Multivariate analysis of functional metagenomes

*Elizabeth A. Dinsdale[1], Robert A. Edwards[1,2,3], Barbara Bailey[4], Imre Tuba[5], Sajia Akhter[6], Katelyn McNair[6], Robert Schmieder[6], Naneh Apkarian[7], Michelle Creek[8], Eric Guan[9], Mayra Hernandez[4], Catherine Isaacs[10], Chris Peterson[7], Todd Regh[11], Vadim Ponomarenko[4]

Author Affiliations:

[1]Department of Biology, San Diego State University, 5500 Campanile Dr, San Diego, 92182

[2]Department of Computer Science, San Diego State University, 5500 Campanile Dr, San Diego, 92182

[3]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439

[4]Department of Mathematics and Statistics, San Diego State University, 5500 Campanile Dr, San Diego, 92182

[5]Department of Mathematics and Statistics, San Diego State University, Imperial Valley, 720 Heber Ave, Calexico, CA 92231

[6]Computational Science Research Center, San Diego State University, 5500 Campanile Dr, San Diego, 92182

[7]Pomona College, 550 North College Avenue, Claremont, CA 91711-6301

[8]Chapman University, One University Drive, Orange, CA 92886

[9]Torrey Pines High School, 3710 Del Mar Heights Road, San Diego, CA 92130-1316

[10]San José State University, One Washington Square, San José, Ca 95192

[11]Southern Oregon University, 1250 Siskiyou Boulevard Ashland, OR 97520

*Corresponding author:

Department of Biology, San Diego State University, 5500 Campanile Dr, San Diego, CA 92182

Email: elizabeth_dinsdale@hotmail.com

Tel: 619 594 5623

Fax: 619 594 5676

Keywords: metagenomics, statistics, microbiology, random forest, canonical discriminant analysis, principal component analysis

Abstract

Metagenomics had been established as a major tool for the description of microbial and viral communities. The shear magnitude of data generated in each metagenome makes identifying key difference in the function and taxonomy between communities difficult to elucidate. Here were present 7 statistical analyses that could be used to compare and contrast the metabolic functions of microbes (or viruses) within and between 10 environments. Random forests provided a robust and enlightening description of both the clustering of metagenomes and the metabolic processes that were important in separating microbial communities from different environments. All analysis identified that the presence of phages genes within the microbial community was a predictor whether the microbial community was host associated or free living. Genes that are contributed by phage are a key contributors to the success of microbes invasion and survival within a eukaryotic host.

Introduction

Vast communities of microbes occupy every environment, consuming and producing compounds that shape the local geochemistry. Over the last several years sequence based approaches have been developed for the large-scale analysis of microbial communities. This technique, typically called metagenomics, involves extracting and sequencing the DNA *en masse*, and then using high performance computational analysis to associate function with sequence.

Most of the focus in metagenomics has been on single environments such as coral atolls[1,2], cow intestines[3], ocean water[4], and microbiolites[5]. Early work compared extremely different environments, like soil and water[6]. More recently, the Human Microbiome Project has expanded our understanding of the microbes inhabiting our own bodies, comparing samples from the same site among and between individuals[7-9]. Previously, we demonstrated that analysis of functional diversity in metagenomes could differentiate the microbial processes occurring in multiple environments[10]. That study utilized the only publicly available metagenomes at that time: 45 microbial samples and 32 viral samples. The raw DNA sequences were compared to the SEED subsystems[11], and the normalized counts of the number of sequences in each subsystem in each metagenome were used as the input. That provided a raw data set with 23 response variables and 82 observations or samples. In that first study, a canonical discriminant analysis was used to separate the important metabolisms occurring in each metagenome. Subsequently, an analysis of the nucleotide composition of metagenomes used a principal components analysis to discriminate different environments. That work showed that dinucleotide frequency was preferred over other orders (tri-, tetra-, etc). Dinucleotide abundances provide 16 variables that were used to separate 86 samples[12].

There are a wide range of statistical tools that can be applied to multivariate data like metagenomes, however different tools provide different information and vary in their strengths and weaknesses. Here we provide an overview of different statistical techniques that can be used to compare and contrast metagenomes from different environments, and to discover how functional groups can differentiate between and within environments. We briefly introduce each statistical method and describe its ability to separate metagenomes across environmental space. This analysis recapitulated the discriminating power of metagenomes to identify differences in functional potential both between and within environments. A unique signature represented each environment: for example, the abundance of phage proteins was the major discriminator between host-associated microbial environments and free-living microbes. Subtle differences between open and coastal marine environments were associated with differences in the abundance of photosynthetic proteins. Cofactors, vitamins, and stress related proteins were consistently found in higher abundance in environments where the conditions for microbial survival were potentially unstable, such as hydrothermal springs. Each of these differences provides a clue for detailed microbiological analysis of communities.

Results

At the time of analysis, 212 metagenomes were selected from the set of publicly available data. They were classified as coming from ten different environments. The annotations provided composition data for 27 different functional groupings (subsystems). All of the raw data used in this study is provided as Supplemental Online Material and maybe downloaded from http://edwards.sdsu.edu/research/REU2009SupplementalMaterial/. In any statistical analysis it is important to keep the number of response variables less than the number of observations to ensure support for the conclusions that are made. In this study, we used ten classifications (the environments), 27 response variables (the functional groups), and 212 observations (the metagenomes).

Common statistical techniques were used to explore the relationship between the metagenomes, environments, and subsystems (**Supplementary Fig. 1**). In general, statistical methods can be divided into two broad categories: supervised techniques and unsupervised techniques. Supervised techniques require that the samples be separated into predetermined groups before the analysis begins, and those groups are used as part of the analytical methods. In this case, the metagenome samples were grouped according to the environment where the sample was collected. In contrast, unsupervised techniques do not require *a priori* knowledge of the group separations, but the groups are generated by the statistical technique. In the unsupervised cases we compare the resultant groups to the original sampled environment.

When categorizing data, many statistical methods are prone to over-fitting the data – reading more into the data than is really there. In general, increasing the size of the data sets, using similar group sizes to even out the distribution of the data, and limiting the number of groups to be much less than the number of variables avoids over-fitting the data. Sample size considerations are particularly relevant to metagenomic data analysis, due to the nature of the data. There are thousands of proteins identified in each metagenome, but at the time of analysis there were less than 300 publicly available samples, meaning that there were many less samples than potential variables. Combining the proteins into functional groupings reduces the number of variables to be less than the number of samples available (subsystems were used here, but other groups like COGs are also widely used for metagenome analysis[13]).

We begin by assessing the clustering of the metagenomes, and test whether the clusters chosen to reflect the environmental signals are statistically supported (*K*-means, decision trees, and random forests). We then move on to methods to explore and visualize the underlying structure of the data (multi-dimensional scaling, linear discriminant, principal components, and canonical discriminant analysis). Obviously statistical analysis is not a linear process, and many of the techniques were influenced by the results from previous (or subsequent) analysis. Although this discussion attempts to maintain a linear structure for readability, that is not always possible or appropriate. Detailed methods and source code for all of these operations are given in the Supplemental Online Material.

K-means Clustering

The most straightforward method to cluster the data is grouping the data into related sets. *K*-means clustering aims to classify observations into *K* groups by partitioning observations into clusters in order to minimize the sum of squared distances from each observation to the mean of its assigned group. The *K*-means algorithm starts by randomly selecting a specified number of means and groups observations by assigning each one to the mean it is closest to in distance. The group means are then recalculated using the observations, replacing the previous means. The observations are reassigned to a group based on the distance between the value and the mean of the group. The algorithm iterates until the groups stabilize. The algorithm will converge to a local minimum, but not necessarily to a global minimum, therefore it is necessary to initialize and run the analysis many times.

Varying the number of groups (*K*) will result in different solutions to the *K*-means algorithm. The sum of squares in general decreases as *K* increases, since the larger the number of groups the more choices are available when assigning an observation to a group, hence a better fitting choice can typically be found. For this reason, selecting *K* with the smallest sum of squares over fits the data. In fact, when K is the number of observations or more, each observation will form a group by itself and the sum of squares will be 0, but this does not give any useful information about the data. A plot of the sum of squares versus values of *K* is useful for determining an optimal value of *K* (**Supplemental Fig. 2a**). *K* is often selected where the plot has an "elbow" (a steep drop at a specific *K* followed by a gradual decline towards zero). However, with metagenomic data, it was not unusual for the plot to appear rounded rather than have a single unique elbow (**Supplemental Fig. 2a**). Therefore, an alternative optimization using silhouettes[14] was used. Ideally, each observation is much closer to the mean of its group than to the mean of any other group. The silhouette of an observation is the difference between its distance from the closest of the *K* means and the second closest, divided by its distance from the second closest mean. In the best possible case, the observation is close to its own mean and not very close to the second best mean, i.e. its silhouette is close to 1. The set of all silhouettes (one for each observation) for *K* from 1 to 10 is shown in **Supplemental Fig. 2b**. For each value of *K* we calculate the average silhouette width, and use *K* that optimizes the width of the silhouettes. We found a maximum at *K=6*, with another smaller optimal width with *K=10* (**Supplemental Fig. 2c**).

The *K*-means algorithm was most useful for identifying outliers, which could be removed from future analysis as required. *K=6* groups, identified two broad categories, 1) the aquatic group cluster and 2) the human, terrestrial animal associated and mat community cluster (**Supplemental Table 1**). The remaining four groups were small and consisted of outliers. Therefore, the K-means showed broad patterns in the data.

Classification Trees

Decision trees can be used to group microbes that show similar metabolic functions. A supervised decision tree constructs a classification tree by identifying variables and decision rules that best distinguish between predefined classes (supervised). If the response variable is continuous, instead of predefined classes, a regression tree can be constructed which predicts the average value of the response variable. Since our data has predefined classes, we will only consider classification trees for the analysis. Trees are invariant under monotonic transformations of the predictor variables because constructing a tree uses binary partitions of the data and thus most variable scaling is unnecessary.

The construction of a supervised tree attempts to minimize the mixing of the different predefined classes within a leaf (called the node impurity). At each branching point, the algorithm chooses a single variable and a value of this variable which splits the node so that the impurity is minimized. There are several criteria for measuring the impurity of a node, such as the Gini index (described below in the section on random forests), the twoing criterion (splitting the classes into two groups and calculating the mixing of the groups), the deviance (the likelihood of each split), or the misclassification rate (based on a known classification)[15-18]. In general, trees are a balance between classification strength and model complexity with the overall goal of maximizing prediction strength and minimizing over-fitting. Often a large tree is grown that over fits the data, and pruning and cross validation are used to select the most appropriate subtree of that original tree[16].

To cross validate a tree, the data set is divided into *k* randomly selected groups of near equal size. A large tree is built using the data points in only *k−1* groups and pruned to give a sequence of subtrees. The tree and subtrees are used to predict the classes of the remaining data points, and these predictions are compared against the actual classes of those data points. The misclassification rate and the cross-validated deviance estimate are computed for each tree, and the process is repeated for each group. This *k*-fold cross-validation procedure[16] is typically repeated many times, so that different subsets are selected in each trial. The misclassification and deviance values for each tree size are averaged over the repetitions, and the subtree that minimized the standard error in the misclassification rate or the lowest average deviance is selected. Trees constructed using cross-validation tools are typically less susceptible to over fitting than other forms of classification. *K*-fold cross validation is particularly appropriate for metagenomic data where there may be few samples in each environmental group and as many samples as possible should be used to identify the right tree.

Unlike *K*-means clustering, decision tree classification provides information about the variables that drive the separation. The best classification tree using all the variables was determined by 500 runs of 10-fold cross-validation, which selected a nine-leafed tree. This classification tree (**Fig. 1**) demonstrated that phage proteins separated the host associated microbial communities and the majority of free-living communities. In particular, and as has been shown before[13,19], the host associated communities and some microbial communities from the fresh water and hypersaline environment characteristically had more phage proteins.

Harsh environments (such as hypersaline aquatic environments) had more co-factors, vitamins and pigments. Within the marine realm, the coastal and deep water samples had, as expected, fewer photosynthetic proteins than the open water samples, but the photosynthetic potential of the reefs was mixed, as seen before[2]. Photosynthetic potential also aided the identification of stratification in the mat microbial communities by depth, a separation that was supported by metabolism that occurs in microaerobic or anoxic conditions.

Random Forests

While decision trees are useful classification tools, they lack robustness: small changes in the data, such as adding one more sample, can yield dramatically different results. The random forests[20] technique generates a large ensemble of trees, by choosing a random subset of the original data with replacement (bootstrapping), and using a user-defined number of variables selected at random from all of the variables at each node splitting. The resulting ensemble of trees (the random forest) is then used with a majority voting rule to decide which metagenomes belong to which groups. The computation is not excessive: a random forest with one thousand trees trained on two hundred metagenome datasets was computed in a few seconds. The random forest is typically used to separate data into predefined groups (a *supervised* random forest). A subset of the data and variables is used to generate the trees, and thus the approach can predict the environment to which a metagenome belongs. The random forest does not produce branching rules like a single classification tree because the trees in the random forest all differ from one another.

Sampling the data with replacement generates a new dataset to grow each tree in the forest – a process called *bagging* (*bootstrap aggregating*). The metagenomes that are chosen at least once during the sampling process are considered *in-bag* for the resulting tree, while the remaining metagenomes are considered *out-of-bag* (OOB). Upon mature growth of the forest, each metagenome will be out-of-bag for a subset of the trees: that subset is used to predict the class of the metagenome. If the predicted class does not match the original given class, the OOB error is increased[20]. A low OOB error means the forest is a strong predictor of the environments that the metagenomes come from. Misclassifications contributing to the OOB errors are displayed in a *confusion matrix*. The rows in the confusion matrix represent the classes of the metagenomes and the columns represent the classes predicted by the subsets of the trees for which each metagenome was OOB. Each class error, weighted for class size, contributes to the single OOB error (**Supplemental Table 2**). The OOB error and a confusion matrix are used to judge the misclassification error and clarify where the errors occur, while the variable importance measure allows for identifying which variables are best at discriminating among groups.

In an unsupervised random forest, the metagenome data is classified without *a priori* class specifications. Synthetic classes are generated randomly and the trees are grown. Despite synthetic classes, similar metagenomes will end up in the same leaves of trees due to the tree branching process, and the *proximity* of two metagenomes is measured by the number of times they appear on the same leaf. The proximity is normalized so that a metagenome has proximity of one with itself and 1−proximity is a dissimilarity measure[21]. The strength of the clustering detected this way may be measured by a "partitioning around the medoids" (PAM) analysis[22]. Conceptually similar to the *K*-means clustering described above, PAM picks *K* metagenomes called medoids, then creates clusters around them by assigning each metagenome to whichever medoid it is closest to using the dissimilarity measure above as a kind of distance. The algorithm looks for whichever *K* metagenomes minimize the sum of the distances between all metagenomes and their assigned medoids. The result can be visualized using a multi-dimensional scaling (MDS; see below) plot.

For a supervised random forest, variable importance measures such as mean decrease in accuracy and mean decrease in the Gini coefficient also provide biological insight (**Supplemental Fig. 4**). These two values indicate which variables contributed the most to generating strong trees. These can then be used to generate single trees with branching rules or used in other visualization analysis such as canonical discriminant analyses (CDA; see below, where we use mean decrease in Gini in the CDA).

The mean decrease in accuracy that a variable causes is determined during the OOB error calculation phase. The values of a particular variable are randomly permuted among the set of out-of-bag metagenomes. Then the OOB error is computed again. The more the accuracy of the random forest decreases due to the permutation of values of this variable, the more important the variable is deemed[20]. The mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest[17]. Each time a particular variable is used to split a node, the Gini coefficients for the child nodes are calculated and compared to that of the original node. The Gini coefficient is a measure of homogeneity from 0 (homogeneous) to 1 (heterogeneous). The decreases in Gini are summed for each variable and normalized at the end of the calculation. Variables that split nodes into nodes with higher purity have a higher decrease in Gini coefficient.

Overall, the photosynthesis and phage groups were the most important in separating the data sets, and a break occurred between these two variables and the remaining variables, suggesting that just these two measures could be used to grossly classify the metagenomes (**Supplemental Fig. 4**). The next break appeared after the eighth variable in the mean decreasing accuracy plot. These eight variables were thus chosen for the CDA analysis described below. The misclassification rate of the random forest analysis was 31 % (**Supplemental Table 3**).

Multiple dimensional scaling

Multidimensional scaling directly scales objects based on the similarities or dissimilarities between them[23]. MDS tries to project the proximity measures of the metagenomes as determined by another technique, such as *K*-means, random forests, etc. to a lower-dimensional Euclidean space. For the random forests, the similarity was measured as the number of times two metagenomes appeared on the same leaf in the trees, and the distance between two samples on the plot represents that value. The MDS plots are colored either by the five PAM groupings (**Fig. 2a**), or the ten predefined environments (**Fig. 2b**). In this analysis, the microbes from human and animal hosts separated from the other samples along the first dimension while the aquatic and mat communities separated along the second dimension. MDS can be used as a data reduction process as well as a visualization tool, since the values for each sample calculated for the plots can be used as variables in subsequent analyses.

Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a supervised statistical technique that aims to separate the data into groups based on hyperplanes and describe the differences between groups by a linear classification criterion that identifies decision boundaries between groups. The advantages of LDA are the ability to visualize the data and obtain a statistically robust analysis of the classification ability, while the disadvantages include the requirements for (i) normal distribution of the data, (ii) near even size of the groups, and (iii) a linear relationship between the datasets, all of which are unlikely with metagenomics data.

The LDA over all 27 functional group variables separated the data (**Supplemental Fig. 4**), and showed that the human and terrestrial animal associated metagenomes separated from a cluster consisting of all of the aquatic samples except the hypersaline community. The mat samples separated distinctly from the other clusters. Leave one out cross-validation can be used to judge how well an LDA acts as a classifier for new data[24]. The LDA analysis performed worst of all the classification techniques, with an error rate of 0.36 (i.e. 36 % of the samples were misclassified). Most of the misclassified samples were from the marine environment that fell among the large aquatic cluster.

Principal Component Analysis

Principal component analysis (PCA) is a statistical technique that reduces the number of variables that account for most of the variance of the data. PCA selects linear combinations of the original variables sorted so that each accounts for as much of the sample variance as possible while being orthogonal to the previous ones. Such linear combinations of the variables are called the principal components. Obviously there are exactly as many principal components as original variables; 27 in our case. The goal of PCA is to explain as much of the variance as possible in the first few components. A plot with 27 variables is difficult to interpret, so we used the variables with the top eight largest variances for the analysis. There was a natural separation in the magnitude of the variances after eight variables. **Figure 3** shows a PCA plot with respect to the first two principal components that depicts each metagenome colored by its environment.

The positioning of the data on the plane is strongly influenced by the number of sequences associated with DNA metabolism, cell division, and amino acid metabolism in one direction, and virulence and RNA metabolism in the other, with cofactor metabolism important in both directions. The metagenomes did not separate particularly well with this analysis, however human and terrestrial animal associated samples clustered above aquatic samples. The first two dimensions of the PCA did not provide good resolution of the nuances within an environment, explaining only 38 % of the variance (compared to 91 % of the variance for the CDA, below). The inability of the first two dimensions of the PCA to explain the variance suggests that the variance in our data was inherently multi-dimensional. A subset of the variable genes could be identified and analyzed by a PCA to increase resolution of the analysis; the random forests and CDA analysis provide non-biased methods of identifying those features.

Canonical Discriminant Analysis

Canonical Discriminant Analysis (CDA) is a dimension reduction statistical tool, similar to principal component analysis (PCA) and linear discriminant analysis (LDA). The most important aspect of the CDA is that, unlike PCA and LDA, it is used to separate data into preassigned categories, in this case the preassigned environments. Specifically, the algorithm finds one fewer axes than classes of data. The axes are uncorrelated linear functions that best separate the data classes. Like a supervised random forest, the goal of the CDA is to understand which variables are responsible for differentiating between the groups.

The CDA identifies variation between classes: the first canonical component is the linear combination of variables that has maximum multiple correlation with the classes[25]. The second canonical component is obtained by finding the linear combination uncorrelated with the first canonical variable that has the highest possible multiple correlation with the classes. The process is repeated until the maximum number of canonical components is obtained. Ideally, only the first two or three canonical components are needed to adequately separate distinct groupings. A fundamental difference between PCA and CDA is in the covariance matrix: in the former the covariance matrix displays the variance between individual samples, while in the latter it displays the variance between classes. The covariance matrix has to be full rank, which requires at least as many individual samples as variables. This analysis used 212 samples and 27 variables. First, the unsupervised random forest (above) was used to identify the eight most important variables. Then CDA was used to identify the best linear combinations of these variables. To visualize the CDA the canonical scores of the data points were plotted along new axes (the canonical components), vectors represent the influence of each variable, and group centroids aid the visualization (**Fig. 4**).

CDA is efficient at separating preassigned classes, but to measure the accuracy of the assignments, a misclassification error was computed. There were no available error estimation functions for canonical discriminant analysis in R, so an analysis was devised operating on the leave-one-out principal like those described above: for a data set with *n* samples, the function completes *n* canonical discriminant analyses, each time leaving out a different sample. The canonical scores are computed for the left out sample by comparing the group to which it is closest with its preassigned class. That comparison is then used to compute a scale-invariant distance (the minimal Mahalanobis distance[26]) for each group. A slightly different approach based on bootstrapping and combining CDA with LDA was also developed and reported similar results. Code for both functions is provided in the Supplemental Online Material.

PCA dimension reduction uses a small number of orthogonal linear combinations of the variables to explain the variance of the data. CDA explained a large amount of the variance (91 %) compared with 38 % in the PCA, showing the importance of a key set of metabolic process occurs in each environment. However, CDA has two main drawbacks: (i) the metagenomes are placed into predefined groups and thus are subject to observer bias; and (ii) the canonical components are linear combinations that best separate the groups, so CDA is prone to over fitting.

The CDA conducted on the data using the eight most important functional subsystems identified in the random forest analysis (**Fig. 4**) showed that the host associated microbial communities were separated from the other environments by the abundance of the phage and dormancy sequences. The length of the lines in the plot are proportional to the importance of that variables in separating the data. The harsh hydrothermal springs were again associated with the need for co-factors while the photosynthetic potential of the microbes provided separation between the coastal and open water metagenomes and reflects the reduced amount of primary production conducted by the microbes in the coastal areas and increased amount of photosynthesis conducted by the microbes in the open water regions of the ocean. Membrane transport, protein and nitrogen metabolism were also important in separating the aquatic and host-associated metagenomes, but to a lesser extent.

The misclassification rate of the CDA was 39.7 % when used with the eight most important variables identified by the random forest, but the misclassification rate increased to 45 % when the eight variables with the largest variance were used (as with the PCA).

Discussion

Metagenomic data provides a wealth of information about the functional potential of microbial communities, but the vastness of the data makes it difficult to discern the patterns that are important discriminators. A range of clustering and classification techniques were applied to metagenomic data to analyze the data, and the multiple analyses conducted on the data demonstrated the stability of the metabolic profiles in describing the difference between environments. The results show that a mixture of methods provides the most effective analysis of the data: *K*-means was used to identify outliers, random forests to identify the most important variables, and either a classification tree or CDA to test the relevance of the environment to genomic content.

Each of the analyses, except the PCA separated the microbial samples into three broad groups (based on the biomes from where they were isolated): the human and animal associated samples, the microbial mats, and the aquatic samples. The LDA provided little additional separation, but the combination of random forests and CDA demonstrated that phage activity is a major separator of host associated microbial communities and free-living or environmental microbial communities, suggesting that the phages are playing different ecological roles within each environment. In free-living microbial communities, phages are major predators and generally show similar diversity to their hosts. In host associated microbial communities, phages are more diverse suggesting that they may provide specific genes to increase host survival[13]. These techniques also showed that the mat communities separated from both the animal associated metagenomes and the aquatic samples because of the vitamin and co-factor metabolism, suggesting a role for secondary metabolism associated with growth in this extreme environment. The dominant metabolism that separated the aquatic samples was photosynthesis: not surprisingly, samples from deep in the ocean, and some of the impacted reef sites, do not have much photosynthesis, while photosynthesis abounds at unaffected reefs and surface waters of the open ocean. Although only the one or two most abundant phenotypes in each sample were described here, the statistical analyses also reveal less obvious separations among the data, and unraveling the role of microbes in the global geobiology is an important goal for post-metagenomic studies.

Any analysis of biological data must be informed by the underlying biological system(s) being analyzed, however the goal of statistical analysis is to highlight differences between samples whether or not we (currently) know how to explain them. Those differences that are statistically significant, yet lack a thorough or convincing biological explanation are to be embraced as new avenues for research and those statistical techniques should not be merely dismissed as poor approaches[27].

We hope that the statistical tools described here will help microbial ecologists understand their data in more detail, and help them parse out the important and interesting nuances that separate different environmental samples.

Acknowledgements

This work was supported by the NSF REU grant from the NSF Division of Mathematical Sciences (NSF DMS-0647384) to VP (Research Experience for Undergraduates Student Participants: NA, MC, EG, MH, KI, CP, TR, REU Instructors: IT, BB, RAE, ED). Students conducted all of the analysis and wrote scripts for R. EAD was supported by a grant from the NSF Transforming Undergraduate Education and Science (NSF TUES-104453). RAE was supported by the PhAnToMe grant from the NSF Division of Biological Infrastructure (NSF DBI-0850356).

References

1. Wegley, L., Edwards, R., Rodriguez-Brito, B., Liu, H. & Rohwer, F. Metagenomic analysis of the microbial community associated with the coral Porites astreoides. *Environmental microbiology* **9**, 2707–2719 (2007).
2. Dinsdale, E.A. et al. Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS One* **3**, 1584 (2008).
3. Brulc, J.M. et al. Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proc Natl Acad Sci U S A* **106**, 1948-1953 (2009).
4. Angly, F.E. et al. The marine viromes of four oceanic regions. *PLoS Biol* **4**, e368 (2006).
5. Breitbart, M. et al. Metagenomic and stable isotopic analyses of modern freshwater microbialites in Cuatro Cienegas, Mexico. *Environmental microbiology* (2008).at <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18764874>
6. Tringe, S.G. et al. Comparative metagenomics of microbial communities. *Science* **308**, 554-7 (2005).
7. Kurokawa, K. et al. Comparative Metagenomics Revealed Commonly Enriched Gene Sets in Human Gut Microbiomes. *DNA Res* **14**, 169-181 (2007).
8. Turnbaugh, P.J. et al. The human microbiome project. *Nature* **449**, 804-10 (2007).
9. Turnbaugh, P.J. et al. A core gut microbiome in obese and lean twins. *Nature* **457**, 480-484 (2009).
10. Dinsdale, E.A. et al. Functional metagenomic profiling of nine biomes. *Nature* **452**, 629–632 (2008).
11. Overbeek, R. et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* **33**, 5691-702 (2005).
12. Willner, D., Thurber, R.V. & Rohwer, F. Metagenomic signatures of 86 microbial and viral metagenomes. *Environ. Microbiol* **11**, 1752-1766 (2009).
13. Reyes, A. et al. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334-338 (2010).
14. Rousseeuw, P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20**, 53–65 (1987).
15. Breiman, L. Technical note: Some properties of splitting criteria. *Mach Learn* **24**, 41-47 (1996).
16. Breiman, L., Friedman, J., Stone, C.J. & Olshen, R.A. *Classification and Regression Trees*. (Chapman and Hall/CRC: 1984).
17. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. (Springer: 2009).
18. De'ath, G. & Fabricius, K.E. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* **81**, 3178–3192 (2000).
19. Oliver, K.M., Degnan, P.H., Hunter, M.S. & Moran, N.A. Bacteriophages encode factors required for protection in a symbiotic mutualism. *Science* **325**, 992 (2009).
20. Breiman, L. Random forests. *Machine learning* **45**, 5–32 (2001).

21. Shi, T. & Horvath, S. Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics* **15**, 118–138 (2006).
22. Marden, J.I. Multivariate Statistical Analysis Old School.
23. Quinn, G.P. & Keough, M.J. *Experimental Design and Data Analysis for Biologists*. (Cambridge University Press: 2002).
24. Venables, W.N. & Ripley, B.D. *Modern Applied Statistics with S*. (Springer: 2002).
25. Campbell, N.A. & Atchley, W.R. The geometry of canonical variate analysis. *Systematic Zoology* 268–280 (1981).
26. De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D.L. The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems* **50**, 1–18 (2000).
27. Kuczynski, J. et al. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat. Methods* **7**, 813-819 (2010).

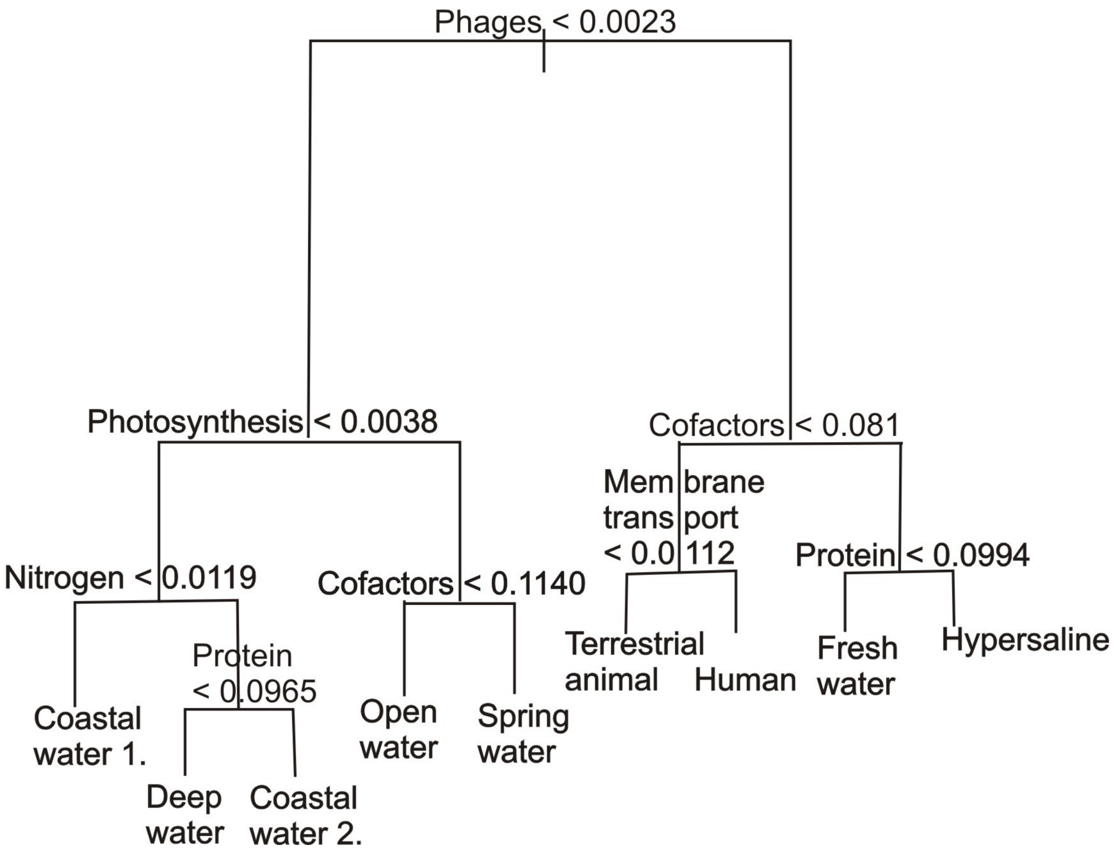
Figure Captions

Fig. 1. A classification tree of the separation of metagenomes from different environments based on the abundance of the subsystems in each environment. The abundances are normalized as described in the supplemental methods. The tree has been pruned to only show the eight most important variables.

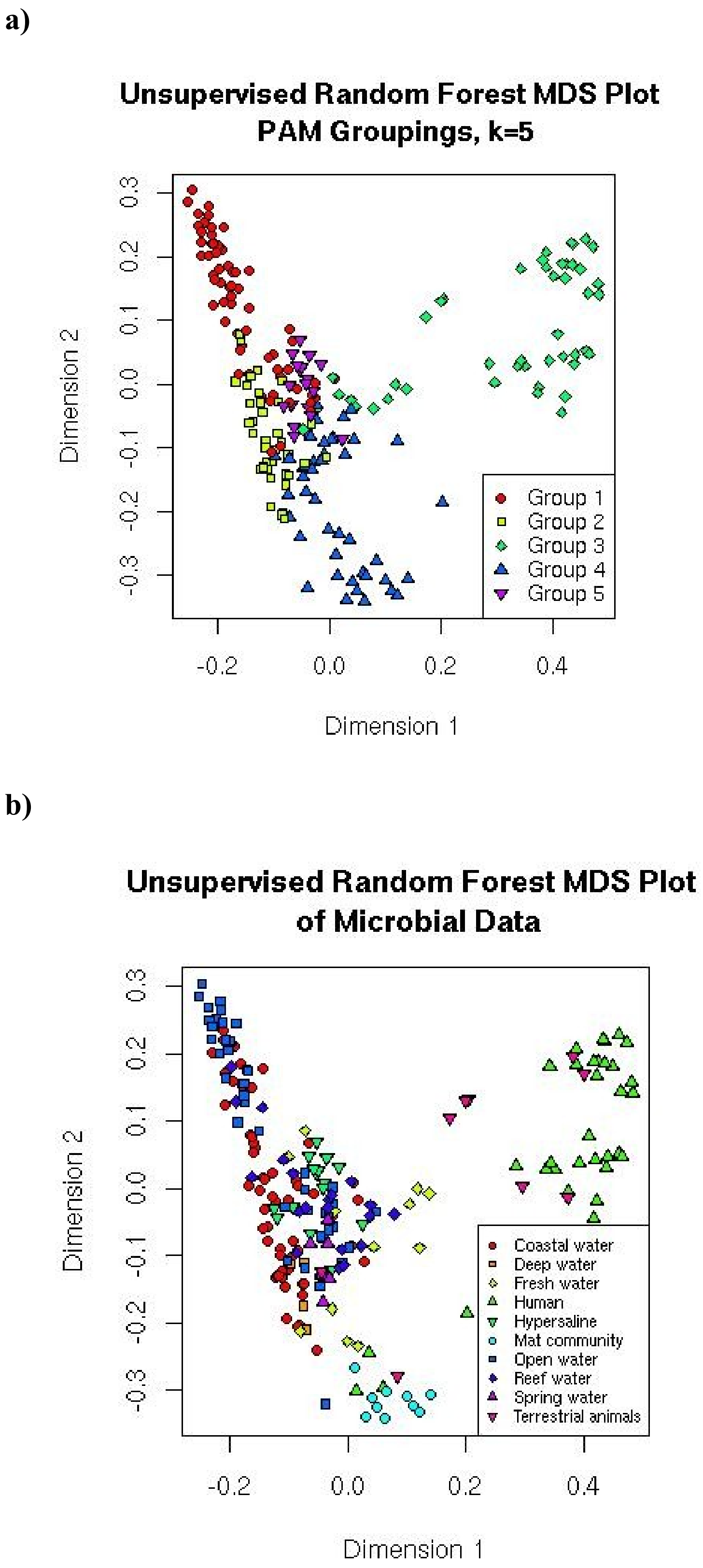
Fig. 2. Multiple dimensional scale plot of the distances calculated from the unsupervised random forest. The distances are the number of times the samples appear on the same leaf of the tree, and the MDS has scaled them so that they plot projects those distances into two dimensions. Colored by (a) the five PAM groupings suggested by the random forest (see text); or (b) the original environments the samples came from.

Fig. 3. Principal component analysis of the 212 metagenomes using the top eight variables identified from the random forest analysis. The samples are colored and shaped by the environment where they came from. The samples are largely aligned on a 45° plane from virulence-DNA metabolism to amino acids-cofactors.

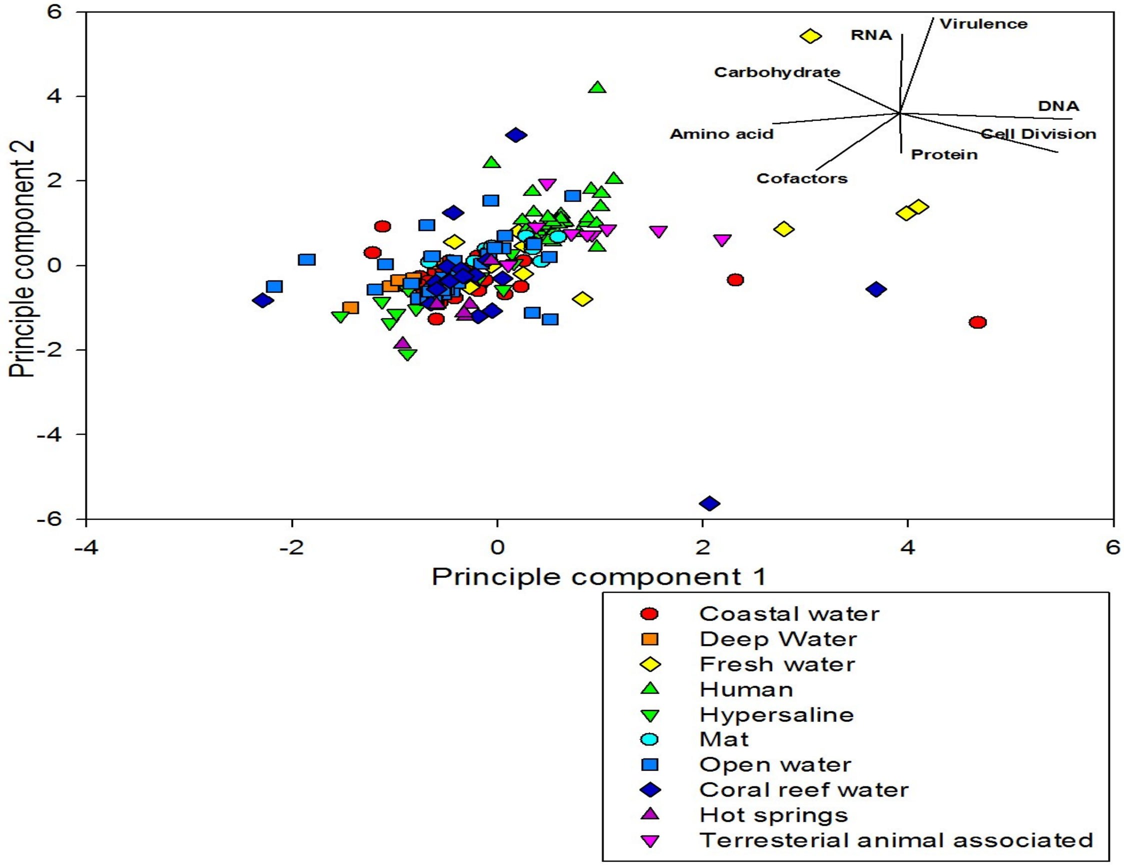
Fig. 4. Canonical discriminant analysis of the 212 metagenomes using the top eight variables identified from the random forest analysis. The plot shows the separation in the host associated microbial communities and the free living communities. The analysis explained 91 % of the variance, suggesting that metagenomes can be discriminated by the metabolic potential. Line depict the h-plot of important metabolic processes and the points are the centroid or mean for the 10 environments.



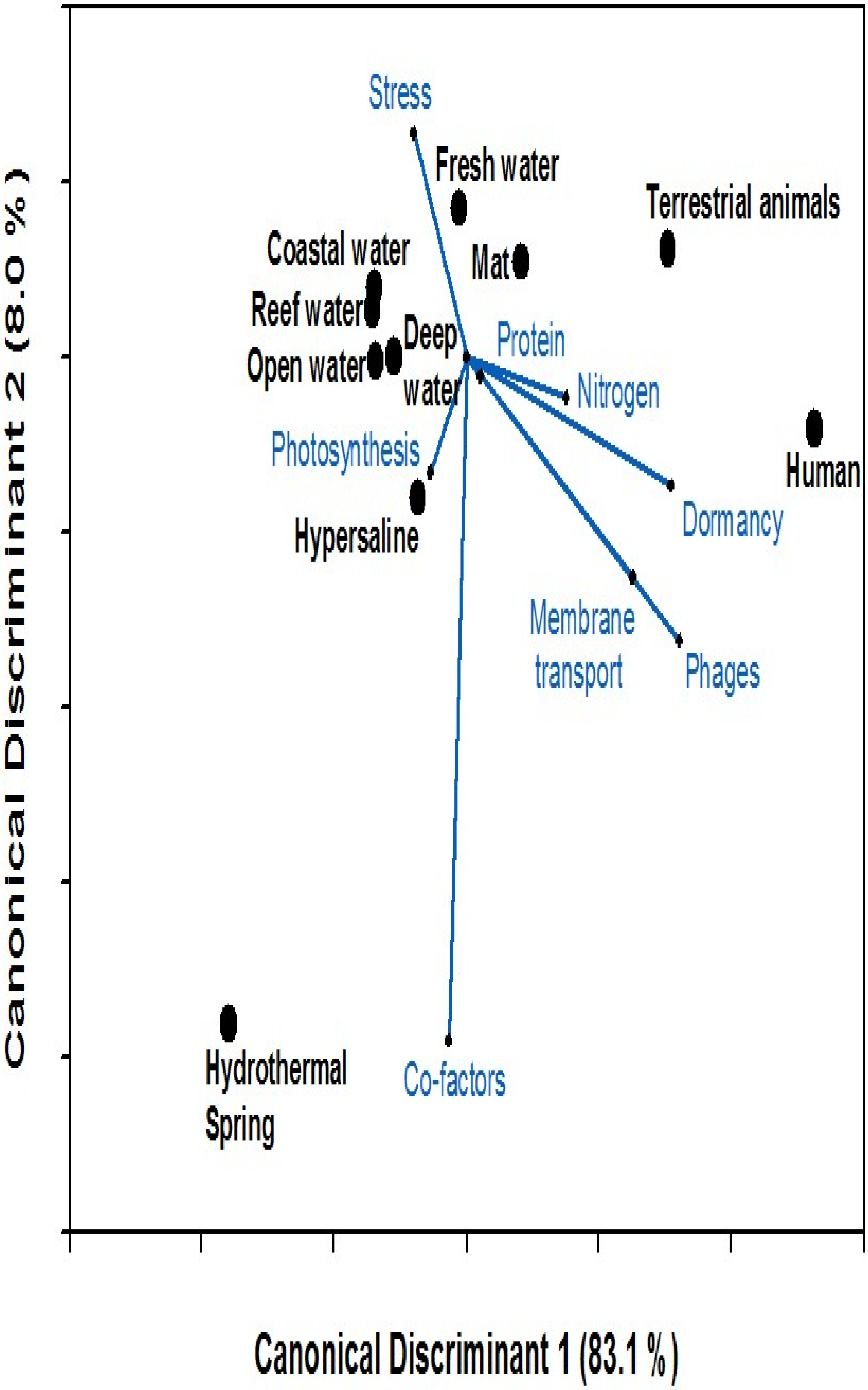
Dinsdale et al. Fig. 1



Dinsdale et al. Fig. 2



Dinsdale et al., Fig. 3


Stress
Fresh water
Terrestrial animals
Coastal water
Mat
Reef water
Deep water
Protein
Open water
Nitrogen
Photosynthesis
Human
Hypersaline
Dormancy
Membrane transport
Phages
Hydrothermal Spring
Co-factors
Canonical Discriminant 2 (8.0 %)
Canonical Discriminant 1 (83.1 %)

Dinsdale et al. Fig. 4

Multivariate analysis of functional metagenomes

*Elizabeth A. Dinsdale[1], Robert A. Edwards[1,2,3], Barbara Bailey[4], Imre Tuba[5], Sajia Akhter[6], Katelyn McNair[6], Robert Schmieder[6], Naneh Apkarian[7], Michelle Creek[8], Eric Guan[9], Mayra Hernandez[4], Catherine Isaacs[10], Chris Peterson[7], Todd Regh[11], Vadim Ponomarenko[4]

Supplemental Online Material

Detailed Methods

Metagenomes

Publicly available metagenomes were selected from the Edwards Lab metagenome database (http://edwards.sdsu.edu/mymgdb/) {Schmieder, Edwards, Unpublished}. All samples were annotated using the real-time K-mer based annotation system using a 10-amino acid word size and a requirement for at least two words per protein (http://edwards.sdsu.edu/rtmg). This approach, described elsewhere, {ref: Edwards, Overbeek, Disz, Olson} uses signature K-mers to identify the functions encoded in the metagenome sample. The K-mer based approach allows all of the samples to be annotated against the same core database, and for the annotations to be updated whenever required. The K-mer based annotation provides the number of sequences for each function, subsystem, and two level hierarchy in the subsystems ontology { PMID: 21421023 }. Counts were normalized by the total number of hits to account for the different sample sizes of each metagenomes and to yield percent composition by function. The functional hierarchies *clustering-based subsystems* and *experimental subsystems* were removed from the data, leaving 27 first level functional hierarchies or functional families. The metagenomes were classified as belonging to ten different environments: hypersaline; mat community (from Solar Salterns); hydrothermal springs; human associated; other terrestrial animal associated; freshwater; and marine. Because of the abundance of marine samples (for example, because of the Global Ocean Survey data), these samples were further sub-divided into four groups: open ocean, coastal water, deep water, and coral-reef associated samples.

Supplemental Online Figures and Tables

Supplemental Table 1. The samples present in each of the clustered identified by the *K*-means analysis with *K* of six. This was chosen because the silhouette analysis suggested that six clusters were the most appropriate (**Supplemental Fig. 2**). There were 33 human, 9 terrestrial animal, 10 mat community, 42 open water, 20 reef water, 60 coastal water, 5 deep water, 7 fresh water, 15 hypersaline, 6 hot spring samples in total.

Cluster	Number of metagenomes	Original metagenome classification
1	52	31 human 5 terrestrial animals 6 mat community Water samples: • 4 open • 3 reef • 2 coastal • 1 fresh
2	1	1 reef water sample
3	1	1 reef water sample
4	3	1 human 1 fresh water 1 reef water
5	149	4 mat 4 terrestrial animals 1 human Water samples: • 56 coastal • 5 deep • 15 hypersaline • 6 spring • 38 open • 13 reef • 7 fresh
6	6	Water samples: • 2 coastal • 3 fresh • 1 reef

Supplementary Table 2: Tree size and average deviance from a series of tree cross-validation experiments.

Tree Size	Average CV Deviance
1	152.014
2	122.432
3	102.636
4	99.642
6	92.762
8	92.970
9	92.812
14	95.848
16	98.342
17	98.622

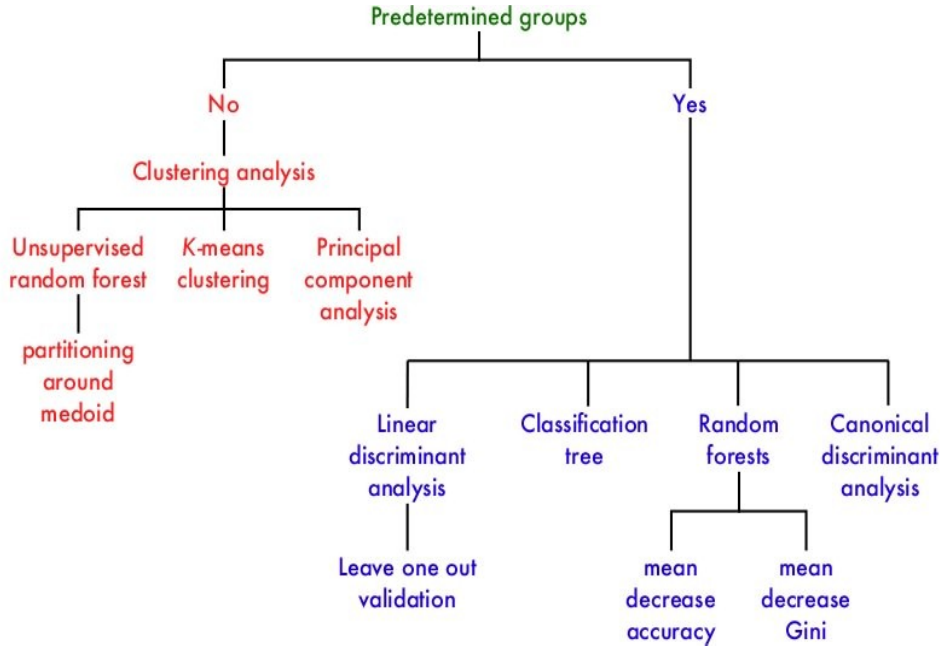
Supplemttaty Table 3. The group that each metagenome was assigned to by the random forest analysis.

Initial classification	Classification from the random forest								
	Mixed marine	Deep water	Coastal marine	Open marine	Spring water	Terrestrial animals	Human associated	Fresh water	Hyper-saline
Freshwater	3				1	1			
Open marine	6	1	1	31					2
Spring water	1				5				
Coastal marine	6	1	43	8	2				
Terrestrial animal						5 cow 2 mice	3 mice 1 fish		
Human associated	1		1				32		
Mat community	4	1						4	
Deep marine		4	1						
Reef water	4	1		15					
hypersaline	4				1				9
Total	29	8	47	44	8	8	36	10	11

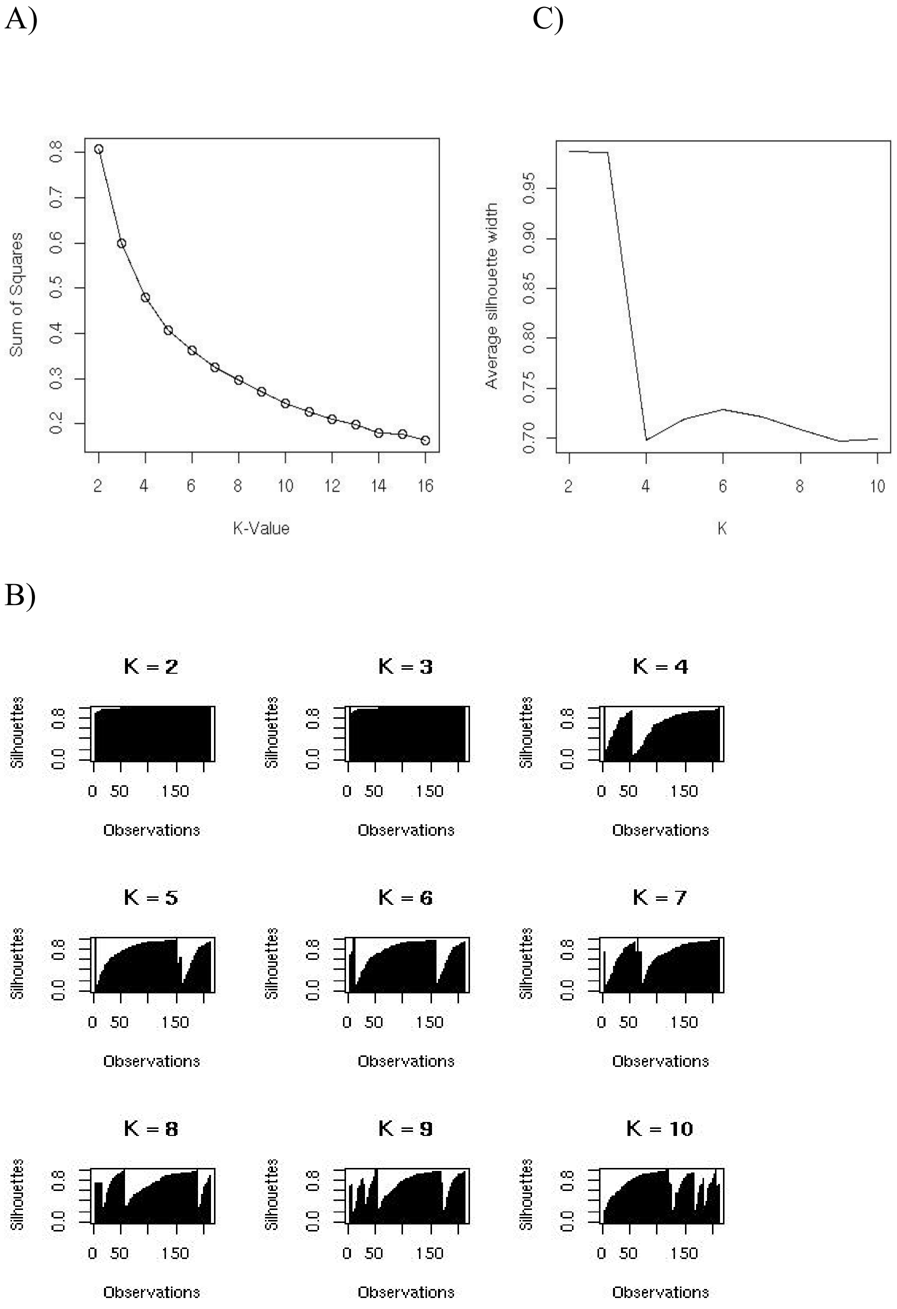
Supplementary Table 4. The missclassification table generated by the canonical discrimant analysis.

	coastal	deep	fresh	human	hypersaline	mat	open	reef	spring	Terrestrial animal	Class error
coastal	9.820	0.000	0.301	0.391	0.000	0.226	0.962	0.009	0.127	0.160	0.181
deep	0.990	0.004	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.995
fresh	0.816	0.000	0.433	0.231	0.000	0.235	0.081	0.028	0.160	0.075	0.783
human	0.000	0.000	0.207	6.268	0.000	0.457	0.014	0.051	0.000	0.000	0.104
hypersaline	1.231	0.000	0.000	0.000	1.485	0.000	0.283	0.000	0.000	0.000	0.504
mat community	0.382	0.000	0.000	0.004	0.000	1.613	0.000	0.000	0.000	0.000	0.193
open	4.377	0.009	0.033	0.448	0.169	0.349	2.410	0.169	0.014	0.018	0.698
reef	1.509	0.009	0.283	0.429	0.000	0.226	1.117	0.235	0.023	0.377	0.994
spring	0.047	0.000	0.000	0.000	0.000	0.000	0.113	0.004	0.834	0.000	0.165
terrestrial	0.287	0.000	0.108	1.193	0.000	0.216	0.000	0.000	0.000	0.193	0.903

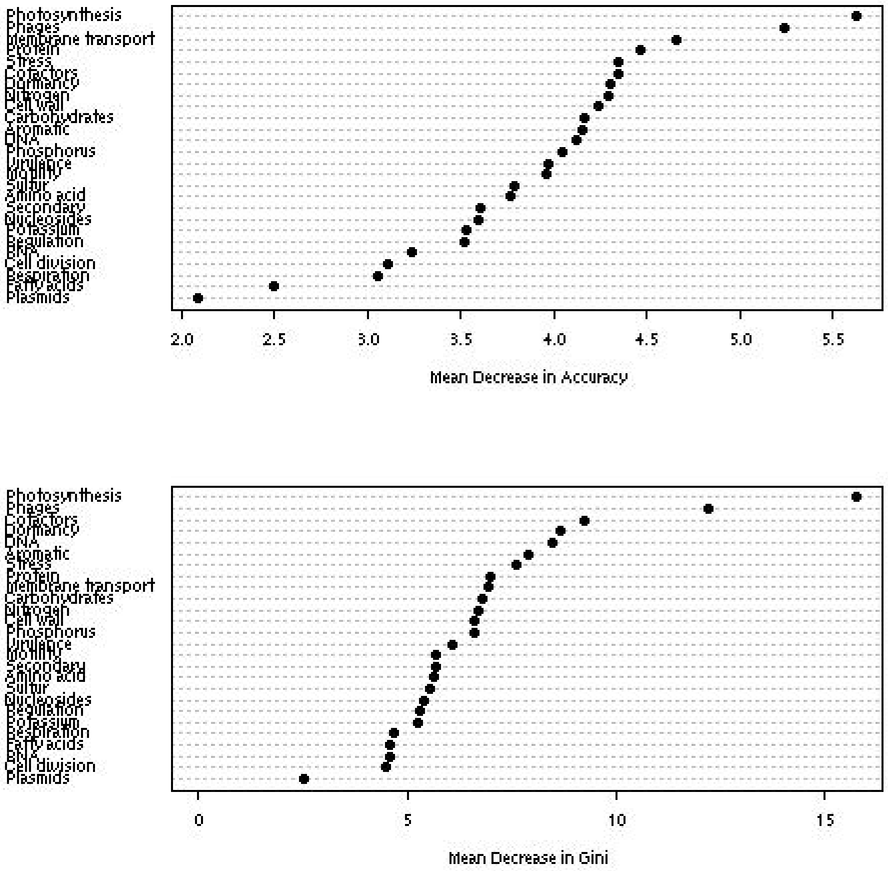
Supplemental Figure 1. A diagram of the relationship between the seven statistical methods evaluated.



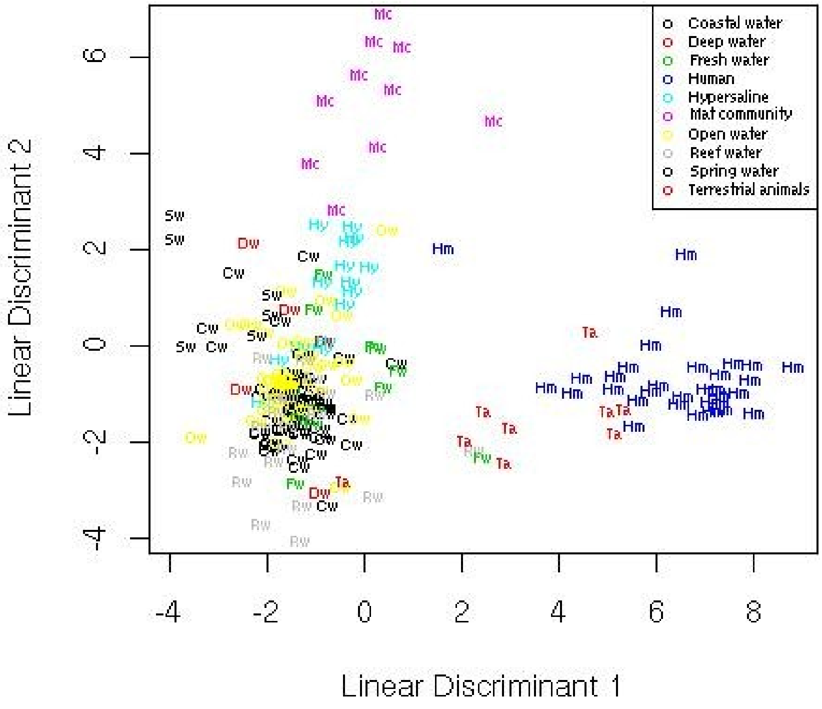
Supplemental Fig. 2. **(a)** The sums of squares and *K*-value used to identify the number of groups that the samples should be split into. No clear elbow was evident, therefore silhouette plots were use to examine the data. **(b)** A silhouette plot showing how it creates metagenomic groups in the data. The most favorable grouping number is where the average silhouette width is nearest to one. **(c)** The variation of average silhouette width and *K*. There is a peak at *K*=6 and an uptick at *K*=10.



Supplemental Fig. 3 Mean decrease in (a) accuracy and (b) Gini determined by the random forest analysis for the variables.



Supplemental Fig. 4. Linear discriminant analysis of the environmental samples.


Linear Discriminant 1
Linear Discriminant 2
○ Coastal water
○ Deep water
○ Fresh water
○ Human
○ Hypersaline
○ Mat community
○ Open water
○ Reef water
○ Spring water
○ Terrestrial animals