STATISTICAL APPROACHES FOR SPATIAL-TEMPORAL DYNAMICS OF MICROBIAL COMMUNITIES IN THE RED SEA

by

Paul Bjorndahl

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
December 2019

I dedicate this thesis to my husband John Kyle Sleigh for his continuous support.

TABLE OF CONTENTS

LIST OF TABLES iv
LIST OF FIGURES v
ABSTRACT vi
LIST OF ABBREVIATIONS USED vii
ACKNOWLEDGEMENT viii
CHAPTER 1 INTRODUCTION 1
1.1 Background, Motivation and Data 1
1.2 Statistical Frameworks 4
CHAPTER 2 REVIEW OF INFERENCE FRAMEWORKS 8
2.1 Bayesian Inference of Microbial Communities 8
2.2 Non-negative Matrix Factorization 11
2.3 Structural Topic Models 13
2.3.1 Variational Inference 15
2.4 Subsampling Ranking and Forward Selection 16
CHAPTER 3 RESULTS AND INSIGHTS 18
3.1.1 Determining the Number of Assemblages for BioMiCo 18
3.1.2 Determining the Number of Assemblages for NMF 21
3.1.3 Determining the Number of Assemblages for STM 25
3.2 Comparative Analysis of Assemblages 30
3.3 Prediction 40
3.4 Assemblage Characteristics 47
3.4.1 2015 Assemblage Associations 52
3.4.2 2016 Assemblage Associations 58
CHAPTER 4 CONCLUSION AND FUTURE WORK 62
BIBLIOGRAPHY 65

LIST OF TABLES

Table 3.1 BioMiCo Classification Testing Prediction..... 42

Table 3.2 Supervised NMF Prediction..... 44

Table 3.3 All Taxa Prediction..... 44

Table 3.4 SuRF Taxa Prediction..... 45

Table 3.5 SuRF and Predominant Selected Taxa..... 46

LIST OF FIGURES

Figure 3.1	BioMiCo assemblage distributions.....	21
Figure 3.2	Residual sum of squares (RSS) vs number of assemblages.....	22
Figure 3.3	Explained variance vs number of assemblages (K values).....	23
Figure 3.4	Cophenetic correlation coefficients vs K value.....	24
Figure 3.5	Diagnostic plots for 2015 STM.....	28
Figure 3.6	Diagnostic plots for 2016 STM.....	29
Figure 3.7	Diagnostic plots for STM with cyanobacteria.....	30
Figure 3.8	Bray-Curtis dissimilarity among 2015 samples.....	31
Figure 3.9	Bray-Curtis dissimilarity among 2016 samples.....	32
Figure 3.10	Bray-Curtis dissimilarity of 2015 samples labelled by assemblage	34
Figure 3.11	Bray-Curtis dissimilarity of 2016 samples labelled by assemblage	36
Figure 3.12	K-means clusters of 2015 assemblage weights.....	37
Figure 3.13	K-means clusters of 2016 assemblage weights.....	38
Figure 3.14	Hierarchical clustering of seasonal assemblages.....	39
Figure 3.15	Heterotrophic and cyanobacteria profiles.....	41
Figure 3.16	2015 NMF assemblage succession.....	48
Figure 3.17	2015 STM assemblage succession.....	49
Figure 3.18	2016 NMF assemblage succession.....	50
Figure 3.19	2016 STM assemblage succession.....	51
Figure 3.20	Correlations for 2015 environmental covariates.....	57
Figure 3.21	Correlations for 2015 pathways.....	58
Figure 3.22	Correlations for 2016 environmental covariates.....	60
Figure 2.23	Correlations plots for 2016 pathways.....	61

ABSTRACT

Changing environmental factors impact the biological structure and function of marine microbial communities. There is a need for modelling samples of microbial communities as a mixture of sub-communities. Sub-communities are modelled to be consistent with the observed distribution and abundance of taxa because (i) there is extensive among-sample variation in microbial abundances, and (ii) there are important ecosystem processes that appear to be carried out by the collective activities of microbial species. Three statistical frameworks are applied to identify a robust consensus of sub-community structure from the Gulf of Aqaba, Red Sea. Assemblages of taxa are derived from different methods to capture sub-community spatial-temporal dynamics and can be used to train predictive models of important environmental categories. Assemblages and their test predictions are compared to each other and to results from variable selection of individual microbial taxa. Assemblages also demonstrate characteristics associated with distinct physiochemical features of the ecosystem.

In the first chapter, a number of motivating questions are addressed for densely sampled spatial and temporal microbial communities. A real marine microbiome dataset from the Gulf of Aqaba (Station A, Red Sea) is presented, including how it is sampled by depth and time in the environment. Two Bayesian approaches and one maximum-likelihood method are introduced as frameworks to identify a robust consensus of heterotrophic community structure.

In the second chapter, details about how assemblages of co-occurring taxa are inferred from different methods are reviewed. Predictive models can be used with assemblage proportions learned in a training year to classify covariate categories in a test year. Another variable selection method is described for comparison of its predictive accuracy to the assemblage methods.

In the third chapter, statistical approaches are discussed and evaluated for determining the number of assemblages in different models. Assemblage distributions are shown to capture latent time and depth dynamics. Dominant assemblages are examined in distinct environments by comparing their compositions across methods to show that spatial-temporal dynamics produce similar subcommunities. The predictive accuracies of assemblage mixture weights for classifying environment features are compared. Assemblages are also shown to reproduce correlations with other biotic and abiotic covariate vectors over separate years. Covariates have ecological interpretations that characterize assemblages as predominantly associated with features of the ecosystem like water column stratification, seasonality and algal blooms. Further, correlation analysis is conducted to gather evidence for a functional profile of assemblages based on significant associations with important metabolic pathways. The results discussed reinforce the characterization of assemblages with specific traits.

Conclusions and directions for future work are presented in the fourth chapter

LIST OF ABBREVIATIONS USED

Ammonia Oxidizing Archaea (AOA)

Amplicon Sequence Variants (ASVs)

Bayesian Inference of Microbial Communities (BioMiCo)

Chlorophyll Fluorescence (Chl Fluor)

Generalized Linear Model (GLM)

Kullback-Leibler Divergence (KL)

Non-negative Matrix Factorization (NMF)

Prochlorococcus (Pro)

Random Forest (RF)

Structural Topic Models (STM)

Subsampling Ranking and Forward Selection (SuRF)

Synechococcus (Syn)

Temperature (Temp)

Total Organic Nitrogen (TON)

ACKNOWLEDGEMENT

I wish to express my gratitude to Joseph P. Bielawski, Hong Gu and Katherine A. Dunn. Their guidance and encouragement made this work possible. I want to thank Toby Kenney and Andrew Irwin, the readers of this thesis, for their valuable comments and suggestions for improvement. I would also like to thank my friend Hayley Mills for being a source of extreme positive reinforcement.

CHAPTER 1 INTRODUCTION

1.1 BACKGROUND, MOTIVATION AND DATA

The relationship between marine microorganisms and their environment is complex and dynamic. Variation in microbial communities in response to one or more environmental factors may drive critical marine ecosystem processes like nitrification (Zeglin 2015). Conversely, microbial taxonomic and functional dynamics that are conserved over longitudinal studies of the water column may mediate changes or predict patterns along physiochemical gradients in space and time (Faust et al. 2015).

We apply statistical models to investigate such microbial community dynamics from the Red Sea over two years. The Red Sea is a seawater inlet of the Indian Ocean ideally suited for investigating complex adaptations to changing conditions in the ocean due to its unusually high temperature, salinity, solar irradiance and anthropogenic pressures (Haroon et al. 2016). Thompson et al. (2017) have analyzed how microbial taxonomic diversity and functional variation across environmental gradients is explained by physiochemical parameters. We shift focus to community relationships with environmental covariates and metagenomic functions.

Classical statistical methods have limited value and are often inappropriate for modelling communities from microbiome samples. Samples are highly complex, being comprised of overlapping mixtures of species from different communities. The number of variables, frequently referred to as species or taxa, is extremely large (often thousands or tens of thousands); and the matrices of abundance information are sparse. Previous work has identified indicator species or broad taxonomic groups with an effect on disease states in the human gut microbiome (e.g. pathogens). However, in a marine setting we

want to make inferences about whole community structures that transition over time and space (depths). Models that can learn different subcommunity dynamics are a better fit for predicting environmental features of interest than just isolated species. This approach allows for subcommunity structures to contribute to a model for the inter-dependencies between the microbiome and stable or changing environmental factors.

The data were collected at Station A (29° 28′ N, 34° 55′ E), which is an open-water site in the middle of the Gulf of Aqaba, Red Sea. Sequence data for the taxonomic marker gene (16S) was obtained for 106 seawater samples from 2015, and for 136 seawater samples from 2016. Those samples were collected from multiple depths (0, 20 40, 60, 80, 100, 140, 200 and 400 m) approximately every two weeks between March and June, with more limited sampling in February and September. Environmental DNA was extracted from each water sample. Three samples did not have sufficient DNA to obtain 16S sequence data; all three were obtained in 2015 at 400 m. Approximately 450bp of the V4-V5 hypervariable region of 16S was sequenced for each of the remaining samples. Following extensive processing of the raw sequence reads (completed by Dr. Katherine A. Dunn), the data were used to infer the presence of species-level taxonomic units.

For each sample, paired-end sequencing reads were assembled and validated according to size. Sequences were screened for quality, and chimeric sequences were removed. Open reference picking was performed on all remaining sequences using the QIIME pipeline (Caporaso et al. 2010), and utilizing sumaclust for de novo OTU picking. Three samples had less than 10,000 reads after all of the processing; two collected in 2015 at 400m and one collected in 2016 at 100 m. The remaining samples (235) had at least 10,278 reads and were retained for further analyses (maximum 42,692, mean 18,042

and median 17,818). Sequence reads were assigned taxonomic status, where possible, from kingdom to species. All of this processing work was done by Dr. Katherine A. Dunn.

Overall, for samples were taken over the two years 2015 and 2016, there were nine different depths and 14 or 16 time points for each year respectively. In 2015 there were a total of 103 samples and 2414 heterotrophic taxa with non-zero total abundance in all samples (Amplicon Sequence Variants or ASVs). In 2016 there were a total of 136 samples and 2474 heterotrophic taxa with non-zero total abundance in all samples. In addition to time and depth, the other environmental covariates that were measured concurrently were: *Synechococcus* and *Prochlorococcus* (cyanobacteria), nitrite (NO_2), nitrate (NO_3), phosphate (PO_4), total organic nitrogen (TON), water density, temperature, pressure, salinity, irradiance, chlorophyll fluorescence and oxygen. Cyanobacteria cell concentrations were measured using a flow cytometer (within 1 day of collection). Metagenomic data was also sequenced for a subset of samples, and it provides additional information about the abundance of functional gene families and metabolic pathways in the environment where those samples were collected.

There are many studies in marine microbial ecology about phototrophs like cyanobacteria (Sieradzki et al. 2018). So we are motivated to model heterotrophic communities to gain insight into the interactions between connected groups of heterotrophs and autotrophic microorganisms. The hypothesis is that heterotrophic subcommunity compositional and functional differences are influenced by, and predictive of, spatial-temporal habitats and ecosystem-scale processes like cyanobacteria blooms (Ren et al. 2017). Addressing these broad biological objectives requires that we address

the following range of research questions, each with unique statistical issues: 1) Can we reduce the dimensionality of the taxa variables down to latent subcommunity structures? 2) Do inferred subcommunities really represent biologically meaningful features and do they capture inter-dependencies in the environment? 3) How do underlying heterotrophic communities change over seasons and in response to the water column stability/variability at different depths? 4) How are subcommunities associated with other biotic and abiotic factors; and what are their relative contributions before, during and after cyanobacteria blooms? 5) Do specific subcommunity dynamics improve predictability of states of interest like seasons, depth or blooms?

1.2 Statistical Frameworks

Microbial taxonomic units are typically inferred as species- or stain-level units via the sequencing of DNA amplicons (e.g., 16S gene) from environmental samples of DNA (Callahan et al. 2017). This method is employed because it is fast, relatively inexpensive, and the majority of microbial diversity is uncultivable in the laboratory (Callahan et al. 2017). Sets of such co-occurring taxa (amplicon sequence variants, or ASVs) are then inferred to belong to community structures, called assemblages, by one of several different statistical methods. As opposed to looking at individual taxa, the approach of modelling assemblages of ASVs gives us a working hypothesis: Can we treat assemblages resolved from statistical models as subcommunities? Assemblages have the advantage of providing variable selection in the taxa space, and simplifying the relationship between conditionally rare or abundant taxa and ecological features (Logares et al. 2015). Functional trait-based associations, as derived from metagenomic data, offer

the possibility of interpreting microbial assemblages in terms of functionally-coherent sub-communities (Boon et al. 2014, Webb et al. 2010).

With this in mind, I seek to establish a consensus of ASV assemblage structure and environmental distribution across three diverse methodologies. I review each method and assess differences in their inference of assemblages. All three methods have previously been used on simulated and real-world data sets. They are applied to densely sampled spatial-temporal data to compare how similarly they fit the data. First, BioMiCo (Shafiei et al. 2015) is a Bayesian inference method in which samples are modelled by a hierarchical mixture of multinomial distributions with Dirichlet priors applied to the parameters of the distributions at each level. The model is supervised with labels for one or more environments (*e.g.*, depth or season) and trained on a set of samples to: (*i*) learn how to explain and differentiate environments through its mixture of various assemblages and (*ii*) assign appropriate environment values to new test samples of unknown label. Based on the mixture weights learned from the training samples, BioMiCo computes the posterior probability that a test sample originated from a microbiome belonging to any of the label values that the model was trained on.

Second, non-negative matrix factorization (NMF) is used to split a data matrix of samples by taxa counts into the product of two matrices so that each column of one matrix describes an assemblage of ASVs. The columns of the other matrix contain linear coefficients for each assemblage corresponding to each sample. NMF can likewise be supervised by class labels (Cai et al. 2017) and the assemblage distributions over ASVs used to predict the label of a new sample according to its microbial composition.

Third, another hierarchical Bayesian inference method, based on the structured topic model (STM), is applied to microbiome data such that community-level ASV content and prevalence is modeled in place of document-level topics (Roberts et al. 2016). STM parameterizes its prior distributions with continuous or discrete covariate information from sample environments. STM leverages such metadata to improve its posterior probability estimates and allows for covariance among assemblages. STM is computationally much faster than BioMiCo and about the same speed as NMF because it does not use an MCMC process to sample from the posterior distribution. Instead it approximates the posterior through an optimization technique called variational inference. However, STM is not currently able to model the assemblage weights of a test set of ASVs given the assemblage distribution from a model fitted on a training set – STM is in principle an unsupervised method.

The number of and composition of assemblages is both a research objective and a critical aspect of each method; thus, statistical insights into how to determine the number of assemblages are discussed. Assemblage distributions are characterized by the mixture weights (proportions) of ASVs. In the case of BioMiCo and STM these are resolved as the posterior probability distribution of the ASVs for a given assemblage. As these are hierarchical models, each sample has an inferred assemblage distribution (also resolved as posterior probabilities under BioMiCo and STM). The empirical success of using such distributions to accurately classify samples (*e.g.*, Shafiei et al. 2015, Cai et al. 2017) supports the interpretation of assemblages as natural subcommunities. This study strengthens this interpretation with the novel finding that assemblages inferred by using

different methods are more closely clustered together when they are from the same environment.

A primary objective of this study is to investigate how community structure (as inferred from assemblage mixture weights) is associated with covariate information about the ecosystem. Biologically meaningful associations are tested via the ability of off-the-shelf linear and non-linear algorithms to accurately predict feature classes of season, depth and cyanobacterial blooms. I also test how the assemblage proportions assigned for each sample are associated with covariate information about the ecosystem by performing regressions and correlation analysis. In addition to taxonomic data processed as ASV counts, functional profiling was performed using the HUMAnN2 pipeline by Dr. Katherine A. Dunn (Franzosa et al. 2018). Combinations gene fragments present in samples can be mapped to known enzyme-encoding genes to infer abundances of metabolic pathways. Review of the literature provided insight into which metabolic pathways from metagenomes were especially important for analysis of marine environments (Thompson et al. 2017). I conducted Spearman rank-based and robust regression with relevant pathway abundances to discover any strong correlations with important functional traits. The results are summarized and used to build evidence for assemblage associations.

CHAPTER 2 REVIEW OF INFERENCE FRAMEWORKS

2.1 BAYESIAN INFERENCE OF MICROBIAL COMMUNITIES (BIOMICO)

BioMiCo developed by Shafiei et al. (2015) builds a predictive model of latent subcommunity mixtures in distinct environments. Samples represent K pre-specified environmental communities, which are modeled as mixtures of L microbial assemblages. Assemblages are in turn a mixture of the ASVs observed in an environment. Only fixed environment labels and ASV counts within samples are observed. The model training phase learns the ASV contributions to assemblages, and the assemblage contributions to environmental communities as two latent levels of community structure. The relative contribution of the k^{th} environment to the n^{th} sample is modeled through the latent variable $\pi_{nk} \sim Dirichlet(\alpha_\pi)$. The relative contribution of each of L assemblages to environment k is $\theta_{kl} \sim Dirichlet(\alpha_\theta)$ for k = 1 ⋯ K. Each assemblage is composed of a mixture of T different ASVs. The relative contribution of ASV i to assemblage l is $\phi_{li} \sim Dirichlet(\alpha_\phi)$ for l = 1⋯ L. Symmetric Dirichlet priors are used because there is no knowledge to favour a particular ASV, assemblage or environment. Sparse Dirichlet priors are employed to minimize variance and maximize interpretability of the posterior distributions (Shafiei et al. 2015). The hyper-parameters for the priors (α_π, α_θ and α_ϕ) are given initial values then learned from a Metropolis-within-Gibbs sampling scheme. The community structure of samples is not known, so the prior variables are inferred from the data.

Training a model starts with a matrix of samples by ASV counts that represents the distribution of ASVs in each sample. Let W_{ni} be the observed data of ASV i in sample n. The environment and assemblage assignments for ASV i in sample n are denoted by X_{ni} and Z_{ni}, respectively. The distributions of assemblage (Z) and environment (X) assignments for each ASV given the data and priors represent the mixing of ASVs in assemblages and the mixing of assemblages in environments. BioMiCo uses Gibbs sampling to draw samples from the posterior distribution of Z and X. For each ASV in each sample, it draws the assemblage and environment assignments (Z_{ni} and X_{ni}, respectively) of this ASV given the current assemblage and environment assignments of all the other ASVs in all samples except the ith ASV in the nth sample, denoted by Z_{-ni} and X_{-ni}. The conditional distribution of interest is given by:

$$P(X_{ni} = k, Z_{ni} = l \mid X_{-ni}, Z_{-ni}, W, \alpha_\pi, \alpha_\theta, \alpha_\phi) = \frac{\alpha_\pi + C_n^k}{\sum_k(\alpha_\pi + C_n^k)} \times \frac{\alpha_\theta + C_k^l}{\sum_l(\alpha_\theta + C_k^l)} \times \frac{\alpha_\phi + C_{W_{ni}}^l}{\sum_W(\alpha_\phi + C_{W_{ni}}^l)} \quad (2.1)$$

Where the $C_{W_{ni}}^l$ term is the number of times ASV W_{ni} is assigned to the lth assemblage. C_k^l is the number of times an ASV in the kth environment factor is assigned to the lth assemblage. C_n^k is the number of times an ASV in the nth sample is assigned to the kth environment factor. These values are then normalized and used to draw new assignment values for X_{ni} and Z_{ni} which are immediately updated for use in the next iteration. The Gibbs sampler goes through the data ASV by ASV, and reassigns each ASV to an assemblage using the above posterior probability equation. In a true model of the community, the posterior distribution can be interpreted as iteratively assigning ASVs greater probability in assemblages where they are more common. Likewise assemblages

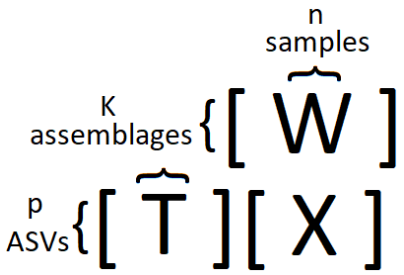
are iteratively assigned greater probability in environments where they are more common.

Statistical validation of the model is conducted to estimate generalization error for new data sets. Samples are divided into training and testing sets. The model is applied to the training set supplied with distinct labels for a feature of the environment like season or depth. In the training phase, the mixture weights for ASVs within assemblages and assemblages within environment factor values shared across multiple samples are obtained. In the testing phase the objective is to sample the posterior distribution of environment assignments X_{test} given the observed ASV distribution of the test data, W_{test}. The trained model also gives us the environment and assemblage assignments for the training ASV data. The posterior distribution of X_{test} and Z_{test} is sampled jointly, that is: $P(X_{test}, Z_{test} \mid X_{train}, Z_{train}, W_{test}, \alpha_{\phi}, \alpha_{\pi}, \alpha_{\theta})$ and marginalized over the assemblage assignments to obtain the posterior probability of each environment factor assignment: $P(X_{test} \mid X_{train}, Z_{train}, W_{test}, \alpha_{\phi}, \alpha_{\pi}, \alpha_{\theta})$.

In the testing phase, Gibbs sampling is used similarly to the training process. However, it is not necessary to iterate over the samples in the training set. The count variables C_w^l and C_k^l from the training phase are carried forward so it is not necessary to run the MCMC for as many iterations. The model predicts the environment factor contributions to each test sample and then each sample can be classified by discrete assignment to the environment that has its maximum posterior probability. When validating the model the observed environment labels for each test sample are known. Prediction accuracy is measured as the percent of the factor values that are correctly predicted for test samples.

2.2 Non-negative Matrix Factorization (NMF)

NMF factors a matrix into the product of two matrices of smaller dimension such that all entries are non-negative. It has been used for image recognition, signal processing and computational biology because the non-negativity constraint allow for decomposition into additive parts (Lee et al. 1999, Brunet et al. 2004, Gaujoux et al. 2010). Microbial abundance data are counts or proportions which are naturally non-negative. Although methodologically different from BioMiCo, the motivation for using NMF is likewise to find a parts based representation (mixing weights) of the sampled communities from different environments. More formally, given a non-negative p×n matrix X, X is approximated by TW, where T is a non-negative p×k matrix referred to as the type (assemblage) matrix and W is a non-negative k×n weight matrix. Each column of X is approximated by a non-negative linear combination of the columns of T. k is the number of assemblages.



Choosing k such that (p + n) × k << np, reduces the dimensionality of the ASV (taxa) space significantly. Each column in T describes an assemblage (composition of ASVs) and each column in W contains the linear coefficients for the corresponding sample columns in X.

The community within a sample is thus approximated by a mixture of the assemblages. The key idea is that elements of X are modelled as independent Poisson

observations given their mean in the matrix TW. ASV count data is treated as a Poisson sample from a weighted mean of assemblages. T and W are computed by maximizing the Poisson log-likelihood of the data given by:

$$L(T,W) = \sum_{i,j}\left(X_{i,j}\log(TW)_{i,j} - (TW)_{i,j}\right) \qquad (2.2)$$

The supervised implementation of NMF assumes a Poisson distribution for generating the observations X from T and W to maximize the likelihood (2.2). Unsupervised NMF can estimate the factor matrices T and W by minimizing an objective loss function F, for example the Kullback-Leibler divergence (KL), which measures the quality of the distributional approximation of TW to X. The KL divergence still uses an underlying Poisson distribution (Eisen et al. 1998). NMF iteratively calculates matrices T and W to minimize $F: \min KL(TW \parallel X)$ through optimization. The context of this objective function F for unsupervised NMF are relevant in section 3 when the results of assessing the appropriate number of assemblages (K) under NMF for the Red Sea data are discussed.

In our analysis, X is the matrix of ASV counts for each sample. NMF is supervised by depth or season classes using both a 60m depth and day 100 (early April) cutoff to separate class labels. The goal is to find the latent structures within the different classes. First, separately identify the assemblages in each class and then combine them into a single matrix of assemblages. For example, at the cutoff threshold suppose X has two depth labels surface and deep, X=($X^{(Surface)}$, $X^{(Deep)}$). From sample classes $X^{(Surface)}$ and $X^{(Deep)}$ calculate the non-negative type matrices $T^{(Surface)}$, $T^{(Deep)}$ and weight matrices $W^{(Surface)}$, $W^{(Deep)}$ by NMF. These type matrices are combined together and the type matrix for the whole data is: T= ($T^{(Surface)}$,$T^{(Deep)}$) (Cai et al. 2017). T is non-negative

since each of its component matrices are non-negative. T is fixed and the weight vectors in W associated with different samples are independent. In order to maximize the Poisson log-likelihood for all the data X, the Poisson log-likelihood is maximized for each sample. So to calculate the weight matrix W, a non-negative Poisson regression of each sample in X on T is performed. The details of this procedure are given in Cai et al. 2017 Appendix A.

As a result of supervised NMF a matrix of assemblage mixing weights for each sample is obtained as well as a matrix of ASV composition of each assemblages. These mixture proportions represent the latent structure of subcommunities in all the data samples that best distinguish class labels.

2.3 Structural Topic Models (STMs)

STMs are probabilistic mixed membership models originally developed for text mining. STM implements Bayesian inference to discover latent topics (assemblages) based on word (ASV) counts. The original terminology for STMs has analogous components for microbiome data that will subsequently be used. Samples are referred to in the literature as documents, the corpus of documents being the collection of samples taken. Words translate to individual ASVs and topics are assemblages. Further, STMs allow for metadata information from other sample measurements to improve estimation of the assemblage weights. For example, depth and time values are used as covariates to influence the proportion of each sample devoted to an assemblage. This component of the model is referred to as assemblage prevalence (instead of topic prevalence). Categorical covariate can also optionally (although not utilized here) be incorporated to affect the

rates of ASV occurrence (i.e. species frequency) in an assemblage. This component is referred to as assemblage content**.**

Like BioMiCo, STMs are a two-level hierarchy of mixture distributions. An assemblage is defined as a mixture over ASVs where each ASV has a probability of belonging to an assemblage. A sample is a mixture over assemblages, meaning that a single sample can be composed of multiple assemblages.

Formally, the setup and generative process of the model is given by the following:

- D are samples indexed by d ϵ {1,…,D}
- n ϵ {1, … , N_d} are indices for ASVs in each sample d
- K are assemblages indexed by k ϵ {1,…,K}
- $\boldsymbol{w}_{d,n}$ are the observed ASVs
- P is the number of covariates incorporated into the model

1) For each ASV n in a sample d, an assemblage is assigned from a multinomial distribution with parameter θ_d. $z_{d,n}$ indicates the assigned assemblage.

$$\text{Assemblage assignments: } \boldsymbol{z}_{d,n} \sim Multinomial_K(\boldsymbol{\theta}_d)$$

2) Given the assemblage assignment, a specific ASV ($w_{d,n}$) is chosen from a corresponding multinomial distribution over ASVs. The appropriate multinomial distribution parameter is denoted $Bz_{d,n}$.

$$\text{Draw each ASV: } \boldsymbol{w}_{d,n} \sim Multinomial_V(\boldsymbol{B}_{\boldsymbol{z}_{\mathrm{d,n}}})$$

3) Metadata covariates are represented by a D×P matrix X of prevalence covariates whose respective rows are denoted x_d.

4) A logistic Normal prior distribution controls the proportion of ASVs in a sample that are attributed to different assemblages. This logistic Normal prior has a mean vector μ_d parameterized as a linear model of the metadata covariates.

$$\text{Assemblage proportions: } \boldsymbol{\theta}_d \sim LogisticNormal_{K-1}(\boldsymbol{\Gamma}'\boldsymbol{x}'_d,\ \Sigma)$$

$\mu_d = \Gamma' x_d'$ where $\Gamma = [\gamma_1|\ldots|\gamma_K]$ is a matrix of coefficients for the sample covariates.

$$\gamma_k \sim Normal\_p(0, \sigma_k^2 I_p)$$

This is how metadata are incorporated into the model by allowing the vectors of ASV proportions allocated to assemblages to vary as a function of covariates.

The logistic normal distribution incorporates a covariance structure among the assemblage proportions. This is advantageous since naturally some microbial subcommunities may be highly correlated. However, this ability to model correlations between assemblages and leverage covariates sacrifices some computational efficacy. Inference becomes complicated because the logistic Normal prior is not conjugate with the multinomial likelihood of the generative model. So approaches like Gibbs sampling are not possible. The posterior is instead approximated using Variational Expectation-Maximization with a Laplace approximation to the non-conjugate part of the model (Dempster et al. 1977; Liu 1994; Meng and Van Dyk 1997; Blei and Lafferty 2007; Wang and Blei 2013). The variational inference approach is briefly described.

2.3.1 Variational Inference

Consider a generic model with observations $\mathbf{x} = x_{1:n}$ and latent variables $z_{1:m}$ (such as the hierarchical sample-assemblage proportions θ, and assemblage-ASV proportions β). The inference problem is to compute the posterior:

$$p(z \mid x) = \frac{p(z, x)}{p(x)}$$

In variational inference the posterior is approximated by proposing a flexible family of distributions for the latent variables. An iterative optimization procedure updates the latent variable distributions to find the member of the proposed family that minimizes the Kullback-Leibler (KL) divergence to the conditional posterior. Essentially, variational inference approximates the posterior through optimization in place of traditional numerical methods that approximate the posterior by simulating sample draws from a target distribution (Roberts et al. 2016).

The proposed conditional sample-assemblage distribution is in the exponential family. The conditional distribution of the observed data given the sample-assemblage distribution is also proposed to be in the exponential family. Substituting these exponential family distributions into the iterative optimizing updates for the latent variables does not yield a closed form expression. To solve this problem a Laplace approximation is used in the optimization calculation by taking a quadratic Taylor expansion around the maximum of each update formula (Wang et al. 2013). Applying an approximation that is not analytically intractable helps make variational inference more efficient which is another reason why STMs are computationally faster than Gibbs sampling. STMs have been appropriated here to model large microbiome data. A comprehensive review of variational inference is given in Blei et al. (2018).

2.4 SUBSAMPLING RANKING AND FORWARD SELECTION (SURF)

SuRF (Liu et al. 2019) is a sparse variable selection strategy for identifying key biomarker ASVs. SuRF is a two part procedure. First, a large number of stratified

subsamples are generated. For example, each subsample may contain 90% of the data representing balanced sample classes of season or depth strata. Given a response variable, LASSO regressions are performed on the subsamples and the taxa predictors selected by each LASSO are recorded. A list of all recorded ASVs is ranked by the number of times they are selected in each subsample. The ordering of ASVs determines the strength of association with the response variable.

Second, forward selection is applied to the list of candidate variables consisting of all the ranked ASVs based on a p-value from the null distribution calculated by a permutation test at each step. At each step the predictor variables not yet selected by the model so far are permuted to randomize their relationship to the response variable, so that the correlation structure among predictor variables is preserved. The largest log likelihood ratio (LR) statistic is recorded for each permutation of the candidate data to get a null distribution for the maximum log LR statistic. The value at the $(1 - \alpha)$ percentile of this null distribution is the critical value used to determine forward selection. That is, the original unpermuted candidate ASVs are added one at a time to the current regression model based on a conditional test that the current model is correct. At each step the first log LR statistic greater than the critical value is selected. The whole process is then repeated with new permutations and a new null distribution until none of the LR statistics for the ranked candidate taxa exceed the critical value. The log LR statistic does not follow a χ^2 distribution because multiple predictors are tested at each stage. SuRF can agglomerate ASVs at higher taxonomic levels but here only ASVs at the lowest taxa level possible were selected.

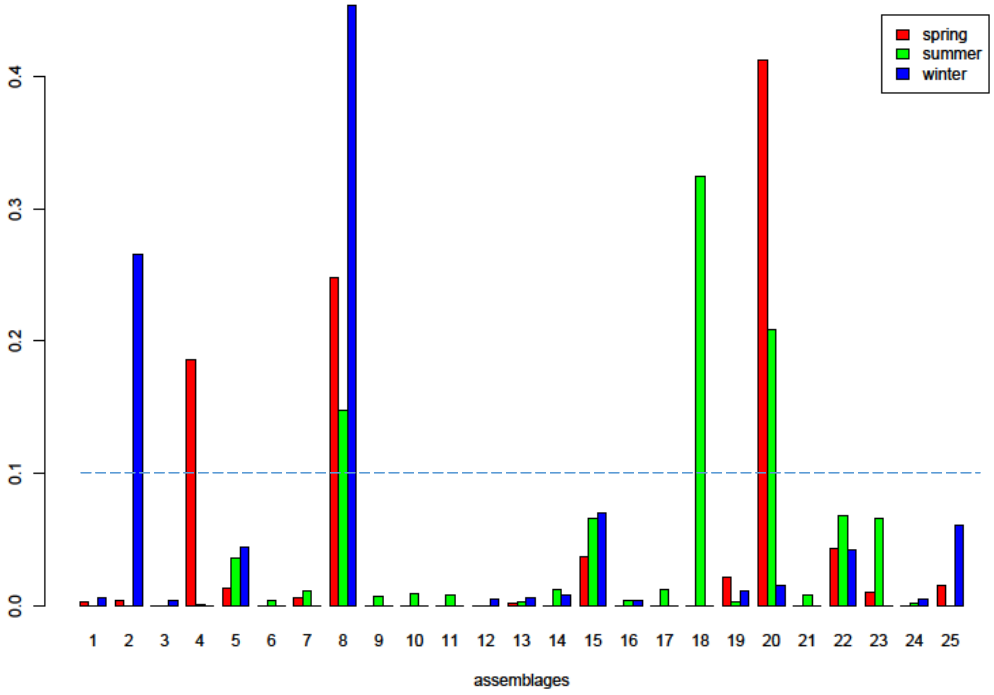
CHAPTER 3 RESULTS AND INSIGHTS

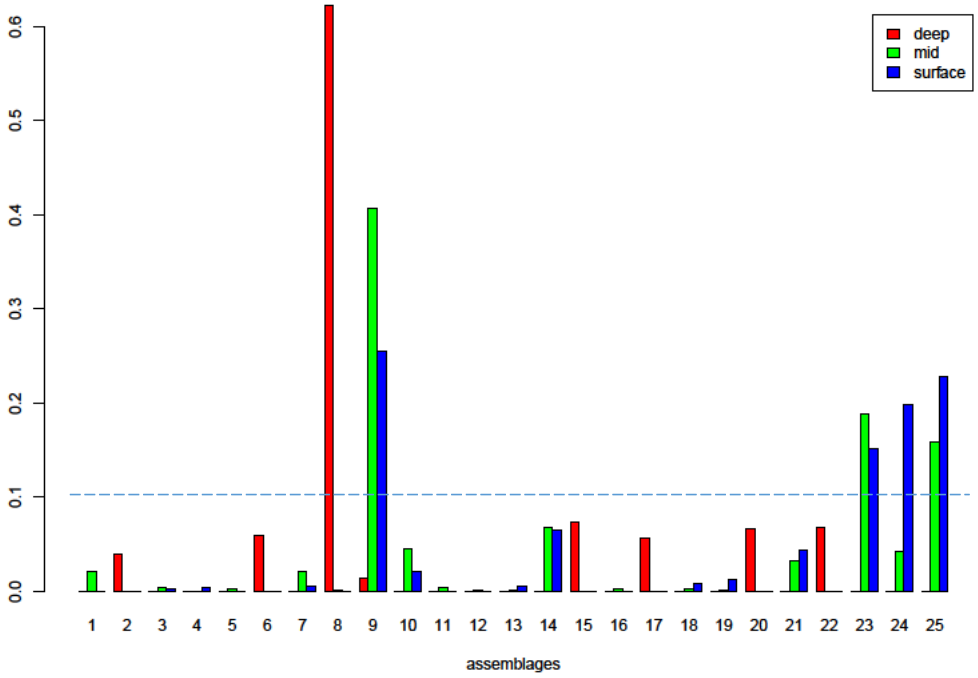
The results in this section are only for the heterotrophic community taxa unless explicitly stated that phototrophs were included. As mentioned in the introduction the heterotrophs are primarily of interest but in several cases analysis was also run with phototrophic cyanobacteria included. The optimal number of assemblages (K) is chosen based on a threshold for assemblage posterior probabilities inferred from BioMiCo, quality of fit diagnostics for unsupervised NMF and STM and cross validation tests for supervised NMF. The agreement on a K value among methods produces a readily comparable number of assemblages. The distributions of assemblages from each method are compared in terms of their environmental interpretations and taxa composition. Finally covariate categories are predicted from assemblage proportions and assemblages are characterized by associated physiochemical vectors and functional traits.

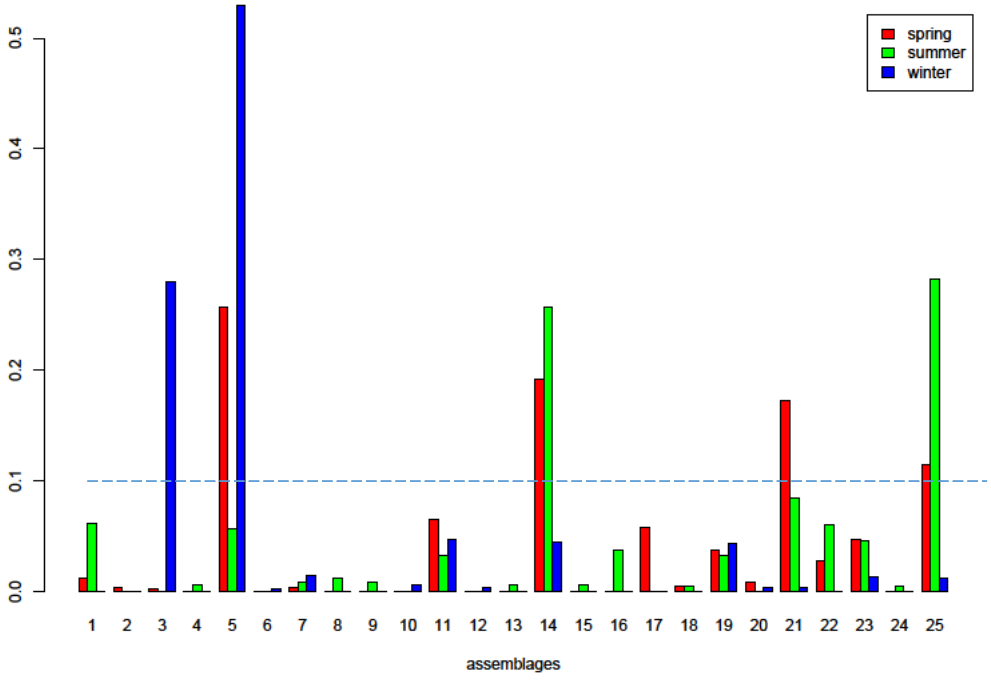
3.1.1 Determining the Number of Assemblages for BioMiCo

In this study BioMiCo was supervised by environment factors of sample season (winter, spring, summer) and depths (surface depths 60m or above, middle depths 80 – 200m and deepest depth 400m). The first 2500 iterations of the Markov chain Monte Carlo (MCMC) for the training phase were considered “burn-in” and were discarded. Following the burn-in, the MCMC was run for a minimum of 2,000 iterations for each sample draw. Multiple chains were run with 20 samples drawn for each and the number of assemblages (K) set to 25. The model assigned assemblage posterior probabilities to assemblages in each environment. Assemblages that occurred at greater than 10% contribution to an environment were considered predominant assemblages. BioMiCo

consistently identified 5 predominant assemblages for the heterotrophic communities in both years and when supervised by either season or depth factors (Figure 3.1). The following plots show the posterior probability contribution of assigned assemblage to each environment. BioMiCo was also run on the entire microbial community including cyanobacteria. In that case 6 predominant assemblages were found for 2015 and between 5 and 7 for 2016 communities depending on which environment factors were trained.




deep
mid
surface
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
assemblages


spring
summer
winter
0.0
0.1
0.2
0.3
0.4
0.5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
assemblages

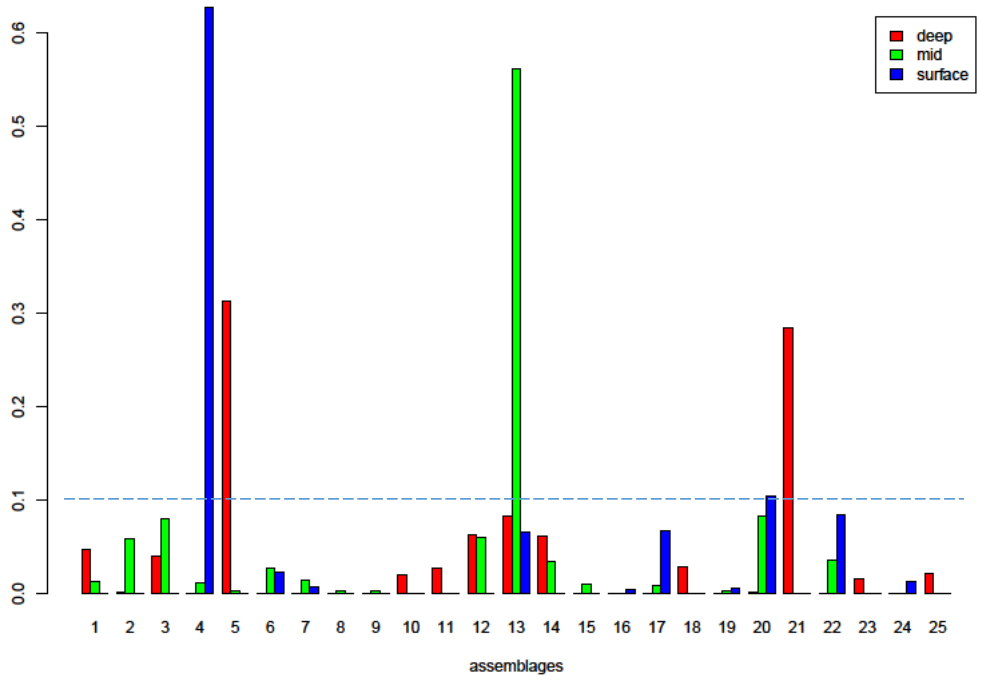


Figure 3.1 BioMiCo assemblage distributions showing contributions to seasonal and depth environments. The first two plots are for 2015 and the last two are for 2016. Dotted lines are at the 10% predominance threshold.

3.1.2 Determining the Number of Assemblages for NMF

For unsupervised NMF several quality measures have been proposed for determining the number of assemblages K (factorization rank of T, the type matrix) for best model fit. For example, the sparseness of the sample over assemblage weight matrices modelled over multiple runs of NMF in both years was highest for K = 5 types.

Hutchins (2008) proposed K should the smallest value where the marginal residual sum of squares (RSS) from a loss function F (KL divergence) presents an inflection (elbow) point in its curve over a range of K values. Frigyesi et al. (2008) proposed the smallest value of K for which the decrease in the RSS remains larger than the decrease of the RSS from randomizing the observed counts of each ASV to destroy

the community connection to samples. Figure 3.2 shows these residual quality measures for a range of K values along the x-axis and indicate K = 5 assemblages is appropriate for the 2015 and 2016 community data. The randomized data is shown by the dotted line with triangle points plotted at each K. The smallest k given these quality measures was chosen to control the model complexity and avoid overfitting the data by supposing too many columns of T (subcommunities).

The explained variance (evar) from the data X and the NMF estimates TW evaluates how well the model reconstructs the data. The inflection point at K = 5 of the evar plot suggests that the marginal improvement in explained variance of the model fit is not worth the added complexity for more than 5 assemblages (Figure 3.3). Fewer assemblages are also desirable for comparing their biological interpretations across different methods.

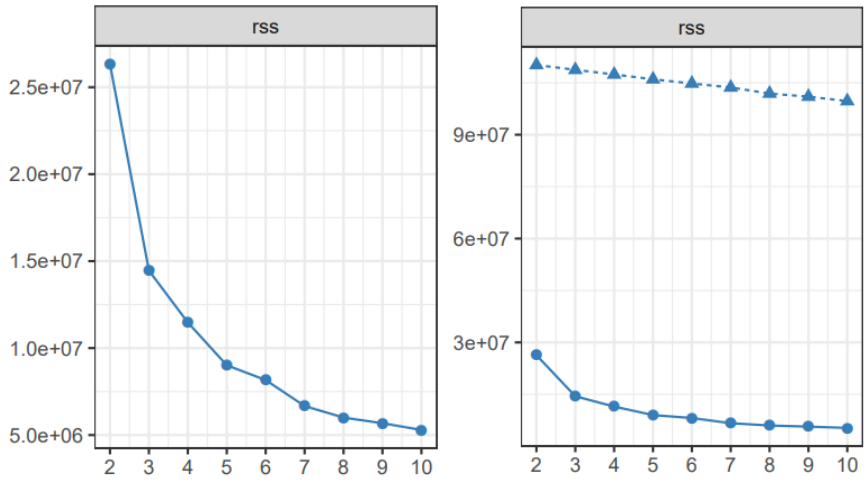


Figure 3.2 Residual sum of squares (RSS) vs number of assemblages (K values) and comparison to RSS for randomized data (▲)

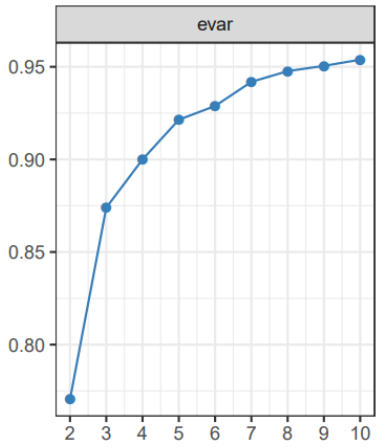


Figure 3.3 Explained variance vs number of assemblages (K values)

Another criterion for K is the cophenetic correlation coefficient. The NMF weight matrix clusters samples together based on their weights in each assemblage. NMF calculates an n × n connectivity matrix C where $C_{ij} = 1$ if sample i and sample j are clustered together by NMF and 0 otherwise. The cumulative average of matrices C over a sequence of NMF runs yields an n × n matrix of empirical probabilities that each pair of samples, i and j, are clustered together. For each K value the multiple runs of NMF are carried out on perturbed subsets of the original data produced from subsampling the data with and without replacement (. The pairwise average matrix is called the consensus matrix $\overline{C}$, and is used to measure sample similarity. Here 50 runs of NMF were used to determine the consensus matrix. The cophenetic coefficient measures the correlation of the sample distances induced from the consensus matrix and the sample distances from hierarchical clustering of samples using these same distances. Brunet (2004) proposed K should be the smallest value of K after which the cophenetic begins decreasing. Multiple runs of NMF again indicate 5 assemblages based on this criteria (Figure 3.4). However, this criteria should not be the only or foremost quality measure. Residuals were

considered first because Frigyesi et al. (2008) concluded that the cophenetic correlation reports the ability of factorization into K assemblages to classify samples into K classes. This would not be ideal if, for example, the true number of assemblages was 5 but k = 5 produced 2 cluster classes. Then the cophenetic would report a small value for the correct number of assemblages.

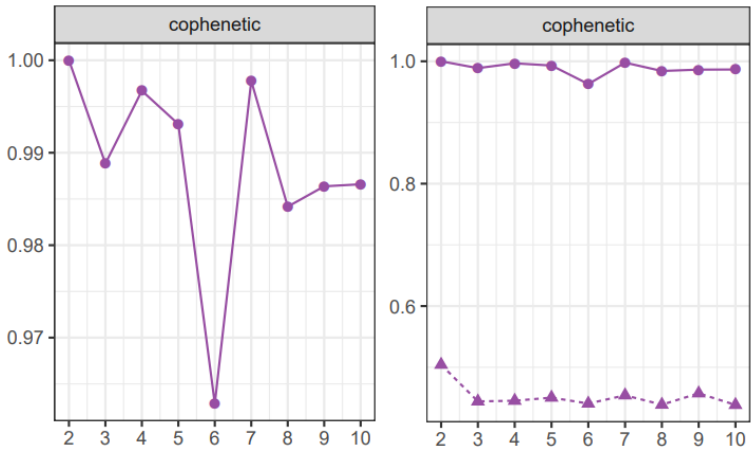


Figure 3.4 Cophenetic correlation coefficients vs K value (left) and comparison with cophenetic coefficients calculated from randomized data over the same K range on the right (▲).

For supervised NMF the number of types (assemblages) for each class of samples (e.g. higher and lower depths) is chosen to best discriminate the separate classes from each other. A sequence of K values in a range from 2 to 10 was used to choose the number of assemblages for each class separately through the cross-validation procedure described in Cai et al. (2017) Appendix B. Briefly, if supervised by two classes, fit an NMF model on training folds from one class and compare the deviances on the test fold from that class with the deviances on a fold from the other class using a Wilcoxon Rank-Sum test. Ranking these deviances gives us a statistic (Z-value) for how different the

classes are for a given k. The objective is to choose K such that the deviances from the two classes are best separated (Cai et al. 2017).

The number of types for each seasonal and depth class was chosen separately through the cross-validation procedure described. This yielded 5 types (assemblages) total, 2 corresponding to a winter mixed water column and 3 for a spring-summer stratified water column, before and after day 103 of both years (early April). There were also 5 types total when supervised by water column depth classes, specifically 2 for depths 0-60m and 3 80-400m. Supervised NMF determined there were 5 assemblages present in each of the 2015 and 2016 heterotrophic communities. When ASV counts included the phototrophs, NMF supervised by the same class labels for depth and season produced 3 assemblages in each class for a total of 6 microbial assemblages. For all heterotrophic community models a total of 5 assemblages were therefore chosen for modelling with NMF. It was encouraging that multiple NMF evaluation metrics produced a number of assemblages that was consistent with the number of assemblages assigned by BioMiCo.

3.1.3 Determining the Number of Assemblages for STM

For STM a data driven search was conducted for the appropriate number of assemblages by calculating diagnostic metrics and plotting the results over a range of K values. First, held-out likelihood analysis was performed, where half the ASVs in a subset of samples are held out and the model is trained (Wallach et al. 2009). The sample-level latent variables (assemblage distributions) are used to evaluate the probability of the second half of ASVs in the heldout portion of samples (Roberts et al.

2016). A K value is determined where the likelihood is maximized locally for the smallest manageable number of assemblages in the range. Second, with each K specified for a model, the multinomial distribution gives a dispersion (variance σ^2) of the residuals. K is determined so that the residuals are not overdispersed. This residual analysis is based on Taddy (2013) and in practice a larger dispersion suggests that the current number of latent assemblages do not account for the variance. This provides rough evidence that more assemblages might be needed. K was chosen from an inflection point in the residuals plot where adding more assemblage complexity does not improve the dispersion.

Another criterion for K borrowed from text analysis called semantic coherence (or coherence) was considered for estimating STMs. Coherence was proposed in the context of probabilistic topic models by Mimno et al. (2011). The *a priori* reasoning is that pairs of ASVs belonging to the same ecological niche community will more likely co-occur within a sample of that community. ASV pairs from different communities will less likely occur together. The key assumption of this thought process is that in an assemblage of random taxa it is likely that very few ASVs will co-occur. Coherence for different K gives us a measure of how well frequently co-occurring taxa are dominant in each of K assemblages. Coherence is greatest when the most probable ASVs in an assemblage are frequently present together.

Let $\mathcal{D}(v)$ be the sample frequency of ASV v (i.e. the number of samples with at least one occurrence of v). Let $\mathcal{D}(v_i, v_j)$ be the co-sample frequency of ASVs v_i and v_j, which is the number of samples with one or more ASVs v_i and at least one ASV v_j. For

each of the M most probable ASVs in an arbitrary individual assemblage k (lowercase distinct from total assemblages K), the semantic coherence of assemblage k is defined by:

$$\mathcal{C}_k = \sum_{i=2}^{M} \sum_{j=1}^{i-1} \log\left(\frac{\mathcal{D}(v_i, v_j) + 1}{\mathcal{D}(v_j)}\right) \qquad (3.2)$$

Coherence compares the frequency of ASV v_j in samples that already contain v_i to the frequency of v_j in the whole of the sample communities. If v_j is not more probable in samples containing v_i then the coherence should be close to zero. There may be ASVs that are so ubiquitous and correlated with each other that they are the largest contributing (predominant) taxa in an assemblage. Such an assemblage would not be very meaningful since it captures an overly-general baseline of ASVs in the environment. Coherence will be high when there are few assemblages dominated by very frequently occurring ASVs. So it is not ideal to simply seek to maximize coherence. The FREX metric developed in Bischof et al. (2012) and Airoldi et al. (2016) balances the exclusivity of ASVs to each proposed assemblage with their rates of occurrence within assemblages.

Diagnostic values were plotted by the number of assemblages specified in a range from k = 3 to k = 10 for models with prevalence prior parameterized by date and depth covariates. The x-axis is the number of assemblages tried. Over multiple runs of STM the optimal points in the local range of K values showed that K= 5 assemblages was the smallest number after which the marginal improvement in most metrics began to decrease. The exclusivity and FREX measures were also monotonically increasing with K. It was heuristically determined that K = 5 for exclusivity as well (Figure 3.5 and 3.6).

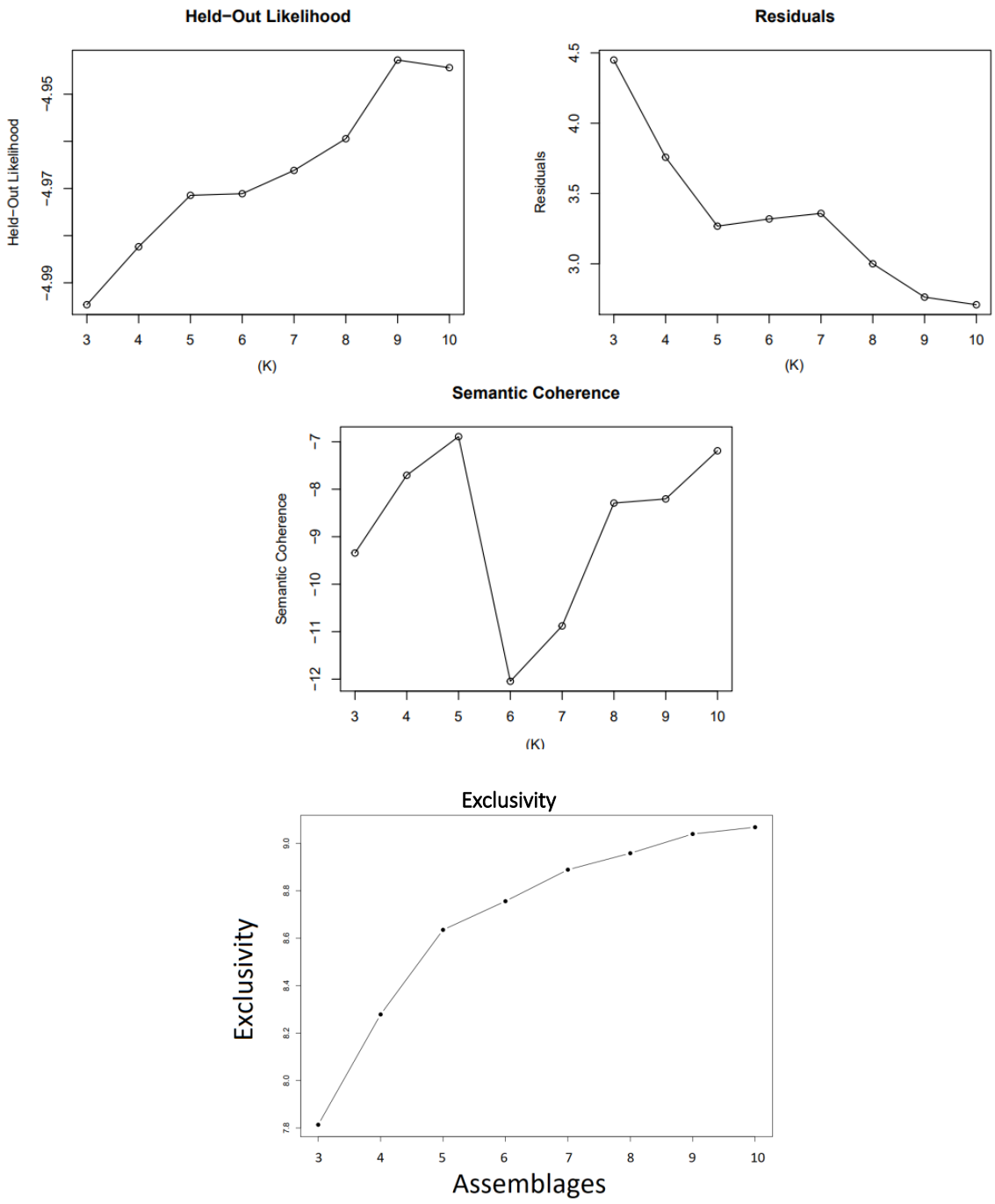


Figure 3.5 Diagnostic plots for 2015 STM runs over a range of K values to determine the optimal number of assemblages

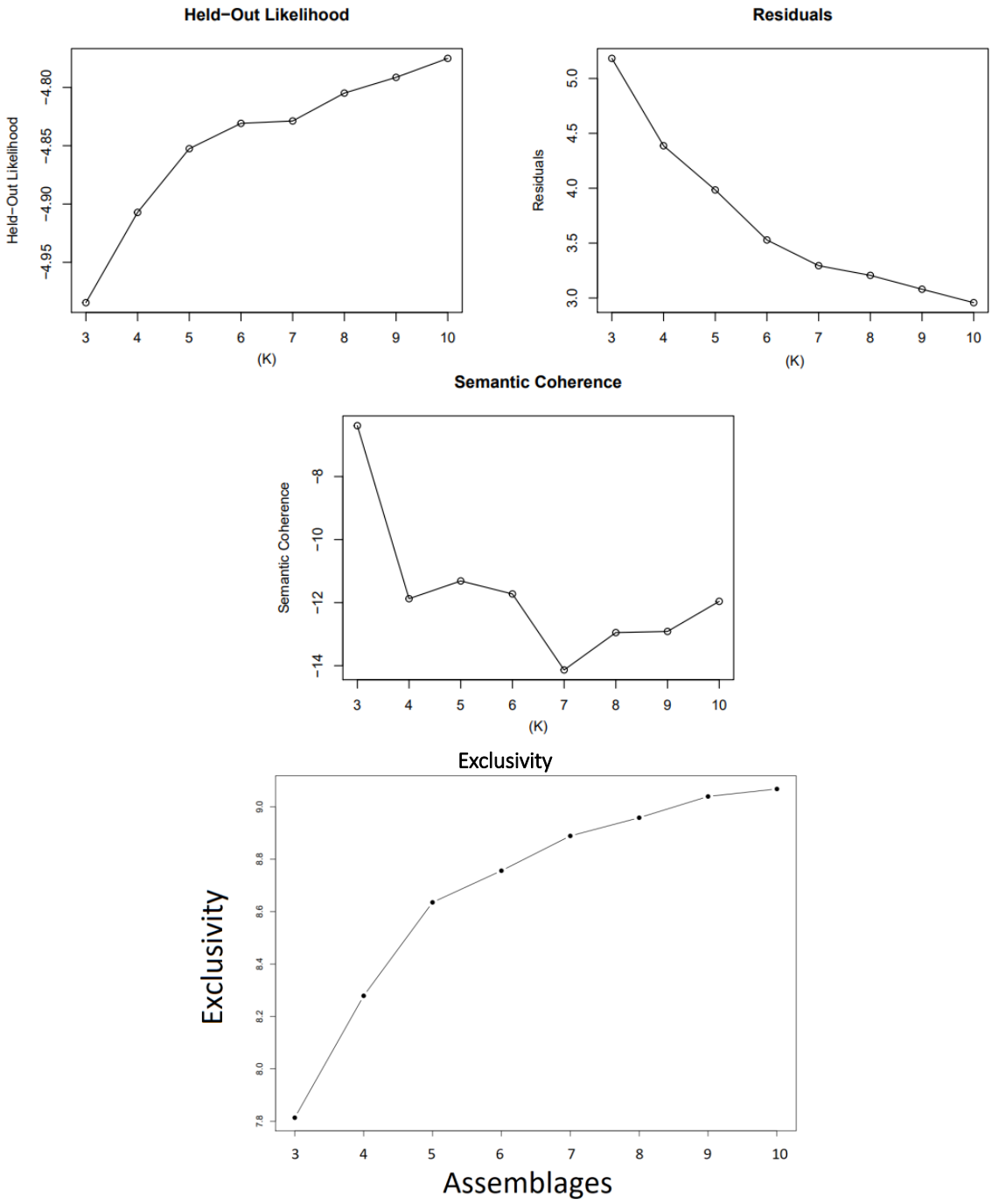


Figure 3.6 Diagnostic plots for 2016 STM runs over a range of K values to determine the optimal number of assemblages

For comparison, the same metrics run on the entire microbial community with cyanobacteria resulted in 6 assemblages as the optimal number (Figure 3.7). BioMiCo and NMF resolved an additional 6th assemblage when the cyanobacteria were included as well.

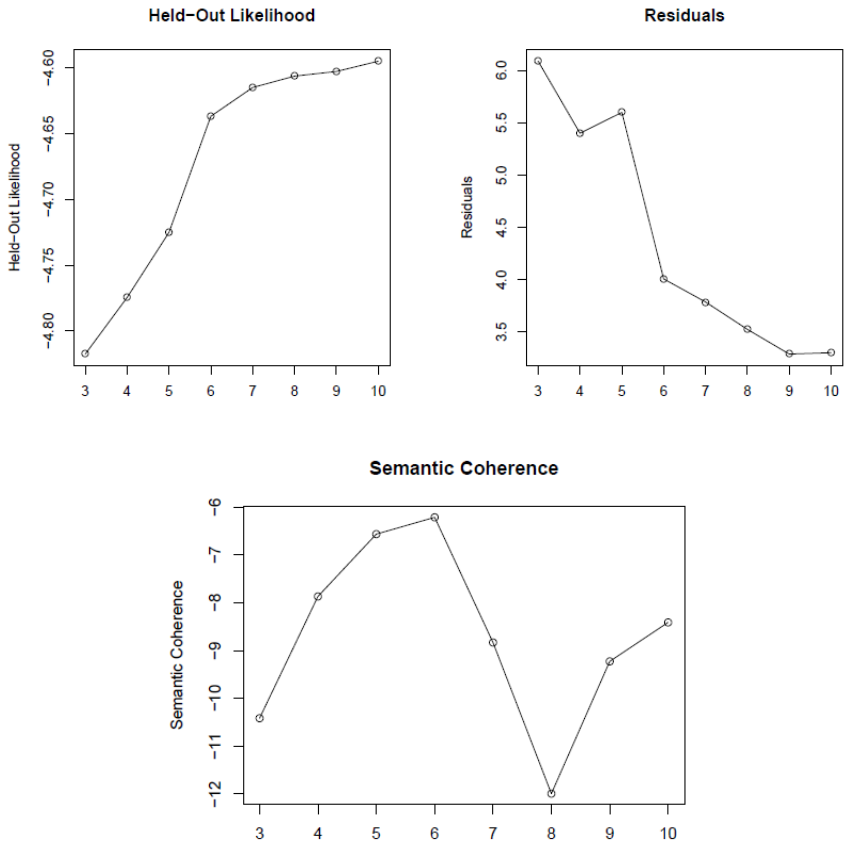


Figure 3.7 Diagnostic plots for STM runs over a range of K values to determine the optimal number of assemblages for all taxa including cyanobacteria.

3.2 Comparative Analysis of Assemblages

At the sample level, Bray-Curtis dissimilarity summarizes the pairwise difference in compositional abundances of taxa for samples over both years. Here it is applied to just the heterotrophic ASVs. The Bray Curtis distance is defined as:

$BC_{ij} = 1 - \frac{2F_{ij}}{S_i + S_j}$,where i and j are two samples, F_{ij} is the sum of only the lesser frequencies for each ASV found in both samples and S_i is the total sum of frequencies of all taxa in sample i. This assessment of pairwise community dissimilarity suggests community structure varies along dimensions of season and depth. Samples within seasonal and depth groups had more similar taxonomic compositions. Non-metric Multidimensional Scaling (NMDS) projection using the dissimilarity matrix reveals compositional divergence in two dimensions (Figure 3.8 and 3.9). The goodness-of-fit of these 2-dimensional non-parametric plots for the actual multidimensional space of the samples was measured by the stress. Stress measures how well the ranking of observed dissimilarities correlates with the ranking of ordination distances. Stress reported here was < 0.05 for both 2015 and 2016.

Figure 3.8 Bray-Curtis dissimilarity among 2015 samples showed compositional divergence along dimensions of season and depth. Note that the summer samples with larger second NMDS dimension (NMDS2) were from higher depths where the water column was warmer. More of the summer samples that were closer to the spring and winter samples (lower NMDS2 values) were from deeper, colder depths in the water column.

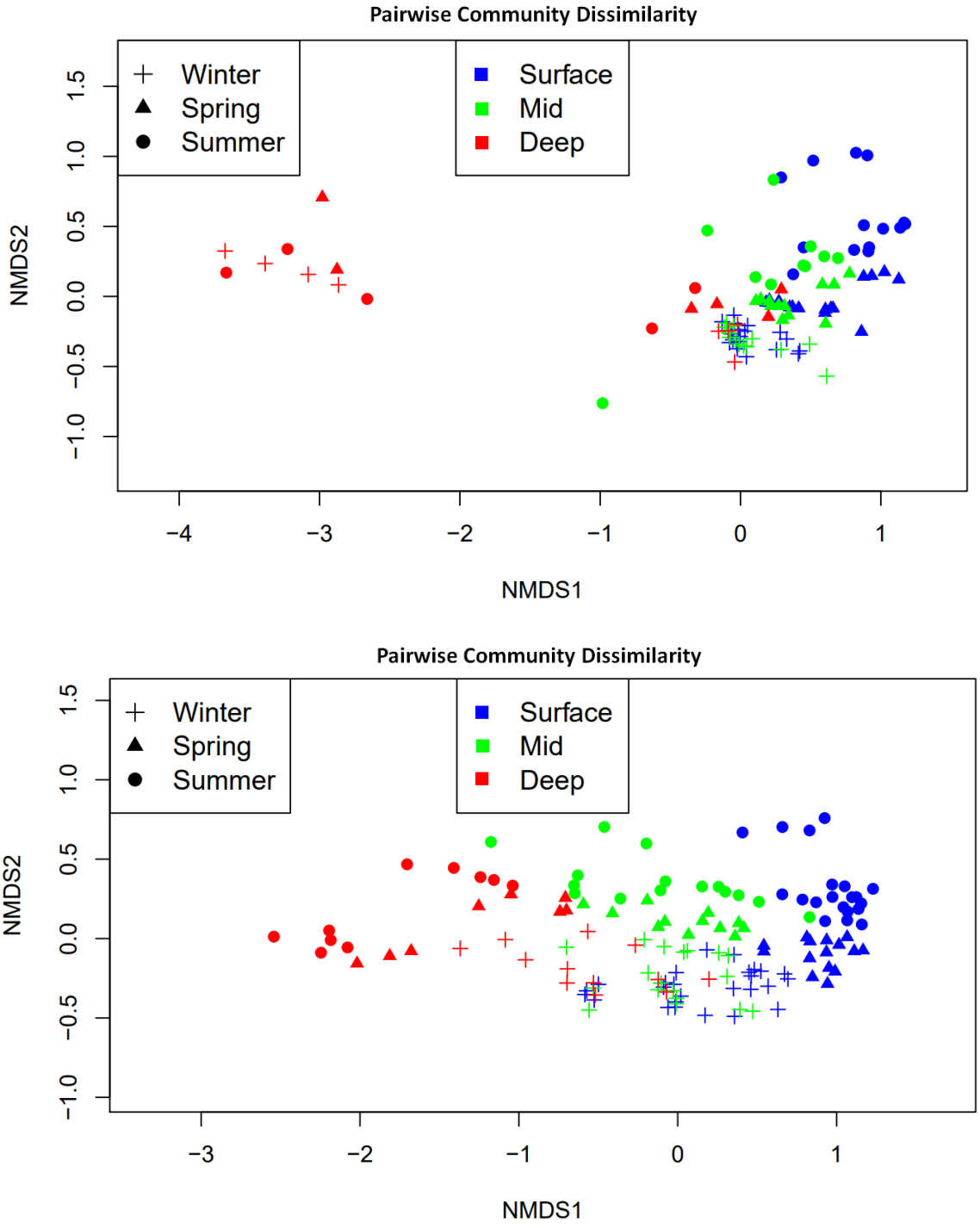


Figure 3.9 Bray-Curtis dissimilarity among 2016 samples showed compositional divergence along dimensions of season and depth. There was more overlap among samples from different depth groups (i.e. smaller variance among NMDS1 values) during the winter season (lower NMDS2 values). This reflects greater taxonomic homogeneity in the mixed water column of the winter compared to the progressively stratified spring and summer depth layers.

Samples from the Bray-Curtis plots were labelled by their membership in the assemblage with the greatest contribution, that is, the assemblage for which the sample had the largest mixing weight. Samples labeled by assemblage in this way showed how ASV composition of assemblages captured the structure of seasonality and depth. When the other environmental covariates observed at the sample time and depth points are projected onto the NMDS dimensions, we see that separate assemblages were related to specific physical vectors (Figures 3.10 and 3.11). For example, in 2015, a distinct heterotrophic assemblage from NMF and STM (A1) was dominant among samples at lower depths and higher concentrations of NO_3 and PO_4. NMF and STM assemblage 3 (A3) was more predominant along an increasing gradient of NO_2 whereas the neighbouring A5 was more predominant at higher oxygen and *Synechococcus* concentrations. A4 was associated with increased *Prochlorococcus* during a bloom at later time points and higher temperatures. In 2016, NMF A2 was dominant at higher temperatures and later time points in the summer and was less associated with *Prochlorococcus* which had an earlier bloom that year more in sync with *Synechococcus*.

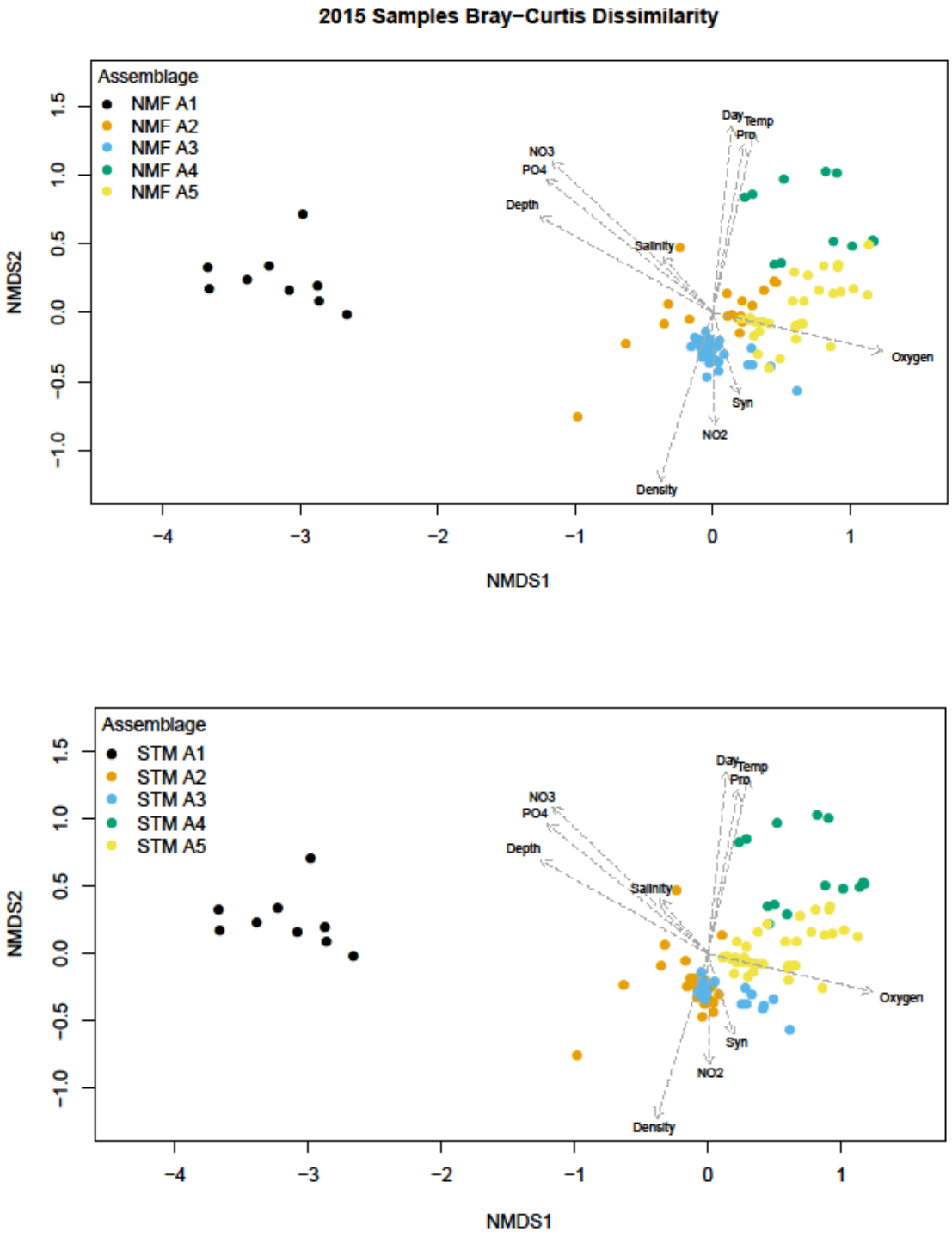
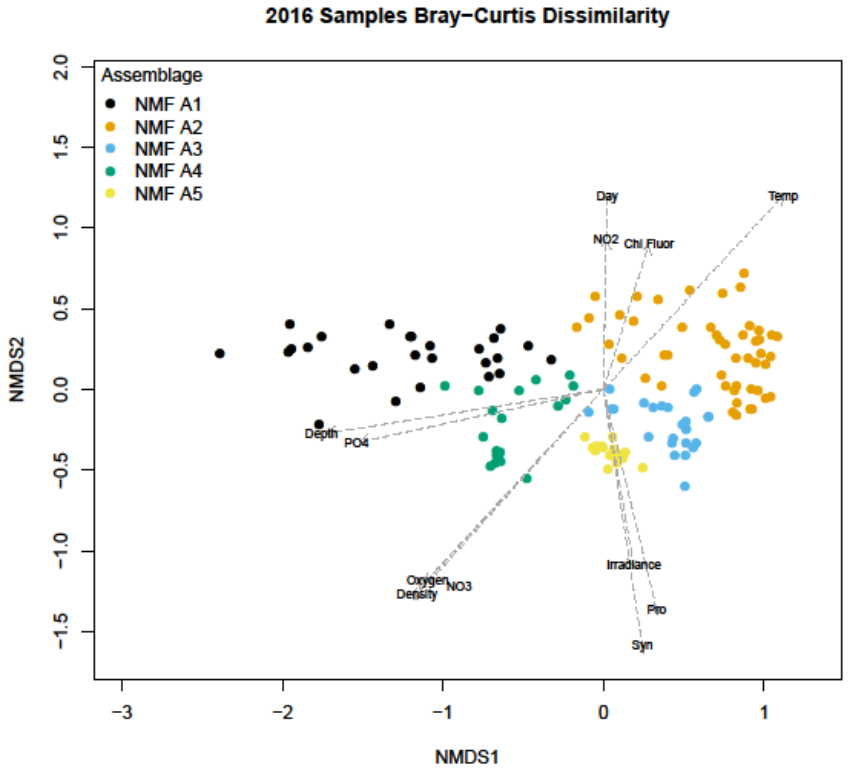


Figure 3.10 Bray-Curtis dissimilarity of 2015 samples labelled by NMF (top) and STM (bottom) assemblage membership. Different colours in each plot represent the separate but corresponding assemblages. Syn, Pro and Temp are used as short forms for *Synechococcus*, *Prochlorococcus* and temperature vector labels.


2016 Samples Bray−Curtis Dissimilarity
Assemblage
NMF A1
NMF A2
NMF A3
NMF A4
NMF A5
Day
Temp
NO2
Chl Fluor
Depth
PO4
Irradiance
Oxygen
Density
NO3
Pro
Syn
NMDS1
NMDS2

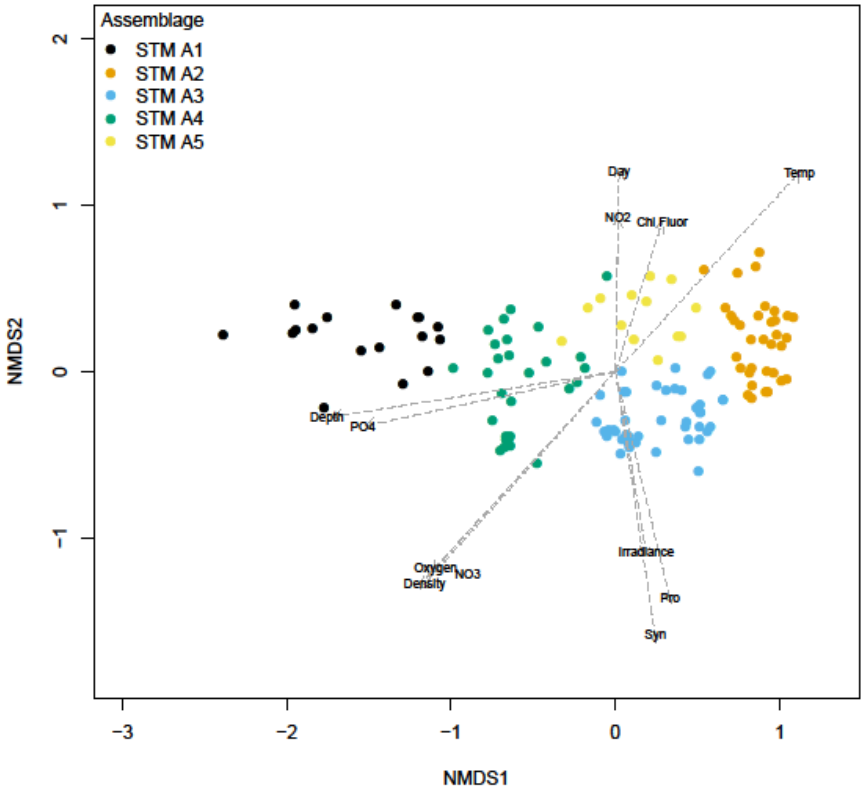

Assemblage
STM A1
STM A2
STM A3
STM A4
STM A5
Day
Temp
NO2
Chl Fluor
Depth
PO4
Irradiance
Oxygen
Density
NO3
Pro
Syn
NMDS1
NMDS2

Figure 3.11 Bray-Curtis dissimilarity of 2016 samples labelled by NMF (top) and STM (bottom) assemblage membership. Different colours in each plot represent the separate but corresponding assemblages. Syn, Pro, Temp and Chl Fluor are used as short forms for *Synechococcus*, *Prochlorococcus*, temperature and chlorophyll fluorescence vector labels.

When comparing the plots for NMF and STM in figures 3.9 and 3.10 there is similarity in the separation of assemblages over environment factors for both methods. However, there is some variability in what samples belong to assemblages from different methods. For example, in the last two plots of 2016, STM assemblage A3 overlaps NMF A3 and A5. In contrast, NMF A1 in 2016 overlaps STM A1 and A4.

Cluster analysis also showed structure according to season and depth. K-means clustering was conducted with the weights or mixing probabilities of sample distribution over assemblages. An appropriate number of clusters was chosen based on the marginal improvements in total within cluster sum of squares and the local maximum average silhouette width for different numbers of clusters. Samples were plotted by time, depth and log transformed phototroph levels (Figures 3.12 and 3.13 log-log plots). Observed points were then labelled by cluster membership. The goal was to see if the heterotrophic assemblage weights clustered in patterns that were associated with spatial-temporal and cyanobacteria dynamics. K-means clustering of all assemblage weights showed that the mixing proportions cluster samples around specific seasonal and depth environments as well as cyanobacteria concentrations. There were distinct clusters at the highest and lowest concentrations of the cyanobacteria. Mixing weight clusters from NMF and STM similarly separated samples by season and depth.

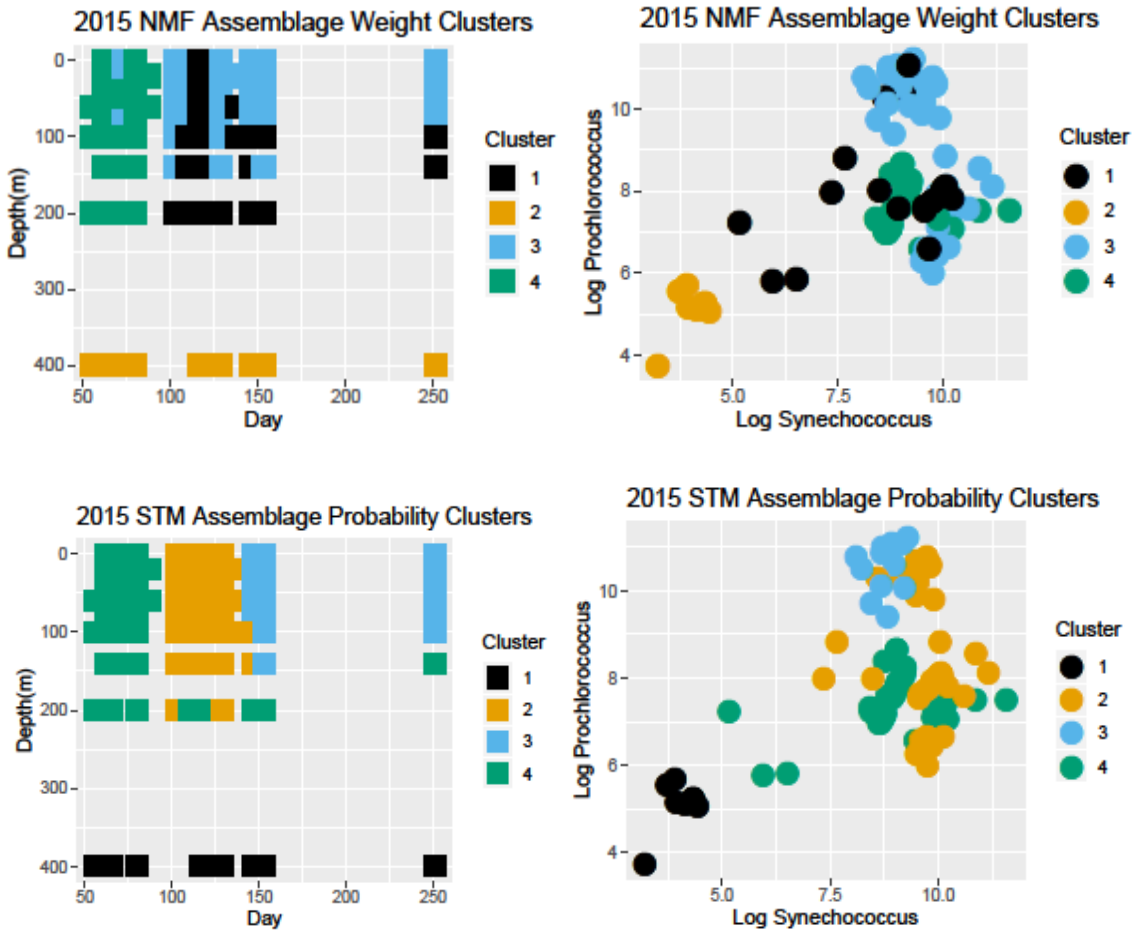
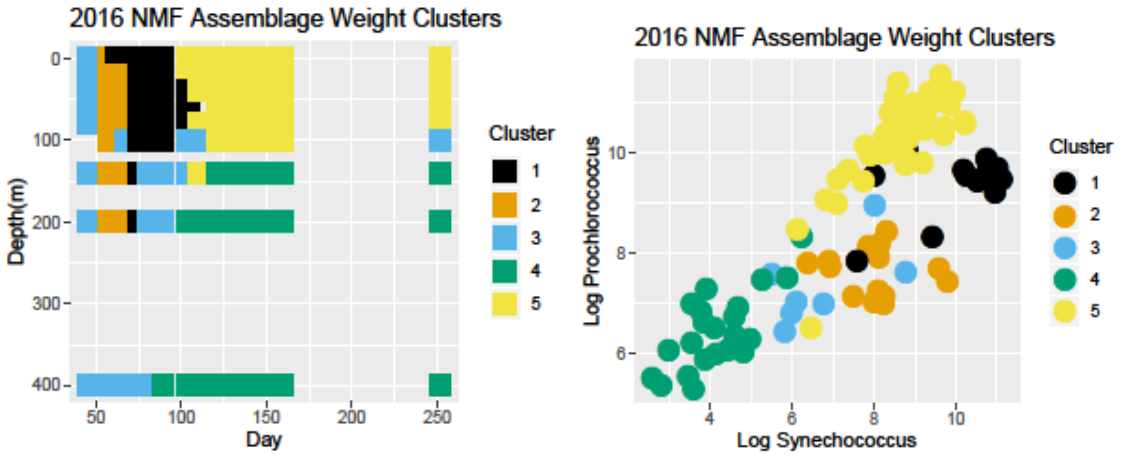


Figure 3.12 K-means clusters of 2015 assemblage weights for NMF and STM. Cluster colours have no correspondence with colours from previous plots.



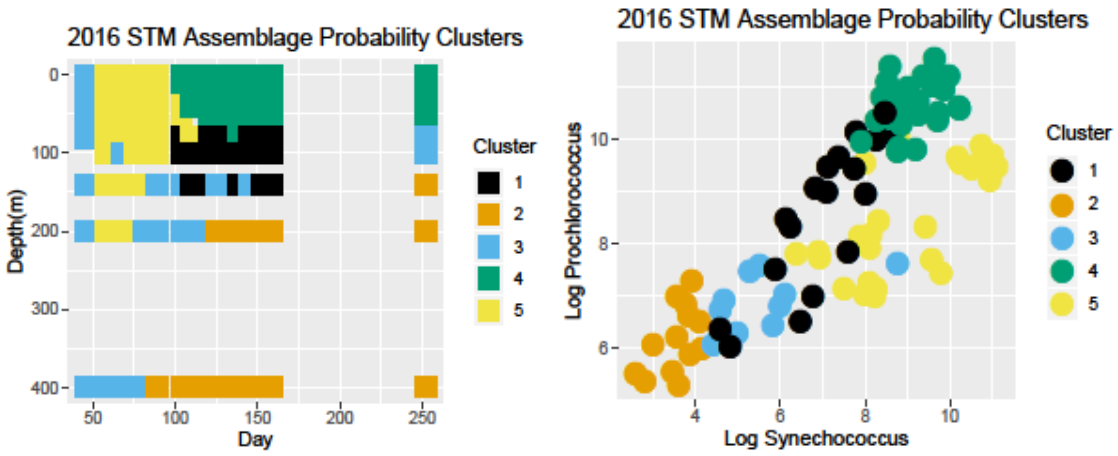


Figure 3.13 K-means clusters of 2016 assemblage weights for NMF and STM. Cluster colours have no correspondence with colours from previous plots.

At the ASV level, assemblages from all three methods that were associated with the same seasons and depths were compared. ASVs that had greater than 10% mixing proportion in an assemblage composition were selected as predominant taxa. Bray-Curtis dissimilarity among assemblages from different methods was used to examine the divergence of assemblage composition. The mixing proportions of predominant ASVs were treated in the same way as the abundance of taxa in a sample. The largest contributing taxa identified from BioMiCo assemblages were used as a baseline composition for comparison across methods and environments. Hierarchical clustering with an average linkage was then used to show corresponding assemblage similarity in a dendrogram.

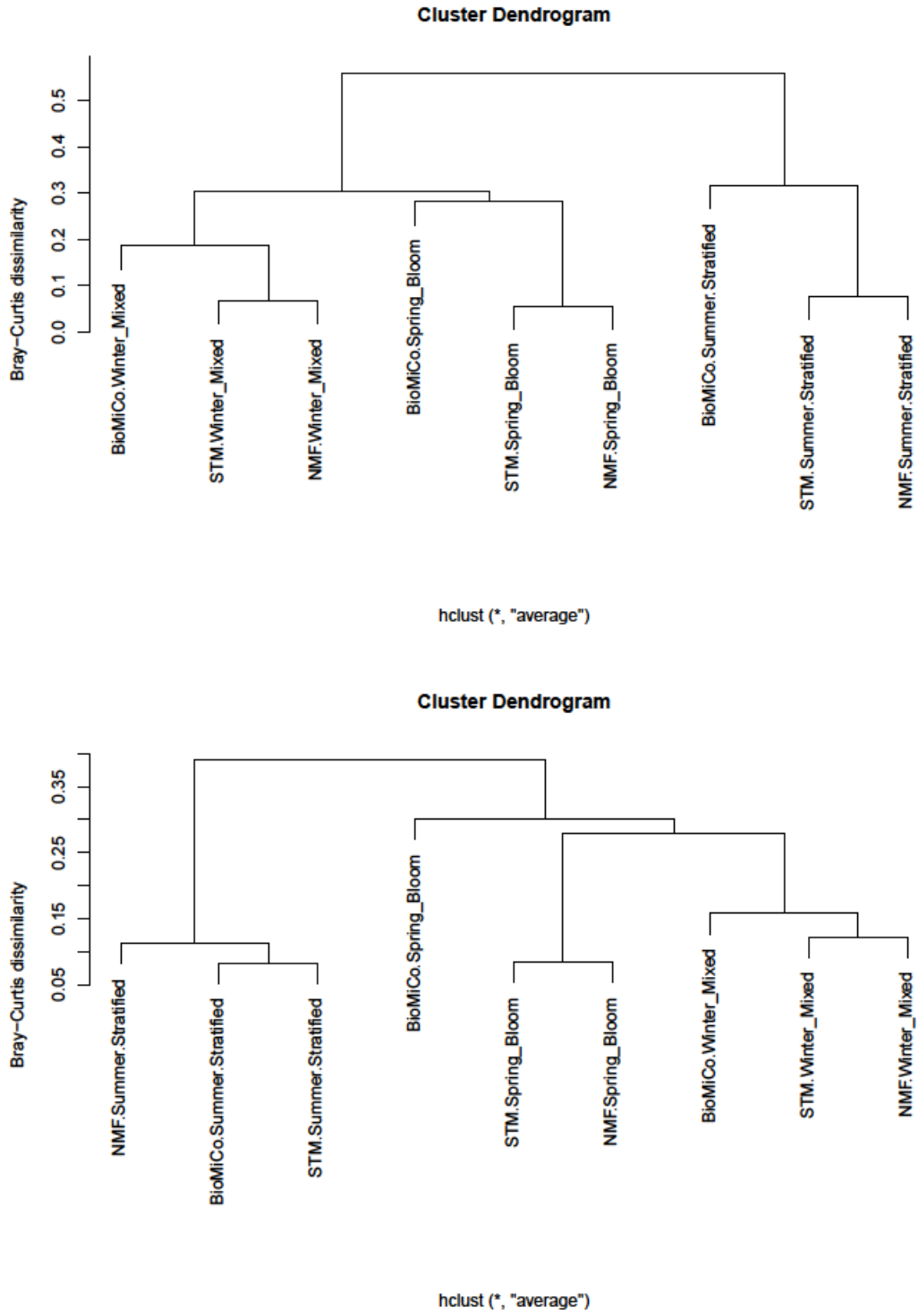
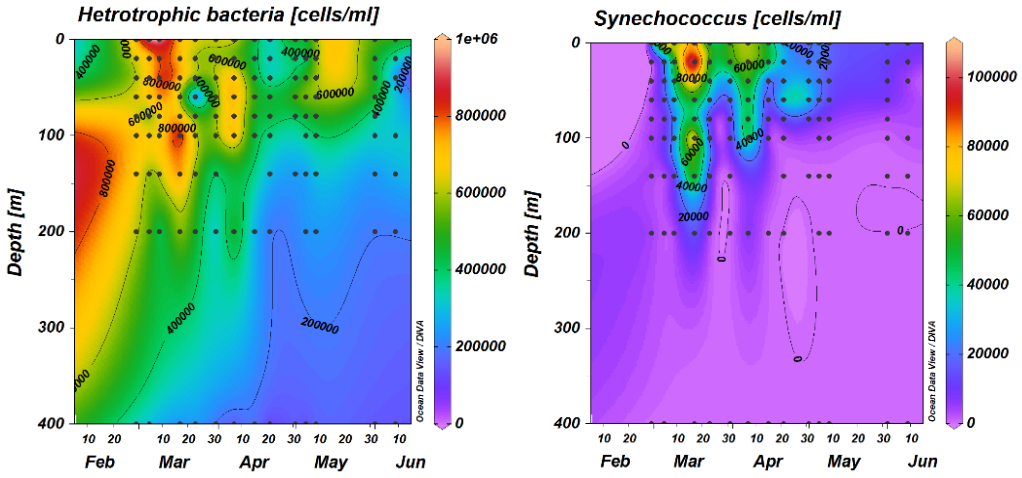


Figure 3.14 Agglomerative hierarchical clustering with an average linkage of seasonal assemblages from BioMiCo, NMF and STM for 2015 (top) and 2016 (bottom).

Assemblages within distinct environments are more closely clustered together even when inferred from different methods. There is a stronger consensus established of composition between NMF and STM, although BioMiCo assemblages were still very similar with respect to spatial-temporal dynamics.

3.3 PREDICTION

Since there are two separate years of data, models could be trained on each year and tested on the alternate year. Same year training and testing accuracy was based on leave-one-out cross validation. Testing allows us to quantify how well assemblages captured generalized microbial patterns predictive of season, depth and phototrophic concentrations. Only heterotrophic assemblages were modelled without being supervised by any cyanobacteria related covariate classes. Profiles of heterotrophic bacteria and cyanobacteria cell concentration over time and depth were examined to decide how to split training and testing samples into classes of distinct environments. Figure 3.14 shows example bacteria profiles.



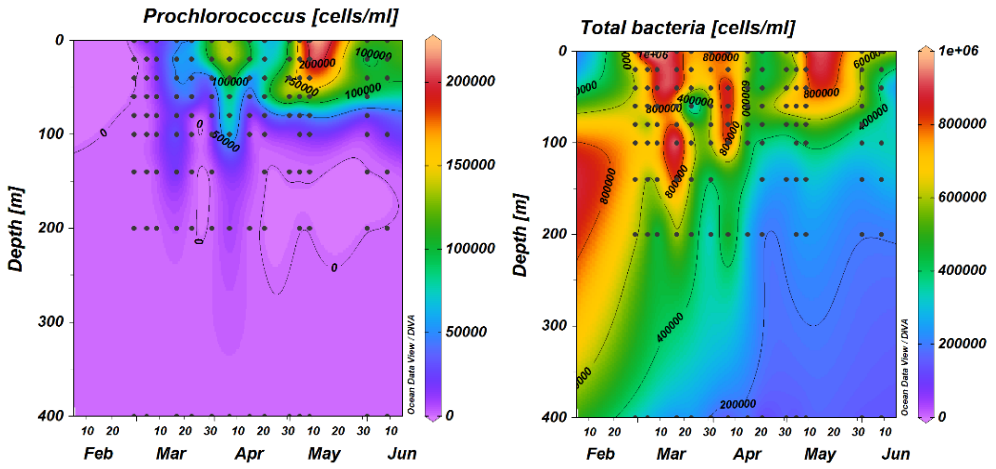


Figure 3.15 Heterotrophic and cyanobacteria profiles over time and depth.

Sample classes were split the same way for all prediction methods (BioMiCo, NMF, SuRF, etc.) and produced classes that were naturally almost balanced by seasonal stratification of the warming water column, water depth strata and low vs high cyanobacteria cell concentrations (as proxy for bloom vs non-bloom states). Seasonal class splits for 103 total samples in 2015 and 136 total samples in 2016 were 55/48 samples (winter mixed/spring-summer stratified water column) and 80/56 samples respectively. Depth class splits were 54/49 samples (surface-higher depth/mid-deep depth) in 2015 and 75/61 samples in 2016. For *Synechococcus* the split of observations in low/high concentration classes were 44/51 in 2015 and 50/51 in 2016. For *Prochlorococcus* the splits of observations in classes were 47/48 in 2015 and 58/43 in 2016. The total numbers of observations in each year for the cyanobacteria were fewer than the total number of heterotrophic samples because some date and depth data points were missing measurements of *Synechococcus* and *Prochlorococcus* cell concentration.

As discussed previously for BioMiCo, the posterior distribution of the environment and assemblage assignments in the test set year, X_{test} and Z_{test}, are sampled jointly given the training year assignments and test year sample ASV counts: $P(X_{test}, Z_{test} \mid X_{train}, Z_{train}, W_{test}, \alpha_\phi, \alpha_\pi, \alpha_\theta)$. Marginalizing over the assemblage assignments yields test year environment probabilities. Samples are classified by their maximum posterior environment probability and this prediction is compared to the observed environments of the test year. The results for seasonal and water column classes for BioMiCo are summarized in Table 3.1.

Table 3.1 BioMiCo prediction accuracy for training and testing on each year. Class error rates are indicated respectively in parentheses after each accuracy percentage.

BioMiCo Classification Testing Accuracy			
Train	**Test**	**Season** (mixed/stratified water column)	**Depth** (surface-higher/mid-deep)
2015	2016	79% (35.7% / 8.7%)	83.8% (16%/16%)
2016	2015	72.8% (10% / 39%)	66% (28%/39%)

The type matrix T_{train} estimated by supervised NMF gives the ASV composition of assemblages for a particular training year. The k × 1 weight vectors for each sample in the weight matrix, W, are independent so to estimate the test year weights W_{test}, the Poisson log-likelihood is maximized for the test data X_{test} using the training type matrix. This is equivalent to performing non-negative Poisson regression of each sample column in X_{test} on the ASVs of assemblages in T_{train}. The test year weight matrices W_{test} were calculated. Generalized linear models (GLMs) and Random Forests (RFs) were trained on sample-over-assemblage proportions from each year. Testing was carried out on the alternate year's W_{test} to classify season, depth, Synechoccoccus and *Prochlorococcus* categories. The best training and testing classification accuracies achieved were recorded in Table 3.2.

Training and testing classification accuracies were calculated using all ASVs for comparison and these results are displayed in Table 3.3. ASV counts were also permuted in training data sets (randomized row entries of observed sample-by-ASV matrix) in order to destroy any connection between communities and covariates. Permuting ASV counts reduced prediction accuracies to random chance (approximately 50%) confirming that there is a community signal in the NMF mixture weights that captures a true predictive relationship between subcommunities and environment.

We performed prediction using Random Forests and logistic regression trained on all assemblage proportions as well as specific assemblages associated with particular covariates. For example, as will be seen in the next section, *Synechococcus* was strongly associated with NMF A5 and *Prochlorococcus* was associated with A4. Cyanobacteria cell concentration categories were predicted with training accuracies of approximately 83% for 2015 and 89% for 2016 (leave-one-out validation with RFs). *Prochlorococcus* produced about the same result of 80% leave-one-out training accuracy for 2015 but 95% for 2016. The training accuracy was the same or slightly lower when only a subset of associated assemblages was trained. Training on 2016 and testing on 2015 *Prochlorococcus* classes yielded a better prediction accuracy of 84% for associated assemblages compared to 81% from all assemblages. Testing on a 2015 subset of *Synechococcus* related assemblages had classification of 70% up from 65% for all assemblages. All these results for training and testing accuracy based on supervised NMF test matrices for 2015 and 2016 are recorded alongside predictions in Table 3.2.

Table 3.2 Training and testing accuracies for classification using supervised NMF assemblages. Parentheses for cyanobacteria classes show prediction accuracies from training on subset of characteristic assemblages.

Supervised NMF Classification Accuracy					
Train	**Test**	**Season**	**Depth**	***Synechococcus***	***Prochlorococcus***
2015	2015	96%	88%	83% (82% with A1, A5)	80% (80% with A3, A4)
2015	2016	75%	85%	65% (70% with A1, A5)	88% (85% with A3, A4)
2016	2016	98%	83%	89% (83% with A1, A3)	95% (90% with A1, A2)
2016	2015	75%	67%	72% (62% with A1, A3)	81% (84% with A1, A2)

Table 3.3 Training and testing accuracies for classification using all ASVs

All Taxa Classification Accuracy					
Train	**Test**	**Season**	**Depth**	***Synechococcus***	***Prochlorococcus***
2015	2015	94%	74.5%	88% (2%/21%)	86% (10%/17%)
2015	2016	80%	81%	69% (19.6%/42%)	90% (11%/8%)
2016	2016	93%	82.7%	88 (9%/14%)	97% (4%/2%)
2016	2015	68%	66%	72 (27%/27%)	65% (25%/44.7%)

BioMiCo and NMF results were comparable or outperformed classification based on all ASV taxa in test cases of season and depth classes, except testing accuracy for season in 2016. In that instance the prediction accuracy was still well within a margin of error, all taxa only performed approximately one percentage point better than BioMiCo. The advantage of assemblage representations for prediction of environmental features is clear. Not only do assemblages intrinsically capture latent subcommunity relationships and perform the same or better for classification but they reduce the dimension and variance of noisy ASV data.

SuRF was run with 50 subsamples and 100 permutations for sparse selection of the best predictor taxa in each year at a p-value of 0.05. The selected taxa produced by

SuRF were recorded; GLMs and RFs were trained on those taxa with the same response variable classes already used. We then tested the linear and non-linear models with the predictor taxa observations in the test year. Results are reported in Table 3.4 with the number of taxa predictors in parentheses. Some SuRF prediction did not perform as well as BioMiCo or NMF assemblage proportions. Considering that only 1 to 4 taxa out of approximately 2500 were used for prediction the accuracies seemed commensurate. SuRF did not leverage any season or depth supervising factors either. It could be argued that assemblages that fuse community distributions over taxa with spatial-temporal information provide a better model than even the most rigorously selected predictor species. Table 3.5 contains lists of SuRF taxa sorted by response variable used to select predictors. The highest contributing ASVs identified within NMF assemblages are also given in table 3.5. Many top contributing (predominant) assemblage taxa correspond with SuRF explanatory ASVs, notably Thermoplasmata Marine Group II and Alphaproteobacteria SAR11 Clade. This correspondence supports the connection between assemblage proportions and environmental covariates.

Table 3.4 Training and testing classification accuracies using SuRF selected taxa for seasonal, depth, and cyanobacteria response variables.
* An important issue to note is that the best predictor taxa identified in 2016 were not present in 2015. So it was not possible to train on 2016 SuRF taxa and then test on those same taxa in 2015 because the test counts were all zero.

SuRF Classification Accuracy					
Train	**Test**	**Season**	**Depth**	***Synechococcus***	***Prochlorococcus***
2015	2015	96% (2 taxa)	75% (2 taxa)	90% (1 taxa)	86% (3 taxa)
2015	2016	62.5% (2 taxa)	70.5% (2 taxa)	67% (1 taxa)	81% (3 taxa)
2016	2016	98% (2 taxa)	92% (4 taxa)	87% (1 taxa)	98% (3 taxa)
2016	2015	*N/A	66% (4 taxa)	72% (1 taxa)	71% (3 taxa)

Table 3.5 Comparing lists of taxa selected by SuRF and top predominant ASVs identified by NMF. Response category refers to the specific response variable that SuRF was run with to select taxa for each year.

Response Category	SuRF Predictor Taxa	NMF Assemblage Top Predominant Taxa
2015		
Seasonality	- Alphaproteobacteria SAR11 clade - Thermoplasmata Marine Group II	- Candidatus Nitrosopelagicus - Thermoplasmata Marine Group II - Alphaproteobacteria SAR11 clade - Nitrosopumilaceae - Candidatus Actinomarina - Flavobacteriaceae NS5 marine group
Depth Strata	- Alphaproteobacteria Rhodospirillales Magnetospiraceae - Verrucomicrobiae Arctic97B-4 marine group	
Synechococcus	- Flavobacteriaceae Formosa - Alphaproteobacteria Rhodobacterales Rhodobacteraceae	
Prochlorococcus	- Alphaproteobacteria Rhodobacterales Rhodobacteraceae - Flavobacteriaceae NS4 marine group - Parvibaculales PS1 clade	
2016		
Seasonality	- Flavobacteriales NS9 marine group - Verrucomicrobiae Arctic97B-4 marine group	- Candidatus Nitrosopelagicus - Thermoplasmata Marine Group II - Alphaproteobacteria SAR11 clade Clade I - Parvibaculales OCS116 clade - Candidatus Actinomarina - Gammaproteobacteria SAR86 clade
Depth Strata	- Alphaproteobacteria Puniceispirillales SAR116 clade - Verrucomicrobiae Pedosphaerales Pedosphaeraceae - Pirellulaceae Rhodopirellula - Flavobacteriales Cryomorphaceae NS10 marine group	
Synechococcus	- Gammaproteobacteria SAR86 clade - Flavobacteriaceae NS5 marine group	
Prochlorococcus	- Alphaproteobacteria SAR11 clade	

	- Flavobacteriaceae NS5 marine group - Gammaproteobacteria SAR86 clade	

3.4 ASSEMBLAGE CHARACTERISTICS

In section 3.2 NMDS projection plots showed covariates projected on sample dissimilarity and assemblage membership. This initially indicated which assemblages might be related to environmental vectors of cyanobacteria concentration, nitrogen compounds or density of the water column. Time series plots of the assemblage proportions at different depths were the next step towards characterizing heterotrophic subcommunity succession. In Figures 3.15 to 3.18 below, NMF and STM proportions are plotted over time and facetted for each depth.

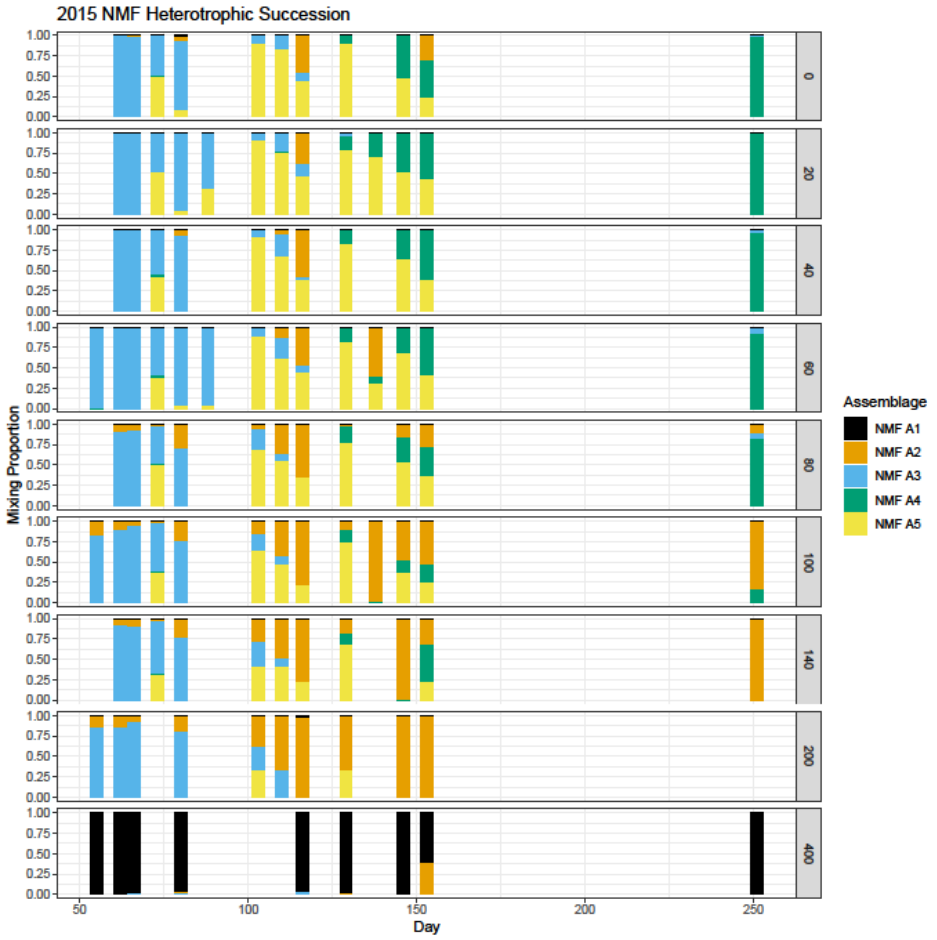


Figure 3.16 Bar plots of 2015 NMF assemblage weights over time and depth.

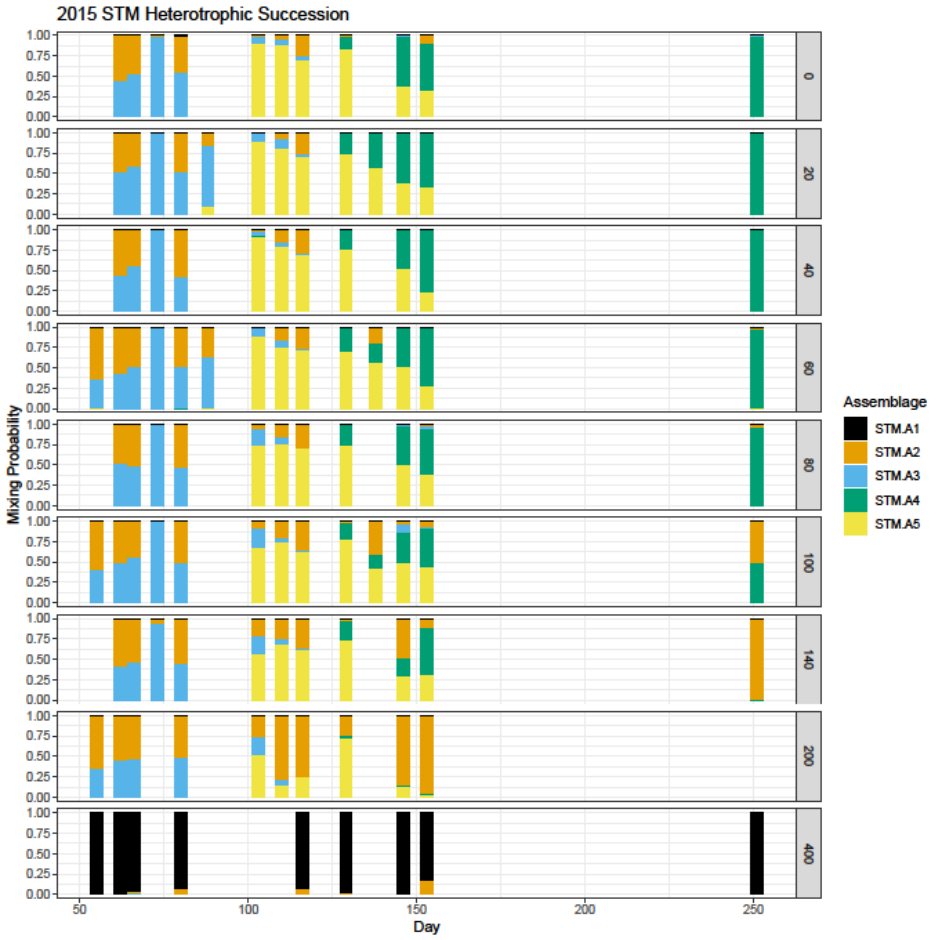


Figure 3.17 Bar plots of 2015 STM assemblage weights over time and depth.

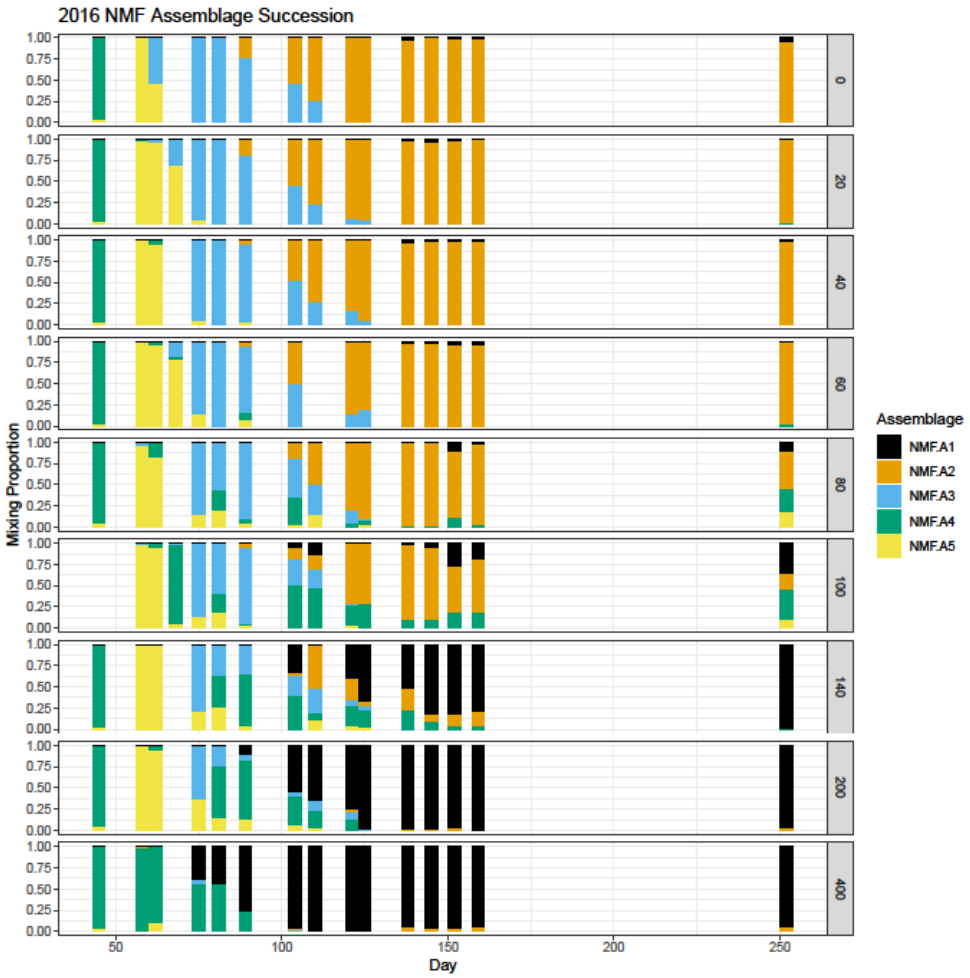


Figure 3.18 Bar plots of 2016 NMF assemblage weights over time and depth.

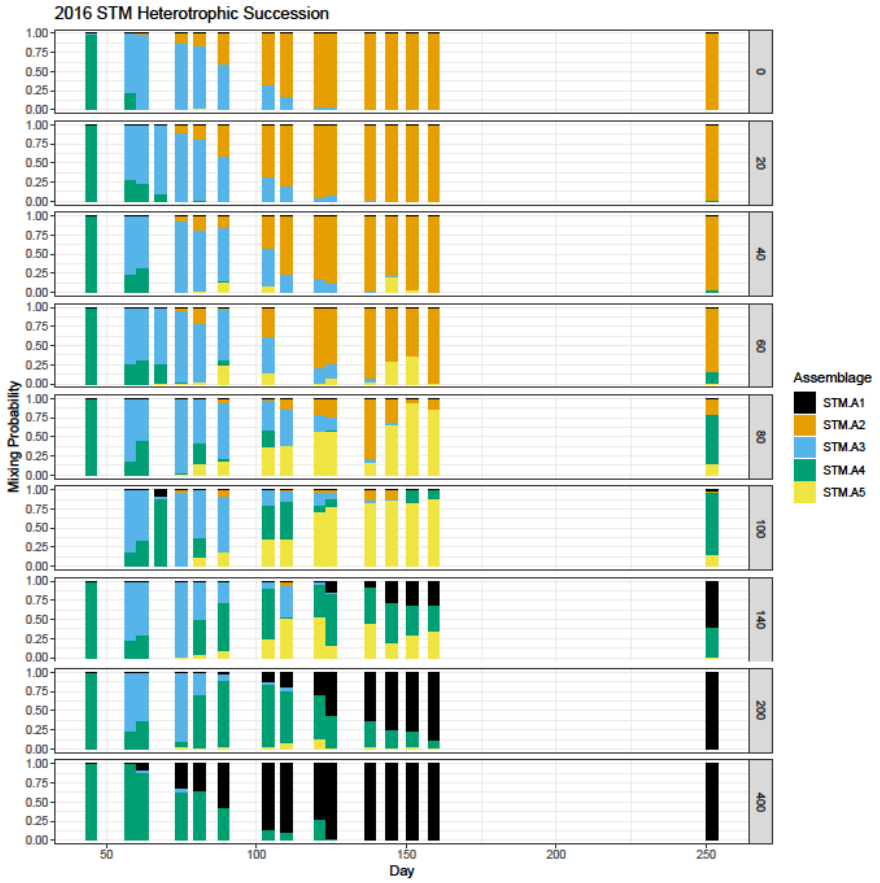


Figure 3.19 Bar plots of 2016 STM assemblage weights over time and depth.

Figures 3.16 to 3.19 show the succession of assemblages over seasons and with changes in water column variation and stability at different depths. The assemblage weights are not just used for prediction. They give information about how the community transitions in response to spatial-temporal changes and these plots of community transition show methodological similarities.

Spearman rank-based correlation analysis was performed with sample over assemblage proportions on the other biotic and abiotic covariates measured. Factors that had significant correlation coefficients with p-values less than 0.01, or marginally significant less than 0.05, suggested the strength of relationships between assemblages

and different explanatory covariates. The results are summarized in sections 3.4.1 and 3.4.2.

As mentioned in the introduction of the data (section 1.1), both taxonomic and metabolic pathway abundances were processed. Assemblage weights were highly correlated with particularly important pathway abundances (as referenced for example in Caspi et al. 2018). Spearman correlation and robust regression was also performed with the assemblage proportions and the most relevant pathways. A Wald test was used to assess significance of coefficients for robust regressions. Robust regression was done as an extra check on associations because there appeared to be many outliers in the pathway abundance data. Examples of correlation results are presented in Figures 3.20 to 3.23. In many cases, significant relationships were found with functional potential that reinforced the ecological interpretation of assemblages as subcommunities. Detailed heterotrophic assemblage associations are discussed in the next sections for separate years because some covariates and pathway measurements were different between 2015 and 2016.

3.4.1 Assemblage Associations in 2015

In 2015 NMF Assemblage 1 (NMF A1) and STM Assemblage 1 (STM A1) accounted for virtually all of the assemblage mixing proportion (posterior probability density) at 400m depth for all seasons. The water column of the Red Sea is permanently stratified below 300 m (Edwards 1987 and Stambler 2005), so a single predominant subcommunity at 400m was expected. Both NMF A1 and STM A3 had smallest p-values (< 0.001) for water density (depth), NO_2, NO_3, TON, PO_4, oxygen and salinity. These assemblages had significant negative regression coefficients with photosynthetic and

aerobic respiration pathways: PHOTOALL-PWY: oxygenic photosynthesis, PWY-101: photosynthesis light reactions, PWY-241: C4 photosynthetic carbon assimilation cycle NADP-ME type, PWY-7117: C4 photosynthetic carbon assimilation cycle PEPCK type, PWY-3781: aerobic respiration I (cytochrome c) and PWY-7279: aerobic respiration II (cytochrome c) (yeast). According to Gilbert et al. (2010) photosynthetic pathways can be greatly affected in winter seasons and when daylight is more restricted below the photic zone in deeper depths. Respiratory metabolism genes have also been observed to be more abundant at night and in less oxygenated water (Gilbert et al. 2010). Other significant ($p < 0.05$) pathways were PWY490-3: nitrate reduction VI (assimilatory) and PWY-3661: glycine betaine degradation I. Spearman and robust regression found similar correlations with these assemblages and their most predominant taxa including *Nitrosopumilaceae*. *Nitrosopumilaceae* are chemolithoautotrophs that grow by oxidizing inorganic nitrogen compounds including ammonia. Some of these species are able to use urea as a source of ammonia for storage, assimilation and nitrification (Könneke et al. 2005). Sinking organic compounds from surface waters pass through the middle water column and become locked into a deeper water stratum. This alters the chemical environment with increased abundance of nitrogen and phosphate molecules influencing the community at these depth. Other major contributing species in these assemblages were part of the known deep sea *Thermoplasmata Marine Group II*. For another example, regression also found correlations with PWY-3661: glycine betaine degradation I. This pathway is a potential catabolic mechanism for osmoregulation. Cells would experience increased osmotic stress at deeper depths like 400m where salinity and solute concentrations are higher (Ren et al. 2017). The association of these assemblages with this pathway

response to osmotic stress helps confirm our characterization of NMF A1 and STM A3 as a deep water subcommunity.

NMF A2 and STM A2 contributions were highest at mid to deeper depths 60-400m throughout the spring, day 80-85, 110-115 and 130-140. Both assemblages showed similar mixing proportions over the same seasonal periods. Their significant regression p-values were for nitrite (NO_2), *Prochlorococcus* (STM only), temperature (STM only) and water density. The most predominant ASV in NMF/STM A2 was *Candidatus Nitrospelagicus*. *Nitrospelagicus* are ammonia oxidizing archaea (AOA) which would be expected to be present in an assemblage associated with the products of nitrification. Regressions found significant correlations with PWY-3661: glycine betaine degradation I ($p < 0.05$) and PWY-5505: L-glutamate and L-glutamine biosynthesis. In microorganisms the latter pathway and its products are used for ammonia assimilation and glutamate also serves as a storage form of ammonia (Caspi et al. 2018). The confluence of these correlations distinguished these assemblages as being associated with deeper water nitrogen metabolism processes during the spring blooms.

NMF A3 and STM A3 from 2015 were positively associated with NO_2 and negatively associated with *Prochlorococcus* and temperature (p-values < 0.001) with greatest significant seasonal contributions in the mixed water column (0-200m) of the winter and early spring, especially prior to day 75 (February to mid-March). *Parvibaculales OCS116* was a highly contributing taxon negatively correlated with PWY-7198: pyrimidine deoxyribonucleotides de novo biosynthesis IV which produces ammonia. *Candidatus Nitrospelagicus* was also predominant, parallel to the previous nitrification associated assemblages discussed in the spring at lower depths (NMF A2,

STM A2). Another taxonomic class exclusive to NMF A3 and STM A5 was Gammaproteobacteria HOC36, to which ammonia oxidizers are known to belong. There is evidence then that NMF and STM A3 captured a subcommunity related to nitrification in the winter and the onset of the first spring *Synechococcus* bloom.

NMF and STM A4 had largest mixing proportions at the surface and upper photic zone (0-60m) during the late spring and summer (after day 125), in contrast to greater NMF/STM A2 proportions below 60m. The assemblages were significantly correlated with *Prochlorococcus*, NO_2, temperature, irradiance and density as well as pathways: PHOTOALL-PWY: oxygenic photosynthesis, PWY-101: photosynthesis light reactions, PWY-3781: aerobic respiration I (cytochrome c) and PWY-5505: L-glutamate and L-glutamine biosynthesis. A4 had a negative regression coefficient for Synechococcus suggesting a negative correlation. Since in 2015 *Synechococcus* and *Prochlorococcus* bloom cycles were negatively correlated with each other, it was not surprising that a *Prochlorococcus* assemblage correlation would be negatively correlated with *Synechococcus*. Robust regression also found local correlations with chlorophyllide a biosynthesis I (aerobic, light-dependent) and Pwy-5505 also related to ammonia assimilation and metabolic growth from a nitrogen source. Predominant heterotrophs were *Candidatus Actinomarina* and taxa belonging to the *Alphaproteobacteria SAR11* clade which are known to be dominant in surface water (Cram et al. 2015). It would be expected to find these taxa and pathways in the photic zone in the late spring and summer during *Prochlorococcus* blooms. Linear correlations with pathways related to the characteristics of the assemblage supported the association of both A4s with *Prochlorococcus*.

NMF A5 and STM A5 were characterized by surface depths (0-60m) during the spring *Synechococcus* blooms. Synechococccus and depth had the lowest p-values followed by NO_2, PO_4, TON and oxygen. Correlations were also significant with the following photosynthetic and metabolic respiratory pathways: PHOTOALL-PWY: oxygenic photosynthesis, PWY-101: photosynthesis light reactions, PWY-241: C4 photosynthetic carbon assimilation cycle, NADP-ME type, PWY-7117: C4 photosynthetic carbon assimilation cycle, PEPCK type, PWY-3781: aerobic respiration I (cytochrome c), PWY-7279: aerobic respiration II (cytochrome c) (yeast) and PWY490-3: nitrate reduction VI (assimilatory). The most predominant distinct heterotrophs included Candidatus Actinomarina and Rhodobacteraceae *Roseovarius*. Assemblage dynamics and correlations with pathways common for phototrophs characterized this assemblage as a subcommunity present at surface depths during the spring blooms of *Synechococcus*.

When considering the entire community with cyanobacteria, the models assigned 6 assemblages with different compositions, dynamics and associations than the heterotropic community assemblages. The additional 6th assemblage had mixing proportions greatest at mid depths of 60-200m and 60-100m in the late spring after day 125 (early May) and summer. The 6th NMF assemblage had smallest p-values (< 0.001) from regression on *Prochlorococcus*, temperature and oxygen concentration. The 6th STM assemblage likewise had smallest p-values for *Prochlorococcus* and oxygen, but also NO3, PO4, and TON – capturing the deeper water predominant association with sinking nutrients. Since Prochlorococcus ASVs were included in the community model

for this 6th assemblage, regression of Prochlorococcus on the assemblage was not very informative. This is the why the heterotrophs were modelled separately.

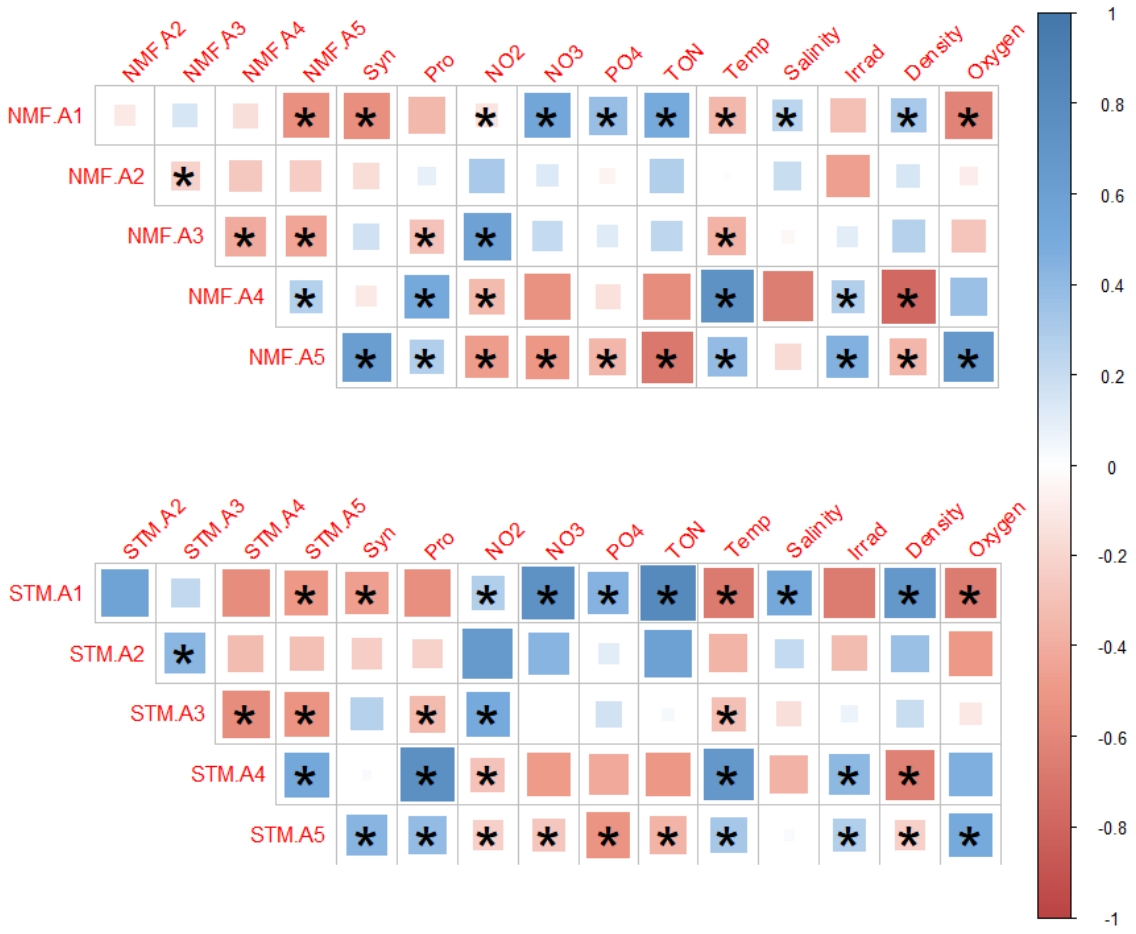


Figure 3.20 Correlation plots for 2015 NMF and STM assemblages with environmental covariates. Asterisks indicate significant correlations at the 0.01 level. The colour bar and square sizes indicate the strength and direction of correlation.

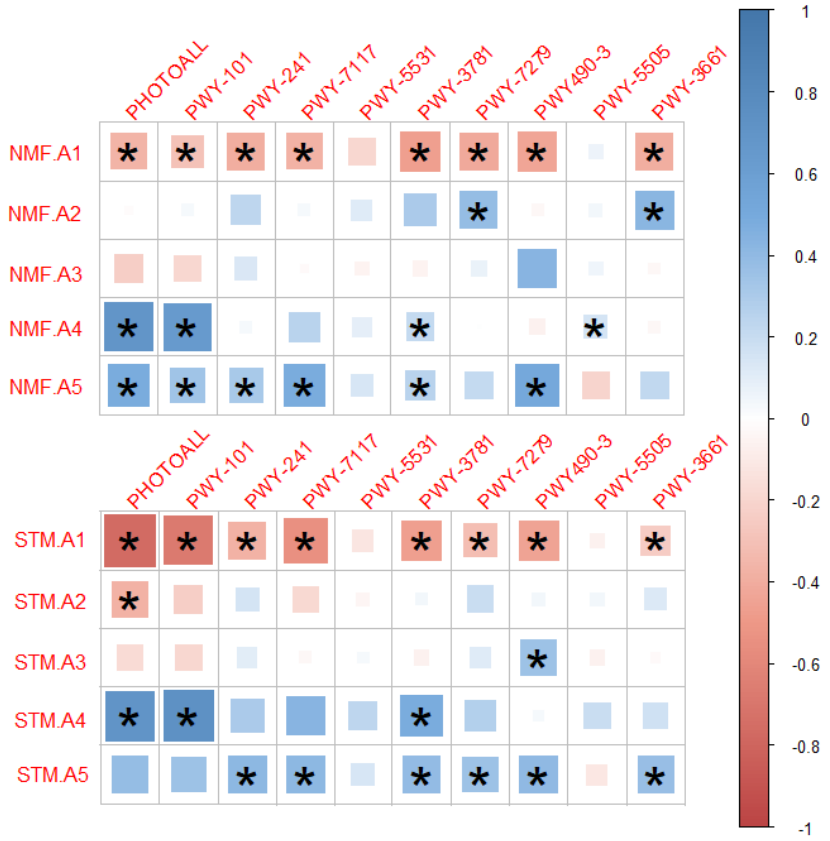


Figure 3.21 Correlation plots for 2015 NMF and STM assemblages with important metabolic pathways. Asterisks indicate significant correlations at the 0.01 level. The colour bar and square sizes indicate the strength and direction of correlation. Correlations between assemblages are not shown because the correlation analysis was conducted with a subset of samples for which pathway abundances were available. There were more samples, and therefore, more informative assemblage weights available for environmental factors.

3.4.2 Assemblage Associations in 2016

In 2016 NMF A1 and STM A1 had similar deep water associations as their 2015 counterparts except that seasonality and irradiance were significant in 2016 and salinity was no longer significant. Pathway regressions analogous to 2015 results were present except glycine betaine degradation was not significant. The 2016 deepest water assemblage was predominant later in the seasonality of the year as seen in Figure 3.17

and 3.18 and negatively associated with cyanobacteria. NMF A2 and STM A2 were predominant in the spring and later summer at upper and surface depths above 100m. These assemblages had strongest positive correlations with *Prochlorococcus* and temperature and strong negative associations with nitrite, nitrate and phosphate. Significant positive correlations with pathway abundances for PHOTOALL-PWY: oxygenic photosynthesis and PWY-101: photosynthesis light reactions supported the association with the more synchronized cyanobacteria blooms in 2016. NMF A3 and STM A3 were predominant at depths above 200m and before day 100 (early spring). The most significant correlations were a positive association with *Synechococcus* and a negative association with phosphate. Both assemblages were correlated with photosynthetic and aerobic respiration pathways as well as a significant strong positive correlation with PWY 490-3 nitrate reduction. NMF A4, A5 and STM A4 showed greatest contribution in winter at all depths with a sustained predominance at 400m and strongest positive correlation with nitrite. These assemblages were also negatively associated with cyanobacteria concentrations and temperature. Some marginally significant negative correlations with photosynthesis pathways existed as well as stronger association with PWY-7279: aerobic respiration II (cytochrome c) (yeast). STM A5 was predominant at mid depths (80-100m) and in later spring day 125-150. The assemblage was associated with nitrite, salinity, irradiance, chlorophyll and weakly with osmoregulation PWY-3661.

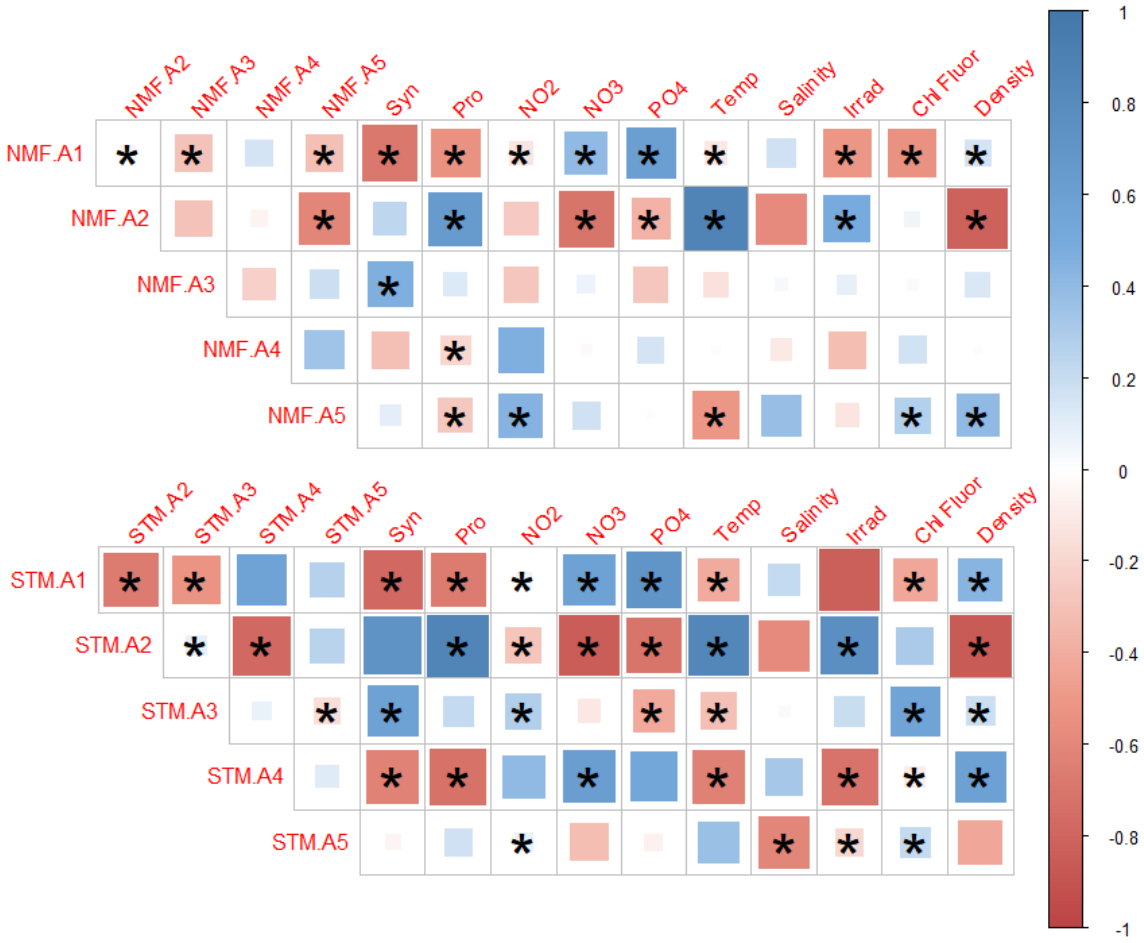


Figure 3.22 Correlation plots for 2016 NMF and STM assemblages with environmental covariates.

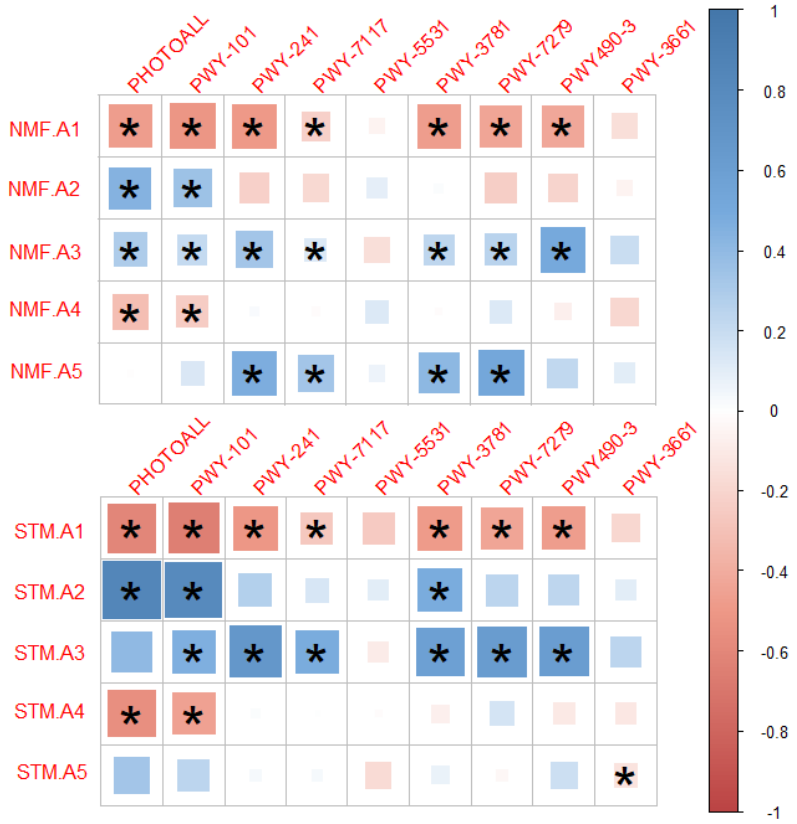


Figure 3.23 Example correlations plots for 2016 NMF and STM assemblages with important metabolic pathways. Correlations between assemblages appear different than in Figure 3.22 because the correlation analysis was conducted with a subset of samples for which pathway abundances were available. There were more samples, and therefore, more informative assemblage weights available for environmental factors.

CHAPTER 4 CONCLUSION AND FUTURE WORK

The application of methods for estimating subcommunities as distributions over a sample collection of species was described in this thesis. Spatial-temporal metadata was incorporated to model assemblages which capture environmental dynamics that interact with microbiome communities. This approach represents an ongoing shift to analysis of interrelated communities and their roles in marine ecosystems (see Fuhrman et al. 2006, Sieradzki et al. 2018, Bálint et al. 2016 and Ren et al. 2017, for other examples). Further, the ability to reduce the variable space of over 2500 taxa down to 5 assemblages was demonstrated for sample communities from winter through to the late summer at depths from the surface down to 400m; this is a necessary capability for all studies of this type. Assemblages derived from methods based on very different model formalisms showed empirically similar distributions over ASVs and 70-90% Bray-Curtis similarity (Figure 3.13 Dendrograms). The mixture of heterotrophic assemblages showed variability structured by changes in seasonality and depth of the water column. Assemblages were correlated with distinct biotic and abiotic factors, thus reinforcing the interpretation that varying heterotrophic assemblage proportions might help represent environmental dynamics. The effects of heterotrophs on cyanobacteria and vice versa were of special importance. Predictive models based on heterotrophic assemblages indeed improved classification of season, water column strata and both *Synechococcus* and *Prochlorococcus* bloom states. Some assemblages characterized by their associations with cyanobacteria proved to be even better predictors of bloom categories. Prediction with the SuRF method of selecting indicator taxa was informative of the advantages and

disadvantages of modelling subcommunities of marine microbiomes compared to the perspective of individual heterotrophic biomarker taxa.

The results related to the Red Sea gave rise to a lot of insights and questions. Several directions for future work continuing with this data are proposed, as well as with new marine microbiomes:

1. A next step of analysis should ascertain if and how heterotrophic communities are driving or responding to blooms, or both. The same methods employed here have been applied to analyze completely separate microbial data from a lake ecosystem. That has established two predominant heterotrophic assemblages, one prior to and one during a cyanobacterial and microcystin toxin bloom in that environment. At each depth the Red Sea data was not sampled at enough time points (only 11-16 dates) nor at evenly spaced indices, which meant time series analysis was not possible. Interpolating between time points would have introduced substantial bias and resulted in many false positive correlations. Anything beyond linear correlations was not appropriate for the data. With that said, more densely sampled time series data could help reveal locally lagged cross-correlations between assemblages and covariates or metabolic pathways of interest. A time delay lag between two time series can potentially be evidence for a causal driving relationship.
2. The processing of the Gulf of Aqaba (Red Sea) data requires a reference library to identify ASVs from 16S sequences, and that library contained human gut taxa as well as marine microbes. A reference library of marine microbial genomes specifically from the Red Sea would facilitate better classification confidence of

species. Haroon, Thompson et al. (2016) reported such a database assembled from metagenomes of microorganisms in the Red Sea at a wide variety of depths and locations that could serve this purpose.

3. Further functional analysis of the Gulf of Aqaba (Red Sea) metagenomic data is required to link the taxonomic communities to subnetworks of substrate-product metabolic reaction pairs. BiomeNet developed by Shafei et al., (2014) is an unsupervised mixed membership hierarchical framework that models the latent biochemical networks of pathways for exactly this purpose. It is more generally our goal to develop new supervised methods for relating microbial communities to specific substrate-product reactions.
4. Ultimately we want to provide deeper insight into statistical methods for modelling subcommunities and how they are applied to real, and highly complex, data like that from the of the Gulf of Aqaba (Red Sea). We need to develop and refine methods for estimating how subcommunities overlap in functional traits and environmental niches. An evaluation of model performance should include an objective measure of subcommunity robustness, generalizability to other environments and variability of community composition and dynamics.

BIBLIOGRAPHY

[1] E. M. Airoldi and J. M. Bischof. A regularization scheme on word occurrence rates that improves estimation and interpretation of topical content (with discussion). Journal of American Statistical Association, 2016.

[2] E.M. Airoldi. J.M. Bischof. Improving and Evaluating Topic Models and Other Models of Text. Journal of the American Statistical Association; 111(516), 1381–1403, 2016.

[3] E. M. Airoldi and J. M. Bischof. Summarizing topical content with word frequency and exclusivity. In International Conference on Machine Learning, volume 29, Edinburgh, Scotland, UK, 2012.

[4] M. Bálint, M. Bahram, A.M. Eren, K. Faust, J.A. Fuhrman, B. Lindahl, R.B. O'Hara, M. Öpik, M.L. Sogin, M. Unterseher, L. Tedersoo. Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. FEMS Microbiol. Rev. 40 686–700, 2016.

[5] E Boon, CJ Meehan, C Whidden, DH Wong, MG Langille, RG Beiko. Interactions in the microbiome: communities of organisms and communities of genes. *FEMS Microbiol Rev*; 38(1):90–118, 2014.

[6] E Boon, CJ Meehan, C Whidden, DH Wong, MG Langille, RG Beiko. Interactions in the microbiome: communities of organisms and communities of genes. *FEMS Microbiol Rev*; 38(1):90–118, 2014.

[7] D. M. Blei, Alp Kucukelbir, Jon D McAuliffe. Variational Inference: A Review for Statisticians. Journal of the American Statistical Association, Vol. 112, Iss. 518, 2018.

[8] D. M. Blei and J. D. Lafferty. A correlated topic model of science. AAS, 1(1):17–35, 2007.

[9] Jean-Philippe Brunet, P. Tamayo, T.R. Golub, J.P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the National Academy of Sciences of the United States of America vol. 101, 12, 2004.

[10] Y. Cai, H. Gu, T. Kenney. Learning Microbial Community Structures with Supervised and Unsupervised Non-negative Matrix Factorization. Microbiome volume 5; Article number: 110, 2017.

[11] B. Callahan, P. McMurdie, S. Holmes. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J 11, 2639–2643, 2017.

[12] Ron Caspi, Richard Billington, Carol A Fulcher, Ingrid M Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Peter E Midford, Quang Ong, Wai Kit Ong, Suzanne Paley, Pallavi Subhraveti, Peter D Karp. The MetaCyc database of metabolic pathways and enzymes, Nucleic Acids Research, Volume 46, Issue D1, Pages D633–D639, 2018.

[13] J. Cram, C. Chow, R Sachdeva. Seasonal and interannual variability of the marine bacterioplankton community throughout the water column over ten years. ISME J 9, 563–580, 2015.

[14] A. P. Dempster, N. Laird, D.B. Rubin. Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Association, 39:1–38, 1977.

[15] F. J. Edwards. Climate and oceanography, in Key Environments: Red Sea. Pergamon, Oxford, U. K, pp. 45–69, 1987.

[16] Naoki Egami, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, Brandon M. Stewart. How to Make Causal Inferences Using Texts, Appendix A.5.1, 2018.

[17] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, David Botstein. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America vol. 95, 25: 14863-8, 1998

[18] K. Faust, L. Lahti, D. Gonze, W.M. de Vos, J. Raes. Metagenomics meets time series analysis: unraveling microbial community dynamics. Current Opinion in Microbiology 25, 56–66, 2015.

[19] E.A. Franzosa, L.J. McIver, G. Rahnavard. L.R. Thompson, M. Schirmer, G. Weingart, K.S. Lipson, R. Knight, J.G. Caporaso, N. Segata, C. Huttenhower. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods; 15(11):962–968, 2018.

[20] J. A. Fuhrman, I. Hewson, M.S. Schwalbach, J.A. Steele, M.V. Brown, and S. Naeem. Annually reoccurring bacterial communities are predictable from ocean conditions. Proceedings of the National Academy of Sciences of the United States of America, 103(35), 13104–13109, 2006.

[21] Renaud Gaujoux and Seoighe Cathal. A flexible R package for nonnegative matrix factorization. BMC bioinformatics vol. 11 367, 2010.

[22] A. Gelman and J. Hill. Data analysis using regression and multilevel/hierarchical models, volume 3. Cambridge University Press New York, 2007.

[23] A. Gelman, X.L. Meng, H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. Statistica sinica, 6(4):733–760, 1996.

[24] Jack A. Gilbert , Dawn Field, Paul Swift, Simon Thomas, Denise Cummings, Ben Temperton, Karen Weynberg, Susan Huse, Margaret Hughes, Ian Joint, Paul J. Somerfield, Martin Mühling. The taxonomic and functional diversity of microbes at a temperate coastal site: a 'Multi-Omic' study of seasonal and diel temporal variation. PLoS ONE; 5:e15545, 2010.

[25] Mohamed F. Haroon, Luke R. Thompson, Donovan H. Parks, Philip Hugenholtz, Ulrich Stingl. A catalogue of 136 microbial draft genomes from Red Sea metagenomes. Scientific data vol. 3 160050, 2016.

[26] M. Könneke, A.E. Bernhard, J.R. de la Torre, C.B. Walker, J.B. Waterbury, D.A. Stahl. Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature 437 543–546, 2005.

[27] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, in: Advances in neural information processing systems, pp. 556–562, 2001.

[28] D. Lee, H. Seung. Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791, 1999.

[29] J.S. Liu. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. Journal of the American Statistical Association, 89(427):958– 966, 1994.

[30] Lihui Liu, Hong Gu, Johan Van Limbergen, Toby Kenney. SuRF: a New Method for Sparse Variable Selection, with Application in Microbiome Data Analysis. arXiv:1909.06439, 2019

[31] C. Lucas, R. Nielsen, M.E. Roberts, B.M. Stewart, A. Storer, D. Tingley. Computer assisted text analysis for comparative politics. Political Analysis, 23(2):254–277, 2015.

[32] Stefano Monti, Pablo Tamayo, Jill Mesirov, Todd Golub. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data Machine Learning, 52, 91–118, 2003

[33] X.L. Meng and D. Van Dyk. The em algorithm—an old folk-song sung to a fast new tune. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 59(3):511–567, 1997.

[34] D. Mimno and D Blei. Bayesian checking for topic models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 227–237. Association for Computational Linguistics, 2011.

[35] D. Mimno, H.M. Wallach, E. Talley, M Leenders, A. McCallum. Optimizing semantic coherence in topic models. In: Proceedings of the conference on empirical methods in natural language processing. Stroudsburg, PA: Association for Computational Linguistics, pp. 262–272, 2011.

[36] D. Newman, J.H. Lau, K. Grieser, T. Baldwin. Automatic evaluation of topic coherence. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100–108. Association for Computational Linguistics, 2010.

[37] A. Pascual-Montano, J.M. Carazo, K. Kochi, D. Lehmann, R.D. Pascual-marqui. Nonsmooth nonnegative matrix factorization (nsNMF) IEEE Trans Pattern Anal Mach Intell. 28:403–415, 2006

[38] Z. Ren, F. Wang, X. Qu, J.J. Elser, Y. Liu, L. Chu. Taxonomic and Functional Differences between Microbial Communities in Qinghai Lake and Its Input Streams. Front Microbiol; 8:2319, 2017.

[40] Margaret Roberts, Stewart Brandon, E. M. Airoldi. A model of text for experimentation in the social sciences. Journal of the American Statistical Association. 2016..

[42] Margaret Roberts, Stewart Brandon, Dustin Tingley, C. Lucas, J. Leder-Luis, S. Gadarian, B. Albertson, D. Rand. Structural topic models for open-ended survey responses. American Journal of Political Science, 58(4):1064–1082, 2014.

[43] Margaret Roberts, Stewart Brandon, Dustin Tingley. Navigating the local modes of big data: The case of topic models. In Data Analytics in Social Science, Government, and Industry. Cambridge University Press, New York, 2015.

[44] M. Shafiei, K.A. Dunn, E. Boon, S.M. MacDonald, D.A. Walsh, H. Gu, J.P. Bielawski. BioMiCo: a supervised Bayesian model for inference of microbial community structure. Microbiome, 3, 8, 2015.

[45] E.T. Sieradzki, J.A. Fuhrman, S. Rivero-Calle, L. Gómez-Consarnau. Proteorhodopsins dominate the expression of phototrophic mechanisms in seasonal and dynamic marine picoplankton communities. PeerJ; 6:e5798, 2018

[46] N. Stambler. Bio-optical properties of the northern Red Sea and the Gulf of Eilat (Aqaba) during winter 1999, J. Sea Res., 54(3), 186– 203, 2005

[47] M. Taddy. Multinomial Inverse Regression for Text Analysis. Journal of the American Statistical Association, 108(503), 755–770, 2013.

[48] L.R. Thompson, G.J. Williams, M.F. Haroon, A. Shibl, P. Larsen, J. Shorenstein, R. Knight, U. Stingl. Metagenomic covariation along densely sampled environmental gradients in the Red Sea. ISME J 138-151, 2017.

[49] G.F. Triantafyllou, G. Yao, K.P. Petihakis, D. Tsiaras, E. Raitsos, I. Hoteit. Exploring the Red Sea seasonal ecosystem functioning using a three-dimensional biophysical model. J. Geophys. Res. Oceans, 119, 1791– 1811, 2014.

[50] H.M. Wallach, I. Murray, R. Salakhutdinov, D Mimno. “Evaluation Methods for Topic Models.” In Proceedings of the 26th Annual International Conference on Machine Learning, pp. 1105–1112. ACM, 2009.

[51] C. Wang and D. Blei. Variational inference in nonconjugate models. Journal of Machine Learning Research, 14:1005–1031, 2013.

[52] C.T. Webb, J.A. Hoeting, G.M. Ames, M.I. Pyne. Poff N. LeRoy. A structured and dynamic framework to advance traits-based theory and prediction in ecology. Ecol Lett; 13:267–283, 2010.

[53] L.H. Zeglin. Stream microbial diversity in response to environmental changes: review and synthesis of existing research. Frontiers in Microbiology6, 2015.