Methods in Description and Validation of Local Metagenetic Microbial Communities

Authors:

David Molik, Integrated Biomedical Sciences, University of Notre Dame, dmolik@nd.edu

Michael Pfrender, Biology, University of Notre Dame, mpfrende@nd.edu

Scott Emrich, Computer Science and Engineering, University of Notre Dame, semrich@nd.edu

Abstract:

1. We propose minhash (as implemented by MASH) and NMF as alternative methods to estimate similarity between metagenetic samples. We further describe these results with cluster analysis and correlations with independent ecological metadata.
2. Species and kmer abundance information is used to determine similarities and create clusters to better understand how communities interact, as well as relate to known environmental variables, such as Ph and Soil Conductivity.
3. We use cluster silhouettes to assess various approaches for clustering metagenetic samples as well as anova to uncover links between metagenetic samples and the known environmental variables.
4. By analyzing data from the Atacama desert and determining the relationship between ecological factors and group membership, we show the applicability of these methods.

Introduction

How microbiome communities, and in a broader context local communities, are determined, described, and validated is a matter of some debate (Holyoak et al. 2005). While Principal Components Analysis (PCA) is the most common computational approach with the most divisive components considered the most ecologically relevant, PCA is biased towards components that have the most variance (Parsons et al. 2009), and is not necessarily useful for factor analysis, or determining underlying variables where many observed variables may reflect a few unobserved variables (Jolliffe 1986). A common description of community structure beyond the initial group assignment is also often lacking in PCA. Here we present two alternative computational methods to determine the grouping of metagenetic samples: k-mer minhash sketching and Non-Negative Matrix Factorization (NMF). NMF can be paired with k-means to estimate the number of groups present and has the benefit of determining the most important feature driving inferred relationships. Minhash sketches can be used to quickly estimate similarities between whole samples in an alignment-free approach, i.e., OTUs do not need to be generated first. While minhash and NMF are used here to cluster metagenetic samples based on inferred relationships, note that NMF focuses on what is distinct (in a cluster) while a minhash implementation (Ondov et al. 2016) is combined with hierarchical methods to infer clusters based on pairwise similarities. For this reason we use Silhouette plots for cluster assessment, which is a measure of how close each point in a cluster is to other clusters. Finally, we use ANOVA to determine the relationships of known environmental factors to the inferred clusters using our new and existing approaches.

One local community can be delineated from another by inferred differences in the species detected within one or more samples (Rusch et al. 2007);(Seshadri et al. 2007). K-mer estimation methods such as MASH or OTU methods such as PCA or NMF should be able to distinguish local communities based on the inference of species. Using these methods, however, does not necessarily provide distinctions between environments or even interrelated communities. Note that the detected abundance of a species in a metagenetic sample may not correlate with its actual abundance in the broader area, making drawing boundaries between local communities using any computational approach difficult. As a result, prior work in community analysis has often relied on additional metadata such as physical barriers and environmental measurements to refine the structure of estimated local communities based on the species observed (Holyoak et al. 2005).

Here, we consider novel, data-driven (unsupervised) approaches for defining communities based on clusters, or different inter-related groups, inferred only from NGS sequence data. These approaches allow us to artificially induce computational cutoffs and, as a result, no prior knowledge/metadata are required to infer potential relationships between samples. Because environmental characteristics can change the viability of a microbial species occupying that area (Hultman et al. 2015);(Gibbons & Gilbert 2015), subsequent comparisons of groupings to independent environmental variables provides a biologically motivated assessment of whether these computationally generated results uncover local communities. To define clusters we introduce MinHash (Ondov et al. 2016) based similarity for determining local community structure, which essentially an approximatation of the Jaccard similarity based on shared speces within samples (see Rusch et al. 2007 and Ondov et al. 2016 for details). We also apply Non-Negative Matrix Factorization (NMF) (Gaujoux & Seoighe, 2010);(Seung & Lee 1999);(Paatero & Tapper 1994) using the nsNMF algorithm (Pascual-Montano et al. 2006) to determine non-shared species based on OTU abundances. Log likelihood statistical analysis of an Atacama desert microbial community indicates that among these *de novo* methods, hierarchical clustering using MinHash similarities has more explicative power than NMF on OTU abundance. This data set is a good choice for this analysis because of the data's wide geographic range and inclusion of environmental variables. For the analysis of the Atacama desert, samples taken from the same sampling location (North/Central/South) were more similar according to alpha diversity (Crits-Christoph et al. 2013); however, we show that other environmental variables can have a statistically higher correlation than sampling location, and specifically that PH, air relative humidity (RH) and soil conductivity best explain observed local communies derived computationally. Combined, these indicate data-driven methods can be directly used to estimate community structure from NGS data.

NMF, MASH, Silhouettes, and ANOVA

Here, NMF (Berman & Plemmons, 1994) and Mash (Ondov et al. 2016) are integral to defining local community structure. NMF—or Non-Negative Matrix Factorization—is method by which to split a matrix into a components, based on the factors that are most important in making that split. For example, for RNA-seq expression analysis, suppose there are ‘k’ known clusters. NMF will break a provided expression matrix (genes by cells or cell tissues) into *k* total clusters while also producing the most important genes for doing so (Yu-Jui, 2016). When applied to observed OTU abundances, NMF will ideally return the most important OTUs to generate a fixed number of community-driven clusters. The power in this method is that different factors may be indicators for each cluster, instead of just the presence or absence of a particular expressed gene or observed species. NMF becomes particularly powerful when paired with k-means (Hartigan & Wong, 1979);(Forgey, 1965), which is a clustering method that can be used to measure how many clusters exist (aka, the ‘fit’). Because NMF combines factor discovery with iterative determination of the total number of clusters, NMF can be a more descriptive alternative to simple PCA-based visualization of the data.

Mash, which is based on MinHash sketching (Broder, 1997), is an alignment-free method by which to estimate the distance between two sequences or sets of sequences. Using this computational method a set of samples can be sequenced and then quickly compared to estimate how similar they are. The resulting pairwise similarity matrix can then be clustered hierarchically, which can be visualized in the form of dendrograms and/or heatmaps. Mash can be run on raw samples if desired at the cost of potentially higher inferred distances. Example hierarchical clustering algorithms are Diana (Struyf et al. 1997);(Kaufman & Rousseeuw, 2009) and McQuitty-WPGMA (McQuitty, 1966).

Using silhouettes (Rousseeuw, 1987);(Handl et al. 2005) and the clustering information derived from NMF we can further describe structure within a cluster. Specifically, silhouette width highlights the ‘belongingness’ of each data point within a cluster; higher averages indicate cluster points are more tightly correlated with each other.

Finally we hypothesize that the local assortment of species is largely determined by the environment in which they live. If so, a change in environment and a corresponding change in observed species should, for the most part, correlate and this correspondence can be tested using both Anova and a mantel test under the right conditions (DeLong, 2013). We also realize that environment itself can correlate with distance, i.e., in the northern hemisphere, northern samples have fewer growing degree days than southern samples. For this reason isolation by distance (IBD) could also manifest as distinct clusters using our computational alternatives just as they would in a traditional PCA analysis.

Results from Atacama Data

Sample clustering based on OTUs was performed using Non-negative matrix factorization (NMF), which determines OTUs that are most informative using linear algebra-based techniques (Ondov et al. 2016);(Seung & Lee 1999);(Paatero & Tapper 1994);(Yu-Jui, 2016). Sample to sample distances were determined based on minhash sketches, which estimate the Jaccard similarity of two samples based on shared subsequences (k-mers). We also determined the OTUs present in these Atacama samples using mothur (see Methods). Given our focus on unsupervised analysis, we processed the mash-based sample distances with multiple clustering methods: K-means (Hartigan & Wong, 1979);(Forgey, 1965), hierarchical (Everitt, 1974);(Hartigan, 1975): Agglomerative and Divisive (Kaufman & Rousseeuw, 2009).

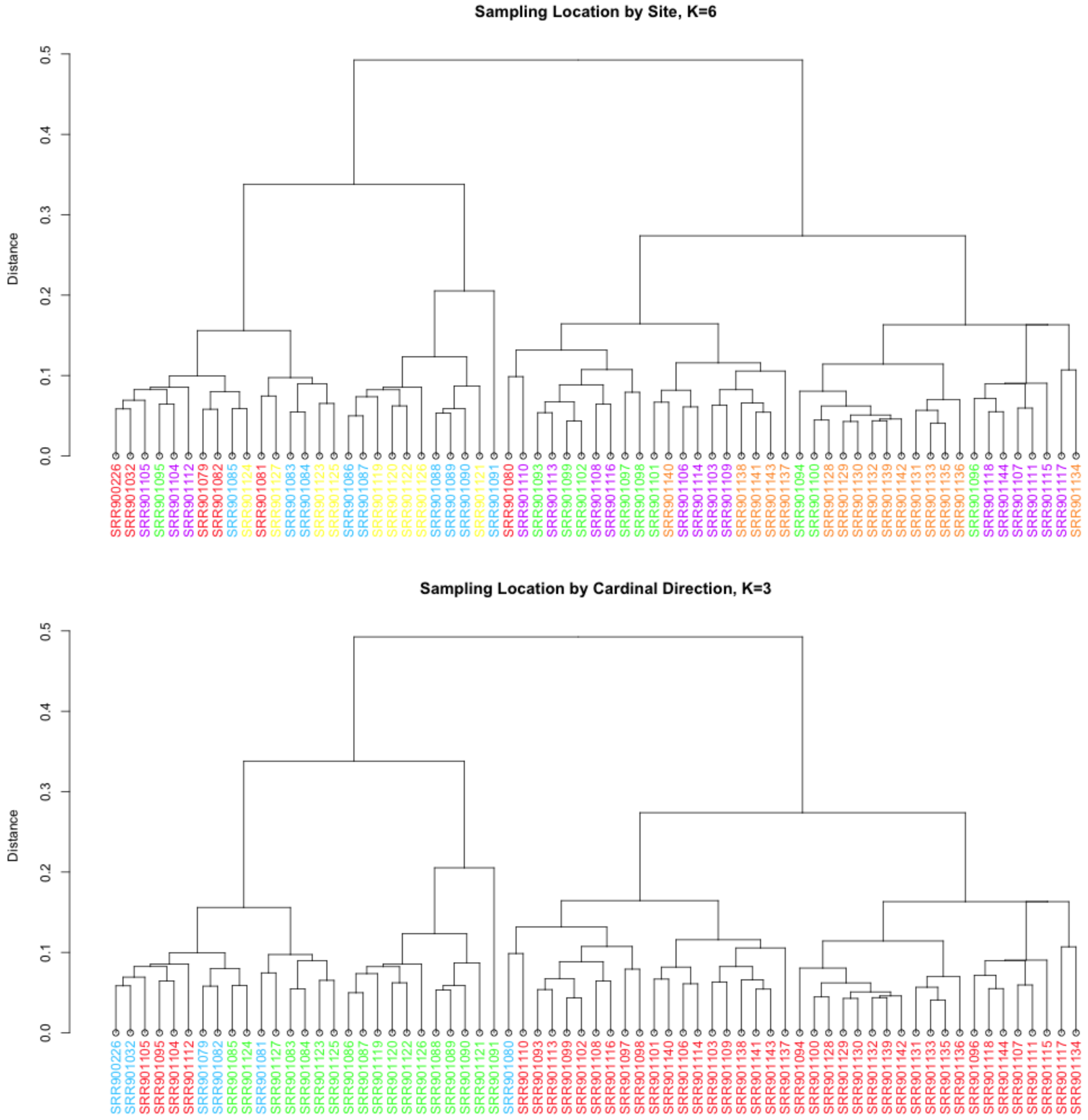


Figure 1. Sample clusterings of Crits-Christoph et al. (2013) data using two measures of distance: site location (top) and cardinal direction (bottom). Dendrograms were generated with Diana and are colored by sampling location (top, 6 total).

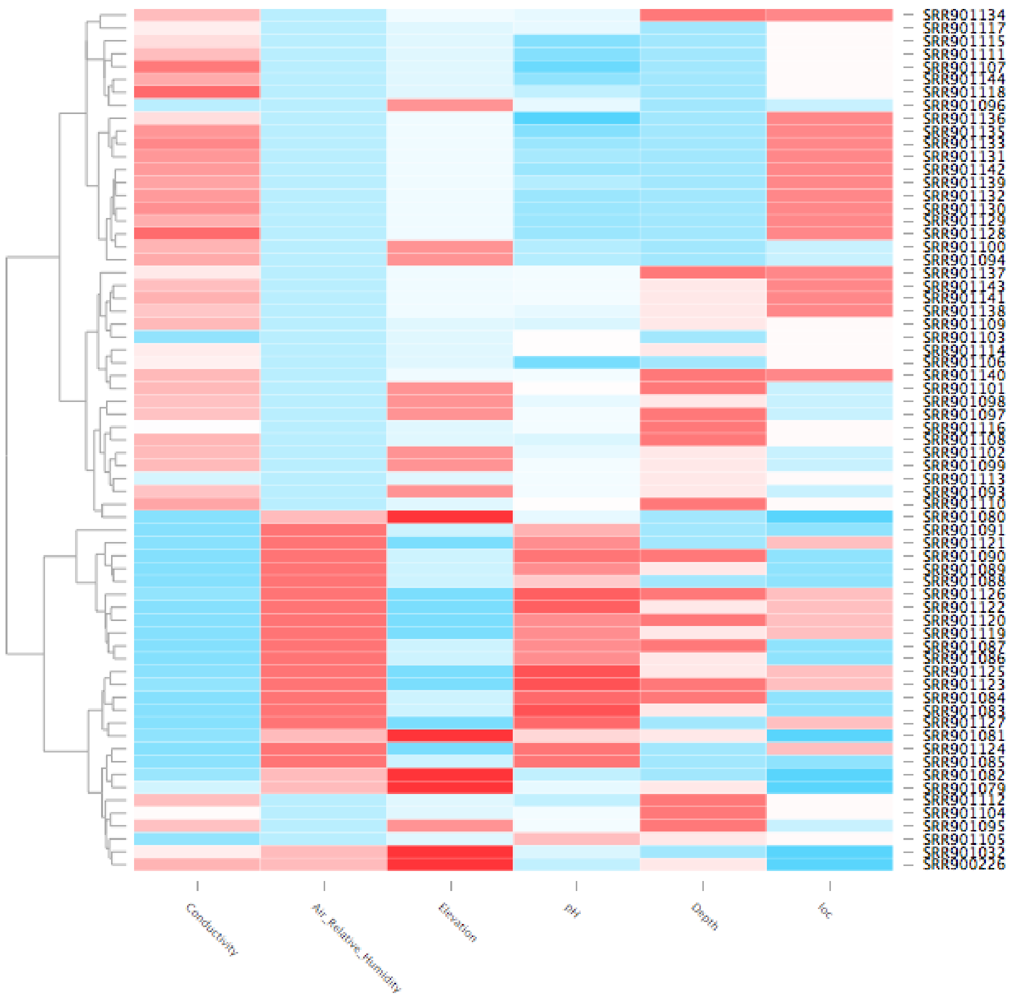


Figure 2. Heatmap of Various Environmental Variables, red indicating low and blue indicating high values. The left hand side is determined by Diana clustering on sample to sample similarities, as in Figure 1.

Although prior work had shown that alpha diversity relationships among Atacama desert samples were driven by geographic location (Crits-Christoph et al. 2013), our preliminary analysis suggested sample to sample similarities based on mash and NMF were better explained by pH, Relative Air Humidity, and Conductivity as well as the previously reported location variable. Note that this “cluster first” computationally focused approach is a departure from previous techniques that draw local communities using external metadata to overcome species dispersion, although the species’ relationships are often defined by interrelated sequence clusters (OTUs).

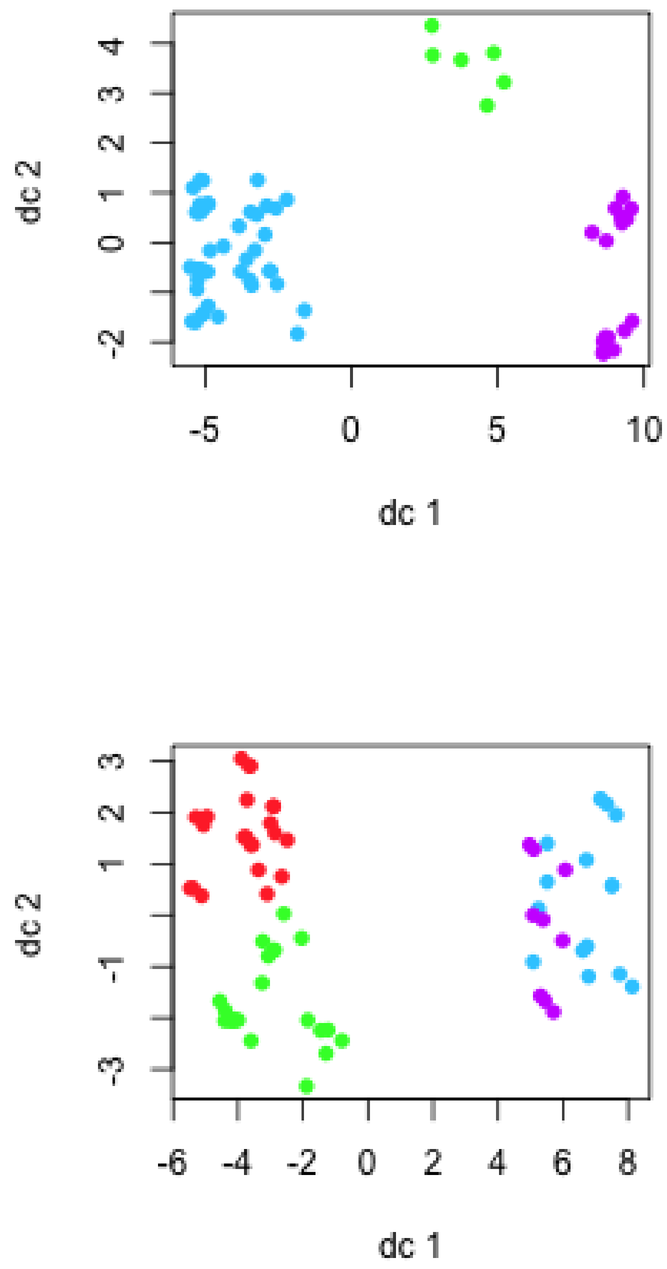


Figure 3. PCA of environmental variables at (top) K=3, and (bottom) K=4, colored according to assigned cluster as determined by K-means (Hartigan & Wong, M 1979).

To be consistent with current practice, we first applied Principal Component Analysis (PCA) using the samples' environmental variables (Air Humidity, Depth, Elevation, Soil Conductivity, and PH) to assess whether there is a ecological basis for observed clusters (Figure 2). We also used Average Silhouette Width (Rousseeuw, 1987), which provides a measure of how dense clusters are, with denser clusters being prefered. Average Silhouette can be used to determine the number of clusters by picking the higher average, in the case of comparing two candidate clusterings.

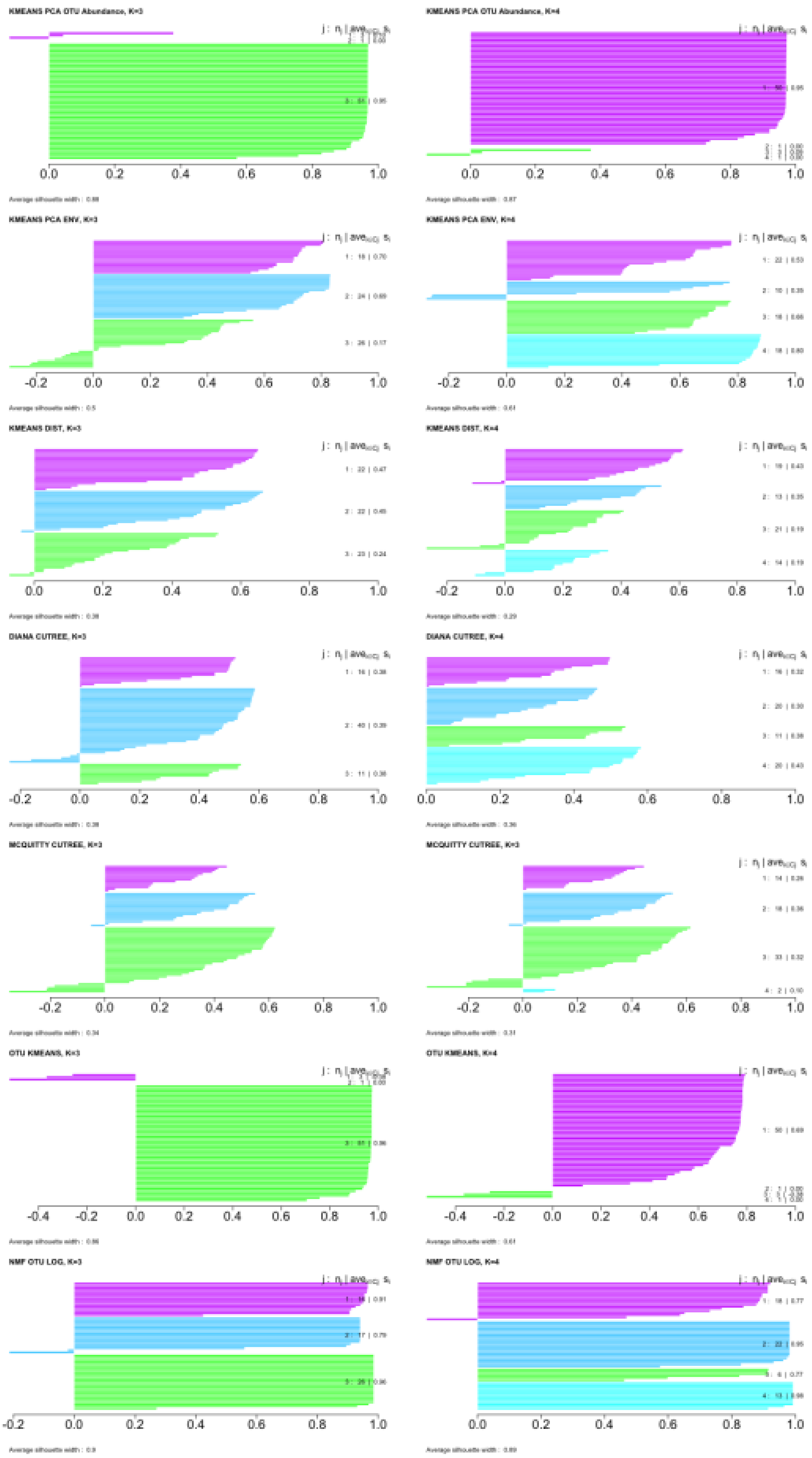

KMEANS PCA OTU Abundance, K=3
j : n_j | $ave_{i\in Cj}$ s_i
3 : 51 | 0.95
0.0 0.2 0.4 0.6 0.8 1.0
Average silhouette width : 0.88
KMEANS PCA OTU Abundance, K=4
j : n_j | $ave_{i\in Cj}$ s_i
1 : 50 | 0.95
0.0 0.2 0.4 0.6 0.8 1.0
Average silhouette width : 0.87
KMEANS PCA ENV, K=3
j : n_j | $ave_{i\in Cj}$ s_i
1 : 18 | 0.70
2 : 24 | 0.69
3 : 26 | 0.17
-0.2 0.0 0.2 0.4 0.6 0.8 1.0
Average silhouette width : 0.5
KMEANS PCA ENV, K=4
j : n_j | $ave_{i\in Cj}$ s_i
1 : 22 | 0.53
2 : 10 | 0.35
3 : 18 | 0.66
4 : 18 | 0.80
-0.2 0.0 0.2 0.4 0.6 0.8 1.0
Average silhouette width : 0.61
KMEANS DIST, K=3
j : n_j | $ave_{i\in Cj}$ s_i
1 : 22 | 0.47
2 : 22 | 0.45
3 : 23 | 0.24
0.0 0.2 0.4 0.6 0.8 1.0
Average silhouette width : 0.38
KMEANS DIST, K=4
j : n_j | $ave_{i\in Cj}$ s_i
1 : 19 | 0.43
2 : 13 | 0.35
3 : 21 | 0.19
4 : 14 | 0.19
-0.2 0.0 0.2 0.4 0.6 0.8 1.0
Average silhouette width : 0.29
DIANA CUTREE, K=3
j : n_j | $ave_{i\in Cj}$ s_i
1 : 16 | 0.38
2 : 40 | 0.39
3 : 11 | 0.38
-0.2 0.0 0.2 0.4 0.6 0.8 1.0
Average silhouette width : 0.38
DIANA CUTREE, K=4
j : n_j | $ave_{i\in Cj}$ s_i
1 : 16 | 0.32
2 : 20 | 0.30
3 : 11 | 0.38
4 : 20 | 0.43
0.0 0.2 0.4 0.6 0.8 1.0
Average silhouette width : 0.36
MCQUITTY CUTREE, K=3
-0.2 0.0 0.2 0.4 0.6 0.8 1.0
Average silhouette width : 0.34
MCQUITTY CUTREE, K=3
j : n_j | $ave_{i\in Cj}$ s_i
1 : 14 | 0.26
2 : 18 | 0.36
3 : 33 | 0.32
4 : 2 | 0.10
-0.2 0.0 0.2 0.4 0.6 0.8 1.0
Average silhouette width : 0.31
OTU KMEANS, K=3
j : n_j | $ave_{i\in Cj}$ s_i
3 : 51 | 0.96
-0.4 -0.2 0.0 0.2 0.4 0.6 0.8 1.0
Average silhouette width : 0.86
OTU KMEANS, K=4
j : n_j | $ave_{i\in Cj}$ s_i
1 : 50 | 0.69
-0.4 -0.2 0.0 0.2 0.4 0.6 0.8 1.0
Average silhouette width : 0.61
NMF OTU LOG, K=3
j : n_j | $ave_{i\in Cj}$ s_i
1 : 16 | 0.91
2 : 17 | 0.79
3 : 26 | 0.96
-0.2 0.0 0.2 0.4 0.6 0.8 1.0
Average silhouette width : 0.9
NMF OTU LOG, K=4
j : n_j | $ave_{i\in Cj}$ s_i
1 : 18 | 0.77
2 : 22 | 0.95
3 : 6 | 0.77
4 : 13 | 0.99
0.0 0.2 0.4 0.6 0.8 1.0
Average silhouette width : 0.89

Figure 4. Left Side, red, k=3, right side, orange, k=4, Top to Bottom: PCA determined clusters on OTU abundance, PCA determined environmental clusters, kmeans on mash distances, diana on mash distances, kmeans on OTU data (euclidean distance), NMF on OTU data, K=3 and k=4 were found to be viable, on the determination that an Average Silhouette Width above .5 was acceptable, while a score above .25 may indicate structure (Rousseeuw, 1987). *The environmentally driven PCA produced viable clusters at both K=3, and K=4, Kmeans on sequence similarity at K=3 weakly indicated structure, Diana clustering weakly indicated structure at K=3 and K=4, and NMF on OTUs produced structured clusters at both K=3 and K=4. All cluster silhouettes show that clustering at 3 or 4 maybe viable, with the exception of K-means on OTUs themselves, in which most samples clustered into a single, large group.*

Because silhouette average width for different clustering methods fell at the best values at either K=4, or at K=3, new clusters were generated at both. At K=3, clusterings were generated by Random Assignment, Non-negative Matrix Factorization based on abundance information, as well as log transformed OTU abundance, the three clusters with least within cluster distances from both the Diana, and from Mcquitty-WPGMA hierarchical clustering. Clusters were also made from Sample PH and from a North, South, or Central location. Since environmental variable mixing was previously reported to be the driver of beta diversity at k=3 (cite Atacama paper), we used environmental variable mixing to also generate clusterings. K=4 clusterings were generated with Random Assignment, Non-negative Matrix Factorization based on abundance information, as well as log transformed abundance information, the four clusters with least within cluster distances from both the Diana, and from Mcquitty-WPGMA hierarchical clustering, as well as from PH. Cluster to Cluster correlations show that Mcquitty-WPGMA is more similar to environmental clusterings; however, all non-random clusterings are more similar to each other than to randomly generated clusterings, indicating all detect some elements of community structure present in the data. Although this analysis has indicated that there was an ecological correlation to computationlly derived clusters, it has not shown which factors, or how those factors affect clustering. Further, skewed species abundances with a few dominant species could make it more difficult to sample rare species at modest sequencing depth; however, because Mash estimates the similarity between two sets, slight stochastic differences in observed abundances should not significantly affect the results relative to traditional OTU approaches that are also subject to sequence depth to uncover OTUs.

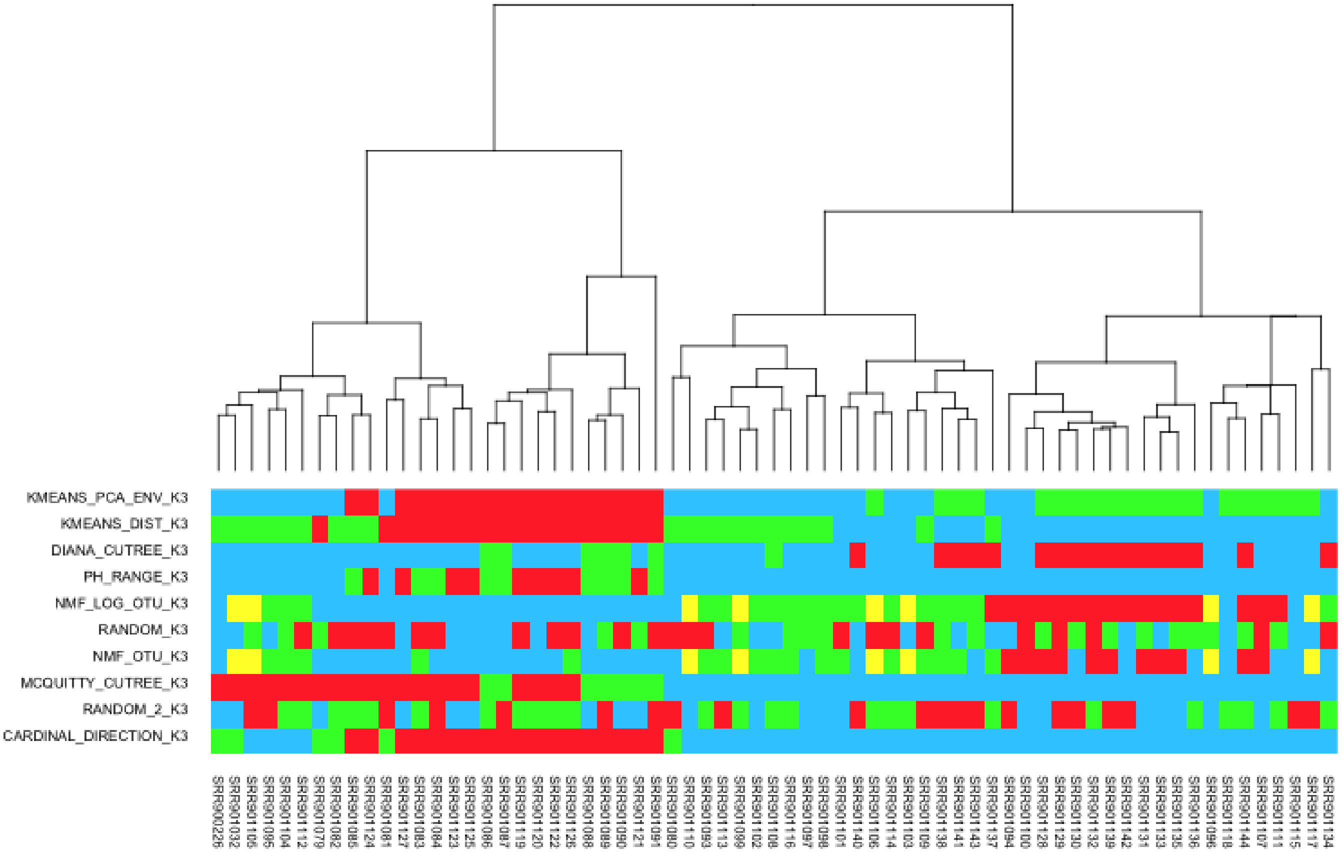


Figure 5. K=3 groupings ordered by a Diana determined dendrogram, yellow means that value was not found. It's more important that samples between belong to the same group than to the same color. Each row represents a grouping with three total groups. If between two grouping a sample has the same color with another sample that means that those two groupings put that sample in the same group. For instance, in K-means on PCA of environmental variables there is is a large red section, and NMF on log OTU many of those samples that were

marked red are now marked blue, that is an indication that these two methods have grouped these samples into their own clusters.

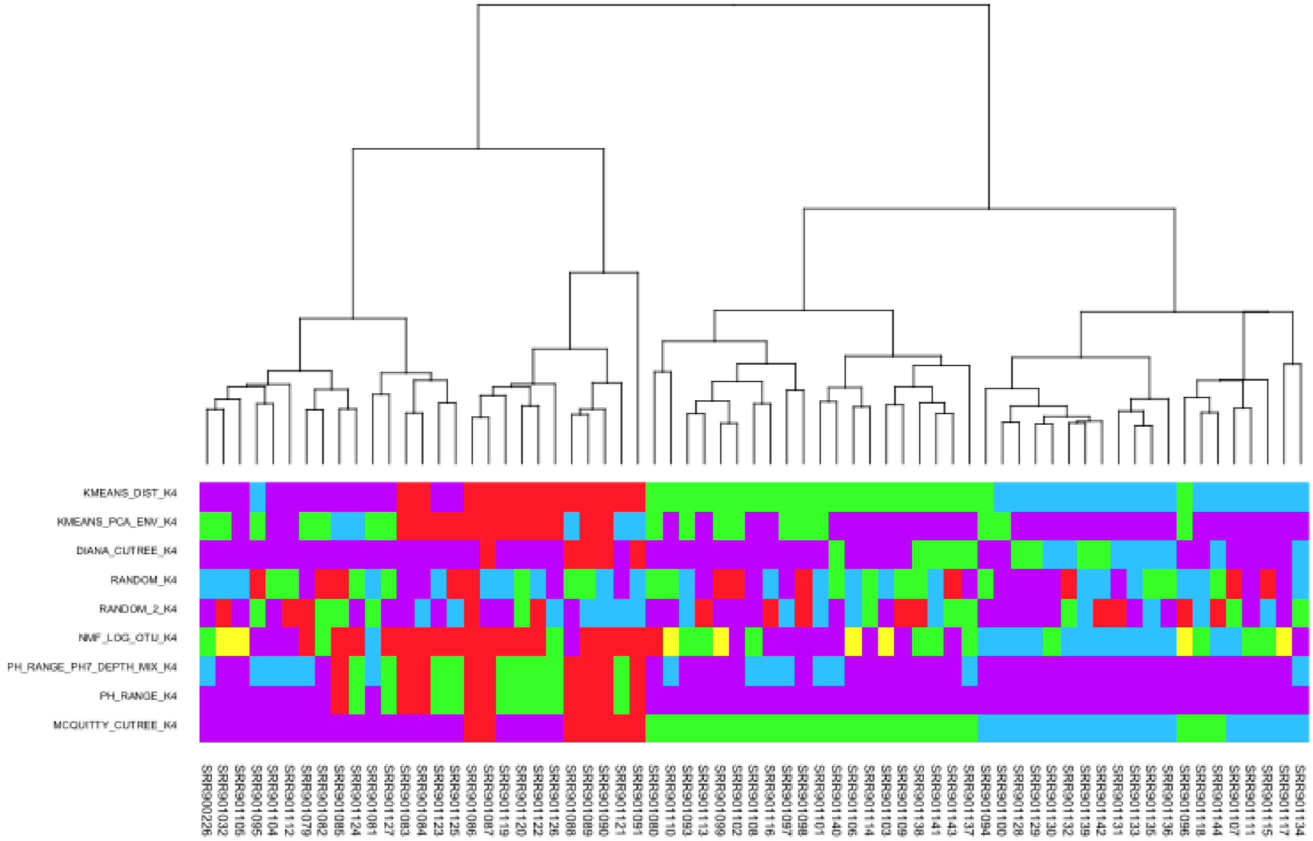
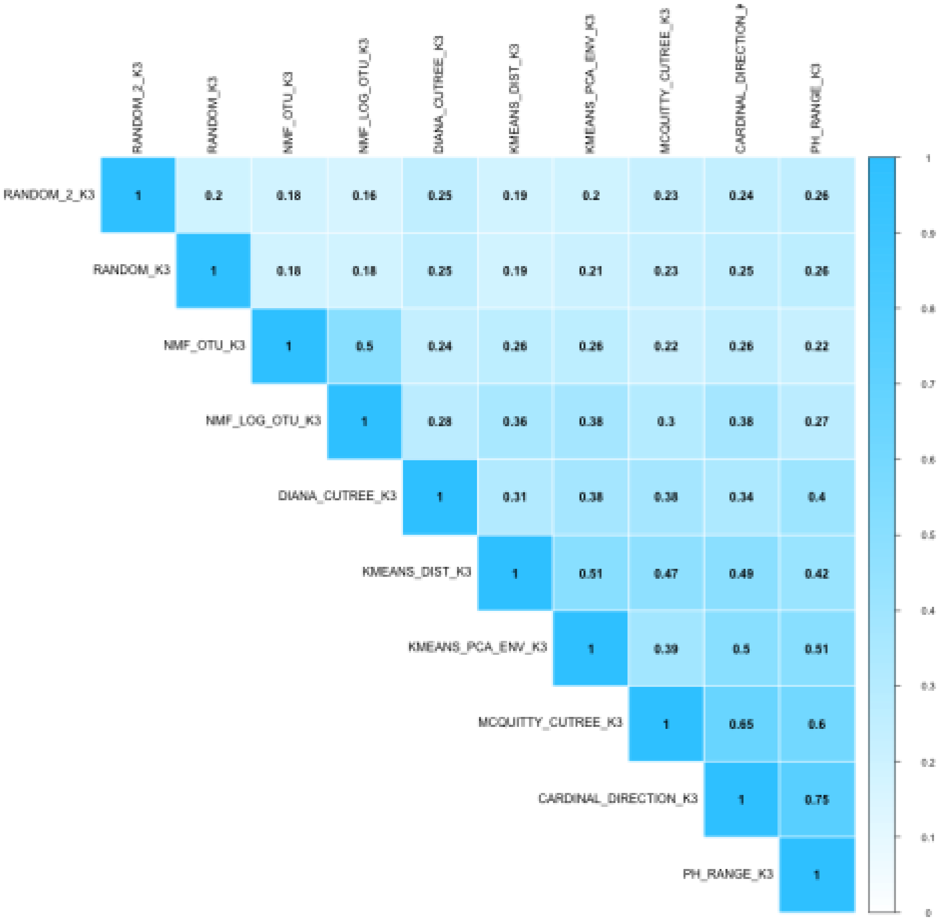


Figure 6. K=4 groupings ordered by a Diana determined dendrogram, yellow means that value was not found. It's more important that samples between belong to the same group than to the same color. Each row represents a grouping with three total groups. If between two grouping a sample has the same color with another sample that means that those two groupings put that sample in the same group.



175 *Figure 7. represents cluster similarities between each cluster, cluster similarity jaccard algorithm was used,*

176 *k=3 clusterings are shown.*

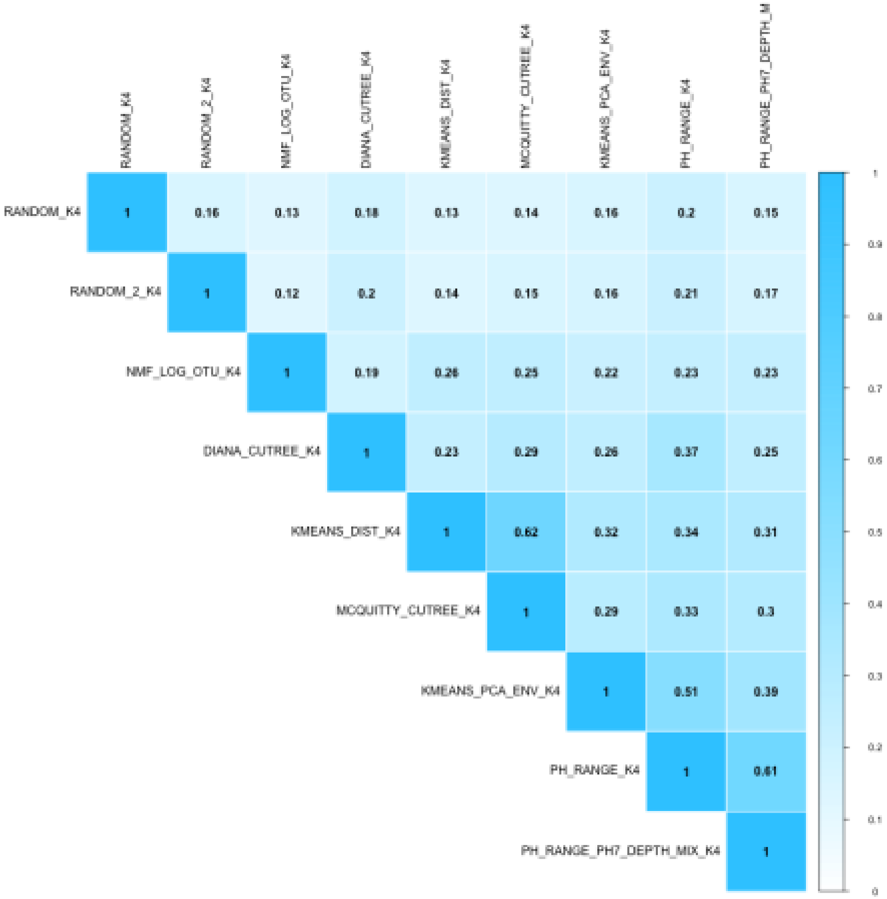


Figure 8. represents cluster similarities between each cluster, cluster similarity jaccard algorithm was used, k=4 clusterings are shown.

ANOVA Results	**MCQUITTY_CUTREE_K3**		**NMF_LOG_OTU_K3**		**MCQUITTY_CUTREE_K4**		**NMF_LOG_OTU_K4**	
Variable	P-value	Signf.	P-value	Signif.	P-value	Signif.	P-value	Signif.
pH	6.97E-13	***	1.13E-12	***	3.99E-14	***	2.41E-07	***
Elevation	0.14448		0.015637	*	0.000823	***	0.02558	*
Conductivity	0.00573	**	7.32E-05	***	0.398154		0.02377	*
Air_Relative_Humidity	5.03E-07	***	0.000361	***	0.006096	**	0.01114	*

Depth	0.08345	.	0.310562		0.004934	**	0.02009	*
pH with Conductivity	0.81559		0.026234	*	0.005361	**	0.00579	**
Elevation with Conductivity	0.50011		0.784539		0.14726		0.01553	*
Elevation with Air_Relative_Humidity	0.01035	*	0.375278		0.029011	*	0.69145	
Conductivity with Air_Relative_Humidity	0.00156	**	0.927365		0.022087	*	0.04518	*
pH and Depth	0.1124		0.99117		0.039508	*	0.01784	*
Conductivity with Air_Relative_Humidity with Depth	0.03048	*	0.371664		0.133212		0.91445	
pH with Elevation with Conductivity with Depth	0.22824		0.425284		0.777819		0.03665	*
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1								
8 observations deleted due to missingness in NMF Analyses, 1 observations deleted due to missingness in Mcquitty Analyses								

The clusterings were modeled by ANOVA, and after calculating a log likelihood test, we found that for both K=3 and K=4 Mcquttity hierarchical clustering, followed by NMF on OTUs, were the most significant and therefore best corresponded to the environmental data. For McQuitty hierarchal clustering, PH and Elevation were found to have the most significance, however, since the elevation was the same for all of the samples of any given sampling site, since elevation is highly correlated with sampling location there maybe some other variable that is being indirectly measured, also highly correlated with sampling location. For NMF on log abundance PH , Conductivity, and Relative humidity of the air were found to be most significant; however, because relative humidity of each sampling site was the same, it is unknown whether relative humidity of the air was the contributing factor or some other, unknown variable, that also differed from site to site was a factor.

McQuitty clustering has a .65 similarity with the Cardinal Direction, and similarly High similarities with other environmentally determined groupings. We also see that both McQuitty and NMF have high p-values with some environmental variables in anova, with Ph being particularly significant in both McQuitty and NMF, and to a lesser extent Relative Air Humidity being significant as well, and that the sample similarities within these groupings are high. This shows that OTU based methods and distance based methods produce similar results, if driven by slightly different environmental variables, and is getting at the underlying structure of the local communities.

As per the clustering silhouettes, some of the methods, Diana and NMF, work better at four clusters, while McQuitty and K-means did better at three. The most explicative results, as per ANOVA, NMF on log OTU abundance and McQuitty slightly disagree on which environmental variables have the most importance, but PH and Relative Air Humidity can be seen across all four ANOVAs.

Concluding Remarks and Recommendations

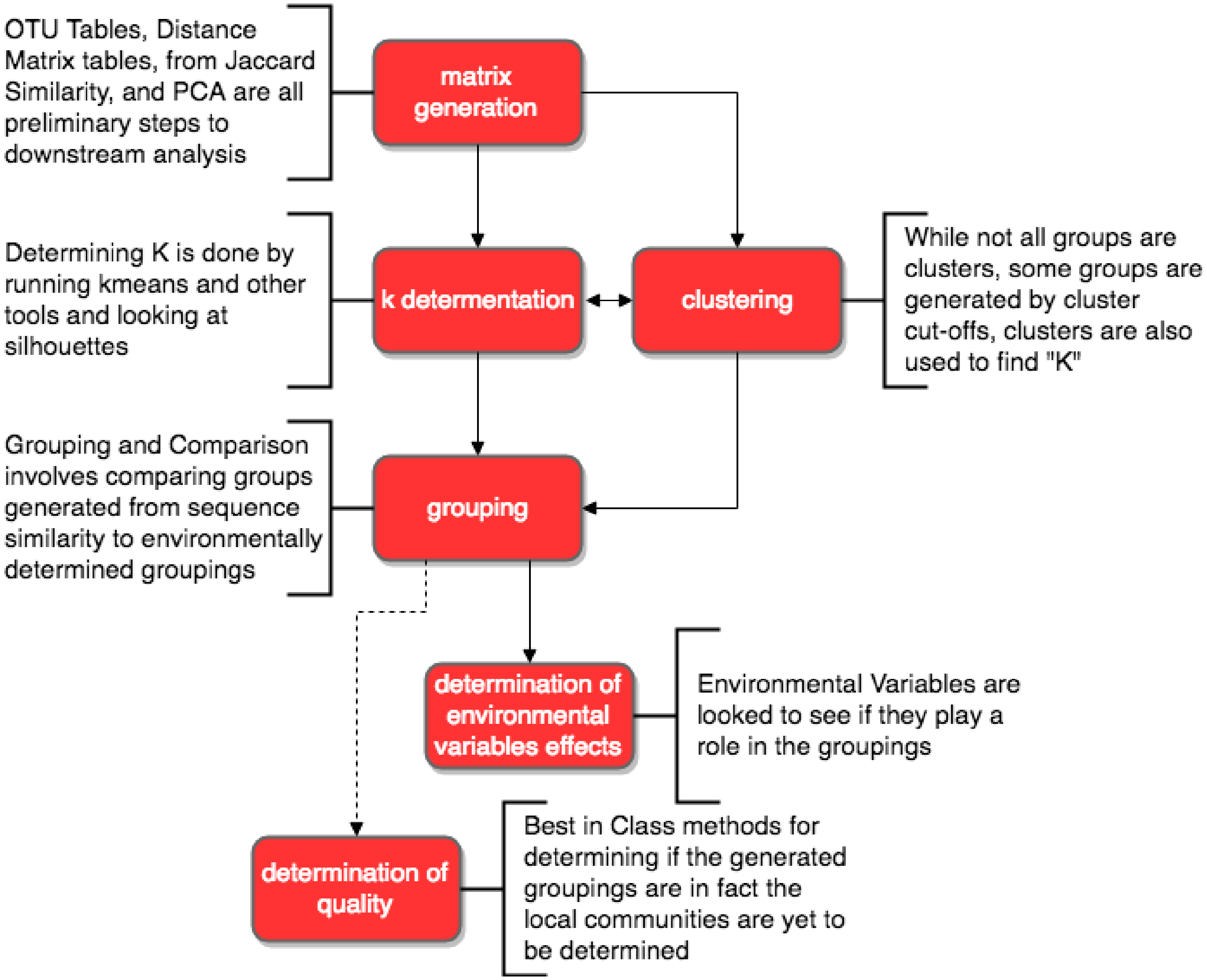

OTU Tables, Distance Matrix tables, from Jaccard Similarity, and PCA are all preliminary steps to downstream analysis
matrix generation
Determining K is done by running kmeans and other tools and looking at silhouettes
k determentation
clustering
While not all groups are clusters, some groups are generated by cluster cut-offs, clusters are also used to find "K"
Grouping and Comparison involves comparing groups generated from sequence similarity to environmentally determined groupings
grouping
determination of environmental variables effects
Environmental Variables are looked to see if they play a role in the groupings
determination of quality
Best in Class methods for determining if the generated groupings are in fact the local communities are yet to be determined

Figure 9. Workflow for Determining, Describing, and Validating Atacama data.

As a general workflow (figure 9), after sample collection either OTUs abundances are generated, or sample to sample distances are calculated by comparing their contained trimmed sequences. In the case of the sample to sample distances a distance matrix is generated that can be clustered though hierarchical or other means, and in the case of OTU abundances NMF or K-means is better suited. We calculated pairwise distance on both shared sequences and on OTUs, and then clustered OTUs and shared sequences via K-means, and for shared sequences diana clustering was also utilized, and for NMF was also utilized for OTU abundance. The groupings can then be checked for the influence of independent variables, through a statistical model, in this case anova, which was run on clusters, the 'anova' function from the R 'stats' package was used, and LogLik from the R 'stats' package was used to compare Log-Liklihoods. Clusters were compared to each other using both RAND and Jaccard similarity cluster evaluation methods, as well as a wilcox test (Hollander et al. 2013);(Bauer, 1972).

The Atacama data used here is from SRA:SRA091062, Bioproject ID: PRJNA208226, which was thought of as three clusters of data, aligning with sampling site: North, Central, and South. Atacama was chosen for its previous environmental analysis, geographically distinct sampling sites, and curated metadata.

Mothur was used to process Raw files for OTU analysis as per non-shhh (Quince et al. 2009) 454 SOP: https://www.mothur.org/wiki/454_SOP. for sequence similarity distance Mothur was used to filter samples based quality scores, as per the shhh and trimming portion of the mothur 454 SOP. Initial NMF analysis (figure 10) was done with “sake” (https://github.com/naikai/sake), which was originally created to analyze gene expression data, was here utilized to look at OTU abundance data, at k=3 both log transformed and non-log transformed data was utilized, the nsNMF NMF algorithm NMF algorithm was used and the the NMF tool was run at 350 runs, at k=4 only log transformed data was run, with the nsNMF (Pascual-Montano et al. 2006) , NMF algorithm at 350 runs. nsNMF was chosen for its design to deal with perceived sparseness in the data. The R 'cluster_similarity' function from the 'clusteval' package was used for Jaccard and RAND similarities, while 'wilcox.test' function from the R 'stats' package was used for the wilcox test. wpgma was chosen for Agglomerative clustering because clusters were expected to be of unequal size, as unweighted hierarchical methods can become distorted when large and small groups are compared, and a clear contrast to centroid clustering, as like k-means, was desired. The R 'hclust' function was used from the 'stats' package was used for agglomerative clustering. Diana, from the R ‘cluster’ pacakge was used for divisive hierarchical clustering, in agglomerative hierarchical clustering samples are combined until all samples are in the same cluster, whereas in divisive hierarchical clustering all samples start in the same cluster and then are partitioned into daughter clusters.. And further analysis and figure analysis was done with the caret (https://cran.r-project.org/package=caret), clusteval (https://cran.r-project.org/package=clusteval), cluster (https://CRAN.R-project.org/package=cluster), corrplot (https://CRAN.R-project.org/package=corrplot), d3heatmap (https://CRAN.R-project.org/package=d3heatmap), fpc (https://CRAN.R-project.org/package=fpc), gplots (https://CRAN.R-project.org/package=gplots), and NMF (https://CRAN.R-project.org/package=NMF) R packages.

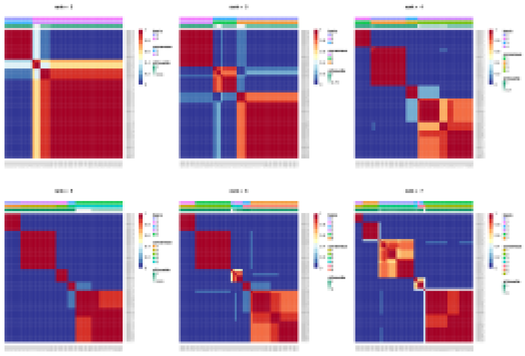


Figure 10. Correlations of various number of clusters (k) based on K-Means from k=2 to k=7.

References

Holyoak, M., Leibold, M.A., Holt, R.D. (2005). *Metacommunities : spatial dynamics and ecological communities*. University of Chicago Press.

Parsons, K.J., Cooper, W.J., Albertson, R.C., Lundrigan, B. & Jr, G. (2009). Limits of Principal Components Analysis for Producing a Common Trait Space: Implications for Inferring Selection, Contingency, and Chance in Evolution (I. Dworkin, Ed.). *PLoS ONE*, **4**, e7957.

Jolliffe, I.T. (1986). Principal Component Analysis and Factor Analysis. pp. 115–128. Springer, New York, NY.

Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S. & Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, **17**, 132.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J.A., Hoffman, J.M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J.E., Li, K., Kravitz, S., Heidelberg, J.F., Utterback, T., Rogers, Y.-H., Falcón, L.I., Souza, V., Bonilla-Rosso, G., Eguiarte, L.E., Karl, D.M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M.R., Strausberg, R.L., Nealson, K., Friedman, R., Frazier, M. & Venter, J.C. (2007). The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific (N.A. Moran, Ed.). *PLoS Biology*, **5**, e77.

Seshadri, R., Kravitz, S.A., Smarr, L., Gilna, P. & Frazier, M. (2007). CAMERA: A Community Resource for Metagenomics. *PLoS Biology*, **5**, e75.

Hultman, J., Waldrop, M.P., Mackelprang, R., David, M.M., McFarland, J., Blazewicz, S.J., Harden, J., Turetsky, M.R., McGuire, A.D., Shah, M.B., VerBerkmoes, N.C., Lee, L.H., Mavrommatis, K. & Jansson, J.K. (2015). Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature*, **521**, 208–212.

Gibbons, S.M. & Gilbert, J.A. (2015). Microbial diversity--exploration of natural ecosystems and microbiomes. *Current opinion in genetics & development*, **35**, 66–72.

Gaujoux, R. & Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, **11**, 367.

Seung, H.S. & Lee, D.D. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.

Paatero, P. & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, **5**, 111–126.

Pascual-Montano, A., Carazo, J.M., Kochi, K., Lehmann, D. & Pascual-Marqui, R.D. (2006). Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**, 403–415.

Crits-Christoph, A., Robinson, C.K., Barnum, T., Fricke, W., Davila, A.F., Jedynak, B., McKay, C.P. & DiRuggiero, J. (2013). Colonization patterns of soil microbial communities in the Atacama Desert. *Microbiome*, **1**, 28.

Berman, A. & Plemmons, R. (1994). *Nonnegative matrices in the mathematical sciences*.

Ho, Yu-Jui. (2016). Single-cell RNA-Seq Analysis and Klustering Evaluation. URL https://github.com/naikai/sake [accessed 24 April 2017]

Hartigan, J. & Wong, M. (1979). Algorithm AS 136: A k-means clustering algorithm. *Statistical Society. Series C (Applied Statistics)*.

Forgey, E. (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classification. *Biometrics*.

Broder, A. (1997). On the resemblance and containment of documents. *Compression and Complexity of Sequences 1997*.

Struyf, A., Hubert, M. & Rousseeuw, P. (1997). Integrating robust clustering techniques in S-PLUS. *Computational Statistics & Data*.

Kaufman, L. & Rousseeuw, P. (2009). *Finding groups in data: an introduction to cluster analysis*.

McQuitty, L.L. (1966). Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data. *Educational and Psychological Measurement*, **26**, 825–831.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53–65.

Handl, J., Knowles, J. & Kell, D.B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, **21**, 3201–3212.

DeLong, E.F. (2013). *Microbial metagenomics, metatranscriptomics, and metaproteomics*.

Everitt, B. (1974). Cluster analysis: An SSRC review of recent research.

Hartigan, J. (1975). Clustering algorithms.

Hollander, M., Wolfe, D. & Chicken, E. (2013). *Nonparametric statistical methods*.

Bauer, D. (1972). Constructing confidence sets using rank statistics. *Journal of the American Statistical Association*.

Quince, C., Lanzén, A., Curtis, T.P., Davenport, R.J., Hall, N., Head, I.M., Read, L.F. & Sloan, W.T. (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods*, **6**, 639–641.