DTU

DTU Health Technology
Bioinformatics

Ordination methods

Gisle Vestergaard
Associate Professor
Section of Bioinformatics
Technical University of Denmark
gisves@dtu.dk

Menu

- What is ordination?
- What are we using ordination for?
- Principal Coordinates Analysis (PCoA) & Principal Components Analysis (PCA)

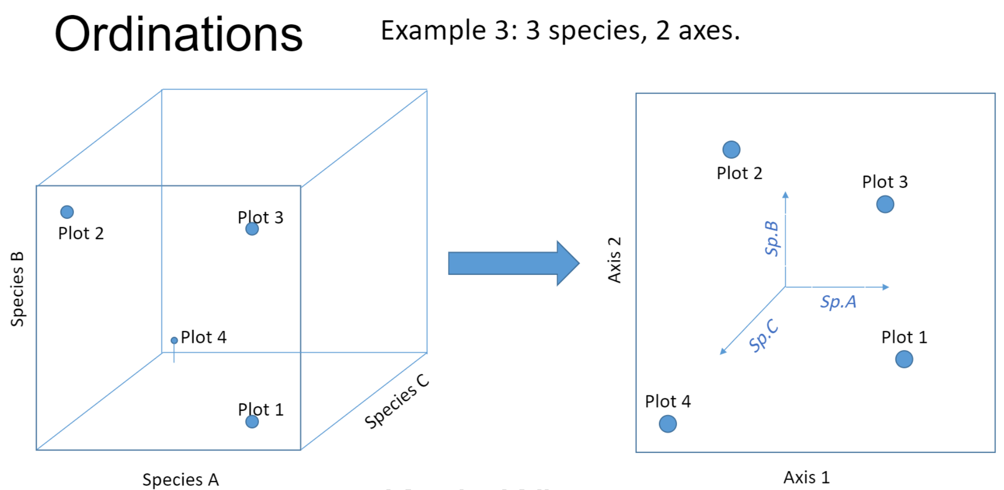
Many types of ordination

- Becoming part of the clergy
- *Nonmetric Multidimensional Scaling (NMDS)*
- Correspondence Analysis (CA)
- Many more
- We will focus on Principal Coordinates Analysis (PCoA)

Ordination in ecology

- A statistical technique in which data from a large number of sites or populations are represented as points in a multidimensional space.

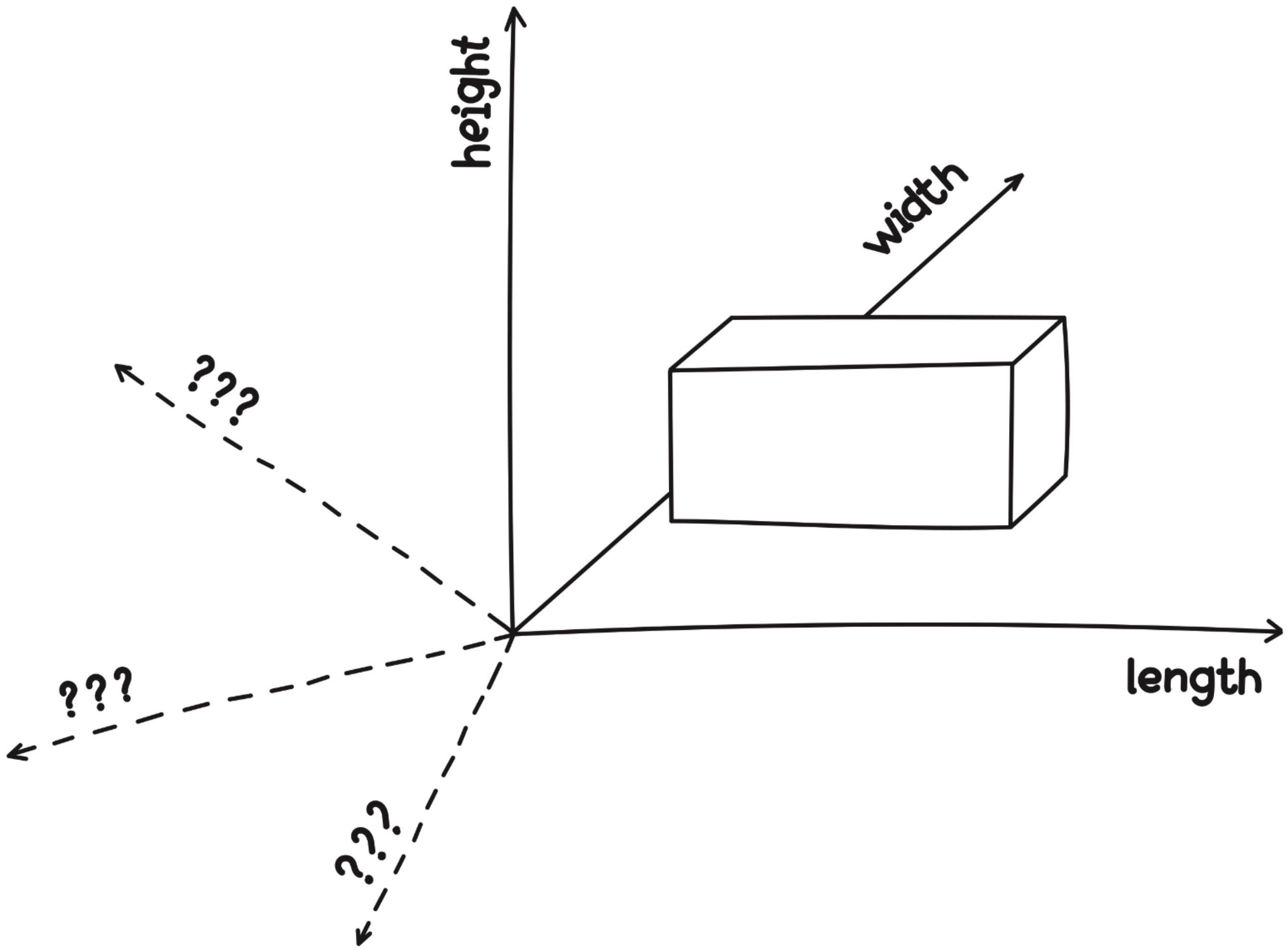
- 2 OTUs/ASVs/Genes = 2 axes
- 3 OTUs/ASVs/Genes = 3 axes
- n OTUs/ASVs/Genes = n axes
- Squish into 2 or 3 dimensions



Kevin Wilcox

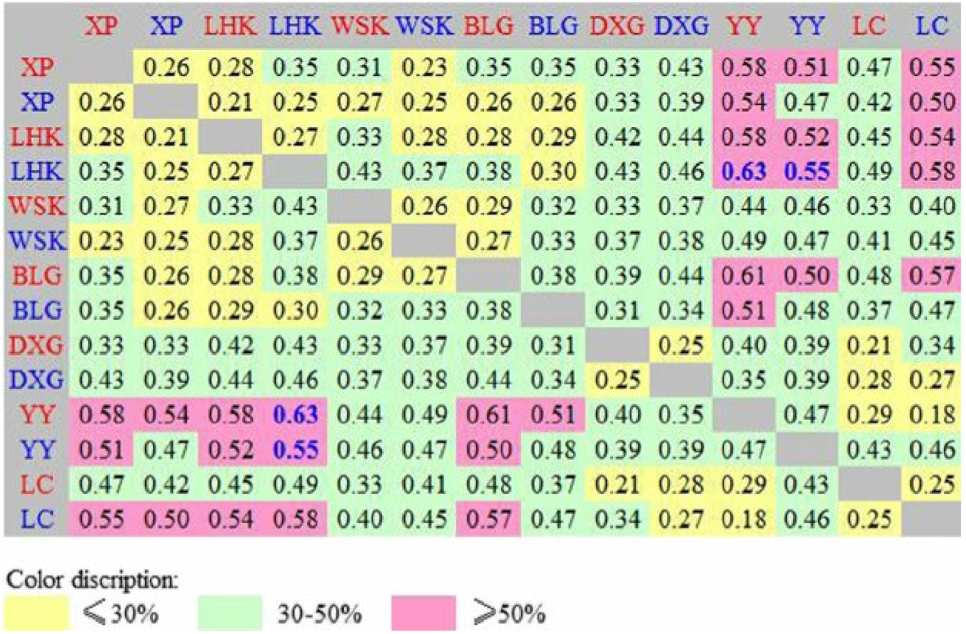
- Mainly used for exploratory analysis and NOT hypothesis testing

N-space



Distance matrix

- Pairwise distances



	XP	XP	LHK	LHK	WSK	WSK	BLG	BLG	DXG	DXG	YY	YY	LC	LC
XP		0.26	0.28	0.35	0.31	0.23	0.35	0.35	0.33	0.43	0.58	0.51	0.47	0.55
XP	0.26		0.21	0.25	0.27	0.25	0.26	0.26	0.33	0.39	0.54	0.47	0.42	0.50
LHK	0.28	0.21		0.27	0.33	0.28	0.28	0.29	0.42	0.44	0.58	0.52	0.45	0.54
LHK	0.35	0.25	0.27		0.43	0.37	0.38	0.30	0.43	0.46	**0.63**	**0.55**	0.49	0.58
WSK	0.31	0.27	0.33	0.43		0.26	0.29	0.32	0.33	0.37	0.44	0.46	0.33	0.40
WSK	0.23	0.25	0.28	0.37	0.26		0.27	0.33	0.37	0.38	0.49	0.47	0.41	0.45
BLG	0.35	0.26	0.28	0.38	0.29	0.27		0.38	0.39	0.44	0.61	0.50	0.48	0.57
BLG	0.35	0.26	0.29	0.30	0.32	0.33	0.38		0.31	0.34	0.51	0.48	0.37	0.47
DXG	0.33	0.33	0.42	0.43	0.33	0.37	0.39	0.31		0.25	0.40	0.39	0.21	0.34
DXG	0.43	0.39	0.44	0.46	0.37	0.38	0.44	0.34	0.25		0.35	0.39	0.28	0.27
YY	0.58	0.54	0.58	**0.63**	0.44	0.49	0.61	0.51	0.40	0.35		0.47	0.29	0.18
YY	0.51	0.47	0.52	**0.55**	0.46	0.47	0.50	0.48	0.39	0.39	0.47		0.43	0.46
LC	0.47	0.42	0.45	0.49	0.33	0.41	0.48	0.37	0.21	0.28	0.29	0.43		0.25
LC	0.55	0.50	0.54	0.58	0.40	0.45	0.57	0.47	0.34	0.27	0.18	0.46	0.25	

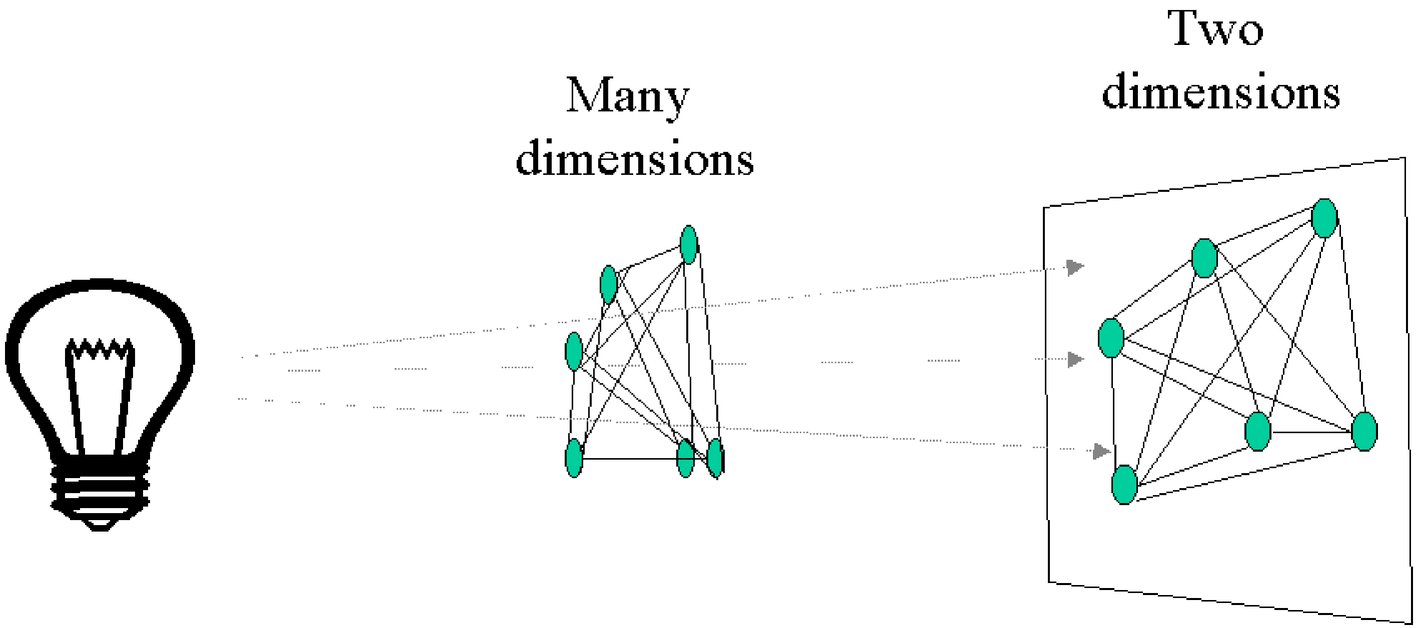
Color discription:

🟨 ⩽30% 🟩 30-50% 🟥 ⩾50%

Principal Coordinates Analysis - PCoA

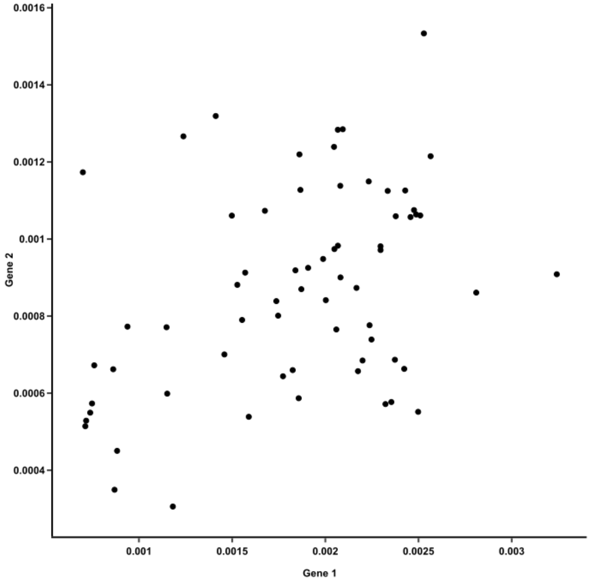
- Principal Coordinates Analysis (PCoA) incorporates the distance between all samples
- Displays a image (often 2D) with matrix rotated so first axis shows maximal variance
  - The maximum variation uncorrelated with axis 1 is projected on the second axis

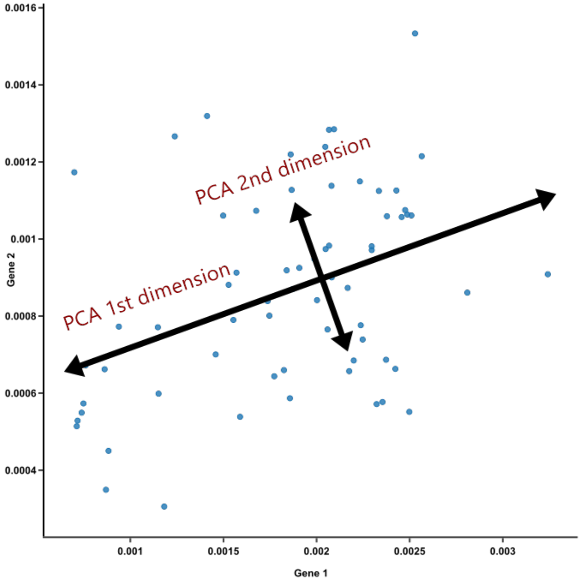
- Principal Components Analysis (PCA) is a PCoA here the distance metric is Euclidean



Principal Components Analysis - PCA

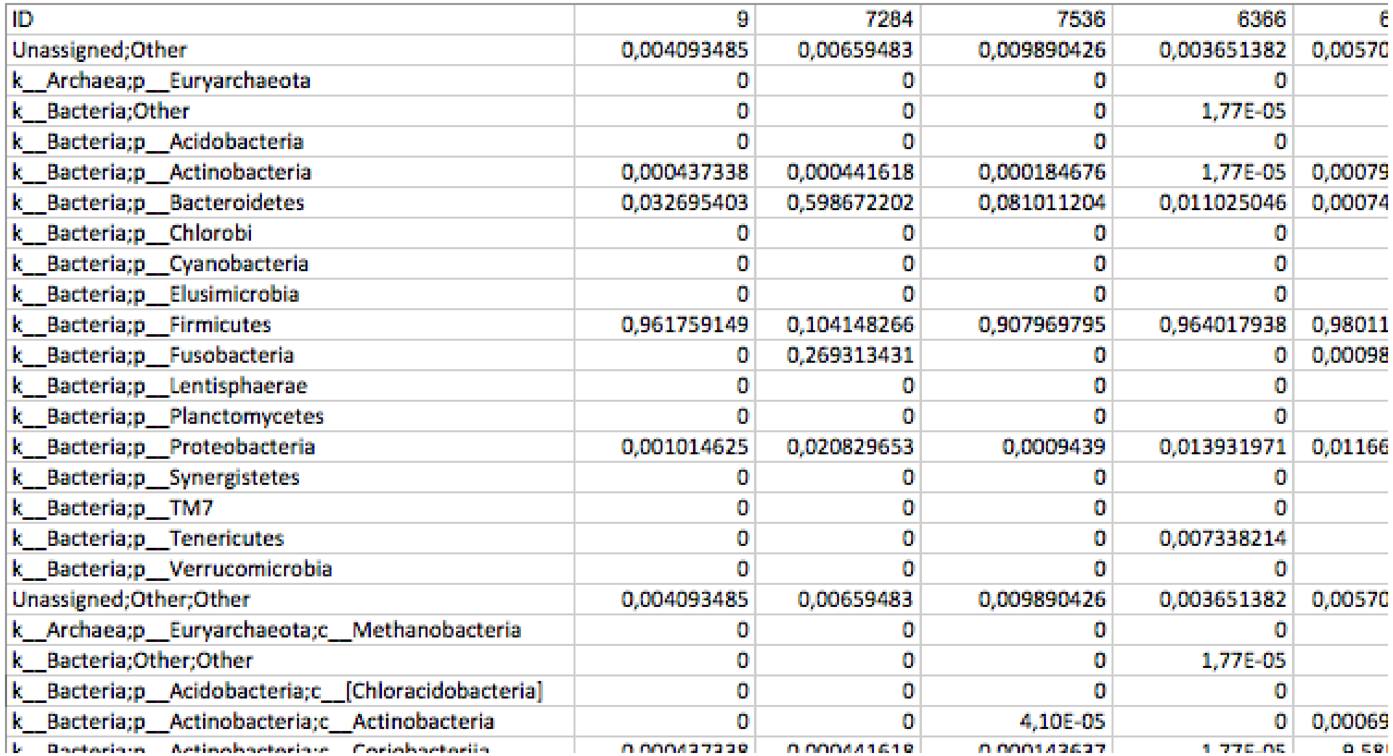
- Principal Components Analysis (PCA) is a rotation of the original data matrix
- Maximum variance is projected or "extracted" along the first axis
  - The maximum variation uncorrelated with axis 1 is projected on the second axis
  - The maximum variation uncorrelated with the first and second axis is projected on the third axis
  - Etc.





All those zeroes…

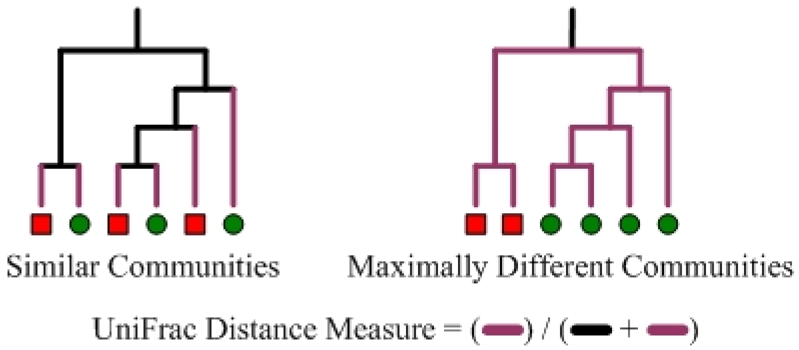
- Excess of zero skews results



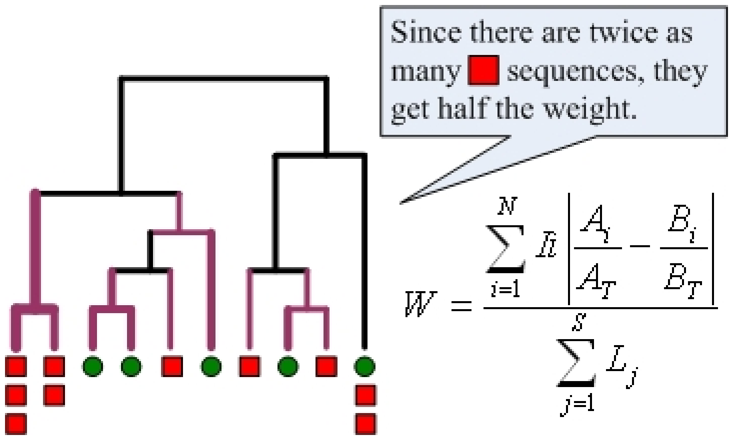
ID	9	7284	7536	6366	6
Unassigned;Other	0,004093485	0,00659483	0,009890426	0,003651382	0,00570
k__Archaea;p__Euryarchaeota	0	0	0	0	
k__Bacteria;Other	0	0	0	1,77E-05	
k__Bacteria;p__Acidobacteria	0	0	0	0	
k__Bacteria;p__Actinobacteria	0,000437338	0,000441618	0,000184676	1,77E-05	0,00079
k__Bacteria;p__Bacteroidetes	0,032695403	0,598672202	0,081011204	0,011025046	0,00074
k__Bacteria;p__Chlorobi	0	0	0	0	
k__Bacteria;p__Cyanobacteria	0	0	0	0	
k__Bacteria;p__Elusimicrobia	0	0	0	0	
k__Bacteria;p__Firmicutes	0,961759149	0,104148266	0,907969795	0,964017938	0,98011
k__Bacteria;p__Fusobacteria	0	0,269313431	0	0	0,00098
k__Bacteria;p__Lentisphaerae	0	0	0	0	
k__Bacteria;p__Planctomycetes	0	0	0	0	
k__Bacteria;p__Proteobacteria	0,001014625	0,020829653	0,0009439	0,013931971	0,01166
k__Bacteria;p__Synergistetes	0	0	0	0	
k__Bacteria;p__TM7	0	0	0	0	
k__Bacteria;p__Tenericutes	0	0	0	0,007338214	
k__Bacteria;p__Verrucomicrobia	0	0	0	0	
Unassigned;Other;Other	0,004093485	0,00659483	0,009890426	0,003651382	0,00570
k__Archaea;p__Euryarchaeota;c__Methanobacteria	0	0	0	0	
k__Bacteria;Other;Other	0	0	0	1,77E-05	
k__Bacteria;p__Acidobacteria;c__[Chloracidobacteria]	0	0	0	0	
k__Bacteria;p__Actinobacteria;c__Actinobacteria	0	0	4,10E-05	0	0,00069
k__Bacteria;p__Actinobacteria;c__Coriobacteriia	0,000437338	0,000441618	0,000143637	1,77E-05	9,58

Unifrac

- Distance metric
  - The distance between the two samples is calculated as (the sum of "unshared" branch lengths)/(the sum of all tree branch lengths)
  - i.e. the fraction of total branch length which is unshared.
- Weighted Unifrac incorporates abundance
- Has good separation power (Thorsen et al., 2016)



Unweighted



Since there are twice as many 🟥 sequences, they get half the weight.

$$W = \frac{\sum_{i=1}^{N} l_i \left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right|}{\sum_{j=1}^{S} L_j}$$

Example