

1 **Methods in Description and** 2 **Validation of Local Metagenetic** 3 **Microbial Communities**

4 **Authors:**

5 David Molik, Integrated Biomedical Sciences, Galvin Life Sciences, University of Notre Dame, Notre
6 Dame, IN 46556, dmolik@nd.edu

7
8 Michael E. Pfrender, Department of Biological Sciences and Environmental Change Initiative, Galvin
9 Life Sciences, University of Notre Dame, Notre Dame, IN 46556, mpfrende@nd.edu

10
11 Scott Emrich, Computer Science and Engineering, Fitzpatrick Hall, University of Notre Dame, Notre
12 Dame, IN 46556, semrich@nd.edu

13 **Abstract:**

- 14 1. We propose MinHash (as implemented by MASH) and NMF as alternative methods to estimate similarity
15 between metagenetic samples. We further describe these results with cluster analysis and correlations
16 with independent ecological metadata.
- 17 2. Using sample to sample similarities based on MinHash similarities we use hierarchal clustering to
18 generate clusters, simultaneously we generate groups based on NMF, and we compare groups generated
19 from the MinHash similarity derived clusters and from NMF to those determined by the environment,
20 looking to Silhouette Width for an assessment of the quality of the cluster.
- 21 3. We analyze existing data from the Atacama Desert to determine the relationship between ecological
22 factors and group membership, and using the generated groups from MASH and NMF we run an
23 ANOVA to uncover links between metagenetic samples and known environmental variables such as pH
24 and Soil Conductivity.

25 **Introduction**

26 How microbial communities, and in some broader context local communities, are determined, described, and
27 validated is a matter of some debate (Holyoak et al. 2005). Principal Components Analysis (PCA) is the most
28 common computational approach used to assess patterns of community. A typical metagenetic experimental design
29 being the sequencing of gene regions that have gone through PCR from aquatic or soil samples, sequences would
30 then be made OTUs, and PCA applied. PCA is biased towards components that have the most variance (Parsons
31 et al. 2009). Communities also can be delineated from another by inferred differences in the identity and
32 abundance of species detected within one or more samples (Rusch et al. 2007; Seshadri et al. 2007). Here we
33 present two such alternative computational methods: MinHash (Broder, 1997) sketching and Non-Negative
34 Matrix Factorization (NMF) (Seung & Lee 1999). NMF can be paired with k-means to estimate the number of
35 groups (i.e., potential local communities) present and has the benefit of determining the most important feature
36 driving inferred relationships. MinHash sketches can be used to quickly estimate similarities between whole
37 samples in an alignment-free approach, i.e., OTUs do not need to be generated first. While MinHash and NMF are
38 used here to cluster metagenetic samples based on inferred relationships, note that NMF focuses on what is
39 distinct (in a cluster) while a MinHash implementation (Ondov et al. 2016) is combined with hierarchical methods
40 to infer clusters based on pairwise similarities.

41 Because detected abundances of a species in metagenetic samples may not correlate with its actual abundance
42 in a broader area, drawing boundaries between actual local communities using any computational approach can be
43 difficult. As a result, prior work in community analysis has often relied on metadata such as physical barriers and
44 environmental measurements to refine the structure of estimated local communities based on the species observed
45 (Holyoak et al. 2005). We propose similar metadata-driven analysis using Silhouette plots for cluster assessment,
46 which is a computational measure of how close each point in a cluster is to other clusters. Finally, we use
47 ANOVA statistical tests to determine association of known environmental factors to the inferred clusters using
48 our new and existing approaches. The advantage of these novel, data-driven (unsupervised) approaches for
49 defining communities is that it allows us to artificially induce computational cutoffs, and, as a result, no prior
50 knowledge/metadata are required to infer associations. Because environmental characteristics can change the

51 viability of a microbial species occupying that area (Hultman et al. 2015; Gibbons & Gilbert 2015), subsequent
52 comparisons of groupings to independent environmental variables provides a biologically motivated assessment
53 of whether these computationally generated results uncover local communities.

54 To assess our new approach we have chosen Atacama desert microbial community because of the data's
55 wide geographic range and inclusion of environmental variables. Log-likelihood statistical analysis of an indicates
56 that among these *de novo* methods applied to Atacama data, hierarchical clustering using MinHash similarities
57 has more explicative power than NMF on OTU abundance (see supplemental). In the previously reported analysis
58 of this Atacama desert dataset samples taken from the same sampling location (North/Central/South) were more
59 similar according to alpha diversity (Crits-Christoph et al. 2013); however, we show that other environmental
60 variables can have a statistically higher correlation than sampling location, and specifically that pH, air relative
61 humidity (RH) and soil conductivity best explain observed local communities derived computationally.
62 Combined, these results indicate data-driven methods can be directly used to estimate community structure from
63 NGS data.

64 **Methods**

65 To define clusters we introduce MinHash (Ondov et al. 2016) based similarity for determining local
66 community structure, which is essentially an approximation of the Jaccard similarity based on shared species
67 within samples (see Rusch et al. 2007 and Ondov et al. 2016 for details). We also apply Non-Negative Matrix
68 Factorization (NMF) (Gaujoux & Seoighe, 2010);(Seung & Lee 1999);(Paatero & Tapper 1994) using the nsNMF
69 algorithm (Pascual-Montano et al. 2006) to determine non-shared species based on OTU abundances.

70 NMF—or Non-Negative Matrix Factorization—is method by which to split a matrix into a component based
71 on the factors that are most important in making that split. For example, for RNA-seq expression analysis,
72 suppose there are 'k' known clusters. NMF will break a provided expression matrix (genes by cells or cell
73 tissues) into *k* total clusters while also producing the most important genes for doing so (Yu-Jui, 2017). When

74 applied to observed OTU abundances, NMF will ideally return the most important OTUs to generate a fixed
75 number of clusters. The power in this method is that different factors may be indicators for each cluster, instead of
76 just the presence or absence of a particular observed species. NMF becomes particularly powerful when paired
77 with k-means (Hartigan & Wong, 1979);(Forgey, 1965), which is a clustering method that can be used to measure
78 how many clusters exist (aka, the ‘fit’). Determining factors in NMF can be done with Non-Negative coefficients
79 while PCA has orthogonal vectors with positive and negative coefficients and since NMF combines factor
80 discovery with iterative determination of the total number of clusters, NMF can be a more descriptive alternative
81 to simple PCA-based visualization.

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

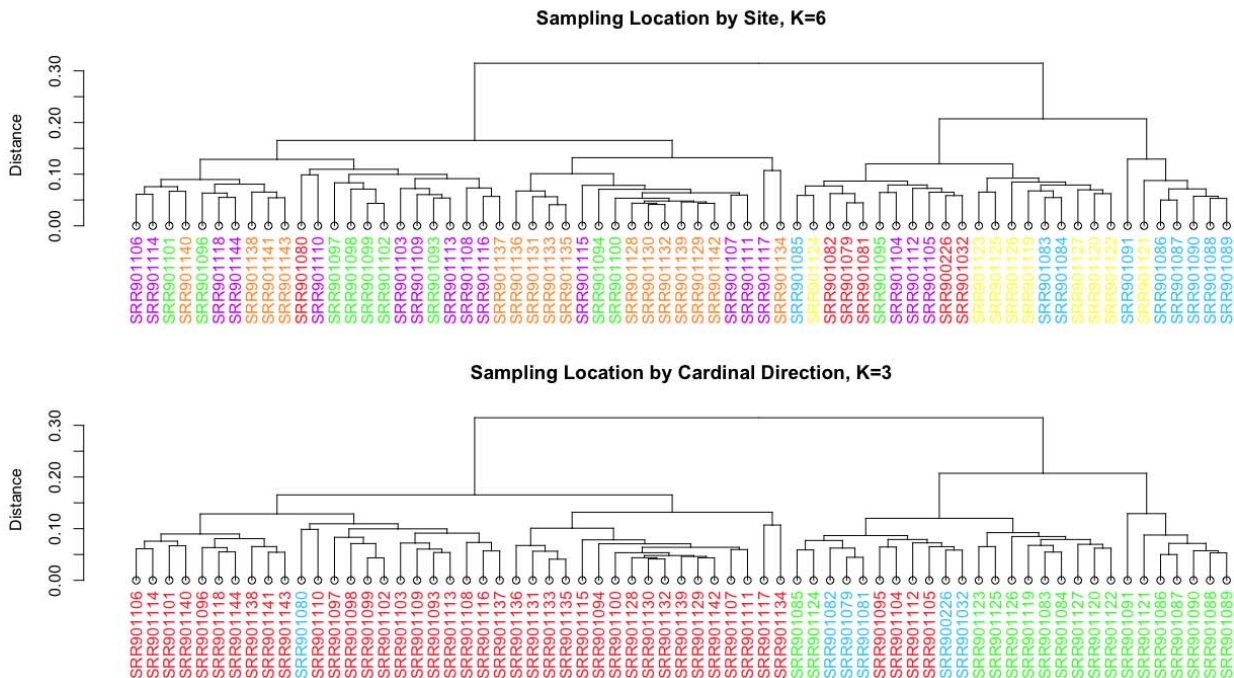
89 Using Silhouette Widths (Rousseeuw, 1987);(Handl et al. 2005) and the clustering information derived from
90 NMF we can further describe structure within a cluster. Specifically, Silhouette Width highlights the
91 ‘belongingness’ of each data point within a cluster; higher averages indicate cluster points are more tightly
92 correlated with each other. Silhouettes are a tool to show how much overlap there is between clusters, or how
93 consistent or distinct they are, similar to looking for distinct clusters in PCA plots.

94 Finally, we hypothesize that the local assortment of species is largely determined by the environment in
95 which they live. If so, a change in environment and a corresponding change in observed species should, for the
96 most part, correlate and this correspondence can be tested using both ANOVA and a mantel test under the right
97 conditions (DeLong, 2013). We also realize that environment itself can correlate with distance, i.e., in the

98 northern hemisphere, northern samples have fewer growing degree days than southern samples. For this reason
99 isolation-by-distance (IBD) could also manifest as distinct clusters using our computational alternatives just as
100 they would in a traditional PCA analysis.

101 **Results**

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



110

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

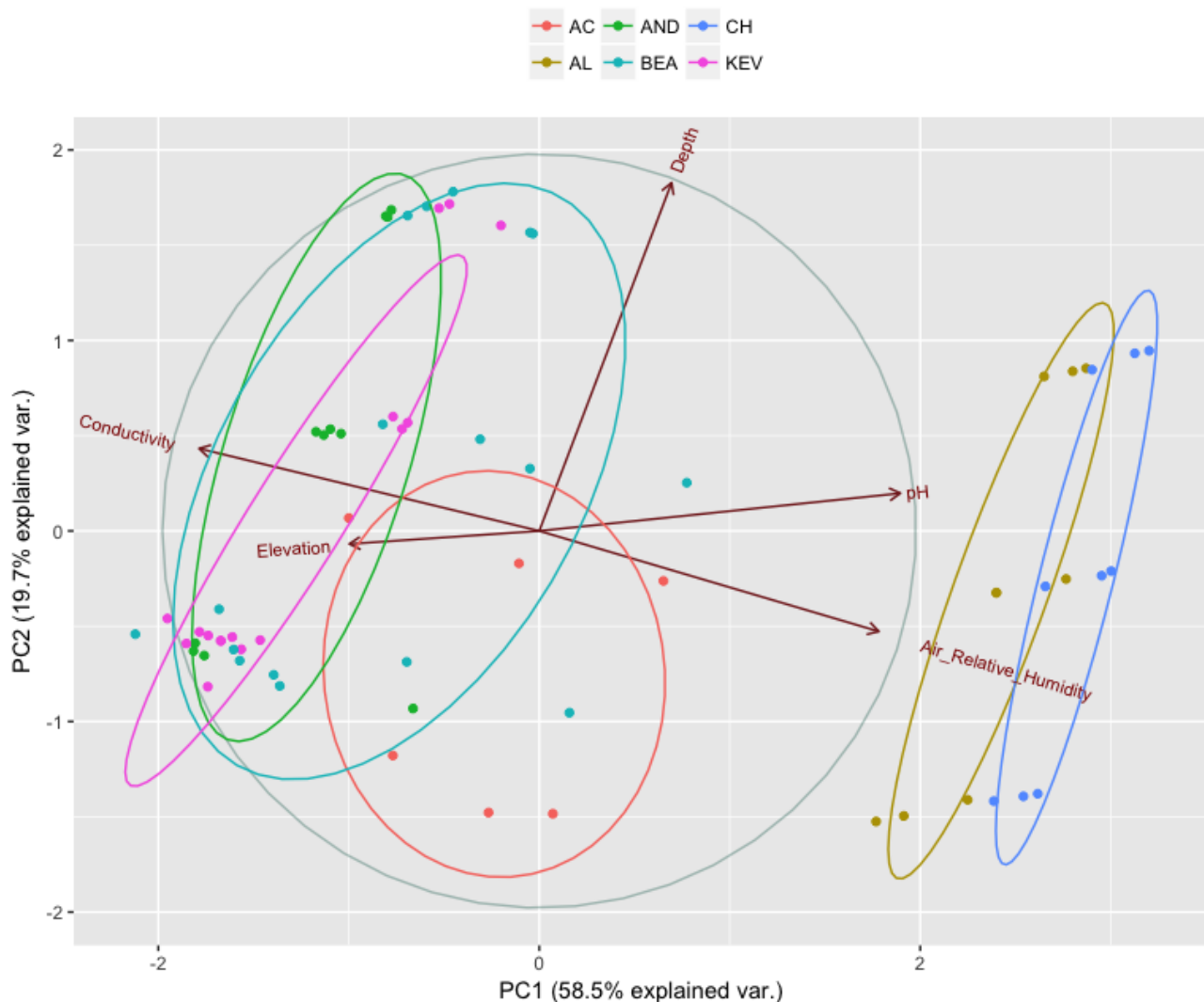


114

115

116 *Figure 2. Heatmap of Various Environmental Variables, scaled in color from low (blue) to high (red), color*
117 *scale per column. The left hand side is determined by McQuitty clustering algorithm on sample to sample*
118 *similarities, as in Figure 1.*

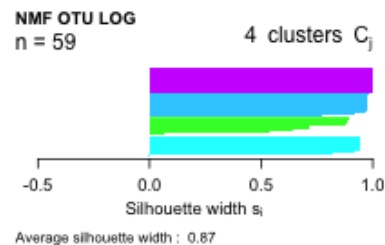
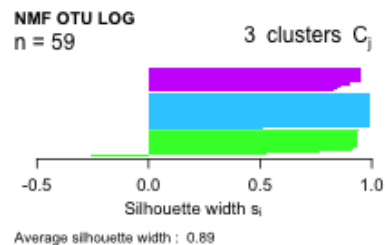
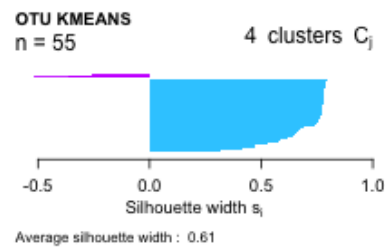
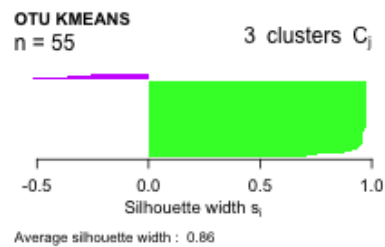
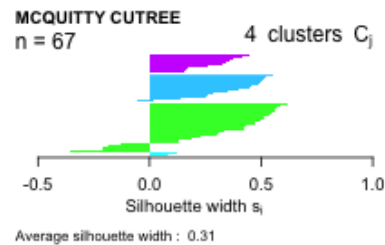
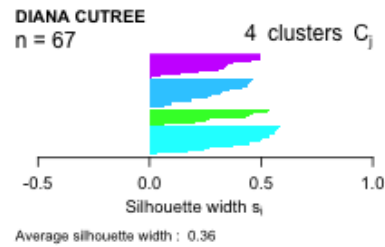
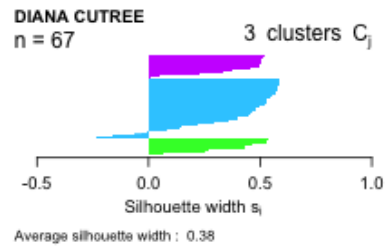
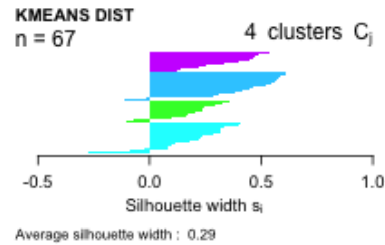
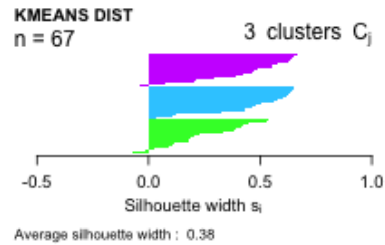
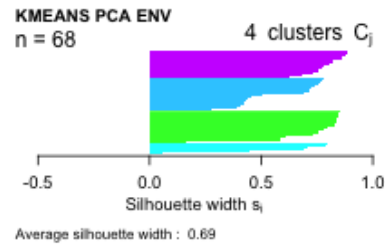
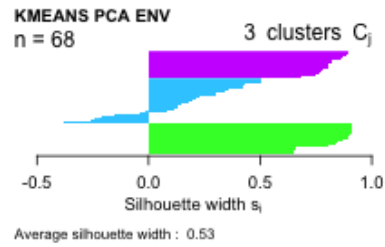
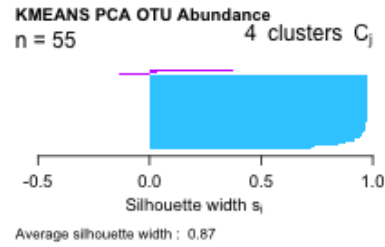
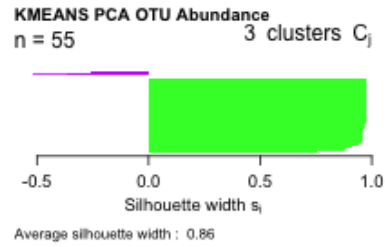
119 Although prior work had shown that alpha diversity relationships among Atacama desert samples were driven
120 by geographic location (Crits-Christoph et al. 2013), our preliminary analysis suggested sample to sample
121 similarities based on mash and NMF were better explained by pH, Relative Air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



125

126 *Figure 3. PCA (Hartigan & Wong, M 1979) of environmental variables, colored according to sampling site.*

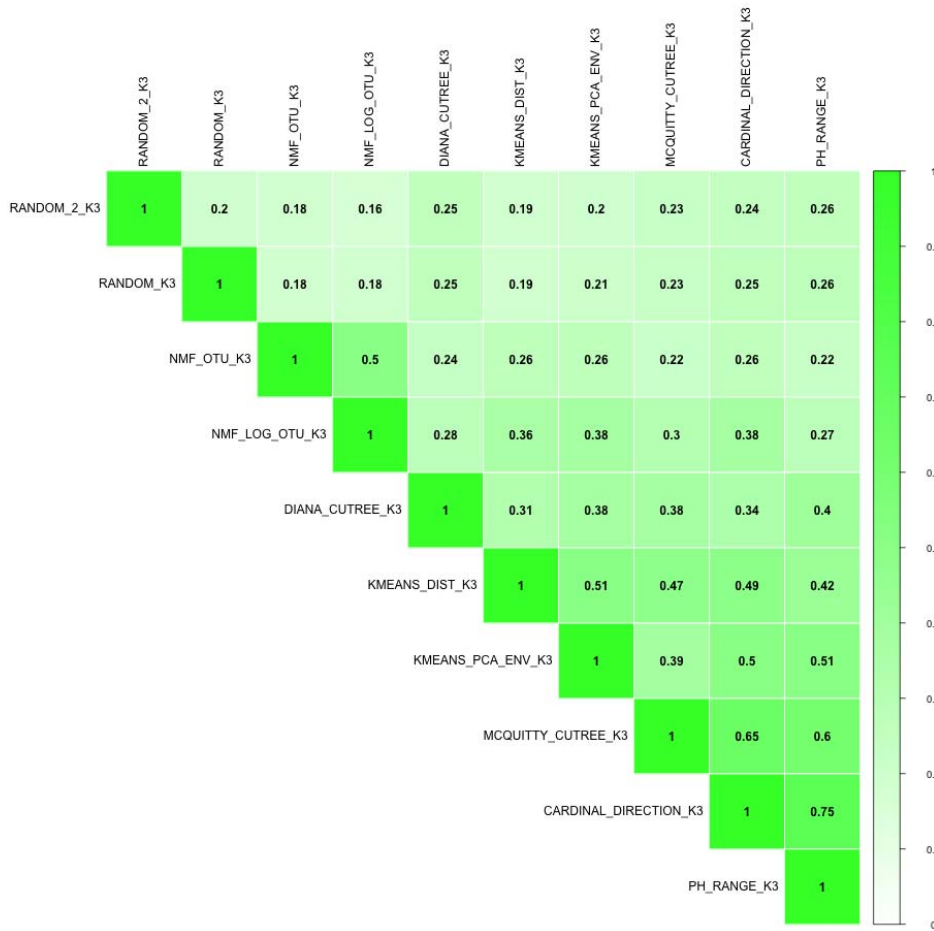
127 To be consistent with current practice, we first applied Principal Component Analysis (PCA) using the
128 samples' environmental variables (Air Humidity, Depth, Elevation, Soil Conductivity, and PH) to assess whether
129 there is a ecological basis for observed clusters (Figure 2). We also used Average Silhouette Width (Rousseeuw,
130 1987), which provides a measure of how dense clusters are, with denser clusters being preferred. Average
131 Silhouette can be used to determine the number of clusters by picking the higher average, in the case of
132 comparing two candidate clusterings.



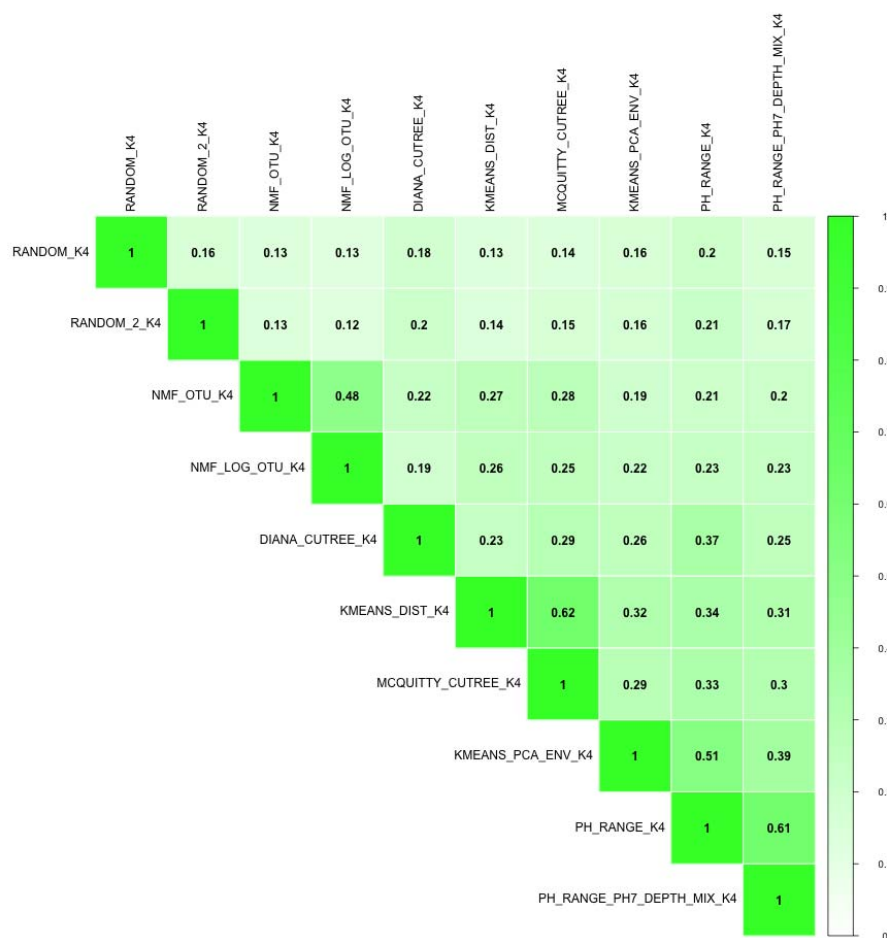
134 *Figure 4. Left side has 3 clusters, while right side utilizes 4 clusters. Top to Bottom: PCA determined clusters*
135 *on OTU abundance, PCA determined environmental clusters, kmeans on mash distances, diana on mash*
136 *distances, kmeans on OTU data (euclidean distance), NMF on OTU data, K=3 and k=4 were found to be viable,*
137 *on the determination that an Average Silhouette Width above 0.5 was acceptable. A score above 0.25 may*
138 *indicate structure (Rousseeuw, 1987). The environmentally driven PCA produced viable clusters at both K=3,*
139 *and K=4, Kmeans on sequence similarity at K=3 weakly indicated structure, Diana clustering weakly indicated*
140 *structure at K=3 and K=4, and NMF on OTUs produced structured clusters at both K=3 and K=4. All Silhouette*
141 *Widths show that clustering at 3 or 4 maybe viable, with the exception of K-means on OTUs, in which most*
142 *samples clustered into a single, large group. Not all samples were viable in each method, “n=” indicates number*
143 *of samples utilized in each method.*

144 Because Silhouette Width Average for different clustering methods fell at the best values at either K=4, or at
145 K=3, new clusters were generated at both. At K=3, clusterings were generated by Random Assignment, Non-
146 negative Matrix Factorization based on abundance information, as well as log transformed OTU abundance, the
147 three clusters with least within cluster distances from both the Diana, and from Mcquitty-WPGMA hierarchical
148 clustering. Clusters were also made from Sample PH and from a North, South, or Central location. Since
149 environmental variable mixing was previously reported to be the driver of beta diversity at k=3 (Crits-Christoph
150 et al. 2013), we used environmental variable mixing to also generate clusterings. K=4 clusterings were generated
151 with Random Assignment, Non-negative Matrix Factorization based on abundance information, as well as log
152 transformed abundance information, the four clusters with least within cluster distances from both the Diana, and
153 from Mcquitty-WPGMA hierarchical clustering, as well as from PH. Cluster to Cluster correlations show that
154 Mcquitty-WPGMA is more similar to environmental clusterings; however, all non-random clusterings are more
155 similar to each other than to randomly generated clusterings, indicating all detect some elements of community
156 structure present in the data. Although this analysis has indicated that there was an ecological correlation to
157 computationally derived clusters, it has not shown which factors, or how those factors affect clustering. Further,
158 skewed species abundances with a few dominant species could make it more difficult to sample rare species at

159 modest sequencing depth; however, because Mash estimates the similarity between two sets, slight stochastic
160 differences in observed abundances should not significantly affect the results relative to traditional OTU
161 approaches that are also subject to



162 *Figure 5. represents cluster similarities between each cluster, cluster similarity jaccard algorithm was used,*
163 *k=3 clusterings are shown.*



164

165 *Figure 6. represents cluster similarities between each cluster, cluster similarity jaccard algorithm was used,*

166 *k=4 clusterings are shown.*

167

ANOVA Results	MCQUITTY_CUTREE_K3		NMF_LOG_OTU_K3		MCQUITTY_CUTREE_K4		NMF_LOG_OTU_K4	
	P-value	Signif.	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								

168

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

178 McQuitty clustering has a .65 similarity with the Cardinal Direction, and similarly high similarities with other
179 environmentally determined groupings. We also see that both McQuitty and NMF have high p-values with some
180 environmental variables in anova, with Ph being particularly significant in both McQuitty and NMF, and to a
181 lesser extent Relative Air Humidity being significant as well, and that the sample similarities within these

182 groupings are high. This shows that OTU based methods and distance-based methods produce similar results, if
183 driven by slightly different environmental variables, and is getting at the underlying structure of the local
184 communities.

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

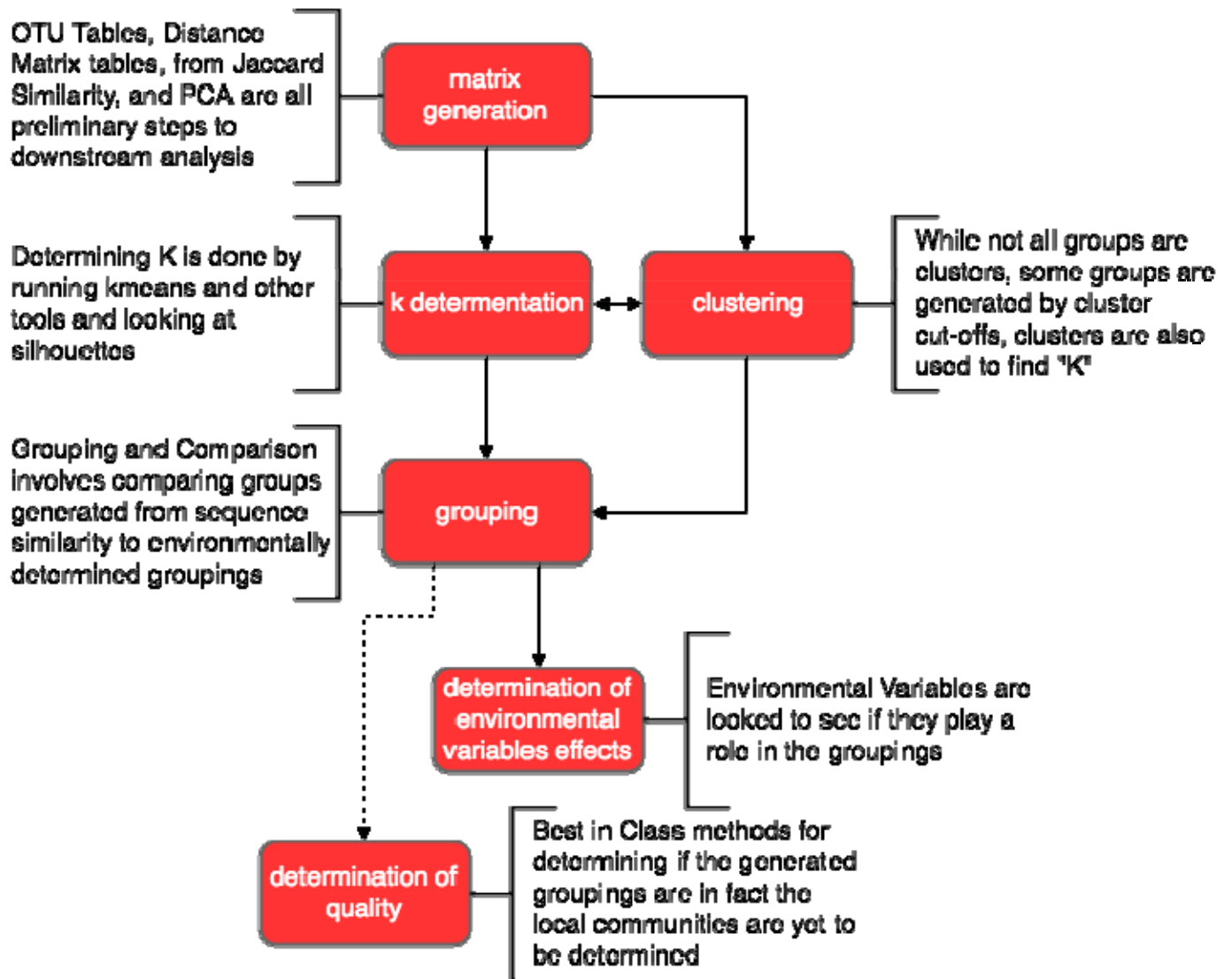
189 ***Discussion***

190 How communities are determined, or even how they are considered, is up for debate. Are communities
191 composed of nearly homogenous samples or are they composed of a mix of different kinds becomes an important
192 question that drives experimental design (Holyoak et al. 2005). If the expectation is that samples should be nearly
193 homogenous, determining communities algorithmically via sketch-based clustering is possible. In this study,
194 however, we observed that samples' clusters derived from MASH had low cluster to cluster similarities relative to
195 the derived clusters using OTUs/Mothur, even with quality trimmed data supplied to MASH. One likely
196 explanation is MASH-driven clustering is being affected by sequences not deemed as OTU sequences either
197 because of low coverage, contamination, insufficient length, or some combination therein in the read data
198 published previously. Even so, ANOVA analysis on the derived clusters showed that some measured
199 environmental variables were consistently significant while others depended on the clustering method used, with
200 MASH and NMF having the high associations. When considering NMF as an alternative this may make some
201 sense given that it is a factorization method focused separating two clusters, while traditional PCA would focus
202 on variables with the most divergence between samples. As such NMF factors can be intuitively understood and
203 can allow for an overlap in basis components (Gaujoux & Seoighe, 2010) making it a viable choice in
204 determining communities.

205 Finally, we believe that abundance derived *de novo* clusters are useful specifically because they group
206 samples without prior knowledge of geospatial or other overriding factors. This is powerful to assess associations
207 between species abundance, communities, and environmental variables (inc. geospatial) without requiring a more
208 complex statistical model.

209 ***Concluding Remarks***

210



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

212

213

214 As a general workflow (figure 9), after sample collection either OTUs abundances are generated, or sample to
215 sample distances are calculated by comparing their contained trimmed sequences. In the case of the sample to
216 sample distances a distance matrix is generated that can be clustered though hierarchical or other means, and in
217 the case of OTU abundances NMF or K-means is better suited. We calculated pairwise distance on both shared

218 sequences and on OTUs, and then clustered OTUs and shared sequences via K-means, and for shared sequences
219 diana clustering was also utilized, and for NMF was also utilized for OTU abundance. The groupings can then be
220 checked for the influence of independent variables, through a statistical model, in this case anova, which was run
221 on clusters, the 'anova' function from the R 'stats' package was used, and LogLik from the R 'stats' package was
222 used to compare Log-Likelihoods. Clusters were compared to each other using both RAND and Jaccard similarity
223 cluster evaluation methods, as well as a wilcox test (Hollander et al. 2013);(Bauer, 1972).

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

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

243 clusters.. And further analysis and figure analysis was done with the caret ([https://cran.r-](https://cran.r-project.org/package=caret)
244 [project.org/package=caret](https://cran.r-project.org/package=caret)), clusteval (<https://cran.r-project.org/package=clusteval>), cluster ([https://CRAN.R-](https://CRAN.R-project.org/package=cluster)
245 [project.org/package=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
246 packages.

249 Supplemental files ([https://github.com/status-five/Methods-in-Description-and-Validation-of-Local-](https://github.com/status-five/Methods-in-Description-and-Validation-of-Local-Metagenetic-Microbial-Communities/releases/tag/v1.0)
250 [Metagenetic-Microbial-Communities/releases/tag/v1.0](https://github.com/status-five/Methods-in-Description-and-Validation-of-Local-Metagenetic-Microbial-Communities/releases/tag/v1.0)) (Molik, 2018)

251 **References**

252

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

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

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

260 Broder, A.Z. On the resemblance and containment of documents. *Proceedings. Compression and*
261 *Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pp. 21–29. IEEE Comput. Soc.

262 Seung, H.S. & Lee, D.D. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*,
263 **401**, 788–791.

- 264 Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S. & Phillippy, A.M.
265 (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, **17**,
266 132.
- 267 Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J.A.,
268 Hoffman, J.M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C.,
269 Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J.E., Li, K., Kravitz, S., Heidelberg, J.F.,
270 Utterback, T., Rogers, Y.-H., Falcón, L.I., Souza, V., Bonilla-Rosso, G., Eguiarte, L.E., Karl, D.M.,
271 Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M.R.,
272 Strausberg, R.L., Neelson, K., Friedman, R., Frazier, M. & Venter, J.C. (2007). The Sorcerer II Global
273 Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific (N.A. Moran, Ed.).
274 *PLoS Biology*, **5**, e77.
- 275 Seshadri, R., Kravitz, S.A., Smarr, L., Gilna, P. & Frazier, M. (2007). CAMERA: A Community Resource
276 for Metagenomics. *PLoS Biology*, **5**, e75.
- 277 Hultman, J., Waldrop, M.P., Mackelprang, R., David, M.M., McFarland, J., Blazewicz, S.J., Harden, J.,
278 Turetsky, M.R., McGuire, A.D., Shah, M.B., VerBerkmoes, N.C., Lee, L.H., Mavrommatis, K. &
279 Jansson, J.K. (2015). Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes.
280 *Nature*, **521**, 208–212.
- 281 Gibbons, S.M. & Gilbert, J.A. (2015). Microbial diversity--exploration of natural ecosystems and
282 microbiomes. *Current opinion in genetics & development*, **35**, 66–72.
- 283 Gaujoux, R. & Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. *BMC*
284 *Bioinformatics*, **11**, 367.
- 285 Paatero, P. & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal
286 utilization of error estimates of data values. *Environmetrics*, **5**, 111–126.

- 287 Pascual-Montano, A., Carazo, J.M., Kochi, K., Lehmann, D. & Pascual-Marqui, R.D. (2006). Nonsmooth
288 nonnegative matrix factorization (nsNMF). *IEEE Transactions on Pattern Analysis and Machine*
289 *Intelligence*, **28**, 403–415.
- 290 Crits-Christoph, A., Robinson, C.K., Barnum, T., Fricke, W., Davila, A.F., Jedynak, B., McKay, C.P. &
291 DiRuggiero, J. (2013). Colonization patterns of soil microbial communities in the Atacama Desert.
292 *Microbiome*, **1**, 28.
- 293 Berman, A. & Plemmons, R. (1994). *Nonnegative matrices in the mathematical sciences*.
- 294 Ho, Yu-Jui , Naishitha Anaparthi, David Molik, Toby Aicher, Ami Patel, James Hicks, Molly G. Hammell.
295 (2017). *SAKE (Single-cell RNA-Seq Analysis and Klustering Evaluation) Identifies Markers of*
296 *Resistance to Targeted BRAF Inhibitors in Melanoma Cell Populations*, Preprint: bioRxiv,
- 297 Hartigan, J. & Wong, M. (1979). Algorithm AS 136: A k-means clustering algorithm. *Statistical Society.*
298 *Series C (Applied Statistics)*.
- 299 Forgey, E. (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classification.
300 *Biometrics*.
- 301 Broder, A. (1997). On the resemblance and containment of documents. *Compression and Complexity of*
302 *Sequences 1997*.
- 303 Struyf, A., Hubert, M. & Rousseeuw, P. (1997). Integrating robust clustering techniques in S-PLUS.
304 *Computational Statistics & Data*.
- 305 Kaufman, L. & Rousseeuw, P. (2009). *Finding groups in data: an introduction to cluster analysis*.
- 306 McQuitty, L.L. (1966). Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data.
307 *Educational and Psychological Measurement*, **26**, 825–831.

- 308 Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.
309 *Journal of Computational and Applied Mathematics*, **20**, 53–65.
- 310 Handl, J., Knowles, J. & Kell, D.B. (2005). Computational cluster validation in post-genomic data analysis.
311 *Bioinformatics*, **21**, 3201–3212.
- 312 DeLong, E.F. (2013). *Microbial metagenomics, metatranscriptomics, and metaproteomics*.
- 313 Everitt, B. (1974). Cluster analysis: An SSRC review of recent research.
- 314 Hartigan, J. (1975). Clustering algorithms.
- 315 Hollander, M., Wolfe, D. & Chicken, E. (2013). *Nonparametric statistical methods*.
- 316 Bauer, D. (1972). Constructing confidence sets using rank statistics. *Journal of the American Statistical*
317 *Association*.
- 318 Quince, C., Lanzén, A., Curtis, T.P., Davenport, R.J., Hall, N., Head, I.M., Read, L.F. & Sloan,
319 W.T. (2009). Accurate determination of microbial diversity from 454 pyrosequencing data.
320 *Nature Methods*, **6**, 639–641.
- 321 Molik , David. (2017). *status-five/Methods-in-Description-and-Validation-of-Local-Metagenetic-*
322 *Microbial-Communities: Initial Release*. Zenodo, 10.5281/zenodo.1164897.

323