Statistical analysis of metagenomes

Liz Dinsdale

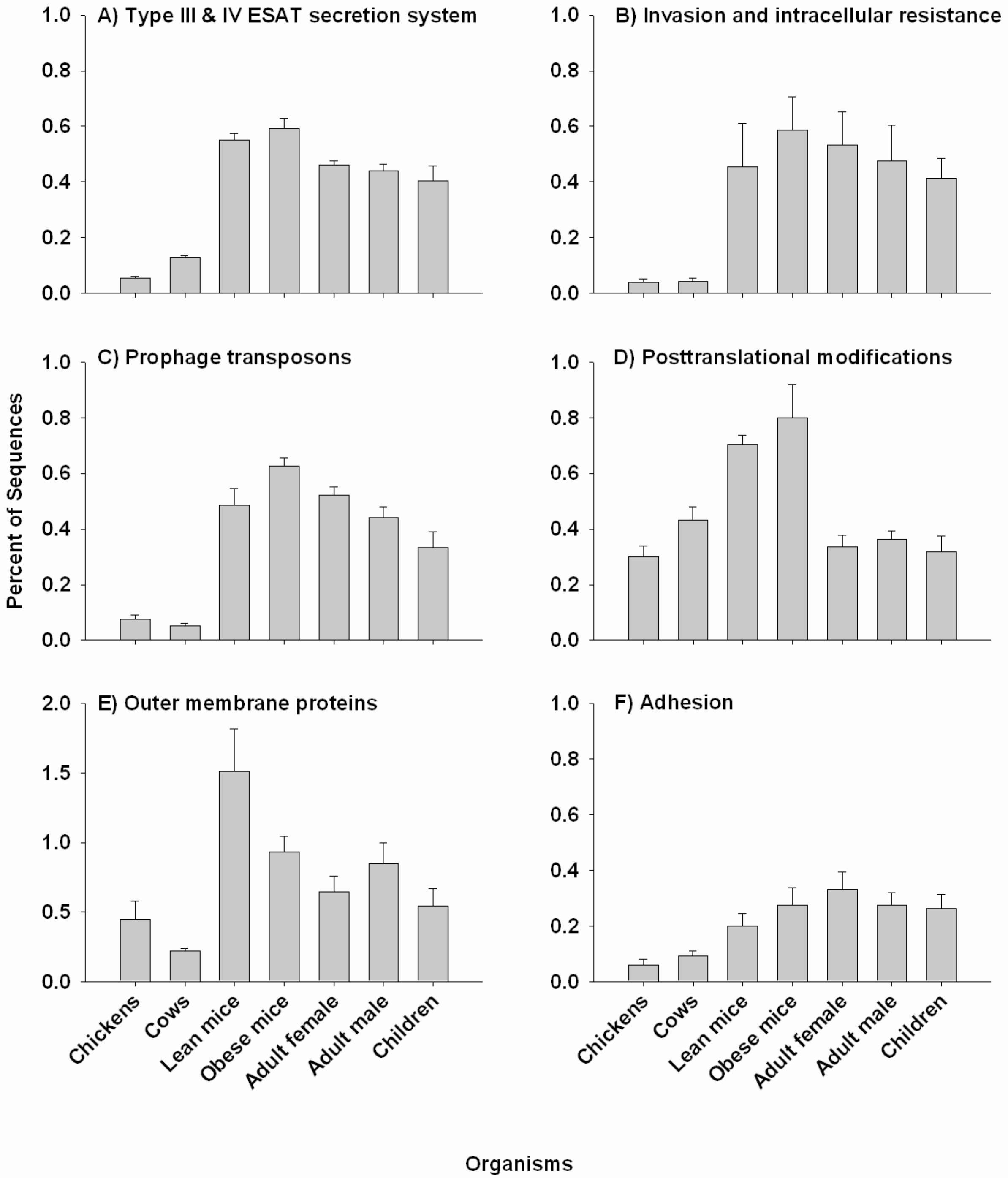
Visualizing the data

- Graph –
  - all data – investigate any points that are outliers (they may be incorrectly entered into the computer)
  - Mean and standard errors (se=standard deviation/square root (number of replicates))
  - Variables by self or in combinations (see if something jumps out at you and is worth investigation)

- Descriptions – grouped data

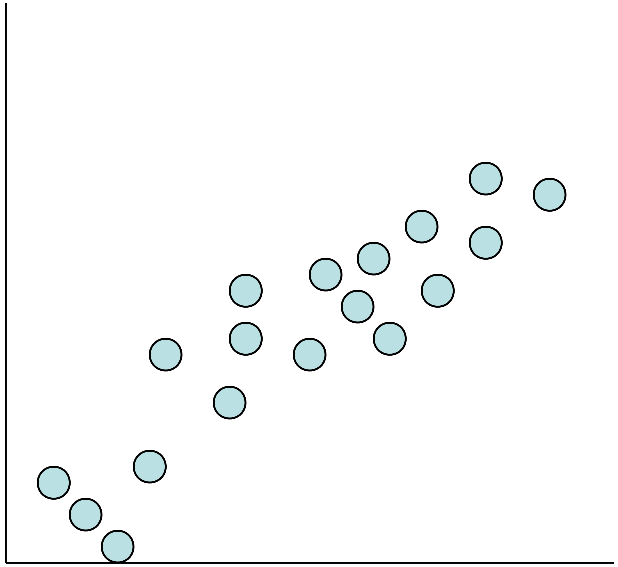
- Statistics – raw data

Mean – useful for describing difference between groups

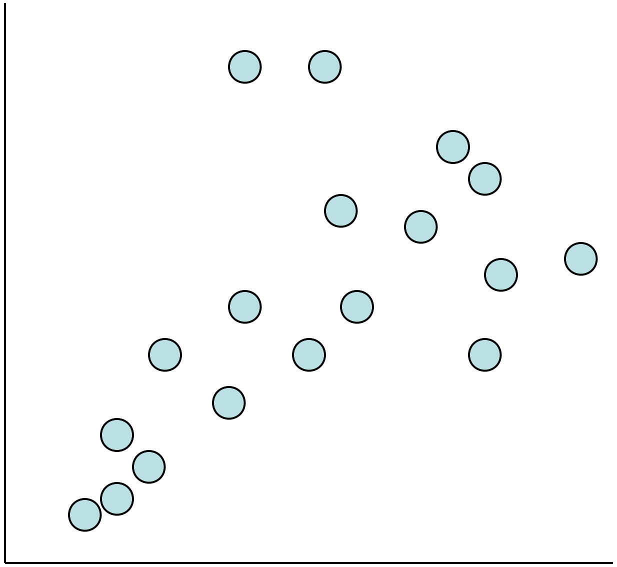


Assumptions of the data

- Normality – ie: typical bell shaped curve
- Homogeneity of variance



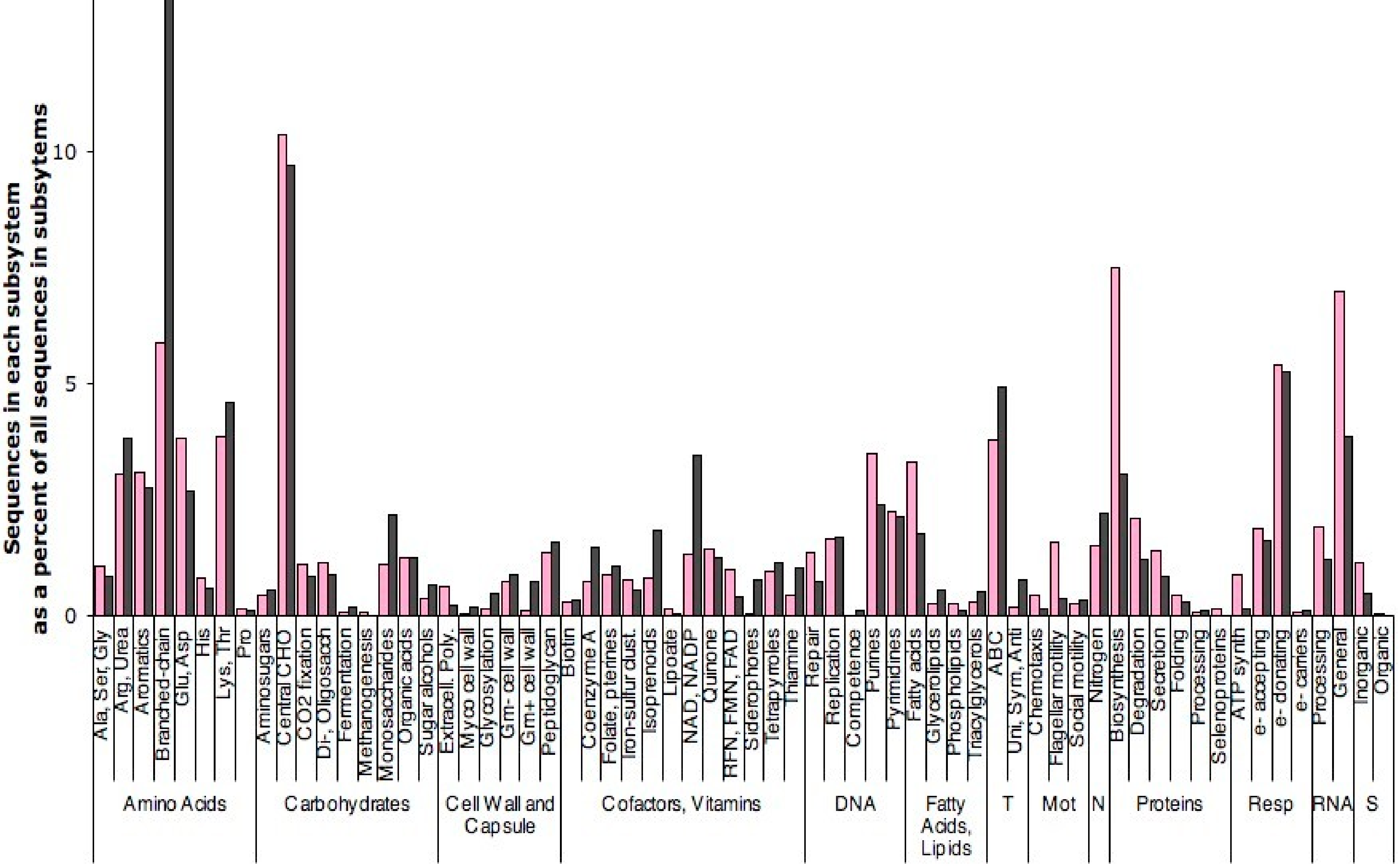
Homogeneous



Heterogeneous

- Are variables correlated

Subsystems in a mine



Pairwise comparisons

- Great for sample a versus sample b
- Need to worry about chance and probability.
- Simple tests, like t-test, g-test (assume data normal)
- Non-parametric tests don't assume normality

XIPE for two samples

- Sample 10,000 proteins from site 1
- Count frequency of each subsystem
- Repeat 20,000 times

- Repeat for sample 2

- Combine both samples
- Sample 10,000 proteins 20,000 times
- Build 95% CI

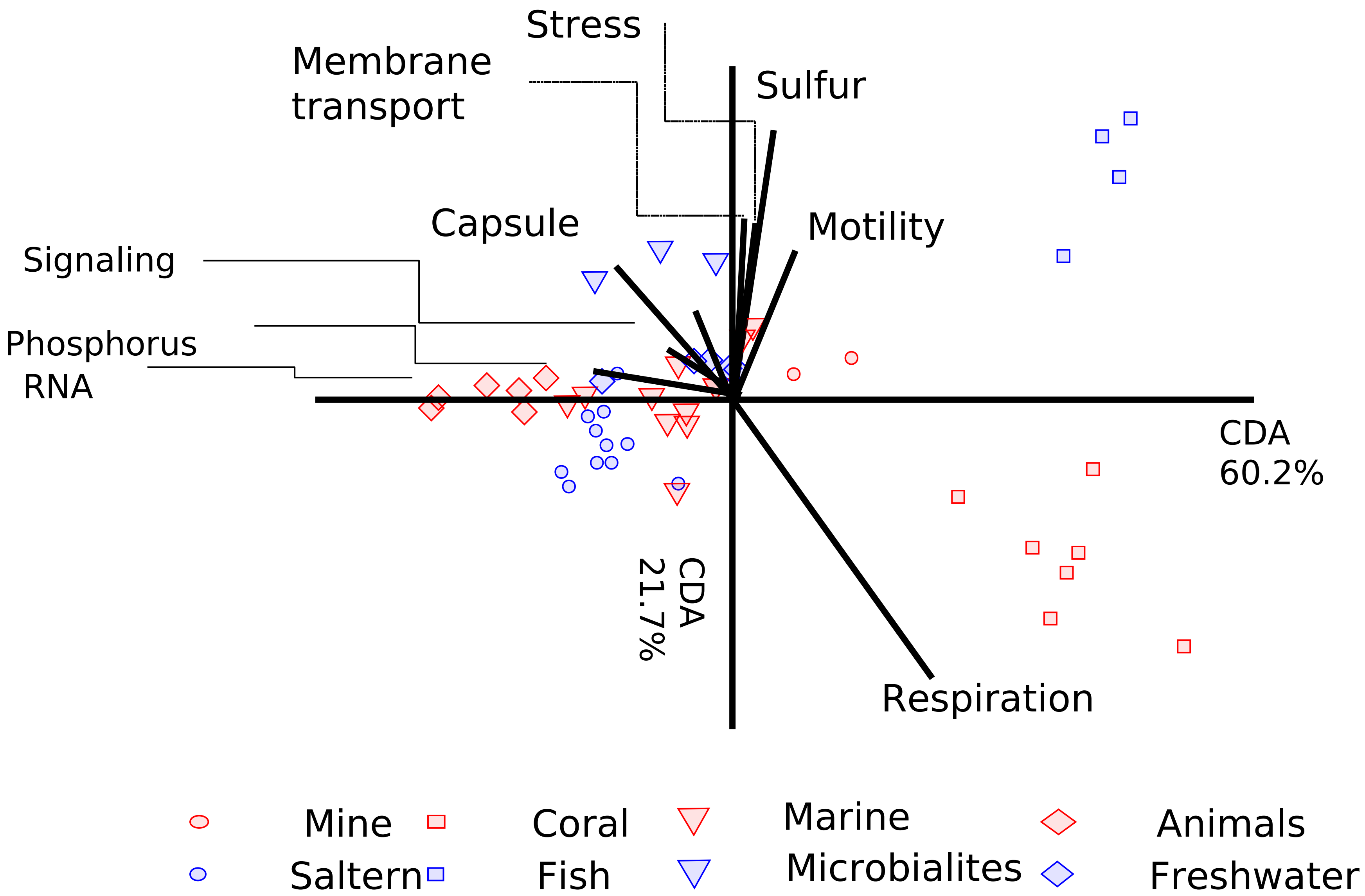
- Compare medians from sites 1 and 2 with 95% CI

Rodriguez-Brito et al., (2006)

Canonical Discriminant Analysis

- Classification technique –
  - metagenomes divided into groups
  - Variables – percent metabolic pathway or taxonomic group
  - Assumptions must have more than one metagenome in each group
  - Variables must not be overly correlated
  - Trying to classify metagenomes on the predictor variables – trying to make groupings

From Sequences To Environments



Dinsdale et al, Nature 2008

Multi-dimensional scaling

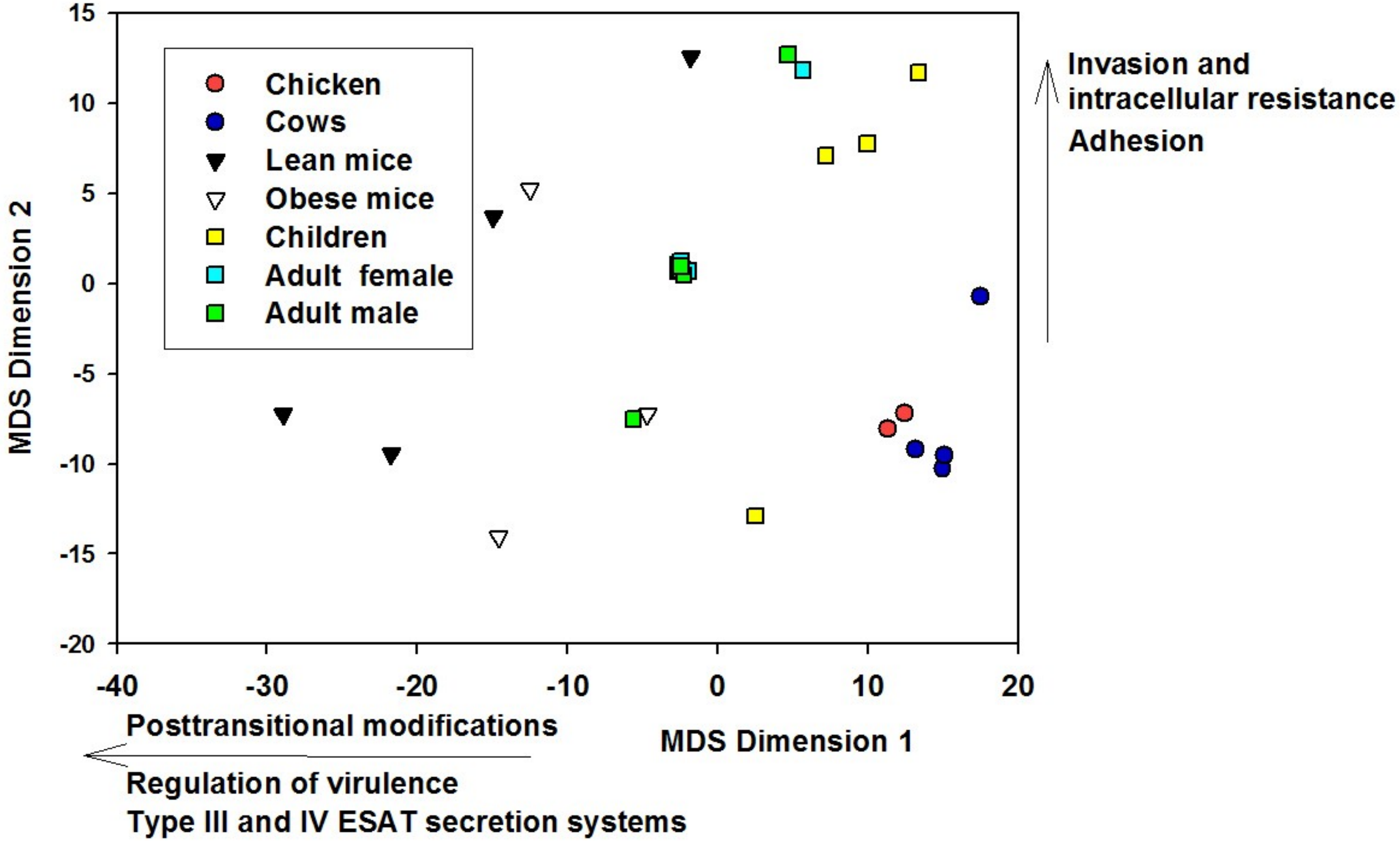
- Finds structure in a set of measurements

- Can use data from multiple sources, eg metabolic and taxonomic in same analysis

- Variables must not be too different in scale, i.e. not dollars compared to years

- Few assumptions on the data

MDS- difference in virulence systems



Invasion and intracellular resistance

Adhesion

Posttransitional modifications

Regulation of virulence

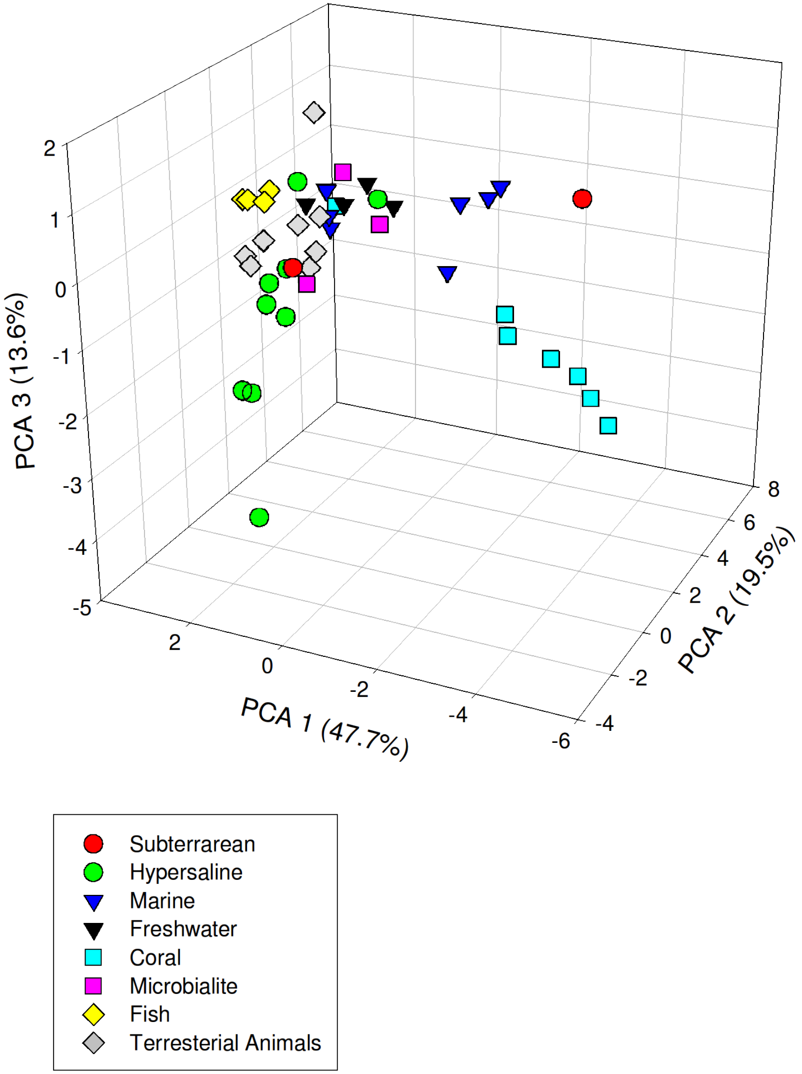
Type III and IV ESAT secretion systems

Qu et al 2008 Plos One 3(8)

Principal Component Analysis

- Data reduction – good when you have lots of variables
- Looking for the factor that is explaining the variation in the data
- Metagenomes are not grouped prior to analysis
- Normal data, unique variables i.e. they do not overlap

PCA to identify if dinucleotides are distributed by environment



Willner et al 2009

BACK!