A peer-reviewed version of this preprint was published in PeerJ on 24 March 2020.

McGhee JJ, Rawson N, Bailey BA, Fernandez-Guerra A, Sisk-Hackworth L, Kelley ST. 2020. Meta-SourceTracker: application of Bayesian source tracking to shotgun metagenomics. PeerJ 8:e8783 https://doi.org/10.7717/peerj.8783

Meta-SourceTracker: Application of Bayesian source tracking to shotgun metagenomics

Jordan J McGhee [1], **Nick Rawson** [2], **Barbara A Bailey** [2], **Antonio Fernandez-Guerra** [3], **Scott T Kelley** Corresp. [4]

[1] Bioinformatics and Medical Informatics Program, San Diego State University, San Diego, California, United States

[2] Department of Mathematics and Statistics, San Diego State University, San Diego, California, United States

[3] Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany

[4] Department of Biology, San Diego State University, San Diego, California, United States

Corresponding Author: Scott T Kelley
Email address: skelley@sdsu.edu

Background. Microbial source tracking methods are used to determine the origin of contaminating bacteria and other microorganisms, particularly in contaminated water systems. The Bayesian SourceTracker approach uses deep-sequencing marker gene libraries (16S ribosomal RNA) to determine the proportional contributions of bacteria from many potential source environments to a given sink environment simultaneously. Since its development, SourceTracker has been applied to an extensive diversity of studies, from beach contamination to studying human behavior.

Methods. Here, we developed metagenomic-SourceTracker (mSourceTracker), an expanded SourceTracker approach for shotgun metagenomic datasets. We tested mSourceTracker using sink samples from coastal marine environment metagenomes and source environment metagenomes collected from freshwater, marine, soil, sand and gut environments. We also implemented features for determining the stability of source proportion estimates using new techniques that split metagenomic data for domain-specific analyses (i.e., Bacteria, Archaea, Eukarya and viruses). The added features allow users to visualize the precision of mSourceTracker and assess ways to optimize performance.

Results. Our results found mSourceTracker to be highly effective at predicting the composition of known sources using shotgun metagenomic libraries. In addition, we showed that different taxonomic domains sometimes presented highly divergent pictures of source origins. These findings indicated that applying mSourceTracker to separate domains may provide a deeper understanding of the microbial origins of complex, mixed-source environments, and further suggested that certain domains may be preferable for tracking specific sources of contamination.

META-SOURCETRACKER: APPLICATION OF BAYESIAN SOURCE TRACKING TO SHOTGUN METAGENOMICS

Jordan J. McGhee[1], Nick Rawson[2], Barbara A. Bailey[2], Antonio Fernandez-Guerra[3], Scott T. Kelley[4]

[1] Bioinformatics and Medical Informatics Program, San Diego State University, San Diego, CA, USA
[2] Department of Mathematics and Statistics, San Diego State University, San Diego, CA, USA
[3] Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany
[4] Department of Biology, San Diego State University, San Diego, CA, USA

Corresponding Author:
Scott T. Kelley[3]
5500 Campanile Drive, San Diego State University, San Diego, CA, 92104-4614 USA
Email address: skelley@sdsu.edu

Abstract

Background. Microbial source tracking methods are used to determine the origin of contaminating bacteria and other microorganisms, particularly in contaminated water systems. The Bayesian SourceTracker approach uses deep-sequencing marker gene libraries (16S ribosomal RNA) to determine the proportional contributions of bacteria from many potential source environments to a given sink environment simultaneously. Since its development, SourceTracker has been applied to an extensive diversity of studies, from beach contamination to studying human behavior.

Methods. Here, we developed metagenomic-SourceTracker (mSourceTracker), an expanded SourceTracker approach for shotgun metagenomic datasets. We tested mSourceTracker using sink samples from coastal marine environment metagenomes and source environment metagenomes collected from freshwater, marine, soil, sand and gut environments. We also implemented features for determining the stability of source proportion estimates using new techniques that split metagenomic data for domain-specific analyses (i.e., Bacteria, Archaea, Eukarya and viruses). The added features allow users to visualize the precision of mSourceTracker and assess ways to optimize performance.

Results. Our results found mSourceTracker to be highly effective at predicting the composition of known sources using shotgun metagenomic libraries. In addition, we showed that different taxonomic domains sometimes presented highly divergent pictures of source origins. These findings indicated that applying mSourceTracker to separate domains may provide a deeper understanding of the microbial origins of complex, mixed-source environments, and further suggested that certain domains may be preferable for tracking specific sources of contamination.

Introduction

Microbes are found in every environment, from the depths of the Pacific Ocean to the hostile conditions of the Atacama Desert. Most microbes co-exist with other microbes in biofilms or in complex dynamic communities referred to as 'microbiomes' (e.g., the gut microbiome) that include hundreds or thousands of different microbial species, many of which play critical roles in animal health and ecosystem function. While much is known about the species composition of microbial communities, less is understood about how they form in the first place and how microbes move among different ecosystems. Understanding the origins of microbial communities is particular important for tracking routes of contamination, such as polluted water systems, but also has important implications for understanding microbiome development and ecosystem function.

Microbial source tracking (MST) approaches have been developed to determine the source origins of particular microbes, with their primary use being the study of bacterial contamination of municipal water (Liu et al., 2018) and freshwater (streams, rivers and lakes) (Newton et al., 2013; Staley et al., 2018) and coastal ocean waters. Standard MST approaches track microbial sources via one or more key bacterial strains or species previously linked to a specific source (e.g., *E. coli* strains only found in cow feces). Traditional MST methods rely on techniques such as culture isolation and PCR with species-specific primers. Other MST approaches have relied on patterns of multiple antibiotic-resistance and carbon utilization profiles (Joyce M. Simpson, Jorge W. Santo Domingo & Reasoner, 2002; Scott et al., 2002). More recently, improvements in next-generation sequencing (NGS) technologies has resulted in NGS being widely in all aspects of microbiology including MST (van Dijk et al., 2014; Martin et al., 2018).

The widely-used SourceTracker program has provided one of the most powerful and effective methods for using NGS data to perform MST. This program uses a combination of Bayes' theorem and Gibb's sampling to analyze data from large bacterial 16S rRNA marker-gene NGS libraries. Unlike previous MST methods which use individual microbes to identify routes of colonization and contamination, SourceTracker uses data from hundreds or thousands of species, and allows simultaneous estimation of the proportion of multiple source environments contributing to a given sink environment, including an estimate of unknown sources (Knights et al., 2011). For example, in a study of bacterial assemblages on restroom surfaces, the researchers used SourceTracker to estimate the relative proportion of skin, feces and soil contributing to each specific sink sample (Flores et al., 2011). At the time of this writing, SourceTracker had been cited over 300 times with a surprising diversity of applications, including identifying individual organisms within the same species based on their microbiomes, determining which body sites contribute most to contamination of built environments and detecting sources of early gut colonization among others (Flores et al., 2011; Hewitt et al., 2013; Hyde et al., 2016; Chen et al., 2018; Kapono et al., 2018). Other applications included applying SourceTracker to forensic analysis and to study human behavior (Lax et al., 2015; Bik et al., 2016). SourceTracker was

designed for use with bacterial 16S rRNA marker genes and has primarily been used with these data. However, it has been applied to a few shotgun metagenomic studies, including one that tracked the source origins of antibiotic resistance gene markers (Baral et al., 2018). While more expensive and computationally intensive, shotgun metagenomic data allows for a much broader potential array of microbial diversity (bacteria, archaea, eukaryotes and viruses) to be used in microbial source tracking.

In this study, we developed and tested a metagenomic-SourceTracker (mSourceTracker) approach to determine the source origins of microbial samples. The goal was to test the utility of mSourceTracker for shotgun metagenomic datasets with samples of known origins and provide tools for determining the reliability of proportion estimates. Specifically, we used samples collected from coastal marine environments, which are commonly a mix of different sources due to runoff from freshwater environments and contamination from land debris. In addition, we applied mSourceTracker to entire metagenomes but also to each organismal group separately. Our results showed that mSourceTracker provides a robust approach for microbial source tracking with metagenomic datasets and also demonstrated how mSourceTracker could provide deeper taxon-specific biological insights into the movements of microbes among ecosystems.

Materials & Methods

Data Collection

Metagenome sequence libraries were obtained from samples collected from coastal marine water, fresh water, human gut (feces), sand and soil environments from multiple studies. These environments were chosen as likely sources of microorganisms to be found in coastal marine waters which tend to have runoff from rivers and possibly contaminated with sewage. A total of 223 samples were used for this study; 110 coastal marine samples, 30 freshwater samples, 64 soil samples, 6 sand samples, and 13 gut samples (see Supplemental Table 1 for details).

Taxonomic separation of metagenomic data

Taxonomic abundances were generated for all samples using the k-mer approach implemented in the Kaiju ver. 1.5.0 program (Menzel, Ng & Krogh, 2016). Kaiju produces estimated taxonomic abundances primarily at the genus level. To determine the domain of each genus (Archaea, Bacteria, Eukarya or virus), we wrote a programming script in python3.6 using the URL:

http://taxonomy.jgi-psf.org/tax/sc_name/{}

to extract the full taxonomic lineage information given the genus name.

For example, given the genus Salmonella, the URL:

http://taxonomy.jgi-psf.org/tax/sc_name/Salmonella

returns the string:

sk:Bacteria;p:Proteobacteria;c:Gammaproteobacteria;o:Enterobacterales;f:Enterobacteriaceae;g:Salmonella

The "genus" names within the Kaiju output that did not return a lineage from the URL were manually searched in NCBI. The domain information was then added to a dictionary within our script and later compiled into a single data frame. For domain specific source tracking analysis, the taxonomic abundances for each sample were separated by domain and written into corresponding data frames. Species and count numbers from each sample were merged using species name with all previously processed samples within each domain-specific data frame. Once all samples were processed through our pipeline, the assembled data frames for each domain were then written to an output table formatted to be used with mSourceTracker.

Simulation data for mSourceTracker analysis

To test the prediction accuracy of mSourceTracker, 14 samples from the coastal marine environment were defined as sinks and all remaining samples from every environment defined as sources. The analysis was performed on the combined dataset and each domain separately, with a default rarefaction limit of 1000. Samples which did not have a minimum count of 1000 for any given kingdom file were removed from all datasets. Proportions for sink samples were compared using 10 and 100 draws. The number of chains was set at 5 for all comparisons. The same mapping file was used for comparing differences in proportions between kingdom datasets. The number of chains was held at 5 and the number of draws were variable so as to keep the chain differences below 5%.

mSourceTracker: diagnostic add-on feature

Gibbs sampling data computed from the SourceTracker 'envcounts' array was written to a temporary output file along with source and sink ID's. When the diagnostic function was called using the command '--diagnostics', the Gibbs data file was read and placed into an array using numpy (van der Walt, Colbert & Varoquaux, 2011). The array was split based on the number of chains and number of draws defined by user inputs. Here, we use the standard Markov Chain Monte Carlo terminology of "draw" and "chain". However, it should be noted that the original SourceTracker codebase uses the term "restart" to refer to an MCMC draw, and "draw" for an MCMC chain. Array data was multiplied by the alpha1 preset to convert numbers into respective proportion values. Each chain produces a moving average via Gibbs sampling over the number of draws selected. The script then calculates the difference between the maximum and minimum chains. If the proportion value of any two chains differs by a default value of 5%, or by user defined parameters, all chains are exported onto a single line graph per sample for each

environment. Each line represents a single chain and the legend displays the proportion estimate of each chain for the given sample and environment. A single text output table displays the absolute differences between the maximum and minimum chains for all samples in each environment.

Random Forest classifier

Sample names in the feature tables used for mSourceTracker analysis were converted to name of the source for which they were collected. One source was then selected to be tested with all other sources being categorized as 'other' to identify features important in classifying the selected environment. Training and testing sets were randomly created at approximately a 3:1 ratio respectively. Random Forest is an ensemble learning method which classifies by the votes of its component trees. Using scikit-learn Random Forest classifier (Pedregosa et al., 2011) we fit and classified the data using 500 trees. Random state was set at 0 and the out-of-bag score was made True. The classifier was run multiple times to ensure there were no important features returned due to overfitting or other errors. The 10 most important features were graphed for each environment based on relative importance utilizing pandas (McKinney, 2010) and matplotlib (Hunter, 2007). Confusion matrices and statistics for the Random Forest classifier were also produced using scikit-learn modules. This process was performed for each of the organismal domains.

Results

Effects of parameter adjustments on the accuracy and precision of mSourceTracker

In order to apply the Bayesian approach of mSourceTracker to metagenomics data, we downloaded a total of 223 samples from 5 clearly identified environments (Table A1). These sources include 110 coastal marine samples, 30 freshwater samples, 64 soil samples, 6 sand samples, and 13 gut samples. Fourteen of the coastal marine samples were chosen to be used as "sink" samples. The k-mer based Kaiju analysis identified a total of 5725 taxa across all sample from the four major taxonomic domains of life. Of these, 88.8% of them were bacteria sequences, while eukaryotes and archaea comprised ~9.0% and ~1.9% of our metagenomic sequences, respectively. Virus-matching sequences comprised just ~0.3% of all the samples.

Since previous studies indicated that adjusting SourceTracker's default parameters (e.g., number of restarts) with 16S data led to more stable estimates of source proportions (Henry et al., 2016), we determined how the proportional composition for our sink samples would be affected if we adjusted the default parameters for metagenomic samples. Fig. 1A shows an example of how the number of draws can affect proportion estimates. Because the estimates are a moving average, increasing the number of draws to 100 resulted in a decreased variability between each chain. Longer chains converge to a more stable estimate over time. We also increased the number of chains to 5 so we could compare multiple independent proportional estimates in a single run. As indicated in Fig. 1B, with only 10 draws, source proportion estimates among the different chains could vary considerably but increasing the number of draws to 100 resulted in convergence of the chains. Analysis of 14 source proportion estimate from 223 samples using 5 chains with 10 draws per chain, found an average of 3±2% difference between the two most different estimates. However, there were instances in which proportion estimates of the most different draws even after 10 draws, in some samples, were as much as 15-20% different (data not shown). Increasing the number of restarts to 100 dramatically minimized the differences among chains.

Figure 1C and D show the dramatic improvement and how increasing the number of draws and chains reduces the variability in the source proportion estimates for metagenomic samples. As mentioned previously, each draw is dependent on prior draws and a single chain runs the risk of getting caught in a local-maxima in the target distribution and returning an inaccurate estimation. More draws reduce the likelihood this phenomenon could affect the final estimations because draws are averaged together for each environment. For all subsequent testing we adjusted the default parameters in mSourceTracker to minimize the range between chains such that the biggest difference between chains would be less than 5%.

mSourceTracker analysis of separated domains

Once we established the best general parameters for mSouceTracker, we then compared the results of combined mSourceTacker analysis to single-domain analyses of the same samples. Figure 2 shows the results for a single coastal marine sink sample in which mSourceTracker had been run on the combined species and each specific domain. The results of the bacteria alone most closely resemble the proportions we get from the combined metagenomic data. This is likely because the Gibbs sampling approach used to estimate the proportions would tend to pick bacteria taxa, since the bulk of the sequences from the metagenomic data (88.4%) were bacterial. In this particular sample "coastal marine" comprised the largest source proportion in both the combined and bacteria fractions. Archaea and eukaryotes also indicated a high level of coastal marine, but the proportions were much lower. In contrast, the overwhelming majority of our viral sequences were determined to be from a freshwater environment, while the archaea indicated a high proportion of sand and gut sources and the eukaryotes were almost evenly split in this particular sample between the coastal marine, freshwater, sand and soil origins.

Figure 3 shows the estimated proportions for all 14 of the coastal marine sink samples broken down by domain. In these 14 samples, bacteria averaged 50.2% from the coastal marine environment. Eukaryote samples averaged 42.8% coastal marine with the remaining composition being evenly distribute among the other 4 environments. Archaea samples were approximately split between the coastal marine (38.6%) and sand (30.4%) environments. Most virus samples were dominantly estimated to be from freshwater despite being coastal marine samples. Virus samples average composition was 77.6% freshwater and just 0.075% coastal marine.

Random Forest of environments by domain

We used Random Forest to determine, which organisms were best at classifying samples into each environment seen in figure 4. Confusion matrices for Random Forest performed on each environmental condition are seen in figure 5. Accuracy scores remained above 95% for all classifications and the out-of-bag error was below 5% for most samples (Table 1). Several of these organisms which were determined to be an important feature were found in literature to be closely associated with that environment. This further demonstrates how mSourceTracker can be advantageous as certain taxa may be better suited for determining particular sources of contamination.

Discussion

Our results demonstrated not only the effectiveness of mSourceTracker with metagenomic datasets but also that the taxonomic diversity of metagenomic samples can potentially lend deeper insight into the mixed-source origins of complex environmental samples. The mSourceTracker analysis of the complete 14 test sink metagenomic libraries consistently revealed the biggest sources to be coastal marine, though the proportions varied considerably from sample to sample (Fig. 2; data not shown). Domain-specific mSourceTracker analysis, on the other hand, often revealed patterns remarkably distinct from the combined taxa set (Fig. 2, 3). The bacterial source origins typically mirrored the full libraries, likely because the bacteria were the most abundant in all the samples. However, the other domains could be unique. For instance, mSourceTracker analysis of just the identified viruses mainly identified freshwater as the primary contributor to the sink diversity; according to the virus data, freshwater contributed as much as 94% of the diversity in some samples (Fig. 3). Archaea-specific analysis typically identified both coastal marine and sand as more or less equal contributors, while the eukarotic suggested more even distribution among coastal marine, freshwater, sand, soil and even gut (22% in one sample).

The fact that domain-specific mSourceTracker analysis resulted in different source proportion estimates has two important ramifications. First, it shows that mSourceTracker can be used to identify the environmental sources of a particular group of organisms. For instance, one may conclude that, for a given sample, 75% of the viruses present originated from freshwater, while half of the bacteria were marine in origin, and 28% of the Eukarya came from soil runoff. Such results provide novel, sample-specific insight into the movement and origins of the organisms in that environment, which could be especially useful in understanding the complexity of contamination patterns or dispersal among biomes. One could also easily imagine splitting not only by domain, but also by specific phylogenetic groups (e.g., methanogens or the proteobacteria) or even multiple independent datasets (e.g., untargeted chemical or metabolic datasets). This is similar in principle to the approach taken by previous research to study the origins of antibiotic resistance markers (Gou et al., 2018; Baral et al., 2018; Li, Yin & Zhang, 2018). The identification of distinct origin sources for different taxonomic groups in the same "sink" samples is not without precedent in the literature. For example, a previous marker-gene study of restroom environments found that the fungi appeared to have radically different origins (plants and soils) than the bacteria from the same samples (human skin and gut) (Gibbons et al., 2015; Fouquier, Schwartz & Kelley, 2016). Other studies have shown very different patterns of diversity and abundance among different 'omics datasets, indicating this is a rule rather than the exception (Bikel et al., 2015; Guirro et al., 2018; Cocolin et al., 2018).

The second important ramification is that the diversity of environmental origins among the taxonomic domains indicates that particular taxonomic lineages may be better than others for tracking particular sources of contamination. For example, to study the input of freshwater into

the coastal marine environment the viruses may be superior to the bacteria, while eukaryotes may be better for tracking soil inputs. The Random Forest analysis identified a significant number of new taxa that were highly indicative of particular environments (Fig. 4). For every domain in every environment, we were able to identify certain features (taxa) that contributed significantly to the classification of the environment. In the future, such taxa could be used singly or in combination to detect particular types of contamination. This is the same principle used by culture-based source-tracking that tracks fecal contamination using strains of *E. coli* (Ravaliya et al., 2014). Recently, Stachler and Bibby (2014) proposed using sequences of crASSphage as a highly specific indicator of human fecal contamination (Stachler & Bibby, 2014).

One important caveat of mSourceTracker method is the general challenge of identifying taxa from metagenomic datasets. It is well known that much of the sequences from metagenomic datasets are not currently identifiable because databases are incomplete. Unlike 16S, it is not possible to put all the sequences from a library into a phylogenetic context, so many of them remain unknown and not currently useful in mSourceTracker analysis. As databases grow, this problem should diminish. The other issue is one of identification itself. There are many methods of identifying reads from metagenomic libraries, both alignment and k-mer based, and sometimes they can give very different results for the same samples (Quince et al., 2017). We expect that this may have a profound effect in some cases, and future research should look at the importance identification algorithms and databases on mSourceTracker results. Finally, in order for mSourceTracker to be broadly applicable, it will be critical to have many more metagenomes from "pure" environmental datasets. Large environmental collections such as the Earth Microbiome Project make it easy to find 16S ribosomal RNA libraries for any given environment, and it is relatively cheap to create many libraries in any given study and the analysis is easy to perform on a laptop. As the costs of sequencing continues to decline and the computational power and number of available data sets increases, the mSourceTracker approach will become increasingly tractable and commonplace.

Conclusions

In this study, we demonstrated three findings: (1) mSourceTracker is a straightforward and effective method for determining source proportions using shotgun metagenomic datasets; (2) Our chain convergence tests and visualizations allow researcher to identify when estimates do not converge, which mainly occurred when source datasets had poor taxonomic coverage; and (3) The purposeful domain-specific subdivision of metagenomic datasets has the potential to lend powerful new biological insights into the source and movement of microorganisms among environments. While our analyses demonstrated mSourceTracker's utility and potential, the results are only as good as the input data allow (the "garbage in, garbage out" rule). All inferences based on metagenomics data are dependent on the extent and quality of existing

databases and the effectiveness on taxonomic identification approaches. Methods other than Kaiju and more extensive databases could certainly produce different results and hopefully reduce the proportion of unknowns in the estimates. We also note that some of our source sample sets of metagenomes were small; increasing the sample size, purity and number of source datasets could also have a significant impact on interpretations. Investigation of all these parameters is beyond the scope of this study, which is focused on mSourceTracker development and proof of principle. However, such factors should be taken into account in future studies.

Acknowledgements

We thank E. Dinsdale for helpful insights and comments on the manuscript. We also thank P. Torres for guidance with the original SourceTracker and R. Edwards for allowing us an account on the anthill computer cluster at San Diego State University.

Figure Legends

Figure 1. Effects of increasing Markov chain length and number on estimating source proportions. (A) Comparison between ten and one-hundred draws for a single Markov chain. The same coastal marine sample was used to create both chains. (B) Convergence of five independent Markov chains for a single sample using either ten or one-hundred draws per chain. Chains represent the estimated proportions for a given "sink" from a single environment or "source". (C) Absolute percent differences between the two Markov chains with the highest and lowest average proportions over all draws for each environment or "source". The same coastal marine sample was used with five chains and either 10 or 100 draws. (D) Proportions per source for the same single sink sample after 10 and 100 draws respectively.

Figure 2. Taxon-dependent source proportion estimates in a single metagenome sample. Graphs represent the estimated proportions from each "source" or environment for a single coastal marine "sink". The middle pie chart "Meta" represents the estimated proportion contributed by 5 potential source environments and unknown based on the entire metagenome. The other pie charts depict the estimated proportions based on the bacterial, viral, archaea and eukaryal members of the sample (see Methods). The number of chains was held at 5 and the number of draws were variable so as to keep the chain differences below 5%.

Figure 3. Taxon-dependent source proportion estimates for 14 different coastal marine samples. Metagenome data was separated into taxa groups (see Methods) and multiple coastal marine samples were designated as "sinks". Heatmaps produced by SourceTracker represent the proportions from each of the source environments for sink samples. SourceTracker default number of chains was changed to 5, and number of draws were adjusted per taxa group so absolute values between any 2 Markov chains did not exceed 5%.

Figure 4. Random Forest analysis of each organismal group across 4 environments. Random Forest was used to determine which species were important in classifying samples as belonging to a certain environment. Each source was run against all other source classified as "other" and the data was randomly divided into testing and training subsets at approximately a 3:1 ratio. 500 estimators were used each time and the 10 most important features were graphed based on relative importance.

Figure 5. Confusion matrix for organismal groups across 4 environments. Heat map of confusion matrices for Random Forest analysis for each domain. Data was randomly split into training and testing set at approximately a 3:1 and run using 500 estimators. Graphs display predicted source (x-axis) vs. the true source for (y-axis) for each sample in the testing data set. The magnitude of the color represents the number of samples tested for that condition.

References

Baral D, Dvorak BI, Admiraal D, Jia S, Zhang C, Li X. 2018. Tracking the Sources of Antibiotic Resistance Genes in an Urban Stream during Wet Weather using Shotgun Metagenomic Analyses. *Environmental Science & Technology* 52:9033–9044. DOI: 10.1021/acs.est.8b01219.

Bik HM, Maritz JM, Luong A, Shin H, Dominguez-Bello MG, Carlton JM. 2016. Microbial Community Patterns Associated with Automated Teller Machine Keypads in New York City. *mSphere* 1:e00226-16. DOI: 10.1128/mSphere.00226-16.

Bikel S, Valdez-Lara A, Cornejo-Granados F, Rico K, Canizales-Quinteros S, Soberón X, Del Pozo-Yauner L, Ochoa-Leyva A. 2015. Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. *Computational and Structural Biotechnology Journal* 13:390–401. DOI: 10.1016/J.CSBJ.2015.06.001.

Chen X, Xu J, Ren E, Su Y, Zhu W. 2018. Co-occurrence of early gut colonization in neonatal piglets with microbiota in the maternal and surrounding delivery environments. *Anaerobe* 49:30–40. DOI: 10.1016/J.ANAEROBE.2017.12.002.

Cocolin L, Mataragas M, Bourdichon F, Doulgeraki A, Pilet M-F, Jagadeesan B, Rantsiou K, Phister T. 2018. Next generation microbiological risk assessment meta-omics: The next need for integration. *International Journal of Food Microbiology* 287:10–17. DOI: 10.1016/J.IJFOODMICRO.2017.11.008.

van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. 2014. Ten years of next-generation sequencing technology. *Trends in Genetics* 30:418–426. DOI: 10.1016/J.TIG.2014.07.001.

Flores GE, Bates ST, Knights D, Lauber CL, Stombaugh J, Knight R, Fierer N. 2011. Microbial Biogeography of Public Restroom Surfaces. *PLoS ONE* 6:e28132. DOI: 10.1371/journal.pone.0028132.

Fouquier J, Schwartz T, Kelley ST. 2016. Rapid assemblage of diverse environmental fungal communities on public restroom floors. *Indoor Air* 26:869–879. DOI: 10.1111/ina.12279.

Gibbons SM, Schwartz T, Fouquier J, Mitchell M, Sangwan N, Gilbert JA, Kelley ST. 2015. Ecological succession and viability of human-associated microbiota on restroom surfaces. *Applied and environmental microbiology* 81:765–73. DOI: 10.1128/AEM.03117-14.

Gou M, Hu H-W, Zhang Y-J, Wang J-T, Hayden H, Tang Y-Q, He J-Z. 2018. Aerobic composting reduces antibiotic resistance genes in cattle manure and the resistome dissemination in agricultural soils. *Science of The Total Environment* 612:1300–1310. DOI: 10.1016/J.SCITOTENV.2017.09.028.

Guirro M, Costa A, Gual-Grau A, Mayneris-Perxachs J, Torrell H, Herrero P, Canela N, Arola L. 2018. Multi-omics approach to elucidate the gut microbiota activity: Metaproteomics and metagenomics connection. *ELECTROPHORESIS* 39:1692–1701. DOI: 10.1002/elps.201700476.

Henry R, Schang C, Coutts S, Kolotelo P, Prosser T, Crosbie N, Grant T, Cottam D, O'Brien P, Deletic A, McCarthy D. 2016. Into the deep: Evaluation of SourceTracker for assessment of faecal contamination of coastal waters. *Water Research* 93:242–253. DOI: 10.1016/J.WATRES.2016.02.029.

Hewitt KM, Mannino FL, Gonzalez A, Chase JH, Caporaso JG, Knight R, Kelley ST. 2013.

Bacterial Diversity in Two Neonatal Intensive Care Units (NICUs). *PLoS ONE* 8:e54703. DOI: 10.1371/journal.pone.0054703.

Hunter JD. 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* 9:90–95. DOI: 10.1109/MCSE.2007.55.

Hyde ER, Navas-Molina JA, Song SJ, Kueneman JG, Ackermann G, Cardona C, Humphrey G, Boyer D, Weaver T, Mendelson JR, McKenzie VJ, Gilbert JA, Knight R. 2016. The Oral and Skin Microbiomes of Captive Komodo Dragons Are Significantly Shared with Their Habitat. *mSystems* 1:e00046-16. DOI: 10.1128/mSystems.00046-16.

Joyce M. Simpson, Jorge W. Santo Domingo * and, Reasoner DJ. 2002. Microbial Source Tracking: State of the Science. DOI: 10.1021/ES026000B.

Kapono CA, Morton JT, Bouslimani A, Melnik A V., Orlinsky K, Knaan TL, Garg N, Vázquez-Baeza Y, Protsyuk I, Janssen S, Zhu Q, Alexandrov T, Smarr L, Knight R, Dorrestein PC. 2018. Creating a 3D microbial and chemical snapshot of a human habitat. *Scientific Reports* 8:3669. DOI: 10.1038/s41598-018-21541-4.

Knights D, Kuczynski J, Charlson ES, Zaneveld J, Mozer MC, Collman RG, Bushman FD, Knight R, Kelley ST. 2011. Bayesian community-wide culture-independent microbial source tracking. *Nature Methods* 8:761–763. DOI: 10.1038/nmeth.1650.

Lax S, Hampton-Marcell JT, Gibbons SM, Colares GB, Smith D, Eisen JA, Gilbert JA. 2015. Forensic analysis of the microbiome of phones and shoes. *Microbiome* 3:21. DOI: 10.1186/s40168-015-0082-9.

Li L-G, Yin X, Zhang T. 2018. Tracking antibiotic resistance gene pollution from different sources using machine-learning classification. *Microbiome* 6:93. DOI: 10.1186/s40168-018-0480-x.

Liu G, Zhang Y, van der Mark E, Magic-Knezev A, Pinto A, van den Bogert B, Liu W, van der Meer W, Medema G. 2018. Assessing the origin of bacteria in tap water and distribution system in an unchlorinated drinking water system by SourceTracker using microbial community fingerprints. *Water Research* 138:86–96. DOI: 10.1016/J.WATRES.2018.03.043.

Martin TC, Visconti A, Spector TD, Falchi M. 2018. Conducting metagenomic studies in microbiology and clinical research. *Applied Microbiology and Biotechnology* 102:8629–8646. DOI: 10.1007/s00253-018-9209-9.

McKinney W. 2010. Data structures for statistical computing in python. In: *Proceedings of the 9th Python in Science Conference (SCIPY 2010)*. 51–56.

Menzel P, Ng KL, Krogh A. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications* 7:11257. DOI: 10.1038/ncomms11257.

Newton RJ, Bootsma MJ, Morrison HG, Sogin ML, McLellan SL. 2013. A Microbial Signature Approach to Identify Fecal Pollution in the Waters Off an Urbanized Coast of Lake Michigan. *Microbial Ecology* 65:1011–1023. DOI: 10.1007/s00248-013-0200-9.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. 2017. Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology* 35:833–844. DOI: 10.1038/nbt.3935.

Ravaliya K, Gentry-Shields J, Garcia S, Heredia N, Fabiszewski de Aceituno A, Bartz FE, Leon JS, Jaykus L-A. 2014. Use of Bacteroidales microbial source tracking to monitor fecal

contamination in fresh produce production. *Applied and environmental microbiology* 80:612–7. DOI: 10.1128/AEM.02891-13.

Scott TM, Rose JB, Jenkins TM, Farrah SR, Lukasik J. 2002. Microbial source tracking: current methodology and future directions. *Applied and environmental microbiology* 68:5796–803. DOI: 10.1128/aem.68.12.5796-5803.2002.

Stachler E, Bibby K. 2014. Metagenomic Evaluation of the Highly Abundant Human Gut Bacteriophage CrAssphage for Source Tracking of Human Fecal Pollution. *Environmental Science & Technology Letters* 1:405–409. DOI: 10.1021/ez500266s.

Staley C, Kaiser T, Lobos A, Ahmed W, Harwood VJ, Brown CM, Sadowsky MJ. 2018. Application of SourceTracker for Accurate Identification of Fecal Pollution in Recreational Freshwater: A Double-Blinded Study. *Environmental Science & Technology* 52:4207–4217. DOI: 10.1021/acs.est.7b05401.

van der Walt S, Colbert SC, Varoquaux G. 2011. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering* 13:22–30. DOI: 10.1109/MCSE.2011.37.

Figure 1

Effects of increasing Markov chain length and number on estimating source proportions.

(A) Comparison between ten and one-hundred draws for a single Markov chain. The same coastal marine sample was used to create both chains. (B) Convergence of five independent Markov chains for a single sample using either ten or one-hundred draws per chain. Chains represent the estimated proportions for a given “sink” from a single environment or “source”. (C) Absolute percent differences between the two Markov chains with the highest and lowest average proportions over all draws for each environment or “source”. The same coastal marine sample was used with five chains and either 10 or 100 draws. (D) Proportions per source for the same single sink sample after 10 and 100 draws respectively.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27869v1 | CC BY 4.0 Open Access | rec: 23 Jul 2019, publ: 23 Jul 2019

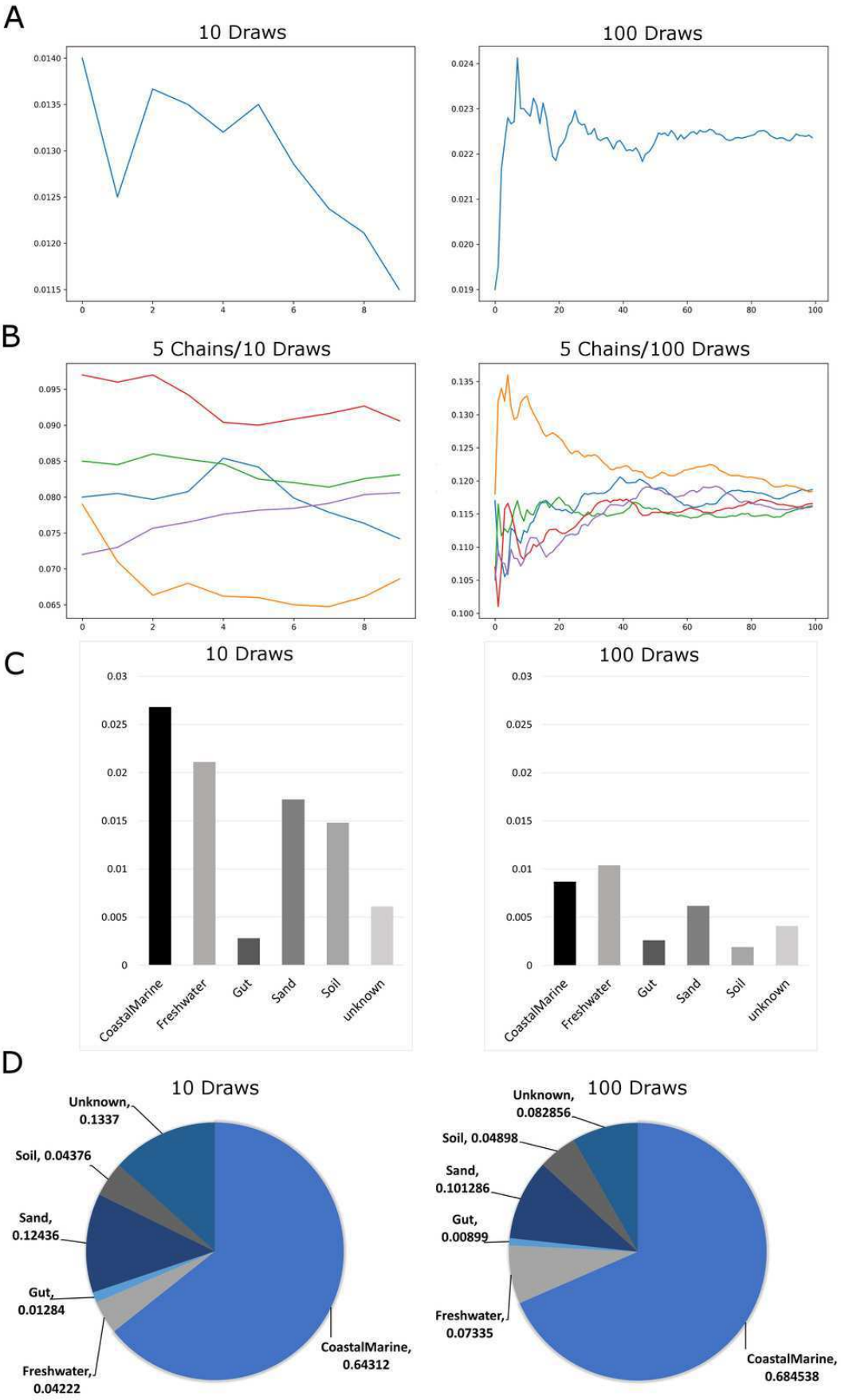


Figure 2

Taxon-dependent source proportion estimates in a single metagenome sample.

Graphs represent the estimated proportions from each “source” or environment for a single coastal marine “sink”. The middle pie chart “Meta” represents the estimated proportion contributed by 5 potential source environments and unknown based on the entire metagenome. The other pie charts depict the estimated proportions based on the bacterial, viral, archaea and eukaryal members of the sample (see Methods). The number of chains was held at 5 and the number of draws were variable so as to keep the chain differences below 5%.

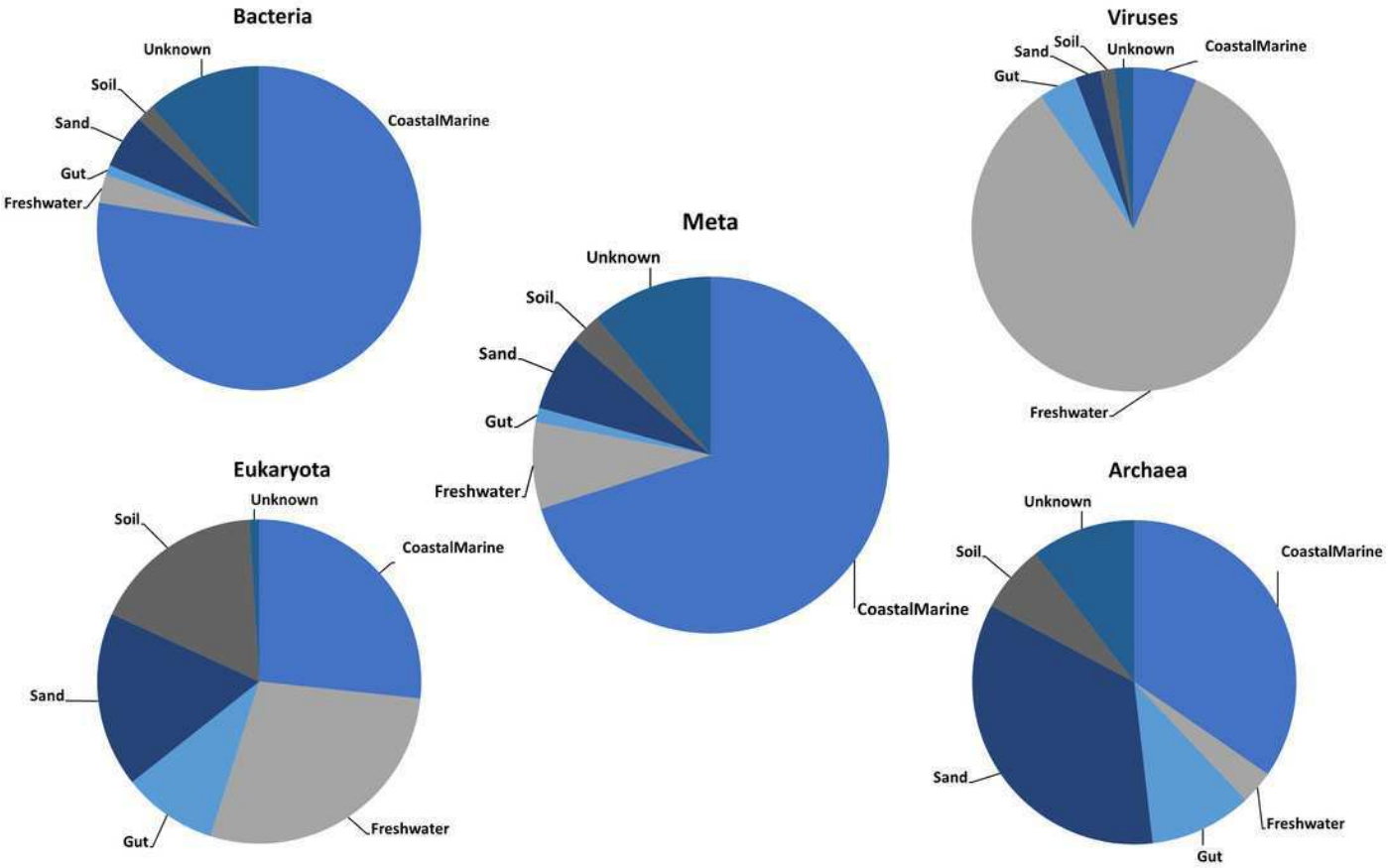


Figure 3

Taxon-dependent source proportion estimates for 14 different coastal marine samples.

Metagenome data was separated into taxa groups (see Methods) and multiple coastal marine samples were designated as “sinks”. Heatmaps produced by SourceTracker represent the proportions from each of the source environments for sink samples. SourceTracker default number of chains was changed to 5, and number of draws were adjusted per taxa group so absolute values between any 2 Markov chains did not exceed 5%.

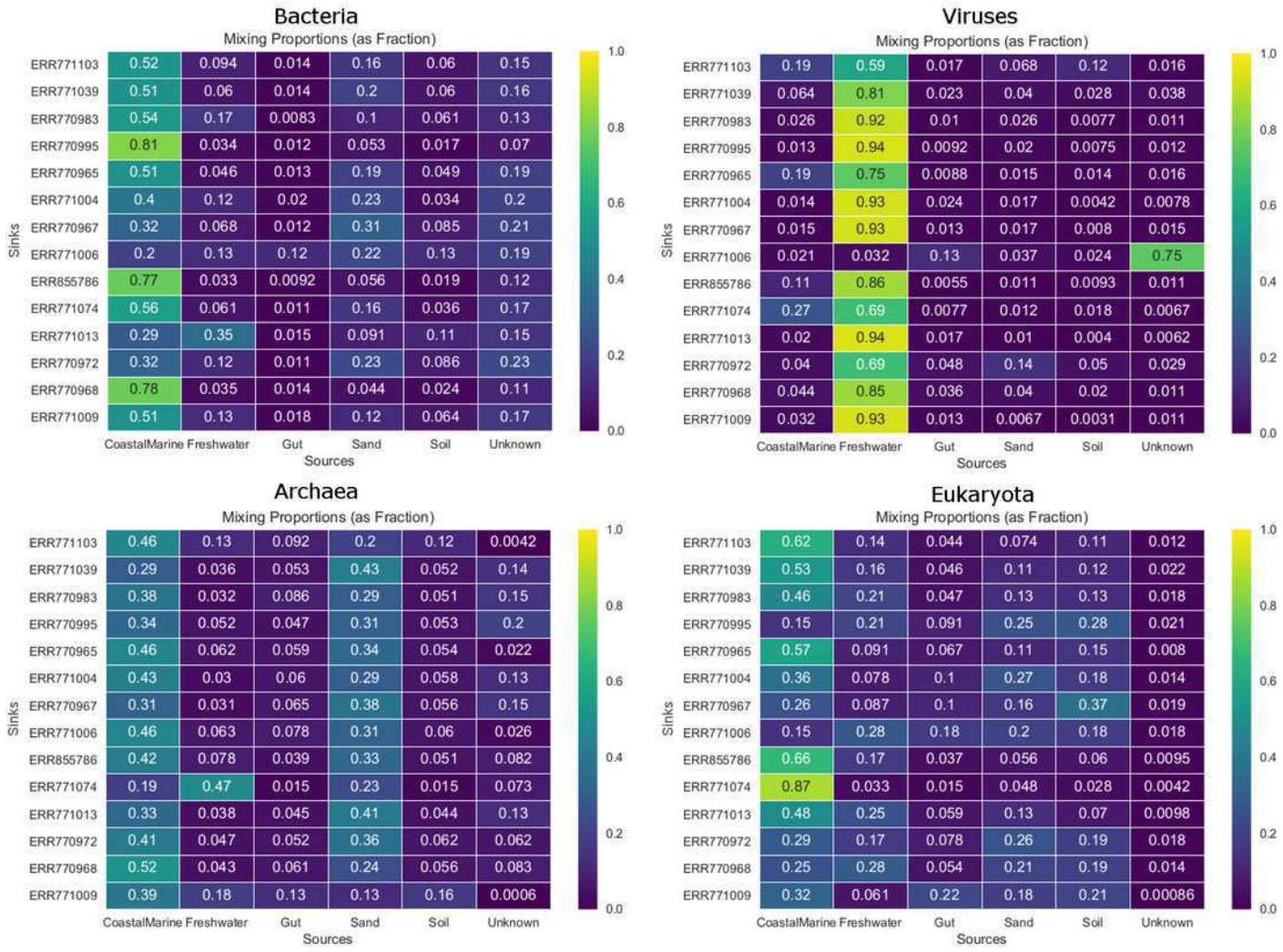


Figure 4

Random Forest analysis of each organismal group across 4 environments. Random Forest was used to determine which species were important in classifying samples as belonging to a certain environment.

Each source was run against all other source classified as “other” and the data was randomly divided into testing and training subsets at approximately a 3:1 ratio. 500 estimators were used each time and the 10 most important features were graphed based on relative importance.

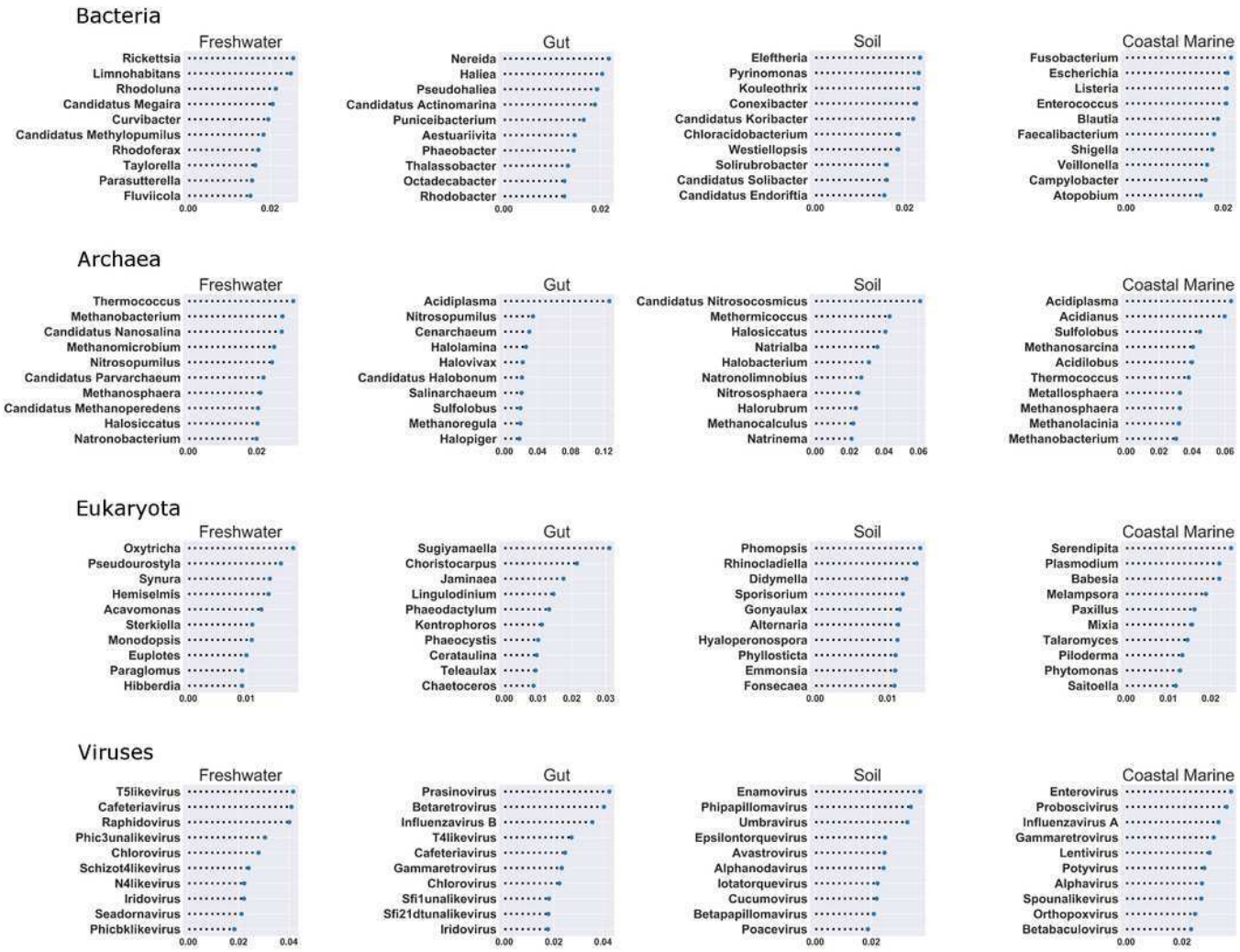
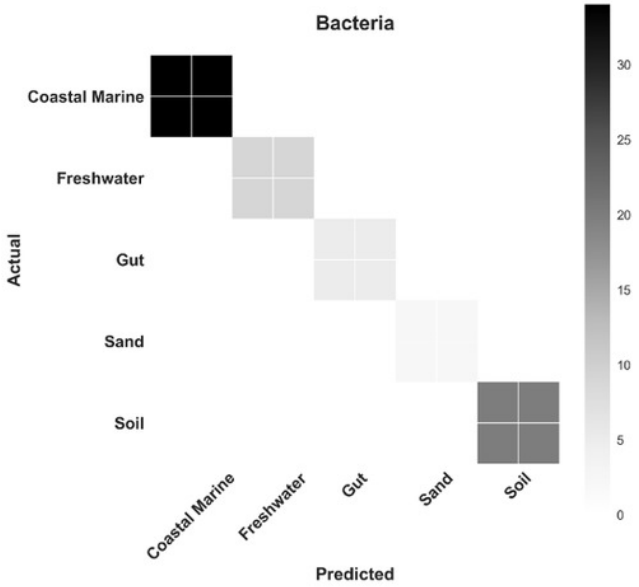


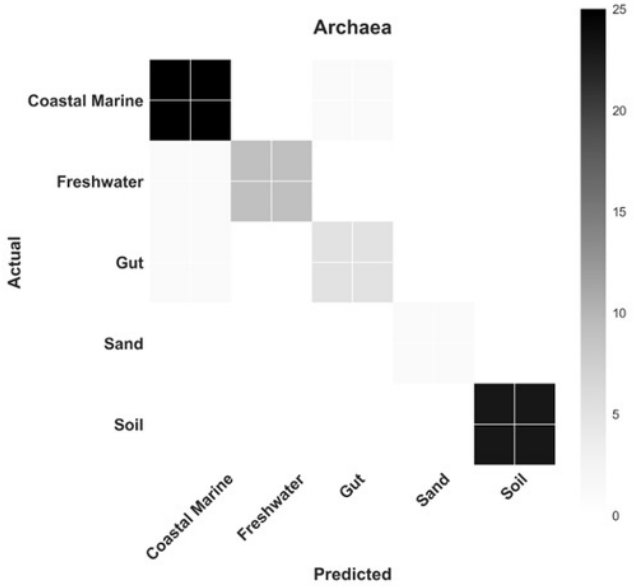
Figure 5

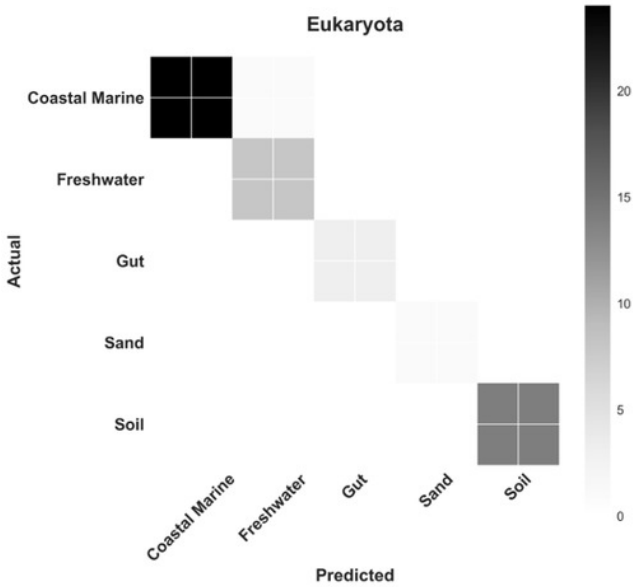
Confusion matrix for organismal groups across 4 environments. Heat map of confusion matrices for Random Forest analysis for each domain.

Data was randomly split into training and testing set at approximately a 3:1 and run using 500 estimators. Graphs display predicted source (x-axis) vs. the true source for (y-axis) for each sample in the testing data set. The magnitude of the color represents the number of samples tested for that condition.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27869v1 | CC BY 4.0 Open Access | rec: 23 Jul 2019, publ: 23 Jul 2019







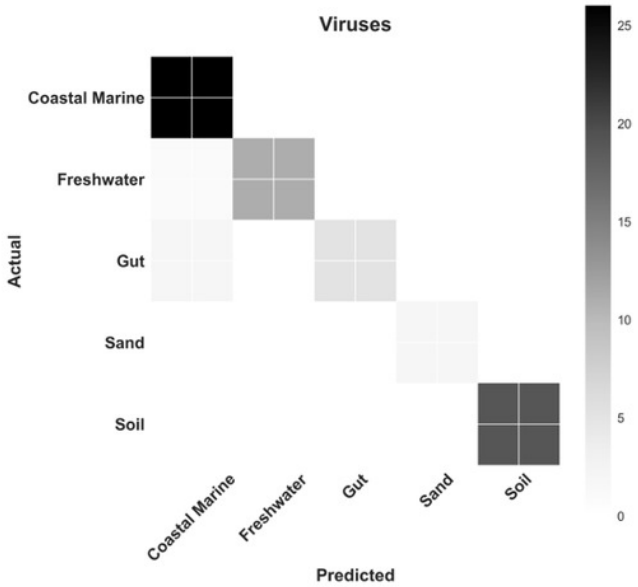


Table 1(on next page)

Random Forest results by taxonomic group and environment type.

The columns show the accuracy and out-of-bag error for predicting each environment type.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27869v1 | CC BY 4.0 Open Access | rec: 23 Jul 2019, publ: 23 Jul 2019

Table 1. Random Forest results by taxonomic group and environment type.

	Environment Type			
Domain	**Gut**	**Coastal Marine**	**Freshwater**	**Soil**
Bacteria				
Accuracy	0.982	0.984	0.97	0.985
OOB error	0.022	0.026	0.052	0.0174
Archaea				
Accuracy	0.966	0.977	0.966	0.98
OOB error	0.0497	0.062	0.049	0.021
Eukaryota				
Accuracy	0.955	0.967	0.959	0.984
OOB error	0.047	0.062	0.066	0.034
Viruses				
Accuracy	0.964	0.927	0.96	0.965
OOB error	0.054	0.061	0.058	0.027

1