# Multivariate Statistics in Marine Genomics

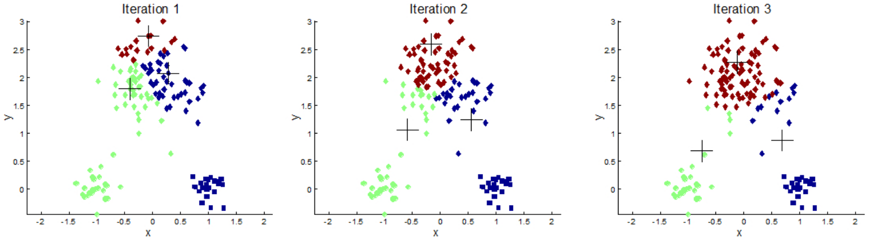
In addition to gigantic fasta files of sequence data, we often have a wealth of environmental data to take into account (like temperature, depth, salinity, pH, location, just to name a few). It can be really hard to figure out which factors are important when there are so many. This is to just give you a general idea of the logic behind some common analyses seen in the literature and give you a better understanding to approach environmental genomics data.

## What is multivariate statistics? How can it help me in marine functional genomics?

Simply put, multivariate statistics deals with observations made over many different variables. Multivariate statistics extracts information about multiple variables and how they might be related to each other, and to the dependent variable. A search or optimization algorithm is employed to find combinations of variables that best explain the variation in the observed data. They reduce model complexity while preserving predictive power. Unfortunately, these algorithms tend to transform the variables in ways that are not very easy to interpret, multivariate analyses are less useful in trying to explain biological phenomenon than predict them.

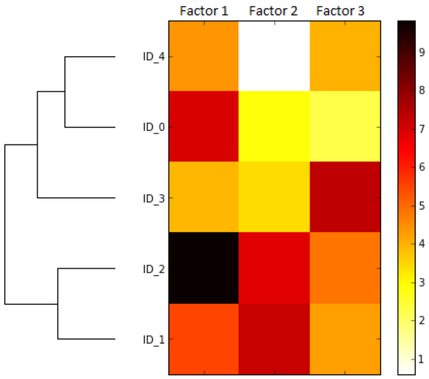
## Cluster analysis

As the name implies, cluster analyses create groups based on how much the observations ‘resemble’ each other. Common cluster techniques include neural networks and nearest neighbor algorithms. These methods are useful when there are no pre-defined categories or groups in the dataset, pattern recognition, and complexity reduction.



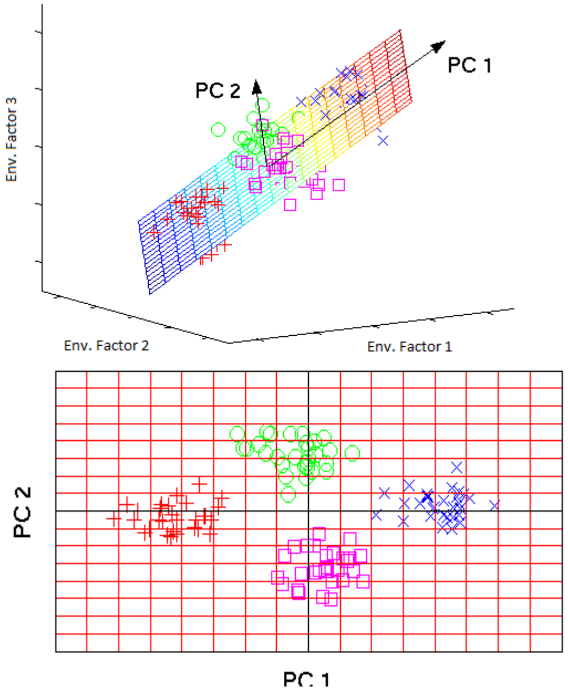
k-means clustering creates groups in steps, altering the factors until the algorithm is satisfied with how separate the groups are. The final combination of factors that best separates the data can identify patterns and sources of variation.

Hierarchical clustering is oftentimes represented as a tree and/or heat map, where similar observations in factors (e.g. environmental parameters, gene expression, etc.) are grouped together.



## Ordination

A broad class of analysis, ordination includes many types of analysis that may be familiar to genomics students such as PCA or NMDS. These methods ‘compress’ multiple variables into combinations that account for the greatest amount of variance in the data. They create ‘new’ X and Y coordinates based on those combined variables, which can emphasize variation and bring out strong patterns.



In this PCA example, the original data (top) is plotted in 3D based on 3 different environmental factors. The raw data looks like it has some patterns, but it’s a little unclear.

The PCA (bottom) combines the 3 factors to reduce it to 2D plot, based on principal components 1 and 2 (PC1 and PC2). There’s now a much more striking pattern in the data. Additionally, we now only have to test 2 factors (PC1 and PC2) rather than 3.

In this example, we only used three factors, but in reality you can compress any number of factors together into just 2 or three PCs, which allow you to focus on those factors that contribute the most to the data, and ignore the rest.