Multivariate Analysis of Metagenomes – An Undergraduate REU Story

Apkarian, N.[1], Creek, M.[2], Guanz, E.[3], Hernandez, M.[4], Isaacs, K.[5], Peterson, C.[1], Regh, T.[6], Edwards, R.A.[7], Bailey, B.[8], Salamon, P.[8], Tuba, I.[9], and E. A. Dinsdale[10]

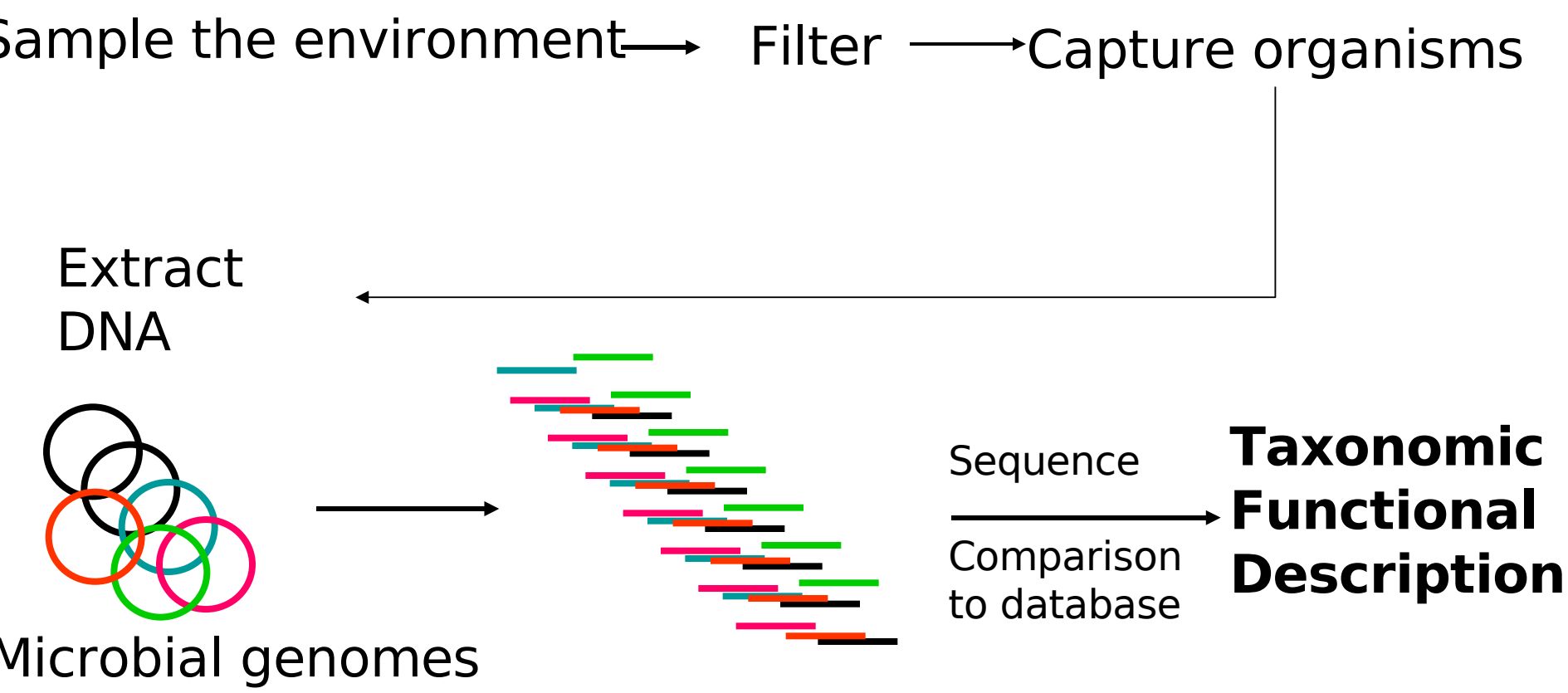
CSRC Colloquium

23rd Oct, 2009

Microbes, Microbes, Microbes

- Everywhere
- Individual health
- Ecosystem health
- Global “Health” – Biogeochemistry
- Problem – small, difficult to grow, hard to study
- Solution – extract and sequence DNA - Metagenomics

Metagenomics



Metagenomics successfully describes microbes across multiple environments

- Coral reef degradation
- Microbialite growth
- Coral metabolism
- Shark skin
- Virulence metabolism associated with different mammals
- Functional profiling across environments

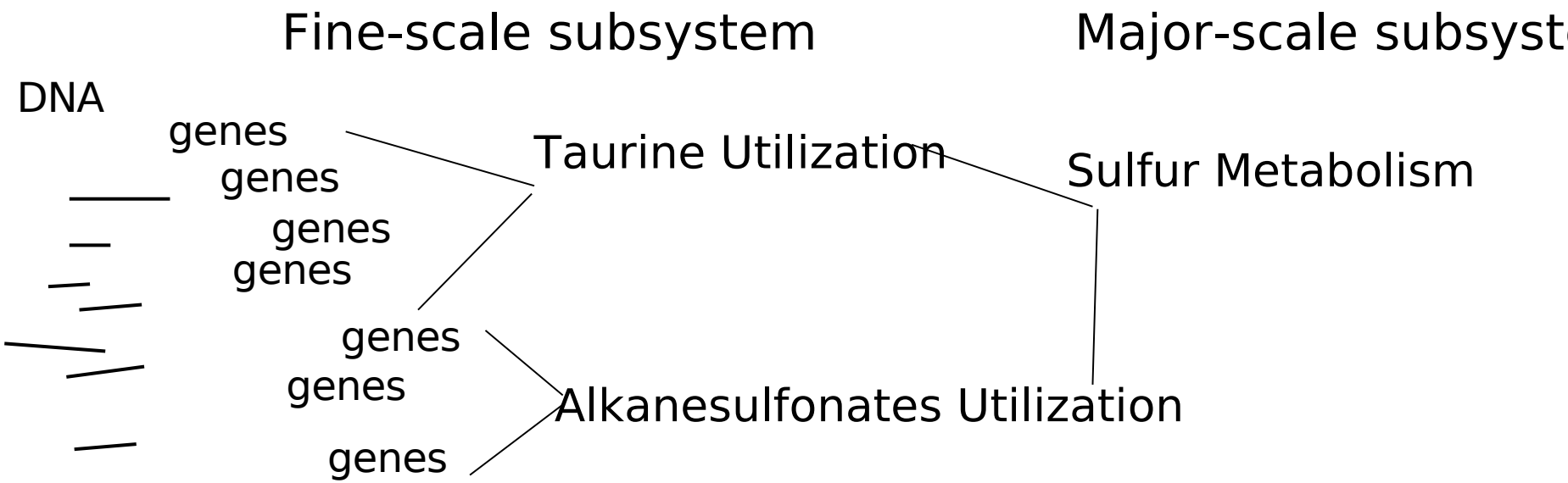
Problem – too successful!

2000 metagenomes – 20 billion sequences

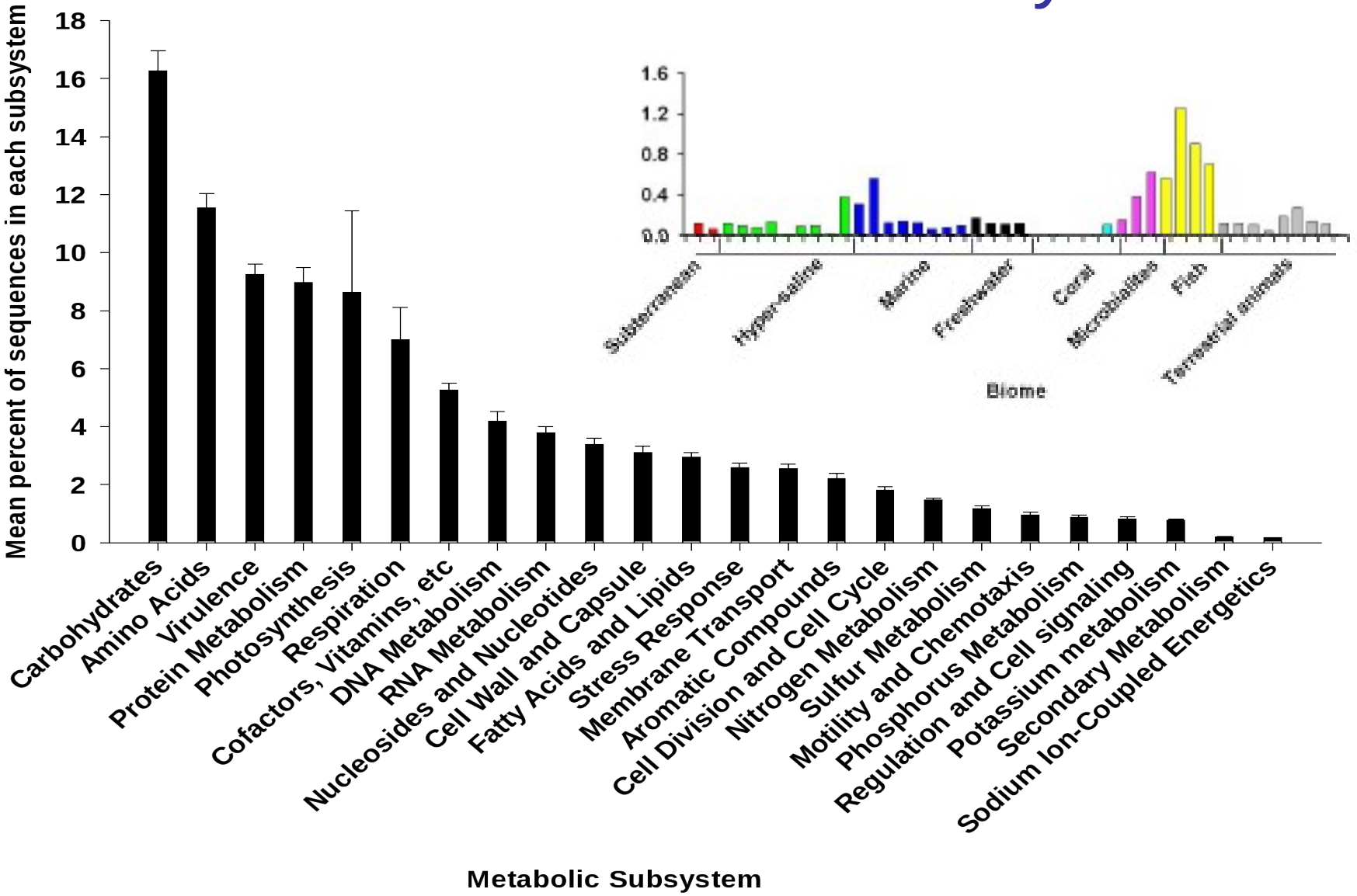
Metabolic functions across metagenomes from different environments

- 200 metagenomes
  - Marine, humans, terrestrial animals, springs, hyper-saline, microbial mats,
  - All over the world
  - Range of sequence lengths and therefore sequencing technologies
  - 27 major metabolic pathways
  - Analyzed against a single database

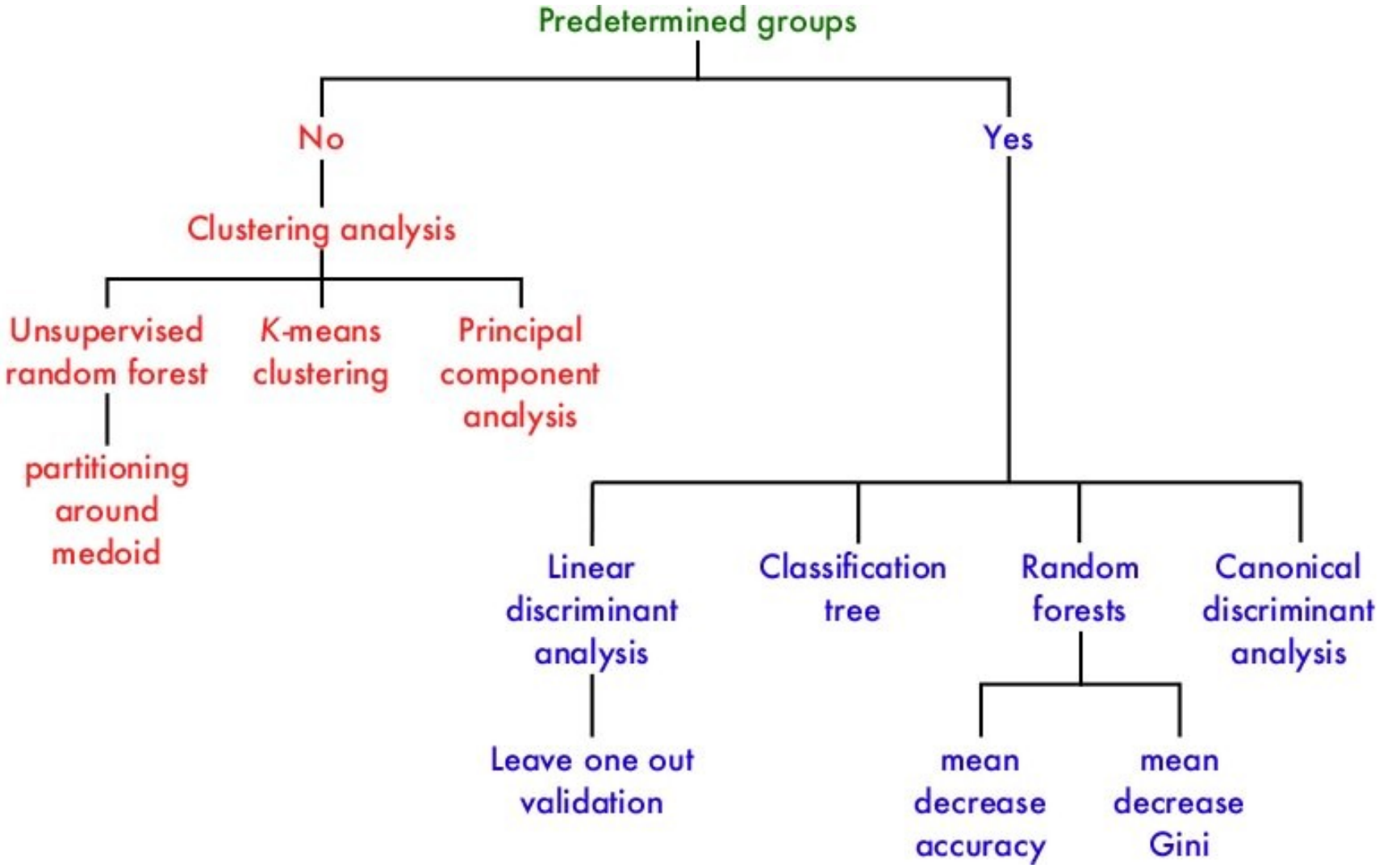
Real-time metagenomic functional database



The sequences showed a range of abundances across the subsystems



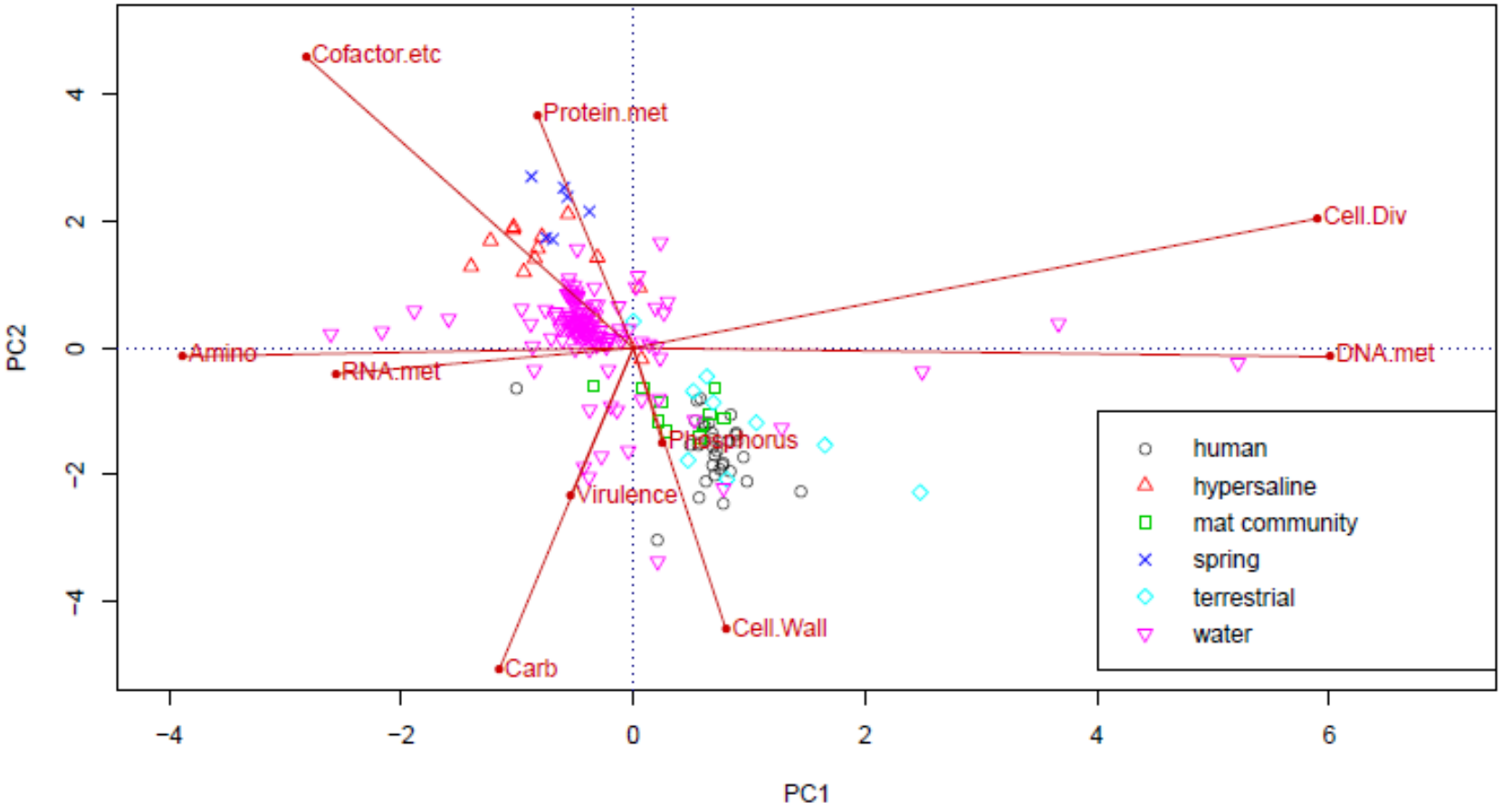
What Multivariate statistics should be used?



Principal component analysis

- Orthogonal linear combination of variables in a dataset that maximizes variance
- Variance is information
- Dimensional reduction that preserves variance
- PCA graphs plot the transformation of the data over the first 2 principal components
- PCA graphs visually identify clusters and outliers

PCA – shows the metagenomes in 2D space



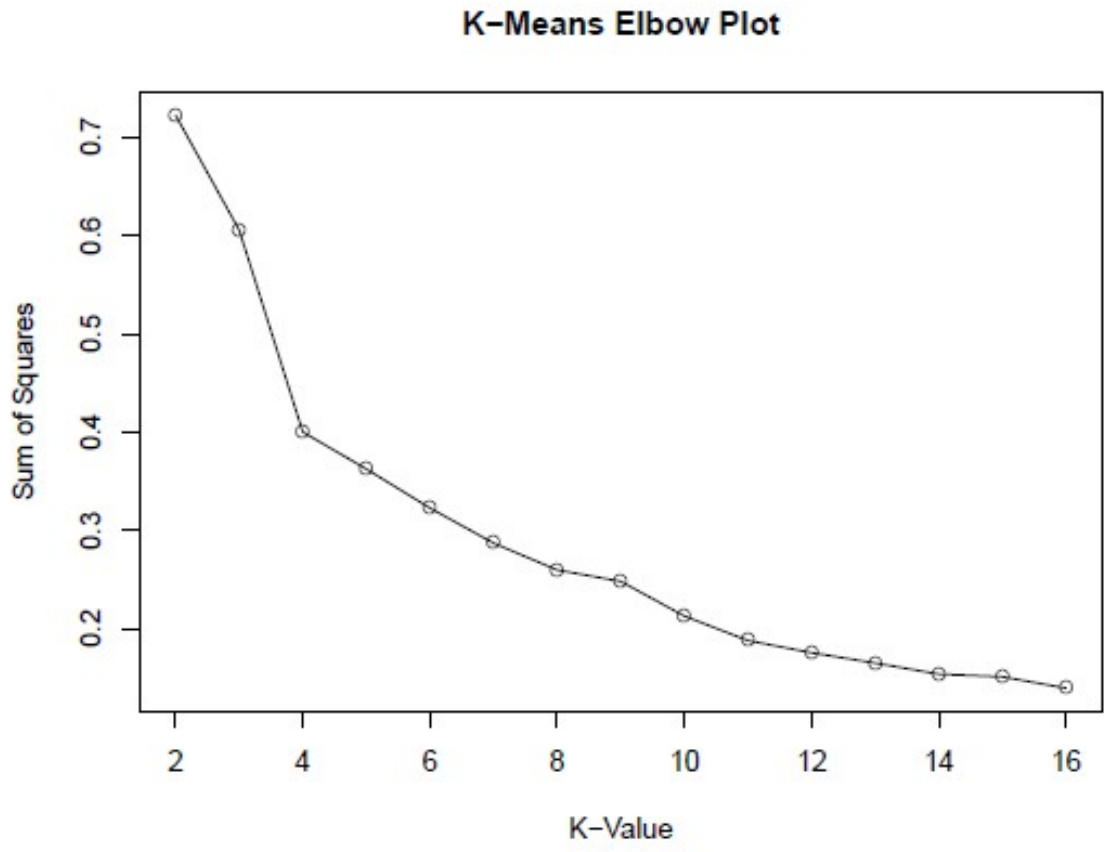
PCA and Metagenomics: Summary

- **Advantages**
  - Reduces dimensions while preserving variance
  - Visual clustering
  - Unsupervised
  - Information about variables
- **Disadvantages**
  - Reducing dimensions may greatly reduce variance
  - Ineffective with too much data
  - Strictly linear

K - means

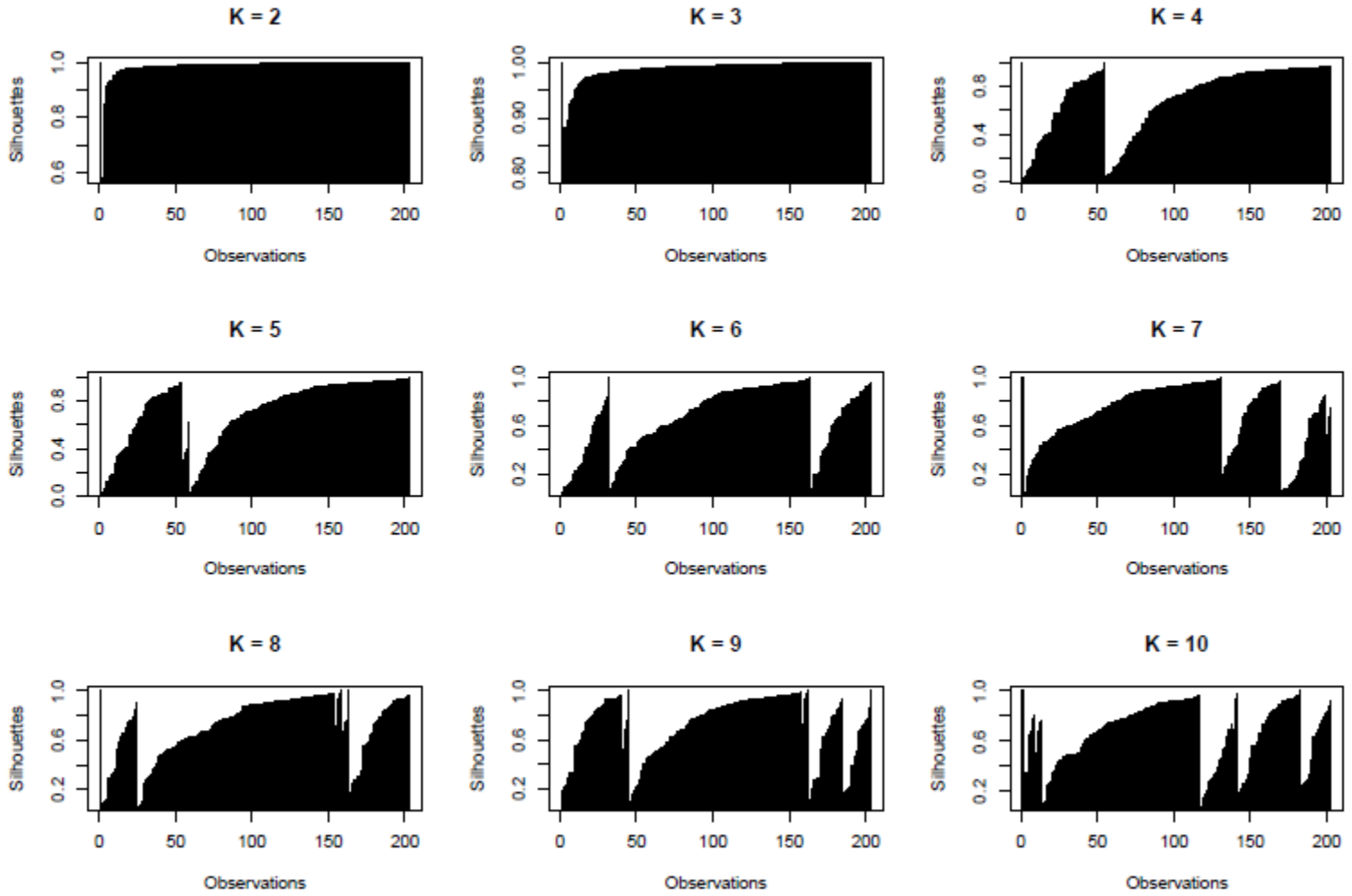
- Classifies observations into K clusters
- Minimize the sum of squared distances from each observation to the mean of its assigned group
- Group means are calculated and randomly replaced
- Repeating algorithm
- Initialize numerous times

Difficult to find clear elbow – which would identify the number of groups



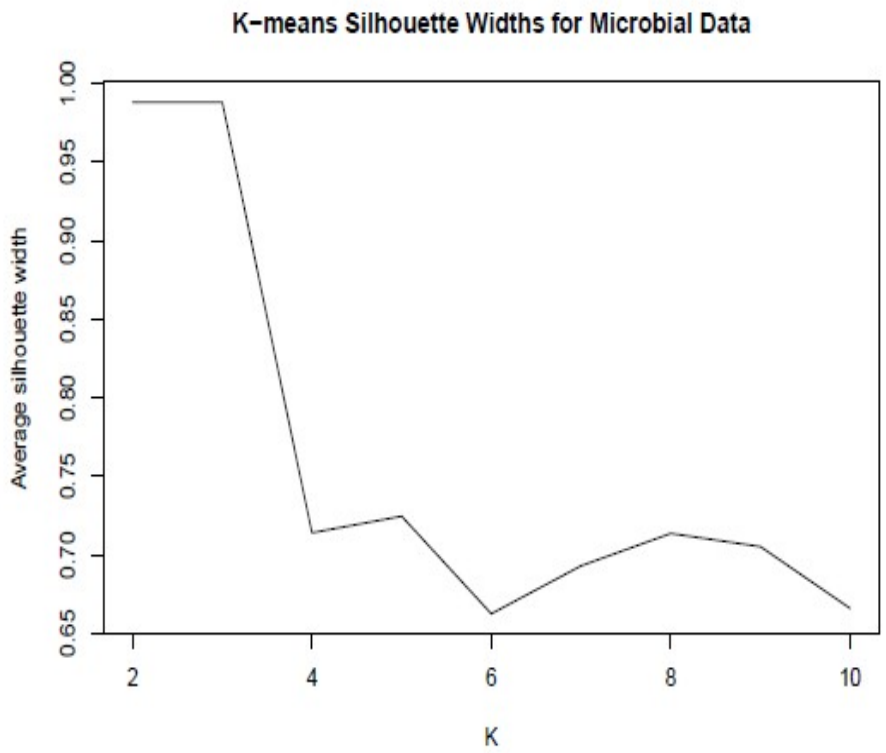
Silhouette plot

Average silhouette width



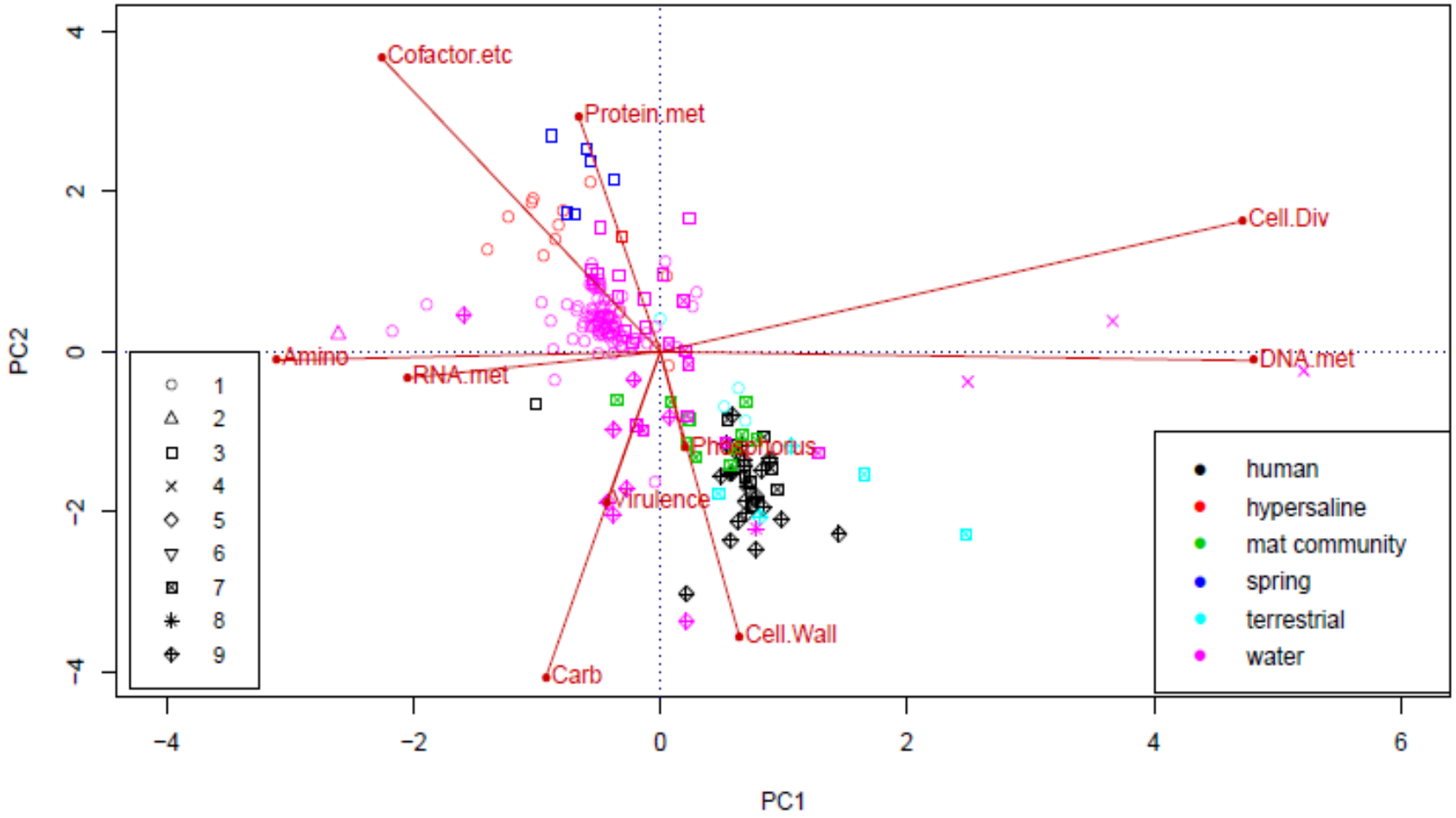
Observations - metagenomes

K- mean output



Cluster	Environmental data within the cluster
1	Hyper-saline (except 1) and water samples
2	1 water sample
3	All spring samples, 1 human (outlier), 1 water sample
4	3 water samples
5	1 water sample
6	1 water sample
7	All mat community samples, 5 human, 3 terrestrial animals
8	1 water sample
9	Remaining human, 2 terrestrial animals, water

9 clusters visualized in the PCA



K-means obtaining groups

Advantages

- Non-biased classification
- Tool to check assumed groupings
- Paired with visualization methods (PCA)
- Aids in the identification of outliers by partitioning them into separate groups

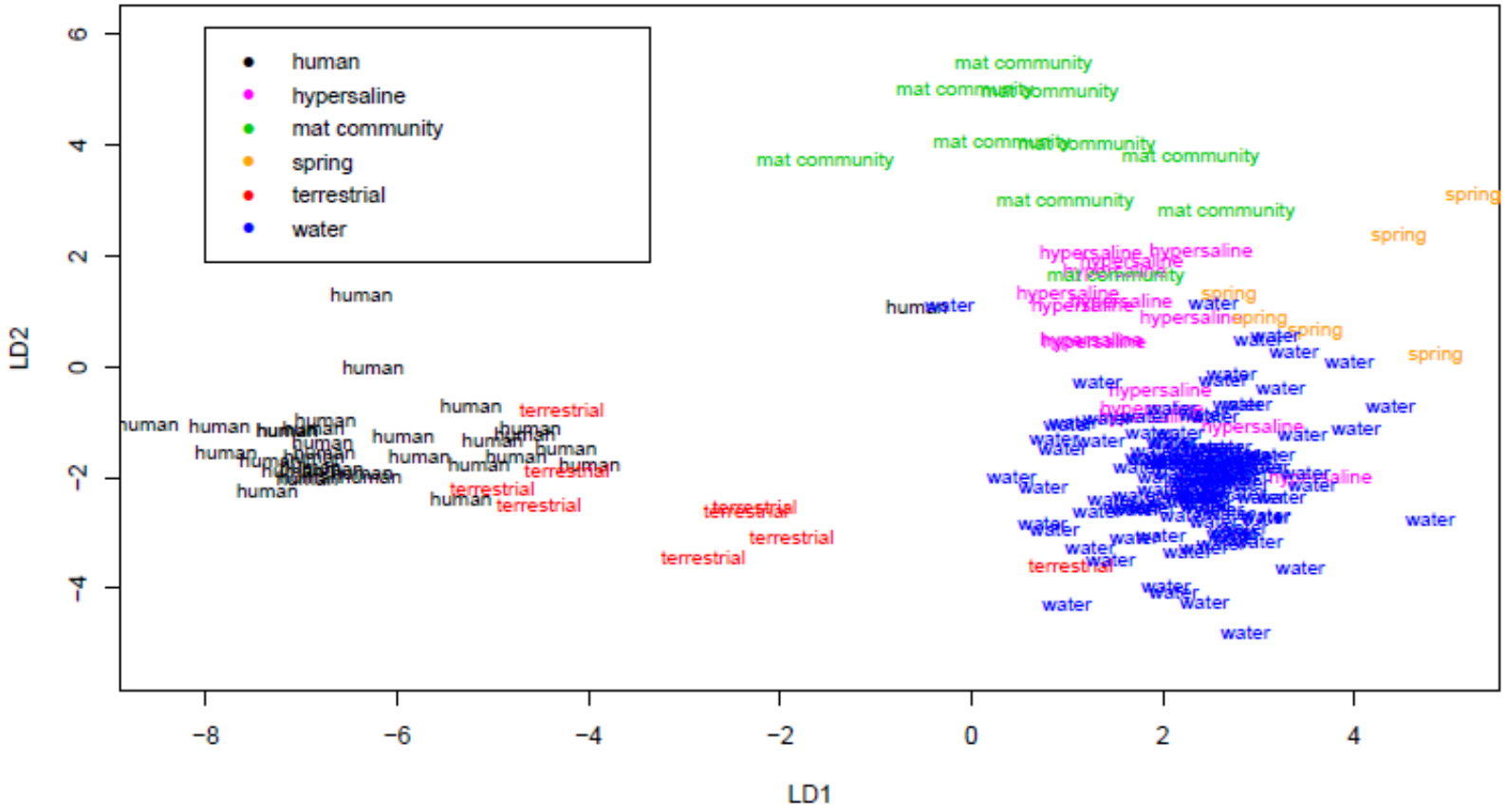
Disadvantages

- Frequently there is no indication of a best K
- May produce clusters that are not identifiably meaningful
- If an outlier is not partitioned into its own group, it could highly influence the selection of means

Linear discriminant analysis

- Constructs linear combinations of the variables in a manner that best separates the given groups.
- These functions are hyperplanes that cut through the space in the dataset.
- Used to classify new data.
- Plots of the data projected into the 2-D space formed by two discriminant functions
- Validity of an LDA as a classifier is judged using “leave one out” cross-validation.

All 27 variables



LDA

Advantages

- Useful tool for visualizing the separate groups.
- Biological insights
- Judge the validity of the LDA “leave one out” cross-validation

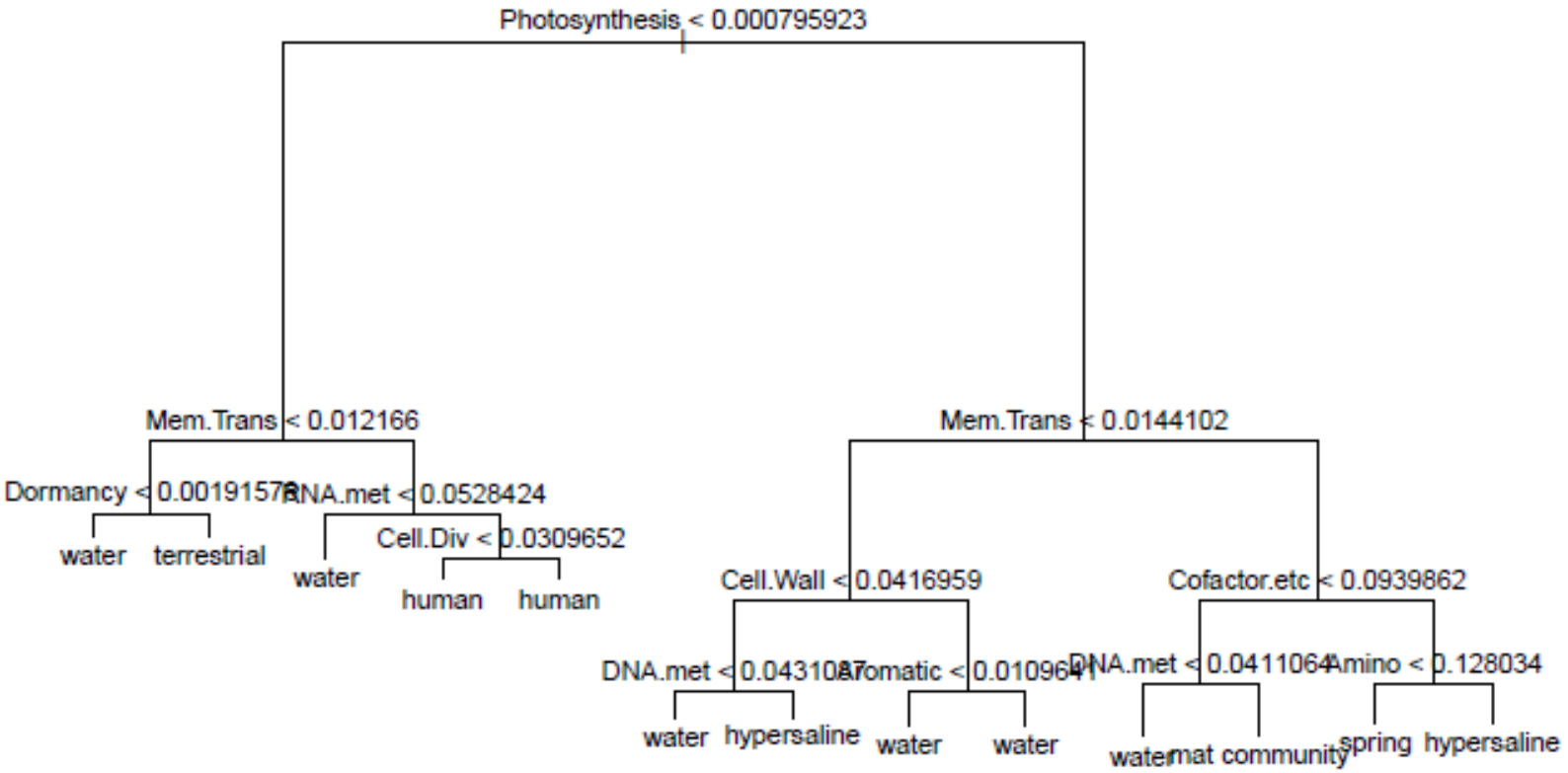
Disadvantages

- Two Assumptions
  - Groups normal distribution
  - Each group’s conditional distribution has the same covariance matrix
  - Linear functions may not the best way to discriminate - Quadratic,

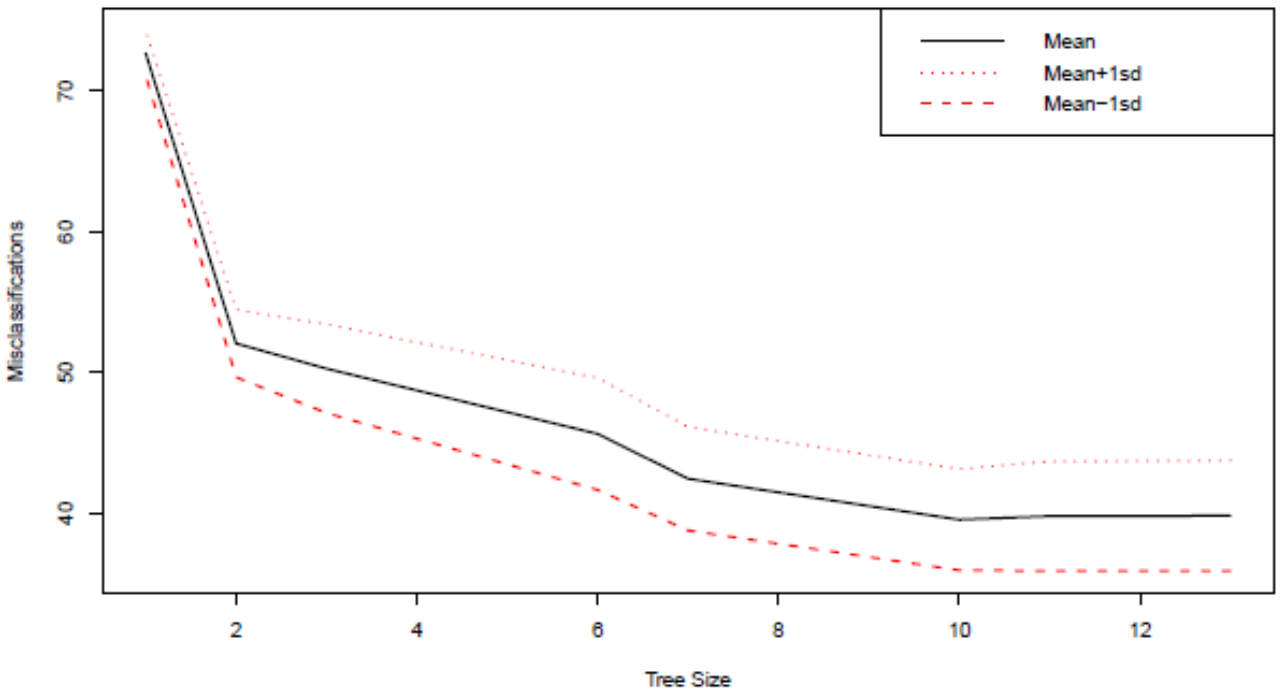
TREES

- Implementations of two types of statistical procedures: classification and regression
- Graphical models relating variables to data classes
- Representations of rectangular divisions in a sample space
- Tools for variable selection and prediction

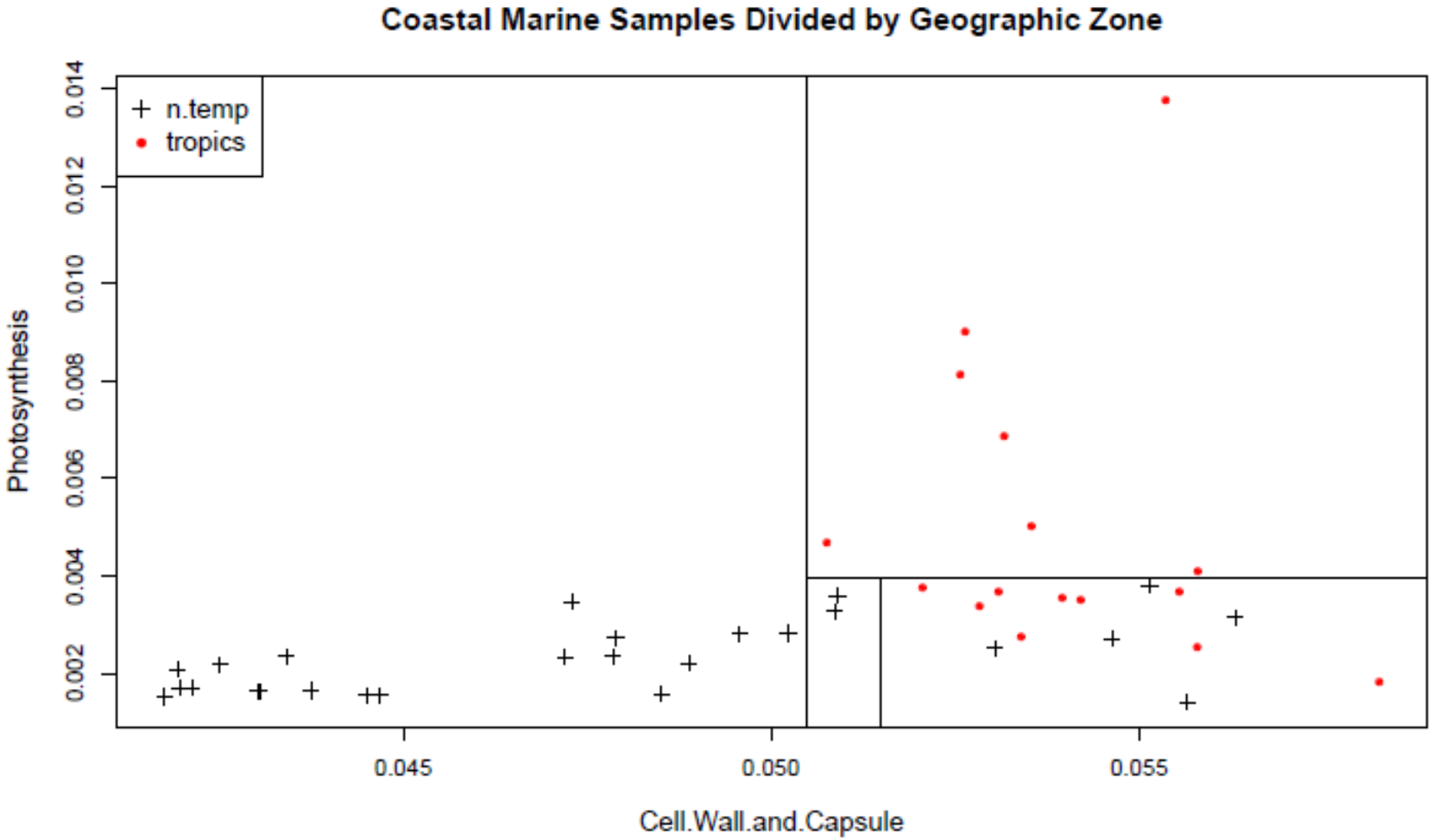
Groups and describes the point at where the data splits



Variance to identify correct tree size



Rectangle plots – visualize distribution of data across two variables



Trees

Advantages:

- Do not require variable scaling (trees are invariant under monotonic transformations of the predictor variables)
- Can be cross-validated and fit with model selection algorithms and variable selection tools
- Model selection tools limit the risk of overfitting

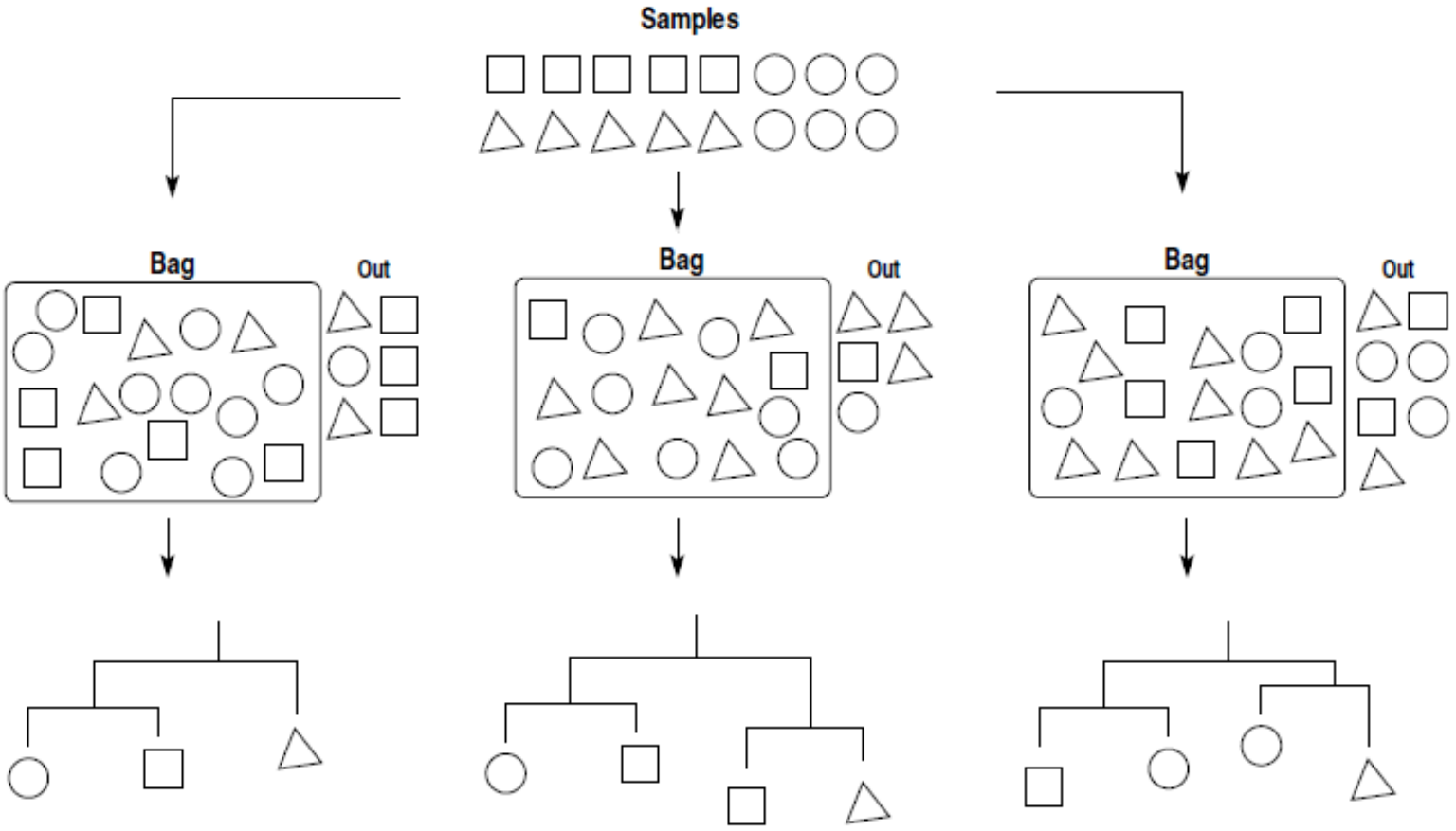
Disadvantages:

- Lack stability with respect to small changes in the data set
- Do not account for linear combinations between variables
- Have limited utility as methods of variable selection

Random forest

- Composed as a set of trees
- Random subset of all the training metagenomes using **b**ootstrap **ag**gregation (bag) – sampling with replacement.
- Un-sampled metagenomes in each set are called out-of-bag.
- Each node in each tree is determined from a random subset of all the variables.
- Instead of classifying new data by tree branching rules, Random Forest classifies by vote of its component trees.

Out of Bag classification



Flexibility in Random forest analysis

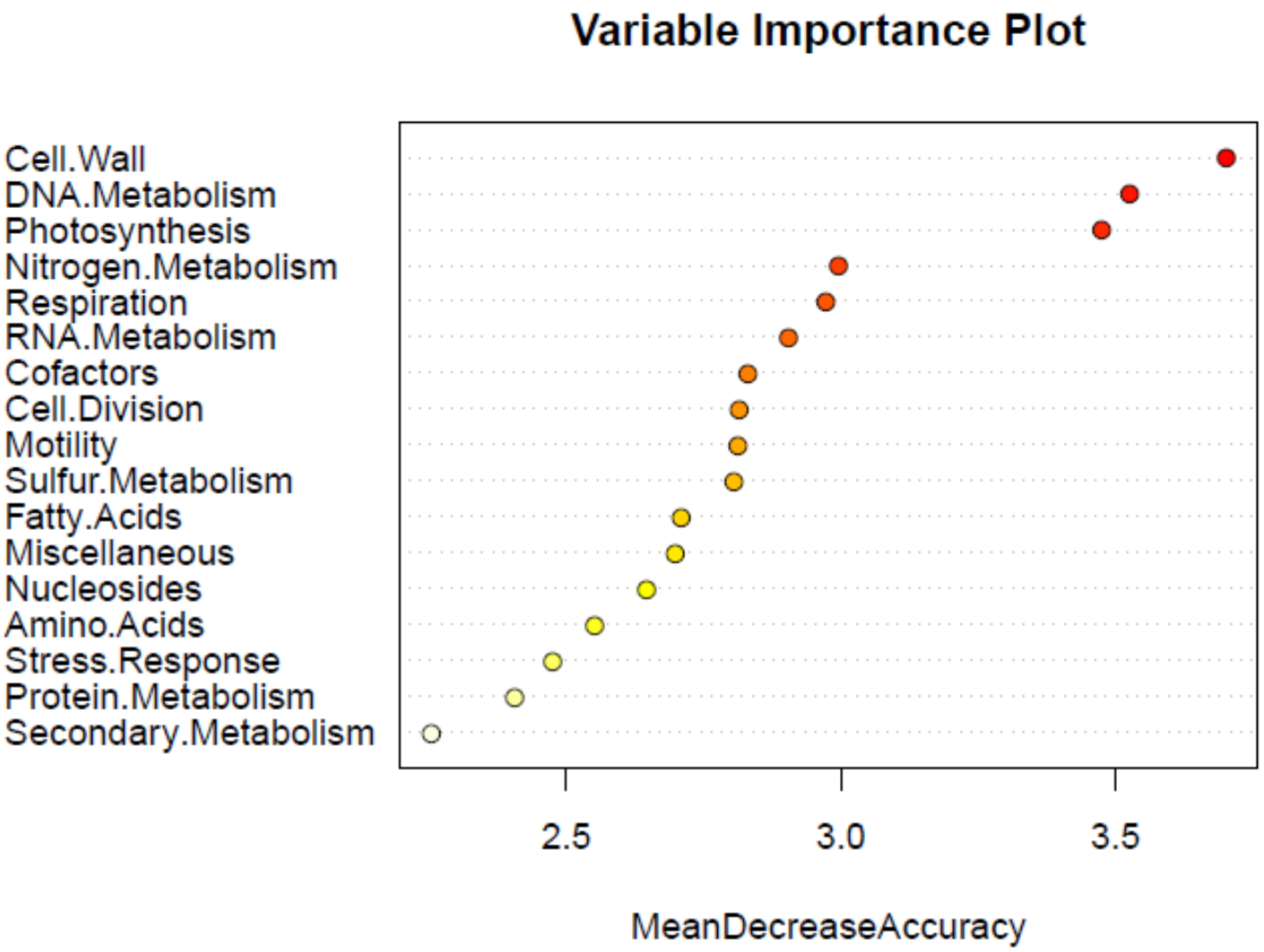
Supervised:

- In a supervised Random Forest, groupings for the training data are input to the algorithm.
- Estimated classification error is computed using out-of-bag data.

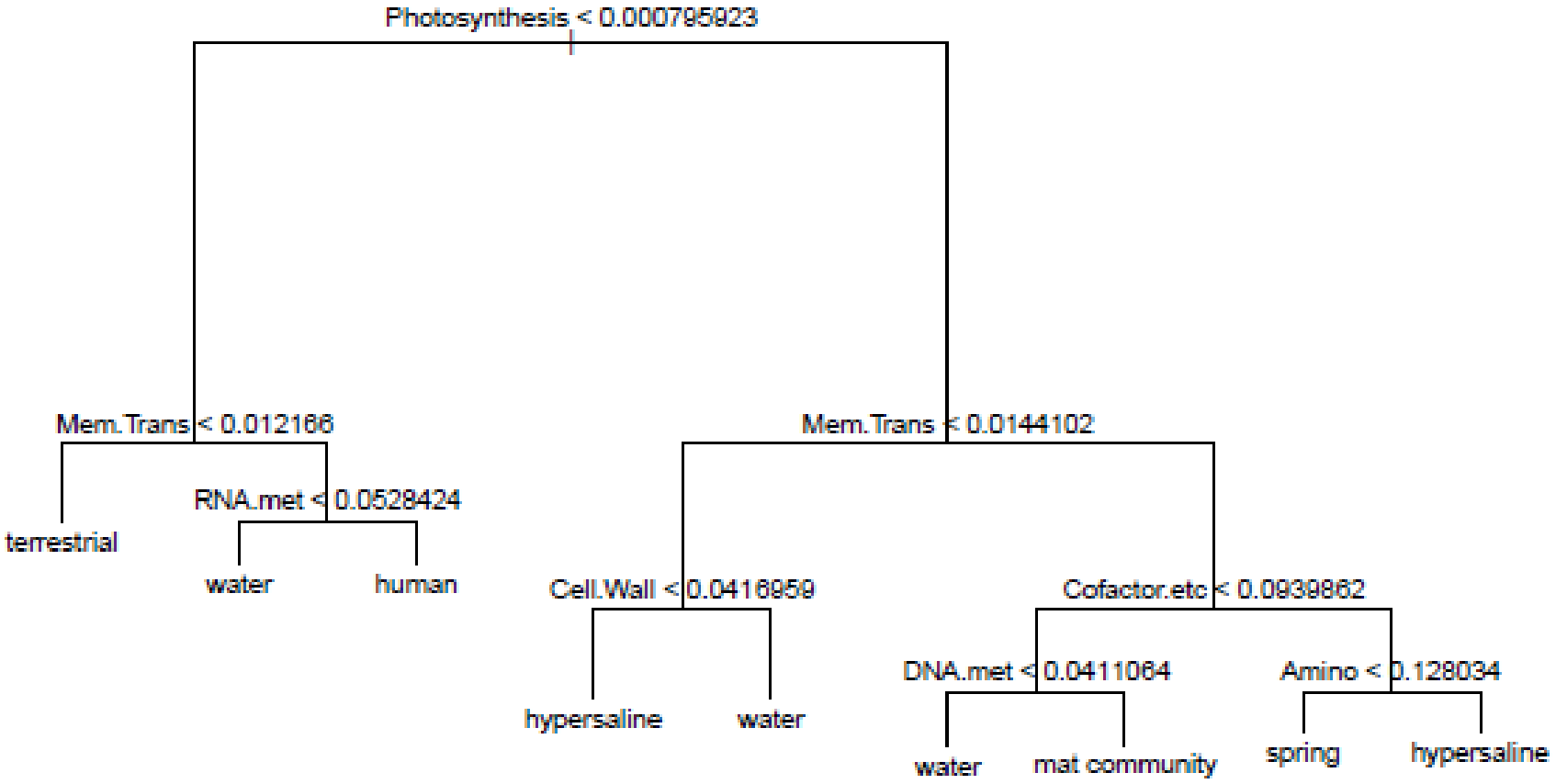
Unsupervised:

- In an unsupervised Random Forest, groupings for training data are not given.
- The Random Forest algorithm creates random synthetic groupings instead.
- While generating the trees, similar data will be difficult to separate, despite different synthetic groups.

Visualize important variables



Clustering and the level at which the variation occurs



Random forests

Supervised Random Forests

Advantages:

- More robust than single trees
- Prediction error automatically generated
- Yields variable importance measures

Disadvantages:

- No branching rules, predicts as a black box
- Error may be deceptive when class sizes vary greatly

Unsupervised Random Forests with PAM

Advantages:

- Clusters emerge without bias from initial groupings
- Varying group sizes do not have detrimental effect

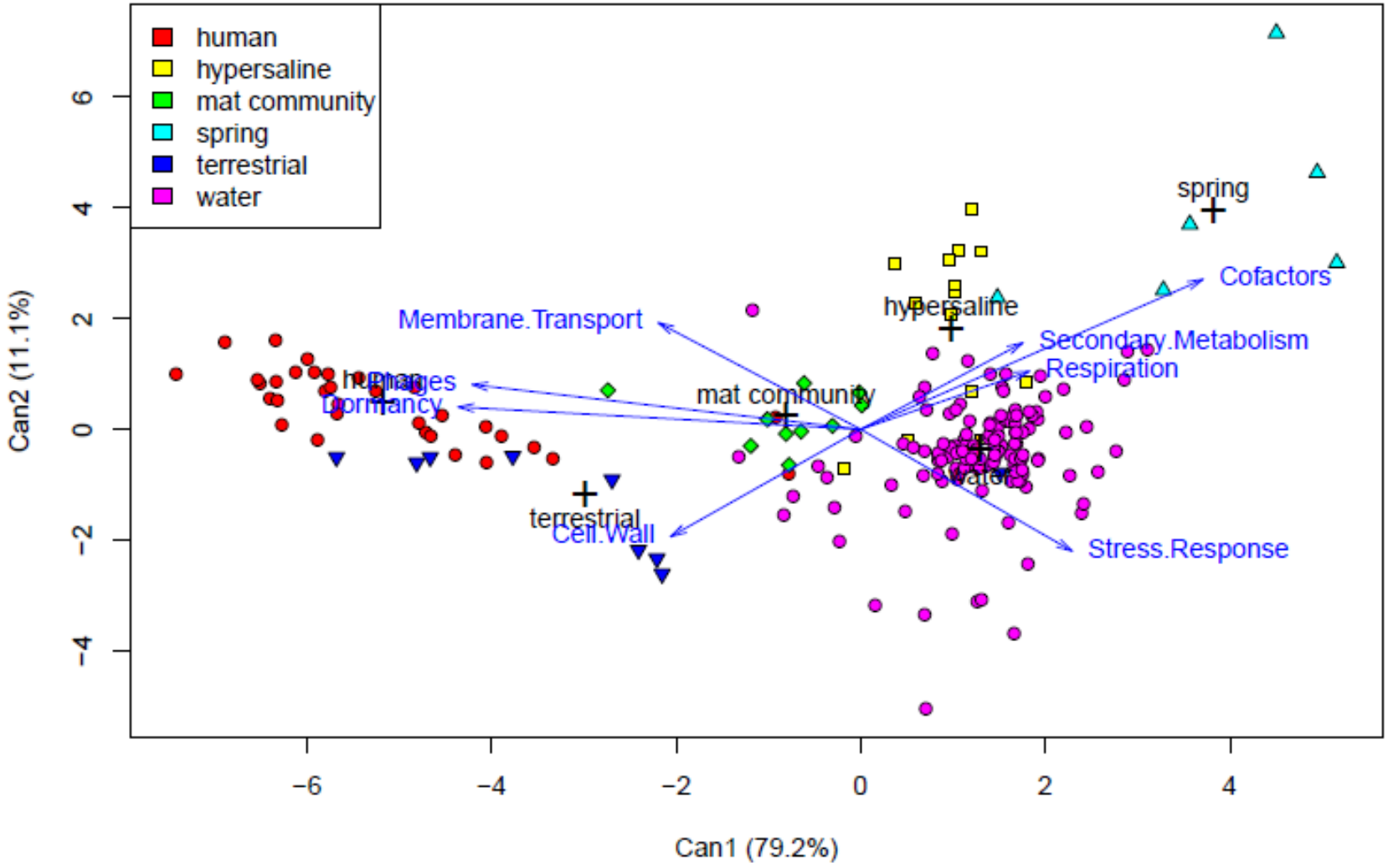
Disadvantages:

- No variable importance or weighting

Canonical discriminant analysis

- Finds axes in k-dimensional space that best separate the given classes
- Canonical components - Uncorrelated linear functions that best explain the variance between classes
- Importance of variables to differentiating groups
- Classification and prediction abilities

Separation across environments



CDA

Advantages:

- Excellent classifier
- Clear visualization
- Can be combined with other techniques
- Prediction

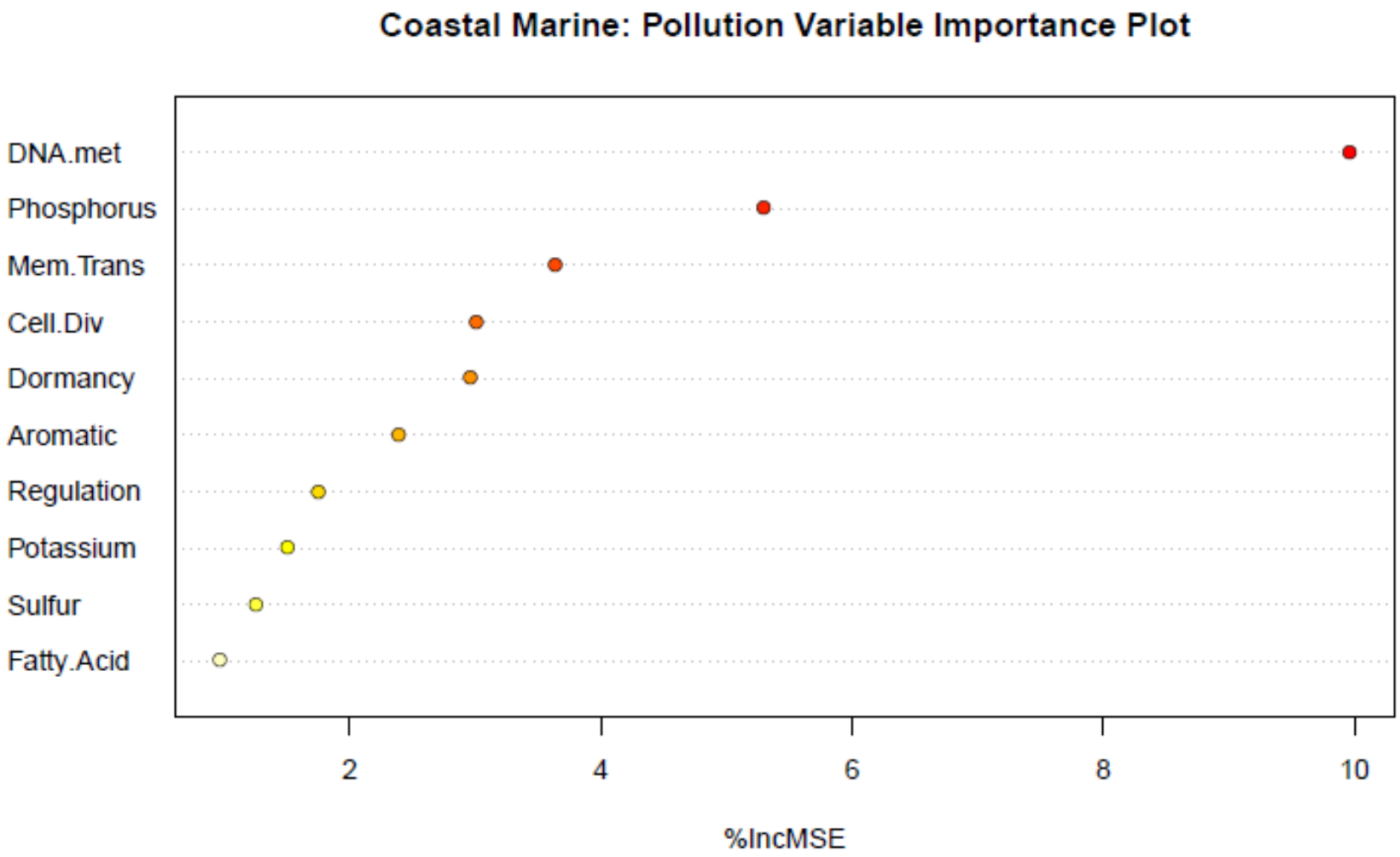
Disadvantages:

- Overfitting/Artificial separation of classes
- Sample size restrictions
- Supervision bias
- Confounding variables

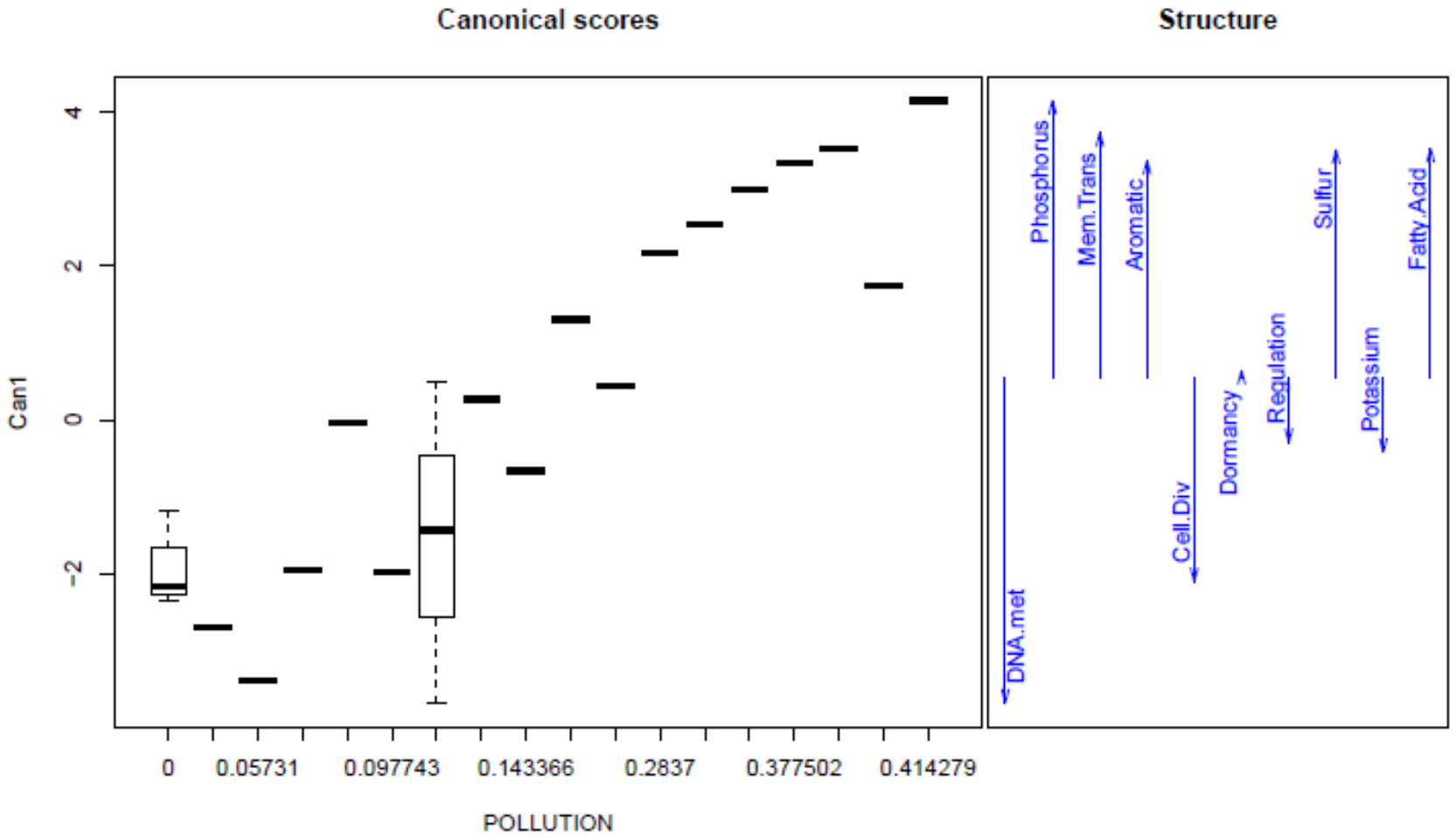
Process linking Maths and Biology

- Students – developed new methods
- Put the Maths and statistics into practical use
- Real data unknown outcomes
- Inspired to go further
  - Metabolisms from fresh to hypersaline
  - Viral and microbial relationship
  - Internal and external metabolism
  - Metabolisms with human activity

Open ocean with Pollution index



Metabolic processes changed with pollution



Summary

- Identified combinations of statistics
- Presentations – Hawaii, Brazil, San Diego,
- Student presentations – Biomedical school at Stanford
- Paper and book chapter in preparation

Participants: Apkarian, N.[1], Creek, M.[2], Guanz, E.[3], Hernandez, M.[4], Isaacs, K.[5], Peterson, C.[1], Regh, T.[6], Edwards, R.A.[7], Bailey, B.[8], Salamon, P.[8], Tuba, I.[9], and E. A. Dinsdale[10]

1. Pomona College, 333 N. College Way, Claremont, CA 91711, USA
2. Chapman University, One University Drive, Orange, CA 92866, USA
3. Torrey Pines High School, 3710 Del Mar Heights Road, San Diego, CA 92130, USA
4. Computational Sciences, San Diego State University, 5500 Campanile Dr. San Diego, Ca 92182, USA
5. San Jose State University, One Washington Square, San José, Ca 95192, USA
6. Southern Oregon University, 1250 Siskiyou Boulevard, Ashland, OR 97520, USA
7. Computer Sciences, San Diego State University, 5500 Campanile Dr. San Diego, Ca 92182, USA
8. Mathematics and Statistics, San Diego State University, 5500 Campanile Dr. San Diego, Ca 92182, USA
9. Mathematics and Statistics, Imperial Valley Campus, San Diego State University, 720 Heber Ave, Calexico, Ca 92182, USA
10. Biology Dept, San Diego State University, 5500 Campanile Dr. San Diego, Ca 92182, USA