Project report: Inferring interactions between microbial species from gene expression data

Janet B. Matsen
Department of Chemical Engineering
University of Washington
Seattle, WA 98195
jmatsen@uw.edu

Serena X. Liu
Department of Genome Sciences
University of Washington
Seattle, WA 98195
selenay@uw.edu

Abstract

Previous research has shown that for environments containing methane, methanotrophic (methane-utilizing) bacteria cooperate with methylotrophic (methanol-utilizing) bacteria in a species-dependent manner [1, 2]. We aim to better understand these partnerships by exploring data gathered from experiments using cultures of complex populations isolated from Lake Washington sediment. Our code is available at github.com/JanetMatsen/ML_meta-omics.

1 Introduction

Methane is a potent greenhouse gas. Changes in atmospheric methane levels have profound effects on the environment. In freshwater lake environments, dynamic cycling of methane is mediated by aerobic methanotrophic bacteria, which are capable of metabolizing methane and typically make up only a small fraction of the total population of organisms. Previous research has suggested that in environments containing methane, methanotrophic bacteria may act cooperatively with non-methanotrophic methylotrophic bacteria [1, 2]. Methylotrophic bacteria cannot metabolize methane directly, but can metabolize methanol, an intermediate metabolite of methane metabolism.

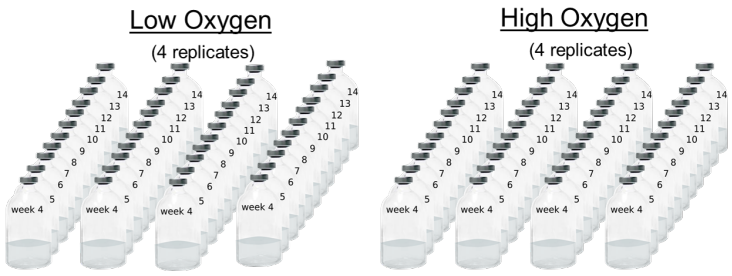
Cooperative behavior between methanotrophs and methylotrophs is species-dependent. However, the particular physiological interactions and constraints that determine which species may pair cooperatively with each other are not yet understood. Using data gathered on complex populations collected from the sediment of Lake Washington, we aim to investigate genetic factors that may contribute to the synergistic cooperation between methanotrophs and methylotrophs.

2 Data

2.1 Data collection

For this project, we worked with RNA-sequencing (RNA-seq) data collected from natural methane-oxidizing microbial communities grown in lab environments. Eight initial bottles were inoculated from natural sediment samples originating from Lake Washington; these communities contain a high level of microbial diversity. These initial samples were then cultured in the lab with methane and either high or low oxygen levels. The lab cultures were serially transferred (a portion of the culture transferred to a new bottle for continued growth) for 14 weeks, and separate metagenomic sample measurements were taken each week for eleven weeks, beginning at week 4 (Fig. 1b). Each serial transfer series thus has 11 sample measurements, resulting in 88 total samples.

(a) Sediment corer used to gather the Lake Washington samples



(b) Diagram of the experimental setup: 8 independent sediment samples were collected from Lake Washington. 4 were cultured in low oxygen conditions, and 4 in high oxygen conditions. Samples were serially transferred weekly, and metagenome RNA-seq data was collected for weeks 4-14 for each sample.

Figure 1: Equipment and experimental setup.

For each sample, RNA was collected and sequenced using the Illumina platform, and the resulting sequencing reads were mapped to 55 bacterial isolate genomes to obtain read counts for each gene (as a measure of relative gene expression). Relative species abundances were inferred from 16S DNA sequencing. Because there were some issues with aligning the RNA-sequencing (RNA-seq) data for 5 samples, our final data set contained 83 samples.

2.2 Data challenges & modeling considerations

As is common with biological experiments, we had a very small sample size (83) relative to the number of features ($\sim$200,000 genes across dozens of species). This limited our power for modeling and necessitated feature selection to reduce our feature space. These considerations guided both our data pre-processing steps and our choice of model, described below.

3 Methods

3.1 Data pre-processing

Prior to analyzing our data, we first applied a number of filtering steps aimed at removing redundant or uninformative features from our dataset. After splitting our data into two sets, one containing data for methanotroph genes and the other containing data for methylotroph genes, we aggregated the gene expression data within each set based on gene name. For example, if gene A were expressed in 15 different species, our original dataset would include 15 separate features corresponding to that gene. Since we were interested in broader community trends of gene expression, we collapsed these fifteen entries into a single entry for gene A that corresponded to the sum of reads for gene A across all species.

Next, we removed any genes in our dataset that had zero variance, since these features are not informative for probing differences between samples. We also filtered out genes which were annotated as either ”hypothetical” or having ”unknown function,” since these entries are a large source of noise and are uninformative for determining the gene functions important for observed trends. After performing all of these filtering steps, we ended up with a total of 4,563 methanotroph genes and 9,324 methylotroph genes (compared to the original 44,210 methanotroph and 149,137 methylotroph genes), reducing the total number of features by an order of magnitude. Finally, we normalized our expression data to have mean 0 and unit variance.

3.2 Exploratory analysis via PCA

In our preliminary analysis, we sought to get a sense for how our data were distributed, and how the different samples were related to each other. In particular, since our data was comprised of samples from 8 serial transfer series, we wanted to see how much our data depended on time. If the data were strongly time-dependent, it would be difficult to treat the individual samples as independent observations in downstream analysis.

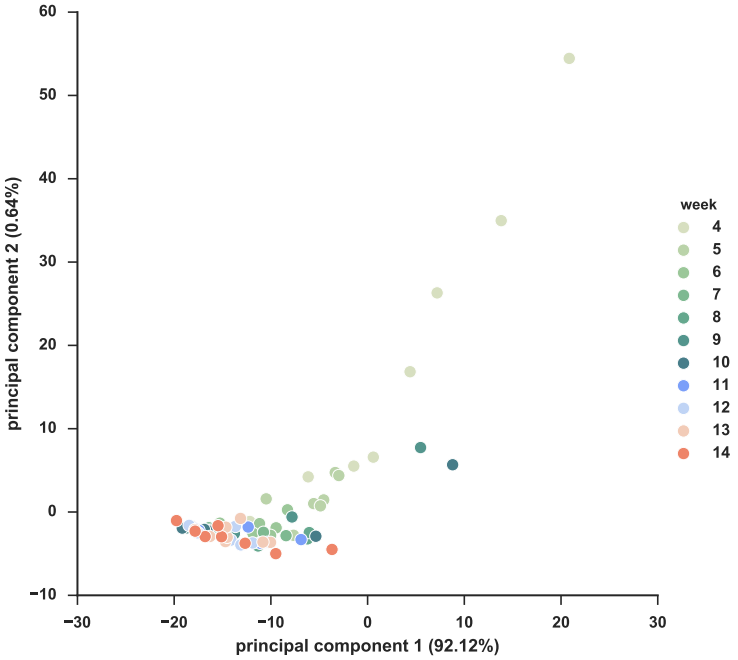


Figure 2: PCA analysis of gene expression data for samples collected from Lake Washington. Axes show the percent of variance explained by the first and second principal components. Individual sample points are colored by time collected after initial culture (in weeks).

Fig. 2 shows our data visualized in a two-dimensional space defined by the first and second principal components. From this analysis, we can see that while the data are somewhat correlated with time (in particular, the earlier timepoints are separated from the later timepoints). However, samples collected at the same timepoint from different culture series are generally tightly correlated with each other, suggesting that there were not significant batch effects between independent biological samples. This is important since our analysis implicitly assumes that each sample is an independent observation, and strong batch effects would have violated this assumption.

As it is, we did not explicitly address the time-dependency of our data in the analysis described below. However, methods to apply our chosen model (Canonical correlation analysis, or CCA) to multivariate time series data have been described in the literature [3], and such methods could be explored in future analyses of our data.

3.3 Sparse CCA

Canonical correlation analysis (CCA) is a classical statistics method dating back to 1936 [4] that has commonly been used to investigate multivariate relationships between groups of variables [5]. Given two sets of feature observations on the same set of samples, $\mathbf{X}_1$ and $\mathbf{X}_2$, where $\mathbf{X}_1$ is $n \times p_1$ and $\mathbf{X}_2$ is $n \times p_2$, CCA seeks $\mathbf{w}_1 \in \mathbb{R}^{p_1}$ and $\mathbf{w}_2 \in \mathbb{R}^{p_2}$ such that the correlation between $\mathbf{X}_1\mathbf{w}_1$ and $\mathbf{X}_2\mathbf{w}_2$ is maximized. That is, it finds

$$\max_{\mathbf{w}_1, \mathbf{w}_2} {\mathbf{w}_1}^\top {\mathbf{X}_1}^\top \mathbf{X}_2 \mathbf{w}_2$$

subject to the constraint ${\mathbf{w}_1}^\top {\mathbf{X}_1}^\top \mathbf{X}_1 \mathbf{w}_1 = {\mathbf{w}_2}^\top {\mathbf{X}_2}^\top \mathbf{X}_2 \mathbf{w}_2 = 1$.

CCA is useful for multivariate analysis and identifying relationships among complex data. In our case, we were interested in investigating the relationships between the relative gene expression profiles of methanotrophs and methylotrophs. We wished to identify genes that might be responsible for cooperation between (and thus co-occurrence) of microbes from these two species classes. Since

we wanted to look at the relationships between two groups of variables (methanotroph genes and methylotroph genes), CCA seemed like a good fit for our problem. Indeed, upon searching the literature, we found that CCA had been successfully applied to similar problems in ecology and genomics [6, 7].

However, even after our data pre-processing steps, we still had thousands of gene features for methanotrophs and methylotrophs, and only 83 samples. Standard CCA is not well defined when $n < \min(p_1, p_2)$, so we used a L1-penalized variant of CCA described by Witten et al. [5]. This "sparse CCA" algorithm seeks to find

$$\max_{\mathbf{w}_1, \mathbf{w}_2} {\mathbf{w}_1}^\top {\mathbf{X}_1}^\top \mathbf{X}_2 \mathbf{w}_2$$

subject to the constraints $||\mathbf{w}_1||^2 \leq 1, \quad ||\mathbf{w}_2||^2 \leq 1, \quad ||\mathbf{w}_1||_1 \leq c_1, \quad ||\mathbf{w}_2||_1 \leq c_2$, where c_1 and c_2 are the regularization parameters for $\mathbf{w}_1$ and $\mathbf{w}_2$, respectively. Having small c_1 and c_2 results in sparse $\mathbf{w}_1$ and $\mathbf{w}_2$, where many elements of $\mathbf{w}_1$ and $\mathbf{w}_2$ are zero.

For our analysis, we used the implementation of sparse CCA provided by Witten et al. in the R package PMA [5]. An R script was written that runs CCA for a single set of data and specified penalty values. This was called from Python using the subprocess module during cross-validation and for training the final model on the full dataset.

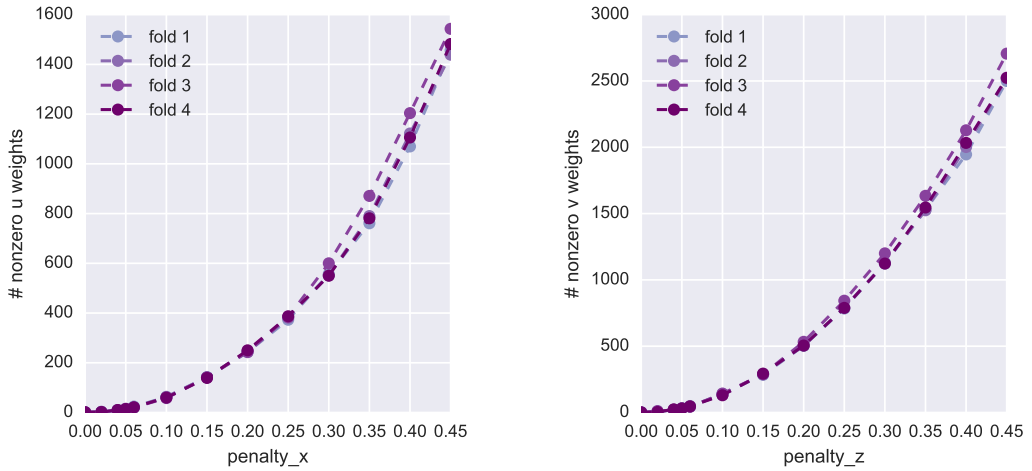
4 Results

4.1 Cross-validation to select optimal hyperparameters

As the first step in our sparse CCA analysis, we aimed to determine the optimal regularization parameter values (c_1, c_2). Since we were data limited with respect to the number of samples, we chose to perform 4-fold cross-validation instead of holding out part of our data as an untouched test set. To identify the best regularization parameters for our dataset (while still avoiding overfitting to our data), we tested a grid of values for c_1 ("penalty_x", the regularization parameter for methanotroph genes) and c_2 ("penalty_z", the regularization parameter for methylotroph genes). In the PMA package, the two sets of feature observations were called $\mathbf{X}$ and $\mathbf{Z}$ ($\mathbf{X}_1$ and $\mathbf{X}_2$ from before), and the corresponding weight vectors were called $\mathbf{u}$ and $\mathbf{v}$ ($\mathbf{w}_1$ and $\mathbf{w}_2$ from before).

For each c_1, c_2 value pair, we performed 4-fold cross-validation (training on $\frac{3}{4}$ of the data, and validating on the remaining $\frac{1}{4}$, then repeating for all four possible combinations of folds) and recorded the mean validation fold correlation between $\mathbf{Xu}$ (methanotroph gene projections) and $\mathbf{Zv}$ (methylotroph gene projections) for that pair of regularization weights.

As expected from the theory, smaller regularization weight parameters correspond to more stringent regularization and result in sparser weight vectors (Fig. 3).



(a) Number of nonzero elements in $\mathbf{u}$ vs. "penalty_x" (regularization weight for $\mathbf{u}$)

(b) Number of nonzero elements in $\mathbf{v}$ vs. "penalty_z" (regularization weight for $\mathbf{v}$)

Figure 3: Weight vector sparsity increases as regularization weights decrease in magnitude.

We also observed that the effects of the regularization parameters c_1 and c_2 on the sparsity of $\mathbf{u}$ and $\mathbf{v}$ appeared to be mostly independent of each other (Fig. 4). That is, a given c_1 penalty corresponds to a certain level of sparsity in $\mathbf{u}$ regardless of the paired c_2 value, and vice versa.

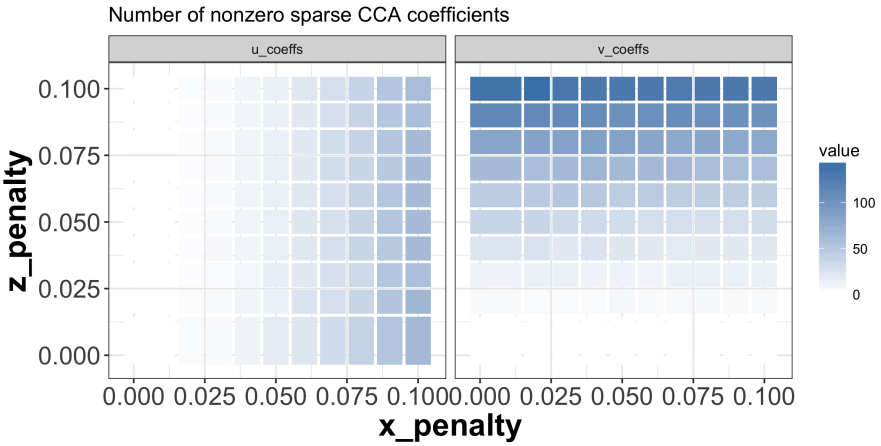


Figure 4: Demonstration of (mostly) independent effect of penalty_x and penalty_z on the number of nonzero coefficients in the u and v vectors. Darker blue indicates more nonzero weights.

After completing this process for each pair of regularization values, we then selected the c_1, c_2 pair with the highest validation fold correlation value to be our regularization weights for our full analysis on the whole dataset. Plots of training correlation and validation correlation against the regularization weights are shown in Fig. 5. As seen in Fig. 5c and Fig. 5d, we get the highest mean validation fold correlation at $c_1 = 0.15$ and $c_2 = 0.15$.

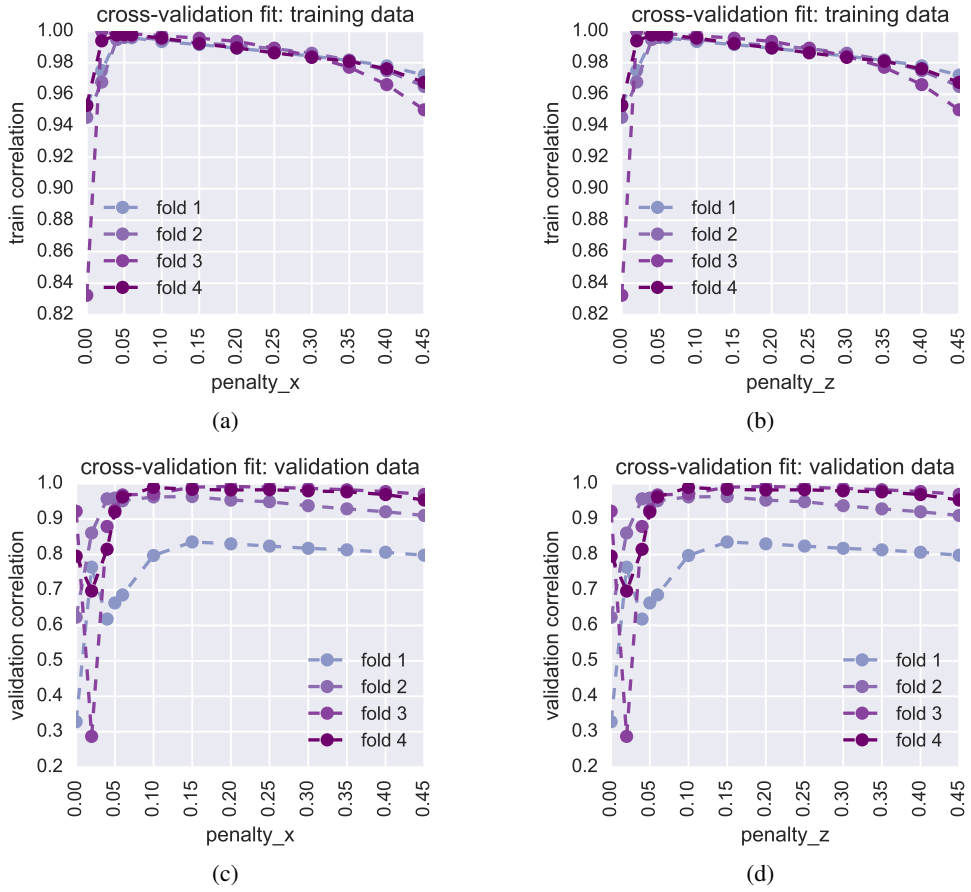


Figure 5: Hyperparameter tuning for the sparse CCA regularization weights using 4-fold cross-validation. Plots of training correlation between $\mathbf{Xu}$ and $\mathbf{Zv}$ against "penalty_x" (a) and "penalty_z" (b), and plots of validation correlation against "penalty_x" (c) and "penalty_z" (d).

4.2 Preliminary identification of genes important for methanotroph-methylotroph interaction

Performing sparse CCA on the full dataset using regularization weights $c_1 = c_2 = 0.15$ resulted in weight vectors with 156 nonzero elements for $\mathbf{u}$ and 317 nonzero elements for $\mathbf{v}$. Of these, we further investigated the top 10 genes in each set with the largest magnitude weights. Individual gene weights and literature-annotated functions are summarized in Table 1 and Table 2.

Table 1: Top 10 genes for methanotrophs (largest magnitude).

Gene	CCA weight	Description
Protein-disulfide isomerase	-0.195800	catalyzes the formation and breakage of disulfide bonds between cysteine residues within proteins as they fold
bacterial peptide chain release factor 1 (bRF-1)	-0.193264	directs the termination of translation at specific codons
Mono-oxygenase ydhR	-0.181711	monooxygenase that may be used for metabolism of aeromatic compounds
Right handed beta helix region	-0.178729	N/A (DNA motif, not a gene)
Predicted glycosyltransferases	-0.168887	biosynthesis of disaccharides, oligosaccharides and polysaccharides
Nitric oxide reductase subunit C	-0.164595	participates in nitrogen metabolism and in the microbial defense against nitric oxide toxicity.
Soluble methane monooxygenase reductase apoprotein	-0.158424	oxidation of the C-H bond in methane
Quinone oxidoreductase, YhdH/YhfP family	-0.154480	oxidoreductase for quinones (derivatives of aromatic compounds)
Sulfate adenylyltransferase subunit 2	-0.150052	first intracellular reaction of sulfate assimilation
SRSO17 transposase	-0.148082	enzyme transposition, the movement of genes from one position on the chromosome to another

Table 2: Top 10 genes for methylotrophs (largest magnitude).

Gene	CCA weight	Description
YfhO	-0.145618	transmembrane protein of unknown function
precorrin-6A reductase	-0.142571	oxidoreductase protein, important for porphyrin and chlorophyll metabolism
Acetyl-CoA carboxylase	-0.136985	biotin-dependent enzyme, important for fatty acid biosynthesis
Uncharacterized protein involved in exopolysaccharide biosynthesis	-0.129443	transmembrane protein, exact function unclear
amylovoran biosynthesis glycosyltransferase AmsB	-0.127814	involved in biosynthesis of amylovoran (virulence factor); also functions as a glycosyl transferase
Glycine cleavage system T protein (aminomethyltransferase)	-0.125660	catalyzes the reversible oxidation of glycine
Ectoine hydroxylase-related dioxygenase, phytanoyl-CoA dioxygenase (PhyH) family	-0.125171	involved in biosynthesis of ectoine (compound important for handling extreme osmotic stress)
dimethylhistidine N-methyltransferase	-0.124951	enzyme (catalyzes transfer of a methyl group between amino acid derivatives
N-hydroxyarylamine O-acetyltransferase	-0.116586	catalyzes acetyl-CoA-dependent N-acetylation of aromatic amines
ATP-, maltotriose- and DNA-dependent transcriptional regulator MalT	-0.111049	transcription factor for unknown gene targets

5 Discussion & future directions

Previous research has shown that in nature, methane-utilizing methanotrophs have significant interactions with methanol-utilizing methylotrophs. We applied sparse CCA to RNA-seq data collected from complex microbial populations in order to identify influential genes expressed by methanotrophs and methylotrophs, and the effects of these genes on the expression profiles of the opposite group. The top genes (Tables 1, 2) included a mixture of genes linked to methane metabolism (monooxygenase) and methanol metabolism. There are also many genes tied to other functions, including biosynthesis. It is likely that some of the top genes predicted correspond to differences in oxygen conditions across samples, a variable we were not able to account for in the scope of this work.

With CCA, caution must be taken to ensure that the identified inter-group correlations are not primarily driven by rare (but tightly correlated) features. In the context of this dataset, these features would be genes with infrequent and low magnitude expression levels, but which co-occurred in the dataset with very high correlation. One potential solution for mitigating the impact of such genes is to add a pseudocount to all of the gene expression data. These pseudocounts disproportionately affect the correlations of rarer genes, so it lessens the impact of the rare but highly correlated outliers.

Given additional time, we would update our CCA analysis code to reduce the ability for rare noisy features to receive nonzero CCA weights. First we would include summaries of the expression levels for each gene. This would allow systematic removal of features below a (tuneable) threshold of expression. Second, we would add an argument that allows pseudo-counts to be added to the entire dataset. The magnitude of the added pseudo-counts would need to be tuned with consideration for the overall distribution of expression levels in the samples.

Acknowledgments

We would like to thank David Beck for providing us with the RNA-seq dataset, and for his guidance and support throughout this project. We would also like to acknowledge Ludmila Chistoserdova for some biological discussions and Maria Hernandez for the wet lab efforts that lead to our data.

References

1. Oshkin, I. Y. *et al.* Methane-fed microbial microcosms show differential community dynamics and pinpoint taxa involved in communal response. *The ISME Journal* **9,** 1119–1129 (2015).
2. Beck, D. A. *et al.* A metagenomic insight into freshwater methane-utilizing communities and evidence for cooperation between the Methylococcaceae and the Methylophilaceae. *PeerJ* **1,** e23 (2013).
3. Min, W. & Tsay, R. S. On canonical analysis of multivariate time series. *Statistica Sinica,* 303–323 (2005).
4. Hotelling, H. Relations between two sets of variates. *Biometrika* **28,** 321–377 (1936).
5. Witten, D. M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics,* kxp008 (2009).
6. Witten, D. M. & Tibshirani, R. J. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology* **8,** 1–27 (2009).
7. Ramette, A. Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology* **62,** 142–160 (2007).