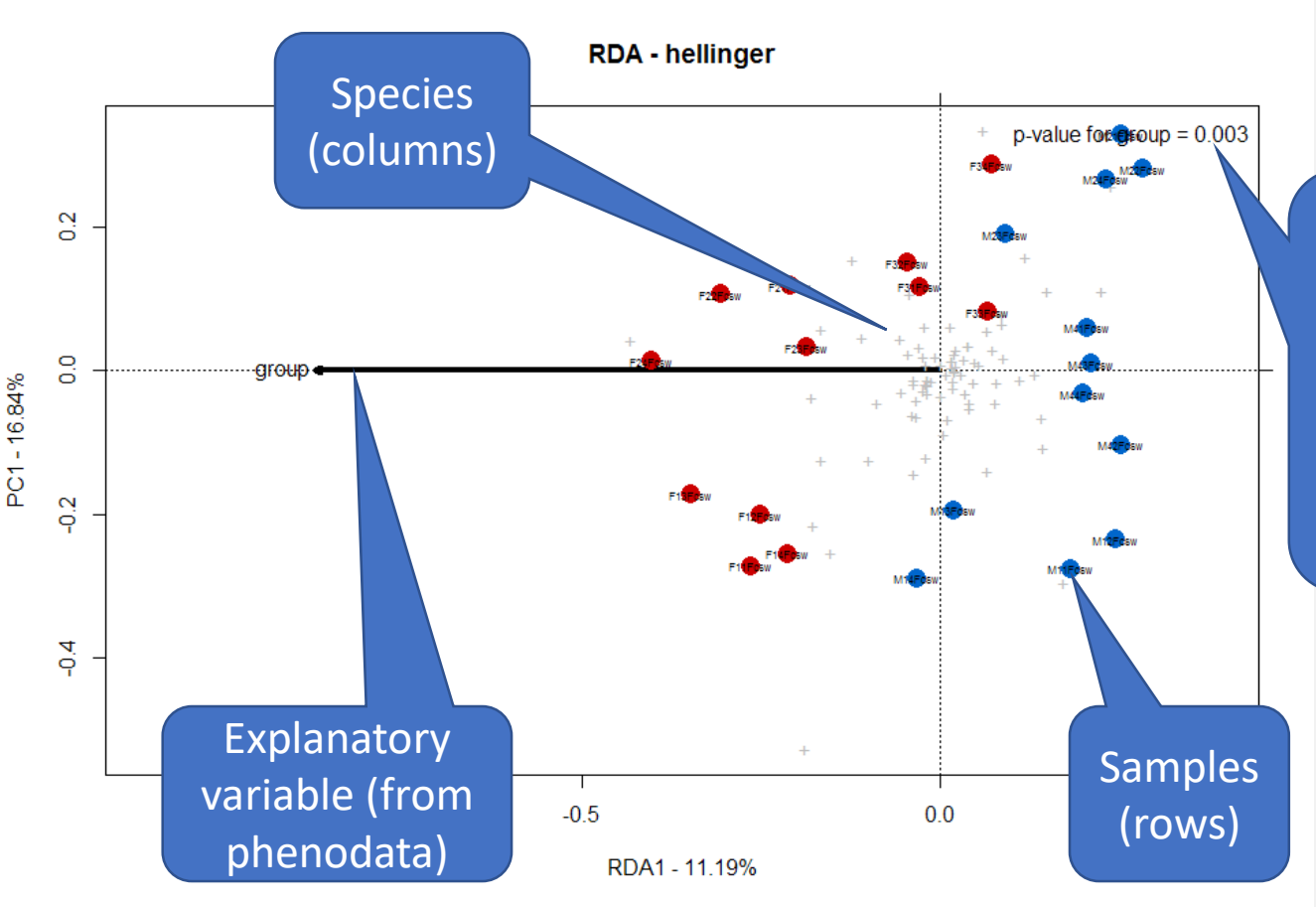


By Pierre Legendre

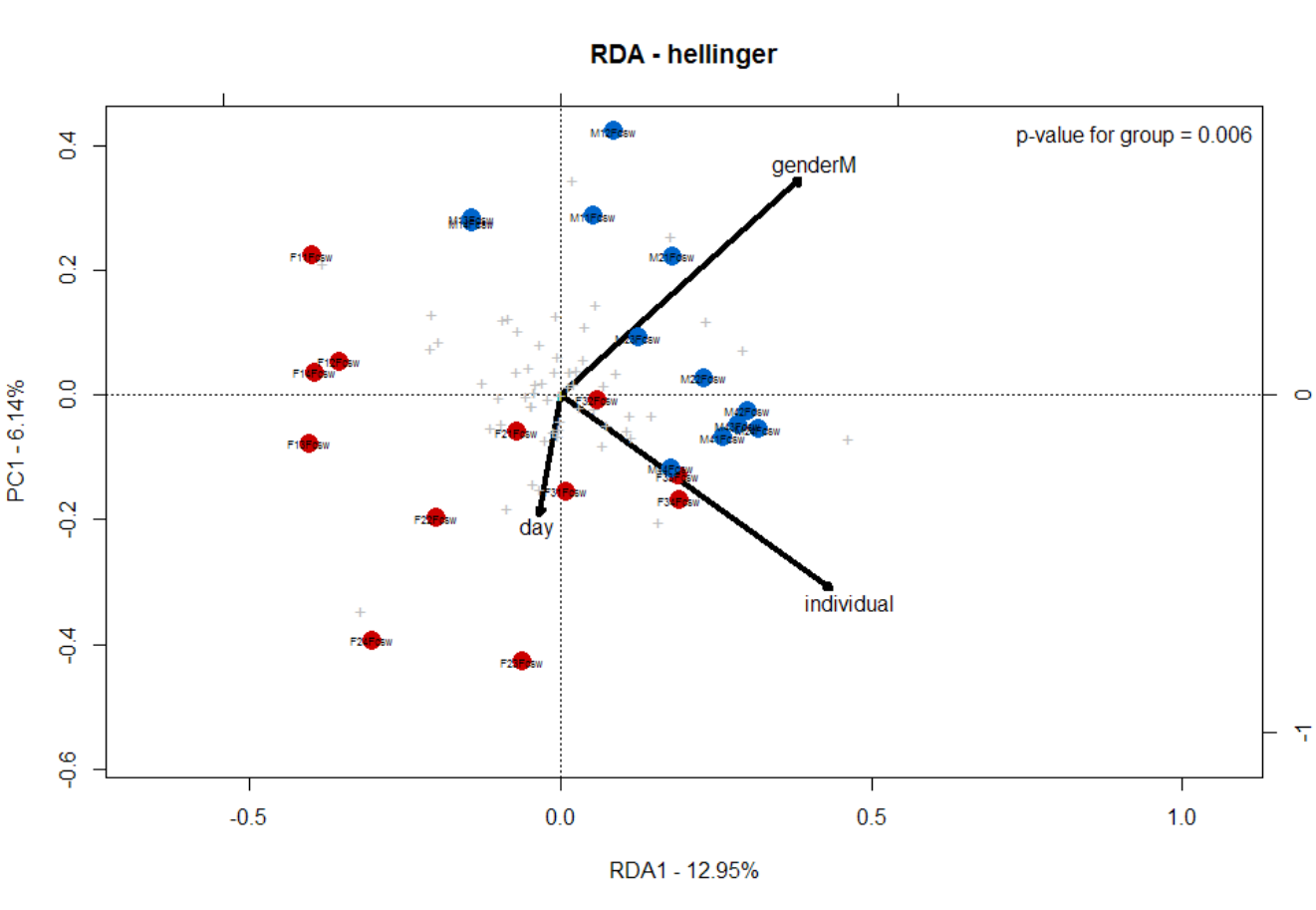
RDA


RDA - hellinger
p-value for group = 0.003
group
PC1 - 16.84%
RDA1 - 11.19%
0.2
0.0
-0.2
-0.4
-1.0
-0.5
0.0
0

RDA


RDA - hellinger
p-value for group = 0.003
group
PC1 - 16.84%
RDA1 - 11.19%
0.2
0.0
-0.2
-0.4
-0.5
0.0
Species (columns)
P-value for explanatory variable (resampling based)
Explanatory variable (from phenodata)
Samples (rows)

RDA with several explanatory variables



Statistical analyses

Statistical analyses - diversity

- Contributed diversity
  - alpha
    - diversity inside an area or ecosystem (species richness)
  - beta
    - diversity between ecosystems
  - gamma
    - overall diversity of all ecosystems in a particular area
- Diversity can be measured with different indexes, such as Shannon entropy or just the count of species (but the species count is dependent on the sampling depth, which can be checked using the rarefaction curves)

Statistical analyses – comparing groups

- Do the groups differ in species composition?
  - Permutational Multivariate Analysis of Variance Using Distance Matrices
  - Multivariate homogeneity of groups dispersions (variances)
  - Analysis of Molecular Variance
    - Based on a (euclidean) distance matrix between sequences
    - Distances (or their variance, to be more exact) are partitioned according to a grouping variable into a within group and between groups variance (this is similar to standard one-way ANOVA)

Indicator species approach

- What are the taxa that differentiate between the group in a best possible way?

Variance analyses

Analysis of Molecular Variance

- Excoffier, L., Smouse, P. E. and Quattro, J. M. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.
- A distance matrix is partitioned similarly to a usual ANOVA partitioning of the sum of squares
- Spits out a p-value for the explanatory variables

Analysis of Molecular Variance

```
        Analysis of Molecular Variance

Call: amova(formula = dd ~ group, nperm = 9999)

             SSD        MSD df
group  1632.0833 1632.08333  1
Error 20133.0000  915.13636 22
Total 21765.0833  946.30797 23

Variance components:
        sigma2 P.value
group  59.7456  0.0998
Error 915.1364

Variance coefficients:
 a
12
```

Permutational Multivariate Analysis of Variance Using Distance Matrices

- Anderson, M.J. 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, **26**: 32–46.
- In a sense offers an alternative to AMOVA especially for crossed and nested factors
- Spits out a p-value for the explanatory variables

Permutational Multivariate Analysis of Variance Using Distance Matrices

```
Call:
adonis(formula = dd ~ group, permutations = 9999)

Permutation: free
Number of permutations: 9999

Terms added sequentially (first to last)

          Df SumsOfSqs  MeanSqs F.Model       R2 Pr(>F)
group      1   1632.08 1632.083 1.78343 0.074986 0.0992 .
Residuals 22  20133.00  915.136         0.925014
Total     23  21765.08                  1.000000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Multivariate homogeneity of groups dispersions (variances)

- Anderson, M.J. (2006) Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* **62**, 245–253.
- The betadisper is a multivariate analogue of Levene's test for homogeneity of variances.
- Non-euclidean distances between objects and group centroids are handled by reducing the original distances to principal coordinates.
- This procedure has latterly been used as a means of assessing beta diversity.

Multivariate homogeneity of groups dispersions (variances)

```
Analysis of Variance Table

Response: Distances
           Df   Sum Sq Mean Sq F value  Pr(>F)
Groups      1  103.709 103.709 1.28169 0.26977
Residuals 22 1780.152  80.916
```

Contribution diversity approach

Contribution diversity approach

- Additive diversity partitioning deals with the relation of mean alpha and the total (gamma) diversity.
- Although alpha diversity values often vary considerably. Thus, contributions of the sites to the total diversity are uneven.
- This site specific contribution is measured by contribution diversity components. A unit that has e.g. many unique species will contribute more to the higher level (gamma) diversity than another unit with the same number of species, but all of which common.
- Lu, H. P., Wagner, H. H. and Chen, X. Y. 2007. A contribution diversity approach to evaluate species diversity. *Basic and Applied Ecology*, 8, 1–12.
- From the R/vegan help page

Contrib. diversity approach

	alpha	beta	gamma
Bacteroides	0.275862069	0.783346093	1.059208162
Clostridiales_unclassified	0.275862069	0.783346093	1.059208162
Lachnospiraceae_unclassified	0.275862069	0.783346093	1.059208162
Faecalibacterium	0.275862069	0.783346093	1.059208162
Ruminococcaceae_unclassified	0.275862069	0.783346093	1.059208162
Bacteria_unclassified	0.264367816	0.742208767	1.006576583
Roseburia	0.264367816	0.764537316	1.028905132
Blautia	0.252873563	0.703703020	0.956576583
Alistipes	0.241379310	0.655078225	0.896457536
Firmicutes_unclassified	0.206896552	0.572894317	0.779790869

- Contribution of every taxon to alpha, beta and gamma diversity

- alpha
  - diversity inside an area or ecosystem (species richness)
- beta
  - diversity between ecosystems
- gamma
  - overall diversity of all ecosystems in a particular area

Indicator species approach

Dufrene-Legendre Indicator Species Analysis

- Dufrene, M. and Legendre, P. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecol. Monogr. 67(3):345-366.
- Calculates the indicator value (fidelity and relative abundance) of species in clusters or types.
- Gives p-values for taxa separating the specified groups

Dufrene-Legendre Indicator Species Analysis

	cluster	indicator_value	probability
Roseburia	1	0.6933	0.02720272
Lachnospiraceae_unclassified	1	0.6221	0.04910491
Subdoligranulum	2	0.6548	0.01540154
Ruminococcaceae_unclassified	2	0.6350	0.03330333
Prevotella	2	0.6171	0.01610161
Coprococcus	2	0.6000	0.01700170
Oscillibacter	2	0.5903	0.04190419

Sum of probabilities = 53.1752175217522
Sum of Indicator Values = 21.58
Sum of Significant Indicator Values = 4.41
Number of Significant Indicators = 7
Significant Indicator Distribution

1 2
2 5

Indicator Species Analysis Minimizing Intermediate Occurrences

- Aho, K., D.W. Roberts, and T.W.Weaver. 2008. Using geometric and non-geometric internal evaluators to compare eight vegetation classification methods. J. Veg. Sci. In press.
- Calculates the constancy (fractional occurrence of each species in every type), and then calculates twice the the sum of the absolute values of the constancy - 0.5, normalized to the number of clusters (columns).

Indicator Species Analysis Minimizing Intermediate Occurrences

	isa
Bacteroides	1.00
Clostridiales_unclassified	1.00
Lachnospiraceae_unclassified	1.00
Faecalibacterium	1.00
Ruminococcaceae_unclassified	1.00
Bacteria_unclassified	0.92
Roseburia	0.92
Varibaculum	0.92
Actinomycetales_unclassified	0.92
Ornithinimicrobium	0.92