Environment International 124 (2019) 312–319

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Metabarcoding and machine learning analysis of environmental DNA in ballast water arriving to hub ports

William A. Gerhard, Claudia K. Gunsch*

Duke University, Department of Civil and Environmental Engineering, 121 Hudson Hall, Durham, NC 27708-0287, United States

ARTICLE INFO

Handling Editor: Frederic Coulon

Keywords:
Ballast water
Microbiome
Environmental DNA
Biomarker
Machine learning
High throughput sequencing

ABSTRACT

While ballast water has long been linked to the global transport of invasive species, little is known about its microbiome. Herein, we used 16S rRNA gene sequencing and metabarcoding to perform the most comprehensive microbiological survey of ballast water arriving to hub ports to date. In total, we characterized 41 ballast, 20 harbor, and 6 open ocean water samples from four world ports (Shanghai, China; Singapore; Durban, South Africa; Los Angeles, California). In addition, we cultured *Enterococcus* and *E. coli* to evaluate adherence to International Maritime Organization standards for ballast discharge. Five of the 41 vessels – all of which were loaded in China – did not comply with standards for at least one indicator organism. Dominant bacterial taxa of ballast water at the class level were Alphaproteobacteria, Gammaproteobacteria, and Bacteroidia. Ballast water samples were composed of significantly lower proportions of Oxyphotobacteria than either ocean or harbor samples. Linear discriminant analysis (LDA) effect size (LEfSe) and machine learning were used to identify and test potential biomarkers for classifying sample types (ocean, harbor, ballast). Eight candidate biomarkers were used to achieve 81% (k nearest neighbors) to 88% (random forest) classification accuracy. Further research of these biomarkers could aid the development of techniques to rapidly assess ballast water origin.

1. Introduction

The volume of total ballast water discharges to ports in the United States has grown significantly over the last decade (Gerhard and Gunsch, 2018). This increase in ballast water discharge may provide additional opportunity for the accidental co-discharge, and introduction, of aquatic invasive species (Bax et al., 2003; Carlton, 2001). Following a ballast-associated cholera outbreak in Peru during the 1990s, a renewed focus was placed on the role of ballast water management in preventing accidental microbial introduction (McCarthy and Khambaty, 1994; Ruiz et al., 2000).

Ballast water is a known vector for the global proliferation of pathogens (Aguirre-Macedo et al., 2008; Drake et al., 2005, 2007; Ruiz et al., 2000). Recent research has discovered that ballast water may also serve as a vector for the global movement of antibiotic resistance genes (ARGs) (Ng et al., 2018). The World Health Organization has called the global proliferation of antibiotic resistance the greatest risk to human health in the 21st century (WHO, 2014). The possibility of ARG translocation in ballast water further necessitates the need for effective ballast water management and regulation. As a result, there has been a push by researchers and regulators to develop tools for rapid measurement of vessel compliance prior to discharge (Drake et al., 2014; Egan et al., 2015; Emami et al., 2012; Fykse et al., 2012).

Effectively advising the development of ballast water management techniques requires the use of advanced characterization technologies. One such characterization technology is high throughput sequencing (HTS), which has made large improvements in cost and accuracy over the last decade (Czaplicki and Gunsch, 2016; Shokralla et al., 2012). These improvements have brought HTS into the mainstream of environmental science, and it is used to augment analysis of environmental questions from bioremediation to water quality to the impacts of air pollution (Adar et al., 2016; Gwin et al., 2018; Lefèvre et al., 2018; Staley et al., 2013). DNA in environmental matrices is often referred to as environmental DNA or eDNA and can be used to analyze community ecology dynamics that may not be feasible to examine via other methods (Brady, 2007; Li et al., 2018; Stoeck et al., 2018; Thomsen and Willerslev, 2015). In addition, HTS has been shown to be well-suited to identifying novel biomarkers for classification (Tan et al., 2015).

Abbreviations: ASV, amplicon sequence variants; DADA, Divisive Amplicon Denoising Algorithm; EEZ, exclusive economic zone; HTS, high throughput sequencing; IMO, International Maritime Organization; kNN, k nearest neighbors; LDA, linear discriminant analysis; MPN, most probable number; NMDS, non-metric multidimensional scaling; OOB error, out-of-bag error; OTU, Operational Taxonomic Unit; PCoA, Principal Coordinates Analysis

* Corresponding author.
E-mail address: ckgunsch@duke.edu (C.K. Gunsch).

https://doi.org/10.1016/j.envint.2018.12.038
Received 17 October 2018; Received in revised form 17 December 2018; Accepted 17 December 2018
Available online 17 January 2019

The application of HTS to examine environmental DNA in ballast water is still a developing field with relatively little published research. Recent studies have used HTS to examine different portions of the ballast microbiome, including the 16S rRNA gene, eukaryotic 18S rRNA gene, and viral fractions (Brinkmeyer, 2016; Darling et al., 2018; Kim et al., 2015; Lymperopoulou and Dobbs, 2017; Ng et al., 2015). Previously published HTS studies have examined relatively small sample sizes, with the largest eukaryotic 18S rRNA gene study examining 39 ballast water samples (Darling et al., 2018), the largest bacterial 16S rRNA gene study examining 17 ballast water samples (Lymperopoulou and Dobbs, 2017), and the largest viral study examining five ballast water samples (Kim et al., 2015). In addition, all of these studies were performed on ballast water arriving in a single port.

The goals of the present study were to: 1) Perform ballast water metabarcoding analysis on a larger sample size and geographic scale than previously published studies; 2) Utilize bioinformatic analyses for less-characterized matrices; and 3) Identify biomarkers that may be useful for rapid assessment of ballast water origin. To accomplish these goals, we characterized 41 ballast, 20 harbor and 6 open ocean water samples via HTS. Samples were gathered from four different countries (United States, China, Singapore, South Africa), which allowed intra-study bacteriome comparison of ballast arriving to different ports for the first time. Second, we utilized amplicon sequence variants (ASVs) rather than Operational Taxonomic Units (OTUs) for classification. This type of analysis has been shown to be better suited for less characterized matrices and may be ideal for analyzing ballast or ocean water (Callahan et al., 2017). Third, machine learning was performed to identify candidate bacteria for classification of water type (ocean, harbor, ballast) that may be useful as potential biomarkers for ballast water exchange in future research.

2. Materials and methods

2.1. Site selection

We selected worldwide shipping hubs as sampling sites as these locations are likely to be ecologically important and may have the most potential to impact other areas via ballast water translocation. In addition, preference was placed on sites near partner institutions with the resources to collaborate on this project, and port access must be negotiable in advance of sampling trips via existing industry partnerships. As a result of these criteria, there were no ports along the Atlantic Ocean or in Europe that were included in this study. The ports included in this study, the reason for including them, the sample abbreviation, and their reference notation throughout the paper are the following: 1) Los Angeles/Long Beach, CA – Busiest harbor along the Pacific Coast of the United States, CA, United States; 2) Singapore – Busiest transshipment port in the world, S, Singapore; 3) Durban, South Africa – Busiest port in sub-Saharan Africa, SA, South Africa; and 4) Shanghai, China – Busiest port in the world, CN, China. All of the international research sites are hub cities along global shipping linkages (Wang and Wang, 2011).

2.2. Sample collection

Over a two-year period from September 2015 to August 2017, a total of 20 harbor and 41 ballast water samples were collected from four different ports and analyzed (Table 1). Ballast samples were collected by opening the ballast tank manhole and lowering a 1.2 L Kemmerer sampler to the water below (Wildco, Yulee, Florida). Three distinct 1.2 L samples were drawn from the ballast tank at approximately 1 m below the surface and stored in autoclaved glass bottles on ice for transport to the laboratory. Harbor water samples were collected immediately next to the docks with moored vessels at a depth of approximately 1 m. Harbor samples were collected and analyzed using the same methods as ballast water samples. In addition to harbor and ballast water samples, six open ocean samples were collected from a sailboat in the South China Sea between Singapore and Jakarta, Indonesia. These samples were also collected using the same techniques and depth as ballast and harbor water samples. The specific coordinates of open ocean samples can be found in the Supplementary material (Table S1). Preparation for molecular analyses was performed on sailboat samples using a temporary lab and samples were stored in the ship's freezer until they could be transported to a lab for further analysis.

Table 1
Ballast and harbor samples collected and included in this study.[a]

	Ballast	Harbor	Total
LA/Long Beach, CA	23	7	30
Singapore	7	4	11
Durban, South Africa	4	4	8
Shanghai, China	7	5	12
Total	41	20	61

[a] Excludes six open ocean samples collected in the South China Sea (n = 67).

2.3. Culture-based analyses

Culture-based analyses were performed on all harbor and ballast water samples within 12 h of collection; however, resource limitations at sea prevented their application to ocean samples. The combined analyses required 20 mL of each sample. Total coliform and *E. coli* most probable number (MPN) per 100 mL were measured using IDEXX Colilert® (Westbrook, ME USA) according to the manufacturer's protocol with the slight modification of diluting the sample 10× to account for higher salinity (Microbial Contaminants Method 9223, 2005). Intestinal *Enterococcus* MPN per 100 mL was calculated using IDEXX Enterolert® (Westbrook, ME USA) according to the manufacturer's protocol. Both of these tests are EPA approved methods for quantification of indicator bacteria in water. These tests have been used in geographically-remote laboratory settings and ballast water research previously (Gerhard et al., 2017; Ng et al., 2018).

2.4. Preparation for molecular analysis

One liter of each 1.2 L triplicate was filtered through 0.45 µm polycarbonate filter paper using a vacuum pump within 12 h of collection. The resulting filter was stored in a −20 °C freezer prior to transport to Duke University for DNA extraction and analysis. DNA extraction was performed on all filtered samples using the MoBio PowerSoil® DNA Isolation Kit (Carlsbad, CA USA) according to the manufacturer's protocol. PowerSoil® was chosen instead of PowerWater® because prior studies show that the former can effectively prevent inhibition (Cox and Goodwin, 2013). In addition, studies have shown similar recovery with the two MoBio kits when using filter homogenized samples (Kaevska and Slana, 2015).

The extraction resulted in 100 µL DNA extract volumes for each of the triplicates. One 100 µL volume was used for HTS analysis of the 16S rRNA gene. The other two 100 µL elution volumes were transferred into two single-use aliquots of 20 µL for 16S rRNA gene sequencing and one reserve aliquot of 60 µL for future analysis to be determined.

2.5. High throughput sequencing

Illumina MiSeq amplicon sequencing was performed on all samples. The 16S rRNA gene sequencing primers 314F (5′-CCTACGGGAGGCAG CAG-3′) and 807R (5′GGACTACHAGGGTATCTAAT-3′) that cover the V3 and V4 rRNA regions were used to amplify the bacterial fraction of extracted DNA according to previous protocols (Prosdocimi et al., 2013). Samples were prepared using the Illumina workflow for 16S rRNA gene analysis (16S Metagenomic Sequencing Library Preparation,

Illumina Inc.). Samples were normalized, pooled, and run on a paired-end MiSeq platform using V3 sequencing technology.

Raw sequencing reads were processed in R using the DADA2 bioinformatics package according to a previously described pipeline (Callahan et al., 2016). Briefly, the Divisive Amplicon Denoising Algorithm (DADA) uses a model-based approach to correct amplicon errors without constructing OTUs. Rather than retaining individual reads, analysis using DADA2 generates a set of ASVs based upon the raw reads that can then be matched against a reference database. This allows for combination of multiple sequencing runs and reduces computational load when working with large datasets. Though the use of exact sequence variants results in slightly different values for diversity and richness metrics, values strongly correlate to OTU-based analysis. Comparison of samples using each of these methods is likely to yield similar results (Glassman and Martiny, 2018). In this study, we used SILVA v132 as the reference database for 16S rRNA gene sequences.

2.6. Physical-chemical parameters and sample metadata

Temperature, pH, and salinity were measured using a YSI Professional Plus handheld multi-parameter instrument when possible. Access limitations prevented the use of a YSI to gather physical-chemical parameters in China. Instrument calibration problems resulted in data for some samples to be removed from consideration in this study. Of samples collected in sites when a YSI was available, physical-chemical parameters were gathered for 39 of 55 samples and 25 of 34 ballast water samples.

2.7. Data analysis

All downstream data analysis of 16S rRNA gene data was performed using R in the phyloseq package (McMurdie and Holmes, 2013). Rarefaction was performed on all samples. In addition, Shannon's index and Simpson's index were calculated for all samples using the phyloseq and vegan packages (Oksanen et al., 2017). ASV data were transformed to relative abundance of each sample. Further data visualizations that were not available through the phyloseq or DADA2 packages were performed using the ggplot2 package within R (Wickham, 2009).

Concentration of ASVs in different samples was compared using the Wilcoxon rank-sum test in each sample type (ocean, ballast, harbor) and location (California, China, Singapore, South Africa). Comparisons were performed by combining samples into two groups. For example, testing if the concentration of California samples differed from all other samples was performed by creating one group of California samples and a second group composed of all other study sites before running the Wilcoxon rank-sum test. The concentration of *E. coli* and *Enterococcus* was compared by converting raw concentration to log scale and performing ANOVA and the Tukey test to examine differences across sample type and location.

Cluster analysis was used to analyze community similarities between samples. The analysis was performed with non-metric multidimensional scaling (NMDS) based on Bray-Curtis dissimilarities at the phylum, class, and order levels. Edges were assigned to samples with > 0.70 similarity. Environmental vectors and factors were fit to the ordination plots using the envfit function in the vegan R package (Oksanen et al., 2017). The environmental vectors included in this analysis were temperature, pH, salinity, and ballast residence time in the ballast tank. Samples without all four environmental variables were not included in the vector analysis. In addition, PCoA was performed and visualized using unweighted UniFrac distance between samples.

2.8. Machine learning to identify potential markers

Though this study is the largest high throughput sequencing study of ballast water to date (41 ballast samples), the total sample size ($n = 67$) is still relatively small to generate a machine learning model. As a result, machine learning was performed with 67-fold cross-validation, withholding one sample per cross-validation. Limitations of this approach are discussed in the discussion section.

Linear discriminant analysis (LDA) effect size (LEfSe) analysis was performed to select potential markers for classifying samples by sample type (ballast, harbor, ocean) (Segata et al., 2011). A random forest model using 5000 trees was generated using all potential markers and performance metrics were recorded (Breiman et al., 2018). The variable with the smallest contribution to classification accuracy was removed, and the model was rerun with performance metrics recorded. Iterative reduction of the variables input to the model was performed to determine the minimum number of variables required to achieve a lower out-of-bag (OOB) error than the original model with all 19 variables. The robustness of the selected variables from the random forest method was assessed by using three other machine learning methods: 1) Penalized linear regression (Goeman et al., 2018); 2) multinomial logistic regression (Ripley and Venables, 2016); and 3) k nearest neighbors (Ripley and Venables, 2015).

3. Results

3.1. 16S rRNA gene analysis

3.1.1. Reads, rarefaction, sample coverage, and diversity indices

Sequencing resulted in a total of approximately 22 million raw Illumina reads across the 67 samples. Approximately 19 million reads remained after filtering, approximately 17 million reads were tabled, and approximately 15 million reads remained after removing chimeric sequences. These reads corresponded to 52,838 unique ASVs. Rarefaction curves were calculated and approached a plateau in all samples (Supplementary data Fig. S1). The range of unique ASVs per sample were: ballast 440 (CA13) to 2993 (S17); harbor 569 (CA21) to 4339 (CN3); and ocean 136 (S5) to 700 (S8). Shannon's index ranges were: ballast 2.61 (S19) to 6.42 (S17); harbor 4.52 (CA39) to 6.61 (CN3); and ocean 3.13 (S2) to 5.02 (S8). Simpson's index ranges were: ballast 0.744 (S19) to 0.995 (CA35); harbor 0.941 (SA7) to 0.994 (CN10); and ocean 0.856 (S2) to 0.982 (S8). The diversity values for all samples are shown in the Supplementary data (Table S2).

3.1.2. Relative abundance of major amplicon sequence variants

At the phylum level, bacterial communities in ballast tanks were dominated by Proteobacteria (11.2%, CA35 to 71.7%, S19), Bacteroidetes (0.8%, SA6 to 51.9%, CA13), and Actinobacteria (0.2%, SA6 to 20.8%, CN9). In harbor water samples, the major ASVs were Proteobacteria (12.2%, CN5 to 59.5%, S12), Bacteroidetes (3.0%, CN5 to 35.8%, CA27), and Cyanobacteria (2.5%, CN10 to 27.1%, S16). In ocean water samples, the major ASVs were Proteobacteria (19.9%, S2 to 73.5%, S5), Cyanobacteria (4.8%, S8 to 58.5%, S2), and Bacteroidetes (4.6%, S5 to 12.7%, S8). The relative abundance of Proteobacteria in ballast water was significantly higher than in other sample types ($p = 0.045$). In addition, the relative abundance of Cyanobacteria in ballast water was significantly lower than in other sample types ($p < 0.0001$). All phyla with $> 5\%$ relative abundance were visualized to examine for trends (Fig. 1A).

At the class level, bacterial communities in ballast water were dominated by Alphaproteobacteria (9.0%, CA13 to 51.4%, CA31), Gammaproteobacteria (7.7%, CN1 to 70.6%, S19), and Bacteroidia (6.7%, SA4 to 58.7%, CA13). In harbor samples, the bacterial community was predominantly composed of Alphaproteobacteria (12.8%, CN5 to 34.2%, SA7), Gammaproteobacteria (11.8%, S16 to 54.4%, S12), and Bacteroidia (5.1%, CN5 to 38.0%, CA27). In ocean samples, the major classes were Gammaproteobacteria (7.0%, S2 to 65.6%, S5), Oxyphotobacteria (5.5%, S8 to 59.9%, S2), and Alphaproteobacteria (10.9%, S3 to 24.6%, S7). The relative abundance of certain classes in the ballast water microbial community had significant differences much like the relative abundance of certain phyla. For example,

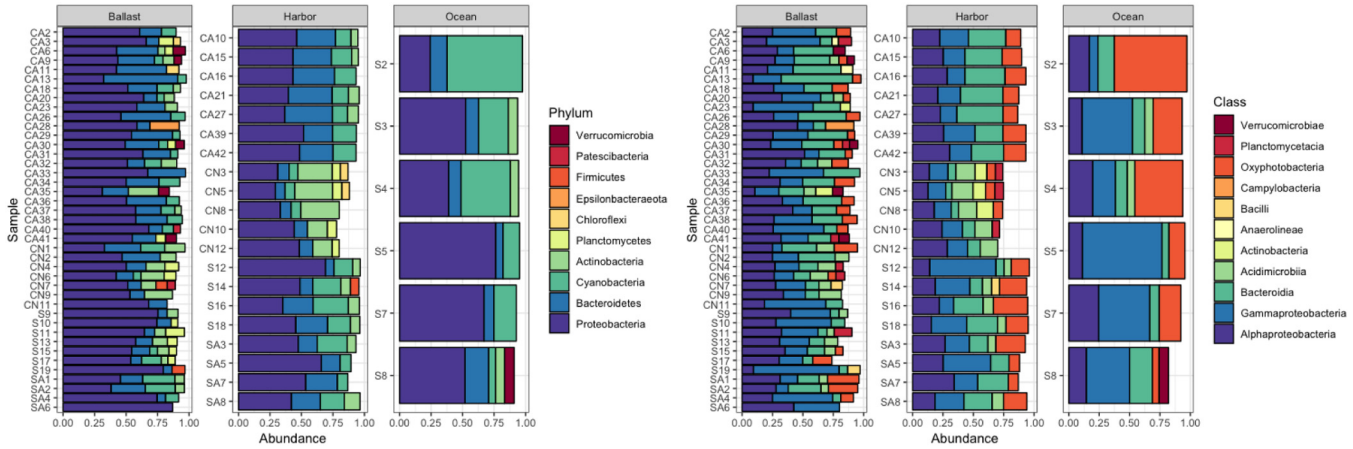

Fig. 1. (A, left) Relative abundance of ASVs in all samples with relative abundance > 0.05 at the phylum level. (B, right) Relative abundance of ASVs in all samples with relative abundance > 0.05 at the class level.

Oxyphotobacteria composed a significantly higher proportion of the population in ocean and harbor water than ballast water ($p = 0.012$). The composition of the ballast water bacterial community when analyzed at the class level looked similar across all locations; however, the composition of classes in the harbor water bacterial community in China appeared to differ from California, Singapore, and South Africa. All classes with > 5% relative abundance were visualized to assess the existence of trends (Fig. 1B).

3.1.3. Multivariate analysis

Cluster analysis of the relative abundance of ASVs in each sample at the phylum level revealed a trend by location when samples were faceted by sample type. This segmentation was especially evident in harbor water samples (Fig. 2A). In addition, ballast and harbor water samples collected in the same location tended to cluster together in the un-faceted analysis. A pattern of clustering despite the lack of faceting was most easily observed in the California samples (Fig. 2B). Network projections using Bray-Curtis dissimilarities revealed edges by location (Fig. 3). Harbor water samples from different sites were distinct from one another, because harbor samples from two different sites never had a Bray-Curtis dissimilarity value < 0.70. A non-metric multidimensional scaling (NMDS) plot identified possible clustering by location. Temperature may explain some variation ($r^2 = 0.48$, $p = 0.004$); however, it was likely confounded by location. Explanatory environmental variables such as pH ($r^2 = 0.13$, $p = 0.29$), salinity ($r^2 = 0.05$, $p = 0.66$), and residence time of ballast water in ballast tanks ($r^2 = 0.14$, $p = 0.27$) did not have significant correlations to sample clustering (Supplementary data Fig. S2). PCoA analysis depicts California samples as marginally separate from all other sample locations, regardless of sample type; however, there was significant overlap among other sample locations (Fig. 4). There was some differentiation of Chinese samples, including a cluster of three Chinese harbor water samples separate from all other samples; however, Chinese ballast water samples retained apparent overlap with many other sample types and locations.

3.2. Indicator organisms

Culture-based analysis of indicator organisms identified five vessels that exceeded IMO Regulation D-2 for *E. coli* (one), *Enterococcus* (three), or both (one) (Supplementary data Table S3). The number of vessels arriving at each port that exceeded proposed regulations were: Shanghai, China (two of seven); Singapore (three of seven); Los Angeles/Long Beach, California (zero of 24); and Durban, South Africa (zero of four). The IMO Regulation D-2 Ballast Water Performance Standard includes standards for acceptable concentrations of indicator microbes. The regulation includes language requiring *E. coli* concentration < 250 colony forming units per 100 mL and intestinal *Enterococci* concentration < 100 colony forming units per 100 mL. All of the ballast samples in this study that exceeded IMO standards were loaded in the Chinese EEZ or Chinese ports. Some harbor water samples also exceeded IMO Regulation D-2 for *E. coli* (one), *Enterococcus* (three),

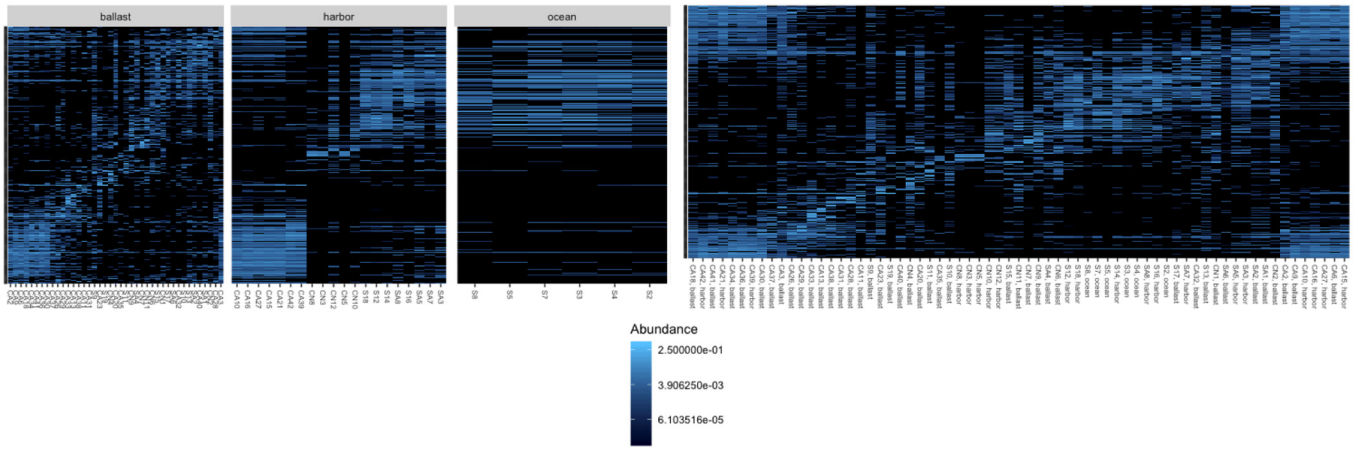

Fig. 2. (A, left) Phylum-level clustered heatmap of relative abundance of ASVs among samples faceted by sample type. (B, right) Phylum-level clustered heatmap of relative abundance of ASVs.

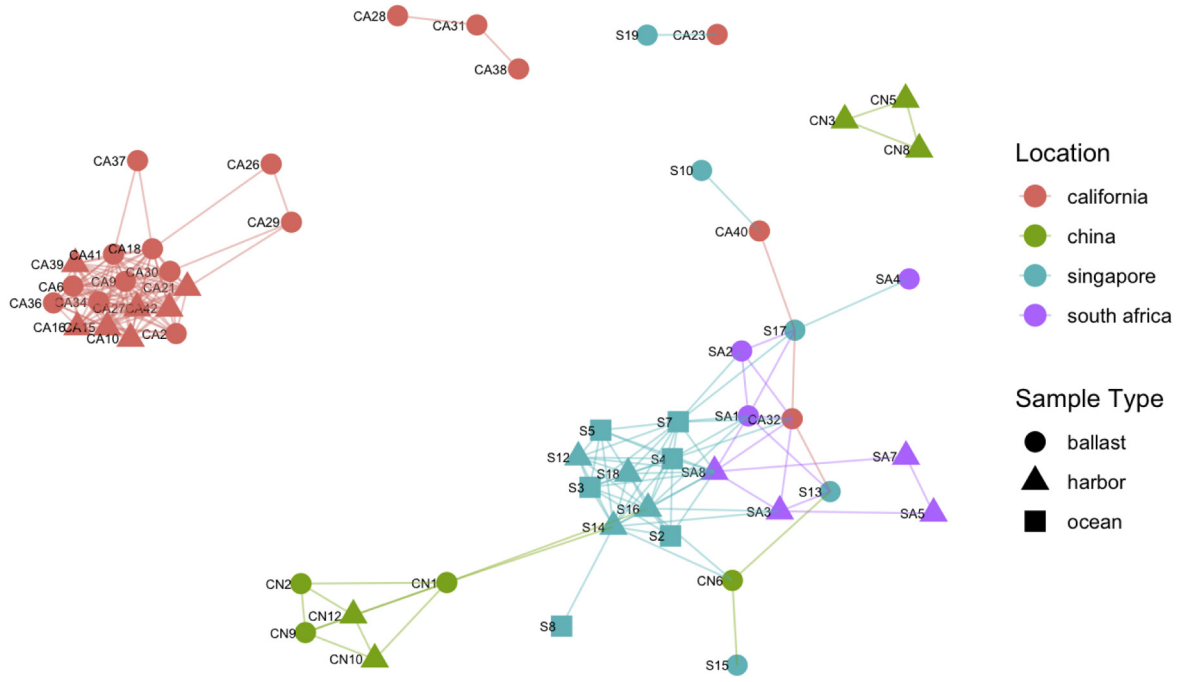

Fig. 3. Network projection of Bray-Curtis dissimilarity values < 0.70 between linked samples calculated on relative abundance of ASVs.

or both (three). The number of harbor samples exceeding the regulation segmented by location were: Shanghai, China (five of five); Singapore (one of four); Los Angeles/Long Beach, California (zero of seven); and Durban, South Africa (one of four).

ANOVA tests were performed to examine for differences in variance among the sample locations and types. In ballast water, there were significant differences between locations in the log transformed concentration of total coliforms ($p < 0.001$), *Enterococcus* ($p = 0.001$), and *E. coli* ($p = 0.003$). Tukey's post hoc test identified that lower concentrations of the tested indicator organisms were typically observed in vessels arriving to the United States or South Africa when compared to vessels arriving in China or Singapore; however, this finding was not always statistically significant. In harbor water, there were significant differences between locations in the log-transformed concentration of total coliforms ($p = 0.002$), *Enterococcus* ($p = 0.001$), and *E. coli* ($p = 0.001$). Tukey's post hoc test indicated that there was a significantly higher concentration of all tested indicator bacteria in Chinese harbor water when compared to United States or Singapore harbor water ($p < 0.05$); however, there was not a significant difference between Chinese and South African harbor water.

3.3. Machine learning and classification biomarker identification

LEfSe identified 19 potential markers to indicate sample type (ballast, harbor, ocean) with linear discriminant analysis (LDA) values > 10^4 (Supplementary data Fig. S3). After iterative removal and assessment of out-of-bag (OOB) error, the best random forest model used eight variables and had an OOB error of 11.94% (Fig. 5). Based on the iterative reduction, these eight variables (taxonomic level) in decreasing order of importance are: 1) HIMB11 (genus); 2) Actinobacteria (class); 3) Cyanobium_PCC-6307 (genus); 4) Sulfurimonas (genus); 5) Marivivens (genus); 6) Thiomicrospirales (order); 7) Gilvibacter

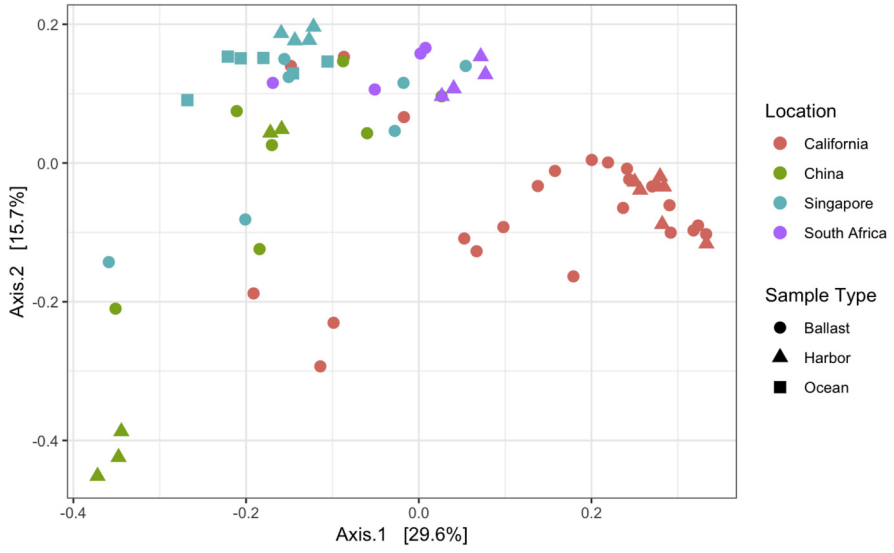

Fig. 4. PCoA of all samples calculated using unweighted UniFrac on relative abundance of ASVs.

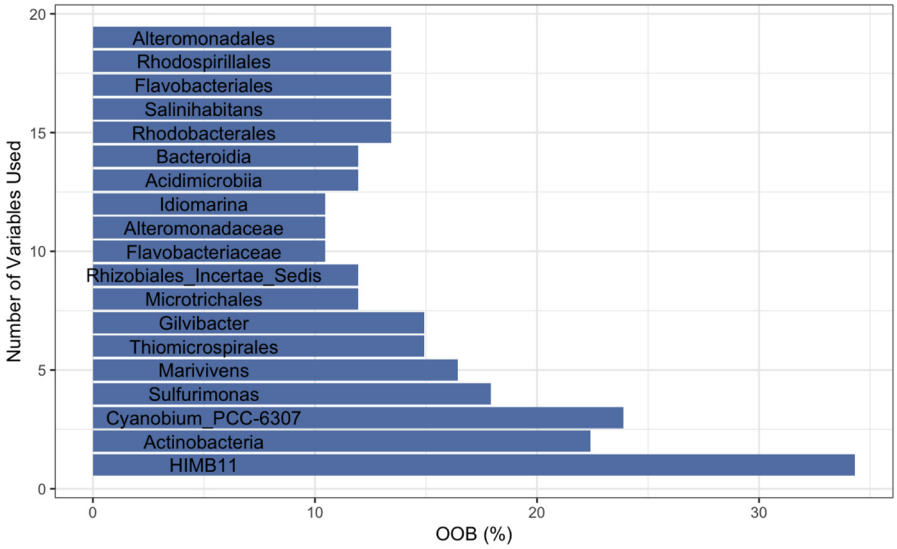

Fig. 5. Random forest out-of-bag error when using different markers. Iterative removal of variables was performed to select the best model. The variable removed at each step is annotated on the figure.

(genus); and 8) Microtrichales (order).

Additional machine learning models were generated using the eight variables identified by the random forest to assess the robustness of these indicators to additional machine learning methods. The k nearest neighbors (kNN) approach required fewer than six neighbors to be used so that ocean samples (n = 6) could be correctly identified. The best performance was achieved when one neighbor was used (accuracy = 0.806), and the worst performance occurred when five neighbors were used (accuracy = 0.657). The penalized linear regression model had a maximum accuracy when the L2 value was 0.01 (accuracy = 0.776); however, this was not significantly better than the performance of the unpenalized linear regression model (accuracy = 0.731). Rapid loss of accuracy was observed when penalties > 0.01 were applied to either L1 or L2 (Supplementary data Fig. S4). In addition, the multinomial regression model performed well without penalties. The model is improved through iteration, so the lowest accuracy was observed on the initial iteration (accuracy = 0.239) and the maximum accuracy was observed with 250 iterations (accuracy = 0.866). A slight decline in accuracy was recorded after 250 iterations as accuracy dropped to 0.851; however, the accuracy stabilized at this value despite additional iterations.

4. Discussion

4.1. 16S rRNA gene analysis

4.1.1. Relative abundance of major amplicon sequence variants

The present study shares several findings with previously published research. Similar to the existing literature, our study suggests that Alphaproteobacteria, Gammaproteobacteria, Bacteroidetes, and unclassified Bacteria dominate the bacterial assemblages of the environmental DNA in ballast water (Lymperopoulou and Dobbs, 2017). In California, South Africa, and Singapore, ballast water samples had a wider range of alpha diversity values than harbor water samples, but the median was approximately the same between sample types (Table S1). This was not the case in China, where harbor water samples consistently had higher alpha diversity (Table S1). This finding is similar to previous research, including findings that the ballast water alpha diversity of a tank loaded in Chinese ports or the Chinese exclusive economic zone (EEZ) had a higher alpha diversity score (Ng et al., 2015).

This study identified higher relative abundance of Cyanobacteria in harbor water and ballast water with low residence time than was previously reported (Lymperopoulou and Dobbs, 2017). This finding may be associated with many variables, including sample time or sample method. In addition, a much higher number of unique ASVs were found in this research than OTUs found in all previous studies using HTS of 16S rRNA genes (Brinkmeyer, 2016; Lymperopoulou and Dobbs, 2017; Ng et al., 2015). This could be explained by the larger sample size and wider range of sample types as previously discussed. Furthermore, the pipeline that generates ASVs will not cluster similar sequences until later in the analysis, thereby resulting in more richness reported in the initial stages of the pipeline.

4.1.2. Multivariate analysis

Clustering of harbor and ballast samples by sampling site is a surprising result, because ballast water collected in a study site often originates in another locale. Location-specific clustering may be related to the route travelled by the vessels. Vessels arriving to South Africa, China, and Singapore were typically arriving after short journeys through the South China Sea or the Indian Ocean; whereas, vessels arriving to California typically crossed through the North Pacific with ballast water exchange or treatment occurring while underway. In addition, major ocean currents near Singapore, China, and South Africa flow from the equator towards the poles, while the major ocean current along the California coast flows from the Arctic to equator. This difference may result in different physical, chemical, and microbial characteristics of water along the route and in the harbor. Ocean microbial community richness metrics have been previously observed to change with latitude (Fuhrman et al., 2008), so different locations of ballast water exchange may explain some of the variation in microbial community dynamics of ballast water arriving at different sites. Similarly, the microbial community of harbor water likely depends on several characteristics such as turbidity, temperature, and salinity. Harbor water samples from different sites never had a Bray-Curtis dissimilarity value < 0.70, so the harbors included in this study often had quite different bacterial communities when compared to one another. The impact of bacterial community dynamics at the site of ballast uptake is not well understood and should be further examined in future research.

4.1.3. Amplicon sequence variants vs. OTUs

ASVs may be better suited than OTUs for analysis of less characterized matrices, such as ocean or ballast water, because ASVs are a DNA sequence that is determined independently of a reference

database, which carries some intrinsic biological meaning (Callahan et al., 2017). While we intended to characterize the amplicons as both ASVs and OTUs to allow comparison of results, initial efforts were abandoned because of the higher computational demand of OTU-based analysis which led to unreasonably long lengths of analysis times. Our experience was similar to that of others who reported that ASV based analysis reduces the computational demands of a data set when compared to OTUs (Callahan et al., 2017). Overall, we identified a higher number of ASVs than the number of OTUs that were previously reported (Lymperopoulou and Dobbs, 2017). There are several possible explanations for why our numbers may be higher. First, the present study examined the V3–V4 region of the 16S rRNA gene. Though the primers and region were selected based on the Illumina standard protocol for amplicon sequencing, the lack of a large overlap region in the forward and reverse reads likely increased the number of ASVs reported. Second, a higher number of ASVs would still be expected in this study compared to previous research, because this study included a much larger set of samples (61 vs. a maximum of 19 in previous 16S rRNA gene studies) that were collected in the open ocean and ballast water from a wider geographic area (Lymperopoulou and Dobbs, 2017). The larger number of samples as well as the diversity of samples may have contributed to a greater number of observed microorganisms. Third, ASV analysis is more likely than closed-reference OTU analysis to identify high variation, because biological variation not present in the reference database will be lost during OTU assignment (Callahan et al., 2017).

4.2. Indicator organisms

The finding of ballast tanks exceeding IMO standards has been previously reported (Altug et al., 2012). The presence of pathogens has been reported in up to 48% of tanks in previous research (Burkholder et al., 2007). A slightly lower rate of tanks with possible pathogens was found herein (16 of 41 or 39%). However, this number may have been higher if we had included as many pathogens as those tested in the previous study (Burkholder et al., 2007).

The presence of indicator organisms in ballast water is not surprising given their global spread. When present, the observed concentrations were often not problematic according to the standards set forth by the IMO. It is important to note that at least one ballast water sample with concentrations of either *E. coli* or *Enterococcus* that exceeded the standards was collected in both Singapore and China. Further, China, Singapore, and South Africa all had at least one harbor water sample with concentrations of *E. coli* or *Enterococcus* exceeding the standards. Receiving ballast water with concentrations of indicator organisms greater than the defined thresholds may pose a risk to human and environmental health; however, the direct risks posed by ballast water with high concentrations of indicator organisms are difficult to elucidate when the receiving harbor also intermittently has values above these thresholds.

In addition, harbors with concentrations above the recommended standards require ballast water treatment systems to actively reduce concentrations prior to discharge. This poses an additional challenge when compared to ballast loaded in zones with indicator organism concentrations below the IMO Regulation D-2 standards. In this study, all ballast water samples above the thresholds were loaded in the Chinese EEZ. This finding highlights the fact that certain ports may be hot spots for microbial activity and may serve as hubs for microbial translocation. The reasons for this observation may be regulatory (e.g. lack of industry oversight, differential waste management) or geographic (e.g. high turbidity, warmer climate). Further research should be performed to understand the role of individual ports in the global proliferation of microorganisms through ballast.

4.3. Machine learning and classification biomarker identification

The lowest OOB error in a random forest machine learning model was achieved by using eight of the 19 variables identified by LEfSe as possible markers for classification. The kNN and penalized regression machine learning models using the eight markers identified from the random forest also performed well, which suggests that the selected markers are robust to machine learning method and may be useful biomarkers for water type differentiation between ballast, harbor, and ocean water. There was not sufficient sample size for a holdout test set, which may bias the accuracy and OOB error values to indicate a better model than reality. However, the performance of these machine learning models could be further improved with additional tuning. Further research applying machine learning to larger sample sizes would be useful to generate more accurate real-world performance metrics. In addition, future research should be performed to further assess the reliable capability of markers identified herein for accurately classifying ballast, harbor, and ocean water.

5. Conclusion

Ballast water had significantly different relative abundances of some bacterial taxa compared to harbor and ocean water. This difference may allow for the use of biomarkers to rapidly assess water origin from ballast tanks, harbors, and ports. Further refinement of this approach may allow for classification of ballast water as originating from a harbor or a port. The approach described herein may serve as a useful proof-of-concept for a machine learning and biomarkers-based approach to classification. As with other machine learning based studies, additional work to increase the sample size will likely lead to a reduction in the number of necessary biomarkers needed to achieve the desired accuracy thresholds for classification and improve the real-world applicability of this model.

Acknowledgements

This work was supported by the National Science Foundation (NSF) under grants no. DGE 1545220, no. OISE 1243433, no. OISE 1713717, and no. OISE 1614161. Further support was provided through a Graduate Student Training Enhancement Grant from Duke University Interdisciplinary Studies, and a Graduate Research and Training Award from the Duke University Center for International and Global Studies.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2018.12.038.

References

Adar, S.D., Huffnagle, G.B., Curtis, J.L., 2016. The respiratory microbiome: an underappreciated player in the human response to inhaled pollutants? Ann. Epidemiol. 26, 355–359. https://doi.org/10.1016/j.annepidem.2016.03.010.

Aguirre-Macedo, M.L., Vidal-Martinez, V.M., Herrera-Silveira, J.A., Valdés-Lozano, D.S., Herrera-Rodríguez, M., Olvera-Novoa, M.A., 2008. Ballast water as a vector of coral pathogens in the Gulf of Mexico: the case of the Cayo Arcas coral reef. Mar. Pollut. Bull. 56, 1570–1577. https://doi.org/10.1016/j.marpolbul.2008.05.022.

Altug, G., Gurun, S., Cardak, M., Ciftci, P.S., Kalkan, S., 2012. The occurrence of pathogenic bacteria in some ships' ballast water incoming from various marine regions to the Sea of Marmara, Turkey. Mar. Environ. Res. 81, 35–42. https://doi.org/10.1016/j.marenvres.2012.08.005.

Bax, N., Williamson, A., Aguero, M., Gonzalez, E., Geeves, W., 2003. Marine invasive alien species: a threat to global biodiversity. Mar. Policy 27, 313–323. https://doi.org/10.1016/S0308-597X(03)00041-1.

Brady, S.F., 2007. Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules. Nat. Protoc. 2, 1297–1305. https://doi.org/10.1038/nprot.2007.195.

Breiman, L., Cutler, A., Liaw, A., Wiener, M., 2018. randomForest: Breiman and Cutler's Random Forests for Classification and Regression.

Brinkmeyer, R., 2016. Diversity of bacteria in ships ballast water as revealed by next

generation DNA sequencing. Mar. Pollut. Bull. 107, 277–285. https://doi.org/10.1016/j.marpolbul.2016.03.058.
Burkholder, J.A.M., Hallegraeff, G.M., Melia, G., Cohen, A., Bowers, H.A., Oldach, D.W., Parrow, M.W., Sullivan, M.J., Zimba, P.V., Allen, E.H., Kinder, C.A., Mallin, M.A., 2007. Phytoplankton and bacterial assemblages in ballast water of U.S. military ships as a function of port of origin, voyage time, and ocean exchange practices. Harmful Algae 6, 486–518. https://doi.org/10.1016/j.hal.2006.11.006.
Callahan, B.J., Mcmurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J., Holmes, S.P., 2016. HHS Public Access 13. pp. 581–583. https://doi.org/10.1038/nmeth.3869.DADA2.
Callahan, B.J., McMurdie, P.J., Holmes, S.P., 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 11, 2639–2643. https://doi.org/10.1038/ismej.2017.119.
Carlton, J.T., 2001. The scale and ecological consequences of biological invasions in the World's oceans. In: Invasive Species and Biodiversity Management, pp. 195.
Cox, A.M., Goodwin, K.D., 2013. Sample preparation methods for quantitative detection of DNA by molecular assays and marine biosensors. Mar. Pollut. Bull. 73, 47–56. https://doi.org/10.1016/j.marpolbul.2013.06.006.
Czaplicki, L.M., Gunsch, C.K., 2016. Reflection on molecular approaches influencing state-of-the-art bioremediation design: culturing to microbial community fingerprinting to omics. J. Environ. Eng. 142. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001141.
Darling, J., Martinson, J., Gong, Y., Okum, S., Carney, K.J., Ruiz, G., 2018. Ballast water exchange and invasion risk posed by intra-coastal vessel traffic: an evaluation using high throughput sequencing. Environ. Sci. Technol. https://doi.org/10.1021/acs.est.8b02108.
Drake, L.A., Meyer, A.E., Forsberg, R.L., Baier, R.E., Doblin, M.A., Heinemann, S., Johnson, W.P., Koch, M., Rublee, P.A., Dobbs, F.C., 2005. Potential invasion of microorganisms and pathogens via “interior hull fouling”: biofilms inside ballast water tanks. Biol. Invasions 7, 969–982. https://doi.org/10.1007/s10530-004-3001-8.
Drake, L.A., Doblin, M.A., Dobbs, F.C., 2007. Potential microbial bioinvasions via ships' ballast water, sediment, and biofilm. Mar. Pollut. Bull. 55, 333–341. https://doi.org/10.1016/j.marpolbul.2006.11.007.
Drake, L.A., Tamburri, M.N., First, M.R., Smith, G.J., Johengen, T.H., 2014. How many organisms are in ballast water discharge? A framework for validating and selecting compliance monitoring tools. Mar. Pollut. Bull. 86, 122–128. https://doi.org/10.1016/j.marpolbul.2014.07.034.
Egan, S.P., Grey, E., Olds, B., Feder, J.L., Ruggiero, S.T., Tanner, C.E., Lodge, D.M., 2015. Rapid molecular detection of invasive species in ballast and harbor water by integrating environmental DNA and light transmission spectroscopy. Environ. Sci. Technol. 49, 4113–4121. https://doi.org/10.1021/es5058659.
Emami, K., Askari, V., Ullrich, M., Mohinudeen, K., Anil, A.C., Khandeparker, L., Burgess, J.G., Mesbahi, E., 2012. Characterization of bacteria in ballast water using MALDI-TOF mass spectrometry. PLoS One 7. https://doi.org/10.1371/journal.pone.0038515.
Fuhrman, J.A., Steele, J.A., Hewson, I., Schwalbach, M.S., Brown, M.V., Green, J.L., Brown, J.H., 2008. A latitudinal diversity gradient in planktonic marine bacteria. Proc. Natl. Acad. Sci. 105, 7774–7778. https://doi.org/10.1073/pnas.0803070105.
Fykse, E.M., Nilsen, T., Nielsen, A.D., Tryland, I., Delacroix, S., Blatny, J.M., 2012. Real-time PCR and NASBA for rapid and sensitive detection of *Vibrio cholerae* in ballast water. Mar. Pollut. Bull. 64, 200–206. https://doi.org/10.1016/j.marpolbul.2011.12.007.
Gerhard, W.A., Gunsch, C.K., 2018. Analyzing trends in ballasting behavior of vessels arriving to the United States from 2004 to 2017. Mar. Pollut. Bull. 135, 525–533. https://doi.org/10.1016/j.marpolbul.2018.07.001.
Gerhard, W.A., Choi, W.S., Houck, K.M., Stewart, J.R., 2017. Water quality at points-of-use in the Galapagos Islands. Int. J. Hyg. Environ. Health 220, 485–493. https://doi.org/10.1016/j.ijheh.2017.01.010.
Glassman, S.I., Martiny, J.B.H., 2018. Broadscale ecological patterns are robust to use of exact sequence variants versus operational taxonomic units. MSphere 3, 1–5. https://doi.org/10.1128/mSphere.00148-18.
Goeman, J., Meijer, R., Chaturvedi, N., Lueder, M., 2018. Penalized: L1 (Lasso and Fused Lasso) and L2 (Ridge) Penalized Estimation in GLMs and in the Cox Model.
Gwin, C.A., Lefevre, E., Alito, C.L., Gunsch, C.K., 2018. Microbial community response to silver nanoparticles and Ag+ in nitrifying activated sludge revealed by ion semiconductor sequencing. Sci. Total Environ. 616–617, 1014–1021. https://doi.org/10.1016/j.scitotenv.2017.10.217.
Kaevska, M., Slana, I., 2015. Comparison of filtering methods, filter processing and DNA extraction kits for detection of mycobacteria in water. Ann. Agric. Environ. Med. 22, 429–432. https://doi.org/10.5604/12321966.1167707.
Kim, Y., Aw, T.G., Teal, T.K., Rose, J.B., 2015. Metagenomic investigation of viral communities in ballast water. Environ. Sci. Technol. 49, 8396–8407. https://doi.org/10.1021/acs.est.5b01633.
Lefèvre, E., Bossa, N., Gardner, C.M., Gehrke, G.E., Cooper, E.M., Stapleton, H.M., Hsu-Kim, H., Gunsch, C.K., 2018. Biochar and activated carbon act as promising amendments for promoting the microbial debromination of tetrabromobisphenol A. Water Res. 128, 102–110. https://doi.org/10.1016/j.watres.2017.09.047.
Li, F., Peng, Y., Fang, W., Altermatt, F., Xie, Y., Yang, J., Zhang, X., 2018. Application of environmental DNA metabarcoding for predicting anthropogenic pollution in rivers. Environ. Sci. Technol. https://doi.org/10.1021/acs.est.8b03869.
Lymperopoulou, D.S., Dobbs, F.C., 2017. Bacterial diversity in ships' ballast water, ballast-water exchange, and implications for ship-mediated dispersal of microorganisms. Environ. Sci. Technol. 51, 1962–1972. https://doi.org/10.1021/acs.est.6b03108.
McCarthy, S.A., Khambaty, F.M., 1994. International dissemination of epidemic *Vibrio cholerae* by cargo ship ballast and other nonpotable waters. Appl. Environ. Microbiol. 60, 2597–2601.
McMurdie, P.J., Holmes, S., 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8, e61217.
Microbial Contaminants Method 9223, 2005. Standard Methods for the Examination of Water and Wastewater. Standard Methods.
Ng, C., Le, T.-H., Goh, S.G., Liang, L., Kim, Y., Rose, J.B., Yew-Hoong, K.G., 2015. A comparison of microbial water quality and diversity for ballast and tropical harbor waters. PLoS One 10, e0143123. https://doi.org/10.1371/journal.pone.0143123.
Ng, C., Goh, S.G., Saeidi, N., Gerhard, W.A., Gunsch, C.K., Gin, K.Y.H., 2018. Occurrence of *Vibrio* species, beta-lactam resistant *Vibrio* species, and indicator bacteria in ballast and port waters of a tropical harbor. Sci. Total Environ. 610–611, 651–656. https://doi.org/10.1016/j.scitotenv.2017.08.099.
Oksanen, A., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., Mcglinn, D., Minchin, P.R., Hara, R.B.O., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., 2017. vegan: Community Ecology Package.
Prosdocimi, E.M., Novati, S., Bruno, R., Bandi, C., Mulatto, P., Giannico, R., Casiraghi, M., Ferri, E., 2013. Errors in ribosomal sequence datasets generated using PCR-coupled “panbacterial” pyrosequencing, and the establishment of an improved approach. Mol. Cell. Probes 27, 65–67. https://doi.org/10.1016/j.mcp.2012.07.003.
Ripley, B., Venables, W., 2015. class: Functions for Classification.
Ripley, B., Venables, W., 2016. nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models.
Ruiz, G.M., Rawlings, T.K., Dobbs, F.C., Drake, L.A., Mullady, T., Huq, A., Colwell, R.R., 2000. Global spread of microorganisms by ships. Nature 408, 49–50.
Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., Huttenhower, C., 2011. Metagenomic biomarker discovery and explanation. Genome Biol. 12. https://doi.org/10.1186/gb-2011-12-6-r60.
Shokralla, S., Spall, J.L., Gibson, J.F., Hajibabaei, M., 2012. Next-generation sequencing technologies for environmental DNA research. Mol. Ecol. 21, 1794–1805. https://doi.org/10.1111/j.1365-294X.2012.05538.x.
Staley, C., Unno, T., Gould, T.J., Jarvis, B., Phillips, J., Cotner, J.B., Sadowsky, M.J., 2013. Application of Illumina next-generation sequencing to characterize the bacterial community of the Upper Mississippi River. J. Appl. Microbiol. 115, 1147–1158. https://doi.org/10.1111/jam.12323.
Stoeck, T., Frühe, L., Forster, D., Cordier, T., Martins, C.I.M., Pawlowski, J., 2018. Environmental DNA metabarcoding of benthic bacterial communities indicates the benthic footprint of salmon aquaculture. Mar. Pollut. Bull. 127, 139–149. https://doi.org/10.1016/j.marpolbul.2017.11.065.
Tan, B., Ng, C., Nshimyimana, J.P., Loh, L.L., Gin, K.Y.-H., Thompson, J.R., 2015. Next-generation sequencing (NGS) for assessment of microbial water quality: current progress, challenges, and future opportunities. Front. Microbiol. 6. https://doi.org/10.3389/fmicb.2015.01027.
Thomsen, P.F., Willerslev, E., 2015. Environmental DNA – an emerging tool in conservation for monitoring past and present biodiversity. Biol. Conserv. 183, 4–18. https://doi.org/10.1016/j.biocon.2014.11.019.
Wang, C., Wang, J., 2011. Spatial pattern of the global shipping network and its hub-and-spoke system. Res. Transp. Econ. 32, 54–63. https://doi.org/10.1016/j.retrec.2011.06.010.
WHO, 2014. Antimicrobial resistance: global report on surveillance. Bull. World Health Organ. https://doi.org/10.1007/s13312-014-0374-3.
Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis.