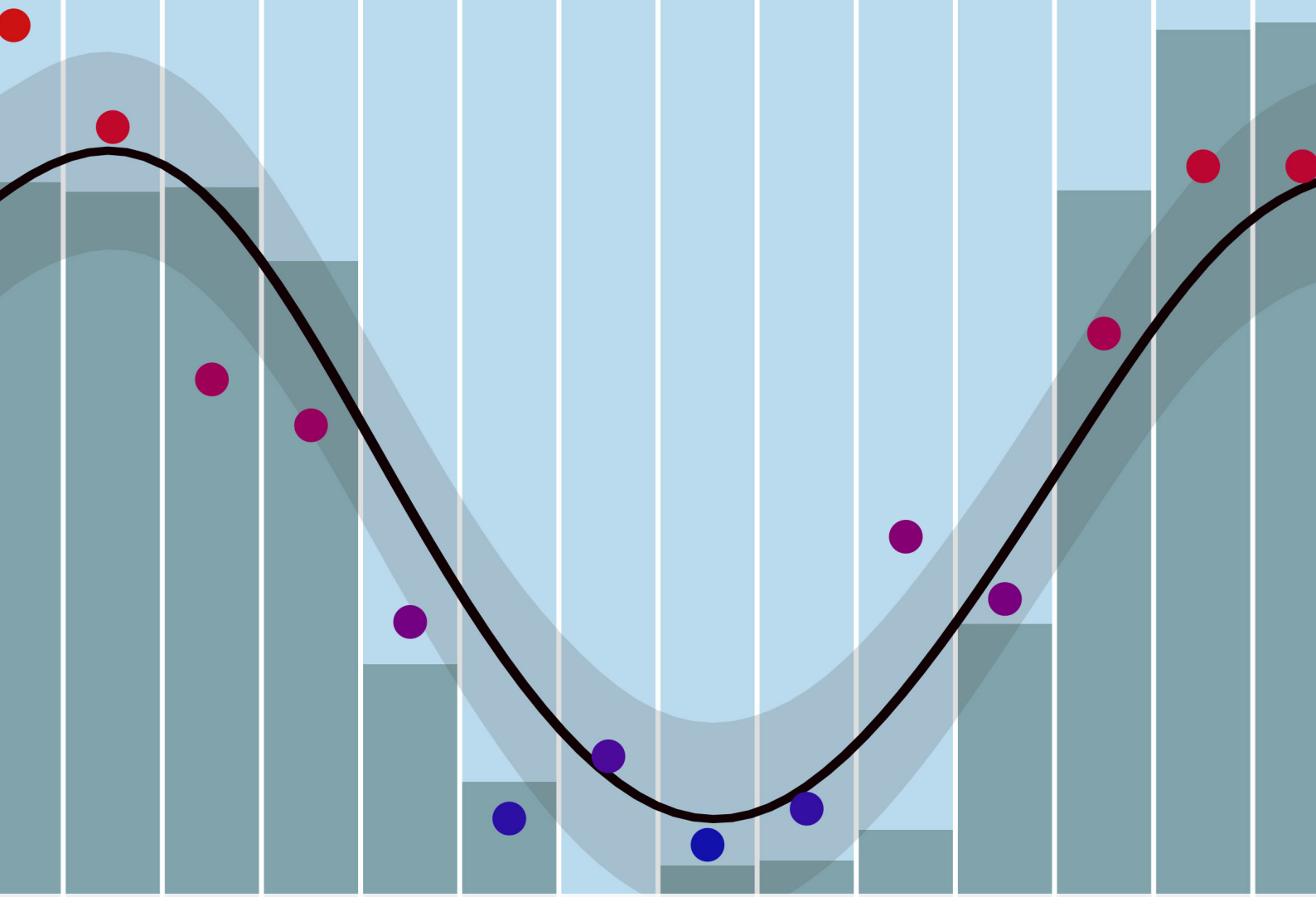
NEW INSIGHTS INTO MICROBIAL ECOLOGY THROUGH SUBTLE NUCLEOTIDE VARIATION

EDITED BY: A. Murat Eren, Mitchell Sogin and Loïs Maignien
PUBLISHED IN: Frontiers in Microbiology



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88919-988-4
DOI 10.3389/978-2-88919-988-4

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

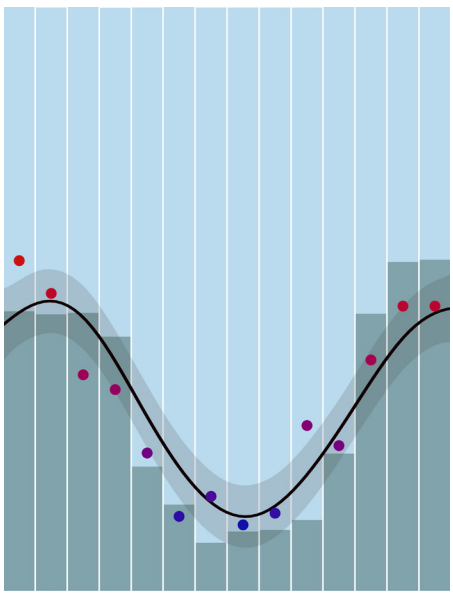
Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

NEW INSIGHTS INTO MICROBIAL ECOLOGY THROUGH SUBTLE NUCLEOTIDE VARIATION

Topic Editors:
A. Murat Eren, University of Chicago & Marine Biological Laboratory, USA
Mitchell Sogin, Marine Biological Laboratory, USA
Loïs Maignien, Marine Biological Laboratory, USA & Université de Bretagne Occidentale, France



Two SAR11 oligotypes with more than 99% sequence identity at the 16S ribosomal RNA gene-level fluctuates throughout the year as a function of water temperature at Cape Cod, MA, USA.
Cover image by A. Murat Eren

The 16S ribosomal RNA gene commonly serves as a molecular marker for investigating microbial community composition and structure. Vast amounts of 16S rRNA amplicon data generated from environmental samples thanks to the recent advances in sequencing technologies allowed microbial ecologists to explore microbial community dynamics over temporal and spatial scales deeper than ever before. However, widely used methods for the analysis of bacterial communities generally ignore subtle nucleotide variations among high-throughput sequencing reads and often fail to resolve ecologically meaningful differences between closely related organisms in complex microbial datasets. Lack of proper partitioning of the sequencing data into relevant units often masks important ecological patterns.

Our research topic contains articles that use oligotyping to demonstrate the importantance of high-resolution analyses of marker gene data, and providides further evidence why microbial ecologists should open the "black box" of OTUs identified through arbitrary sequence similarity thresholds.

Citation: Eren, A. M., Sogin, M., Maignien, L., eds. (2016). New Insights into Microbial Ecology through Subtle Nucleotide Variation. Lausanne: Frontiers Media. doi: 10.3389/978-2-88919-988-4

Table of Contents

frontiers in Microbiology

EDITORIAL
published: 24 August 2016
doi: 10.3389/fmicb.2016.01318

Editorial: New Insights into Microbial Ecology through Subtle Nucleotide Variation

A. Murat Eren[1,2]*, *Mitchell L. Sogin*[2] *and Loïs Maignien*[2,3]

[1] Department of Medicine, The University of Chicago, Chicago, IL, USA, [2] Marine Biological Laboratory, Josephine Bay Paul Center, Woods Hole, MA, USA, [3] Laboratory of Microbiology of Extreme Environnments, UMR 6197, Institut Européen de la Mer, Université de Bretagne Occidentale, Plouzane, France

Keywords: oligotyping, minimum entropy decomposition, 16S rRNA gene, microbial ecology, high resolution

The Editorial on the Research Topic

New Insights into Microbial Ecology through Subtle Nucleotide Variation

Characterizing the community structure of naturally occurring microbes through marker gene amplicons has gained widespread acceptance for profiling microbial populations. The 16S ribosomal RNA (rRNA) gene provides a suitable target for most studies since (1) it meets the criteria for robust markers of evolution, e.g., both conserved and rapidly evolving regions that do not undergo horizontal gene transfer, (2) microbial ecologists have identified widely adopted primers and protocols for generating amplicons for sequencing, (3) analyses of both cultivars and environmental DNA have generated well-curated databases for taxonomic profiling, and (4) bioinformaticians and computational biologists have published comprehensive software tools for interpreting the data and generating publication-ready figures. Since the initial descriptions of high-throughput sequencing of 16S rRNA gene amplicons to survey microbial diversity, we have witnessed an explosion of association-based inferences of interactions between microbes and their environment.

Despite these advances, the field of microbial ecology faces numerous technical challenges. Sampling and storage strategies, DNA extraction protocols, limitations of the so called "universal" PCR primers, random sequencing errors, and the identification of ecologically relevant units can bias interpretations of observations based on 16S rRNA gene data. Although microbiologists comprehend most of these challenges, the need for handling large number of sequences, and to partition these complex data into appropriate proxies for environmental genomes caught almost everyone off-guard.

De novo clustering of short reads into operational taxonomic units (OTUs) based on "pairwise sequence similarities" quickly became the primary way to partition sequencing data into ecological units as this approach significantly out-performed analyses that relied strictly upon taxonomy. On the other hand, as random sequencing errors can dramatically increase the number of mismatches between two aligned reads, the underlying principle of most *de novo* clustering algorithms that rely on the edit distance was prone to inflating the diversity estimations. The use of 97% sequence similarity threshold emerged as a *de facto* standard, and has successfully reduced the impact of erroneous OTUs on diversity estimations. However, the computational convenience this arbitrary threshold offers has been at the expense of accurate ecological inference, as 3% OTUs are often phylogenetically mixed, and inconsistent (Koeppel and Wu, 2013; Eren et al., 2014; Nguyen et al., 2016).

Oligotyping (Eren et al., 2013) proposes an alternative way to decompose marker gene amplicons. It first considers the entire sequencing data to identify variable nucleotide positions,

OPEN ACCESS

Edited by:
Matthias Hess,
University of California, Davis, USA

Reviewed by:
George Tsiamis,
University of Patras, Greece

***Correspondence:**
A. Murat Eren
meren@uchicago.edu

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *30 May 2016*
Accepted: *09 August 2016*
Published: *24 August 2016*

Citation:
Eren AM, Sogin ML and Maignien L (2016) Editorial: New Insights into Microbial Ecology through Subtle Nucleotide Variation. Front. Microbiol. 7:1318. doi: 10.3389/fmicb.2016.01318

and then utilizes only those positions that show significant variation to partition reads into oligotypes. The identification of variable nucleotide positions in the oligotyping workflow relies on Shannon entropy (Shannon, 1948), which is a measure of information uncertainty (Jost, 2006). The association between the measured entropy and the diversity of nucleotides at a given nucleotide position in a dataset of sequences allows the identification of nucleotide positions that likely carry phylogenetically important signal. The departure from pairwise sequence alignments, and the use of entropy-based decomposition strategy, makes it possible to resolve closely related but distinct taxa that differ by as little as one nucleotide at the sequenced region.

Our research topic contains original research and method papers that employs oligotyping of microbial community data to investigate ecological questions in divergent environments including the human oral cavity (Mark Welch et al.), mammalian guts (Menke et al.), deep-sea sediments (Buttigieg and Ramette), as well as freshwater (Newton and McLellan), sewage (Fisher et al.), marine (Delmont et al.), and soil (Turlapati et al.) ecosystems. In a study that cuts across multiple environments, Schmidt et al. uses oligotyping to investigate the Vibrio ecology in environmental, as well as host- and substrate-associated habitats. Ramette and Buttigieg implements an R package for entropy-based decomposition procedures, and their software library contains additional approaches, such as the "broken stick model" procedure to identify low-abundance oligotypes that could be generated by chance alone, and a "one-pass entropy profiling" approach to efficiently identify those OTUs whose decomposition into oligotypes would most likely explain concealed diversity (Ramette and Buttigieg). Finally, Utter et al. reconcile the individuality, stability, and variability of the oral microbial communities in the context of "spatial structure" of microbes in dental plaque by combining high-resolution depiction of microbial community data with high-resolution imaging of multi-taxa microbial consortia in the human oral cavity (Mark Welch et al., 2016).

Most articles in this collection demonstrate the importance of high-resolution analyses, and provide further evidence that reveals the need to open the "black box" of OTUs in microbial ecology. Doing so not only allows finer representation of the microbial diversity in a wide range of ecosystems, but also improves the ecological signal for downstream analyses that aim to infer correlations (McLellan and Eren, 2014; Reveillaud et al., 2014; Eren et al., 2015; Kleindienst et al., 2015).

While oligotyping demonstrates the efficacy of an entropy-based concept to partition closely related taxa, the algorithm minimum entropy decomposition suggests that the use of information theory can be generalized to analyze entire sets of marker gene data (Eren et al., 2014; Ramette and Buttigieg). The ideal result of a properly partitioned marker gene dataset will have the minimum number of units that contains minimum entropy (i.e., none of the nucleotide positions in final units will have entropy that exceeds the expected error rate of the sequencing device), which in fact can be achieved through multiple ways. Indeed, the search for algorithms that can provide single-nucleotide resolution without relying on arbitrary percent similarity thresholds is not limited to entropy-based approaches: studies that aim to address the same issue include distribution-based clustering (Preheim et al., 2013), cluster-free filtering (Tikhonov et al., 2015), Swarm (Mahé et al., 2015), and recently introduced DADA2 (Callahan et al., 2016).

Potential new directions for a more accurate depiction of microbial communities through marker gene amplicons come with new questions. What should microbial ecologists do with all the data they have generated, and plan to generate during the years to come? What are the computational and ecological issues that will need to be addressed for new methods to be more accessible in the field? Although our collection does not promise answers to these questions, we hope it will further stimulate the community of microbial ecologists and the developers of widely used software platforms to move beyond the use of OTUs that require arbitrary percent similarity cut-offs.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

AME was supported by the University of Chicago and the Marine Biological Laboratory collaboration award.

REFERENCES

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi: 10.1038/nmeth.3869

Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., et al. (2013). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* 4, 1111–1119. doi: 10.1111/2041-210X.12114

Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H., and Sogin, M. L. (2014). Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J.* 9, 968–979. doi: 10.1038/ismej.2014.195

Eren, A. M., Sogin, M. L., Morrison, H. G., Vineis, J. H., Fisher, J. C., Newton, R. J., et al. (2015). A single genus in the gut microbiome reflects host preference and specificity. *ISME J.* 9, 90–100. doi: 10.1038/ismej.2014.97

Jost, L. (2006). Entropy and diversity. *Oikos* 113, 363–375. doi: 10.1111/j.2006.0030-1299.14714.x

Kleindienst, S., Seidel, M., Ziervogel, K., Grim, S., Loftis, K., Harrison, S., et al. (2015). Chemical dispersants can suppress the activity of natural oil-degrading microorganisms. *Proc. Natl. Acad. Sci. U.S.A.* 112, 14900–14905. doi: 10.1073/pnas.1507380112

Koeppel, A. F., and Wu, M. (2013). Surprisingly extensive mixed phylogenetic and ecological signals among bacterial operational taxonomic units. *Nucleic Acids Res.* 41, 5175–5188. doi: 10.1093/nar/gkt241

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. *Peer J.* 3:e1420. doi: 10.7717/peerj.1420

Mark Welch, J. L., Rossetti, B. J., Rieken, C. W., Dewhirst, F. E., and Borisy, G. G. (2016). Biogeography of a human oral microbiome at the micron scale. *Proc. Natl. Acad. Sci. U.S.A.* 113, 201522149. doi: 10.1073/pnas.1522149113

McLellan, S. L., and Eren, A. M. (2014). Discovering new indicators of fecal pollution. *Trends Microbiol.* 22, 697–706. doi: 10.1016/j.tim.2014.08.002

Nguyen, N.-P., Warnow, T., Pop, M., and White, B. (2016). A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *npj Biofilms Microbiomes* 2, 16004. doi: 10.1038/npjbiofilms.2016.4

Preheim, S. P., Perrotta, A. R., Martin-Platero, A. M., Gupta, A., and Alm, E. J. (2013). Distribution-based clustering: using ecology to refine the operational taxonomic unit. *Appl. Environ. Microbiol.* 79, 6593–6603. doi: 10.1128/AEM.00342-13

Reveillaud, J., Maignien, L., Eren, A. M., Huber, J. A., Apprill, A., Sogin, M. L., et al. (2014). Host-specificity among abundant and rare taxa in the sponge microbiome. *ISME J.* 8, 1198–1209. doi: 10.1038/ismej.2013.227

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Technol. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Tikhonov, M., Leach, R. W., and Wingreen, N. S. (2015). Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *ISME J.* 9, 68–80. doi: 10.1038/ismej.2014.117

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 22 April 2016
doi: 10.3389/fmicb.2016.00564

Individuality, Stability, and Variability of the Plaque Microbiome

Daniel R. Utter[1,2]*, *Jessica L. Mark Welch*[2,3] *and Gary G. Borisy*[2]

[1] *Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA,* [2] *Department of Microbiology, The Forsyth Institute, Cambridge, MA, USA,* [3] *Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA, USA*

Dental plaque is a bacterial biofilm composed of a characteristic set of organisms. Relatively little information from cultivation-independent, high-throughput analyses has been published on the temporal dynamics of the dental plaque microbiome. We used Minimum Entropy Decomposition, an information theory-based approach similar to oligotyping that provides single-nucleotide resolution, to analyze a previously published time series data set and investigate the dynamics of the plaque microbiome at various analytic and taxonomic levels. At both the genus and 97% Operational Taxonomic Unit (OTU) levels of resolution, the range of variation within each individual overlapped that of other individuals in the data set. When analyzed at the oligotype level, however, the overlap largely disappeared, showing that single-nucleotide resolution enables differentiation of individuals from one another without ambiguity. The overwhelming majority of the plaque community in all samples was made up of bacteria from a moderate number of plaque-typical genera, indicating that the overall community framework is shared among individuals. Each of these genera fluctuated in abundance around a stable mean that varied between individuals, with some genera having higher inter-individual variability than others. Thus, at the genus level, differences between individuals lay not in the identity of the major genera but in consistently differing proportions of these genera from mouth to mouth. However, at the oligotype level, we detected oligotype "fingerprints," a highly individual-specific set of persistently abundant oligotypes fluctuating around a stable mean over time. For example, within the genus *Corynebacterium*, more than a dozen oligotypes were detectable in each individual, of which a different subset reached high abundance in any given person. This pattern suggests that each mouth contains a subtly different community of organisms. We also compared the Chinese plaque community characterized here to previously characterized Western plaque communities, as represented by analyses of data emerging from the Human Microbiome Project, and found no major differences between Chinese and Western supragingival plaque. In conclusion, we found the plaque microbiome to be highly individualized at the oligotype level and characterized by stability of community membership, with variability in the relative abundance of community members between individuals and over time.

Keywords: human microbiome, 16S rRNA, community dynamics, oral microbiota, community ecology

OPEN ACCESS

Edited by:
Angel Angelov,
Technische Universität München, Germany

Reviewed by:
Jean-Baptiste Ramond,
University of Pretoria, South Africa
Robin Anderson,
United States Department of Agriculture/Agricultural Research Service, USA
Anna Edlund,
UCLA School of Dentistry, USA

***Correspondence:**
Daniel R. Utter
dutter@g.harvard.edu

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *12 January 2016*
Accepted: *04 April 2016*
Published: *22 April 2016*

Citation:
Utter DR, Mark Welch JL and Borisy GG (2016) Individuality, Stability, and Variability of the Plaque Microbiome. Front. Microbiol. 7:564. doi: 10.3389/fmicb.2016.00564

INTRODUCTION

Understanding the baseline stability or variability of the human microbiota is important for evaluating the health significance of perturbations from baseline that may occur during disease, dietary change or antibiotic treatment. A major research effort, the Human Microbiome Project (HMP, http://hmpdacc.org/) was established to provide an integrated overview of the microbial communities that share our bodies. This and other 16S rRNA gene-based studies used high throughput, large-scale cross-sectional sampling and documented a tremendous range of compositional variability between individuals and even between oral sites (Nasidze et al., 2009; Zaura et al., 2009; Segata et al., 2012; Eren et al., 2014; Xu et al., 2014). Recent studies have emphasized both that the normal human microbiota, once established, can remain stable for months or even years (Faith et al., 2013; David et al., 2014) and that the microbiota can be highly variable over short time scales (Gajer et al., 2012; Flores et al., 2014). Thus, a full understanding of the meaning of stability or variability requires connecting the measure of stability both to commonly used community analysis metrics and to a more complete analysis of the organismal composition of the community.

Most studies of stability have drawn primarily upon summary community metrics, typically involving diversity metrics and/or distance metrics to quantify and relate communities over time. One popular analytic tool, UniFrac, provides a phylogeny-based distance metric for comparison of community composition (Lozupone and Knight, 2005). Although the metric is general in nature, it is typically used with taxonomic assignments based on Ribosomal Database Project (RDP, http://rdp.cme.msu.edu/) classification at the genus level or phylotypes defined at the operational taxonomic unit (OTU) level of >97% sequence identity. These analyses provide a useful overview and allow comparison of complex data sets, but do not address the stability or variability of microbial communities at lower taxonomic levels, even when there is signal in the sequenced region of the marker gene to distinguish closely related but distinct members.

The oral cavity provides an excellent test bed for exploring questions of microbiome stability because of its accessibility and the existence of a well-curated and annotated Human Oral Microbiome Database (HOMD, http://homd.org). The HMP, in addition to gut, skin and vaginal sites, included sampling of 9 different sites in the oral cavity. Initial analysis at the genus level characterized some of the similarities and differences among the oral microbial communities (Segata et al., 2012). However, relatively few studies have investigated oral microbial dynamics and, in general, they have been analyses of community composition via summary metrics at the genus or 97% OTU level.

Several studies have emphasized the stability of microbial communities. Costello et al. (2009) analyzed oral samples, saliva and tongue dorsum, on two successive days 3 months apart and showed that variation was less within individuals than between individuals, suggesting stability, and was less over 24 h than over 3 months. Stahringer et al. (2012) analyzed saliva from 82 individuals, found no systematic change in beta diversity over 5- and 10-year intervals, and concluded that the salivary microbiome showed long-term stability. David et al. (2014) analyzed the saliva of a single individual daily for a year and found community stability over periods of months. Cameron et al. (2015) analyzed the saliva of 10 subjects at 2-month intervals for a year, found no significant community differences over the year and, therefore, concluded stability.

In contrast, other studies have emphasized the variability of microbial communities over their stability. Ding and Schloss (2014) analyzed oral HMP data at 2 or 3 time points between 30 days and 1 year. They analyzed a range of body sites and found gut and vagina to be most stable whereas the oral cavity was reported to be least stable. A study of the tongue dorsum community from two individuals with daily sampling over a year emphasized the temporal variability in the tongue dorsum community, documenting drastic shifts in relative abundance of community members at daily time scales (Caporaso et al., 2011). A detailed study by Flores et al. (2014) analyzed multiple body sites, including the tongue dorsum, of 85 subjects weekly over 3 months. Their results pointed to variability of the microbial communities and that individuals differed in the degree to which their microbiomes were variable. Thus, oral microbial dynamics have been characterized paradoxically by the seemingly contradictory qualities of stability and variability.

In an effort to deepen understanding of the oral microbiome beyond summary measures of community membership, we have evaluated microbiome composition and temporal dynamics at the species or sub-species level, using high-resolution analysis of sequence data. Recently, we used an information theory-based approach called oligotyping (Eren et al., 2013) to re-analyze the HMP 16S rRNA gene sequence data of the entire oral microbiome at the single-nucleotide level. We compared the observed sequences to the curated HOMD (Dewhirst et al., 2010) and found that most sequences were exact or near-exact matches to known oral taxa. Some oral microbes differed from one another by only a few nucleotides in the sequenced region of the 16S rRNA gene, but the single-nucleotide resolution of oligotyping made it possible to distinguish them from one another. Some of these closely related sequences matched the same reference taxon in HOMD but were abundant at different oral sites. We interpret these distinctive distributions as demonstrating a level of ecological and functional diversity not previously recognized (Eren et al., 2014). We then applied oligotyping to the tongue dorsum by re-analyzing the Caporaso et al. (2011) data set. We identified a persistent core tongue dorsum microbiome but with rapidly (daily) fluctuating proportions of the characteristic taxa (Mark Welch et al., 2014): Some oligotypes were stable for months but underwent abrupt transitions to alternate oligotypes within days. However, it remained an open question whether this finding was specific to the tongue dorsum community, and the closely related salivary microbiome, or whether other oral communities exhibited similar community dynamics.

Recently, Jiang et al. (2015) published a high-quality data set of the plaque microbiome of eight different Chinese individuals over a period of 3 months. Their objective was to define a "dynamic core microbiome," the set of taxa present at all time points in all individuals. Consequently, they pooled the data from all eight individuals at each time point. However, the same data,

when kept un-pooled, provides an opportunity to investigate questions of stability and variability of the plaque microbiome within individuals, as well as to compare these samples to those of the Human Microbiome Project, collected in the United States. Here we re-analyze this high-quality plaque time series data with single-nucleotide resolution. Our results provide evidence for a unique community "fingerprint" in the plaque of each individual, consisting of a set of organisms and their relative abundances that fluctuate around a stable mean within each individual over the months-long sampling period.

METHODS

Sample Collection and Sequence Acquisition

This is a re-analysis of existing sequence data; procedures for informed consent, institutional approval, and sample collection and sequencing are described in the original publication (Jiang et al., 2015). Briefly, Jiang et al. collected supragingival plaque samples from eight healthy, 25- to 28-year-old Chinese subjects at eight time points: Day 0, Day 1, Day 3, Week 1, Week 2, Week 3, Month 1, and Month 3 for a total of 64 samples. In their study, the V4-V5 hypervariable region of the 16S rRNA gene was sequenced using Roche 454 GS-FLX pyrosequencing; sequence data was stored in the NCBI sequence read archive (SRA) under SRA accession number SRP049987 (http://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP049987; Jiang et al., 2015). The total data available from NCBI consisted of 359,565 reads, with an average of 5618 sequences per sample (SD = 923.8). To eliminate the artificial length variation among reads introduced by the original quality trimming, we re-trimmed each read to 336 nucleotides and removed the reads that were shorter, which reduced the size of the data set by less than 10%. These reads were then aligned against the GreenGenes reference alignment (McDonald et al., 2011; greengenes.lbl.gov/Download/OTUs/gg_otus_6oct2010/rep_set/gg_97_otus_6oct2010_aligned.fasta) using PyNAST version 0.1 (biocore.github.io/pynast/) (Caporaso et al., 2010), and we removed positions from the resulting alignment that consisted only of gap characters. The final data set contained a total of 301,657 reads.

We also downloaded data from Eren et al. (2014), which was part of an oligotyping re-analysis of data from The Human Microbiome Project Consortium (2012). The downloaded data consisted of supplemental data sets, Dataset_S01 (V1-V3 oligotypes) and Dataset_S02 (V3-V5 oligotypes). No re-analysis was done on this data; it was only collapsed to the genus level and reformatted for comparison to both the original Jiang et al. (2015) work and our re-analysis of it.

Minimum Entropy Decomposition and Taxon Assignment

Minimum Entropy Decomposition (MED; Eren et al., 2015) is an automated data analysis algorithm that operates on the same principle as oligotyping (Eren et al., 2013). MED uses high-entropy nucleotide positions to iteratively partition a given collection of reads into *de novo* bins we will refer to as oligotypes. We used the MED pipeline, version 2.0, (Eren et al., 2015) to decompose the data set. The "minimum substantive abundance" criterion (−M) was set to 60, and the "maximum variation allowed" criterion (−V) was set to 3. Minimum entropy decomposition of 301,657 reads generated 333 oligotypes, retaining a total of 227,991 sequences. Of the 73,666 reads removed, 58,591 reads were removed due to the minimum substantive abundance criterion and 15,075 reads were removed due to the maximum variation allowed criterion. A visual breakdown of the taxonomy of reads lost in each step of the process, from alignment to final output, is presented in Supplementary Image 1. Taxonomy was then assigned to each of the 333 oligotypes by querying the representative sequence of each oligotype against the Human Oral Microbiome Database (HOMD) RefSeq v.13.2 (www.homd.org, Dewhirst et al., 2010) using the Global Alignment Search Tool (GAST; Huse et al., 2008). Taxonomic data and abundance by oligotype are presented in Supplementary Data Sheet 1.

Data Processing and Figure Creation

We used R (version 3.2.2; R Core Team, 2015) for all post-MED data analysis. We used the *metaMDS* function in the vegan package (Oksanen et al., 2013) to generate the MDS analysis shown in **Figures 2**, **3**. All figures were created with the *ggplot* function in the ggplot2 package (Wickham, 2009). After generation in R, figures were cleaned and processed for final publication with Inkscape (version 0.91, http://inkscape.org/).

RESULTS

Analysis of Plaque 16S rRNA Gene Sequencing Data with Single-Nucleotide Resolution

We used Minimum Entropy Decomposition (MED; Eren et al., 2015) to re-analyze the time series data generated from the supragingival plaque samples of eight individuals at eight time points over 3 months (Jiang et al., 2015). This data set was of interest because it presented an opportunity to analyze the temporal dynamics of the plaque microbiome at single-nucleotide resolution.

However, before pursuing the question of microbiome dynamics, we had to address a striking difference between the genus-level results of Jiang et al. (2015) and those emerging from the Human Microbiome Project. Jiang and co-authors reported a high percentage of an unusual taxon, *Parascardovia*, and a low percentage of *Actinomyces* in contrast to the Human Microbiome data as re-analyzed by Eren et al. (2014), which showed no *Parascardovia* and substantial amounts of *Actinomyces* (**Figure 1**). Otherwise, the two studies were in general agreement. We asked whether the disparity could have arisen from genetic, cultural, or environmental differences between the two populations or from technical causes such as method of informatics analysis or sequencing strategy. The HMP data was collected and analyzed over two different regions of the 16S RNA gene, V1-V3 and V3-V5. Although these two regions give slightly different abundance values for plaque genera, their results were similar and neither contained *Parascardovia*. The Jiang et al. study sequenced the V4-V5 region, which overlaps

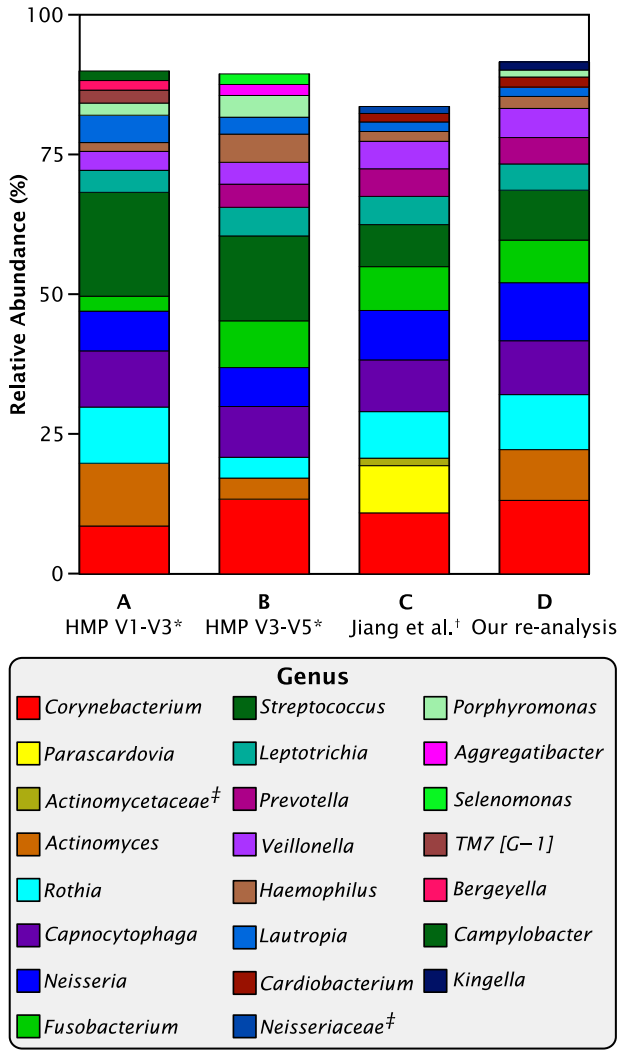


FIGURE 1 | Abundant genera in Chinese and Western plaque. Mean relative abundances of the most abundant 15 taxa in HMP and Jiang et al. data. **(A)** *Eren et al. (2014) re-analysis of Human Microbiome Project V1-V3 supragingival plaque data; **(B)** *Eren et al. (2014) re-analysis of HMP V3-V5 data; **(C)** †data from supplemental table S2 of Jiang et al. (2015); **(D)** our re-analysis of Jiang et al. data. ‡Sequences in table S2 of Jiang et al. (2015) that were labeled to the family level only.

the V3-V5 region of the HMP study. Therefore, the location of the sequenced region was not a sufficient explanation for the disparity.

When we re-analyzed the Jiang et al. sequence data with the MED pipeline and then categorized the resulting oligotypes into genera, the disparity disappeared (**Figure 1**). In our re-analysis of the Jiang et al. data, no sequences were identified as *Parascardovia* and the abundance of *Actinomyces* was within the range of variation of the HMP data. Consequently, we conclude that the disparity between the two data sets is likely to be methodological and does not reflect any genetic, cultural, or environmental difference between the two populations. The genera *Parascardovia* and *Actinomyces* are in the same phylum, Actinobacteria, and it is possible that sequences representing *Actinomyces* were misclassified as *Parascardovia* in Jiang et al.'s original analysis. We evaluated possible mechanisms of error in classification (Supplementary Data Sheet 2) but were unable to identify the exact source of the misclassification. Nevertheless, the salient point for this study is that when the sequence data was processed through the MED pipeline, the disparity between the Jiang et al. and the HMP results disappeared.

Composition of Chinese and Western Supragingival Plaque is Broadly Similar at the Genus Level

Next, we wanted to understand how Chinese plaque composition compared to Western plaque, as represented by the HMP data. As a basis for comparison of the Chinese and the two HMP data sets, we used the relative abundances of the 40 genera shared between all three data sets. These shared genera made up the vast majority of all the sequence reads: non-shared taxa comprised only 3.2% of the V1-V3 and 2.2% of the V3-V5 re-analyzed HMP data (Eren et al., 2014), and 0.9% of the re-analyzed Jiang et al. data. Differential presence or absence of low-abundance, non-shared genera may result from technical differences in experimental design and we reasoned that eliminating them from the analysis would allow for a more parsimonious comparison across studies. Supplementary Data Sheet 3 provides a breakdown of mean relative abundances by genus in each data set.

Comparing the re-analyzed data from Jiang et al. to our analysis of the HMP data revealed that Chinese and American plaque have a similar overall composition when viewed at the genus level (**Figure 2**). On a multi-dimensional scaling (MDS) plot of Bray-Curtis distances, covariance ellipses representing the HMP data from the V1-V3 (blue) and the V3-V5 (cyan) regions of the 16S rRNA gene overlap but do not entirely coincide, showing that similar but not identical communities are recovered when the same DNA sample is analyzed using two different regions of the marker gene. The re-analyzed Jiang et al. data in **Figure 2** (red) also overlaps substantially with the HMP data, showing that within the range of experimental variation there is no detectable difference at the genus level between Chinese and Western plaque.

Intra-Individual Plaque Variability is Less than Inter-Individual Variability

The time series information provided by the Jiang et al. data set provides an unusual opportunity to analyze intra-individual variation in plaque over time relative to inter-individual variation within the same study. However, to evaluate intra-individual temporal dynamics, we needed to establish a level of analytic resolution sufficient to clearly distinguish the plaque microbiota of individuals. **Figure 3** shows multidimensional scaling (MDS) plots at various levels of analytic resolution. Analyses for all samples were carried out based on relative abundances of all taxa. At the phylum level (**Figure 3A**), the plots for individuals largely overlapped one another. Not surprisingly, the overlap was greatly reduced at the genus level (**Figure 3B**). Close inspection of **Figure 3B** shows that samples from most of the individuals

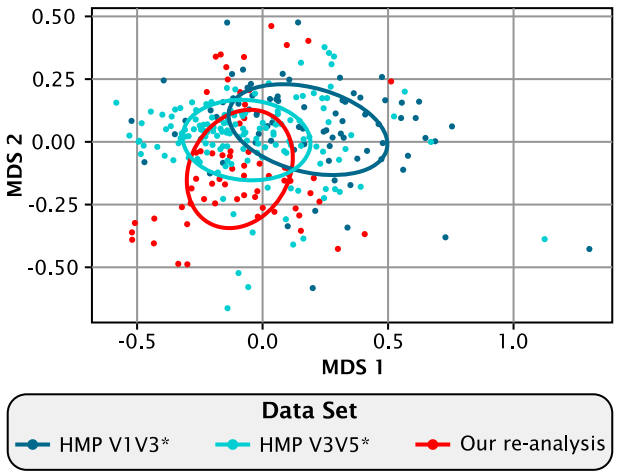


FIGURE 2 | Compositional similarity between Chinese and Western plaque. Multidimensional scaling plot showing the relatedness among the re-analyzed *(Eren et al., 2014) HMP V1-V3 (dark blue), and HMP V3-V5 data sets (light blue), and our re-analysis of the Jiang et al. (2015) data (red). Each dot represents an individual sample. Sample distances were calculated by the Bray-Curtis dissimilarity index using the relative abundances of the 40 genera shared between all three data sets. Ellipses bound the covariance of each data set and are centered on the mean of that set.

overlapped with those from only two or three neighbors, and samples from one individual (E, colored orange) were distinct from all the other individuals. Similar results were obtained at the operational taxonomic unit (OTU) level of 97% sequence identity (**Figure 3C**). Thus, at both the genus and 97%-OTU levels of resolution, individuals showed variation in plaque composition from sample to sample, with the variation for each individual, in most cases, ranging over a relatively small fraction of the total range of variability of the population. However, overlap among individuals still occurred.

Increasing the resolution beyond the genus or 97%-OTU level revealed far more dramatic distinctions between individuals. When the same data was plotted at the oligotype level (**Figure 3D**), the covariance overlap between individuals diminished drastically and individuals occupied largely non-overlapping regions of the plot. The stress function of the MDS plot in **Figure 3D** has a relatively high value of 0.2, but the topology was consistent in 49 out of 50 trials (Supplementary Image 3). This result shows that the single-nucleotide-level resolution of oligotyping reveals individual-level differences that are less apparent at the genus or 97%-OTU level. Taken together, our analysis demonstrates that the level of similarity observed within and between individuals is dependent on the taxonomic resolution at which the data is analyzed and that single-nucleotide resolution enables differentiation of individuals from one another without ambiguity.

Moving from Summary Metrics to Understanding Community Composition

Summary metrics such as MDS plots based on the Bray-Curtis dissimilarity index provide a measure of the degree of overall difference between microbial communities. However, like any summary metric, they inevitably obscure underlying key information. For example, the plots *per se* do not distinguish between differences arising from the presence of the same taxa in differing proportions or from the presence of different taxa. This distinction is of biological importance because it reflects upon the fundamental membership of microbial communities. Understanding the nature of the differences in the plaque community between individuals therefore requires moving from summary metrics to a more detailed analysis of the data itself, specifically the community composition of individual samples.

Deconstructing each sample by analyzing the relative abundance of taxa revealed that most of the plaque community was made up of bacteria from a moderate number of plaque-typical genera. A set of 17 genera was present in every individual at almost every time point and collectively made up between 80 and 99% of each plaque sample (**Figure 4A**). Thus, at the genus level of taxonomic resolution, the bulk of the plaque community was composed of a consistent set of taxa in all individuals, supporting the view of a core temporal plaque microbiome at the genus level.

To assess the stability of taxonomic composition and its consistency across individuals, we used the straightforward metrics of mean, standard deviation, and coefficient of variation. For the 8 samples from each individual, we calculated the mean abundance and its standard deviation for each genus; deviations from the mean are plotted in **Figure 4B** for the 10 genera with the highest mean relative abundances. For most taxa, the fluctuations were within 2 standard deviations from the mean, as expected because the mean and standard deviation are themselves calculated from the eight data points. More interestingly, the mean relative abundance for these major genera was stable over 3 months within each individual. Taxa exhibited shifts in relative abundance over time within an individual, but the shifts were generally fluctuations around an individual mean rather than displaying an increasing or decreasing trend (**Figure 4B**). Further, the mean was relatively constant across individuals for some genera, but in other genera differed substantially from individual to individual. This distinction can be observed visually in **Figure 4B**, as illustrated by the even distribution of *Fusobacterium* (shown in bright green) in contrast to the variability of *Prevotella* (maroon) and *Neisseria* (dark blue). The distinction is also evident in the coefficient of variation. Five relatively constant genera, *Streptococcus*, *Corynebacterium*, *Capnocytophaga*, *Fusobacterium*, and *Actinomyces*, each had a coefficient of variation between 58 and 68% over all 64 samples (**Table 1**) and between 40 and 66% on average within the 8 samples from a single individual (**Table 1**). The more variable genera, such as *Neisseria* and *Prevotella*, had higher coefficients of variation, 126 and 155% respectively across all 64 samples, and 90 and 97% on average within each individual. Some fluctuations were large as a fraction of the overall community, such as the shift of *Prevotella* in individual E from 25 to 4% and back to 21% of the community in consecutive samples, or the shift in *Neisseria* in individual D from 3 to 19% and back to 4% (**Figure 4A**). Interestingly, the two individuals in whom the aerobic *Neisseria* was most abundant, individuals J and K, were also the individuals

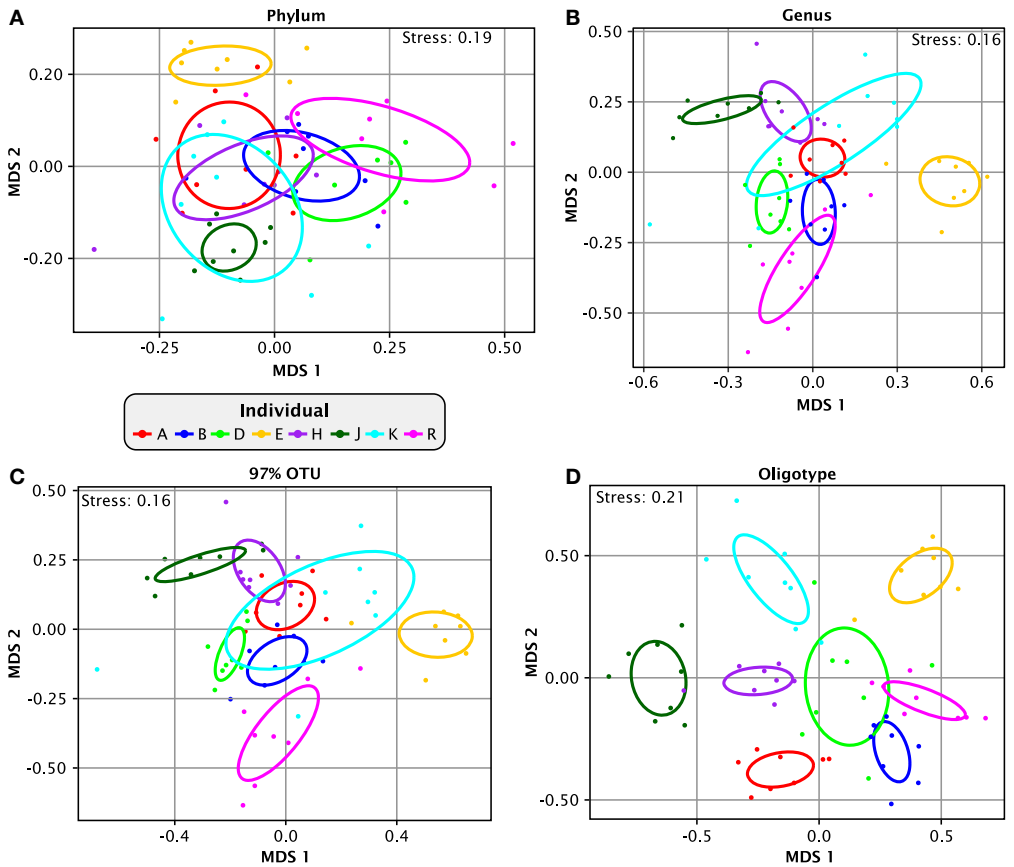


FIGURE 3 | Apparent similarity of samples depends on taxonomic resolution. Multidimensional scaling (MDS) plots of relatedness of samples from different individuals. Bray-Curtis dissimilarity index was calculated based on relative abundances in each individual of all **(A)** phyla, **(B)** genera, **(C)** OTUs at 97% identity and **(D)** oligotypes. For **(C)**, 97% OTUs were created by clustering the oligotype representative sequences using the centroid clustering method according to the Bray-Curtis distances. From this clustered matrix we binned together all oligotypes that were at least 97% similar to any oligotype in the same bin but were not already included in another bin. The ellipses mark the covariance of each individual data set; they are centered on the mean of each individual and colored by individual.

with the lowest fraction of the facultatively aerobic *Streptococcus* (**Figure 4A**). In summary, differences between individuals, as assayed at the genus level, lay not in the identity of the major genera but in consistently differing proportions of these genera from mouth to mouth.

Analysis with single-nucleotide resolution, however, revealed that within certain genera, each individual carried a distinctive set of organisms, as revealed by a distinctive pattern of oligotypes distinguishable by at least one nucleotide in the sequenced portion of their 16S rRNA gene. To illustrate this point, we decomposed *Corynebacterium*, the most abundant genus, into its 24 distinct oligotypes (**Figure 5A**). Between 4 and 19 *Corynebacterium* oligotypes were detectable in a single time point and between 15 and 24 unique oligotypes were detectable in each individual. However, only a handful of these oligotypes reached high abundance in each individual, and these oligotypes tended to maintain that high abundance within the individual over time. Unlike the genus-level analysis, visual comparison of individuals at the oligotype level (e.g., individuals A and H; **Figure 5A**) showed that individuals had strikingly different oligotype profiles, defining "profiles" as the combination of community membership and relative abundance. **Figure 5B** displays the anomaly from the mean for eight *Corynebacterium* oligotypes that were of high relative abundance (mean >10% of the *Corynebacterium*) in any individual. As in the genus-level analysis, the abundance of taxa fluctuated about a mean that was stable for each individual. A stable mean within an individual, but sharp differences between individuals, was confirmed by examination of the coefficient of variation: for the 8 *Corynebacterium* oligotypes of high relative abundance, the mean within-mouth coefficient of variation was 45% in mouths in which that oligotype was >10% of the *Corynebacterium* but was much higher, 173%, across all samples from all individuals, reflecting the consistent abundance of these oligotypes in some mouths and their near-absence in others (**Table 2**).

Similar individual-characteristic results were obtained for most other genera (Supplementary Image 4). Oligotype abundance was highly dynamic, but as with *Corynebacterium*, the oligotypes fluctuated around a stable mean (Supplementary Images 4, 5). However, a few oligotypes, for example oligotypes

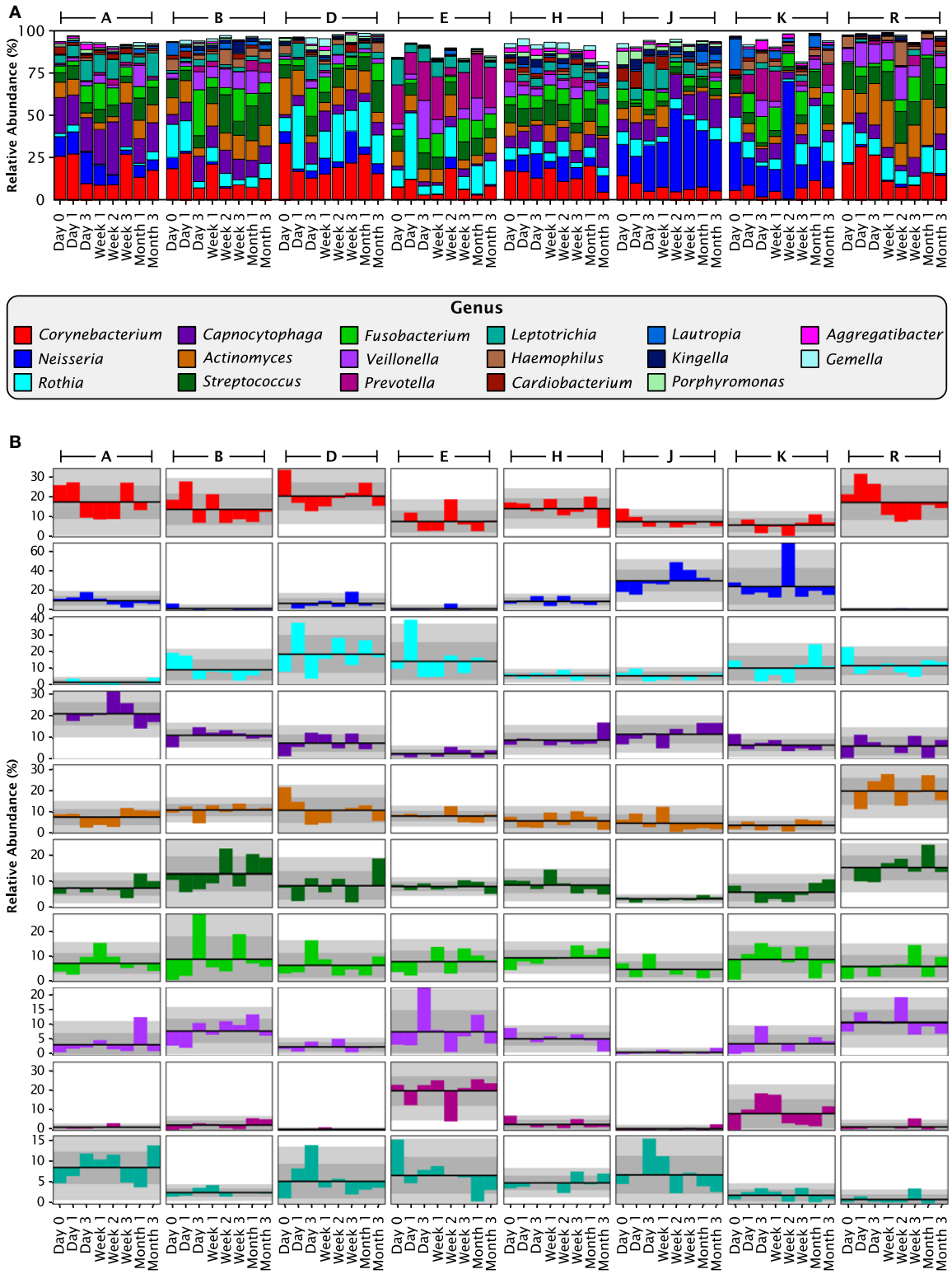


FIGURE 4 | Stable differences between individuals at the genus level. (A) Relative abundances for each individual at each time point, for all 17 genera with greater than 1% mean relative abundance over all 64 samples. Together these genera compose 93.8% of the data set. **(B)** Anomaly from the mean relative abundance for each sample from each individual. The mean relative abundance for an individual is marked by the dark line, and one and two standard deviations by the dark and light gray fields, respectively. Columns represent individuals, and rows represent genera, with colors as in **(A)**.

TABLE 1 | Temporal stability varies between genera.

	A		B		D		E		H		J		K		R		Overall		
	M	CV	M	CV	M	CV	M	CV	M	CV	M	CV	M	CV	M	CV	M	CV	Mean CV
Cor	17.4	48	13.9	56	20.4	34	7.8	70	14.3	35	7.7	41	6.1	58	17.2	50	13.1	60	49
Nei	9.5	54	1.3	169	6.8	80	1.2	190	8.8	46	30.2	36	24.2	78	0.8	71	10.3	126	90
Rot	1.9	85	9.4	65	18.8	60	14.4	77	5.9	35	5.8	46	10.4	71	11.8	48	9.8	84	61
Cap	20.9	25	11.3	25	7.7	52	2.9	64	9.1	37	11.8	36	6.8	40	6.4	68	9.6	64	43
Act	7.7	49	11.1	26	10.9	54	8.2	30	6	58	4.8	87	3.9	57	19.8	32	9	68	49
Str	7.7	38	13	51	8.5	63	8.3	22	8.8	35	3.5	25	6.1	57	15.3	30	8.9	58	40
Fus	7.4	57	9.1	101	6.6	76	8.1	56	9.6	34	4.9	66	8.9	63	6.1	75	7.6	68	66
Vei	3.3	120	7.9	51	2.5	62	7.6	94	5.2	44	0.7	118	3.6	87	10.7	39	5.2	92	77
Pre	1.4	66	2.5	94	0.4	120	19.9	38	2.8	82	0.6	167	8.4	88	1.6	120	4.7	155	97
Lep	8.5	46	2.6	39	5.2	78	6.7	66	4.9	36	6.8	66	1.9	75	0.9	133	4.7	83	67

For each genus (row) the mean relative abundance (M) and coefficient of variation (CV) are shown. The CV is presented as a percentage. In columns A, B, D, E, H, J, K, and R, mean and CV are calculated across the 8 samples from the respective volunteer. The mean and CV columns under the heading "Overall" were calculated from all 64 samples. The Mean CV column is the mean of all the CVs calculated from the individuals, i.e., the mean of [CV(A) ... CV(R)]. Genus abbreviations: Cor, Corynebacterium; Nei, Neisseria; Rot, Rothia; Cap, Capnocytophaga; Act, Actinomyces; Str, Streptococcus; Fus, Fusobacterium; Vei, Veillonella; Pre, Prevotella; Lep, Leptotrichia.

TABLE 2 | Oligotypes are stably abundant within an individual but not between individuals.

	A		B		D		E		H		J		K		R		Overall	
	M	CV	M	CV	M	CV	M	CV	M	CV	M	CV	M	CV	M	CV	M	CV
Cor_01	1.4	42	52.7	**15**	52.2	**25**	35.1	**46**	78.6	**8**	15.3	**56**	28.7	**49**	52.2	**21**	39.5	64
Cor_02	41.5	**34**	0.4	138	1.1	70	0.6	124	0.2	124	35.4	**33**	0.5	98	0	283	9.9	178
Cor_03	2.6	68	2.8	64	6.2	59	24.6	**47**	1.2	120	27.4	**63**	39	**59**	11.3	**58**	14.4	119
Cor_04	29.2	**55**	8.6	67	0.4	173	4.6	107	0.3	264					18.2	**40**	7.7	157
Cor_05	0.3	194	20.8	**40**	11.8	**104**	4.4	158	5.1	81	1.1	141	3.8	74	6	45	6.7	129
Cor_06	0	186	0.1	200	10.8	**32**	11.1	**34**	0.5	108	0.2	186	1.1	102	0.2	119	3	166
Cor_10							0	283	3.7	130	13.4	**72**	0.1	283			2.2	266
Cor_12							0	283	0.1	217			19.4	**56**	0.1	200	2.4	303
Mean of "Overall CV"																		173
Mean of abundant CVs																		45

Mean relative abundance (M) and coefficient of variation (CV) are shown for each Corynebacterium oligotype that comprises a mean of at least 10% of the total Corynebacterium reads in any individual. Mean and CV values are left blank when the oligotype was not detected in any of the 8 samples in that individual. The CV is presented as a percentage. Bolded values indicate CVs for oligotypes with a mean relative abundance of at least ≥10% of the total Corynebacterium reads in that individual. Table headings A, B, D, E, H, J, K, R represent the individual whom the statistics below represent. The mean and CV columns under the heading entitled "Overall" were calculated based the oligotype relative abundances from all 64 samples.

of *Streptococcus* in individuals J and R, *Rothia* in individual B, and *Neisseria* in individual J, broke this generality by showing a transition from low abundance to high abundance, or vice versa (Supplementary Image 5). We may nevertheless conclude that the oligotype-level composition of the microbiota of plaque is distinctive to individuals. When viewed together, oligotype profiles across all genera provided a unique oligotype "fingerprint" for each individual.

Since species designations are the accepted standard within the oral microbiome community, it is important to relate the oligotype patterns to a species-level analysis. Of the 24 *Corynebacterium* oligotypes, 21 represent the species *C. matruchotii* and the remaining 3 oligotypes represent *C. durum* (**Figure 5C**). When individuals were analyzed at species level, all individuals in the data set showed the same two *Corynebacterium* species. For example, comparison of individuals A and H, who have completely distinct *Corynebacterium* compositions at the oligotype level, as shown in **Figure 5A**, showed indistinguishable compositions at the species level with *C. matruchotii* the dominant species and *C. durum* a minor species (**Figure 5C**, Supplementary Data Sheet 1). Thus, some of the distinctions in microbiota between individuals that were so visible at the oligotype level were obscured even at the species level.

The degree to which oligotypes of the 16S ribosomal RNA gene provided better-than-species-level resolution was dependent on the genus. The genus *Capnocytophaga* resembled *Corynebacterium* in having an abundance of oligotypes resulting in a highly distinctive oligotype fingerprint for each individual, even for individuals whose species-level composition was similar (Supplementary Image 4). Within *Streptococcus,* by contrast, the

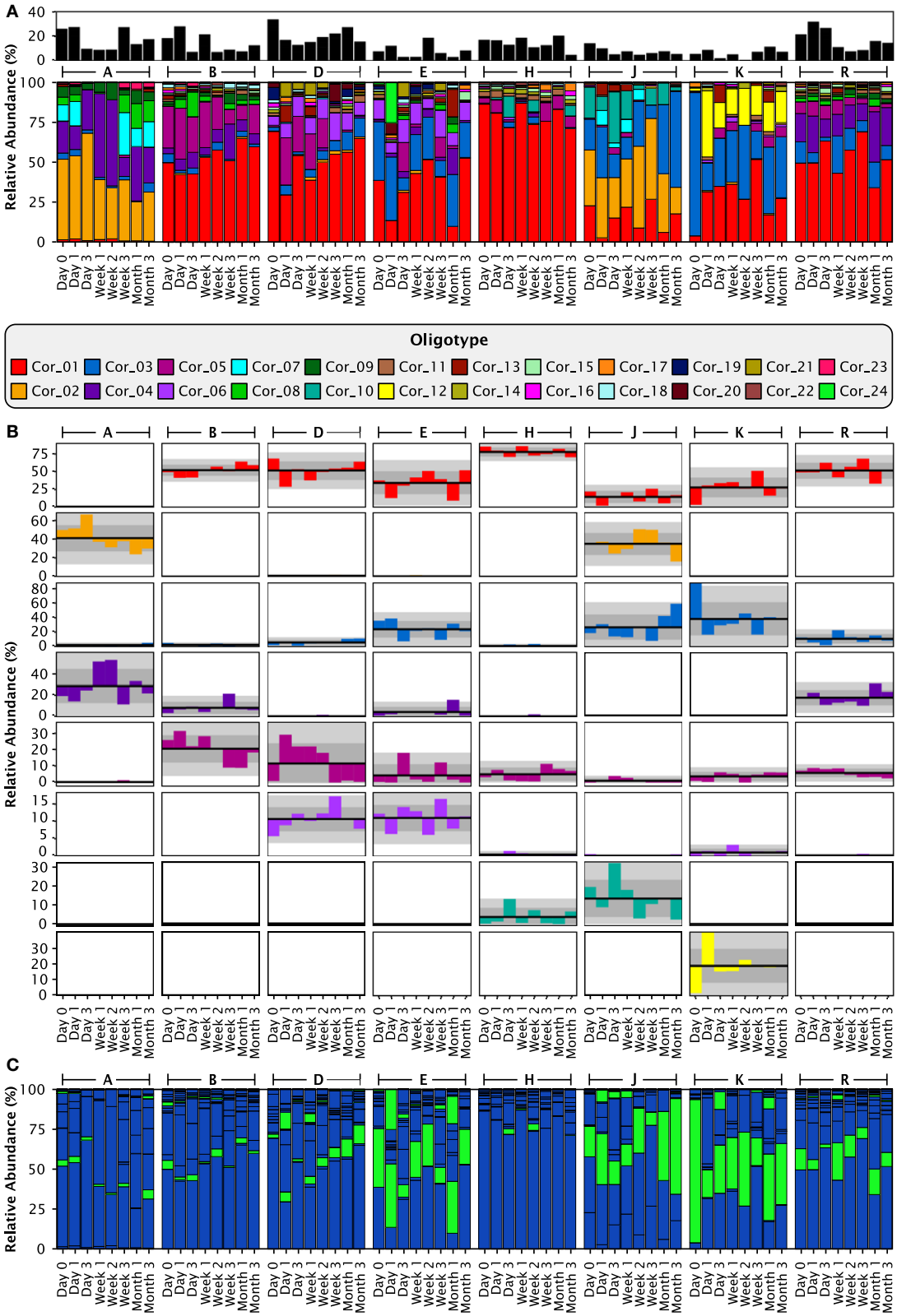


FIGURE 5 | Stable differences between individuals are clear at the oligotype level. (A) Relative abundances of the 24 *Corynebacterium* oligotypes. The smaller, black stackbar shows the *Corynebacterium* abundance relative to all taxa in the sample, while the larger, colored stackbar shows the abundance of each *Corynebacterium* oligotype relative to the total abundance of *Corynebacterium* in each sample. **(B)** Anomaly from the mean relative abundance for each sample from

(Continued)

FIGURE 5 | Continued
each individual. The mean relative abundance for an individual is marked by the dark line, and one and two standard deviations by the dark and light gray fields, respectively. Columns represent individuals and rows represent oligotypes, colored as in **(A)**. **(C)** Exactly the same data and organization as **(A)**, but colored by species instead of by oligotype. *C. matruchotii* oligotypes are colored blue; *C. durum*, green.

major species groups distinguishable by the V4-V5 region were each represented primarily by a single oligotype. In summary, some oligotypes were markers for groups of organisms analogous to species or species groups, whereas other oligotypes provided sub-species level information.

DISCUSSION

Individualized Oligotype Profiles within a Common Framework in Plaque

Our analysis with single-nucleotide resolution of the high-quality Jiang et al. time-series data set showed a common framework of plaque-typical genera and species in each individual, but individual-specific microbiota within this framework. At the oligotype level, individuals were almost entirely distinct from one another, despite the fluctuations within each individual over time. Although the species-level composition of the plaque microbiota is very different from the microbiota inhabiting tongue and saliva (Aas et al., 2005; Eren et al., 2014), nonetheless our time-series results with plaque are in broad agreement with previous high-throughput time-series studies of these other oral sites. Our finding that individuals are more similar to themselves over time than they are to other individuals agrees with the conclusions of Costello et al. (2009), Caporaso et al. (2011), Stahringer et al. (2012) and Cameron et al. (2015) for saliva and tongue. Our demonstration of variability within an individual agrees with the findings of Caporaso et al. (2011), David et al. (2014), Ding and Schloss (2014) and Flores et al. (2014). Thus, our results on plaque are broadly consistent with results on microbial dynamics at other sites in that qualities of both stability and variability are displayed. However, how are these apparently paradoxical qualities to be reconciled?

Looking at the community composition underlying the summary distance metrics, our results showed an oligotype-level "fingerprint" characteristic of each individual, consisting of a set of persistently abundant oligotypes with abundance fluctuating around a stable mean over time. The fluctuation around a stable mean resembles the "stationary dynamics" described by David et al. (2014) for the gut and salivary microbiota of two individuals sampled daily over the course of a year. The drivers of the fluctuations remain unexplained, as does the basis for individual distinctiveness of the overall oligotype composition of plaque.

Clues to the physiological or ecological reasons for both similarity and distinctiveness, however, may be found by considering these time series results in the context of the spatial structure of dental plaque. Recently, we demonstrated the consistent presence in plaque of a "hedgehog" structure, a multi-genus microbial consortium that forms around filamentous corynebacteria (Mark Welch et al., 2016). Interestingly, the taxa that we report here to have relatively constant abundance both within and between individuals—*Corynebacterium*, *Capnocytophaga*, *Fusobacterium*, *Actinomyces*, and *Streptococcus*—are among the major participants in this consortium. *Corynebacterium* forms bush-like (or spiny, hedgehog-like) clusters of filaments, providing the structural framework for the consortium. *Streptococcus* binds in abundance to the distal ends of these filaments, forming an outer shell of "corncob" structures. We hypothesize that this shell of *Streptococcus* alters the local biochemical environment by consuming oxygen and secreting lactate, acetate, carbon dioxide, and peroxide (Ramsey et al., 2011; Zhu and Kreth, 2012). Carbon dioxide-loving taxa such as *Capnocytophaga* and anaerobes or micro-aerophiles such as *Fusobacterium* (Diaz et al., 2000) thrive in positions within the structure that match their metabolic requirements, while *Actinomyces* is frequently found adjacent to the hedgehog structure or intermingled with the *Corynebacterium* filaments at its base. The relatively constant abundance of these taxa in the time series data reported here suggests that there is a limit to the individuality of plaque, in that the consortium taxa are consistently present across individuals.

By contrast, the taxa that were among the more variable in the time series data—*Neisseria* and *Prevotella*—are sporadic participants in the hedgehog structure or are absent from it. Members of the family *Neisseriaceae* were detected sporadically in the hedgehog structure, in or near the aerobic outer shell, while *Prevotella* was generally absent from the hedgehog (Mark Welch et al., 2016). The inter-individual variability of these taxa shown here suggests either functional interchangeability of taxa or different biology of plaque in different individuals. For example, the low abundance of *Streptococcus* in the individuals with unusually high *Neisseria* suggests that in these individuals *Neisseria*, an aerobe, may fill a functional role generally carried out by the facultatively aerobic *Streptococcus*. The wide variation in abundance of the obligate anaerobe *Prevotella*, by contrast, may suggest a difference in plaque physiology between individuals with high-*Prevotella* or low-*Prevotella* communities. Whether such differences might result from host-specific factors (Flores et al., 2014) or chance historical events perpetuated within each host by priority effects is an important topic for further investigation.

Significance of Plaque Microbiota Fingerprints

The physiological or ecological meaning of distinct oligotypes and the overall microbiota fingerprint is likewise an important topic for future study. In our analysis, a 336-nucleotide stretch of the rRNA gene, analyzed with single-nucleotide resolution, acted as a tag for the underlying organism. In the absence of detailed knowledge of the organisms under study, it is difficult to assess how much ecological meaning to assign to these different tags.

However, evidence accumulated over several decades indicates that for some groups of organisms, very small differences in the rRNA sequence represent significant evolutionary distances and divergent ecology. Among the enterobacteria, for example, *E. coli* and several species of *Shigella* and *Salmonella* have 16S rRNA gene sequences that are more than 99% identical (Cilia et al., 1996; Fukushima et al., 2002). The same is true of a number of species within the genus *Bacillus* (Ash et al., 1991; Fox et al., 1992) and, within the oral microbiome, the same is true of the abundant commensal *Streptococcus mitis* and the highly pathogenic *S. pneumoniae* (Denapaite et al., 2010; Kilian et al., 2014). Indeed, enormously important differences in biology and pathogenicity can also occur between strains that are considered members of the same species and have identical or nearly-identical 16S rRNA gene sequences (Böddinghaus et al., 1990; Perna et al., 2001; Jin et al., 2002). These findings suggest that even a single nucleotide difference in the 16S rRNA gene can indicate the presence of significant underlying differences in the genomes and functional roles of organisms.

Alternatively, it is also possible that the different versions of the 16S rRNA gene sequence simply represent population-level variation at a neutral site, and that the organisms possessing one or another of these variants are not physiologically different. The data we present here argue against this possibility. If the organisms represented by these sequences were functionally equivalent, they would be expected to vary in relative abundance in a random walk. The pool of available oligotypes is widely shared; most oligotypes in this analysis were detectable, albeit in low abundance, in most individuals. Yet, in most cases, a random walk did not occur; instead, different oligotypes dominated consistently in different individuals. Thus, the stable oligotype profile within each individual suggests that plaque oligotypes indicate the presence not of neutral variants but of evolutionarily selected, ecologically distinct organisms.

The relationship of oligotypes to species-level groupings is not straightforward. Species-level taxonomy itself is not static, but is continually subject to refinement. *Corynebacterium matruchotii*, for example, is thought to contain cryptic species (Barrett et al., 2001). Nonetheless, species groupings represent current knowledge of the biology of the organisms and can provide meaningful insight. Our analysis revealed some cases in which oligotypes apparently correspond to species-level groups or small clusters of described species, such as within the genus *Streptococcus*. If distinct strains were present with differing gene content and physiology but identical 16S rRNA gene sequences, they were, naturally, indistinguishable by this analysis. In other genera, however, the resolving power of the sequenced region of the 16S rRNA gene is greater (or the scrutiny of the genus by microbiologists has been lesser) and oligotypes identified sub-species-level groups, such as within the genera *Corynebacterium* and *Capnocytophaga*.

Cross-Cultural and Genus-Level Consistency with Individual Variations

Our comparison of plaque 16S rRNA gene sequencing data from China and the United States showed that the plaque microbial community contains the same major genera and spans a similar range of variation in individuals from both cultures. Our results suggest that any systematic differences between Chinese and Western plaque, should they exist, lie in shifts of species composition within abundant genera, or in the presence and abundance of rare genera. In contrast to the lack of ethnic differences we found in the supragingival plaque community, previous studies have reported ethnic distinctions in other oral microbiomes. Mason et al. (2013) studied the saliva and supra- and sub-gingival plaque from four ethnic groups living in America (Chinese, Latino, non-Hispanic whites, and non-Hispanic African Americans) and found significant clustering by ethnicity in sub-gingival and saliva samples but not in supra-gingival plaque. Li et al. (2014) sampled saliva from Africans, Germans, and native Alaskans and found that the African samples differed from the Alaskan and Germans ones by a number of measures, including a high abundance of the genus *Enterobacter*. Takeshita et al. (2014) compared the salivary microbiomes of Koreans and Japanese and found *Neisseria* to be significantly more abundant in Koreans and *Prevotella*, *Fusobacterium*, and *Veillonella* more abundant in Japanese. A study of the oral mucosa of uncontacted Amerindians showed similar overall diversity to the oral microbiomes of developed Americans, but higher proportions of certain taxa including *Prevotella*, *Fusobacterium*, and *Gemella* (Clemente et al., 2015). Park et al. (2015) reported an unusual taxon, the halophilic gamma proteobacterium *Halomonas hamiltonii*, to be abundant in subgingival plaque from healthy Korean volunteers. It may be that the presence of ethnic or cultural signatures varies by oral site with certain locations such as saliva and subgingival plaque being more sensitive to ethnic differences than others. Any such signatures, however, were undetectable in this study of the healthy supragingival plaque community.

Regardless of the question of ethnic signatures, our application of oligotype analysis to supragingival plaque highlights the importance of single-nucleotide analysis of microbial communities. Understanding of the individuality, stability and variability, the habitat and community dynamics, and the physiological or ecological meaning of microbial communities all would be deepened by analysis of sequence data at the highest level of resolution possible.

AUTHOR CONTRIBUTIONS

JMW and GB conceived and designed the work; DU and JMW analyzed the data; and DU, JMW, and GB wrote the paper.

ACKNOWLEDGMENTS

We thank F.E. Dewhirst, The Forsyth Institute for helpful discussions and A.M. Eren, University of Chicago for a critical reading of the manuscript. We thank W. X. Jiang, Y. J. Hu, L. Gao, Z. Y. He, C. L. Zhu, R. Ma, and Z. W. Huang, Shanghai, the researchers who collected the original data set, without whose work our re-analysis would have been

impossible. Our work was supported by National Institutes of Health (NIH) National Institute of Dental and Craniofacial Research Grant DE022586 (to GGB). Additional support was provided by Harvard University's Department of Organismic and Evolutionary Biology graduate program (to DRU).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2016.00564

REFERENCES

Aas, J. A., Paster, B. J., Stokes, L. N., Olsen, I., and Dewhirst, F. E. (2005). Defining the normal bacterial flora of the oral cavity. *J. Clin. Microbiol.* 43, 5721–5732. doi: 10.1128/JCM.43.11.5721-5732.2005

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Ash, C., Farrow, J. A. E., Dorsch, M., Stackebrandt, E., and Collins, M. D. (1991). Comparative analysis of *Bacillus anthracis*, *Bacillus cereus*, and related species on the basis of reverse transcriptase sequencing of 16S rRNA. *Int. J. Syst. Bacteriol.* 41, 343–346. doi: 10.1099/00207713-41-3-343

Barrett, S. L. R., Cookson, B. T., Carlson, L. C., Bernard, K. A., and Coyle, M. B. (2001). Diversity within reference strains of *Corynebacterium matruchotii* includes *Corynebacterium durum* and a novel organism. *J. Clin. Microbiol.* 39, 943–948. doi: 10.1128/JCM.39.3.943-948.2001

Böddinghaus, B., Wolters, J., Heikens, W., and Böttger, E. C. (1990). Phylogenetic analysis and identification of different serovars of *Mycobacterium intracellulare* at the molecular level. *FEMS Microbiol. Lett.* 70, 197–203. doi: 10.1016/S0378-1097(05)80039-4

Brandt, B. W., Bonder, M. J., Huse, S. M., and Zaura, E. (2012). TaxMan: a server to trim rRNA reference databases and inspect taxonomic coverage. *Nucleic Acids Res.* 40, W82–W87. doi: 10.1093/nar/gks418

Cameron, S. J., Huws, S. A., Hegarty, M. J., Smith, D. P., and Mur, L. A. (2015). The human salivary microbiome exhibits temporal stability in bacterial diversity. *FEMS Microbiol. Ecol.* 91:fiv091. doi: 10.1093/femsec/fiv091

Caporaso, J. G., Bittinger, K., Bushman, F. D., DeSantis, T. Z., Andersen, G. L., and Knight, R. (2010). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* 26, 266–267. doi: 10.1093/bioinformatics/btp636

Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., et al. (2011). Moving pictures of the human microbiome. *Genome Biol.* 12:R50. doi: 10.1186/gb-2011-12-5-r50

Cilia, V., Lafay, B., and Christen, R. (1996). Sequence heterogeneities among 16S ribosomal RNA sequences, and their effect on phylogenetic analyses at the species level. *Mol. Biol. Evol.* 13, 451–461. doi: 10.1093/oxfordjournals.molbev.a025606

Clemente, J. C., Pehrsson, E. C., Blaser, M. J., Sandhu, K., Gao, Z., Wang, B., et al. (2015). The microbiome of uncontacted Amerindians. *Sci. Adv.* 1:e1500183. doi: 10.1126/sciadv.1500183

Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I., and Knight, R. (2009). Bacterial community variation in human body habitats across space and time. *Science* 326, 1694–1697. doi: 10.1126/science.1177486

David, L. A., Materna, A. C., Friedman, J., Campos-Baptista, M. I., Blackburn, M. C., Perrotta, A., et al. (2014). Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* 15, 1. doi: 10.1186/gb-2014-15-7-r89

Denapaite, D., Brückner, R., Nuhn, M., Reichmann, P., Henrich, B., Maurer, P., et al. (2010). The genome of *Streptococcus mitis* B6—What is a commensal? *PLoS ONE* 5:e9426. doi: 10.1371/journal.pone.0009426

Dewhirst, F. E., Chen, T., Izard, J., Paster, B. J., Tanner, A. C. R., Yu, W.-H., et al. (2010). The human oral microbiome. *J. Bacteriol.* 192, 5002–5017. doi: 10.1128/JB.00542-10

Diaz, P. I., Zilm, P. S., and Rogers, A. H. (2000). The response to oxidative stress of *Fusobacterium nucleatum* grown in continuous culture. *FEMS Microbiol. Lett.* 187, 31–34. doi: 10.1111/j.1574-6968.2000.tb09132.x

Ding, T., and Schloss, P. D. (2014). Dynamics and associations of microbial community types across the human body. *Nature* 509, 357–360. doi: 10.1038/nature13178

Eren, A. M., Borisy, G. G., Huse, S. M., and Mark Welch, J. L. (2014). Oligotyping analysis of the human oral microbiome. *Proc. Natl. Acad. Sci. U.S.A.* 111, E2875–E2884. doi: 10.1073/pnas.1409644111

Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., et al. (2013). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* 4, 1111–1119. doi: 10.1111/2041-210X.12114

Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H., and Sogin, M. L. (2015). Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J.* 9, 968–979. doi: 10.1038/ismej.2014.195

Faith, J. J., Guruge, J. L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A. L., et al. (2013). The long-term stability of the human gut microbiota. *Science* 341, 1237439–1237439. doi: 10.1126/science.1237439

Flores, G. E., Caporaso, J. G., Henley, J. B., Rideout, J. R., Domogala, D., Chase, J., et al. (2014). Temporal variability is a personalized feature of the human microbiome. *Genome Biol.* 15, 804. doi: 10.1186/s13059-014-0531-y

Fox, G. E., Wisotzkey, J. D., and Jurtshuk, P. Jr. (1992). How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int. J. Syst. Evol. Microbiol.* 42, 166–170. doi: 10.1099/00207713-42-1-166

Fukushima, M., Kakinuma, K., and Kawaguchi, R. (2002). Phylogenetic analysis of *Salmonella*, *Shigella*, and *Escherichia coli* strains on the basis of the gyrB gene sequence. *J. Clin. Microbiol.* 40, 2779–2785. doi: 10.1128/JCM.40.8.2779-2785.2002

Gajer, P., Brotman, R. M., Bai, G., Sakamoto, J., Schütte, U. M. E., Zhong, X., et al. (2012). Temporal dynamics of the human vaginal microbiota. *Sci. Transl. Med.* 4, 132ra52–132ra52. doi: 10.1126/scitranslmed.3003605

Huse, S. M., Dethlefsen, L.es, Huber, J. A., Mark Welch, D., Relman, D. A., and Sogin, M. L. (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet.* 4:e1000255. doi:10.1371/journal.pgen.1000255

Jiang, W.-X., Hu, Y.-J., Gao, L., He, Z.-Y., Zhu, C.-L., Ma, R., et al. (2015). The impact of various time intervals on the supragingival plaque dynamic core microbiome. *PLoS ONE* 10:e0124631. doi: 10.1371/journal.pone.0124631

Jin, Q., Yuan, Z., Xu, J., Wang, Y., Shen, Y., Lu, W., et al. (2002). Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res.* 30, 4432–4441. doi: 10.1093/nar/gkf566

Kilian, M., Riley, D. R., Jensen, A., Brüggemann, H., and Tettelin, H. (2014). Parallel evolution of *Streptococcus pneumoniae* and *Streptococcus mitis* to pathogenic and mutualistic lifestyles. *mBio* 5, e01490–14–e01490–14. doi: 10.1128/mbio.01490-14

Li, J., Quinque, D., Horz, H.-P., Li, M., Rzhetskaya, M., Raff, J. A., et al. (2014). Comparative analysis of the human saliva microbiome from different climate zones: Alaska, Germany, and Africa. *BMC Microbiol.* 14:316. doi: 10.1186/s12866-014-0316-1

Lozupone, C., and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* 71, 8228–8235. doi: 10.1128/AEM.71.12.8228-8235.2005

Mark Welch, J. L., Rossetti, B. J., Rieken, C. W., Dewhirst, F. E., and Borisy, G. G. (2016). Biogeography of a human oral microbiome at the micron scale. *Proc. Natl. Acad. Sci. U.S.A.* 113, E791–E800. doi: 10.1073/pnas.1522149113

Mark Welch, J. L., Utter, D. R., Rossetti, B. J., Mark Welch, D. B., Eren, A. M., and Borisy, G. G. (2014). Dynamics of tongue microbial communities with single-nucleotide resolution using oligotyping. *Front. Microbiol.* 5:568. doi: 10.3389/fmicb.2014.00568

Mason, M. R., Nagaraja, H. N., Camerlengo, T., Joshi, V., and Kumar, P. S. (2013). Deep sequencing identifies ethnicity-specific bacterial signatures in the oral microbiome. *PLoS ONE* 8:e77287. doi: 10.1371/journal.pone.0077287

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., et al. (2011). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* 6, 610–618. doi: 10.1038/ismej.2011.139

Nasidze, I., Li, J., Quinque, D., Tang, K., and Stoneking, M. (2009). Global diversity in the human salivary microbiome. *Genome Res.* 19, 636–643. doi: 10.1101/gr.084616.108

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., et al. (2013). *Vegan: Community Ecology Package. Package version 1.0.1*. Available online at: http://cran.r-project.org/web/packages/vegan/index.html. (Accessed July 10, 2015).

Park, O. J., Yi, H., Jeon, J. H., Kang, S. S., Koo, K. T., Kum, K. Y., et al. (2015). Pyrosequencing analysis of subgingival microbiota in distinct periodontal conditions. *J. Dental Res.* 94, 921–927. doi: 10.1177/0022034515583531

Perna, N. T., Plunkett, G. III, Burland, V., Mau, B., Glasner, J. D., Rose, D. J., et al. (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* 409, 529–533. doi: 10.1038/35054089

R Core Team. (2015). *R: A Language and Environment for Statistical Computing*. Vienna: Foundation for Statistical Computing.

Ramsey, M. M., Rumbaugh, K. P., and Whiteley, M. (2011). Metabolite cross-feeding enhances virulence in a model polymicrobial infection. *PLoS Pathog.* 7:e1002012. doi: 10.1371/journal.ppat.1002012

Segata, N., Haake, S., Mannon, P., Lemon, K. P., Waldron, L., Gevers, D., et al. (2012). Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol.* 13:R42. doi: 10.1186/gb-2012-13-6-r42

Stahringer, S. S., Clemente, J. C., Corley, R. P., Hewitt, J., Knights, D., Walters, W. A., et al. (2012). Nurture trumps nature in a longitudinal survey of salivary bacterial communities in twins from early adolescence to early adulthood. *Genome Res.* 22, 2146–2152. doi: 10.1101/gr.140608.112

Takeshita, T., Matsuo, K., Furuta, M., Shibata, Y., Fukami, K., Shimazaki, Y., et al. (2014). Distinct composition of the oral indigenous microbiota in South Korean and Japanese adults. *Sci. Rep.* 4:6990. doi: 10.1038/srep06990

The Human Microbiome Project Consortium. (2012). A framework for human microbiome research. *Nature* 486, 215–221. doi: 10.1038/nature11209

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer.

Xu, X., He, J., Xue, J., Wang, Y., Li, K., Zhang, K., et al. (2014). Oral cavity contains distinct niches with dynamic microbial communities. *Environ. Microbiol.* 17, 699–710. doi: 10.1111/1462-2920.12502

Zaura, E., Keijser, B. J., Huse, S. M., and Crielaard, W. (2009). Defining the healthy "core microbiome" of oral microbial communities. *BMC Microbiol.* 9:259. doi: 10.1186/1471-2180-9-259

Zhu, L., and Kreth, J. (2012). The role of hydrogen peroxide in environmental adaptation of oral microbial communities. *Oxid. Med. Cell. Longev.* 2012:717843. doi: 10.1155/2012/717843

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 29 September 2015
doi: 10.3389/fmicb.2015.01028

A unique assemblage of cosmopolitan freshwater bacteria and higher community diversity differentiate an urbanized estuary from oligotrophic Lake Michigan

*Ryan J. Newton * and Sandra L. McLellan*

School of Freshwater Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

OPEN ACCESS

Edited by:
Lois Maignien, Université Bretagne Occidentale, France

Reviewed by:
Peter Bergholz, North Dakota State University, USA
Steven Singer, Lawrence Berkeley National Laboratory, USA

***Correspondence:**
Ryan J. Newton, School of Freshwater Sciences, University of Wisconsin-Milwaukee, Great Lakes Research Facility, 600 E. Greenfield Ave., Milwaukee, WI 53204, USA
newtonr@uwm.edu

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *26 June 2015*
Accepted: *10 September 2015*
Published: *29 September 2015*

Citation:
Newton RJ and McLellan SL (2015) A unique assemblage of cosmopolitan freshwater bacteria and higher community diversity differentiate an urbanized estuary from oligotrophic Lake Michigan. Front. Microbiol. 6:1028. doi: 10.3389/fmicb.2015.01028

Water quality is impacted significantly by urbanization. The delivery of increased nutrient loads to waterways is a primary characteristic of this land use change. Despite the recognized effects of nutrient loading on aquatic systems, the influence of urbanization on the bacterial community composition of these systems is not understood. We used massively-parallel sequencing of bacterial 16S rRNA genes to examine the bacterial assemblages in transect samples spanning the heavily urbanized estuary of Milwaukee, WI to the relatively un-impacted waters of Lake Michigan. With this approach, we found that genera and lineages common to freshwater lake epilimnia were common and abundant in both the high nutrient, urban-impacted waterways, and the low nutrient Lake Michigan. Although the two environments harbored many taxa in common, we identified a significant change in the community assemblage across the urban-influence gradient, and three distinct community features drove this change. First, we found the urban-influenced waterways harbored significantly greater bacterial richness and diversity than Lake Michigan (i.e., taxa augmentation). Second, we identified a shift in the relative abundance among common freshwater lineages, where acI, acTH1, *Algoriphagus* and LD12, had decreased representation and *Limnohabitans*, *Polynucleobacter*, and *Rhodobacter* had increased representation in the urban estuary. Third, by oligotyping 18 common freshwater genera/lineages, we found that oligotypes (highly resolved sequence clusters) within many of these genera/lineages had opposite preferences for the two environments. With these data, we suggest many of the defined cosmopolitan freshwater genera/lineages contain both oligotroph and more copiotroph species or populations, promoting the idea that within-genus lifestyle specialization, in addition to shifts in the dominance among core taxa and taxa augmentation, drive bacterial community change in urbanized waters.

Keywords: bacterial community, freshwater, urban ecology, Lake Michigan, oligotyping, bacterioplankton

Introduction

As a result of continued urbanization worldwide and its contribution to deteriorating ecosystem services (Corvalan et al., 2005), the relationship between urban development, biodiversity patterns, and ecosystem dynamics has been the focus of increasing research attention and theoretical development (Grimm et al., 2000; Alberti, 2005; Pickett et al., 2011). In aquatic ecosystems, urbanization alters watershed ecosystem functioning through the movement, magnitude, and content of surface water runoff (Allan, 2004; Alberti et al., 2007; Hale et al., 2015). As a major component of aquatic biological communities, bacteria are critical drivers of energy flow and nutrient recycling (Cotner and Biddanda, 2002), yet we know relatively little about bacterial biodiversity patterns in urban-influenced waterways, whether there are important differences in the bacterial assemblages between urbanized and non-urbanized systems, or whether urban-influenced aquatic environments promote the persistence of organisms that impact human health or well-being (Paerl et al., 2003; Newton et al., 2013; King, 2014).

The effects of urban landscape modification can account for much of the water quality deterioration in urbanized waterways (Brabec et al., 2002), which characteristically have high solute (Booth and Jackson, 1997; Kaushal and Belt, 2012) and nutrient (Carpenter et al., 1998; Wollheim et al., 2005; Hale et al., 2015) loads and high productivity (Correll, 1998). Both the total productivity and the heterogeneity in nutrient resources play a prominent role in structuring species co-existence patterns across all scales of life (Mittelbach et al., 2001; Chase and Leibold, 2002; Jankowski et al., 2014). However, the mechanisms driving these compositional changes in response to increased ecosystem productivity are complex and at minimum depend upon the total resource pool, the balance of resources within this pool, and the richness of competing species for specific resources (Cardinale et al., 2009). Since urbanization results in increased delivery of nutrients to surface waters (Carpenter et al., 1998; Paul and Meyer, 2001), high nutrient concentration is likely one driver of changes in the bacterial assemblage in these systems. For this reason, the patterns of bacterial community assembly across an urbanization gradient may in large part mirror those observed across trophy or primary productivity gradients.

Increased productivity or nutrient load has been shown to relate to changes in the diversity and composition of bacterial communities in freshwater ecosystems (Horner-Devine et al., 2003; Yannarell and Triplett, 2004; Longmuir et al., 2007; Smith, 2007; Kolmonen et al., 2011; Jankowski et al., 2014). Yet a clear relationship between productivity and bacterial diversity or community change has not been identified consistently. For example, bacterial richness was uncoupled to total phosphorus concentration in 100 lakes in Finland (Korhonen et al., 2011) and productivity related variables were not strong predictors of community composition across 30 lakes in Wisconsin, USA when geographic and landscape related variables were considered (Yannarell and Triplett, 2005). Also, several processes have been implicated in driving bacterial community change across aquatic environmental gradients, including: complete community displacement or turnover (Bell et al., 2010), changes in the relative abundance of a few core taxa (Shade et al., 2010), and an increase in the presence of rare or novel taxa that augment a core community (Jankowski et al., 2014; Shade et al., 2014). These varied and sometimes contradictory findings suggest that the relationship between microbial community structure and ecosystem productivity are complex and still poorly defined.

Few studies that examined explicitly the relationship of system productivity and bacterial community change also identified the bacterial types causing the observed change. In one such study, an increased representation of rare and/or novel taxa in more eutrophic conditions were implicated as being responsible for much of the observed community change, but the taxonomic affiliation of these taxa were not considered (Jankowski et al., 2014). Studies involving the distribution and growth traits of common lake taxa have provided some insight into which taxa would be expected to drive changes across productivity/trophy gradients. Specifically, members of the genus *Limnohabitans* and *Flavobacterium* exhibited high maximum growth rates and abundance correlations to high nutrient conditions in lakes (Šimek et al., 2006; Newton et al., 2011a; Neuenschwander et al., 2015), while the freshwater lineages LD12 and acI have slower growth rates and traits indicating a more oligotrophic lifestyle (Šimek et al., 2006; Newton et al., 2011a; Salcher et al., 2011b; Ghylin et al., 2014).

Using an analysis of bacterial community composition along sample transects from the highly urbanized waterways in the Milwaukee estuary to the relatively low urban-impacted waters of Lake Michigan, we assess how the bacterial assemblage differs between these two connected environments. Specifically we evaluate whether processes identified as driving microbial community change in aquatic systems, such as complete community turnover, shifts in the community contribution of common taxa, or taxa augmentation also drive changes in the richness and composition of bacteria across an urbanization gradient. With these data we also identify the taxa responsible for differences in the community assemblages across the urban-influence gradient and evaluate whether there are differential distribution patterns for narrowly-defined sequence-based groups (oligotypes) within several ubiquitous freshwater genera/lineages.

Materials and Methods

Sample Collection and Site Characteristics

All samples analyzed for bacterial community composition were collected from surface waters (0–0.5 m depth) during the ice-off season (April to October) in the waterways of Milwaukee, WI or in Lake Michigan. Each final sample consisted of three surface water samples that were combined, mixed, and subsampled into 1- to 4-l bottles. The samples were collected on 15 separate expeditions spanning the years 2008–2012. See **Figure 1** for a sample map of the collection locations and Supplementary Table 1 for sample metadata. Samples collected in 2008–2010 were described previously (Newton et al., 2013). Sample processing and filtering methods are described in Newton et al. (2011b).

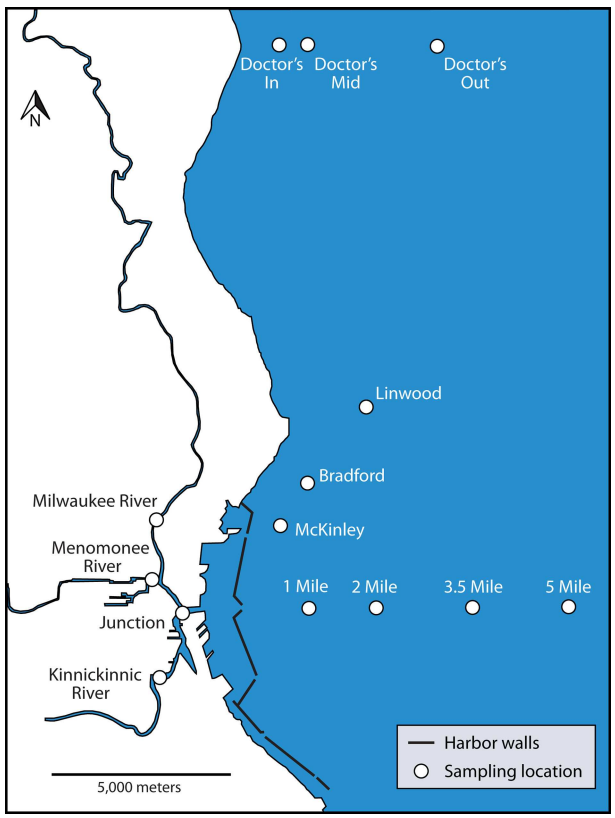


FIGURE 1 | Map of Milwaukee, WI, USA urban estuary and nearshore Lake Michigan. Sampling locations included in this study are indicated with site names.

We characterized the average or "typical" chemical and physical conditions of the waterways using data from the Milwaukee Metropolitan Sewerage District Water Quality Monitoring program housed via the WATERBase database at the University of Wisconsin-Milwaukee (www.waterbase.glwi.uwm.edu/). From these data, we retrieved surface water sample measurements collected on 19 occasions for Lake Michigan and on 31 occasions for the rivers and inner harbor. These samples were limited to the months of June through October for the years 2008–2010, which represents a similar seasonal period and most of the years during which the bacterial community water samples were obtained. Three sample sites (2 mile, Linwood, and Doctors Out) were used to represent Lake Michigan and one sample site each was used to represent each of the rivers and the inner harbor (see **Figure 1** for sample locations). Data was obtained for water temperature, pH, conductivity, suspended solids, total phosphorus, Total Kjeldahl Nitrogen, nitrate/nitrite, and chlorophyll *a* according to the standard protocols listed in the Standard Methods for the Examination of Water and Wastewater (20th ed., 1998)[1]. The median and range for each environmental parameter at each sample site are listed in **Table 1**.

Based on the environmental parameters representing each area and the connection between each waterway to the urban landscape, we grouped the sample locations into two categories: (1) urban-impacted and (2) Lake Michigan, respectively representing high and low impact from urban discharge. The urban-impacted category includes the three rivers and inner harbor samples and the Lake Michigan category includes all samples outside of the harbor break walls (see **Figure 1** for sample locations).

16S rRNA Gene Sequencing and Processing

DNA extraction procedures for all filtered water samples are detailed in Newton et al. (2013). Extracted DNA was used to construct amplicon libraries for high-throughput 16S rRNA gene sequencing targeting either the V6 or V4 to V6 regions (amplified in the reverse direction V6 to V4). Amplicon libraries were sequenced using either the 454 Life Sciences or the illumina® platform. Details for amplicon library construction, sequencing procedures, and post-sequencing quality control methods for the V6 454 platform are described in McLellan et al. (2010), for the V6V4 454 platform in Newton et al. (2013), and for the illumina® V6 platform in Eren et al. (2013b). Sequencing methods for each sample are listed in Supplementary Table 1.

The National Center for Biotechnology Information Sequence Read Archive has archived the raw data under SRA Projects SRP018584 (V6 454), SRP059202 (V6V4 454), and SRP056973 (V6 illumina). Trimmed and quality filtered sequence data are publicly available from the Visualization and Analysis of Microbial Population Structures website (VAMPS; http://vamps.mbl.edu; Huse et al., 2014) under project names SLM_SWG_Bv6, SLM_NIH_Bv6v4, and SLM_NIH2_Bv6.

Dataset Construction

We used the algorithm Global Alignment for Sequence Taxonomy (GAST; Huse et al., 2008) to assign taxonomy to all sequences. A dataset consisting of sequences binned by the most resolved taxonomic assignment down to genus was used in whole community composition comparisons among samples. Analyses using this dataset are termed "taxon-based." We also constructed a second, higher resolution dataset based on closed-reference clustering, where reads are searched against the curated SILVA database (Pruesse et al., 2007) as part of the Visualization and Analysis of Microbial Population Structures (VAMPS; http://vamps.mbl.edu) database (Huse et al., 2014) and then clustered as defined by the best database match for each read (see Huse et al., 2008 for more details). Since reference sequence matches are not identical across sequence regions (V6 vs. V6V4 data), but reference-based clustering provides more narrowly-defined groupings than taxon-based assignments, and therefore a more accurate representation of total bacterial diversity, this dataset was used only for richness and diversity comparisons. Analyses using this dataset are termed "reference-based."

We constructed a third, high-resolution dataset to explore distribution patterns within and among common freshwater

[1]Standard Methods for the Examination of Water and Wastewater 20 th ed., 1998. American Public Health Association (APHA), 1015 15th Street, NW, Washington, DC 20005.

TABLE 1 | Chemical and physical properties of sampled environments[a,b,c].

	Urban estuary				Lake Michigan		
	Junction	MKE	MN	KK	2 mile	Linwood	Doctors park out
Temperature (°C)	19.4 (11.8–24.6)	21.6 (7.5–26.8)	22.1 (12.7–29.1)	19.3 (12.3–25.0)	17.0 (10.4–21.7)	17.5 (8.6–21.5)	16.4 (6.9–22.9)
pH	8.0 (7.5–8.5)	8.2 (7.7–8.6)	7.8 (7.4–8.3)	8.0 (7.6–8.4)	8.4 (8.0–8.7)	8.4 (8.2–8.6)	8.4 (8.2–8.6)
Conductivity (μS/cm)	587 (257–799)	807 (210–896)	700 (336–997)	615 (315–899)	294 (279–341)	285 (274–305)	285 (275–292)
Susp. solids (mg l^{-1})	6 (4–80)	12 (4–100)	8 (3–170)	10 (6–140)	bd (bd–bd)	bd (bd–bd)	bd (bd–bd)
Total P ($\mu g l^{-1}$)	68 (bd–230)	115 (44–290)	100 (bd–280)	82 (42–260)	bd (bd–72)	bd (bd–bd)	bd (bd–25)
TKN (mg l^{-1})	0.60 (bd–1.10)	0.72 (bd–1.60)	0.68 (bd–1.60)	0.62 (bd–1.40)	bd (bd–0.62)	bd (bd–0.83)	bd (bd–0.84)
NO_3/NO_2 (mg l^{-1})	0.71 (0.38–1.10)	0.90 (0.24–1.40)	0.58 (0.27–1.10)	0.71 (0.35–1.20)	0.27 (bd–0.44)	0.26 (bd–0.30)	0.29 (bd–0.31)
Chlorophyll *a* ($\mu g l^{-1}$)	6.0 (0.9–17.8)	6.4 (3.3–34.8)	6.7 (2.1–52.9)	6.0 (1.8–27)	1.2 (0.3–7.2)	0.8 (bd–7.5)	0.4 (0.2–1.8)

[a] *The Median (Range) are listed for each water chemical/physical property measurement.*

[b] *Abbreviations: Susp. Solids, Suspended Solids; Total P, Total Phosphorus; TKN, Total Kjeldahl Nitrogen; bd, below detection.*

[c] *The detection limits are as follows: Suspended Solids 1 mg l^{-1}, Total Phosphorus 20 μg l^{-1}, Total Kjeldahl Nitrogen 0.36 mg l^{-1}, Nitrate/Nitrite 0.20 mg l^{-1}, Chlorophyll a 0.11 μg l^{-1}.*

genera/lineages. This dataset consisted only of amplicons assigned by GAST to the *Actinobacteria* family *Sporichthyaceae* and genus *Aquiluna*, the *Bacteroidetes* genera *Algoriphagus*, *Arcicella*, *Flavobacterium*, *Fluviicola*, and *Sediminibacterium*, the *Proteobacteria* lineage SAR11 and genera *Hydrogenophaga*, *Polynucleobacter*, *Rhodobacter*, *Rhodoferax*, *Sphingopyxis*, and the *Verrucomicrobia* genus *Luteolibacter*. All amplicons assigned to these 14 common freshwater groups were aligned (within-group alignments) using the align.seqs command in mothur (Schloss et al., 2009). After alignment, the non-overlapping sequence from the V6V4 amplicons was trimmed from the 14 alignments using the filter.seqs command in mothur (Schloss et al., 2009). We then conducted a high-resolution oligotyping analysis on the trimmed alignments as described previously (Eren et al., 2013a; oligotyping.org). Oligotyping is a supervised computational method that uses Shannon entropy calculations to identify nucleotide variation in alignments. The entropy calculations are used to select highly variable positions in the alignment, which are then used to parse the data into groups having identical sequences at the defined positions. These highly-resolved groups are known as oligotypes (Eren et al., 2013a). We set the minimum substantive abundance criterion (M) to the lesser of 0.01% of all sequences assigned to each group or 10 and the minimum sample prevalence (s) to 2 for all 14 oligotyping analyses. Oligotypes were deemed to have converged when entropy values within each oligotype were below 0.2 according to the procedures described in Eren et al. (2013a).

For the family *Sporicthyaceae*, reference sequences from each oligotype were compared against the freshwater database from Newton et al. (2011a) to assign a more refined freshwater naming structure. *Sporichthyaceae* oligotypes were resolved to the lineages acI-A, acI-B, acI-C, acSTL, and acTH1 when the representative sequence was identical to or contained a single mismatch to sequences representing only one of the lineages. After splitting the *Sporichthyaceae* into five distinct lineages, our final oligotyping dataset consisted of 18 unique lineages that were used in subsequent analyses. For *Rhodoferax*, reference sequences for each oligotype were also compared against the Newton et al. (2011a) freshwater database and only those sequences identical to or with a single mismatch to sequences representing the *Limnohabitans* lineage were retained. The SAR11 GAST assigned sequences, throughout are referred to as LD12, the freshwater lineage to which these sequences belong.

Data on the distribution of freshwater taxa generated from clone library sequence data as reported in Newton et al. (2011a) were used in a community composition comparative analysis. These data include the relative abundance of common freshwater genera and lineages from the epilimnion of 47 lakes located primarily in North America and Europe, but also including Antarctica, Africa, and China. This database included only studies with data generated from universal bacterial primers and random clone selection for sequencing and for which more than 40 sequences were present (see Newton et al., 2011a for further dataset details).

Statistical Analyses

We conducted all data analyses in the R statistical language (R Core Team, 2013). We used the community analysis package *vegan* (Oksanen et al., 2013) and the Bray-Curtis dissimilarity metric for all community composition comparisons. Non-metric multidimensional scaling (NMDS) and hierarchical clustering were based on Bray-Curtis dissimilarities using the relative abundance of taxon- or reference-based groups, calculated as the sequence count for a group divided by the total sequence counts for a sample (whole community) or the sequence counts for a subset of taxa/lineages from a sample (e.g., common freshwater genera/lineages only). To identify the number of dimensions to include in NMDS analyses, a scree plot was used to identify dimensional convergence for ordination stress and a low dimension analysis ($k = 2$) was compared to a higher dimensionality analysis ($k = 10$) for significant ordination correlation using a Procrustes rotation via the *protest* function. Analysis of Similarity (ANOSIM) statistics (999 permutations) were carried out with the *anosim* function (Oksanen et al., 2013) and were used to test the significance of *a priori* assigned

sample group differentiation. We used the Mann-Whitney *U*-test to examine whether the distribution of measurements for two groups differed significantly (Mann and Whitney, 1947). For most data visualization we used the *ggplot2* R package (Wickham, 2009) or base graphics in R. We constructed heatmaps with the *heatmap.2* function in the *gplots* R package (Warnes et al., 2013).

We used two measures of diversity, inverse Simpson index (Lande, 1996) and the tail statistic (Li et al., 2012) to compare among sample groups. These two metrics differ in their weighting of abundant vs. rare members in a sample. The inverse Simpson metric places more emphasis on the diversity of the most abundant taxa/groups among samples, while the tail statistic places more emphasis on the diversity of more rare community members (Li et al., 2012). We carried out inverse Simpson diversity calculations using the *diversity* function in the *vegan* package (Oksanen et al., 2013) and the tail statistic according to the equation developed by Li et al. (2012).

For all richness and diversity calculations and data comparisons using oligotypes of the common freshwater genera/lineages, we used a subsampled dataset to reduce the artifacts of disproportionate sequencing depth when using non-relativized data. We subsampled randomly once all samples having >30,000 quality-filtered sequences to 30,000 sequences using the R package *plyr* (Wickham, 2011; see Supplementary Table 1 for sequence read counts after subsampling).

To compare the magnitude of a "habitat preference" between the urban estuary waters and Lake Michigan for common freshwater genera/lineages, we used the ratio of the mean relative abundance of each genus/lineage among the urban estuary samples vs. its mean relative abundance in the Lake Michigan samples. To minimize the effect caused by differences in the proportion that these common freshwater bacteria make up in each sample, each genus/lineage relative abundance was calculated as the proportion of sequences in the high-resolution dataset of 18 common freshwater genera/lineages. To minimize the impact of temporal abundance variability for an individual genus/lineage, the relative abundance of each genus/lineage was normalized to the sample with the highest relative abundance within each sample transect.

To identify individual oligotypes that preferentially associated with either the urban-influenced waterways or Lake Michigan, we performed a multinomial species classification using the CLAM test (Chazdon et al., 2011) in the *vegan* R package (Oksanen et al., 2013). This model allowed us to divide oligotypes into the following four categories based on their distribution among samples: oligotypes preferential to urban-influenced waterways, oligotypes preferential to Lake Michigan, oligotypes showing no preferential distribution (generalists), and oligotypes that were too rare to classify with confidence. The CLAM test was performed on the subsampled dataset using an alpha value of 0.01 divided by the total number of oligotypes ($n = 351$), a coverage limit of 30, and a specialization threshold of 0.75. A specialization threshold =0.67 (a supermajority) is considered conservative (Chazdon et al., 2011).

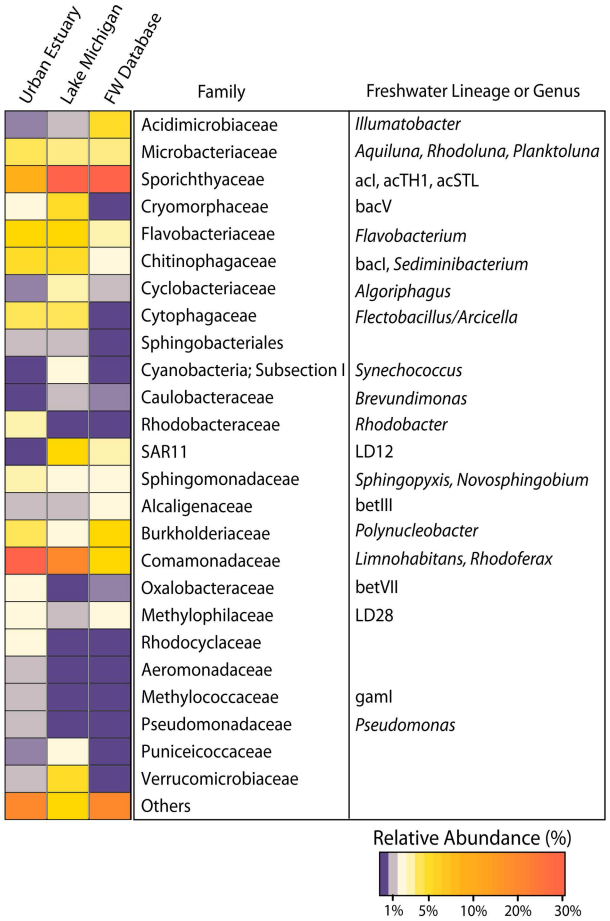


FIGURE 2 | Heatmap indicating the relative abundance of all bacterial families with a mean relative abundance of ≥1% among either all Urban Estuary or all Lake Michigan samples. Common freshwater lineages as described in Newton et al. (2011a) are indicated with their respective family assignments. A previously compiled freshwater dataset is also depicted (FW Database) and consists of bacterial group distributions inferred from whole community 16S rRNA gene amplification and clone library construction across 47 lakes as described in Newton et al. (2011a).

Results

The Lake Michigan Bacterial Community Resembles Other Freshwater Lakes but Differs from Milwaukee's Urban-impacted Waterways

The surface water community in relatively nearshore (<10 km from shore) Lake Michigan is dominated by many of the freshwater bacterial genera and lineages that are common to the surface waters of smaller freshwater lakes (**Figure 2**). On average, the bacterial families in Lake Michigan with the highest number of assigned sequence reads were *Sporichthyaceae,* (28%; freshwater lineages acI, acTH1, and acSTL), *Comamonadaceae,* (13%; freshwater genera *Limnohabitans and Rhodoferax*), *Flavobacteriaceae* (8%; freshwater genera *Flavobacterium*), SAR11 (7%; freshwater lineage LD12), and *Verrucomicrobiaceae* (5%). The families *Sporichthyaceae, Comamonadaceae, and*

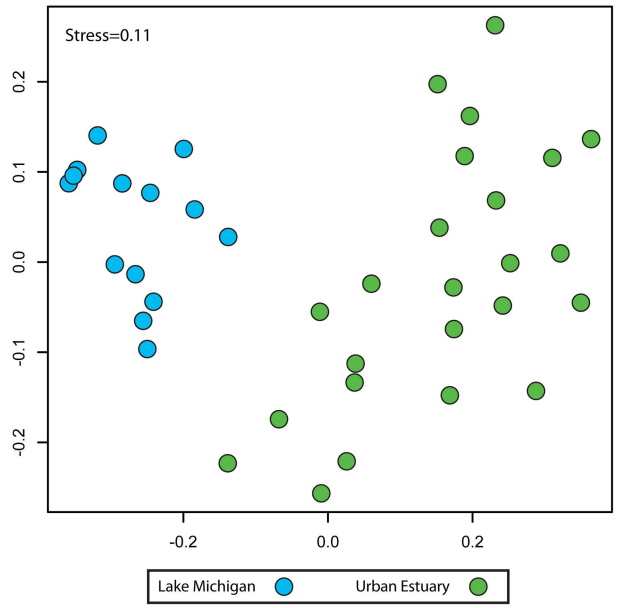


FIGURE 3 | Non-metric multidimensional scaling plot indicating the community composition relationships (Bray-Curtis similarity) between the Urban Estuary (green) and Lake Michigan (blue) samples. Community composition is based on the grouping of sequences by taxonomic assignment to genus and compiled as the relative contribution of each taxon to the community.

Flavobacteriaceae were the only bacterial families that averaged ≥5% of the reads in samples from the urban-impacted waterways. In addition to these common freshwater lineages, the urban impacted waterways also harbored other bacterial families at relatively high abundances (each at ≥2% of the community) that were not common in Lake Michigan, namely, *Oxalobacteraceae* (freshwater lineage betVII), *Rhodocyclaceae*, and *Rhodobacteraceae* (freshwater genera *Rhodobacter*).

NMDS analysis of sequence data binned by taxonomic assignment to genus (taxon-based) indicated the urban-impacted water (rivers and inner harbor) communities were distinct from the bacterial communities of Lake Michigan (**Figure 3**; urban-impacted vs. Lake Michigan; ANOISM $R = 0.80$ $p = 0.001$). Since three different sequencing region/platform combinations were used to create these data, we examined whether this community composition pattern was influenced by the sequencing procedures used (see Supplementary Table 1 for sample details). We found there was a significant, but small proportion of the community variation explained by sequencing procedure (ANOSIM $R = 0.15$, $p = 0.009$), and this variation was distinct from and much smaller than the variation separating the urban-water and Lake Michigan communities (Supplementary Figure 1). Two dimensions were used in the final NMDS ordination calculation, as ordination stress was relatively low (0.11) and additional dimensions did not alter the sample relationship patterns observed (Procrustes test for ordination similarity between $k = 2$ and $k = 10$; $r = 0.801$, $p = 0.001$).

Taxa Augmentation in Urban Waterways

The microbial communities present in the urban waters had higher taxonomic (taxon-based, binned by genus assignment) and reference-based (binned by reference sequence) richness than the communities from Lake Michigan (**Table 2**). The urban water communities also contained higher alpha-diversity levels than the Lake Michigan communities, and this diversity increase was observed with both the inverse Simpson index (reference-based analysis) and the tail statistic (taxon- and reference-based; **Table 2**). Only the taxon-based diversity comparison, using the inverse Simpson test, showed no significant difference between the urban-impacted water communities and Lake Michigan ($p > 0.01$; **Table 2**).

Most of the identified taxa in Lake Michigan were also detected in the urban-impacted waters. For example, of the 1458 taxa identified in at least two samples, only one was present solely in Lake Michigan, while 397 were present solely in the urban-impacted waterway samples. However, these 397 urban-water associated taxa did not typically comprise a large part of the community, contributing on average only 0.14% of the sequence reads in the urban-waterway samples. Together these data indicate an increased distinction between the urban waterways and Lake Michigan as the grouping method used to identify organisms becomes more narrow (from taxon- to reference-based) and as the diversity index puts more weight on more rare organisms (from inverse Simpson to Tail), suggesting a higher number of more closely related (within-genus), but relatively rare organisms in the urban waterways.

Common Freshwater Taxa Exhibit Differential Preference for Urban-impacted vs. Lake Michigan Waters

After examining the whole bacterial community composition differences between the urban-impacted and Lake Michigan waters, we further examined the distribution of 18 common freshwater genera/lineages across four sampling transects. The 18 genera/lineages included: acI-A, acI-B, acI-C, acTH1, acSTL, *Aquiluna*, *Algoriphagus*, *Arcicella*, *Flavobacterium*, *Fluviicola*, *Sediminibacterium*, LD12, *Sphingopyxis*, *Hydrogenophaga*, *Limnohabitans*, *Polynucleobacter*, *Rhodobacter*, and *Luteolibacter*. All genera/lineages were present in all samples ($n = 23$) except for LD12 (22/23) and acI-C (21/23). These 18 genera/lineages comprised on average 44.9 ± 6.9% of the sequence reads in the urban water communities and on average 69.3 ± 2.6% of the Lake Michigan communities.

The relative abundance of the 18 common freshwater lake bacteria genera/lineages (calculated as relative to each other) indicated differential distributions in the urban-impacted waters vs. Lake Michigan (ANOSIM $R = 0.65$ $p < 0.001$), suggesting some common lineages were favored over the others by the conditions present in each environment. We explored whether individual genera/lineages exhibited a "preference," defined as an increased average relative abundance vs. the other common genera/lineages, for either the urban impacted or non-impacted Lake Michigan waters. We found that some genera/lineages were favored by the conditions present in the urban waterways,

TABLE 2 | Diversity comparison between Urban estuary and Lake Michigan samples[a].

	Taxon—Whole community			Reference sequence—Whole community		
Sample environment	**Richness**	**Inverse simpson**	**Tail**	**Richness**	**Inverse simpson**	**Tail**
Urban estuary	432 ± 104	7 ± 6	49 ± 23	2680 ± 1232	68 ± 40	584 ± 463
Lake Michigan	185 ± 25	5 ± 2	16 ± 3	1015 ± 440	37 ± 14	145 ± 85
Mann-Whitney U	400**	277*	400**	379**	325**	370**

[a]*Mean and standard deviation are reported.*
***Indicates significance at $p < 0.01$.*
**Indicates significance at $p < 0.05$.*

while others were favored by the conditions in Lake Michigan (**Figure 4**). The organisms affiliated with the *Actinobacteria* lineages acI-B, acI-C, and acTH1 the *Alphaproteobacteria* lineage LD12, and the *Cytophagia* genus *Algoriphagus* had a strong preference for the conditions in Lake Michigan, while the *Betaproteobacteria* genera *Rhodobacter*, *Polynucleobacter*, and *Limnohabitans* had a strong preference for the urban-impacted waters (**Figure 4**).

Oligotyping Reveals Unique Environmental Distribution Patterns within Common Freshwater Lake Taxa

We used oligotyping to provide a refined sequence-based analysis for each the 18 common freshwater lake genera/lineages (see Materials and Methods for details). The 18 genera/lineages were represented by 351 oligotypes. In contrast to the whole community, the common freshwater lake genera/lineages did not exhibit significant richness or diversity differences ($p > 0.01$) between the urban-impacted and Lake Michigan waters (**Table 3**). These data in conjunction with the whole community diversity differences indicate that a similar level of diversity for common lake bacteria is present across both environments, but in the urban-impacted waterways these common lake community members are augmented with a large number of microorganisms that are uncommon in lake surface waters.

Although the common freshwater genera/lineages oligotype richness and diversity did not differ significantly between the urban-impacted and Lake Michigan samples, there was a significant difference in the distribution of these oligotypes between the two environments (**Figure 5**; ANOSIM $R = 0.90$, $p < 0.001$). A CLAM statistical approach using stringent conditions for environmental specialist determination (see Materials and Methods) indicated 80 of the 351 oligotypes exhibited significantly differentiated distributions between the two environments (51 associated with urban waters and 29 with Lake Michigan; **Figure 6**). The *Actinobacteria* lineages (acIA, acIB, acIC, acTH1) and the genus *Fluviicola* harbored the majority of Lake Michigan favored oligotypes (20; Supplementary Table 2), while the genera *Flavobacterium*, *Hydrogenophaga*, *Limnohabitans*, and *Rhodobacter* harbored a large number of the urban-water favored oligotypes (36; Supplementary Table 2). In several cases, oligotype pairs with one or two nucleotide differences (>97 or >96% identity, respectively) had opposite preferences for the urban waters vs. Lake Michigan (Supplementary Table 2).

Discussion

We observed a strong division between the bacterial community composition in the urban-impacted waterways of the Milwaukee estuary and of those in greater Lake Michigan. This result was not surprising given the numerous differences in the chemical and physical conditions of these two distinct but connected systems. In particular, higher nutrient and particle loads, water temperature, and lower residence time differentiate the sampled urban estuary waters and the waters of oligotrophic Lake Michigan. Nutrient and particle load, residence time, and temperature are all parameters that have been shown to impact the bacterial community makeup of freshwater systems (Lindström et al., 2005; Allgaier and Grossart, 2006; Newton et al., 2011a). Here we did not attempt to distinguish among these parameters as a driving force for community differentiation. Instead, we sought to further our understanding of urban influences on aquatic bacterial communities by identifying how the bacterial assemblages of urbanized waterways differed from those of a connected, but oligotrophic low urban-impacted system. Our study shows that a core pelagic bacterial community is present across this urban-eutrophic to oligotrophic gradient, as at all levels of classification: (1) taxon—genus, 2) sequence—reference-based, and (3) oligotype, the majority of sequence types in the lake were also recovered from the urban estuary. However, large changes in the bacterial assemblages were also present, notably a loss of diversity among taxa/lineages not considered common to lakes during the transition from the estuary to the open lake and a significant composition change both among cosmopolitan freshwater taxa/lineages and for oligotypes within these taxa/lineages.

Taxa Augmentation in Urban-influenced Waters

Our results showed that bacterial richness was higher in the urban waterways, supporting what had been reported in several studies examining bacterial community trends across lake productivity/trophy gradients (Kolmonen et al., 2011; Logue et al., 2012; Jankowski et al., 2014). The bacterial diversity estimates that emphasized more rare community members resulted in a larger diversity disparity between the urban estuary and Lake Michigan habitats, indicating the presence of a much larger pool of rare community members in the urban-influenced

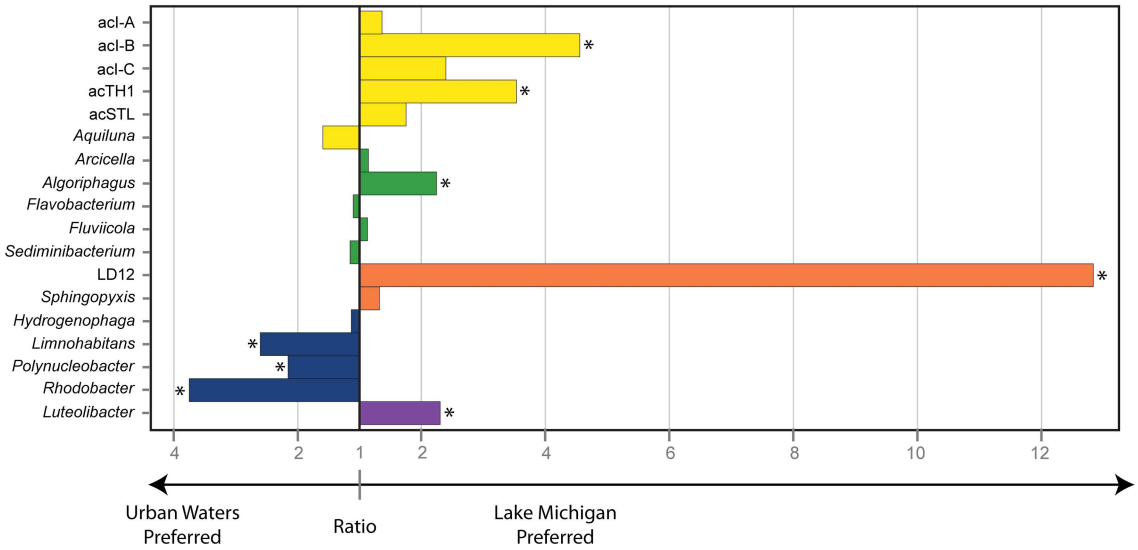


FIGURE 4 | Magnitude of habitat preference between the Urban Estuary waters and Lake Michigan for common freshwater genera/lineages. Habitat preference is determined by the ratio of the mean relative abundance of each genus/lineage among the urban estuary samples vs. its mean relative abundance in the Lake Michigan samples. Bars plotting to the left indicate an urban estuary preference while bars plotting to the right indicate a Lake Michigan preference. A significant association (Mann-Whitney *U*-test, $p \leq 0.01$) with either environment is indicated by an asterisk. Bar color indicates bacterial phylum where yellow, *Actinobacteria*; green, *Bacteroidetes*; orange, *Alphaproteobacteria*; blue, *Betaproteobacteria*; and purple, *Verrucomicrobia*.

TABLE 3 | Oligotype diversity comparisons for common freshwater genera/lineages[a].

Sample environment	Oligotype—Core freshwater		
	Richness	Inverse simpson	Tail
Urban estuary	144 ± 20	14 ± 4	19 ± 4
Lake Michigan	127 ± 18	14 ± 4	16 ± 2
Mann-Whitney U	92	64	96*

[a]*Mean and standard deviation are reported.*
**Indicates significance at* $p < 0.05$.

waterways. Our diversity estimates also indicated this rare pool of organisms was not derived from genotypic variation within the most common freshwater genera/lineages as at our finest scale of organism resolution, the oligotype, there were not on average differences in the richness and diversity between the two environments. Instead, we suggest a typical pelagic freshwater community in the urban estuary was being augmented by a large number of more rare freshwater organisms and/or organisms not found in pelagic lake communities.

In a lake productivity gradient study, Jankowski and coauthors suggested that increased habitat heterogeneity, which is typically associated with higher nutrient lake systems, provides additional resources that allow rare or absent taxa in oligotrophic systems to flourish in more eutrophic waters (Jankowski et al., 2014). Although we did not examine habitat heterogeneity explicitly here, it is likely a contributing factor to the increased diversity in the urban-influenced waterways. In our system, the variation, which is one measure of habitat heterogeneity, in the chemical and physical characteristics of the urban-influenced waters was much higher than that in greater Lake Michigan (**Table 1**). In support of the relationship between high habitat heterogeneity and recruitment of otherwise rare freshwater taxa, all but one taxon (categorized to genus) present in Lake Michigan was also present in at least one urban estuary sample, but nearly 400 taxa were detected only in the urban estuary samples. Also the larger number of oligotypes (51 vs. 29) from the common freshwater genera/lineages that were classified as being "urban-water" vs. "Lake Michigan" specialists may be a reflection of the increased resource diversity in the urban estuary.

It is also likely that surface runoff and stormwater discharge contributed significantly to the increased diversity observed in the urban-impacted waters. Impervious surfaces decrease water infiltration and increase surface runoff, and storm sewers redirect normal water flow. Together, these urban constructions dramatically alter the flow of water into urban surface waters (Brabec et al., 2002; Kaushal and Belt, 2012). In previous work, we estimated that under typical weather conditions, 2–11% of the 16S rRNA genes recovered in Milwaukee estuary samples had an urban environment origin (Fisher et al., 2015). We also found that some of these organisms, including organisms indicating human fecal pollution, were present consistently in the estuary over several years of sampling (Newton et al., 2011b; Fisher et al., 2014). At this time, it is not clear whether these organisms persist because dispersal is frequent enough from the urban environment to overcome local environmental dynamics (i.e., mass effects) or whether the conditions in these urban waterways allow these organisms to have prolonged survival and/or grow (i.e., species sorting; Lindström and Langenheder, 2012). If pathogenic organisms are maintained or proliferate in urban water systems, then these waterways may present a

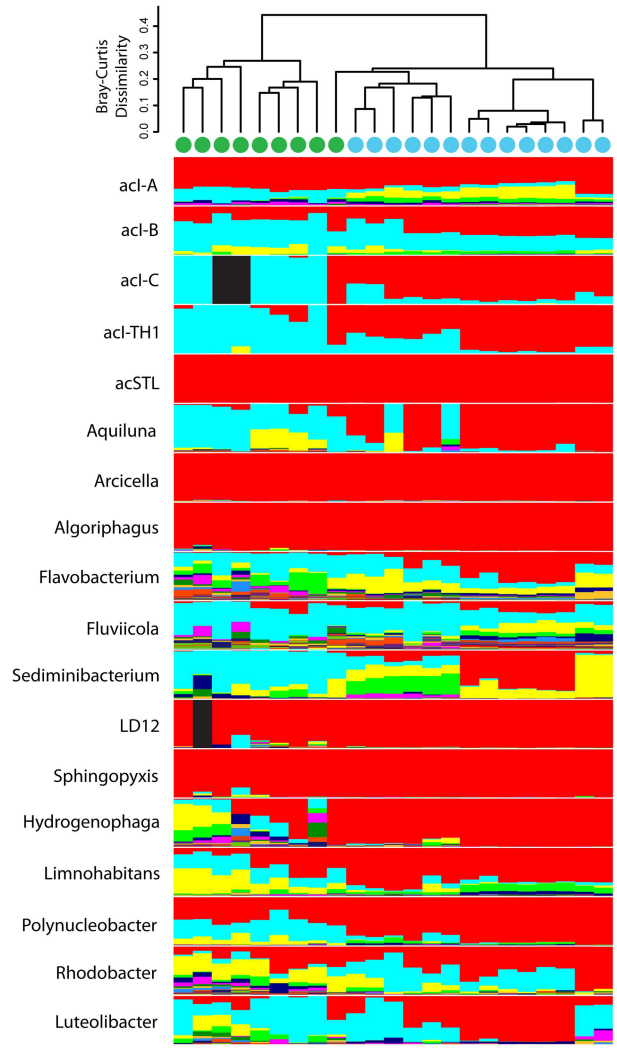


FIGURE 5 | Oligotype composition within each freshwater bacterial genus/lineage among samples is indicated in stacked bar plots. The relationship of the oligotype composition in each sample is depicted with a cluster dendrogram based on Bray-Curtis dissimilarity among samples. Fully black bars indicate no sequences were recovered from that genus/lineage in that sample. Samples were clustered via an unweighted pair group method with arithmetic mean calculation. Samples collected from the urban estuary are labeled with a green circle while those collected from Lake Michigan are labeled with a blue circle.

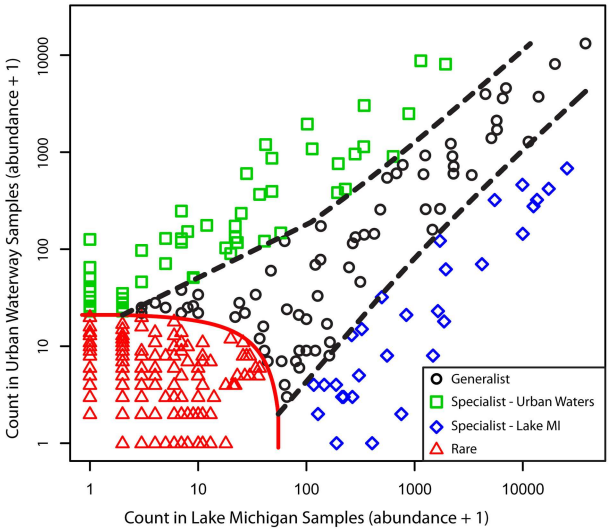


FIGURE 6 | Classification results for oligotype distributions between the urban estuary and Lake Michigan samples. Specialization was set with a threshold of $K = 0.75$, a coverage limit = 30, and $P = 0.01/351$. Classification status is indicated by point color and shape. The specialist and rarity thresholds are indicated by dotted lines.

greater human health risk than previously recognized (Fisher et al., 2014). Our data certainly suggest that the delivery of a large number of foreign, "urban-derived" bacteria may be common in urbanized waterways. This potentially massive immigration combined with the increased habitat heterogeneity in more eutrophic systems, appears to create a significantly more diverse bacterial assemblage in urbanized systems. We also note these data further support the idea that bacterial community assemblage patterns across productivity gradients contrast those for other organisms like fish and zooplankton, which typically exhibit decreased diversity in high productivity systems (Dodson et al., 2000; Barnett and Beisner, 2007; Jankowski et al., 2014).

Core Freshwater Community Shifts

Although we observed differences in the bacterial community composition between the urban estuary and Lake Michigan environments, the whole community analysis approach was not sufficient to identify whether these differences were the result of increased diversity in the urban-influenced waterways or stemmed from a combination of changes among rare and common organisms. Previous work across lake trophic gradients suggests that some bacterial groups are widespread (Jezbera et al., 2011, 2013; Kolmonen et al., 2011; Newton et al., 2011a; Jankowski et al., 2014), which could indicate most of the changes in eutrophic communities result from the increased abundance of rare or absent organisms in oligotrophic systems. Indeed changes in the so-called "conditionally rare taxa" can be a dominant driver of community change across environmental gradients (Shade et al., 2014). However, shifts in the dominant or common community members also frequently drive change in the bacterial community composition across environmental gradients (e.g., Gobet et al., 2010; Shade et al., 2010).

We used 18 ubiquitous freshwater lake genera/lineages to compare change in the composition among dominant freshwater taxa. Although these genera/lineages comprised a large proportion of the community in both environments, they differed in their distribution and generated sample similarity patterns similar to those represented by the whole community. The genera/lineages favored in either the eutrophic or oligotrophic waterways generally matched what is known about the lifestyles of these organisms. The urban estuary favored *Betaproteobacteria* genera including *Limnohabitans*, a genus defined by its fast-growth rates and copiotrophic lifestyle (Šimek et al., 2006; Jezbera et al., 2011), and *Rhodobacter*, a genus frequently abundant in near-shore eutrophic conditions, but less

common in the pelagic low-nutrient freshwater environment (Imhoff, 2006; Newton et al., 2011a). In contrast, the lineages acI and LD12 were favored in oligotrophic Lake Michigan. Both of these lineages are characterized by slower-growth, small cell sizes, and predation avoidance or oligotroph life strategies (Newton et al., 2011a; Salcher et al., 2011b; Ghylin et al., 2014). These results suggest that even at fairly broad taxonomic characterization such as genus or phylogenetic lineage there may be conserved characteristics within some freshwater groups, which contribute to community assembly patterns across urban/trophic gradients.

Within Genus/Lineage Composition Change

Recently, several studies have identified within-genus and with-species organism distribution patterns related to the biological and environmental properties of freshwater habitats. For example, it is now known that the ubiquitous freshwater bacterium *Polynucleobacter necessarius* subspecies *asymbioticus,* members of the genus *Limnohabitans,* and *Flavobacterium* each contain dozens of ribosomal gene sequence variants differentiated in their spatial and temporal distributions by lake characteristics such as pH, conductivity, and dissolved organic matter (Jezbera et al., 2011, 2013; Neuenschwander et al., 2015). Here we used an oligotyping approach to provide both high discriminatory power among closely related sequences (as low as one nucleotide) and to reduce the effects of sequencing errors (Eren et al., 2013a), so that we could better resolve distribution patterns within some of the most common freshwater bacterial genera/lineages. Despite the near ubiquity of the 18 examined freshwater genera/lineages, we observed the greatest community distinction between the urban estuary and Lake Michigan samples when using the higher organism discrimination provided by oligotyping. We also found that 8 of the 18 examined freshwater genera/lineages harbored both oligotypes that were favored in the urban estuary and oligotypes favored in Lake Michigan, including several instances where these opposite distribution patterns occurred among oligotypes with one or two nucleotide differences. It appears diversification is high within many of the ubiquitous freshwater bacterial genera and often includes organisms with distinct advantages over other closely related organisms in either eutrophic or oligotrophic waters. Together these results indicate that in addition to taxa augmentation, and common freshwater genus/lineage life strategy differences, a third mechanism, within-genus diversification, is driving community assemblage differences between the urban-influenced and Lake Michigan waters.

The combination of oligotyping and a habitat classification statistical approach also revealed a number of interesting trends among the common freshwater genera/lineages. The *Bacteroidetes* phylum, especially the genera *Flavobacterium*, *Fluviicola*, and *Sediminibacterium* had especially high oligotype richness, suggesting either the diversity of freshwater organisms associated with these genera is high or that a large number of urban-associated organisms belonging to these genera are delivered via city surface runoff and stormwater. *Flavobacterium* and *Sediminibacterium* had a large number of rare oligotypes, which supports the idea that many of these organisms are immigrants from the urban-environment. However, the *Flavobacterium* genus also contained a large number of oligotypes classified as urban-water specialists. The described diversity within this genus is immense and includes a number of fast-growing, opportunisitic species-like phylotypes (Neuenschwander et al., 2015) that are common in lotic systems (Read et al., 2015), which suggests these organisms should be common in many urban-influenced systems. Interestingly, the most abundant *Flavobacterium* oligotype was a Lake Michigan specialist and the only one of the 16 *Flavobacterium* oligotype specialists that was not urban-water associated.

A number of other genera/lineages were dominated by oligotypes assigned primarily to one of the environmental specialist categories. The commonly noted oligotroph clades acI-A, acI-B, and LD12 (Newton et al., 2011a) contained only Lake Michigan specialists. The genus *Fluviicola,* also contained a large number of Lake Michigan specialist oligotypes, but at this time relatively little is known about this genus (Salcher et al., 2011a). It is unlikely we over-classified oligotypes as specialists, as we chose a conservative criterion for classification (specialization $K = 3/4$; Chazdon et al., 2011). We also found some groups had a high number of oligotypes classified as generalists (e.g., acI-A, *Fluviicola*). It may be that some common freshwater organisms are true euryoecious organisms, resulting in broadly abundant distributions. It is also likely many generalist classifications are the result of our inability to distinguish among organisms with short-read 16S rRNA gene technologies. Recent studies have shown that the 16S-23S internal transcribed spacer (ITS) region, a less conserved bacterial genomic region, was able to identify organism distribution patterns among lakes that were otherwise obscured when examining 16S rRNA gene data (Jezbera et al., 2013; Hahn et al., 2015). The combined results of this study and the previous studies using ITS-based sequence groupings, indicate that more narrowly-defined organismal approaches are necessary to further our understanding of the biogeography and ecology of the ubiquitous freshwater pelagic bacteria.

Technical Considerations

The data in our sequence-based analyses were derived from three different sequencing platforms: 454 V6, 454 V6V4, and illumina V6. The choice of gene amplification conditions and sequencing platform are known to influence the composition of the resultant sequence data (e.g., Wu et al., 2010; Schloss et al., 2011). We also observed an influence of sequencing conditions on our bacterial community composition data (see Supplementary Figure 1 and associated Results Section); however, this influence on the overall community composition and diversity was small in comparison to the influence of the primary environmental gradient examined. Also, in all cases, the dominant freshwater oligotypes were present across all three sequencing platforms (see **Figure 5** for example), which suggested that although our analyses were influenced by the platform used, the differences did not manifest in the loss/gain of dominant freshwater groups. We agree with previous work that the use of a single sequencing platform gives the most robust cross-sample comparisons, but in the case of some meta-analyses, including this one, these data

may not exist. Our data suggest that cross-platform comparisons of 16S rRNA gene data are feasible and can give meaningful results especially when care is taken to quality-control sequence output and strong environmental gradients are examined. We suspect that if a single sequencing platform had been used here, the within-habitat diversity estimates and community composition variation in our data would have decreased and therefore furthered the distinction between the communities in urban-influenced waterways and oligotrophic Lake Michigan.

Conclusions

In our study system, water flows from the urban-impacted Milwaukee estuary into oligotrophic Lake Michigan, and with it, the estuary bacterial assemblage is continuously dispersed into the lake. Despite this direct connection, our examination of the bacterial communities across this environmental gradient revealed quite distinct assemblages. We found Lake Michigan harbors lower bacterial diversity than the urban-impacted estuary, shifts the dominance among common freshwater genera/lineages, and selects for what are likely unique species or populations within many of the common freshwater bacterial lineages. These data support the idea that the oligotrophic lake represents a strong selective force favoring a particular set of cosmopolitan freshwater taxa and largely prevents the successful dispersal of bacteria from the urban environment. It remains to be seen whether smaller but heavily urban-influenced lakes are more likely to contain persistent bacterial populations of urban origin. Either way, it is clear the environmental conditions in these urban waterways impact heavily the composition of the core freshwater community and increase the prevalence of bacteria that are not common to pelagic freshwaters.

The fact that many of the common freshwater genera/lineages harbored both "urban-estuary" and "Lake Michigan" specialists, further suggests the ubiquity of many common freshwater bacteria is a result of large-scale diversification within these groups (e.g., Jezbera et al., 2011; Hahn et al., 2015). Given the "island-like" nature of lakes across the globe and an ongoing desire to understand microbial diversification in natural systems, the study of within-genus or within-species genetic diversification of lake bacteria warrants further exploration. Whether or not urban waterways alter significantly the ecological function of these bacterial communities, select for genetic compositions or functional traits that are distinct from un-impacted surface waters, or contribute to the maintenance and/or proliferation of microorganisms that impact human health or well-being is yet to be determined. Further integration of the microbial components of urban landscapes is needed in the ongoing development of an ecological understanding and theory for urban areas.

Funding

This work was funded by the National Institutes of Health grant R01-AI091829 and MMSD Contract M03029P10 to SM.

Author Contributions

RN and SM conceived the work and analyses. RN carried out the data collection and analyses. RN and SM wrote the paper.

Acknowledgments

We thank the crew of the R/V Neeskay and Jessica VandeWalle, Elizabeth Sauer, and Colin Peake for assistance with lake sampling. Mitch Sogin, Hilary Morrison, and Joseph Vineis provided sequencing support at the Marine Biological Laboratory. We also thank the two reviewers for their contributions to manuscript revision.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.01028

References

Alberti, M. (2005). The effects of urban patterns on ecosystem function. *Int. Reg. Sci. Rev.* 28, 168–192. doi: 10.1177/0160017605275160

Alberti, M., Booth, D., Hill, K., Coburn, B., Avolio, C., Coe, S., et al. (2007). The impact of urban patterns on aquatic ecosystems: an empirical analysis in Puget lowland sub-basins. *Landsc. Urban Plan.* 80, 345–361. doi: 10.1016/j.landurbplan.2006.08.001

Allan, J. D. (2004). Landscapes and riverscapes: the influence of land use on stream ecosystems. *Annu. Rev. Ecol. Syst.* 35, 257–284. doi: 10.1146/annurev.ecolsys.35.120202.110122

Allgaier, M., and Grossart, H. P. (2006). Seasonal dynamics and phylogenetic diversity of free-living and particle-associated bacterial communities in four lakes in northeastern Germany. *Aquat. Microb. Ecol.* 45, 115–128. doi: 10.3354/ame045115

Barnett, A., and Beisner, B. E. (2007). Zooplankton biodiversity and lake trophic state: explanations invoking resource abundance and distribution. *Ecology* 88, 1675–1686. doi: 10.1890/06-1056.1

Bell, T., Bonsall, M. B., Buckling, A., Whiteley, A. S., Goodall, T., and Griffiths, R. I. (2010). Protists have divergent effects on bacterial diversity along a productivity gradient. *Biol. Lett.* 6, 639–642. doi: 10.1098/rsbl.2010.0027

Booth, D. B., and Jackson, C. R. (1997). Urbanization of aquatic systems: degradation thresholds, stormwater detection, and the limits of mitigation. *J. Am. Water Resour. Assoc.* 33, 1077–1090. doi: 10.1111/j.1752-1688.1997.tb04126.x

Brabec, E., Schulte, S., and Richards, P. L. (2002). Impervious surfaces and water quality: a review of current literature and its implications for watershed planning. *J. Plan Lit.* 16, 499–514. doi: 10.1177/088541202400903563

Cardinale, B. J., Hillebrand, H., Harpole, W. S., Gross, K., and Ptacnik, R. (2009). Separating the influence of resource 'availability' from resource 'imbalance' on productivity-diversity relationships. *Ecol. Lett.* 12, 475–487. doi: 10.1111/j.1461-0248.2009.01317.x

Carpenter, S. R., Caraco, N. F., Correll, D. L., Howarth, R. W., Sharpley, A. N., and Smith, V. H. (1998). Nonpoint pollution of surface waters with phosphorus and nitrogen. *Ecol. Appl.* 8, 559–568. doi: 10.1890/1051-0761(1998)008[0559:NPOSWW]2.0.CO;2

Chase, J. M., and Leibold, M. A. (2002). Spatial scale dictates the productivity-biodiversity relationship. *Nature* 416, 427–430. doi: 10.1038/416427a

Chazdon, R. L., Chao, A., Colwell, R. K., Lin, S. Y., Norden, N., Letcher, S. G., et al. (2011). A novel statistical method for classifying habitat generalists and specialists. *Ecology* 92, 1332–1343. doi: 10.1890/10-1345.1

Correll, D. L. (1998). The role of phosphorus in the eutrophication of receiving waters: a review. *J. Environ. Qual.* 27, 261–266. doi: 10.2134/jeq1998.00472425002700020004x

Corvalan, C., Hales, S., McMichael, A., Butler, C., Campbell-Lendrum, D., Confalonieri, U., et al. (2005). *Ecosystems and Human Well-Being*. Vol. 5. Washington, DC: Island Press.

Cotner, J. B., and Biddanda, B. A. (2002). Small players, large role: microbial influence on biogeochemical processes in pelagic aquatic ecosystems. *Ecosystems* 5, 105–121. doi: 10.1007/s10021-001-0059-3

Dodson, S. I., Arnott, S. E., and Cottingham, K. L. (2000). The relationship in lake communities between primary productivity and species richness. *Ecology* 81, 2662–2679. doi: 10.1890/0012-9658(2000)081[2662:TRILCB]2.0.CO;2

Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., et al. (2013a). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* 4, 1111–1119. doi: 10.1111/2041-210X.12114

Eren, A. M., Vineis, J. H., Morrison, H. G., and Sogin, M. L. (2013b). A filtering method to generate high quality short reads using Illumina paired-end technology. *PLoS ONE* 8:e66643. doi: 10.1371/journal.pone.0066643

Fisher, J. C., Levican, A., Figueras, M. J., and McLellan, S. L. (2014). Population dynamics and ecology of Arcobacter in sewage. *Front. Microbiol.* 5:525. doi: 10.3389/fmicb.2014.00525

Fisher, J. C., Newton, R. J., Dila, D. K., and McLellan, S. L. (2015). Urban microbial ecology of a freshwater estuary of Lake Michigan. *Elementa* 3:000064. doi: 10.12952/journal.elementa.000064

Ghylin, T. W., Garcia, S. L., Moya, F., Oyserman, B. O., Schwientek, P., Forest, K., et al. (2014). Comparative single-cell genomics reveals potential ecological niches for the freshwater acI Actinobacteria lineage. *ISME J.* 8, 2503–2516. doi: 10.1038/ismej.2014.135

Gobet, A., Quince, C., and Ramette, A. (2010). Multivariate cutoff level analysis (MultiCoLA) of large community data sets. *Nucleic Acids Res.* 28, e155. doi: 10.1093/nar/gkq545

Grimm, N. B., Morgan Grove, J., Pickett, S. T. A., and Redman, C. L. (2000). Integrated approaches to long-term studies of urban ecological systems. *Bioscience* 50, 571–584. doi: 10.1641/0006-3568(2000)050[0571:IATLTO]2.0.CO;2

Hahn, M. W., Koll, U., Jezberová, J., and Camacho, A. (2015). Global phylogeography of pelagic Polynucleobacter bacteria: restricted geographic distribution of subgroups, isolation by distance and influence of climate. *Enivron. Microbiol.* 17, 829–840. doi: 10.1111/1462-2920.12532

Hale, R. L., Turnbull, L., Earl, S. R., Childers, D. L., and Grimm, N. B. (2015). Stormwater infrastructure controls runoff and dissolved material export from arid urban watersheds. *Ecosystems* 18, 62–75. doi: 10.1007/s10021-014-9812-2

Horner-Devine, M. C., Leibold, M. A., Smith, V. H., and Bohannan, B. J. M. (2003). Bacterial diversity patterns along a gradient of primary productivity. *Ecol. Lett.* 6, 613–622. doi: 10.1046/j.1461-0248.2003.00472.x

Huse, S. M., Dethlefsen, L., Huber, J. A., Mark Welch, D., Relman, D. A., and Sogin, M. L. (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet.* 4:e1000255. doi: 10.1371/journal.pgen.1000255

Huse, S. M., Mark Welch, D. B., Voorhis, A., Shipunova, A., Morrison, H. G., Eren, A. M., et al. (2014). VAMPS?: a website for visualization and analysis of microbial population structures. *BMC Bioinformatics* 15:41. doi: 10.1186/1471-2105-15-41

Imhoff, J. (2006). "The phototrophic alpha-proteobacteria," in *The Prokaryotes*, *3rd Edn.*, Vol. 5, eds M. Dworkin, S. Falkow, K.-H. Schleifer, E. Rosenberg, and E. Stackebrandt (New York, NY: Springer Science), 41–64.

Jankowski, K., Schindler, D. E., and Horner-Devine, M. C. (2014). Resource availability and spatial heterogeneity control bacterial community response to nutrient enrichment in lakes. *PLoS ONE* 9:e86991. doi: 10.1371/journal.pone.0086991

Jezbera, J., Jezberová, J., Brandt, U., and Hahn, M. W. (2011). Ubiquity of *Polynucleobacter necessarius* subspecies *asymbioticus* results from ecological diversification. *Environ. Microbiol.* 13, 922–931. doi: 10.1111/j.1462-2920.2010.02396.x

Jezbera, J., Jezberová, J., Kasalický, V., Šimek, K., and Hahn, M. W. (2013). Patterns of *Limnohabitans* microdiversity across a large set of freshwater habitats as revealed by reverse line blot hybridization. *PLoS ONE* 8:e58527. doi: 10.1371/journal.pone.0058527

Kaushal, S., and Belt, K. (2012). The urban watershed continuum: evolving spatial and temporal dimensions. *Urban Ecosyst.* 15, 409–435. doi: 10.1007/s11252-012-0226-7

King, G. M. (2014). Urban microbiomes and urban ecology: how do microbes in the built environment affect human sustainability in cities? *J. Microbiol.* 52, 721–728. doi: 10.1007/s12275-014-4364-x

Kolmonen, E., Haukka, K., Rantala-Ylinen, A., Rajaniemi-Wacklin, P., Lepistö, L., and Sivonen, K. (2011). Bacterioplankton community composition in 67 Finnish lakes differs according to trophic status. *Aquat. Microb. Ecol.* 62, 241–250. doi: 10.3354/ame01461

Korhonen, J. J., Wang, J., and Soininen, J. (2011). Productivity-diversity relationships in lake plankton communities. *PLoS ONE* 6:e22041. doi: 10.1371/journal.pone.0022041

Lande, R. (1996). Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos* 76, 5–13. doi: 10.2307/3545743

Li, K., Bihan, M., Yooseph, S., and Methé, B. A. (2012). Analyses of the microbial diversity across the human microbiome. *PLoS ONE* 7:e32118. doi: 10.1371/journal.pone.0032118

Lindström, E. S., Kamst-Van Agterveld, M. P., and Zwart, G. (2005). Distribution of typical freshwater bacterial groups is associated with pH, temperature, and lake water retention time. *Appl. Environ. Microbiol.* 71, 8201–8206. doi: 10.1128/AEM.71.12.8201-8206.2005

Lindström, E. S., and Langenheder, S. (2012). Local and regional factors influencing bacterial community assembly. *Environ. Microbiol. Rep.* 4, 1–9. doi: 10.1111/j.1758-2229.2011.00257.x

Logue, J. B., Langenheder, S., Andersson, A. F., Bertilsson, S., Drakare, S., Lanzén, A., et al. (2012). Freshwater bacterioplankton richness in oligotrophic lakes depends on nutrient availability rather than on species-area relationships. *ISME J.* 6, 1127–1136. doi: 10.1038/ismej.2011.184

Longmuir, A., Shurin, J. B., and Clasen, J. L. (2007). Independent gradients of producer, consumer, and microbial diversity in Lake Plankton. *Ecology* 88, 1663–1674. doi: 10.1890/06-1448.1

Mann, H. B., and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* 18, 50–60. doi: 10.1214/aoms/1177730491

McLellan, S. L., Huse, S. M., Mueller−Spitz, S. R., Andreishcheva, E. N., and Sogin, M. L. (2010). Diversity and population structure of sewage−derived microorganisms in wastewater treatment plant influent. *Environ. Microbiol.* 12, 378–392. doi: 10.1111/j.1462-2920.2009.02075.x

Mittelbach, G. G., Steiner, C. F., Scheiner, S. M., Gross, K. L., Reynolds, H. L., Waide, R. B., et al. (2001). What is the observed relationship between species richness and productivity? *Ecology* 82, 2381–2396. doi: 10.1890/0012-9658(2001)082[2381:WITORB]2.0.CO;2

Neuenschwander, S. M., Pernthaler, J., Posch, T., and Salcher, M. M. (2015). Seasonal growth potential of rare lake water bacteria suggest their disproportional contribution to carbon fluxes. *Environ. Microbiol.* 17, 781–795. doi: 10.1111/1462-2920.12520

Newton, R. J., Bootsma, M. J., Morrison, H. G., Sogin, M. L., and McLellan, S. L. (2013). A microbial signature approach to identify fecal pollution in the waters off an urbanized coast of Lake Michigan. *Microb. Ecol.* 65, 1011–1023. doi: 10.1007/s00248-013-0200-9

Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D., and Bertilsson, S. (2011a). A guide to the natural history of freshwater lake bacteria. *Microbiol. Mol. Biol. Rev.* 75, 14–49. doi: 10.1128/MMBR.00028-10

Newton, R. J., Vandewalle, J. L., Borchardt, M. A., Gorelick, M. H., and McLellan, S. L. (2011b). *Lachnospiraceae* and *Bacteroidales* alternative fecal indicators reveal chronic human sewage contamination in an urban harbor. *Appl. Environ. Microbiol.* 77, 6972–6981. doi: 10.1128/AEM.05480-11

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., et al. (2013). *vegan: Community Ecology Package*. R package version 2.0-8. Available online at: http://CRAN.R-project.org/package=vegan

Paerl, H. W., Dyble, J., Moisander, P. H., Noble, R. T., Piehler, M. F., Pinckney, J. L., et al. (2003). Microbial indicators of aquatic ecosystem change: current applications to eutrophication studies. *FEMS Microbiol. Ecol.* 46, 233–246. doi: 10.1016/S0168-6496(03)00200-9

Paul, M. J., and Meyer, J. L. (2001). Streams in the urban landscape. *Annu. Rev. Ecol. Syst.* 32, 333–365. doi: 10.1146/annurev.ecolsys.32.081501.114040

Pickett, S. T. A., Cadenasso, M. L., Grove, J. M., Boone, C. G., Groffman, P. M., Irwin, E., et al. (2011). Urban ecological systems: scientific foundations and a decade of progress. *J. Environ. Manag.* 92, 331–362. doi: 10.1016/j.jenvman.2010.08.022

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., et al. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* 35, 7188–7196. doi: 10.1093/nar/gkm864

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. ISBN 3-90051-0700. Available online at: http://www.r-project.org

Read, D. S., Gweon, H. S., Bowes, M. J., Newbold, L. K., Field, D., Bailey, M. J., et al. (2015). Catchment-scale biogeography of riverine bacterioplankton. *ISME J.* 9, 516–526. doi: 10.1038/ismej.2014.166

Salcher, M. M., Pernthaler, J., Frater, N., and Posch, T. (2011a). Vertical and longitudinal distribution patterns of different bacterioplankton populations in a canyon-shaped, deep prealpine lake. *Limnol. Oceanogr.* 56, 2027–2039. doi: 10.4319/lo.2011.56.6.2027

Salcher, M. M., Pernthaler, J., and Posch, T. (2011b). Seasonal bloom dynamics and ecophysiology of the freshwater sister clade of SAR11 bacteria 'that rule the waves' (LD12). *ISME J.* 5, 1242–1252. doi: 10.1038/ismej.2011.8

Schloss, P. D., Gevers, D., and Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE* 6:e27310. doi: 10.1371/journal.pone.0027310

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09

Shade, A., Chiu, C.-Y., and McMahon, K. D. (2010). Seasonal and episodic lake mixing stimulate differential planktonic bacterial dynamics. *Microb. Ecol.* 59, 546–554. doi: 10.1007/s00248-009-9589-6

Shade, A., Jones, S. E., Caporaso, J. G., Handelsman, J., Knight, R., Fierer, N., et al. (2014). Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *mBio* 5:e01371-14. doi: 10.1128/mBio.01371-14

Šimek, K., Hornák, K., Jezbera, J., Nedoma, J., Vrba, J., Straskrábová, V., et al. (2006). Maximum growth rates and possible life strategies of different bacterioplankton groups in relation to phosphorus availability in a freshwater reservoir. *Environ. Microbiol.* 8, 1613–1624. doi: 10.1111/j.1462-2920.2006.01053.x

Smith, V. H. (2007). Microbial diversity-productivity relationships in aquatic ecosystems. *FEMS Microbiol. Ecol.* 62, 181–186. doi: 10.1111/j.1574-6941.2007.00381.x

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., et al. (2013). *gplots: Various R Programming Tools for Plotting Data*. R package version 2.12.11. Danbury, CT.

Wickham, H. (2009). *ggplot2: Elegant Graphis for Data Analysis*. New York, NY: Springer.

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *J. Stat. Softw.* 40, 1–29.

Wollheim, W. M., Pellerin, B. A., Vorosmarty, C. J., and Hopkinson, C. S. (2005). N retention in urbanizing headwater catchments. *Ecosystems* 8, 871–884. doi: 10.1007/s10021-005-0178-3

Wu, G. D., Lewis, J. D., Hoffmann, C., Chen, Y.-Y., Knight, R., Bittinger, K., et al. (2010). Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. *BMC Microbiol.* 10:206. doi: 10.1186/1471-2180-10-206

Yannarell, A. C., and Triplett, E. W. (2004). Within-and between-lake variability in the composition of bacterioplankton communities: investigations using multiple spatial scales. *Appl. Environ. Microbiol.* 70, 214–223. doi: 10.1128/AEM.70.1.214-223.2004

Yannarell, A. C., and Triplett, E. W. (2005). Geographic and environmental sources of variation in lake bacterial community composition. *Appl. Environ. Microbiol.* 71, 227–239. doi: 10.1128/AEM.71.1.227-239.2005

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 19 December 2014
doi: 10.3389/fmicb.2014.00646

Phaeocystis antarctica blooms strongly influence bacterial community structures in the Amundsen Sea polynya

Tom O. Delmont[1], Katherine M. Hammar[1], Hugh W. Ducklow[2], Patricia L. Yager[3] and Anton F. Post[1]*

[1] Marine Biology Laboratory, Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Woods Hole, MA, USA
[2] Lamont Doherty Earth Observatory, Columbia University, Palisades, NY, USA
[3] Department of Marine Sciences, University of Georgia, Athens, GA, USA

Edited by:
Lois Maignien, University of Western Brittany, France

Reviewed by:
Jean-Baptiste Ramond, University of Pretoria, South Africa
C. N. Shulse, University of Hawai'i at Manoa, USA
Meghan Chafee, Max Planck Institute for Marine Microbiology, Germany

***Correspondence:**
Anton F. Post, Marine Biological Laboratory, Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, 7 MBL Street, Woods Hole, MA 02543, USA
e-mail: apost@mbl.edu

Rising temperatures and changing winds drive the expansion of the highly productive polynyas (open water areas surrounded by sea ice) abutting the Antarctic continent. Phytoplankton blooms in polynyas are often dominated by the haptophyte *Phaeocystis antarctica*, and they generate the organic carbon that enters the resident microbial food web. Yet, little is known about how *Phaeocystis* blooms shape bacterial community structures and carbon fluxes in these systems. We identified the bacterial communities that accompanied a *Phaeocystis* bloom in the Amundsen Sea polynya during the austral summers of 2007–2008 and 2010–2011. These communities are distinct from those determined for the Antarctic Circumpolar Current (ACC) and off the Palmer Peninsula. Diversity patterns for most microbial taxa in the Amundsen Sea depended on location (e.g., waters abutting the pack ice near the shelf break and at the edge of the Dotson glacier) and depth, reflecting different niche adaptations within the confines of this isolated ecosystem. Inside the polynya, *P. antarctica* coexisted with the bacterial taxa *Polaribacter* sensu lato, a cryptic *Oceanospirillum*, SAR92 and *Pelagibacter*. These taxa were dominated by a single oligotype (genotypes partitioned by Shannon entropy analysis) and together contributed up to 73% of the bacterial community. Size fractionation of the bacterial community [$<3\,\mu m$ (free-living bacteria) vs. $>3\,\mu m$ (particle-associated bacteria)] identified several taxa (especially SAR92) that were preferentially associated with *Phaeocystis* colonies, indicative of a distinct role in *Phaeocystis* bloom ecology. In contrast, particle-associated bacteria at 250 m depth were enriched in *Colwellia* and members of the Cryomorphaceae suggesting that they play important roles in the decay of *Phaeocystis* blooms.

Keywords: Amundsen Sea polynya, phytoplankton bloom, *Phaeocystis antarctica*, microbial community structure, mutualism

INTRODUCTION

Phytoplankton blooms account for a significant fraction of marine primary production. Such blooms occur in the open ocean [e.g., by the cyanobacterium *Trichodesmium* (Capone et al., 1997, 2005) or the diatoms *Hemiaulus* and *Rhizoselenia* (Subramaniam et al., 2008)] as well as in the coastal ocean (e.g., *Karenia*, *Pseudonitzschia*), where they can be a nuisance for aquaculture and fisheries. Bloom events continue to intrigue ocean researchers as the physiological underpinnings of their development, duration and demise remain unresolved (Behrenfeld and Boss, 2014). Species like *Trichodesmium* create short-lived (10–20 days) blooms of the rise-and-crash type, whereas blooms of other species may be sustained over considerably longer periods (1–3 months). The haptophyte *Phaeocystis* is a ubiquitous marine phytoplankter that causes blooms in coastal seas. Species contained in this genus have typical geographic distributions with *P. pouchetii* dominating in the Arctic Ocean, *P. globosa* in temperate coastal seas and *P. antarctica* occupying diverse niches in the Southern Ocean, respectively (Schoemann et al., 2005). *Phaeocystis* blooms significantly impact local carbon, nutrient and sulfur cycles (Van Boekel and Stefels, 1993; Yager et al., 2012) and can disturb ecosystems (Chen et al., 1999).

Extensive phytoplankton blooms occur in Antarctic waters (Arrigo and Van Dijken, 2003). In explored Antarctic polynyas (large open water expanses in sea ice), blooms are often dominated by *P. antarctica* (Arrigo et al., 1999; Smith et al., 2000; Yager et al., 2012; Kim et al., 2013). These populations are generally limited by light and iron availability (Martin et al., 1990; Bertrand et al., 2011a; Alderkamp et al., 2012) and bloom formation occurs when environmental conditions become favorable (Zingone et al., 1999; Smith et al., 2003; Vogt et al., 2012). The duration and scale of these favorable conditions are enhanced by rising temperatures and winds (Arrigo et al., 1998; Turner et al., 2005; Yager et al., 2012). *Phaeocystis* blooms occur in the surface mixed layer and they can span much of the austral summer (Arrigo et al., 1999; Wolf et al., 2013). Their populations rapidly draw down CO_2 concentrations to <100 ppm (Arrigo and Van Dijken, 2003; Yager et al., 2012). Thus, *Phaeocystis* blooms supply

organic carbon and nutrients to the food web inside polynyas (Rousseau et al., 2000; Kirchman et al., 2001a; Ducklow, 2003) and provide ecological niches for microbial heterotrophs (e.g., those capable of degrading particle organic carbon). The requirement for a continued supply of essential nutrients and growth factors like vitamins (Bertrand et al., 2011b), along with the removal of metabolites and exudates that negatively affect algal growth, likely influence bloom intensity and duration in most water bodies. The mechanisms by which *Phaeocystis* blooms sustain their activity over time are not well understood and possibly involve important functional interactions with their surrounding microbial community.

Bacteria entertain a wide range of interactions with phytoplankton (Cole, 1982; Doucette, 1995; Croft et al., 2005; Sher et al., 2011), and these interactions in turn may determine the composition of the bacterial community. A succession of bacterial taxa was observed during a phytoplankton bloom in the North Sea and their occurrence patterns were linked to their ability to degrade algal-derived organic matter (Teeling et al., 2012). The final phase of the bloom was shown to favor *Ulvibacter* and *Formosa* dominance during early and mid-stages of the decline, and to *Polaribacter* in the final stages. *Polaribacter* abundances correlate positively with chlorophyll a concentrations in the Southern Ocean and they play an active role in remineralizing organic matter generated from primary production during bloom events (Wilkins et al., 2013b; Williams et al., 2013). Not only is the free-living bacterial community affected by phytoplankton blooms, the bacterial epibionts that reside on algal cells or colonies alter their community structure as a phytoplankton bloom progresses. For example, *Trichodesmium* colonies have an epibiotic bacterial flora that is distinct from the free-living bacterial community (Hmelo et al., 2012). These changes in community structure are in part driven by chemotactic responses of bacterial taxa to phytoplankton exudates like dimethylsulfoniopropionate (Stocker et al., 2008; Seymour et al., 2010; Stocker and Seymour, 2012). Quorum sensing by associated bacteria was shown to enhance phosphate scavenging by *Trichodesmium* (Van Mooy et al., 2011).

P. antarctica blooms consist mostly of colonies that reach a diameter of a few millimeters and can be identified by the naked eye (Carlson et al., 1998; Smith et al., 1998). They form an important interface between the primary producers and their environment. The mucilaginous colony matrix that encapsulates the *Phaeocystis* cells forms a barrier for the exchange of dissolved compounds but it may also provide a habitat for bacterial species. Previous studies suggest that *P. antarctica* blooms affect microbial community structure in their immediate surroundings. An iron-induced phytoplankton bloom study in the Southern Ocean (West et al., 2008) showed that *Roseobacter*, SAR92 and Bacteroidetes dominated the bacterial community inside the bloom, whereas outside the bloom *SAR11*, *Polaribacter* and different *Roseobacter* types were more prevalent. A metagenomic study of coastal waters near the Antarctic Peninsula showed that bacterial communities were dominated by genotypes capable of chemotrophic, photoheterotrophic and aerobic anoxygenic photosynthetic metabolism (Grzymski et al., 2012). These communities were rich in SAR11-like genotypes, but poor in Bacteroidetes and Gammaproteobacteria (Grzymski et al., 2012). The Amundsen Sea polynya (ASP) is dominated by *Polaribacter* spp. (Bacteroidetes) and Oceanospirillales (Gammaproteobacteria) members (Ghiglione et al., 2012; Kim et al., 2013; Richert et al., submitted). These studies also reported different bacterial communities in areas with ice cover as compared to open water samples. Likewise, iceberg melt affects bacterial communities with Gammaproteobacteria dominating deep waters near icebergs and Bacteroidetes dominating elsewhere (Dinasquet et al., submitted). In accordance with findings elsewhere (Piquet et al., 2011), bacterial abundances in the ASP were highly correlated with *Phaeocystis* and diatoms suggesting a close coupling between the phytoplankton and bacterial communities (Kim et al., 2013). Bacterial productivity is not only higher in the open polynya as compared to adjacent water bodies, the bulk of bacterial exoenzyme activity, respiration and production was associated with the size fraction that contains *Phaeocystis* particles (Williams et al., submitted). However, so far no efforts have been made to assess whether members of the bacterial community interact directly with *Phaeocystis*. We hypothesized that the bacterial community is not limited to the biomineralization of organic carbon and nutrients but that specialized members of this community may also entertain interactions with *Phaeocystis* that stimulate bloom formation or act to enhance and perpetuate such blooms.

In an attempt to investigate the occurrence of such interactions we sampled two *P. antarctica* bloom events (2007–2008 and 2010–2011) from the highly productive ASP in west Antarctica (Arrigo and Van Dijken, 2003; Alderkamp et al., 2012; Mills et al., 2012). We targeted the V6 hypervariable region of the 16S rRNA gene with primers that target bacterial and eukaryotic organelle templates. We used this approach to generate large V6 sequence datasets (10^5–10^6 reads per sample) for various locations at the shelf break, inside the polynya and near the Dotson glacier. We coupled the depth and quality of paired-end Illumina sequencing to oligotyping, a sensitive bioinformatics tools to partition conserved genotype clusters within key microbial taxa, revealing an extended diversity. Using different sampling strategies, we discovered a number of bacterial taxa that preferentially associate with *P. antarctica* and their abundance correlated with that of *Phaeocystis*. We also identified different bacterial taxa that may play a specific role in bloom demise and the degradation of *Phaeocystis* biomass at depth.

MATERIALS AND METHODS

Water samples were collected at various sites across the ASP during the austral summers of 2007–2008 (aboard the icebreaker R/V "Oden"; depth profiles) and 2010–2011 (ASPIRE cruise aboard the R/V "Nathaniel B Palmer," horizontal grid of surface samples) (**Figure 1**, **Table S1**). Cruise track, sampling sites and an overview of geochemical and biological properties have been detailed elsewhere (Yager et al., 2012). Additional information can be found in the BCO-DMO database (http://osprey.bco-dmo.org/project.cfm?id=146&flag=view) and in **Table S2**. For the 2010–2011 cruise, water samples (3–10 L) for microbial community sequence analyses were passed over a 20 μm mesh and collected onto 0.2 μm Sterivex membrane filter cartridges by

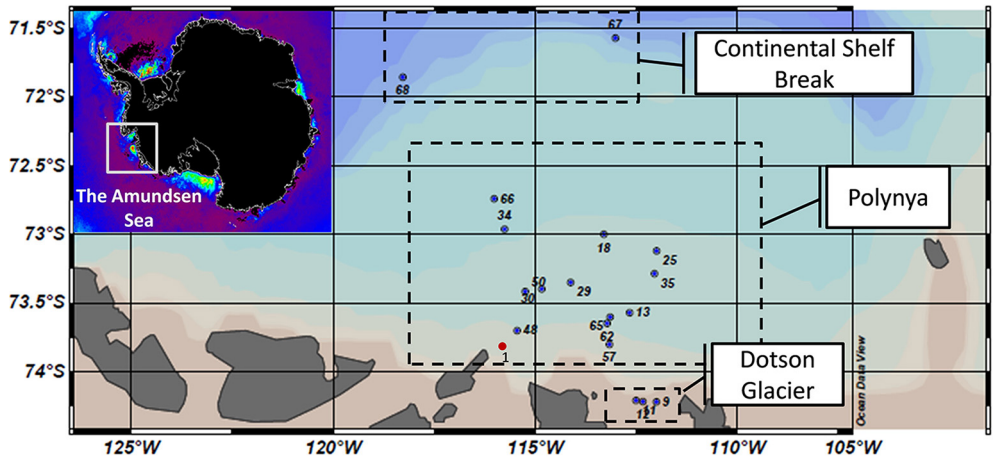


FIGURE 1 | Map of sampling sites across the Amundsen Sea polynya that were visited during the Oden Southern Ocean (OSO) cruise in 2007–2008 (red dot) and the ASPIRE cruise in 2010–2011 (blue dots). Samples were grouped into three major regions: the Continental Shelf Break, the open waters of the polynya and waters abutting the Dotson Glacier. Insert: A map of chlorophyll *a* concentrations (red: high; purple: low) in polynyas abutting the Antarctic continent as estimated by remote sensing (courtesy of Dr. Kevin R. Arrigo, Stanford University). Note that sample numbers are only informative of the station number tracked during the ASPIRE cruise. The corresponding samples are presented in the supplemental material.

pressure filtration (Whatman Masterflex L/S series). Since high biomass caused rapid clogging of the filters, the sampling volumes varied between stations. Two distinct plankton size classes (0.2–3 μm and 3–200 μm) were fractionated for samples collected during the 2007–2008 cruise. This sampling effort (10–20 L) was done along a depth profile that spanned the full water column (**Figure 1**, **Table S1**) and the microbial community analysis was part of the International Consensus for Marine Microbes project. Filters were quickly frozen in the headspace of a LN_2 Dewar and stored at −80°C prior to DNA extraction. We note that the 2007–2008 data were determined on samples from a single depth profile inside the ASP. It was decided to incorporate these data in order to derive first hints regarding the reproducibility of bacterial community compositions that accompany *Phaeocystis* blooms in the ASP and to gain early insights into the bacterial taxa that may associate with *Phaeocystis* colonies and other particles. Metadata of the various samples are presented in **Table S1** and in the supplemental material. DNA extraction was performed using the Puregene kit (Gentra®) after disruption of the cells with lytic enzyme coupled to proteinase K (Sinigalliano et al., 2007). DNA concentrations were quantified using a Nanodrop 2000 instrument (Thermo Fisher Scientific, Wilmington, DE).

The V6 hypervariable region of the 16S rRNA gene (typically 60–65 bp in length) was amplified (25 cycles using HiFi buffer 1X, $MgSO_4$ 2 mM, dNTPs 0.2 mM, combined primers 0.2 mM and four units of platinum HiFi) in triplicate PCR reaction from 10 ng of environmental DNA templates with reverse primer (1046R) "CGACRRCCATGCANCACCT" and the forward primer mix (967F) "CTAACCGANGAACCTYACC," "ATACGCGARGAACCTTACC," "CNACGCGAAGAACCTTANC," and "CAACGCGMARAACCTTACC." PCR cycle conditions were defined as follow: 30 s at 94°C followed by 45 s at 60°C and 1 min at 72°C. The PCR started with 3 min at 94°C and ended with 2 min at 72°C followed by a rapid stepdown to 4°C. Negative controls (no template DNA) were run for each of the index primer combinations in the PCR reactions. V6 amplicon sequences from samples collected during the 2007–2008 R/V "Oden" cruise ($n = 12$) were obtained on a GS-FLX pyrosequencing platform. Sequence reads were subsequently trimmed for low-quality sequences (Huse et al., 2007). For samples collected on the ASPIRE cruise during the 2010–2011 austral summer ($n = 23$), a paired-end sequencing strategy for Illumina Hiseq platform was employed with custom fusion primers described previously (Eren et al., 2013b) targeting the V6 hypervariable region of the 16S rRNA gene. The library design provided a complete overlap of the V6 region, and high-quality V6 reads were generated by requiring a complete match between the two reads of each pair (Eren et al., 2013b). Read sizes of the trimmed datasets are presented in **Table S1**.

Quality-filtered datasets were subsequently annotated using the Global Assignment of Sequence Taxonomy (GAST) pipeline (Huse et al., 2008) using the SILVA 111 database for reference (Quast et al., 2013). The datasets are publically accessible through the VAMPS website (http://vamps.mbl.edu/) under the project names ICM_ASA_Bv6 (2007–2008) and AFP_ASPIR_Bv6 (2011–2012). In order to assess within taxon diversity, reads affiliated to a given genus with GAST were submitted to oligotyping, a computational method for taxonomical partitioning based on Shannon entropy decomposition (Eren et al., 2013a). By utilizing only the nucleotide positions that show high variation, and disregarding the redundant sites with low entropy, oligotyping analysis employs only a fraction of the nucleotide positions across the read length, hence reduces the impact of random sequencing errors while maintaining high sensitivity to discriminate closely related but distinct taxa. STAMP software (Parks and Beiko, 2010) was used to observe taxonomical structure variations inferred from 16S rRNA gene amplicon datasets. Furthermore, principal

component and correspondence analyses were performed using the "R" and Ade4TkGUI software packages (Thioulouse et al., 1997). Box plots were generated using R. A Pearson statistical test was used to study the correlation of specific microbial taxa among the datasets. Finally, One-Way ANOVA tests were used to access the significance of community structure shifts observed between groups of samples.

RESULTS

DISTANT LOCALES IN THE SOUTHERN OCEAN HAVE DISTINCT MICROBIAL COMMUNITIES

During the 2010–2011 austral summer, surface water samples were collected at 18 sites (plus 5 near-bottom water samples for a total of 23 samples) across the ASP (**Figure 1**): 3 sites along the Dotson glacier, 2 sites covered by pack ice at the outer fringe of the polynya near the shelf break and 13 sites across a bloom of *Phaeocystis antarctica* (>10 μg Chl *a*.L^{-1}, O_2-saturation at >400%, pCO_2 100–250 ppm) in the open waters of the polynya. Bacterial community structures (0.2–20 μm) were determined by deep sequencing of 16S rRNA amplicons [60–65 nucleotides of the V6 hypervariable region, $>10^5$ reads per sample (**Table S1**)]. Even though our primer sets amplified some Archaea V6 (and thus provided an approximation of their diversity) we did not include these data here. The diversity and low abundance of Archaea in the ASP center (Archaea/Bacteria ratio of about 1/500) was part of a separate study (Kim et al., 2013). Our observations on the microbial community structures focused on the diversity, relative abundance and distribution of bacterial and eukaryotic taxa.

In a first approach we compared the bacterial communities of four randomly selected ASP surface samples [Ant11, Ant13, Ant14, Ant15 (**Table S1**)] with those determined for four surface water samples from the Antarctic Circumpolar Current (ACC) and four samples near the Antarctic Peninsula (Sul et al., 2013). **Figure S1** shows a heat map comparison of these different sites. It is immediately apparent that the microbial community structures had a high degree of similarity within datasets for each of the three locales in the Southern Ocean, but substantial dissimilarity was noted between datasets. The bacterial communities in the ASP were most dissimilar from those at the two other locales. The Bray-Curtis dissimilarity index for samples inside the ASP was only 0.27 ± 0.13. However, the dissimilarity index increased substantially when comparing different locales: 0.54 ± 0.06 between ASP and ACC; 0.64 ± 0.04 between ASP and waters off the Antarctic Peninsula.

A more detailed visualization of the difference among the bacterial communities from the three locales is presented in **Figure 2**. Whereas the deep samples (most diverse) populated the left two quadrants of both the Principal Component Analysis (PCA) and Correspondence Analyses plots, surface samples (less diverse) organized along the central axis with the exception of the samples that represent the surface waters in the center of the polynya where the *P. antarctica* bloom occurred (**Figure 2**, **Figure S2**). The latter samples all grouped together in a cluster distant from the other samples. The axes in the PCA plot account for about 48% of the total variance. These differences were mainly driven by a dominant contribution of *Pelagibacter* (>10% of total), SAR86, members of the Gamma-Proteobacteria and *Rhodobacteraceae* (each at 1–10%) in the ACC samples. In contrast, bacterial communities in surface waters of the central polynya were dominated by a large group of Flavobacteria, Oceanospirillales, SAR92 and *Ulvibacter* (>1% total) with a minor but significant contribution (<1%) made by *Lutibacter*, *Crocinotimix*, *Roseobacter* and members of the Cryomorphaceae. The microbial community that accompanied the *P. antarctica* bloom was different from those near the Dotson glacier or at the outer fringes of the polynya near the shelf break.

MICROBIAL COMMUNITY STRUCTURE INSIDE THE AMUNDSEN SEA POLYNYA

To better understand differences in microbial communities at different locations in the Amundsen Sea we analyzed the taxonomic composition of Eukarya and Bacteria across the polynya. On average, 1.58 ± 0.72 and 0.015 ± 0.009 μg L^{-1} of DNA were extracted from surface and deep samples, respectively. DNA yields were two orders of magnitude lower in the deeper samples, reflecting the lower biomass levels of this size fraction at depth (**Figure 3A**). Eukarya (>99% phytoplankton taxa) were an important fraction in surface waters (**Figure 3B**), most prominent near the Dotson glacier (62 ± 5.8%). Bacteria made up the bulk of the community in deep samples (94 ± 1.8%) where the number of detected species (576 ± 131) significantly increased ($p < 0.001$) in comparison to surface samples (361 ± 51) (**Figure S3**). Among the Eukarya, haptophyte (>99% *Phaeocystis antarctica*) genotypes dominated the phytoplankton bloom in the ASP (72 ± 7.8%) while diatoms (*Bacillariophyta*) were more abundant near the Shelf break and the Dotson glacier (70 ± 12%, see **Figure 3C**). Note that 16S rRNA gene copy number can vary widely between alga species depending on the number of chloroplasts per cell. Therefore, the ratio of Bacteria/Eukarya and haptophyte/diatoms in each dataset are not necessary representative of the plankton community structure. On the other hand, ratio differences observed for the same populations between samples are more likely to reflect shifts in community structure. In particular, we observed a clear shift from a Proteobacteria dominated bacterial community (73 ± 11.2% vs. 50 ± 8.3%, $p < 0.001$) outside the *P. antarctica* bloom to a Bacteroidetes dominated community (47 ± 8% vs. 18 ± 13.1%, $p < 0.001$) inside the bloom. The abundant *Phaeocystis* populations in central waters of the polynya were accompanied by a bacterial community dominated by Flavobacteria (99 ± 0.8% of total Bacteroidetes), and Proteobacteria. Together they contributed >95% of the V6 reads in each of the polynya samples (**Figure 3D**). Inside the polynya the Proteobacteria were dominated by Gammaproteobacteria (73 ± 6.2%, mostly *Oceanospirillum*-like and SAR92) and Alphaproteobacteria (24 ± 6.2%, mostly *Pelagibacter*) with lesser contributions made by Betaproteobacteria (1.2 ± 1.1%, mostly *Methylophilaceae*), Deltaproteobacteria (0.7 ± 0.4%) and few Epsilonproteobacteria (0.0006 ± 0.004%). We also explored a dataset from another bloom event (R/V Oden cruise, 2007–2008) that was sampled along a single depth profile (**Table S1**) and we compared their microbial communities. Four samples that originated from the upper 100 m of the water column and that had Chl a concentrations of >8 μg L^{-1} all revealed microbial community

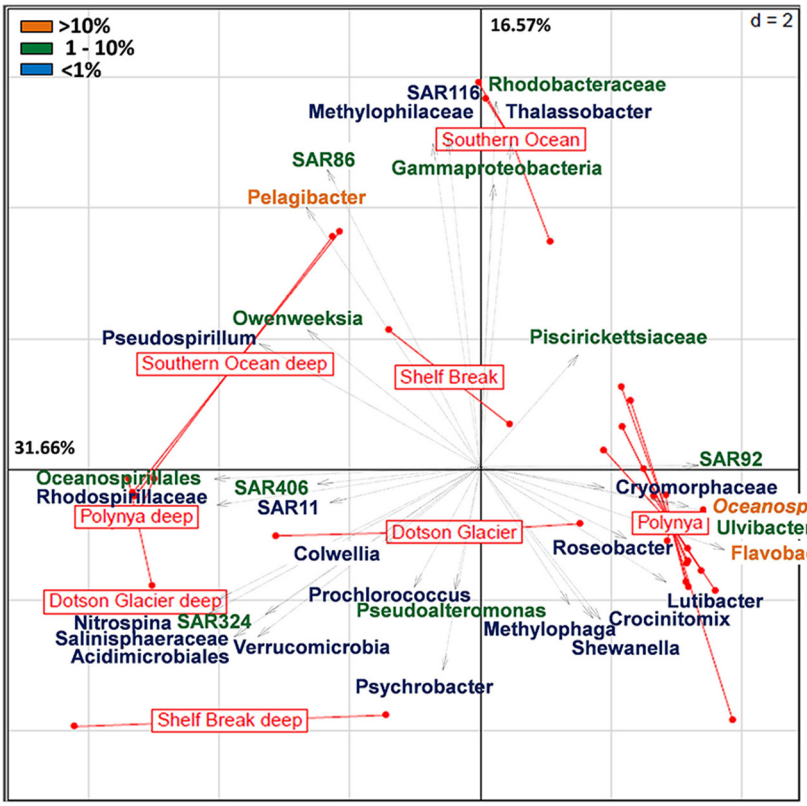


FIGURE 2 | Principal component analysis (euclidean distance) of the relative distribution of the 36 most abundant bacterial taxa across high throughput sequencing datasets from the Southern Ocean (Antarctic Circumpolar Current) and the Amundsen Sea (shelf break, Dotson glacier and polynya). The latter includes surface and deep bacterial communities sampled at the outer rim of the polynya, near the Dotson Glacier and in the open waters of the polynya. Pyrosequencing datasets of the Amundsen Sea polynya representing the 0.2–3 μm size fraction sampled during the 2007–2008 bloom event (4 surface and 3 deep samples; see **Table S1**) were included in the analysis. Orange labels denote taxa that contribute >10% of the bacterial community; green denotes 1–10%; Blue is <1%.

structures that were highly similar to those of the 2010/2011 bloom (see **Figure 2**).

In contrast to the surface samples the bacterial communities in deep samples (**Figure 3D**) were characterized by an increased contribution of Proteobacteria (80 ± 6.1%) and decreasing Bacteroidetes abundances (8.3 ± 2.9%). Both the increase in the prevalence of Proteobacteria and the decrease in Bacteroidetes abundance were significant ($p < 0.01$) in One-Way ANOVA tests. Verrucomicrobia (3.0 ± 1.4%) and Actinobacteria (4.7 ± 1.8%), typically found in deep marine waters (Sogin et al., 2006; Quaiser et al., 2008; Freitas et al., 2012), were significantly more abundant in deep samples (**Figure 3D**). The Proteobacteria were dominated by Gammaproteobacteria (47 ± 14.6%, mostly *Pseudoalteromonas* and *Oceanospirillales*), Alphaproteobacteria (33 ± 15.6%, mostly *Pelagibacter* and *SAR11* related taxa) and Deltaproteobacteria (19 ± 5.7%, mostly *SAR324* and Nitrospina) with a minor contributions made by Betaproteobacteria (0.4 ± 0.3%) and Epsilonproteobacteria (0.2 ± 0.1%) classes. The Delta and Epsilon classes were therefore drastically more represented in deep samples. Three samples from 250 to 785 m depth within the polynya water column and collected during the 2007–2008 bloom event provide similar trends (these samples are part of the "Polynya deep" group in **Figure 2**), with a dominance of *Pelagibacter*, *Oceanospirillales* and *SAR324* genotypes. Altogether, our findings indicate that the surface and deep microbial community structures of the Amundsen Sea polynya were highly similar across the spatial dimensions of the bloom. They were also maintained across temporal scales that exceed bacterial generation times by far. We note that these highly similar bacterial communities were maintained over a 18 day period (19/12/2010 to 05/01/2011), close to the climax of the ∼90 day bloom duration estimated from remote sensing images (Arrigo and Van Dijken, 2003).

OLIGOTYPE DIVERSITY OF BACTERIAL AND EUKARYOTIC TAXA

In order to analyze the bacterial communities in more detail we studied the diversity of selected taxa across the polynya during the 2010–2011 bloom event (**Figures 4–6**). Shannon entropy decomposition or oligotyping (Eren et al., 2013a) was used to track subtle, conserved sequence variations and differentiate genotypes that make up each taxon but differ by as few as 1–7 nucleotides. Among the chloroplast V6 reads, those identified as *Phaeocystis antarctica* were dominated by a single oligotype (>90%) at all locations and depths inside the polynya (**Figure 4**). A few different oligotypes were distinct in surface waters near the shelf break, suggesting the existence of sub-populations of *P. antarctica*

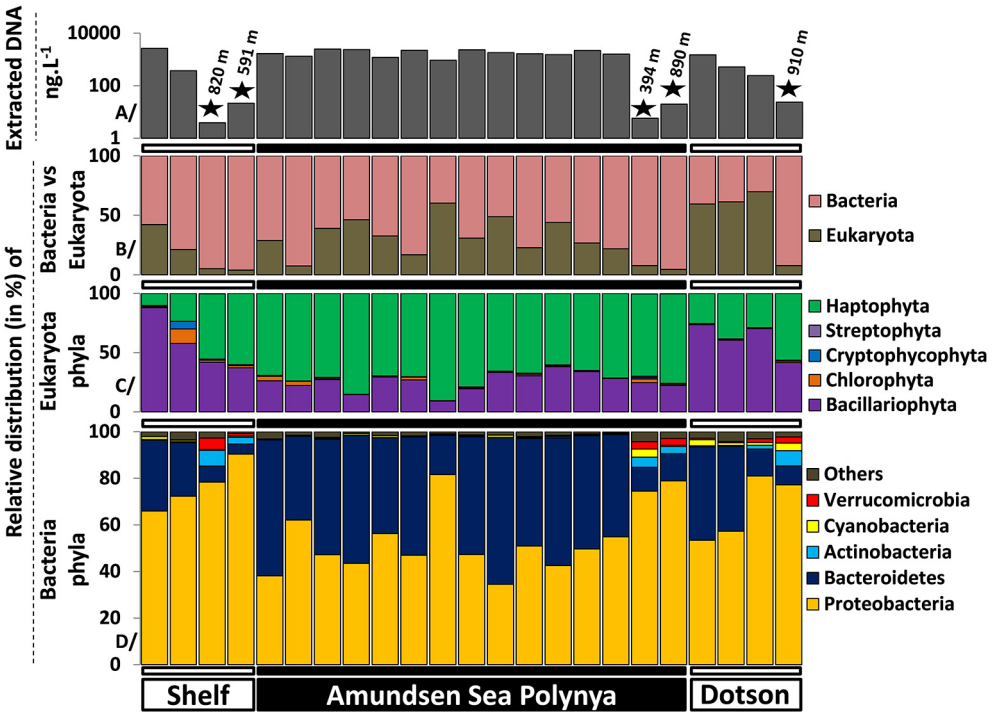


FIGURE 3 | Microbial community structures determined by Illumina sequencing (>10^5 reads per sample) of 16S-V6 rRNA amplicons obtained from surface and deep samples at sites with dense ice cover at the shelf break (Shelf), inside the Amundsen Sea polynya and in open waters adjacent to the Dotson glacier (Dotson) during a *Phaeocystis antarctica* bloom in 2010–2011. (A) (top) denotes the concentration of extracted DNA for each sample as a proxy for microbial biomass; the "★" symbols denote deep (>350 m) samples. The relative contribution of Bacteria vs. Eukaryota (chloroplast 16S) is presented in **(B)**. **(C,D)** present the phytoplankton and bacterioplankton community composition at the phylum level.

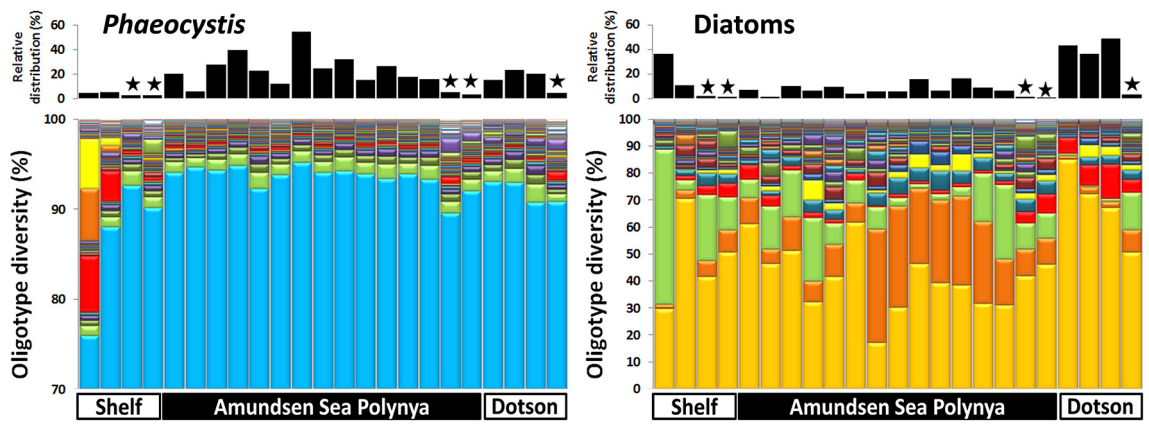


FIGURE 4 | Phytoplankton oligotype diversity of 16S-V6 rRNA amplicon sequences obtained for *Phaeocystis* (left graph) and diatoms (right graph). Surface and deep samples were obtained at sites with dense ice cover at the shelf break (Shelf), inside the Amundsen Sea polynya and in open waters adjacent to the Dotson glacier (Dotson) during a *Phaeocystis antarctica* bloom in 2010–2011. Panels at top of the graphs denote the relative contribution of a taxon within each dataset; the "★" symbols denote deep (>350 m) samples. Oligotypes detected at low abundance ($n < 200$) in the dataset were removed from the analysis.

that do not contribute to bloom formation. Diatom populations were more diverse (**Figure 4**). Among the bacterial V6 reads, the major taxa in surface layers of the polynya were dominated by a single oligotype for SAR92 (97 ± 1.8% of total), *Oceanospirillum*-like bacteria (95 ± 1.4%), *Pelagibacter* (80 ± 7.3%, data not shown) and members of the Flavobacteriaceae (75 ± 14.1%) a family for which V6 sequence do not allow taxonomy assignment below the rank of family (**Figure 5**). These

taxa showed different oligotype diversity patterns in deep samples with the exception of the *Oceanospirillum*-like oligotype diversity that remained strikingly similar despite the strong variation in their relative abundance (0.9–33.9% of the bacterial community) across the datasets. The dominant oligotype related to Flavobacteriaceae could not be resolved taxonomically due to a perfect match between V6 sequences of *Polaribacter* and other members of the Flavobacteriaceae. This oligotype was therefore denoted as *Polaribacter* sensu lato. On the other hand, taxa such as SAR86, *Nitrospina* and *Verrucomicrobia* that were detected in higher abundance outside the bloom (especially in deep samples) were more diverse and lacked a single dominant representative oligotype for either of the niches (**Figure 5**). For SAR86, one oligotype dominated the deep samples while another oligotype was more abundant in the surface waters at the shelf break. The two oligotypes represent bacteria that have so far evaded successful cultivation. *Nitrospina* was equally diverse in all samples except at the shelf break where we observed a few distinct oligotypes. Finally, the diversity among Verrucomicrobia genotypes appeared to be relatively uniform across all datasets with minor variations observed between different locales.

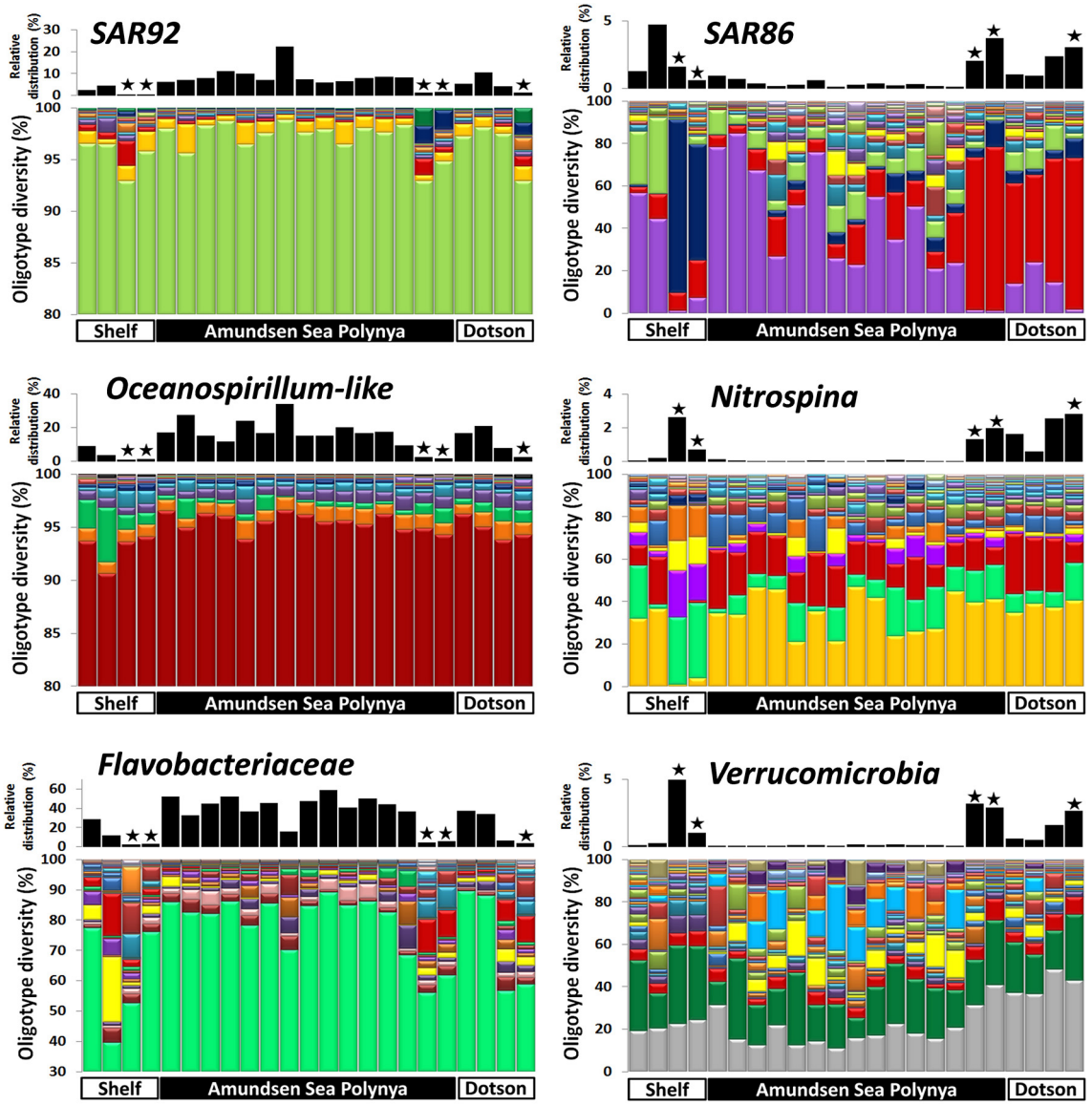


FIGURE 5 | Bacterial oligotype diversity of 16S-V6 rRNA amplicon sequences obtained for taxa that dominated inside (left graphs) and outside (right graph) the *Phaeocystis* surface bloom. Surface and deep samples were obtained at sites with dense ice cover at the shelf break (Shelf), inside the Amundsen Sea polynya and in open waters adjacent to the Dotson glacier (Dotson) during a *Phaeocystis antarctica* bloom in 2010–2011. Panels at top of each graph denote the relative contribution of a taxon within each dataset; the "★" symbols denote deep (>350 m) samples. Oligotypes detected at low abundance ($n < 200$) in the dataset were removed from the analysis.

FREE LIVING vs. PARTICLE ASSOCIATED BACTERIAL TAXA

During the 2007/2008 bloom event in the Amundsen polynya samples from the surface (above 100 m depth, $n = 4$) and at 250 m depth ($n = 1$) were size fractionated. The size fractionation differentiated between bacteria in the 0.2–3 μm (<3-μm) and 3–200 μm (>3-μm) fractions. The <3-μm fraction was thought to be enriched with V6 reads from free-living bacteria, whereas V6 reads in the >3-μm fraction were derived from particle-associated bacteria. The most common particles were *Phaeocystis* solitary cells and colonies along with diatom species that together gave rise to the intense phytoplankton bloom. We observed a substantial increase of *P. antarctica* (from 25 ± 17.1 to 51.2 ± 24.3% of total V6 reads) and a much less pronounced increase in diatom V6 reads (from 1.7 ± 0.8 to 9.5 ± 4%) in the >3-μm fraction of the polynya surface. Presumably, *P. antarctica* was present as single cells in the <3-μm fraction and as small-medium sized colonies in the >3-μm fraction. A taxon-by-taxon comparison of bacterial genotypes between the two size fractions in the surface mixed layer of the polynya identified several taxa that showed a higher relative abundance in either <3-μm or >3-μm fractions (**Figures 6A–C**). For each taxon we calculated the relative enrichment as the ratio of their abundance in the two size fractions using the following equation:

$$\left(\frac{2\text{X}\ (\ >3\ \mu\text{m fraction})}{<3\ \mu\text{m fraction} + >3\ \mu\text{m fraction}} - 1\right) \quad (1)$$

A PCA showed that a large majority of dominant taxa were enriched in the small fraction (**Figure S4**). SAR92 reads were abundant in each fraction but this genus was significantly enriched in the >3-μm fraction. Whereas they contributed ∼13% of the V6 reads in the <3-μm fraction their contribution rose to >35% in the >3-μm fraction (**Figures 6A,B**). Thus, the enrichment ratio of SAR92 (Gammaproteobacteria) was close to 0.5 (**Figure 6C**). Similarly, *Oceanospirillum*-like (Gammaproteobacteria) genotypes were abundant in both the <3-μm and >3-μm fractions, but their enrichment was less pronounced (**Figure 6C**). Several low abundance taxa – a single Firmicute, *Tepidanaerobacter*, and Bacteroidetes genotypes identified as *Haliscomenobacter, Lutibacter*, *Ulvibacter*, and *Cryomorphaceae* showed the same trend as was observed for SAR92 and they were enriched in the >3-μm fraction (**Figure 6C**). In contrast, different *Flavobacteria*, *Oceanospirillales*, and *Piscirickettsiaceae* (Gammaproteobacteria), and *Pelagibacter* (Alphaproteobacteria) together with the low abundance taxa *Acetivibrio* (Firmicutes) and *SAR324* (Deltaproteobacteria), dominated in the <3-μm fraction and they were presumably present mostly as free-living cells (**Figures 6A–C**).

A similar picture of an association of certain bacterial taxa with *Phaeocystis* particles emerged from a further analysis of the 2010/2011 bloom event in the ASP. We observed a significant correlation ($R^2 = 0.75$, $p = 9.5\text{e-}8$) between the relative abundances of *P. antarctica* and those of SAR92 across 23 datasets (**Figure 6D**). Note that relative abundances of SAR92 were normalized to the

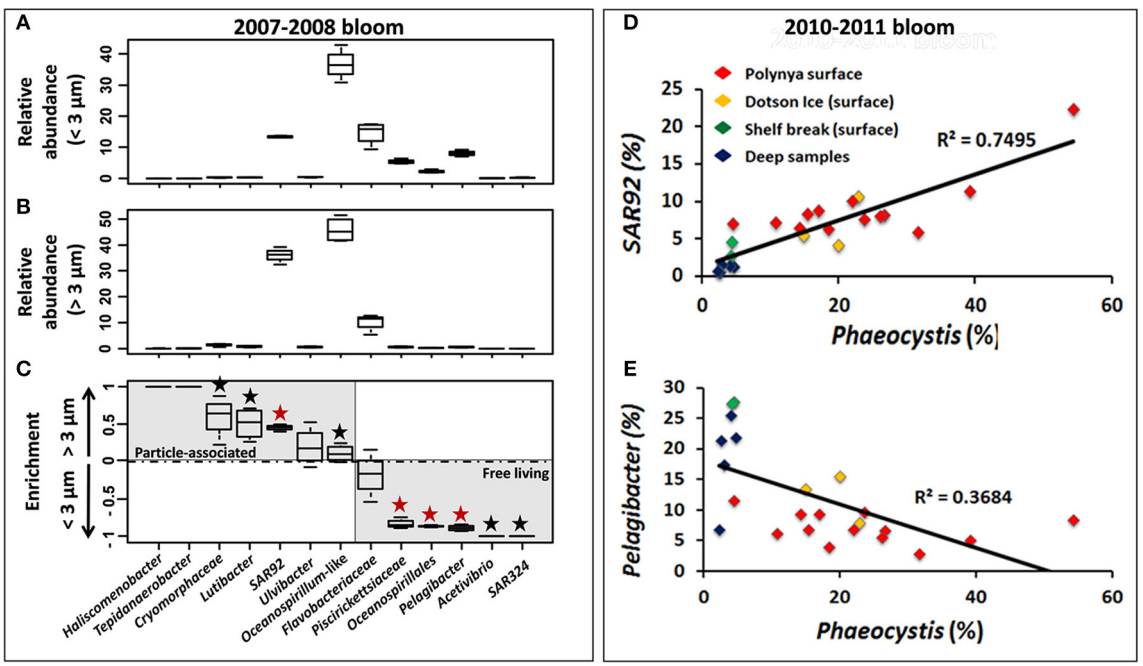


FIGURE 6 | Relative abundance of 13 bacterial taxa in the <3 μm (A) and >3 μm (B) size fractions and preferential enrichment (C) of these taxa in either fraction among *Phaeocystis* bloom samples from surface waters in the Amundsen Sea polynya in 2007–2008. ANOVA test was performed using STAMP software to test the significance of their enrichment in the two size fractions. For each taxa, a black "★" symbol was added when *p*-value score was lower than 0.05 (>95% confidence level). This symbol displayed in red indicates a $p < 0.01$ (>99% confidence level). Correlation between *Phaeocystis* genotypes (percentage of the whole community) and SAR92 **(D)** or *Pelagibacter* **(E)** (percentage of the bacterial community) are displayed across all datasets during the *Phaeocystis* bloom of 2010–2011.

number of bacterial reads in each sample whereas relative abundances of *Phaeocystis* were normalized to total reads (bacteria + chloroplast reads). Assuming a single 16S rRNA gene copy per SAR92 genome (as detected in SAR92 HTCC2207) and the presence of two chloroplasts per alga cell (Moisan et al., 2006), each with two single 16S rRNA gene copy (as detected in the chloroplast genome of *P. antarctica* strain CCMP1374) we estimate an approximate 2:1 occurrence of SAR92 per *Phaeocystis* cell. On the other hand, *Pelagibacter* V6 reads, which were more abundant in the <3-μm fraction (**Figure 6C**), showed a trend of decrease (while not significant) with increasing abundance of *Phaeocystis* (**Figure 6E**), thereby independently confirming the observations made for the 2007/2008 bloom event.

In contrast to their relative abundances in the surface samples, diatoms were more enriched in the >3-μm fraction of the deep sample taken during the 2007/2008 bloom (from 0.8 to 56.6% of the total V6) than *P. antarctica* (from 7 to 21.8%) (data not shown). V6 reads for taxa such as *Pelagibacter*, Oceanospirillales and Piscirickettsiaceae were still more abundant in the <3-μm fraction. However, strong shifts in taxonomic make-up of the >3-μm fraction as compared to the <3-μm fraction occurred (**Figure 7**). SAR92 V6 reads were not enriched in the >3-μm fraction of the sample taken from below the surface mixed layer. Although by no means significant, its relative abundance was slightly higher in the <3-μm fraction (+3.2% of the bacterial community). However, in addition to the omnipresent *Cryomorphaceae* and *Ulvibacter* other taxa became more enriched in the >3 μm fraction. In particular we identified *Colwellia*, *Pseudoalteromonas* and *Cerasicoccus* genotypes that were associated with decaying *Phaeocystis* and/or diatoms, the dominant type of particles in samples below the surface mixed layer. The relative enrichment of these taxa in the >3-μm fraction [as compared to the <3-μm fraction, see Equation (1)] was by a ratio of 0.96 (*Colwellia*), 0.90 (*Pseudoalteromonas*) and 0.98 (*Cerasicoccus*). These relative enrichments were much less pronounced in surface samples with ratios of 0 (±0.09), 0.24 (±0.35), and 0.12 (±0.38) respectively.

DISCUSSION

Whereas persistent blooms of *Phaeocystis antarctica* have been reported for multiple Antarctic polynyas (Arrigo et al., 1999; Smith et al., 2000; Arrigo and Van Dijken, 2003; Alderkamp et al., 2012; Yager et al., 2012) and even in the ACC proper (Alderkamp and van Dijken, pers. comm.), we do not understand all the factors that drive bloom formation and/or support bloom longevity. Previous studies have focused on Fe-limitation of such blooms (Mills et al., 2012) and the role of Fe-supply from glacier melts to polynya surface waters (Alderkamp et al., 2012). Other studies addressed the role of *Phaeocystis* colony formation and control of colony size by grazer populations (Tang et al., 2008). Here we studied the potential for the bacterial flora to play a role in the bloom biology of *P. antarctica*. We have obtained the deepest sequencing of the ASP to date: 10^5–10^6 paired-end (100% overlap) reads for the V6 hypervariable region of 16S rRNA per sample as compared to other studies that report 10^3–10^4 reads for V1 and V3–V4 obtained by pyrosequencing (Kim et al., 2013; Dinasquet et al., submitted; Richert et al., submitted). The complete overlapping sequencing strategy performed here enhanced sequence quality for each V6 read, and so provided highly reliable signatures for the detection of low abundance bacterial populations. Also, using oligotyping we avoided the commonly-used 97% similarity cut-off and partitioned our dataset into homogeneous genotypic units that entail minimal phylogenetic mixture. The single-nucleotide resolution oligotyping achieves allowed us to determine various geographic patterns of bacterial and algal community structure within the confines

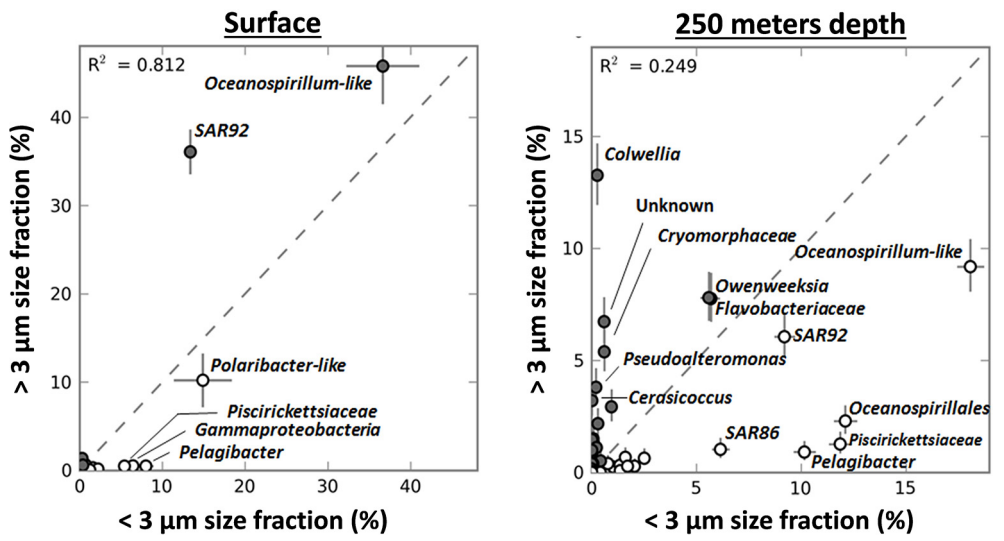


FIGURE 7 | The relative distribution in percentage of genera (based on GAST classification and using STAMP software) in the <3 μm and >3 μm size fractions in samples from surface layers ($n = 8$) and from 250 m depth ($n = 2$) during the *Phaeocystis* bloom in 2007–2008. The coloring of circles reflects the enrichment of the taxa (closed circles when more abundant in the >3 μm size fraction, open circles when more abundant in the <3 μm size fraction).

of this isolated ecosystem. These patterns suggest different niche adaptations. As commonly observed along vertical profiles, depth played a major role in the partitioning of microbial taxa. E.g., diversity of *Nitrospina* genotypes was distinctly different in the deep samples of the shelf break area than elsewhere in the ASP. Oligotyping was also instrumental in identifying partitioning genotype diversity along horizontal gradients, e.g., in determining the diversity of diatom populations (**Figure 4**). These patterns suggest an abundance of niche adaptations for which selective forces and ecological implications are yet to be determined. The main observations from this study are: (1) *P. antarctica* blooms are accompanied by a unique and stable community of free-living bacteria over extended time scales and (2) *Phaeocystis* cells and colonies associate with selected bacterial taxa. We have identified taxa (e.g., SAR92) that accompany productive *Phaeocystis* populations in surface water and different taxa (e.g., *Colwellia*) that are associated with—supposedly decaying—populations of *Phaeocystis* cells at depth, below the illuminated, surface mixed layer.

Antarctic phytoplankton populations under non-bloom conditions typically include *P. antarctica* as one of the dominant species (Yager et al., 2012). Such populations in the ACC, or in coastal waters near the Antarctic Peninsula are accompanied by bacterial populations that are dominated by Proteobacteria, mostly *Pelagibacter* and SAR11-like genotypes (West et al., 2008; Brown et al., 2012; Wilkins et al., 2013a, this study). During bloom events *P. antarctica* often becomes the dominant phytoplankter, most notably in the ASP where such blooms recur annually (Arrigo et al., 1999; Smith et al., 2000; Arrigo and Van Dijken, 2003; Alderkamp et al., 2012; Yager et al., 2012; Kim et al., 2013; Dinasquet et al., submitted). Based on different proxies it has been estimated that the *P. antarctica* blooms contribute >99% of chlorophyll *a*, biomass or cell count. We found that >78% of the V6 reads for chloroplasts (a proxy for relative abundance of *Phaeocystis* cells) were contributed by *P. antarctica* in the <20-μm fraction of polynya samples. Based on chloroplast 16S-V6 we found that the bloom was dominated by a single oligotype. In adjacent waters we discovered a shift in oligotype abundances suggesting that several *Phaeocystis* genotypes did not contribute to the bloom formation in the ASP but had a distinct presence in the waters abutting the polynya.

Since the bulk of the *P. antarctica* bloom was contained in colonies >20-μm this percentage is expected to exceed 78% of the whole phytoplankton community. Whether it is by food web interactions, decay of dead *Phaeocystis* cells, viral lysis, or simple secretion of dissolved organic compounds, *Phaeocystis* blooms can have a large impact on heterotrophic activities and hence shape bacterial communities. We found that members of the Bacteroidetes were most abundant in *Phaeocystis* dominated samples. These observations are in agreement with investigations performed in other locations of the Southern Ocean (Wilkins et al., 2013a; Williams et al., 2013) and support the general standing of this phylum in the specialization of high molecular weight organic matter degradation (Thomas et al., 2011). Members of the Bacteroidetes and Proteobacteria made up 95–97% of the microbial community in the ASP bloom samples. SAR92, *Oceanospirillum-like*, and *Pelagibacter* (Proteobateria), along with *Polaribacter sensu lato* (Bacteroidetes) combined were 73.1% (±5.7) of the bacterial community during the 3 weeks covered by the 2010–2011 cruise. The relative abundances of these taxa are similar to those reported by Kim et al. (2013) for the later stages of the ASP *Phaeocystis* bloom (January–February 2010). In addition, these same taxa dominated the *Phaeocystis* bloom at our sampling site during the summer of 2007–2008. Based on the findings above we suggest that *P. antarctica* blooms are accompanied by stable and distinct microbial communities. Within this community we detected a single, dominant oligotype (>80% of the V6 reads) for each of the dominant taxa (e.g., SAR92, *Oceanospirillum-like*), in contrast with the multiplicity of oligotypes for taxa known from other niches (e.g., SAR86, *Nitrospina*). This observation suggests that specialized ecotypes with conserved genotype signatures co-exist (and possibly interact) with *P. antarctica*. Different phytoplankton species produce different DOM compounds, but closely related species have very similar DOM spectra (Becker et al., 2014). Consistent with this result, blooms of different *Phaeocystis* species have very similar bacterial communities associated with them (Alderkamp et al., 2007) and these bacteria readily degrade labile, presumably low molecular weight carbohydrates produced by these algae (Osinga et al., 1997; Smith et al., 1998; Janse et al., 1999). In addition, polymers excreted by *Phaeocystis* blooms provide a nitrogen rich substrate for heterotrophic bacteria (Solomon et al., 2003) and are expected to induce shifts in microbial community structure. In our study we observed that SAR86 oligotypes (**Figure 5**) associated with *Phaeocystis* blooms were distinct from other SAR86 in adjacent waters with diverse diatom populations as well as in the underlying deep waters. Controlled experimental manipulations and genomic analyses of bacterial metabolisms are needed to better understand the interactions between alga and bacteria and their effects on bacterial community structure.

Biomass produced by *Phaeocystis* blooms is rapidly exported to deeper waters, where cells and colonies become senescent (Ditullio et al., 2000). During the bloom in the ASP in 2010 we detected *Phaeocystis* biomass trapped beneath the surface mixed layer that provides a substrate for microbial degradation. This senescent part of the population was accompanied by a very different microbial community. Contributions by SAR92, Flavobacteriaceae and *Oceanospirillum*-like genotypes were diminished whereas the Gammaproteobacterium SAR86, *Nitrospina* and diverse members of the Verrucomicrobia had become the dominant taxa. The shift in microbial community composition toward Verrucomicrobia and Gammaproteobacteria has been reported for senescent *Phaeocystis* populations (Alderkamp et al., 2007). The increased contribution of *Nitrospina* at depth is likely a result of its role in nitrate formation from ammonium (Luecker et al., 2013) released during the decomposition of senescent *Phaeocystis* populations.

During bloom situations *Phaeocystis* is mostly found as large colonies protected by a semi-permeable membrane (see Schoemann et al., 2005, for a review). In early studies these colonies were thought of as cells within a mucopolysaccharide matrix, but this has been revised to a model where an outer membrane encloses *Phaeocystis* cells within a liquid matrix (Hamm et al., 1999). Indeed, microscopic inspection of *P. antarctica*

colonies showed free-moving and rapidly swimming ciliates within the colony matrix (Delmont, unpublished data). Because of its virtually monotypic blooms and the large cells/colony size *P. antarctica* can be readily enriched by size fractionation. We showed that >3-μm fractions are significantly enriched with the Gammaproteobacteria SAR92. A preliminary estimate indicates that SAR92 and *P. antarctica* cells in surface bloom samples occur in approximately a 2:1 ratio. SAR92 is typically limited by carbon availability and despite carrying proteorhodopsin it does have a photoheterotrophic lifestyle (Stingl et al., 2007). A close association of SAR92 and *P. antarctica* (with SAR92 possibly contained within the colony liquid matrix) could thus be of mutual benefit. Preliminary findings from a metagenome analysis of the 2010-2011 ASP bloom event indicates that SAR92 and other *Phaeocystis* associated bacterial taxa may play a role in sulfur metabolism and iron acquisition via ferrochelatase and siderophore production (Delmont et al., in prep.). SAR92 was abundantly present in phytoplankton bloom samples following a natural occurrence of iron-enrichment in the Southern Ocean (West et al., 2008). This would be especially beneficial if *P. antarctica* would harbor SAR92 within its colony matrix. The fact that this matrix resembles an enclosed aqueous environment allows for rapid diffusion of these secreted compounds and hence efficient usage.

We propose that a mutualistic relationship between *Phaeocystis* and associated bacteria underpins the intensity and longevity of its blooms and thereby sequester substantial amounts of atmospheric carbon dioxide in high-latitude oceans (Smith et al., 1991). Determining the genomic content and activity of associated bacteria will help understanding these mechanisms. Conversely, we determined that a bacterial community dominated by members of the genus *Colwellia* was associated with supposedly senescent *Phaeocystis* cells at depth. Such *Colwellia* are known for the production of extracellular polysaccharides/enzyme complexes involved in the breakdown of high-molecular-weight organic compounds (Methé et al., 2005). Therefore, *Colwellia* and related taxa may play an important role in the recycling of carbon, sulfur and nitrogen by rapidly degrading bloom biomass before their complete sedimentation, a process that can last for more than 8 months (Kirchman et al., 2001b).

ACKNOWLEDGMENTS

This work received financial support from NSF Antarctic Sciences awards ANT-1142095 (Anton F. Post), ANT-0839069 and ANT-0741409 (Patricia L. Yager), and ANT-0839012 (Hugh W. Ducklow). We further acknowledge the support by "Oden Southern Ocean," SWEDARP 2010/2011, a project organized by the Swedish Polar Research Secretariat and National Science Foundation Office of Polar Programs. We are indebted to the Nathaniel B Palmer crew and the Raytheon science crew for their assistance during sampling across the ASP. Drs. Kevin Arrigo and Anne Carlijn Alderkamp (Stanford University) contributed to the maps of the Amundsen Sea shown in **Figure 1**. The 2007/2008 samples were collected with the assistance of K. Bakker and sequenced as part of the International Census of Marine Microbes (ICOMM).We gratefully acknowledge technical assistance provided by A. Murat Eren, Sharon Grim and Joseph H. Vineis during various steps of the sequencing and quality control analyses. We are also grateful to Sara Paver for critical comments and to the reviewers for their constructive remarks.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2014.00646/abstract

Table S1 | Sample designations, metadata and parameters relating to the 16S-V6 sequence datasets used in the analysis underlying Figure 2. Sequence data originated from three different projects and they are publically available on the VAMPS website (http://vamps.mbl.edu/).

Table S2 | Sample designations and environmental metadata relating to the ASPIRE cruise (2010-2011 bloom event).

Figure S1 | Heat map (Bray-Curtis distance) based on the relative distribution of bacterial taxa (based on GAST classification at the genus level) in 3 × 4 samples collected from the Antarctic Circumpolar Current, off the Antarctica peninsula and inside the Amundsen Sea polynya. Dissimilarity between samples is reflected by a color gradient ranging from red (high dissimilarity) to blue (low dissimilarity).

Figure S2 | Correspondence analysis (COA, euclidean distance) of the relative distribution of major bacterial taxa across high throughput sequencing datasets from the Antarctic Circumpolar Current, off the Antarctic Peninsula and the Amundsen Sea polynya. The latter includes surface and deep bacterial communities sampled at the outer rim of the polynya, near the Dotson Glacier and in the open waters of the polynya. Note that the "polynya deep" label is partly covering the "Dotson Glacier deep" label. Pyrosequencing datasets of the Amundsen Sea polynya representing the 0.2–3 μm size fraction sampled during the 2007–2008 bloom event (4 surface and 3 deep samples; see **Table S1**) were included in the analysis.

Figure S3 | (A) represents the number of bacterial species identified in each V6 data from the 2010–2011 bloom event using the Global Assignment of Sequence Taxonomy (GAST) pipeline (Huse et al., 2008) and the SILVA 111 database for reference (Quast et al., 2013). **(B)** represents the number of reads generated for each data set. Datasets represent surface ($n = 18$) and deep samples ($n = 5$).

Figure S4 | Principal component analysis (PCA) of the relative abundance of major bacterial taxa (based on GAST classification) in the <3 μm (A) and >3 μm (B) size fractions of surface and deep samples of the Amundsen polynya during a *Phaeocystis* bloom in 2007–2008 by comparing free living bacteria (0.2–3-μm size fraction, $n = 4$) and alga/particulate associated bacteria (>3-μm size fraction, $n = 4$).

REFERENCES

Alderkamp, A.-C., Buma, A. G., and Van Rijssel, M. (2007). The carbohydrates of *Phaeocystis* and their degradation in the microbial food web. *Biogeochemistry* 83, 99–118. doi: 10.1007/s10533-007-9078-2

Alderkamp, A. C., Mills, M. M., Van Dijken, G. L., Laan, P., Thuróczy, C. E., Gerringa, L. J., et al. (2012). Iron from melting glaciers fuels phytoplankton blooms in the Amundsen Sea (Southern Ocean): phytoplankton characteristics and productivity. *Deep Sea Res. II* 71, 32–48. doi: 10.1016/j.dsr2.2012.03.005

Arrigo, K. R., Robinson, D. H., Worthen, D. L., Dunbar, R. B., Ditullio, G. R., Vanwoert, M., et al. (1999). Phytoplankton community structure and the drawdown of nutrients and CO_2 in the Southern Ocean. *Science* 283, 365–367. doi: 10.1126/science.283.5400.365

Arrigo, K. R., and Van Dijken, G. L. (2003). Phytoplankton dynamics within 37 Antarctic coastal polynya systems. *J. Geophys. Res.* 108, 1–18. doi: 10.1029/2002JC001739

Arrigo, K. R., Worthen, D., Schnell, A., and Lizotte, M. P. (1998). Primary production in Southern Ocean waters. *J. Geophys. Res. Oceans (1978–2012)* 103, 15587–15600. doi: 10.1029/98JC00930

Becker, J. W., Berube, P. M., Follett, C. L., Waterbury, J. B., Chisholm, S. W., Delong, E. F., et al. (2014). Closely related phytoplankton species produce similar suites of dissolved organic matter. *Front. Microbiol.* 5:111. doi: 10.3389/fmicb.2014.00111

Behrenfeld, M. J., and Boss, E. S. (2014). Resurrecting the ecological underpinnings of ocean plankton blooms. *Annu. Rev. Mar. Sci.* 6, 16.11–16.28. doi: 10.1146/annurev-marine-052913-021325

Bertrand, E. M., Saito, M. A., Jeon, Y. J., and Neilan, B. A. (2011b). Vitamin B_{12} biosynthesis gene diversity in the Ross Sea: the identification of a new group of putative polar B_{12}biosynthesizers. *Environ. Microbiol.* 13, 1285–1298. doi: 10.1111/j.1462-2920.2011.02428.x

Bertrand, E. M., Saito, M. A., Lee, P. A., Dunbar, R. B., Sedwick, P. N., and Ditullio, G. R. (2011a). Iron limitation of a springtime bacterial and phytoplankton community in the Ross Sea: implications for vitamin B_{12} nutrition. *Front. Microbiol.* 2:160. doi: 10.3389/fmicb.2011.00160

Brown, M. V., Lauro, F. M., Demaere, M. Z., Muir, L., Wilkins, D., Thomas, T., et al. (2012). Global biogeography of SAR11 marine bacteria. *Mol. Syst. Biol.* 8, 1–13. doi: 10.1038/msb.2012.28

Capone, D. G., Burns, J. A., Montoya, J. P., Subramaniam, A., Mahaffey, C., Gunderson, T., et al. (2005). Nitrogen fixation by *Trichodesmium* spp.: an important source of new nitrogen to the tropical and subtropical North Atlantic Ocean. *Global Biogeochem. Cycles* 19:GB2024. doi: 10.1029/2004GB002331

Capone, D. G., Zehr, J. P., Paerl, H. W., Bergman, B., and Carpenter, E. J. (1997). *Trichodesmium*, a globally significant marine cyanobacterium. *Science* 276, 1221–1229. doi: 10.1126/science.276.5316.1221

Carlson, C. A., Ducklow, H. W., and Hansel, D. A. (1998). Organic carbon partitioning during spring phytoplankton blooms in the Ross Sea polynya and the Sargasso Sea. *Oceanography* 43, 375–386.

Chen, J. F., Xu, N., Jiang, T. J., Wang, Y., Wang, Z. H., and Qi, Y. Z. (1999). A report of *Phaeocystis* globosa bloom in coastal water of Southeast China. *J. Jinan Univ. Nat. Sci. Med. Ed.* 20, 124–129.

Cole, J. J. (1982). Interactions between bacteria and algae in aquatic ecosystems. *Annu. Rev. Ecol. Syst.* 13, 291–314. doi: 10.1146/annurev.es.13.110182.001451

Croft, M. T., Lawrence, A. D., Raux-Deery, E., Warren, M. J., and Smith, A. G. (2005). Algae acquire vitamin B_{12} through a symbiotic relationship with bacteria. *Nature* 438, 90–93. doi: 10.1038/nature04056

Ditullio, G., Grebmeier, J., Arrigo, K., Lizotte, M., Robinson, D., Leventer, A., et al. (2000). Rapid and early export of *Phaeocystis antarctica* blooms in the Ross Sea, Antarctica. *Nature* 404, 595–598. doi: 10.1038/35007061

Doucette, G. J. (1995). Interactions between bacteria and harmful algae: a review. *Nat. Toxins* 3, 65–74. doi: 10.1002/nt.2620030202

Ducklow, H. W. (2003). Seasonal production and bacterial utilization of DOC in the Ross Sea, Antarctica. *Antarct. Res. Ser.* 78, 143–157. doi: 10.1029/078ARS09

Eren, A. M., Maignien, L., Sul, W.-J., Murphy, L. G., Grim, S. L., Morrison, H. G., et al. (2013a). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* 4, 1111–1119. doi: 10.1111/2041-210X.12114

Eren, A. M., Vineis, J. H., Morrison, H. G., and Sogin, M. L. (2013b). A filtering method to generate high quality short reads using Illumina paired-end technology. *PLoS ONE* 8:e66643. doi: 10.1371/journal.pone.0066643

Freitas, S., Hatosy, S., Fuhrman, J. A., Huse, S. M., Welch, D. B. M., Sogin, M. L., et al. (2012). Global distribution and diversity of marine Verrucomicrobia. *ISME J.* 6, 1499–1505. doi: 10.1038/ismej.2012.3

Ghiglione, J.-F., Galand, P. E., Pommier, T., Pedros-Alio, C., Maas, E. W., Bakker, K., et al. (2012). Pole-to-pole biogeography of surface and deep marine bacterial communities. *Proc. Natl. Acad. Sci. U.S.A.* 109, 17633–17638. doi: 10.1073/pnas.1208160109

Grzymski, J. J., Riesenfeld, C. S., Williams, T. J., Dussaq, A. M., Ducklow, H., Erickson, M., et al. (2012). A metagenomic assessment of winter and summer bacterioplankton from Antarctica Peninsula coastal surface waters. *ISME J.* 6, 1901–1915. doi: 10.1038/ismej.2012.31

Hamm, C. E., Simson, D. A., Merkel, R., and Smetacek, V. (1999). Colonies of *Phaeocystis* globosa are protected by a thin but tough skin. *Mar. Ecol. Prog. Ser.* 187, 101–111. doi: 10.3354/meps187101

Hmelo, L. R., Van Mooy, B. A. S., and Mincer, T. J. (2012). Characterization of bacterial epibionts on the cyanobacterium Trichodesmium. *Aquat. Microb. Ecol.* 67, 1–14.

Huse, S. M., Dethlefsen, L., Huber, J. A., Welch, D. M., Relman, D. A., and Sogin, M. L. (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet.* 4:e1000255. doi: 10.1371/journal.pgen.1000255

Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., and Welch, D. M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* 8:R143. doi: 10.1186/gb-2007-8-7-r143

Janse, I., Van Rijssel, M., Ottema, A., and Gottschal, J. C. (1999). Microbial breakdown of *Phaeocystis* mucopolysaccharides. *Limnol. Oceanogr.* 44, 1447–1457. doi: 10.4319/lo.1999.44.6.1447

Kim, J.-G., Park, S.-J., Cha, I.-T., Kim, K.-H., Yang, E.-J., Kim, Y.-N., et al. (2013). Unveiling abundance and distrbituion of planktonic Bacteria and Archaea in a polynya in Amundsen Sea, Antarctica. *Environ. Microbiol.* 16, 1566–1578. doi: 10.1111/1462-2920.12287

Kirchman, D. L., Meon, B., Ducklow, H. W., Carlson, C. A., Hansell, D. A., and Steward, G. F. (2001a). Glucose fluxes and concentrations of dissolved combined neutral sugars (polysaccharides) in the Ross Sea and Polar Front Zone, Antarctica. *Deep Sea Res. II* 48, 4179–4197. doi: 10.1016/S0967-0645(01)00085-6

Kirchman, D. L., Meon, B., Ducklow, H. W., Carlson, C. A., Hansell, D. A., and Steward, G. F. (2001b). Glucose fluxes and concentrations of dissolved combined neutral sugars (polysaccharides) in the Ross Sea and Polar Front Zone, Antarctica. *Deep Sea Res. II* 48, 4179–4197. doi: 10.1016/S0967-0645(01)00085-6

Luecker, S., Nowka, B., Rattei, T., Spieck, E., and Daims, H. (2013). The genome of Nitrospina gracilis illuminates the metabolism and evolution of the major marine nitrite oxidizer. *Front. Microbiol.* 4:27. doi: 10.3389/fmicb.2013.00027

Martin, J. H., Fitzwater, S. E., and Gordon, R. M. (1990). Iron deficiency limits phytoplankton growth in Antarctic waters. *Global Biogeochem. Cycles* 4, 5–12. doi: 10.1029/GB004i001p00005

Methé, B. A., Nelson, K. E., Deming, J. W., Momen, B., Melamud, E., Zhang, X., et al. (2005). The psychrophilic lifestyle as revealed by the genome sequence of Colwellia psychrerythraea 34H through genomic and proteomic analyses. *Proc. Natl. Acad. Sci. U.S.A.* 102, 10913–10918. doi: 10.1073/pnas.0504766102

Mills, M. M., Alderkamp, A. C., Thuróczy, C. E., Van Dijken, G. L., Laan, P., De Baar, H. J., et al. (2012). Phytoplankton biomass and pigment responses to Fe amendments in the Pine Island and Amundsen polynyas. *Deep Sea Res. II* 71, 61–76. doi: 10.1016/j.dsr2.2012.03.008

Moisan, T. A., Ellisman, M. H., Buitenhuys, C. W., and Sosinsky, G. E. (2006). Differences in chloroplast ultrastructure of *Phaeocystis antarctica* in low and high light. *Mar. Biol.* 149, 1281–1290. doi: 10.1007/s00227-006-0321-5

Osinga, R., De Vries, K. A., Lewis, W. E., Van Raaphorst, W., Dijkhuizen, L., and Van Duyl, F. C. (1997). Aerobic degra-dation of phytoplankton debris dominated by *Phaeocystis* sp. in different physiological stages of growth. *Aquat. Microb. Ecol.* 12, 11–19. doi: 10.3354/ame012011

Parks, D. H., and Beiko, R. G. (2010). Identifying biologically relevant differences between metagenomic communities. *Bioinformatics* 26, 715–721. doi: 10.1093/bioinformatics/btq041

Piquet, A. M., Bolhuis, H., Meredith, M. P., and Buma, A. G. (2011). Shifts in coastal Antarctic marine microbial communities during and after melt water-related surface stratification. *FEMS Microbiol. Ecol.* 76, 413–427. doi: 10.1111/j.1574-6941.2011.01062.x

Quaiser, A., López-García, P., Zivanovic, Y., Henn, M. R., Rodriguez-Valera, F., and Moreira, D. (2008). Comparative analysis of genome fragments of Acidobacteria from deep Mediterranean plankton. *Environ. Microbiol.* 10, 2704–2717. doi: 10.1111/j.1462-2920.2008.01691.x

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596. doi: 10.1093/nar/gks1219

Rousseau, V., Becquevort, S., Parent, J. Y., Gasparini, S., Daro, M. H., Tackx, M., et al. (2000). Trophic efficiency of the planktonic food web in a coastal ecosystem dominated by *Phaeocystis* colonies. *J. Sea Res.* 43, 357–372. doi: 10.1016/S1385-1101(00)00018-6

Schoemann, V., Becquevort, S., Stefels, J., Rousseau, V., and Lancelot, C. (2005). *Phaeocystis* blooms in the global ocean and their controlling mechanisms: a review. *J. Sea Res.* 53, 43–66. doi: 10.1016/j.seares.2004.01.008

Seymour, J. R., Simo, R., Ahmed, T., and Stocker, R. (2010). Chemoattraction to dimethylsulfoniopropionate throughout the marine microbial food web. *Science* 329, 342–345. doi: 10.1126/science.1188418

Sher, D., Thompson, J. W., Kashtan, N., Croal, L., and Chisholm, S. W. (2011). Response of *Prochlorococcus* ecotypes to co-culture with diverse marine bacteria. *ISME J.* 5, 1125–1132. doi: 10.1038/ismej.2011.1

Sinigalliano, C. D., Gidley, M., Shibata, T., Whitman, D., Dixon, T., Laws, E., et al. (2007). Impacts of Hurricanes Katrina and Rita on the microbial landscape of the New Orleans area. *Proc. Natl. Acad. Sci. U.S.A.* 104, 9029–9034. doi: 10.1073/pnas.0610552104

Smith, W. O., Jr., Carlson, C. A., Ducklow, H. W., and Hansell, D. A. (1998). Growth dynamics of *Phaeocystis antarctica*-dominated plankton assemblages from the Ross Sea. *Mar. Ecol. Prog. Ser.* 168, 229–244. doi: 10.3354/meps168229

Smith, W. O. Jr, Codispoti, L. A., Nelson, D. M., Manley, T., Buskey, E. J., Niebauer, H. J., et al. (1991). Importance of Phaeocystis blooms in the high-latitude ocean carbon cycle. *Nature* 352, 514–516.

Smith, W. O. Jr., Dennett, M. R., Mathot, S., and Caron, D. A. (2003). The temporal dynamics of the flagellated and colonial stages of *Phaeocystis antarctica* in the Ross Sea. *Deep Sea Res. II* 50, 605–617. doi: 10.1016/S0967-0645(02)00586-6

Smith, W. O. Jr., Marra, J., Hiscock, M. R., and Barber, R. T. (2000). The seasonal cycle of phytoplankton biomass and primary productivity in the Ross Sea, Antarctica. *Deep-Sea Res. II* 47, 3119–3140. doi: 10.1016/S0967-0645(00)00061-8

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., et al. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere. *Proc. Natl. Acad. Sci. U.S.A.* 103, 12115–12120. doi: 10.1073/pnas.0605127103

Solomon, C. M., Lessard, E. J., Keil, R. G., and Foy, M. S. (2003). Characterization of extracellular polymers of *Phaeocystis globosa* and *P. antarctica*. *Mar. Ecol. Prog. Ser.* 250, 81–89. doi: 10.3354/meps250081

Stingl, U., Desiderio, R. A., Cho, J.-C., Vergin, K. L., and Giovannoni, S. J. (2007). The SAR92 clade: an abundant coastal clade of culturable marine bacteria possessing proteorhodopsin. *Appl. Environ. Microbiol.* 73, 2290–2296. doi: 10.1128/AEM.02559-06

Stocker, R., and Seymour, J. R. (2012). Ecology and physics of bacterial chemotaxis in the ocean. *Microbiol. Mol. Biol. Rev.* 76:792. doi: 10.1128/MMBR.00029-12

Stocker, R., Seymour, J. R., Samadani, A., Hunt, D. H., and Polz, M. F. (2008). Rapid chemotactic response enables marine bacteria to exploit ephemeral microscale nutrient patches. *Proc. Natl. Acad. Sci. U.S.A.* 105, 4209–4214. doi: 10.1073/pnas.0709765105

Subramaniam, A., Yager, P. L., Carpenter, E. J., Mahaffey, C., Bjorkman, K., Cooley, S., et al. (2008). Amazon river enhances diazotrophy and carbon sequestration in the tropical North Atlantic Ocean. *Proc. Natl. Acad. Sci. U.S.A.* 105, 10460–10465. doi: 10.1073/pnas.0710279105

Sul, W. J., Oliver, T. A., Ducklow, H. W., Amaral-Zettler, L. A., and Sogin, M. L. (2013). Marine bacteria exhibit a bipolar distribution. *Proc. Natl. Acad. Sci. U.S.A.* 110, 2342–2347. doi: 10.1073/pnas.1212424110

Tang, K. W., Smith, W. O. Jr., Elliott, D. T., and Shields, A. R. (2008). Colony size of *Phaeocystis antarctica* (Prymnesiophyceae) as influenced by zooplankton grazers. *J. Phycol.* 44, 1372–1378. doi: 10.1111/j.1529-8817.2008.00595.x

Teeling, H., Fuchs, B. M., Becher, D., Klockow, C., Gardebrecht, A., Bennke, C. M., et al. (2012). Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. *Science* 336, 608–611. doi: 10.1126/science.1218344

Thioulouse, J., Chessel, D., Dole, S., and Olivier, J.-M. (1997). ADE-4: a multivariate analysis and graphical display software. *Stat. Comput.* 7, 75–83. doi: 10.1023/A:1018513530268

Thomas, F., Hehemann, J.-H., Rebuffet, E., Czjzek, M., and Michel, G. (2011). Environmental and gut bacteroidetes: the food connection. *Front. Microbiol.* 2:93. doi: 10.3389/fmicb.2011.00093

Turner, J., Colwell, S. R., Marshall, G. J., Lachlan-Cope, T. A., Carleton, A. M., Jones, P. D., et al. (2005). Antarctic climate change during the last 50 years. *Int. J. Climatol.* 25, 279–294. doi: 10.1002/joc.1130

Van Boekel, J. S. W., and Stefels, W. H. M. (1993). Production of DMS from dissolved DMSP in axenic cultures of the marine phytoplankton species *Phaeocystis* sp. *Mar. Ecol. Prog. Ser.* 97, 11–18. doi: 10.3354/meps097011

Van Mooy, B. A., Hmelo, L. R., Sofen, L. E., Campagna, S. R., May, A. L., Dyhrman, S. T., et al. (2011). Quorum sensing control of phosphorus acquisition in *Trichodesmium* consortia. *ISME J.* 6, 422–429. doi: 10.1038/ismej.2011.115

Vogt, M., O'Brien, C., Peloquin, J., Schoemann, V., Breton, E., Estrada, M., et al. (2012). Global marine plankton functional type biomass distributions: *Phaeocystis* spp. *Earth Syst. Sci. Data* 4, 107–120. doi: 10.5194/essd-4-107-2012

West, N. J., Obernosterer, I., Zemb, O., and Lebaron, P. (2008). Major differences of bacterial diversity and activity inside and outside of a natural iron-fertilized phytoplankton bloom in the Southern Ocean. *Environ. Microbiol.* 10, 738–756. doi: 10.1111/j.1462-2920.2007.01497.x

Wilkins, D., Lauro, F. M., Williams, T. J., Demaere, M. Z., Brown, M. V., Hoffman, J. M., et al. (2013a). Biogeographic partitioning of Southern Ocean microorganisms revealed by metagenomics. *Environ. Microbiol.* 15, 1318–1333. doi: 10.1111/1462-2920.12035

Wilkins, D., Yau, S., Williams, T. J., Allen, M. A., Brown, M. V., Demaere, M. Z., et al. (2013b). Key microbial drivers in Antarctic aquatic environments. *FEMS Microbiol. Rev.* 37, 303–335. doi: 10.1111/1574-6976.12007

Williams, T. J., Wilkins, D., Long, E., Evans, F., Demaere, M. Z., Raftery, M. J., et al. (2013). The role of planktonic Flavobacteria in processing algal organic matter in coastal East Antarctica revealed using metagenomics and metaproteomics. *Environ. Microbiol.* 15, 1302–1317. doi: 10.1111/1462-2920.12017

Wolf, C., Frickenhaus, S., Kilias, E. S., Peeken, I., and Metfies, K. (2013). Regional variability in eukaryotic protist communities in the Amundsen Sea. *Antarct. Sci.* 25, 741–751. doi: 10.1017/S0954102013000229

Yager, P. L., Sherrell, L., Stammerjohn, S. E., Alderkamp, A.-C., Schofield, O., Abrahamsen, E. P., et al. (2012). ASPIRE: the Amundsen Sea Polynya international research expedition. *Oceanography* 25, 40–53. doi: 10.5670/oceanog.2012.73

Zingone, A., Chrétiennot-Dinet, M. J., Lange, M., and Medlin, L. (1999). Morphological and genetic characterization of *Phaeocystis cordata* and *P. jahnii* (Prymnesiophyceae), two new species from the Mediterranean Sea. *J. Phycol.* 35, 1322–1337. doi: 10.1046/j.1529-8817.1999.3561322.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 06 July 2014; accepted: 07 November 2014; published online: 19 December 2014.

*Citation: Delmont TO, Hammar KM, Ducklow HW, Yager PL and Post AF (2014) Phaeocystis antarctica blooms strongly influence bacterial community structures in the Amundsen Sea polynya. Front. Microbiol. **5**:646. doi: 10.3389/fmicb.2014.00646*

This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology.

frontiers in
MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 05 January 2015
doi: 10.3389/fmicb.2014.00660

Biogeographic patterns of bacterial microdiversity in Arctic deep-sea sediments (HAUSGARTEN, Fram Strait)

***Pier Luigi Buttigieg*[1,2] * *and Alban Ramette*[3†]**

[1] *Hinrichs Lab, Organic Geochemistry Department, MARUM – Center for Marine Environmental Sciences, Bremen, Germany*
[2] *HGF-MPG Bridge-Group for Deep Sea Ecology and Technology, Alfred-Wegener-Institut, Helmholtz-Zentrum für Polar- und Meeresforschung, Bremerhaven, Germany*
[3] *Max Planck Institute for Marine Microbiology, HGF-MPG Bridge-Group for Deep Sea Ecology and Technology, Bremen, Germany*

Edited by:
A. Murat Eren, Marine Biological Laboratory, USA

Reviewed by:
Jennifer F. Biddle, University of Delaware, USA
Christopher Kenneth Algar, Marine Biological Laboratory, USA
Christopher Quince, University of Warwick, UK

****Correspondence:***
Pier Luigi Buttigieg, c/o Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany
e-mail: pbuttigi@mpi-bremen.de

†Present address:
Alban Ramette, Institute of Social and Preventive Medicine, University of Bern, Finkenhubelweg 11, 3012 Bern, Switzerland
e-mail: ramette@ispm.unibe.ch

Marine bacteria colonizing deep-sea sediments beneath the Arctic ocean, a rapidly changing ecosystem, have been shown to exhibit significant biogeographic patterns along transects spanning tens of kilometers and across water depths of several thousand meters (Jacob et al., 2013). Jacob et al. (2013) adopted what has become a classical view of microbial diversity – based on operational taxonomic units clustered at the 97% sequence identity level of the 16S rRNA gene – and observed a very large microbial community replacement at the HAUSGARTEN Long Term Ecological Research station (Eastern Fram Strait). Here, we revisited these data using the oligotyping approach and aimed to obtain new insight into ecological and biogeographic patterns associated with bacterial microdiversity in marine sediments. We also assessed the level of concordance of these insights with previously obtained results. Variation in oligotype dispersal range, relative abundance, co-occurrence, and taxonomic identity were related to environmental parameters such as water depth, biomass, and sedimentary pigment concentration. This study assesses ecological implications of the new microdiversity-based technique using a well-characterized dataset of high relevance for global change biology.

Keywords: HAUSGARTEN, oligotyping, deep sea sediments, Arctic LTER, taxonomic resolution

INTRODUCTION

Ecological analyses are typically concerned with gauging the response of a collection of organisms, grouped into coherent units such as species, to the biotic and abiotic factors affecting them. Establishing meaningful units of bacterial diversity is an ongoing challenge in the microbial sciences (Cohan, 2001, 2002; Kopac and Cohan, 2011; McDonald et al., 2013; Mende et al., 2013) and the nature of these units has been shown to strongly influence the outcomes of ecological analyses (see e.g., Koeppel and Wu, 2014). An approach that has become a standard in microbial ecology relies on the classification of organisms into units based on the level of sequence identity between their 16S rRNA genes. At the more granular end of this classification, organisms that have 16S rRNA gene sequences that are at least 97–98% identical are grouped into operational taxonomic units (OTUs) which are treated as approximations of bacterial 'species' in further analyses. However, it has been shown that the organisms grouped into a single OTU, at times with identical 16S sequences, can show ecologically meaningful genetic and physiological differences, allowing them to colonize distinct niches (e.g., Moore et al., 1998; Hahn and Pöckl, 2005; Coleman et al., 2006).

While alternative differentiae must be sought for organisms with identical 16S genes, the entropy-based method of "oligotyping" (Eren et al., 2013; not to be confused with oligotyping *sensu* Tiercy et al., 1990) offers an approachable means to detect whether position-specific, subtle sequence variation at up to single-nucleotide resolution can reveal coherent, sub-OTU groupings with differential occurrence across samples or responses to environmental factors. This technique has been applied in investigations of human-associated microbes, such as those that compose the oral (Eren et al., 2014a) and gut (Eren et al., 2014b) microbiomes, as well as of aquatic (Eren et al., 2013) and wastewater environments (McLellan et al., 2013), and in the assessment of *Gardnerella vaginalis* diversity (Eren et al., 2011). Such studies have revealed that subtle nucleotide variations can, reproducibly, be associated with distinct environments, hosts, or epidemiological states and encourage the exploration of oligotype-based microdiversity in similar sequenced-based datasets.

Here, we employed oligotyping to reanalyze data from a previous investigation (Jacob et al., 2013) which assessed biogeographic patterns of deep-sea, benthic bacterial diversity at the Long Term Ecological Research (LTER) station, HAUSGARTEN in the Eastern Fram strait (Soltwedel et al., 2005). This LTER comprises two transects, one bathymetric (water depths between ~1000 and ~5500 m) and one latitudinal (at a depth of ~2500 m), intersecting at a central site. At this station, heat- and nutrient-laden Atlantic waters carried by the West Spitsbergen Current flow northward into the Arctic, separated from the cold Eastern Greenland Current by the East Greenland Polar Front. When present, sea ice attenuates light input and, hence, under-ice primary productivity; however, phytoplankton blooms and phytodetritus pulses

occur along melting ice-edges where primary producer communities in the ice are released into the irradiated and meltwater-stabilized water column (Schewe and Soltwedel, 2003; Leu et al., 2011; Boetius et al., 2013). The organic and inorganic detritus supplied to the benthos is of varying composition, either produced in the photic zone of the water column or transported by physical processes such as advection or sea ice rafting (Hebbeln, 2000; Bauerfeind et al., 2009). Due to remineralization processes in the water column, phytodetritus availability decreases with increasing water depth, producing a depth-related gradient in this key component of benthic food supply. Within this system, prokaryotic communities are responsible for over 90% of the respiration performed in a food web sensitive to changes in labile detritus input (van Oevelen et al., 2011). In recent years, notable changes in the system's oceanography, biogeochemistry, and biology have been reported. For example, anomalously warm Atlantic inflows from 2005 to 2007 impacted the composition of the detritus exported to the benthos: reduced export of particulate carbon, zooplankton fecal pellet carbon, and biogenic silica suggested a shift in the composition of phytoplankton communities to favor small, non-siliceous organisms (Piechura and Walczowski, 2009; Lalande et al., 2013). Additionally, changes in Arctic ice dynamics and the loss of multi-year ice – along with its resident, ice-associated communities – are expected to impact biological input to this system, reducing benthic–pelagic coupling (Hop et al., 2006) as observed in other regions of the Arctic (Grebmeier et al., 2006).

Within this context, Jacob et al. (2013) sampled undisturbed sediments along the HAUSGARTEN bathymetric transect (HGI-HGVI; with a depth range of 1284–3535 m along 54 km) and latitudinal transect (N1–N4, HGIV, and S1–S3; 78.608–79.717 N, at a depth of ∼2500 m along 123 km) during July 2009. The authors examined bacterial communities present in the oxic, upper centimeter of the sediment surface. The authors clustered sequences of the 16S rRNA gene's V4–V6 region into OTUs at the conventional sequence identity threshold of 97%. They then derived matrices of OTU relative abundances at each site. Jacob et al. (2013) investigated the response of bacterial diversity, community structure, and spatial turnover across taxonomic levels and found water depth to be a central explanatory parameter, in line with findings on a global scale (Zinger et al., 2011) and in other regions of the Arctic (Bienhold et al., 2012). To assess if subtle nucleotide variation can reveal finer-grained variation in this data, we oligotyped several, abundant OTUs detected in the Jacob et al. (2013) study and (1) examined the degree of separation and/or aggregation of intra-OTU oligotypes across sites, (2) assessed the influence of environmental and spatial variables on oligotype variation, and (3) examined the composition and structure of oligotype association networks, inferred by co-occurrence across both transects. Through these analyses, we aimed to explore oligotyping's potential as a means to enhance the characterization of bacterial diversity at HAUSGARTEN.

MATERIALS AND METHODS

SEQUENCE DATA PROCESSING AND OLIGOTYPING

Sequences obtained by 454 pyrosequencing of the 16S rRNA gene's V4–V6 region (n = 145,938) were previously trimmed and denoised by Jacob et al. (2013) using *mothur* (Schloss et al., 2009). We submitted these trimmed and denoised sequences to the SILVAngs pipeline (v1.0; Quast et al., 2013) using the pipeline's default parameters – save for an OTU clustering threshold of 97% sequence identity – and quality filtering measures. As pyrosequencing-derived reads of varying length were used in this study, alignments were performed by the SILVA incremental aligner (SINA v1.2.10 for ARB SVN [revision 21008]; Pruesse et al., 2012) and OTU classification was performed against the SILVA SSU Ref dataset (release 115). Alignments were examined and terminal regions with poor coverage trimmed in the ARB environment (Ludwig et al., 2004); however, some positions with incomplete but good coverage over all alignment positions were retained. In doing so, we reasoned that if the alignment was to be split among oligotypes in such a way that only valid sequence data was present at a globally incomplete but well-covered position, that position would be a valid target for oligotyping. However, if a resulting oligotype was derived from an incomplete alignment, it was removed from further analysis. The resulting alignments were exported for oligotyping.

Reads belonging to OTUs with total read counts greater than 100 were oligotyped (Eren et al., 2013) to convergence by recursively selecting the alignment position(s) with the greatest entropy for each round of oligotyping. At each step, a round of oligotyping was only performed on alignments which featured at least 21 sequences and included a position with entropy greater than 0.6 (see **Table 1** and Discussion). The oligotyping output was not restricted by any of the software's command line parameters such as the minimum percent, actual, or substantive abundance. Output from the oligotyping software and SILVAngs pipeline were and then imported into the R environment (R Development Core Team, 2014) for further processing and analysis.

DATA PREPARATION

Geographic coordinates were converted from Global Positioning System (GPS) coordinates to Universal Transverse Mercator (UTM) coordinates (i.e., Easting and Northing in m) using the

Table 1 | Entropy in terms of the proportion of deviations from the expected character in a character sequence and the percentage of the dominant character in that sequence.

Entropy	Proportion of alternate characters relative to the dominant character present at an alignment position	Percent occurrence of the dominant character present at an alignment position
0.65	1:5	83.3
0.60	1:6	85.7
0.44	1:10	90.9
0.28	1:20	95.2
0.21	1:30	96.8
0.14	1:50	98.0
0.08	1:100	99.0
0.02	1:500	99.8
0.01	1:1000	99.9

sp (Pebesma and Bivand, 2005) and *rgdal* (Bivand et al., 2013) R packages. Further, all count data were Hellinger transformed prior to applying redundancy analysis (RDA). Environmental variables, comprising pigment, protein, and phospholipid concentrations as well as spatial variables (Easting, Northing, and water depth) were z-scored (i.e., set to zero mean and unit variance).

GENERAL EXPLORATIONS

Simple diagnostic plots were created to (1) illustrate each sampling location's percent contribution of reads to this analysis and illustrate the per location percentage of reads retained (relative to the reads present in all OTUs at that location) following removal of those reads belonging to oligotypes with incomplete alignments (**Figure 1A**), (2) compare the number of reads clustered in a given OTU to the number of unique oligotypes derived from it (**Figure 1B**), and (3) visualize the proportion of oligotypes derived from OTUs across specific higher-order taxa (**Figure 2**).

DETECTING 'RESOLVING' OLIGOTYPES

For each OTU selected for analysis, we calculated the mean "checkerboard" (C) and "togetherness" (T) scores (Stone and Roberts, 1992) of its oligotypes using the R package bipartite (Dormann et al., 2009). High C scores indicate that pairs of oligotypes occur in checkered patterns across samples. That is, one oligotype's presence and absence is repeatedly mirrored by another's in two-by-two units, resembling a similarly sized unit of a checkerboard. High T scores indicate that pairs of oligotypes tend to occur in aggregates across samples, being simultaneously present or absent. Both C and T scores can be high (relative to those calculated from a random distribution of presences and absences) should groups of aggregated oligotypes, the existence of which will increase the average T score of a matrix, form checkered patterns with other groups, increasing the average C score. Based on these distributions, we selected oligotypes with average checkerboard and togetherness scores greater than the third quartile of all scores measured for further investigation. These oligotypes were treated as candidate 'resolving' oligotypes. A resolving oligotype would thus be heterogeneously distributed across sites, but would cluster with other, similarly distributed oligotypes. Hellinger-transformed abundance matrices were visualized as heatmaps with oligotypes grouped by hierarchical cluster analysis (using average linkage) of the corresponding Bray–Curtis dissimilarity matrices.

DETECTING ENVIRONMENTALLY STRUCTURED OLIGOTYPES

We applied RDA as implemented in the R package *vegan* (Oksanen et al., 2013) to Hellinger-transformed oligotype abundance matrices derived from each oligotyped OTU. Forward selection, as described by Blanchet et al. (2008), was used to select explanatory variables across all RDA solutions calculated. The full model's explanatory matrix comprised the following variables: particulate protein concentration, pigment concentration (CPE), Easting, Northing, and water depth. Models associated with a percentage of constrained variation greater than 50% and *P*-values less than 0.05 were investigated further. All *P*-values were corrected for multiple testing using the base R function, p.adjust, employing the method of Benjamini and Hochberg (1995). Variance inflation factors (estimated with *vegan's* vif.cca function) were verified to be <10 to ensure constraints were not multicollinear.

EXPLORING OLIGOTYPE ASSOCIATIONS

Associations between oligotypes were explored using graph theoretic approaches. Only those oligotypes with a total relative abundance greater than one were considered. A graph was created

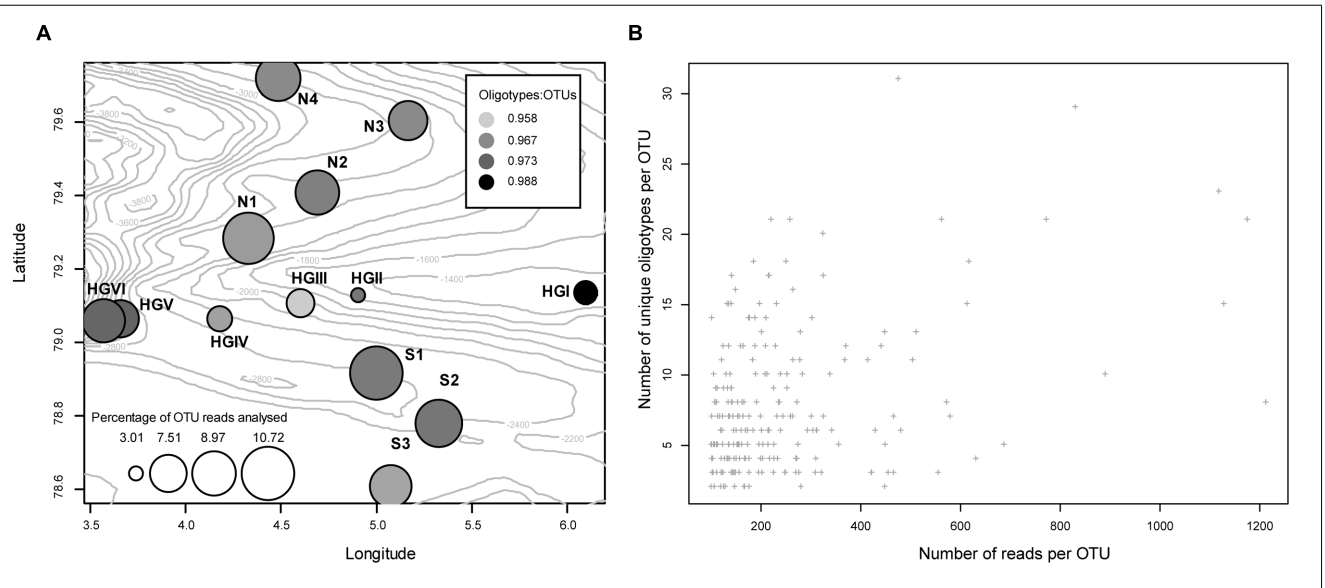


FIGURE 1 | (A) Bubble plot approximating the location of each sampling site (bubble coordinates), the per site, percentage contribution of reads used in this study (bubble size), and the sample-specific ratio of the number of reads that were present in an oligotype to the number of reads in the OTU it was derived from (fill intensity). Numeric values on isobaths indicate the depth of the seafloor in meters below the water surface. **(B)** The total number of reads clustered in a given OTU plotted against the number of oligotypes derived from that OTU (Pearson's $R^2 = 0.39$, $P << 0.01$).

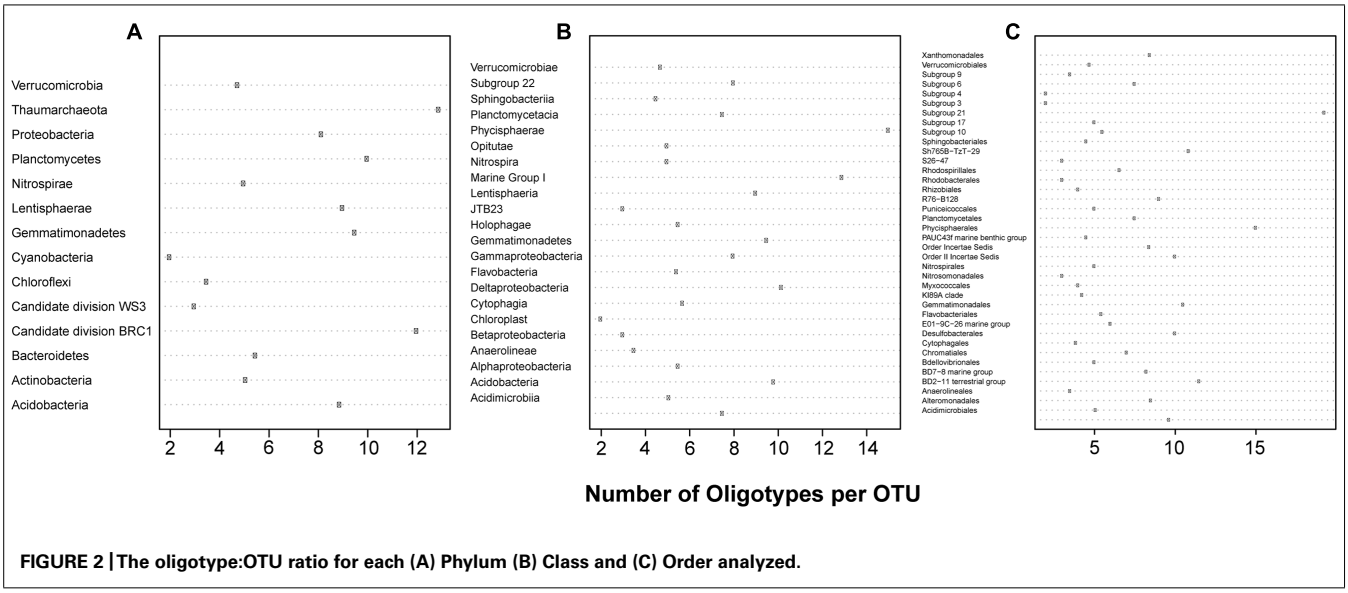


FIGURE 2 | The oligotype:OTU ratio for each (A) Phylum (B) Class and (C) Order analyzed.

with oligotypes as nodes, and edges defined by the value of Whittaker's index of association (IA), as described by Somerfield and Clarke (2013), calculated for each pair of oligotypes. This index is similar to the one-complement of the well-known, asymmetric Bray–Curtis dissimilarity; however, variable (i.e., oligotype) proportions are scaled such that they sum to 100. Consequently, oligotypes with identical percentage abundances across samples have an IA of 100, while those that with no overlapping occurrence across samples have an IA of zero. Significance was assessed by independently permuting ($n = 200$) the sample order in each oligotype abundance vector of the original dataset and recalculating a matrix of IA values. The probabilities of the observed IA values given the permuted values were corrected for multiple testing using the method of Benjamini and Hochberg (1995).

Oligotypes with an IA greater than 85 and an FDR-corrected *P*-value less than 0.05 were linked by an edge and the corresponding IA value was used as an edge weight. The Cytoscape suite (v 3.1.1; Smoot et al., 2011) was used to visualize and analyze the graph object. Node size was scaled by the total abundance of each oligotype (minimum = 2, maximum = 330) and edge width by the value of its weight. The Markov cluster (MCL) algorithm (Enright et al., 2002), as implemented in the *clusterMaker 2* (Morris et al., 2011) Cytoscape 'app,' was used with its default granularity parameter value of 2.5 to identify clusters. As recommended by van Dongen and Abreu-Goodger (2012), the edge weight interval was adjusted from 0.85–1 to 0.001–0.15 to allow better performance of the MCL algorithm.

RESULTS

A total of 19,283 OTUs were generated by the SILVAngs pipeline, of which 95.86% were taxonomically classified. Of these, 217 were represented by at least 100 reads, passed our thresholds for oligotyping, and were used in further analysis. Despite this study targeting bacterial organisms, eight OTUs classified as Thaumarchaeota (Marine Group I) were included in further analyses. Following the oligotyping procedure described above, 1,694 oligotypes were identified, 290 of which were singletons. The minimum, median, and maximum numbers of oligotypes per OTU were 2, 6, and 31, respectively. The oligotyped OTUs represented 14 Phyla, 23 Classes, and 29 Orders (**Figure 2**).

WITHIN-OTU OLIGOTYPE ABUNDANCES SHOW VARIATION ACROSS SAMPLES

Oligotype matrices derived from a total of 25 OTUs possessed average C and T scores above the third quartile of these measures as distributed across all 217 oligotype matrices calculated (i.e., >3.60 and >6.17, respectively; **Table 2**). These scores showed no notable correlation (Pearson's $R^2 = \sim 0.25$, $P = 0.22$). The majority of these OTUs were classified as Acidobacteria or Proteobacteria; however, the highest average C scores belonged to oligotypes of reads assigned to the phyla Gemmatimonadetes and Bacteroidetes, as well as the Candidate division WS3. The highest T scores were observed for reads assigned to the Acidobacteria, Proteobacteria, and Gemmatimonadetes. To illustrate the patterns associated with these average measures, several Hellinger-transformed oligotype abundances were visualized as heatmaps in **Figure 3**.

ASSESSING ENVIRONMENTAL AND SPATIAL EFFECTS ON OLIGOTYPE ABUNDANCES

After performing RDA combined with forward selection, we identified seven OTU-specific oligotype abundance matrices which had greater than 50% of their variation constrained by one or more explanatory variables (**Table 3**). All but one (AHWYC, of the Gammaproteobacteria) featured water depth as an explanatory variable, while porosity, CPE, and a spatial variable were each featured in two models. The triplots of these models, as well as corresponding heatmaps of their Hellinger-transformed oligotype abundances, are displayed in **Figure 4**. Oligotypes, ordinated as bold, red text, show differing responses to the selected explanatory variables. For example, the TGT oligotype of OTU BJCLU (**Figure 4A**) appears in higher relative abundances at shallower

Table 2 | Average checkerboard and togetherness scores for oligotype occurrence matrices generated from selected OTUs, cf. to Figure 3.

OTU ID	Phylum	Class	*n* oligotypes	Mean C score	Mean T score
A83S4	Acidobacteria	Acidobacteria	17	5.16	7.63
DF5XB	Acidobacteria	Acidobacteria	15	4.70	7.01
AS91F	Acidobacteria	Acidobacteria	10	4.60	6.29
EDBYN	Acidobacteria	Subgroup 22	7	5.95	7.19
BTL2B	Acidobacteria	Subgroup 22	9	3.89	9.31
CEL9R	Actinobacteria	Acidimicrobiia	8	4.11	6.18
BRUV2	Actinobacteria	Acidimicrobiia	11	5.98	6.36
DD9DS	Actinobacteria	Acidimicrobiia	6	4.73	9.00
C60MC	Bacteroidetes	Flavobacteria	7	7.05	7.00
DON2B	Candidate division WS3	–	3	7.00	6.33
B177D	Gemmatimonadetes	Gemmatimonadetes	12	8.76	9.39
A5C8S	Planctomycetes	Planctomycetacia	8	4.71	7.21
EMCAY	Proteobacteria	Alphaproteobacteria	6	5.67	7.13
EAFF9	Proteobacteria	Alphaproteobacteria	10	3.71	7.00
BQX8G	Proteobacteria	Deltaproteobacteria	12	6.06	8.02
CMQFL	Proteobacteria	Deltaproteobacteria	15	3.90	6.29
DS0T4	Proteobacteria	Deltaproteobacteria	12	6.30	8.44
DSTJG	Proteobacteria	Deltaproteobacteria	7	5.00	6.76
CA3XY	Proteobacteria	Gammaproteobacteria	7	3.71	7.19
EUGQ5	Proteobacteria	Gammaproteobacteria	8	5.57	9.39
AESJT	Proteobacteria	Gammaproteobacteria	14	6.36	7.65
BQD17	Proteobacteria	Gammaproteobacteria	10	4.60	6.62
AJ3H1	Proteobacteria	Gammaproteobacteria	18	5.90	6.24
EBMBR	Proteobacteria	Gammaproteobacteria	15	5.47	6.49
E00H7	Verrucomicrobia	Verrucomicrobiae	9	4.72	6.64
		Average	**∼10**	**5.34**	**7.31**

sites with higher CPE concentrations while the A oligotype of OTU DTNEI (**Figure 4E**) tends to increase in abundance at increased depth.

EVALUATING OLIGOTYPE-TO-OLIGOTYPE ASSOCIATION

We constructed a network derived from a filtered similarity matrix calculated using Whittaker's IA (see Materials and Methods) which contained 318 nodes (oligotypes) and 1,308 edges (associations; **Figure 5**). A total of 32 connected components (CCs) of varying taxonomic composition were present; however, the network was dominated by a single CC with 225 nodes, while other CCs had between 22 and 2 nodes each. The network had a clustering coefficient of ∼0.28, a density value of ∼0.03, a heterogeneity value of ∼1.33, a centralization value of ∼0.16. Nodes had, on average, ∼8.23 neighbors. Within the largest CC, these values were approximately 0.39, 0.05, 1.07, 0.21, and 11.08, respectively. Additionally the largest CC had scale-free properties with a degree-distribution following a power law: $y = 72.6 \times x^{-1.1}$. Node degree (i.e., the number of edges associated with a given node) ranged from 57 to 1. Of the 10 nodes with the highest degrees (between 39 and 57), five were classified in the Order Gammaproteobacteria (Family: Xanthomonadales), three as Cytophagia, and the remaining two were classified as a Deltaproteobacterium and an Acidobacterium with read abundances between 2 and 68.

The MCL algorithm generated 76 clusters of nodes which included oligotypes belonging to an assortment of taxa and with varying degrees of read abundance (**Figure 6** and **Table 4**). This algorithm resolved the largest CC into several clusters, the largest of which included 72 nodes.

DISCUSSION

In this study, we applied oligotyping to extant sequence data obtained from a unique and dynamic Arctic, deep-sea LTER. While our analyses were primarily exploratory, they indicate that subtle nucleotide variation does indeed provide a new perspective on bacterial diversity at HAUSGARTEN that is not redundant with that derived from OTU-based diversity data. Further, in observing that several of the oligotype abundance matrices derived from

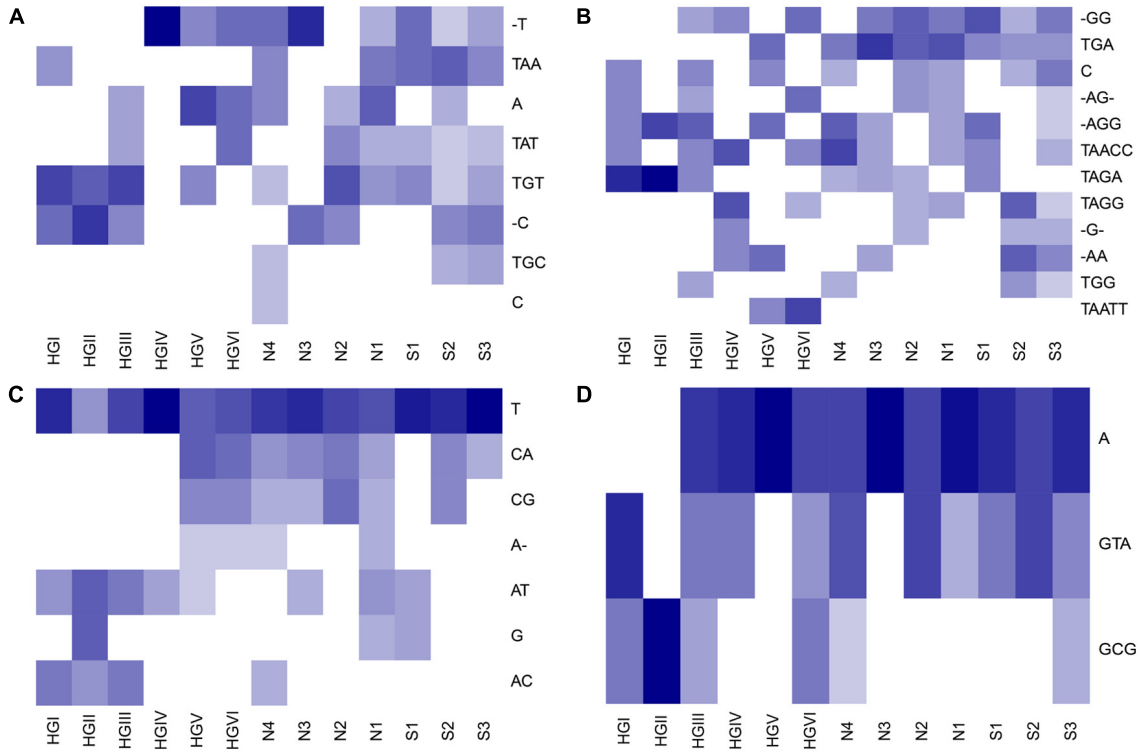


FIGURE 3 | Heatmaps illustrating examples of abundance matrices subject to checkerboard and togetherness score screening. Rows (oligotypes) have been Hellinger transformed and ordered by hierarchical cluster analysis using average linkage and Bray–Curtis dissimilarities. Darker shades indicate higher relative abundance of reads. In the following text, the maximum, untransformed number of oligotype reads in each OTU-derived relative abundance matrix is noted in brackets. The oligotypes for OTUs **(A)** EUGQ5 [8], which had the joint-highest average T score observed; **(B)** B177D [5], which had both the highest average C and T scores observed; **(C)** C60MC [12]; and **(D)** DON2B [9] which both had high average C and T scores are displayed (cf. **Table 2**). AGCT: nucleotides; -: gap.

specific OTUs appear to be structured by environmental or spatial variables (**Table 3** and **Figure 4**), we are encouraged that further application of this technique – particularly in the context of 'omic-centered,' long-term research (see e.g., Davies et al., 2014) – will enhance the likelihood of identifying ecologically meaningful divergence at up to single-base resolution. This, in turn, may aid in the detection of ecotypes (e.g., Moore et al., 1998; Garczarek et al., 2007; Ivars-Martinez et al., 2008) and the concomitant deepening of knowledge surrounding the ecosystems they inhabit. Naturally, the success of such a strategy is directly determined by the selection of an appropriate genetic element, as the 16S gene may, in some cases, have poor resolving power (e.g., Jaspers and Overmann, 2004) and other markers may offer more scope (Lerat et al., 2003; Yilmaz et al., 2011; Mende et al., 2013).

THRESHOLDS FOR OLIGOTYPE DETECTION

In the present case, we limited our analysis to OTUs with high read abundance in order to operate on relatively large alignments which could undergo several rounds of oligotyping. Under this constraint, we noted that the abundance of reads belonging to an OTU does not meaningfully correlate with the number of oligotypes it will be resolved into (**Figure 1B**), which reinforces the notion that understanding nucleotide variation is likely to require specific knowledge of the organisms, evolutionary characteristics, and ecology involved in the diversification processes at work (McDonald et al., 2013). We acknowledge that limiting our analysis to these abundant OTUs precludes the observation of many, potentially important oligotypes; however, we find it prudent to reserve more thorough analysis until a greater body of longitudinal sequence data is amassed at HAUSGARTEN. Repeated observation of oligotypes over time and the evaluation of their variation in the face of environmental variation will provide a far better basis for interpretation.

In addition to focusing solely on abundant OTUs, we only performed a round of oligotyping if an alignment with at least 21 sequences was available and entropy analysis revealed positions with entropy values greater than or equal to 0.6. We acknowledge that our choice of entropy and sequence count thresholds is, ultimately, arbitrary. **Table 1** partly clarifies the nature of our selection: with an entropy value of 0.6, one can expect 85.7% of aligned characters in a given position to be identical, or an alternative character for every six instances of the dominant character. Selecting lower entropies increases the risk of identifying sequencing errors as oligotypes while higher entropy thresholds would decrease the sensitivity of the method. We propose that applying a statistical method to determine a suitable threshold for each execution of the oligotyping procedure may provide a more robust and less

Table 3 | Results of RDA on oligotype abundance matrices derived from selected OTUs.

OTU ID	Class	Order	*n* oligotypes	Model	Constrained variation (%)
BJCLU	Deltaproteobacteria	Bdellovibrionales	5	Y ∼ Easting + CPE + Depth	69
AV4R2	Gammaproteobacteria	Incertae Sedis	4	Y ∼ Depth + Porosity	52
BGP4M	Cytophagia	Cytophagales	3	Y ∼ Depth	70
D3V9F	Gammaproteobacteria	Xanthomonadales	7	Y ∼ Depth	69
DTNEI	Cytophagia	Cytophagales	3	Y ∼ Depth	65
AHWYC	Gammaproteobacteria	Xanthomonadales	5	Y ∼ Porosity + CPE	65
ANOZB	Cytophagia	Cytophagales	4	Y ∼ Depth + Northing	60

Explanatory variables were chosen through forward selection. All models had FDR-corrected P-values of, at most, 0.0098. Y: the response matrix of oligotype relative abundances; CPE: Pigment concentration; Depth: Water depth.

subjective threshold criterion. The broken stick model, commonly used to predict the relative sizes of a randomly fragmented whole, may offer such a solution (Ramette and Buttigieg, 2014).

DETECTING 'RESOLVING' OLIGOTYPES

We attempted to estimate the degree to which reads in an OTU have been distributed across oligotypes such that they may be used to differentiate between sites (i.e., 'resolve' sites based on their distributions) by calculating the average checkerboard (C) and togetherness (T) scores of each OTU-specific oligotype abundance matrix. We used C and T scores as they allowed us to screen for oligotypes with strong, presence-absence-based partitioning and aggregation among sites. This partitioning may be indicative of ecotype partitioning (i.e., competitive exclusion) as observed for other marine bacteria (e.g., Garczarek et al., 2007) and, if observed in repeated studies, may motivate taxon-targeted investigations to determine whether ecotype-level dynamics are in effect. Oligotypes which tend to co-occur at certain sites (e.g., **Figure 3A**, oligotypes TGT and -C at sites HGI–HGIII) may be indicative of subpopulations with similar levels of fitness in those locations. As an example, this screening approach revealed that oligotypes of OTU B177D, from the poorly characterized phylum Gemmatimonadetes, were associated with the highest average C and T scores (**Table 2** and **Figure 3B**). The Gemmatimonadetes have been observed in diverse environments, including soils and aquatic sediments, suggesting a diverse range of metabolic capacities in this phylum (DeBruyn et al., 2011). While confirmation is required, it is not unfounded to hypothesize that such metabolic plasticity may have translated into oligotype-level subpopulations colonizing HAUSGARTEN. Other oligotype matrices with high C and T scores include that of OTU C60MC (**Figure 3C**), classified as a representative of the Bacteroidetes. Apparent depth-related community composition changes within the Bacteroidetes have been observed in the Mediterranean (Díez-Vives et al., 2014), a trend somewhat echoed in our results where several oligotypes (CG, CA, and A-) were absent from shallower sites where others occurred (AT, AC, T, and G). One possible drawback of this approach is that C and T scores are binary measures and are not sensitive to differential abundance in oligotypes that are present in the same site. Thus this approach will not detect patterns which would, for example, indicate that one of a set of oligotypes appears to have greater fitness than others without leading to exclusion. To address this, the application of techniques dealing with abundance-based checkerboard and togetherness measures (Ulrich and Gotelli, 2010) may provide more informative results.

While outside the scope of this OTU-focused study, these results provide motivation to examine the higher-order taxa containing resolving oligotypes – alongside others found to have high C and T scores such as the Acidobacteria, Gamma-, Alpha-, and Deltaproteobacteria – through oligotyping. This will become an especially interesting undertaking as more next-generation sequencing datasets become available from the HAUSGARTEN LTER, enabling the detection of persistent oligotypes in the system and providing motivation for their further study. The natural consequence of confirming recurrent, site-resolving oligotypes is the formulation of hypotheses regarding the drivers of their differentiation in an effort to describe the microbial ecology at this scale.

ENVIRONMENTALLY AND SPATIALLY STRUCTURED OLIGOTYPES

To complement the presence-absence-based checkerboard and togetherness analyses, we employed RDA – a multivariate form of multiple linear regression – to detect linear, abundance-based responses to environmental and spatial explanatory variables. Following our application of RDA and forward variable selection, we observed only seven of the 217 OTUs selected for analysis produced oligotype abundance matrices with greater than 50% explained variation. Indeed, a total of 55 models included at least one explanatory term in the model, while 162 models were trivial (i.e., 'intercept-only' models, featuring no explanatory terms). This result suggests that much of the oligotype-based microdiversity is not structured by the environmental or spatial factors measured; however, it may also imply that variables which are able to account for these responses have been overlooked. Additionally, we accept that our threshold for constrained variation is likely to be harsh for an ecological investigation: due to the sheer complexity of most ecosystems, it is not unusual to explain only a small fraction of the total variation in a response matrix (Cottenie, 2005). Nonetheless, we choose to err on the side of caution and

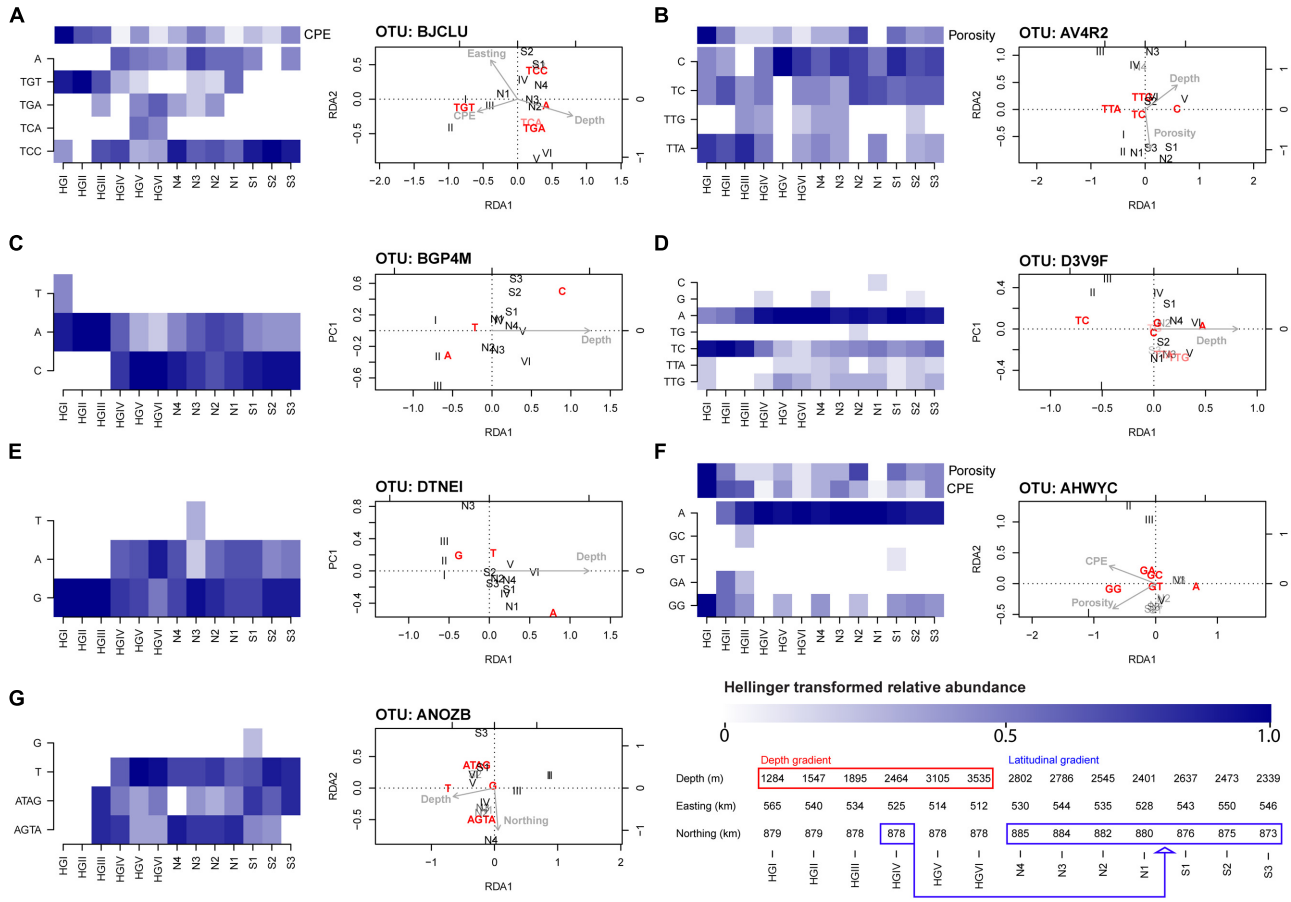


FIGURE 4 | Heatmaps and RDA triplots (type 2 scaling) derived from Hellinger-standardized, OTU-specific oligotype relative abundance matrices. The seven models shown had at least 50% constrained variance and were significant at a *P*-value threshold of 0.05 (FDR-corrected) cf. **Table 3**. The depth and latitudinal gradients sampled are reflected in the sample order (HGI shallowest, HGVI deepest, N4 northernmost, S3 southernmost) and elaborated upon in the bottom right of the figure where the depth and latitudinal gradients are highlighted with red and blue boxes, respectively. Site HGIV, the central site of the intersecting transects, belongs to both transects. See **Figure 1A** and Jacob et al. (2013) for greater detail. When other explanatory variables were featured in the model, an additional heatmap of these variables' z-scored values is included in the panel. Across heatmaps, darker shades indicate higher, Hellinger-transformed relative abundance of reads or higher values of a given explanatory variable. The panels reference oligotype abundance matrices derived from the following OTUs (the maximum, untransformed read abundance across oligotypes in each matrix is noted in brackets): **(A)** BJCLU [18] **(B)** AV4R2 [14] **(C)** BGP4M [16] **(D)** D3V9F [62] **(E)** DTNEI [19] **(F)** AHWYC [39] and **(G)** ANOZB [9]. Explanatory variables are represented by gray text and arrows pointing in their direction of increase, these comprise Porosity (range: 51.8–72.3% volume), CPE (18.86–44.26 μg cm^{-3}), Northing (8727035–8850377 m), Easting (512100–565125 m), and Depth (1284–3535 m). Oligotypes (response variables) are ordinated as bold, red text. Relative to each plot's origin, the position of an oligotype's ordination indicates its direction of increase. Angles between variables indicate their linear correlation, with an angle of 0° indicating perfect positive correlation, 180° indicating perfect negative correlation, and 90° indicating orthogonality. Samples are ordinated as black text. Transparency effects are used to improve visibility in congested regions of the triplots and have no meaning.

report on oligotype matrices strongly structured by our explanatory variables. These results do show, however, that oligotype-level variation reveals patterns that are not evident at the OTU-level and that are related to environmental parameters.

In line with previous findings, water depth prominently featured in the models selected by our methods (**Table 3**). Several OTU-specific oligotypes appear to increase with depth (e.g., oligotype A of OTU D3V9F and oligotype A of OTU DTNEI; **Figures 4D,E** respectively) while others seem to have higher abundances in shallower regions (e.g., oligotype TC of OTU D3V9F; **Figure 4D**) or little response to varying depth (e.g., oligotype G of OTU ANOZB; **Figure 4G**). In several cases, oligotypes within a given OTU appear to show differential responses to depth (e.g., oligotypes A and C of OTU BGP4M and oligotypes TC and A of OTU D3V9F; **Figures 4C,D**, respectively). As discussed above and by Jacob et al. (2013), water depth is likely to act as a proxy variable for numerous depth-related parameters such as pressure or ecosystem composition (e.g., the community composition of larger organisms). Indeed, the negative correlation of depth with benthic phytodetritus concentrations (in our analysis, approximated by CPE concentrations) is reflected in the ordination of oligotypes derived from OTU BJCLU (**Figure 4A**). In this ordination, oligotype TGT appears at shallower sites with higher CPE concentrations, whereas other oligotypes appear to favor deeper

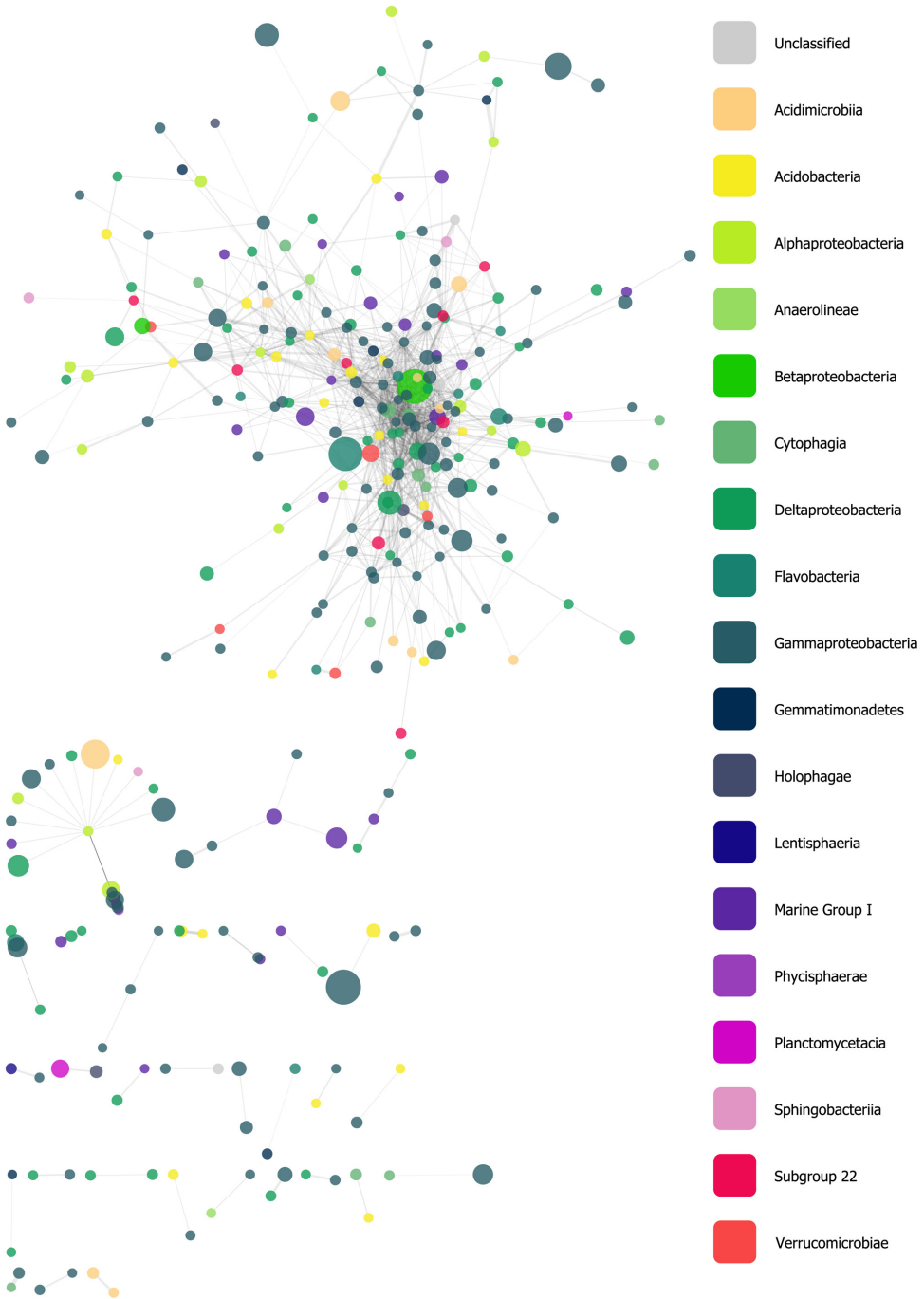


FIGURE 5 | Force-directed, spring-embedded network displaying oligotypes (nodes) with Whittaker's index of association (IA) values greater than 85 (FDR-corrected *P*-values as determined by 200 permutations <0.05), represented as edges. Nodes are color-coded by taxonomic Class. See text for a summary of this network's general statistics.

sites with lower CPE concentrations. Thus, the prominence of depth as an explanatory variable in the RDA models above is unsurprising, but its exact relevance to the oligotypes derived from each OTU analyzed is more difficult to interpret. This provides motivation to design future sampling procedures that would capture a broader suite of depth-related contextual variables in aid of more precise characterization of bacterial community responses across taxonomic scales. Sampling during a natural perturbation which would decouple environmental factors that co-vary with depth (and are thus likely to confound one another in subsequent analyses) may also offer a particularly valuable opportunity to isolate their effects. Additional factors such as porosity, which has

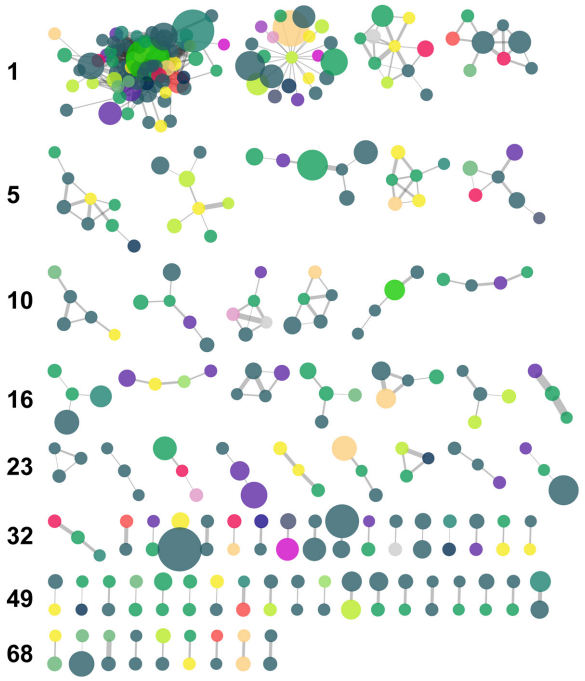


FIGURE 6 | Results of Markov clustering of the graph displayed in Figure 5, with a granularity parameter value of 2.5. Nodes are color-coded by taxonomic Class (see **Figure 5** for key) and their size is proportional to the total relative abundance of the corresponding oligotype. Edge thickness is proportional to the value of the IA between oligotypes. The leftmost cluster of each row is numbered along the left margin.

been observed to co-vary with benthic community structure in other Arctic sediments (Hamdan et al., 2013), and pigment concentration (partially indicative of energy availability in this system and likely associated with the presence of sea ice) are also linked to a bathymetric gradient; however, were not observed to be highly collinear with water depth. Thus, models such as that of OTU AHWYC (**Table 3** and **Figure 4F**) are important inasmuch as they are likely to reflect alternate ecological dynamics, worthy of pursuing in subsequent sampling designs. It is tempting to speculate that oligotypes with differential responses to variables such as depth and CPE concentration represent potential ecotypes. For example, based on their occurrence profiles and ordination by RDA, it may be hypothesized that organisms represented by the TGT oligotype of OTU BJCLU (**Figure 4A**) favor conditions where labile food sources are available (i.e., higher CPE concentrations), while those represented by oligotypes TCC and A are adapted to feeding on more recalcitrant compounds. A similar assertion may be made for oligotypes GG and A derived from OTU AHWYC (**Figure 4F**).

OLIGOTYPE–OLIGOTYPE ASSOCIATIONS

Our network analysis of oligotype associations based on Whittaker's IA revealed a large CC with scale-free properties, a trait that is frequently observed in biological and ecological networks, and several much smaller components (**Figure 5**). The variety of taxa and abundance classes which shared associations in the network and the MCL clusters derived from it (**Figure 6**) is a simple, but informative, result: oligotype associations cross taxonomic boundaries and abundance classes. This implies that oligotype-level variation reveals heretofore uninvestigated sub-OTU co-occurrence patterns that represent, for example, candidate bacterial guilds. Should these associations be validated with independent data (e.g., repeated sampling and sequencing of these HAUSGARTEN sites), they would provide motivation for targeted studies investigating specific sub-OTU microbial interactions. Additionally, the variation of consistently observed, closely associated oligotypes provides a reference against which one is able to identify which contextual parameters are of relevance to the microbial ecology of this rapidly changing ecosystem.

As a final note, we observed several nodes with high degree ($\geq$25), but which corresponded to oligotypes with low abundance ($\leq$5 reads). While true association cannot be ruled out, caution must be exercised in interpreting the associations of 'rare' oligotypes. While we did choose to remove absolute singletons (i.e., oligotypes which only had one read in the entire dataset), we did not use the oligotyping software's parameters to restrict output based on the various abundance measures offered. While this may result in oligotypes generated from sequencing errors contaminating our results, it also prevents false negatives. As stated above, we suggest that the validation of oligotype occurrence through repeated sampling is a more tenable solution to this issue than arguments for or against a given, arbitrary threshold, which may have unpredictable effects on the analysis of count data (as shown in e.g., Gobet et al., 2010).

CONCLUSION

This study adds both to the characterization of the bacterial benthos present at the HAUSGARTEN LTER and to the exploration of oligotyping as a methodology to detect heretofore

Table 4 | Selected characteristics of the five largest MC clusters with reference to the IA network cf. Figure 6.

Cluster	No. of oligotypes	Average and range of oligotype abundance	No. of phyla represented	No. of classes represented
1	72	36.2, 327	8	14
2	22	53.6, 250	6	11
3	10	23.7, 102	3	5
4	8	38.38, 111	4	5
5	8	10.5, 23	3	4

undescribed bacterial microdiversity and ecology. Our results largely confirm previous observations linking responses in microbial community structure to water depth; however, they reveal a finer-grained response that can be both a source and target for new ecological hypotheses. Indeed, oligotypes from within a single OTU were observed to show differential occurrence across sites, respond differently to the explanatory variables analyzed, and associate with oligotypes derived from other OTUs. While work remains to be done in refining this approach and standardizing its application, oligotyping offers a readily applicable means to explore patterns in microbial microdiversity. Sequencing-enabled LTERs and Genomic Observatories (Davies et al., 2012, 2014) are uniquely positioned to evaluate oligotyping and similar methods through repeated sampling and validation and, in the process, have the opportunity to identify distinct microbial subpopulations and ecotypes central to their study site. The value of this capability is especially pronounced in regions undergoing rapid change, where a grasp of microbial responses at fine granularity is desirable.

ACKNOWLEDGMENTS

We are grateful to Christian Quast for his assistance in preparing data exports from the SILVAngs pipeline. Alban Ramette is funded by the Max Planck Society. Pier Luigi Buttigieg's work on this project is supported through the Micro B3 project, funded by the European Union's Seventh Framework Programme (Joint Call OCEAN.2011-2: marine microbial diversity – new insights into marine ecosystems functioning and its biotechnological potential) under the grant agreement no 287589.

REFERENCES

Bauerfeind, E., Nöthig, E. M., Beszczynska, A., Fahl, K., Kaleschke, L., Kreker, K., et al. (2009). Particle sedimentation patterns in the eastern Fram Strait during 2000–2005: results from the Arctic long-term observatory HAUSGARTEN. *Deep Sea Res. Part 1 Oceanogr. Res. Pap.* 56, 1471–1487. doi: 10.1016/j.dsr.2009.04.011

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 57, 289–300. doi: 10.2307/2346101

Bienhold, C., Boetius, A., and Ramette, A. (2012). The energy-diversity relationship of complex bacterial communities in Arctic deep-sea sediments. *ISME J.* 6, 724–732. doi: 10.1038/ismej.2011.140

Bivand, R., Keitt, T., and Rowlingson, B. (2013). *rgdal: Bindings for the Geospatial Data Abstraction Library. R Package Version 0.8–11*. Available at: http://CRAN.R-project.org/package=rgdal

Blanchet, F. G., Legendre, P., and Borcard, D. (2008). Forward selection of explanatory variables. *Ecology* 89, 2623–2632. doi: 10.1890/07-0986.1

Boetius, A., Albrecht, S., Bakker, K., Bienhold, C., Felden, J., Fernández-Méndez, M., et al. (2013). Export of algal biomass from the melting Arctic sea ice. *Science* 339, 1430–1432. doi: 10.1126/science.1231346

Cohan, F. M. (2001). Bacterial species and speciation. *Syst. Biol.* 50, 513–524. doi: 10.1080/10635150118398

Cohan, F. M. (2002). What are bacterial species? *Annu. Rev. Microbiol.* 56, 457–487. doi: 10.1146/annurev.micro.56.012302.160634

Coleman, M. L., Sullivan, M. B., Martiny, A. C., Steglich, C., Barry, K., DeLong, E. F., et al. (2006). Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* 311, 1768–1770. doi: 10.1126/science.1122050

Cottenie, K. (2005). Integrating environmental and spatial processes in ecological community dynamics. *Ecol. Lett.* 8, 1175–1182. doi: 10.1111/j.1461-0248.2005.00820.x

Davies, N., Field, D., Amaral-Zettler, L., Clark, M. S., Deck, J., Drummond, A., et al. (2014). The founding charter of the genomic observatories network. *Gigascience* 3, 2. doi: 10.1186/2047-217X-3-2

Davies, N., Field, D., and Genomic Observatories Network. (2012). Sequencing data: a genomic network to monitor Earth. *Nature* 481, 145. doi: 10.1038/481145a

DeBruyn, J. M., Nixon, L. T., Fawaz, M. N., Johnson, A. M., and Radosevich, M. (2011). Global biogeography and quantitative seasonal dynamics of Gemmatimonadetes in soil. *Appl. Environ. Microbiol.* 77, 6295–6300. doi: 10.1128/AEM.05005-11

Díez-Vives, C., Gasol, J. M., and Acinas, S. G. (2014). Spatial and temporal variability among marine Bacteroidetes populations in the NW Mediterranean Sea. *Syst. Appl. Microbiol.* 37, 68–78. doi: 10.1016/j.syapm.2013.08.006

Dormann, C., Fründ, J., Blüthgen, N., and Gruber, B. (2009). Indices, graphs and null models: analyzing bipartite ecological networks. *Open Ecol. J.* 2, 7–24. doi: 10.2174/1874213000902010007

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584. doi: 10.1093/nar/30.7.1575

Eren, A. M., Borisy, G. G., Huse, S. M., and Mark Welch, J. L. (2014a). Oligotyping analysis of the human oral microbiome. *Proc. Natl. Acad. Sci. U.S.A.* 11, E2875–E2884. doi: 10.1073/pnas.1409644111

Eren, A. M., Sogin, M. L., Morrison, H. G., Vineis, J. H., Fisher, J. C., Newton, R. J., et al. (2014b). A single genus in the gut microbiome reflects host preference and specificity. *ISME J.* 1–11. doi: 10.1038/ismej.2014.97

Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., et al. (2013). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* 4, 1111–1119. doi: 10.1111/2041-210X.12114

Eren, A. M., Zozaya, M., Taylor, C. M., Dowd, S. E., Martin, D. H., and Ferris, M. J. (2011). Exploring the diversity of *Gardnerella vaginalis* in the genitourinary tract microbiota of monogamous couples through subtle nucleotide variation. *PLoS ONE* 6:e26732. doi: 10.1371/journal.pone.0026732

Garczarek, L., Dufresne, A., Rousvoal, S., West, N. J., Mazard, S., Marie, D., et al. (2007). High vertical and low horizontal diversity of *Prochlorococcus* ecotypes in the Mediterranean Sea in summer. *FEMS Microbiol. Ecol.* 60, 189–206. doi: 10.1111/j.1574-6941.2007.00297.x

Gobet, A., Quince, C., and Ramette, A. (2010). Multivariate Cutoff Level Analysis (MultiCoLA) of large community data sets. *Nucleic Acids Res.* 38:e155. doi: 10.1093/nar/gkq545

Grebmeier, J. M., Overland, J. E., Moore, S. E., Farley, E. V., Carmack, E. C., Cooper, L. W., et al. (2006). A major ecosystem shift in the northern Bering Sea. *Science* 311, 1461–1464. doi: 10.1126/science.1121365

Hahn, M. W., and Pöckl, M. (2005). Ecotypes of planktonic actinobacteria with identical 16S rRNA genes adapted to thermal niches in temperate, subtropical, and tropical freshwater habitats. *Appl. Environ. Microbiol.* 71, 766–773. doi: 10.1128/AEM.71.2.766-773.2005

Hamdan, L. J., Coffin, R. B., Sikaroodi, M., Greinert, J., Treude, T., and Gillevet, P. M. (2013). Ocean currents shape the microbiome of Arctic marine sediments. *ISME J.* 7, 685–696. doi: 10.1038/ismej.2012.143

Hebbeln, D. (2000). Flux of ice-rafted detritus from sea ice in the Fram Strait. *Deep Sea Res. Part 2 Top. Stud. Oceanogr.* 47, 1773–1790. doi: 10.1016/S0967-0645(00)00006-0

Hop, H., Falk-Petersen, S., Svendsen, H., Kwasniewski, S., Pavlov, V., Pavlova, O., et al. (2006). Physical and biological characteristics of the pelagic system across Fram Strait to Kongsfjorden. *Prog. Oceanogr.* 71, 182–231. doi: 10.1016/j.pocean.2006.09.007

Ivars-Martinez, E., Martin-Cuadrado, A.-B., D'Auria, G., Mira, A., Ferriera, S., Johnson, J., et al. (2008). Comparative genomics of two ecotypes of the marine planktonic copiotroph *Alteromonas macleodii* suggests alternative lifestyles associated with different kinds of particulate organic matter. *ISME J.* 2, 1194–1212. doi: 10.1038/ismej.2008.74

Jacob, M., Soltwedel, T., Boetius, A., and Ramette, A. (2013). Biogeography of deep-sea benthic bacteria at regional scale (LTER HAUSGARTEN, Fram Strait, Arctic). *PLoS ONE* 8:e72779. doi: 10.1371/journal.pone.0072779

Jaspers, E., and Overmann, J. (2004). Ecological significance of microdiversity: identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. *Appl. Environ. Microbiol.* 70, 4831–4839. doi: 10.1128/AEM.70.8.4831-4839.2004

Koeppel, A. F., and Wu, M. (2014). Species matter: the role of competition in the assembly of congeneric bacteria. *ISME J.* 8, 531–540. doi: 10.1038/ismej.2013.180

Kopac, S., and Cohan, F. M. (2011). "A theory-based pragmatism for discovering and classifying newly divergent bacterial species," in *Genetics and Evolution of Infectious Diseases*, ed. M. Tibayrenc (London: Elsevier), 21–41.

Lalande, C., Bauerfeind, E., Nöthig, E.-M., and Beszczynska-Möller, A. (2013). Impact of a warm anomaly on export fluxes of biogenic matter in the eastern Fram Strait. *Prog. Oceanogr.* 109, 70–77. doi: 10.1016/j.pocean.2012.09.006

Lerat, E., Daubin, V., and Moran, N. A. (2003). From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. *PLoS Biol.* 1:E19. doi: 10.1371/journal.pbio.0000019

Leu, E., Søreide, J. E., Hessen, D. O., Falk-Petersen, S., and Berge, J. (2011). Consequences of changing sea-ice cover for primary and secondary producers in the European Arctic shelf seas: timing, quantity, and quality. *Prog. Oceanogr.* 90, 18–32. doi: 10.1016/j.pocean.2011.02.004

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, et al. (2004). ARB: a software environment for sequence data. *Nucleic Acids Res.* 32, 1363–1371. doi: 10.1093/nar/gkh293

McDonald, D., Vázquez-Baeza, Y., Walters, W. A., Caporaso, J. G., and Knight, R. (2013). From molecules to dynamic biological communities. *Biol. Philos.* 28, 241–259. doi: 10.1007/s10539-013-9364-4

McLellan, S. L., Newton, R. J., Vandewalle, J. L., Shanks, O. C., Huse, S. M., Eren, A. M., et al. (2013). Sewage reflects the distribution of human faecal Lachnospiraceae. *Environ. Microbiol.* 15, 2213–2227. doi: 10.1111/1462-2920.12092

Mende, D. R., Sunagawa, S., Zeller, G., and Bork, P. (2013). Accurate and universal delineation of prokaryotic species. *Nat. Methods* 10, 881–884. doi: 10.1038/nmeth.2575

Moore, L. R., Rocap, G., and Chisholm, S. W. (1998). Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* 393, 464–467. doi: 10.1038/30965

Morris, J. H., Apeltsin, L., Newman, A. M., Baumbach, J., Wittkop, T., Su, G., et al. (2011). clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics* 12:436. doi: 10.1186/1471-2105-12-436

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., et al. (2013). *vegan: Community Ecology Package. R Package Version 2.0-7*. Available at: http://CRAN.R-project.org/package=vegan

Pebesma, E. J., and Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News* 5, 9–13.

Piechura, J., and Walczowski, W. (2009). Warming of the West Spitsbergen Current and sea ice north of Svalbard. *Oceanologia* 51, 147–164. doi: 10.5697/oc.51-2.147

Pruesse, E., Peplies, J., and Glöckner, F. O. (2012). SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* 28, 1823–1829. doi: 10.1093/bioinformatics/bts252

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596. doi: 10.1093/nar/gks1219

Ramette, A., and Buttigieg, P. L. (2014). The R package otu2ot for implementing the entropy decomposition of nucleotide variation in sequence data. *Front. Microbiol.* 5:601. doi: 10.3389/fmicb.2014.00601

R Development Core Team. (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available at: http://www.r-project.org/

Schewe, I., and Soltwedel, T. (2003). Benthic response to ice-edge-induced particle flux in the Arctic Ocean. *Polar Biol.* 26, 610–620. doi: 10.1007/s00300-003-0526-8

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27, 431–432. doi: 10.1093/bioinformatics/btq675

Soltwedel, T., Bauerfeind, E., Bergmann, M., Budaeva, N., Hoste, E., Jaeckisch, N., et al. (2005). HAUSGARTEN: multidisciplinary investigations at a deep-sea, long-term observatory in the Arctic Ocean. *Oceanography* 18, 46–61. doi: 10.5670/oceanog.2005.24

Somerfield, P. J., and Clarke, K. R. (2013). Inverse analysis in non-parametric multivariate analyses: distinguishing groups of associated species which covary coherently across samples. *J. Exp. Mar. Biol. Ecol.* 449, 261–273. doi: 10.1016/j.jembe.2013.10.002

Stone, L., and Roberts, A. (1992). Competitive exclusion, or species aggregation? *Oecologia* 91, 419–424. doi: 10.1007/BF00317632

Tiercy, J. M., Jeannet, M., and Mach, B. (1990). A new approach for the analysis of HLA class II polymorphism: "HLA oligotyping." *Blood Rev.* 4, 9–15. doi: 10.1016/0268-960X(90)90012-H

Ulrich, W., and Gotelli, N. J. (2010). Null model analysis of species associations using abundance data. *Ecology* 91, 3384–3397. doi: 10.1890/09-2157.1

van Dongen, S., and Abreu-Goodger, C. (2012). Using MCL to extract clusters from networks. *Methods Mol. Biol.* 804, 281–295. doi: 10.1007/978-1-61779-361-5_15

van Oevelen, D., Bergmann, M., Soetaert, K., Bauerfeind, E., Hasemann, C., Klages, M., et al. (2011). Carbon flows in the benthic food web at the deep-sea observatory HAUSGARTEN (Fram Strait). *Deep Sea Res. Part 1 Oceanogr. Res. Pap.* 58, 1069–1083. doi: 10.1016/j.dsr.2011.08.002

Yilmaz, P., Kottmann, R., Pruesse, E., Quast, C., and Glöckner, F. O. (2011). Analysis of 23S rRNA genes in metagenomes – a case study from the global ocean sampling expedition. *Syst. Appl. Microbiol.* 34, 462–469. doi: 10.1016/j.syapm.2011.04.005

Zinger, L., Amaral-Zettler, L. A., Fuhrman, J. A., Horner-Devine, M. C., Huse, S. M., Welch, D., et al. (2011). Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLoS ONE* 6:e24570. doi: 10.1371/journal.pone.0024570

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 August 2014; accepted: 13 November 2014; published online: 05 January 2015.

*Citation: Buttigieg PL and Ramette A (2015) Biogeographic patterns of bacterial microdiversity in Arctic deep-sea sediments (Hausgarten, Fram Strait). Front. Microbiol. **5**:660. doi: 10.3389/fmicb.2014.00660*

This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology.

frontiers in MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 07 November 2014
doi: 10.3389/fmicb.2014.00525

Population dynamics and ecology of *Arcobacter* in sewage

Jenny C. Fisher[1], Arturo Levican[2,3], María J. Figueras[4] and Sandra L. McLellan[1]*

[1] School of Freshwater Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI, USA
[2] Laboratorio de Patología de Organismos Acuáticos y Biotecnología Acuícola, Facultad de Ciencias, Universidad Andrés Bello, Viña del Mar, Chile
[3] Interdisciplinary Center for Aquaculture Research (INCAR), Concepción, Chile
[4] Unit of Microbiology, Department of Basic Health Sciences, School of Medicine and Health Sciences, Institut d'Investigació Sanitaria Pere Virgili, University Rovira i Virgili, Reus, Spain

Edited by:
Lois Maignien, University of Western Brittany - Université Bretagne Occidentale, France

Reviewed by:
Raffaella Balestrini, Consiglio Nazionale delle Ricerche, Italy
Brett J. Baker, University of Texas Austin, USA

***Correspondence:**
Sandra L. McLellan, School of Freshwater Sciences, University of Wisconsin-Milwaukee, 600 E. Greenfield Ave., Milwaukee, WI 53204, USA
e-mail: mclellan@uwm.edu

Arcobacter species are highly abundant in sewage where they often comprise approximately 5–11% of the bacterial community. Oligotyping of sequences amplified from the V4V5 region of the 16S rRNA gene revealed *Arcobacter* populations from different cities were similar and dominated by 1–3 members, with extremely high microdiversity in the minor members. Overall, nine subgroups within the *Arcobacter* genus accounted for >80% of the total *Arcobacter* sequences in all samples analyzed. The distribution of oligotypes varied by both sample site and temperature, with samples from the same site generally being more similar to each other than other sites. Seven oligotypes matched with 100% identity to characterized *Arcobacter* species, but the remaining 19 abundant oligotypes appear to be unknown species. Sequences representing the two most abundant oligotypes matched exactly to the reference strains for *A. cryaerophilus* group 1B (CCUG 17802) and group 1A (CCUG 17801^T), respectively. Oligotype 1 showed generally lower relative abundance in colder samples and higher relative abundance in warmer samples; the converse was true for Oligotype 2. Ten other oligotypes had significant positive or negative correlations between temperature and proportion in samples as well. The oligotype that corresponded to *A. butzleri*, the *Arcobacter* species most commonly isolated by culturing in sewage studies, was only the eleventh most abundant oligotype. This work suggests that *Arcobacter* populations within sewer infrastructure are modulated by temperature. Furthermore, current culturing methods used for identification of *Arcobacter* fail to identify some abundant members of the community and may underestimate the presence of species with affinities for growth at lower temperatures. Understanding the ecological factors that affect the survival and growth of *Arcobacter* spp. in sewer infrastructure may better inform the risks associated with these emerging pathogens.

Keywords: oligotyping, *Arcobacter*, sewage, population dynamics, V4V5, Illumina MiSeq

INTRODUCTION

The genus *Arcobacter*, described by Vandamme et al. (1992), belongs to the family *Campylobacteraceae* within the epsilon-Proteobacteria. *Arcobacter* spp. were originally grouped within genus *Campylobacter*, but differ from campylobacters in their ability to grow under aerobic conditions and lower temperatures. The genus *Arcobacter* currently contains 18 species (Levican et al., 2013a; Sasi Jyothsna et al., 2013) isolated from diverse environments (water, plant roots, food) and hosts (humans, poultry, pigs, shellfish) (Collado and Figueras, 2011). Many *Arcobacter* species have been isolated from multiple locations, suggesting that these organisms are metabolically flexible and can survive under an array of environmental conditions.

Three *Arcobacter* species, *A. butzleri*, *A. cryaerophilus*, and *A. skirrowii*, have emerged in recent years as potential human pathogens (Collado and Figueras, 2011). Strains of *A. butzleri* and *A. cryaerophilus* in particular have been isolated from human stool and blood samples, and pathogenicity can range from diarrhea to bacteremia (Figueras et al., 2014). Some *A. butzleri* isolates contain a suite of virulence genes (*cadF*, *ciaB*, *cj1349*, *hecA*, *hecB*, *irgA*, *mviN*, *pldA*, and *tlyA*) (Douidah et al., 2012; Levican et al., 2013a) and can adhere to and invade Caco-2 cells (a gut epithelial cell line) *in vitro* (Levican et al., 2013a). The development of new DNA-based screening methods for clinical samples shows that arcobacters can often be mistaken for *Campylobacter* spp., and therefore, the potential human pathogenicity of these microbes is likely underestimated as is their role in water- and food-borne disease (Collado and Figueras, 2011; Figueras et al., 2014).

Studies of sewage and sewage-contaminated environmental waters reveal that *Arcobacter* spp. are often found in association with raw (untreated) sewage and even treated effluent water (Stampi et al., 1993; Collado et al., 2008, 2010; Cai et al., 2014). The species *A. butzleri* and *A. cryaerophilus* are the most commonly found in isolation studies, and appear to have high genetic diversity within species (Collado et al., 2008, 2010). The species *A. defluvii* and *A. cloacae* have been recently discovered in sewage samples (Collado et al., 2011; Levican et al., 2013b) as well. A culture-independent analysis of sewage using 454 pyrosequencing

showed that *Arcobacter* populations accounted for approximately 4% of sewage bacterial communities, but had low diversity based on V6 pyrotag amplification (Vandewalle et al., 2012). The dominant V6 pyrotag also could not be mapped to a specific species, as this region has relatively low diversity among eight named arcobacters. So while *Arcobacter* appears to be an important component of sewage communities, relatively little is known about the diversity of these organisms or the ecological niche they may occupy in sewer infrastructure.

Here we provide an in-depth, DNA-based analysis of the *Arcobacter* community from 37 sewage samples collected in the US and Spain. The oligotyping approach sorted over 400,000 sequences into ecologically meaningful subgroups and allowed us to track changes in the *Arcobacter* populations across seasons and geography. Our findings reveal potential new species yet to be cultivated and temperature-based trends in the dominant organisms found in sewage.

MATERIALS AND METHODS

SAMPLE COLLECTION AND PROCESSING

We selected a subset of sewage samples from a larger study that contained a complete set of metadata in order to better assess the ecological factors that contribute to the distribution of total *Arcobacter* and also individual species within and among sewage samples (Tables S1, S2). All sewage samples represent a single replicate taken from municipal wastewater treatment facilities: 36 primary influent (untreated) samples collected from 12 facilities in the US on three occasions (August 2012, January 2013, and April 2013) and one sample collected from Reus, Spain in September 2012 (Table S1). These samples represent a range of geographic location, regional climate, and seasonal variation (Table S2).

Technicians at the US sewage treatment plants shipped samples on ice within 24 h of collection to our laboratory in Milwaukee, WI, USA, for processing. A volume of 25 mL of sewage was filtered (0.22 μm, 47 mm S-Pak® Millipore® filters) for each sample and filters were stored at −80°C. The sample from Reus (Spain) was a composite sample collected overnight from 8:00 p.m. to 8:00 a.m. from the inflow of the WWTP of this city. This sample was immediately taken to the laboratory at the Medical School in Reus where it was filtered. DNA was extracted following the protocol described below provided by the Milwaukee laboratory. The DNA was shipped on ice the same day to Milwaukee.

DNA EXTRACTION, AMPLICON SEQUENCING, AND BIOINFORMATIC PROCESSING

We extracted DNA as previously described (Newton et al., 2013). Briefly, the FastSpin Soil DNA kit (MP Biomedicals, Santa Ana, CA) was employed according to the manufacturer's instructions using the material contained in the crushed filters. The DNA purity and concentration was assessed using the NanoDrop® spectrophotometer (Thermo Scientific, Waltham, MA) and by performing an electrophoresis in 1% TAE agarose gel.

The Josephine Bay Paul Center at the Marine Biological Laboratories in Woods Hole, MA, provided Illumina amplicon sequencing. Primers amplified the V4V5 region of the bacterial 16S rRNA gene, and the Illumina MiSeq platform produced the sequence reads. Primers, sequencing protocols, quality control measures, and bioinformatic trimming procedures for Illumina MiSeq are described in detail elsewhere (Morrison et al., 2013). The Global Alignment for Sequence Taxonomy (GAST) software (Huse et al., 2008) assigned taxonomy to our high-quality reads; this study uses only the sequences that mapped to the genus *Arcobacter*. The sequences obtained in this study are available in the National Center for Biotechnology Information (NCBI) Short Read Archive under accession number SRP047513.

OLIGOTYPING

GAST taxonomic classification of sequence reads (Huse et al., 2008) to the genus *Arcobacter* resulted in 408,878 sequences for oligotyping. We implemented the oligotyping pipeline (Eren et al., 2013) to determine ecologically relevant sequence groupings. Gap characters added to the ends of shorter sequences produced sequences of equal length that are required by the analysis pipeline. The "entropy-analysis" script in the oligotyping pipeline calculated the Shannon entropy at each nucleotide along the length of the sequences. The Shannon entropy provides a measure of nucleotide variation at a given position; sites that have A, G, C, and T present in approximately equal proportions among sequences have the highest entropy values, whereas highly conserved sites have a minimum entropy value near zero. Starting with the highest entropy positions along the length of the sequence, we selected 31 positions (4, 41, 54, 55, 56, 57, 58, 65, 70, 78, 85, 105, 112, 114, 115, 118, 120, 128, 130, 133, 159, 203, 212, 226, 250, 287, 301, 308, 332, 335, 343) over a read length of 375 nucleotides until entropy peaks were eliminated in individual oligotypes. We required each oligotype to have a minimum substantive abundance ($-M$, the abundance of the dominant sequence representing the oligotype) of 408 in order to reduce noise in the dataset and to focus our analysis on the more abundant oligotypes. Oligotyping with no noise filtering produced over 3800 oligotypes; elimination of oligotypes with a minimum substantive abundance of less than 408 reads (equivalent to 0.1% of the total abundance of *Arcobacter* sequence reads in the dataset) resulted in 26 oligotypes. Over 90% of sequences (372,028) were retained in the final analysis; discarded sequences were distributed evenly across samples.

STATISTICAL ANALYSES

We used the vegan (Oksanen et al., 2013) and stats packages in R (R Development Core Team, 2012) for statistical analyses. Hierarchical clustering and non-metric multi-dimensional scaling (NMDS) analyses were based on Bray-Curtis dissimilarities, using oligotype matrix-count data as input. We determined the influence of environmental parameters on clusters produced by NMDS using permutation analysis of variance (ADONIS in the vegan package) with 999 permutations. The non-parametric Spearman *rho* correlation coefficients and corresponding *p* values (cor.test in the stats package) were used to determine the relationships between temperature and oligotype proportion.

PHYLOGENETIC ANALYSES

ClustalW in the MEGA5 package (Tamura et al., 2011) aligned DNA sequences. The 16S rRNA sequences of *Arcobacter* reference

strains represented cultivated species (Levican et al., 2013b; Sasi Jyothsna et al., 2013). We aligned the oligotype representative sequences (375 bp in length) to nearly full-length reference sequences, then trimmed to 375 bases for phylogenetic analysis. The V4V5 amplicons overlap the reference sequences from nucleotides 544–928 based on *E. coli* numbering. The Jukes-Cantor method estimated evolutionary distances, and we generated 1000 replicate trees using the Neighbor-Joining algorithm. *Campylobacter jejuni* served as an outgroup.

RESULTS

SAMPLE ENVIRONMENTAL FACTORS AND DISTRIBUTION OF OLIGOTYPES AMONG SAMPLES

The percentage of the total bacterial community that mapped to the genus *Arcobacter* ranged between 0.8 and 19.6% for most samples, and 27/37 samples had more than 5% *Arcobacter*. Two outlier samples contained 73 and 85% *Arcobacter* (**Figure 1A**). Over 40,000 unique sequences were present in the set of 408,878 total *Arcobacter* sequences. Alignments of the nearly complete (>1400 bp) 16S rRNA gene of *Arcobacter* reference strains (Figure S1) allowed the calculation of the Shannon entropy within the V4V5 region compared to the other variable regions; entropy within the V4V5 amplicon sequences is shown as well. The V4V5 region of *Arcobacter* reference sequences contains many high entropy nucleotide positions, although fewer than in the V2 region. Amplicons have similar high entropy nucleotide positions, but very low entropy also occurs uniformly across the length of the amplicon sequences as well. Resolution of high entropy positions produced 26 relatively abundant oligotypes (**Figure 1B**), while the low entropy nucleotide positions represent the high number of unique sequences that derive from microdiversity within the genus but also from sequencing noise. All 26 oligotypes in the noise-filtered analysis had a significant relative proportion (>0.9%) in at least one sample and appeared across different treatment plants and from different collection dates (**Figure 1B**). Oligotypes are numbered based on their total abundance rank within the dataset (i.e., Oligotype 1 had the highest overall abundance). The sewage *Arcobacter* communities from different sites were dominated by 1–3 oligotypes, with extremely high microdiversity in the minor members (represented by the noise-filtered sequences). Overall, the nine most abundant oligotypes within the *Arcobacter* genus accounted for >80% of the total *Arcobacter* sequences in all samples analyzed. The US samples, which were all untreated sewage, contained ~20 oligotypes (19.8 average, 19.5 median), but the Spain sample had only 11 oligotypes, all of which were also found in US samples. In **Figure 1B**, the oligotype distribution within samples is shown with samples grouped by average site temperature and by sample date within each site. Overall, oligotype distribution showed more similar patterns within sites (ADONIS $r^2 = 0.712$, $p < 0.001$),

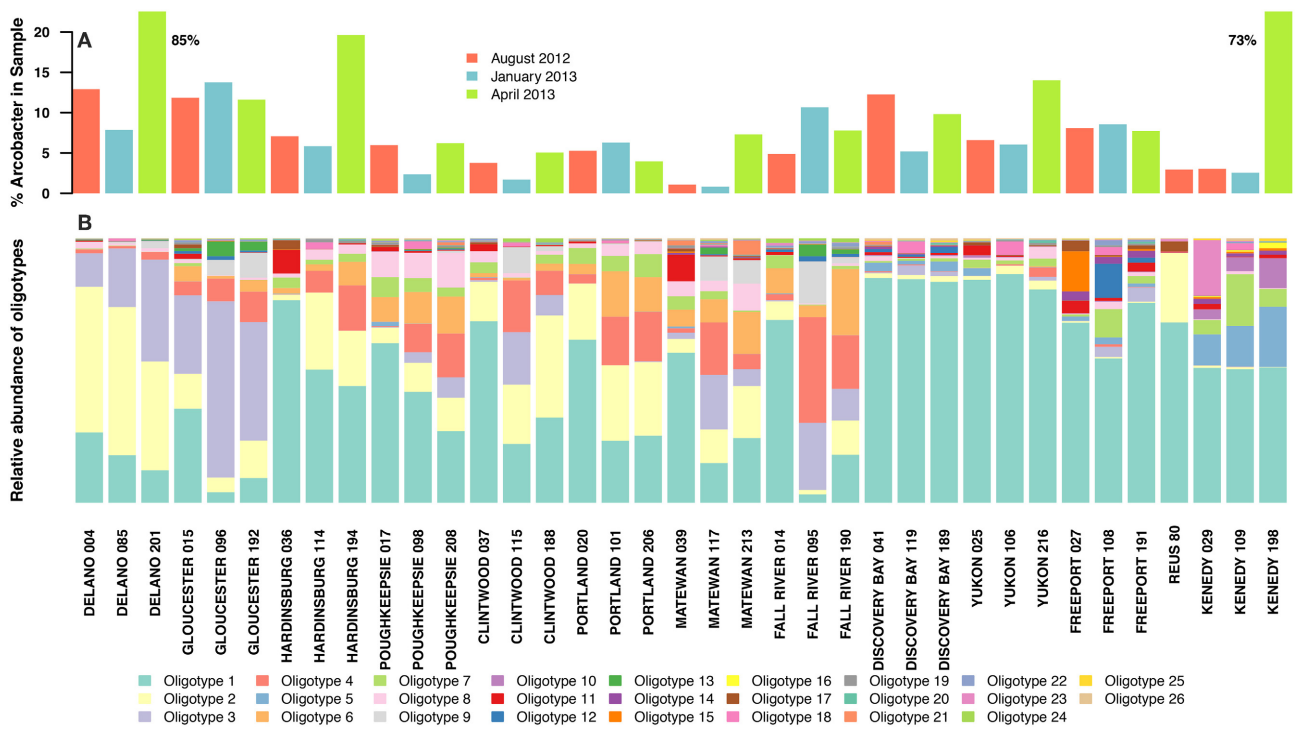


FIGURE 1 | (A) Proportion of sequence reads in each sewage sample that mapped to the genus *Arcobacter*. Samples are color coded by the dates they were sampled: red = August 2012 (September 2012 for the Reus sample), blue = January 2013, green = April 2013. Two samples were outliers with significantly higher *Arcobacter* percentages than the rest of the samples; their values are shown as text next to the bars. **(B)** Proportions of 26 abundant oligotypes generated from sequence reads that mapped to the genus *Arcobacter* using the oligotyping pipeline. Samples are grouped by site, ordered from coldest to warmest average site temperature, then by sample collection date within sites. The legend shows the colors that represent each oligotype in **(B)**; oligotypes are numbered based on the rank of their abundance summed over the whole dataset.

but also appeared to have trends that corresponded to sample temperatures.

TEMPERATURE DYNAMICS OF *ARCOBACTER* OLIGOTYPES

Figure 2A shows a hierarchical clustering analysis based on Bray Curtis dissimilarities of *Arcobacter* population compositions as described by oligotypes. *Arcobacter* populations divided on the basis of the sample temperature, with the division occurring at temperatures higher or lower than 20°C. Samples of similar temperatures taken from the same site also tended to group together. Sample temperatures in the "warm" cluster ranged from 20 to 29.5°C, with two outliers that were 17°C (**Figure 2B**). The "cool" cluster sample temperatures ranged from 9.8 to 19.8°C, with one outlier at 21.2°C.

Oligotype proportion in samples was significantly correlated to temperature for 12 of the 26 oligotypes. Eight of these tended to have a higher proportion with higher temperature (positively correlated), and 4 negatively correlated with higher temperature. The Spearman correlation coefficients and *p* values for all oligotypes are shown in Table S3. No other metadata (total suspended solids, biochemical oxygen demand, total nitrogen, total phosphorus, population size, or average daily flow) had a significant correlation to the proportion of a given oligotype present in a sample (data not shown). The two most abundant oligotypes, which are denoted as "Oligotype 1" and "Oligotype 2" displayed opposing dynamics that coincided with the temperature of a sample; i.e., they correlated positively and negatively with temperature, respectively (**Figure 3**). While Oligotype 1 had by far the highest proportion overall and made up over 50% of almost all the high temperature samples, its relative abundance was notably lower in most lower temperature samples. Oligotype 1 accounted for >80% of the *Arcobacter* sequences represented by oligotypes in the two 17°C samples (Discovery Bay-119 and Yukon-106) that grouped with the "warm" cluster, while the 21.2°C sample from the "cool" cluster (Gloucester-015) had <50% Oligotype 1.

ARCOBACTER SPECIES REPRESENTED BY OLIGOTYPES

Seven of the sequences representing *Arcobacter* oligotypes shared 100% identity with previously characterized *Arcobacter* species based on BLAST comparison of sequences against the NCBI nucleotide database (Table S4). Oligotype 1 and Oligotype 2

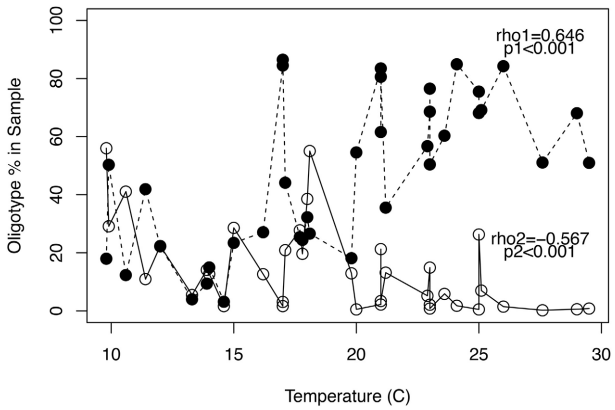


FIGURE 3 | Changes in oligotype proportions with temperature. The two dominant oligotypes based on their relative abundance across all samples showed opposing dynamics with changes in temperature. The proportion of Oligotype (•) increased at temperatures >20°C, while Oligotype (○) proportions decreased above 20°C. The non-parametric correlation coefficient (Spearman's rho) and significance values for the relationships between oligotype proportion and temperature are shown by their respective oligotype. Temperature-proportion correlation coefficients and *p* values for all 26 oligotypes are given in Table S3.

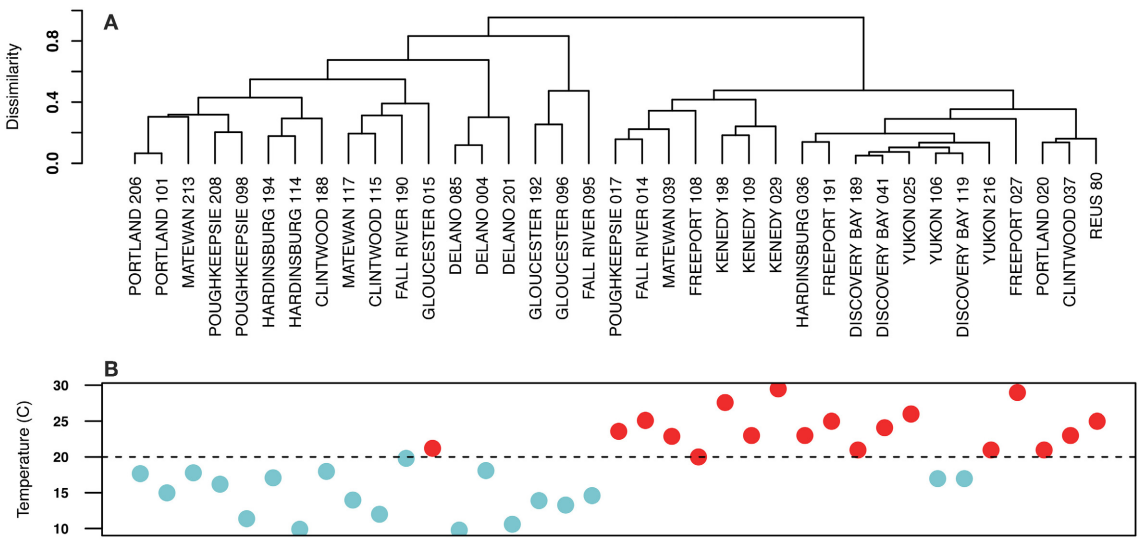


FIGURE 2 | (A) Hierarchical clustering of sewage samples based on Bray Curtis dissimilarities of *Arcobacter* populations. Samples clustered into two distinct groups that were related to the temperature of the sample. The clade of samples on the left was composed of lower temperature samples than the clade on the right. **(B)** Temperatures of the sewage samples in **(A)** are shown below the sample name. Blue dots signify samples with temperatures below 20°C (indicated by the dashed line); red dots represent sample temperatures ≥20°C.

matched exactly to *A. cryaerophilus* subgroups 1B and 1A, whose 16S rRNA genes differ in the V4V5 region by only a single nucleotide. Other oligotypes with 100% identity to the type strains of *Arcobacter* species included Oligotype 4 (*A. suis*), Oligotype 11 (*A. butzleri*), and Oligotype 14 (*A. ellisii*), and Oligotype 17 (*A. cibarius*). Oligotype 23 shared 100% identity with strains of two different species (*A. cloacae* and *A. defluvii*) that were identical in V4V5 region. The majority of oligotypes had no exact matches to the type strains or other strains within cultivated species, although all were at least 98% similar to a characterized *Arcobacter* species. **Figure 4** shows the phylogenetic groupings of the sequences representing the 26 oligotypes in relation to known *Arcobacter* species. Oligotype 3 formed a distinct new clade along with Oligotypes 12 and 13, and Oligotypes 5, 10, 16, and 25 also formed a clade that might represent new species. However, as the phylogenetic analysis was limited to only 375 nucleotides, the groupings of the oligotypes are more illustrative than definitive; full-length sequences would be needed to confirm the true phylogenetic relationships.

DISCUSSION

OLIGOTYPING DISCERNS ECOLOGICALLY RELEVANT PATTERNS WITHIN THE GENUS *ARCOBACTER*

In our study, different *Arcobacter* species were present in higher numbers depending on the sample temperature, similar to a previous study from estuarine water using conventional culturing methods and genetic identification of the isolates (Levican et al., 2014). For instance, Oligotype 11 (*A. butzleri*) was present in almost all samples collected during August but only in a few samples collected during January or April (**Figure 1B**) and correlated positively with higher environmental temperatures (Table S3). Along the same lines, Levican et al. (2014) observed a seasonal distribution of the species *A. butzleri* with a significantly higher recovery during summer. Moreover, in the later study the species *A. cryaerophilus*, *A. skirrowii*, and *A. nitrofigilis* were only isolated from environmental samples when water temperatures were lower (from 7.9°C to 18.2°C); however, the low number of strains recovered did not allow significant correlations to be made between species and either the water temperature or with the culturing approach (Levican et al., 2014). Conversely, the larger dataset used for oligotyping in the present study allowed us to infer a significant correlation between 12 and 26 oligotypes and the environmental temperature.

The fact that the proportions of the most abundant oligotypes varied by site and by temperature suggests that while a set group of organisms may be adapted to the ecological niche represented by a locale, changes in the environment within that system may favor different species at different times. Sewer systems appear to supply a unique niche where *Arcobacter* species thrive. Multiple samples taken from the different WWTPs demonstrate the consistency of community members at each site, but also the seasonal dynamics within populations that occur (Vandewalle et al., 2012). Dominant oligotypes were consistently present in samples collected from the same site, which may indicate a kind of ecological adaptation to general regional conditions such as climate, or more specifically, to the conditions found in particular sewerage systems.

ABUNDANT OLIGOTYPES REPRESENT BOTH CHARACTERIZED AND UNCULTIVATED SPECIES

Oligotyping is often used to compare samples from sites that have obvious ecological differences and where one might expect differentiation of populations based on environmental influences (Eren et al., 2013; Reveillaud et al., 2014). We also used oligotyping to assess the relevant nucleotide signature positions that determine speciation (Eren et al., 2013). Operational Taxonomic Unit analysis can group sequences at a fine level (>97–99% similarity), but the 16S rRNA sequences of known *Arcobacter* species do not vary by a fixed percent. By using changes in the evolutionarily relevant nucleotide positions, we were able to identify significant groupings within the *Arcobacter* genus regardless of the overall degree of sequence similarity. This analysis showed that there are a limited number of dominant ecotypes, despite the high microdiversity within *Arcobacter* sequences.

In a previous study of sewage samples using 454 pyrosequencing of the 60 bp V6 region, a single dominant sequence comprised >80% of the sewage reads that mapped to *Arcobacter* (Vandewalle et al., 2012). The reduced diversity observed in these samples could either be due to a single dominant strain (as was observed for Discovery Bay samples), or because the V6 region is highly conserved among several *Arcobacter* species. The sequence of the dominant V6 pyrotag had a 100% match to *A. cibarius*, *A. cloacae*, *A. cryaerophilus*, *A. defluvii*, *A. skirrowii*, *A. suis*, and *A. venerupis* (data not shown). The genetic information contained within the V4V5 amplicons in this study vs. the V6 region in the previous study allowed better resolution of these reference strains from each other, but still failed to resolve some species (e.g., *A. trophiarum* and *A. thereius*). Amplicon sequencing of the V2 region (which has the highest diversity among named *Arcobacter* species) might therefore reveal even greater *Arcobacter* diversity in sewage, but more importantly, might better resolve the most abundant ecotypes and clarify their relation to known species.

Dominant oligotypes in the sewage samples examined here share 16S rRNA gene sequence identity with named species or ecotypes as well as yet to be characterized, possible new species. Several sewage oligotypes map to arcobacters cultivated from diverse sources: e.g., *A. defluvii* and *A. cloacae* were originally isolated from sewage (Collado et al., 2011; Levican et al., 2013b); *A. venerupis*, *A. suis*, and *A. ellisii* are associated with food products (Figueras et al., 2011; Levican et al., 2012; Hausdorf et al., 2013); and *A. cryaerophilus* and *A. butzleri* are found in diseased animals including humans (Collado and Figueras, 2011).

At least five different groups, each containing multiple oligotypes, appear to represent new uncultivated clades. Eight of the fifteen most abundant oligotype representative sequences had no exact match, and two had closest matches to uncharacterized environmental isolates. The third most abundant had only a 98% match to the closest named species and 100% shared identity with an environmental (non-sewage) isolate. Public sequence databases such as NCBI also contain many 16S rRNA gene sequences of uncultivated and not-yet-described *Arcobacter* species, some of which come from activated sludge and sewage (Collado and Figueras, 2011). Our results demonstrate that oligotyping can be used as a highly reproducible alternative to other sequence grouping methods to elucidate the population diversity

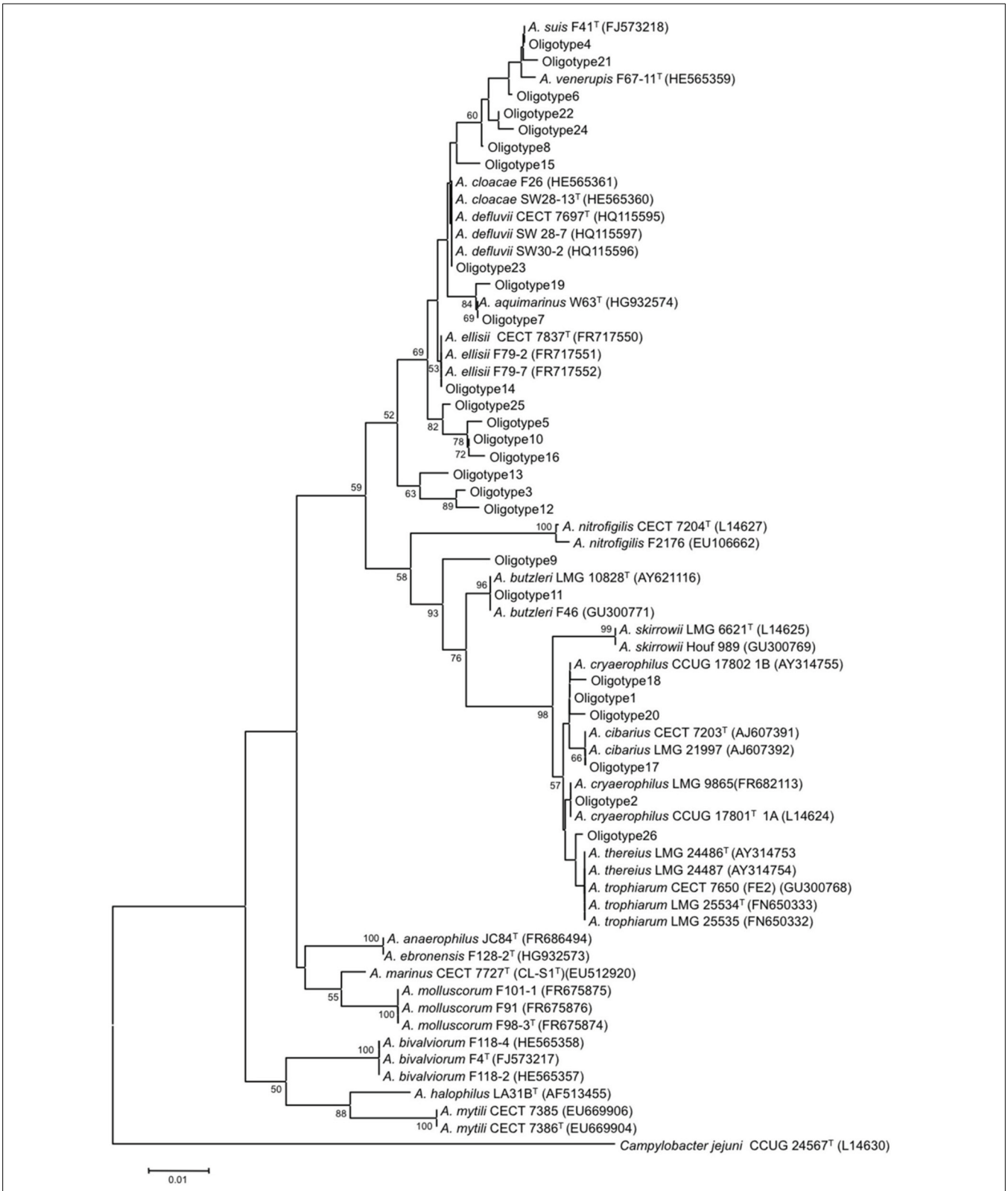


FIGURE 4 | Neighbor-joining tree of *Arcobacter* reference strains and sewage oligotypes. Evolutionary distances were estimated using the Jukes-Cantor algorithm. This tree represents a composite of one thousand replicate trees; branches that occurred >50% of the time are noted at the nodes. The scale bar represents 1% nucleotide substitution. *Campylobacter jejuni* served as the outgroup.

of species within a given genus and may even enable recognition of new species.

The two most abundant oligotypes in this study, Oligotype 1 and Oligotype 2, corresponded with 100% match to two different subgroups of the species *A. cryaerophilus* (subgroups 1B and 1A, respectively). Isolates from subgroup 1B are the second most commonly isolated *Arcobacter* species after *A. butzleri* using traditional culturing and molecular identification methods, but subgroup 1A is rarely recovered in this manner (Collado and Figueras, 2011). Conversely, *A. butzleri*, the predominant species typically recovered from sewage by culturing in other studies (González et al., 2007; Collado et al., 2008, 2010) was only represented by the eleventh most abundant oligotype. The predominance of *A. butzleri* by culturing could be due to the fact that most of studies include an enrichment step that favors the growth of this species; however, this assertion still needs to be experimentally verified (Levican et al., 2014).

IMPLICATIONS FOR CULTURING AND IDENTIFICATION OF NEW *ARCOBACTER* SPECIES

Conventional media and isolation conditions (i.e., incubation temperature, atmosphere, etc.) for the recovery of *Arcobacter* are good for isolating certain species, particularly *A. butzleri* (Collado et al., 2010). However, many uncultured strains appear to be present in sewage as well that are missed by currently used methods (Collado and Figueras, 2011). Regarding the prevalence of *A. cryaerophilus* in different studies, subgroup 1B is much more prevalent than 1A, while both groups have so far been isolated simultaneously only from food products and from animal and human clinical samples (Collado and Figueras, 2011 and references therein). The *A. cryaerophilus* strains isolated from sewage thus far have all belonged to subgroup 1B. It is not clear whether this higher prevalence of subgroup 1B is a consequence of the isolation methods used, or due to specific adaptations of these species to different ecologic niches as observed in the present study.

Different culturing methods and incubation conditions can impact the prevalence and diversity of *Arcobacter* spp. recovered from different sources (food, water, sewage, blood) (Houf et al., 2002; Levican et al., 2014). In fact, the use of direct culturing in parallel to post enrichment cultivation allowed the discovery of the species *A. defluvii* and *A. cloacae* from sewage samples (Collado et al., 2011; Levican et al., 2013b). As previously noted, the large numbers of unclassified sequences indicate that many more potential new species reside in sewage that have yet to be isolated (Collado and Figueras, 2011). Knowing which species or ecotypes grow best at different temperatures may assist in cultivating underrepresented members of the sewage community. Future studies examining both the genetic potential (through genome sequencing) and phenotypic behavior of isolates will help to better determine how these organisms grow and thrive in the sewer systems and how they may impact human health. We lack a full understanding of how the *Arcobacter* spp. in sewage relate to the arcobacters known to be pathogenic to humans and animals. Comparison of sewage oligotypes to the 16S rRNA sequences of clinical isolates may be a first step in this direction, as identification of clinical isolates by sequencing becomes more routine practice (Prouzet-Mauléon et al., 2006; Collado and Figueras, 2011).

OLIGOTYPING TRACKS *ARCOBACTER* POPULATION DYNAMICS

Almost all sewage samples grouped by hierarchical clustering based on *Arcobacter* community similarity separated at a breakpoint of 20°C, which (perhaps not coincidentally) is the temperature that delineates mesophilic bacteria from psychrophilic bacteria (Willey et al., 2008). Only three samples (Gloucester-August 2012, Yukon-January 2013, and Discovery Bay-January 2013) deviated from the group prescribed by their sample temperatures. In many of the sewage samples, the top two most abundant oligotypes had opposite temperature dynamics. Additionally, Oligotype 3, which may represent a new *Arcobacter* species and Oligotype 4 (corresponding to *A. suis*) were nearly absent from the warmest sites. Similar trends in seasonal/temperature-based variation for *Acinetobacter* populations were observed in sewage samples from Milwaukee, WI. The two most abundant *Acinetobacter* V6 pyrotags oscillated in abundance over the course of the year (Vandewalle et al., 2012), and relative proportions of other genera associated with sewage infrastructure (as opposed to the fecal component of sewage) also varied seasonally (Vandewalle et al., 2012).

It is difficult to ascertain how much variation in oligotype distribution is based strictly on sample temperature, as many other factors can contribute to population dynamics. However, observed trends based on both the temperature of the sample at the time of collection and the average site temperature (approximated as the mean of collected sample temperatures) suggest that temperature may contribute significantly to determining, at the very least, the relative proportions of *Arcobacter* species present (D'Sa and Harrison, 2005; Levican et al., 2014). Knowledge of *Arcobacter* population dynamics may lead to a better understanding of risks associated with environmental releases of these organisms in the case of combined/sanitary sewerage overflows (Ashbolt et al., 2010) and provide guidance for better management in food preparations (Van Driessche and Houf, 2008; Kjeldgaard et al., 2009).

SELECTIVE GROWTH OF *ARCOBACTER* SPECIES IN SEWAGE

Arcobacters make up <0.001% of the human gut microbial community (Gevers et al., 2012; Koskey et al., 2014) but they make up a significant portion of sewage samples collected from geographically diverse locations (Shanks et al., 2013; Cai et al., 2014; Koskey et al., 2014). Since *Arcobacter* is found in the human gut, albeit in low proportions, humans may be the source of *Arcobacter* to sewerage systems; however, the sewer pipe environment appears to select for their survival and growth over more dominant gut bacteria, such as the *Lachnospiraceae* (McLellan et al., 2013). Three genera previously recognized to comprise a significant portion of sewage (*Trichococcus, Acinetobacter,* and *Aeromonas*) (Vandewalle et al., 2012) are also present, but in low relative abundance, in human feces (Gevers et al., 2012; Koskey et al., 2014). Thus, the specific ecological conditions present in sewage infrastructure, generally speaking, provide an ideal niche for certain organisms, not only to thrive, but also to maintain diversity within their own populations.

The factors contributing to the survival and growth of different arcobacters are of interest because some species have been identified as emerging human pathogens (Prouzet-Mauléon et al., 2006; Collado and Figueras, 2011; Figueras et al., 2014). Species sharing genetic similarity to pathogenic strains have also been isolated from environmental waters impacted by sewage inputs (Fong et al., 2007; Collado et al., 2010) or detected in impacted waters by molecular analyses (Collado et al., 2008; Lee et al., 2012), suggesting that at least some sewage-based *Arcobacter* species are viable after sewage releases. These organisms can survive or grow at temperatures found in sewerage systems, a range of environmental water temperatures (Levican et al., 2014), and laboratory isolation conditions, which may be ≥30°C (Stampi et al., 1993; Prouzet-Mauléon et al., 2006; Levican et al., 2014). Thus, although different *Arcobacter* species may have relatively high or low optimum growth temperatures, many seem to have a wide range of survival temperatures (D'Sa and Harrison, 2005; Van Driessche and Houf, 2008). If pathogenic *Arcobacter* strains also possess this trait, they may pose a threat even at very low abundance. These findings further stress the need to better understand the genetic crossovers between human and animal pathogenic strains, sewage ecotypes, and cultured isolates in order to ascertain risks associated with *Arcobacter* species.

The results of our study can be used as a general strategy for interpreting sequence data from populations of environmental bacteria, as oligotyping provides high resolution among species, even without full-length sequences. Here we show an approach that allows differentiation of known and unknown species, and also provides information on how environmental factors can modulate the presence and relative abundance of different ecotypes. Oligotyping may provide the means to establish the links between *Arcobacter* communities from food production, water sources, sewage, and diseased humans and animals, in order to better discern patterns of survival, growth, and infection.

ACKNOWLEDGMENTS

This work was funded by a grant to Sandra L. McLellan from the National Institutes of Health (1R01-AI091829) and was also supported in part by funding to María José Figueras from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 311846 and by project AGL2011-30461-C02-02 by the Ministerio de Ciencia e Innovación (Spain). Arturo Levican was supported by a doctoral grant from the Universitat Rovira i Virgili (Spain), CONICYT (Chile) for the Postdoctoral research project FONDECYT 3140296, and by CONICYT/FONDAP grant number 15110027. The authors are solely responsible for the content of this publication. It does not represent the opinion of the European Commission. The European Commission is not responsible for any use that might be made of data appearing therein.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2014.00525/abstract

REFERENCES

Ashbolt, N. J., Schoen, M. E., Soller, J. A., and Roser, D. J. (2010). Predicting pathogen risks to aid beach management: the real value of quantitative microbial risk assessment (QMRA). *Water Res.* 44, 4692–4703. doi: 10.1016/j.watres.2010.06.048

Cai, L., Feng, J., and Zhang, T. (2014). Tracking human sewage microbiome in a municipal wastewater treatment plant. *Appl. Microbiol. Biotechnol.* 98, 3317–3326. doi: 10.1007/s00253-013-5402-z

Collado, L., and Figueras, M. J. (2011). Taxonomy, epidemiology and clinical relevance of the genus *Arcobacter*. *Clin. Microbiol. Rev.* 24, 174–192. doi: 10.1128/CMR.00034-10

Collado, L., Inza, I., Guarro, J., and Figueras, M. J. (2008). Presence of *Arcobacter* spp. in environmental waters correlates with high levels of fecal pollution. *Environ. Microbiol.* 10, 1635–1640. doi: 10.1111/j.1462-2920.2007.01555.x

Collado, L., Kasimir, G., Perez, U., Bosch, A., Pinto, R., Saucedo, G., et al. (2010). Occurrence and diversity of *Arcobacter* spp. along the Llobregat River catchment, at sewage effluents and in a drinking water treatment plant. *Water Res.* 44, 3696–3702. doi: 10.1016/j.watres.2010.04.002

Collado, L., Levican, A., Perez, J., and Figueras, M. J. (2011). *Arcobacter defluvii* sp. nov., isolated from sewage samples. *Int. J. Syst. Evol. Microbiol.* 61, 2155–2161. doi: 10.1099/ijs.0.025668-0

Douidah, L., de Zutter, L., Baré, J., De Vos, P., Vandamme, P., Vandenberg, O., et al. (2012). Occurrence of putative virulence genes in *Arcobacter* species isolated from humans and animals. *J. Clin. Microbiol.* 50, 735–741. doi: 10.1128/JCM.05872-11

D'Sa, E., and Harrison, A. (2005). Effect of pH, NaCl content, and temperature on growth and survival of *Arcobacter* spp. *J. Food Prot.* 68, 18–25.

Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., et al. (2013). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* 4, 1111–1119. doi: 10.1111/2041-210X.12114

Figueras, M. J., Levican, A., Collado, L., Isabel Inza, M., and Yustes, C. (2011). *Arcobacter ellisii* sp. nov., isolated from mussels. *Syst. Appl. Microbiol.* 34, 414–418. doi: 10.1016/j.syapm.2011.04.004

Figueras, M. J., Levican, A., Pujol, I., Ballester, F., Rabada Quilez, M. J., and Gomez-Bertomeu, F. (2014). A severe case of persistent diarrhoea associated with *Arcobacter cryaerophilus* but attributed to *Campylobacter* sp. and a review of the clinical incidence of *Arcobacter*. *New Microbes New Infect.* 2, 31–37. doi: 10.1002/2052-2975.35

Fong, T. T., Mansfield, L. S., Wilson, D. L., Schwab, D. J., Molloy, S. L., and Rose, J. B. (2007). Massive microbiological groundwater contamination associated with a waterborne outbreak in Lake Erie, South Bass Island, Ohio. *Environ. Health Perspect.* 115, 856–864. doi: 10.1289/ehp.9430

Gevers, D., Knight, R., Petrosino, J. F., Huang, K., McGuire, A. L., Birren, B. W., et al. (2012). The human microbiome project: a community resource for the healthy human microbiome. *PLoS Biol.* 10:e1001377. doi: 10.1371/journal.pbio.1001377

González, A., Botella, S., Montes, R., Moreno, Y., and Ferrús, M. (2007). Direct detection and identification of *Arcobacter* species by multiplex PCR in chicken and wastewater samples from Spain. *J. Food Prot.* 70, 341–347.

Hausdorf, L., Neumann, M., Bergmann, I., Sobiella, K., Mundt, K., Froehling, A., et al. (2013). Occurrence and genetic diversity of *Arcobacter* spp. in a spinach-processing plant and evaluation of two *Arcobacter*-specific quantitative PCR assays. *Syst. Appl. Microbiol.* 36, 235–243. doi: 10.1016/j.syapm.2013.02.003

Houf, K., De Zutter, L., Van Hoof, J., and Vandamme, P. (2002). Assessment of the genetic diversity among arcobacters isolated from poultry products by using two PCR-based typing methods. *Appl. Environ. Microbiol.* 68, 2172–2178. doi: 10.1128/AEM.68.5.2172-2178.2002

Huse, S. M., Dethlefsen, L., Huber, J. A., Welch, D. M., Relman, D. A., and Sogin, M. L. (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet.* 4:e1000255. doi: 10.1371/journal.pgen.1000255

Kjeldgaard, J., Jørgensen, K., and Ingmer, H. (2009). Growth and survival at chiller temperatures of *Arcobacter butzleri*. *Int. J. Food Microbiol.* 131, 256–259. doi: 10.1016/j.ijfoodmicro.2009.02.017

Koskey, A., Fisher, J., Eren, A., Ponce Terashima, R., Reis, M., Blanton, R., et al. (2014). *Blautia* and *Prevotella* sequences distinguish human and animal fecal pollution in Brazil surface waters. *Environ. Microbiol.* doi: 10.1111/1758-2229.12189. [Epub ahead of print].

Lee, C., Agidi, S., Marion, J. W., and Lee, J. (2012). *Arcobacter* in Lake Erie beach waters: an emerging gastrointestinal pathogen linked with human-associated fecal contamination. *Appl. Environ. Microbiol.* 78, 5511–5519. doi: 10.1128/AEM.08009-11

Levican, A., Alkeskas, A., Günter, C., Forsythe, S. J., and Figueras, M. J. (2013a). Adherence to and invasion of human intestinal cells by *Arcobacter* species and their virulence genotypes. *Appl. Environ. Microbiol.* 79, 4951–4957. doi: 10.1128/AEM.01073-13

Levican, A., Collado, L., Aguilar, C., Yustes, C., Dieguez, A. L., Romalde, J. L., et al. (2012). *Arcobacter bivalviorum* sp. nov. and *Arcobacter venerupis* sp. nov., new species isolated from shellfish. *Syst. Appl. Microbiol.* 35, 133–138. doi: 10.1016/j.syapm.2012.01.002

Levican, A., Collado, L., and Figueras, M. J. (2013b). *Arcobacter cloacae* sp. nov. and *Arcobacter suis* sp. nov., two new species isolated from food and sewage. *Syst. Appl. Microbiol.* 36, 22–27. doi: 10.1016/j.syapm.2012.11.003

Levican, A., Collado, L., Yustes, C., Aguilar, C., and Figueras, M. J. (2014). Higher water temperature and incubation under aerobic and microaerobic conditions increase the recovery and diversity of *Arcobacter* spp. from shellfish. *Appl. Environ. Microbiol.* 80, 385–391. doi: 10.1128/AEM.03014-13

McLellan, S. L., Newton, R. J., Vandewalle, J. L., Shanks, O. C., Huse, S. M., Eren, A. M., et al. (2013). Sewage reflects the distribution of human faecal *Lachnospiraceae*. *Environ. Microbiol.* 15, 2213–2227. doi: 10.1111/1462-2920.12092

Morrison, H., Grim, S., Vineis, J., and Sogin, M. L. (2013). 16S amplicon fusion primers and protocol for Illumina platform sequencing. *figshare*. doi: 10.6084/m9.figshare.833944. [Epub ahead of print].

Newton, R. J., Bootsma, M. J., Morrison, H. G., Sogin, M. L., and McLellan, S. L. (2013). A microbial signature approach to identify fecal pollution in the waters off an urbanized coast of Lake Michigan. *Microb. Ecol.* 65, 1011–1023. doi: 10.1007/s00248-013-0200-9

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., et al. (2013). *Vegan: Community Ecology Package*. R package version 2.0–8. Available online at: http://CRAN.R-project.org/package=vegan

Prouzet-Mauléon, V., Labadi, L., Bouges, N., Ménard, A., and Mégraud, F. (2006). *Arcobacter butzleri*: underestimated enteropathogen. *Emerg. Infect. Dis.* 12, 307–309. doi: 10.3201/eid1202.050570

R Development Core Team. (2012). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org

Reveillaud, J., Maignien, L., Eren, A. M., Huber, J. A., Apprill, A., Sogin, M. L., et al. (2014). Host-specificity among abundant and rare taxa in the sponge microbiome. *ISME J.* 8, 1198–1209. doi: 10.1038/ismej.2013.227

Sasi Jyothsna, T. S., Rahul, K., Ramaprasad, E. V., Sasikala, C., and Ramana, C. V. (2013). *Arcobacter anaerophilus* sp. nov., isolated from an estuarine sediment and emended description of the genus *Arcobacter*. *Int. J. Syst. Evol. Microbiol.* 63, 4619–4625. doi: 10.1099/ijs.0.054155-0

Shanks, O. C., Newton, R. J., Kelty, C. A., Huse, S. M., Sogin, M. L., and McLellan, S. L. (2013). Comparison of the microbial community structures of untreated wastewaters from different geographic locales. *Appl. Environ. Microbiol.* 79, 2906–2913. doi: 10.1128/AEM.03448-12

Stampi, S., Varoli, O., Zanetti, F., and De Luca, G. (1993). *Arcobacter cryaerophilus* and thermophilic campylobacters in a sewage treatment plant in Italy: two secondary treatments compared. *Epidemiol. Infect.* 110, 633–639. doi: 10.1017/S0950268800051050

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739. doi: 10.1093/molbev/msr121

Vandamme, P., Vancanneyt, M., Pot, B., Mels, L., Hoste, B., Dewettinck, D., et al. (1992). Polyphasic taxonomic study of the emended genus *Arcobacter* with *Arcobacter butzleri* comb. nov. and *Arcobacter skiwowii* sp. nov., an aerotolerant bacterium isolated from veterinary specimens. *Int. J. Syst. Bacteriol.* 42, 344–356.

Vandewalle, J. L., Goetz, G. W., Huse, S. M., Morrison, H. G., Sogin, M. L., Hoffmann, R. G., et al. (2012). *Acinetobacter, Aeromonas* and *Trichococcus* populations dominate the microbial community within urban sewer infrastructure. *Environ. Microbiol.* 14, 2538–2552. doi: 10.1111/j.1462-2920.2012.02757.x

Van Driessche, E., and Houf, K. (2008). Survival capacity in water of *Arcobacter* species under different temperature conditions. *J. Appl. Microbiol.* 105, 443–451. doi: 10.1111/j.1365-2672.2008.03762.x

Willey, J. M., Sherwood, L., Woolverton, C. J., and Prescott, L. M. (2008). *Prescott, Harley, and Klein's Microbiology.* New York, NY: McGraw-Hill Higher Education.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 07 August 2014; accepted: 21 September 2014; published online: 07 November 2014.

*Citation: Fisher JC, Levican A, Figueras MJ and McLellan SL (2014) Population dynamics and ecology of Arcobacter in sewage. Front. Microbiol. **5**:525. doi: 10.3389/fmicb.2014.00525*

This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology.

frontiers in MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 13 November 2014
doi: 10.3389/fmicb.2014.00563

Oligotyping reveals community level habitat selection within the genus *Vibrio*

Victor T. Schmidt[1,2], Julie Reveillaud[1†], Erik Zettler[3], Tracy J. Mincer[4], Leslie Murphy[1] and Linda A. Amaral-Zettler[1,5]*

[1] Marine Biological Laboratory, Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Woods Hole, MA, USA
[2] Department of Ecology and Evolutionary Biology, Brown University, Providence, RI, USA
[3] Sea Education Association, Woods Hole, MA, USA
[4] Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, MA, USA
[5] Department of Earth, Environmental and Planetary Sciences, Brown University, Providence, RI, USA

Edited by:
Rachel Susan Poretsky, University of Illinois at Chicago, USA

Reviewed by:
Patrick K. H. Lee, City University of Hong Kong, China
Migun Shakya, Dartmouth College, USA

***Correspondence:**
Linda A. Amaral-Zettler, Marine Biological Laboratory, Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, 7 MBL St., Woods Hole, MA 02543, USA
e-mail: amaral@mbl.edu

†Present address:
Julie Reveillaud, UMR 6197-Laboratoire de Microbiologie des Environnements Extrêmes (LM2E), Institut Universitaire Européen de la Mer (IUEM), Université de Bretagne Occidentale, Plouzané, France;
CNRS, UMR6197-Laboratoire de Microbiologie des Environnements Extrêmes (LM2E), Institut Universitaire Européen de la Mer (IUEM), Place Nicolas Copernic, Plouzané, France;
Ifremer, Centre de Brest, REM/EEP/LM2E, ZI de la Pointe du Diable, CS 10070, Plouzané, France

The genus *Vibrio* is a metabolically diverse group of facultative anaerobic bacteria, common in aquatic environments and marine hosts. The genus contains several species of importance to human health and aquaculture, including the causative agents of human cholera and fish vibriosis. Vibrios display a wide variety of known life histories, from opportunistic pathogens to long-standing symbionts with individual host species. Studying *Vibrio* ecology has been challenging as individual species often display a wide range of habitat preferences, and groups of vibrios can act as socially cohesive groups. Although strong associations with salinity, temperature and other environmental variables have been established, the degree of habitat or host specificity at both the individual and community levels is unknown. Here we use oligotyping analyses in combination with a large collection of existing *Vibrio* 16S ribosomal RNA (rRNA) gene sequence data to reveal patterns of *Vibrio* ecology across a wide range of environmental, host, and abiotic substrate associated habitats. Our data show that individual taxa often display a wide range of habitat preferences yet tend to be highly abundant in either substrate-associated or free-living environments. Our analyses show that *Vibrio* communities share considerable overlap between two distinct hosts (i.e., sponge and fish), yet are distinct from the abiotic plastic substrates. Lastly, evidence for habitat specificity at the community level exists in some habitats, despite considerable stochasticity in others. In addition to providing insights into *Vibrio* ecology across a broad range of habitats, our study shows the utility of oligotyping as a facile, high-throughput and unbiased method for large-scale analyses of publically available sequence data repositories and suggests its wide application could greatly extend the range of possibilities to explore microbial ecology.

Keywords: oligotyping, *Vibrio* ecology, host-microbe interactions, illumina sequencing, 16S rRNA analysis, plastisphere, aquaculture pathogens, meta-analysis

INTRODUCTION

Vibrio is a ubiquitous, speciose and commercially important bacterial genus with both host associated and free-living representatives. Several species within the genus are pathogenic to humans and animals. *Vibrio cholerae* has caused six historic and one ongoing cholera pandemic, and countless epidemics (Mutreja et al., 2011) including a recent outbreak in Haiti that killed more than 8000 people (Chin et al., 2011). *Vibrio* pathogens are also important to the aquaculture industry, where they inflict costly losses on farmed fish, mollusks and shrimp (Austin and Austin, 2007), limiting the development of an industry poised to help bridge global food gaps and preserve wild fisheries (FAO, 2012). Due to their importance to human and animal welfare, and the ease with which they are cultured, vibrios are relatively well studied, with over 570 publicly available annotated genomes and over 64,000 16S rRNA gene sequences annotated as vibrios in GenBank as of March 2014. *Vibrio* therefore represents an ideal candidate for applying new analytical approaches using pre-existing data to gain further insights into the ecology of the genus.

Making sense of *Vibrio* ecology has been a challenge, owing in part to its complex life history, its capacity to partition resources, and a strong propensity for lateral gene transfer between closely related species (Hunt et al., 2008; Cordero et al., 2012). The complexity of the genus is well illustrated by the diversity of its life histories. On one hand, *Vibrio* has an average of 11 rRNA gene copies, allowing for rapid growth rates under good conditions (Heidelberg et al., 2000), suggestive of r-selected taxa, which can rapidly multiply given favorable conditions (Andrews and Harris,

1986). Conversely, bioluminescent vibrios have formed symbiotic relationships with squid and anglerfish over evolutionary time scales (Ruby and Nealson, 1976), suggestive of a more stable K-selected strategy. Some flexibility between r vs. K strategies may even exist within fine scale taxonomic categories, as environmental conditions such as pH, concentrations of bile, bicarbonate and nutrients may trigger rapid growth within a host in a formally dormant environmental bacterium (Skorupski and Taylor, 2013). To further complicate the ecology of individual *Vibrio* species, recent experiments indicate that disparate species can form socially cohesive groups, taking advantage of their propensity for exchanging genetic elements to confer greater antibiotic resistance among closely related strains, and to likely regulate virulence (Cordero et al., 2012).

Vibrios also seem to be highly variable in habitat preference. Traditionally *Vibrio* life history has been studied in association with multicellular marine hosts, including fish, mollusks, and a wide range of zooplankton (Liston, 1956; Aiso et al., 1968), yet they can also exist in the ambient aquatic environment, associated with plastic particles (Zettler et al., 2013), or phytoplankton blooms (Gilbert et al., 2012). Whether individual *Vibrio* species, or communities of vibrios, are specific to particular habitats is an open question, and distinguishing specialized associations from opportunistic colonization is challenging (Takemura et al., 2014). Host specificity has been observed in other bacterial genera, including *Blautia* (Eren et al., 2014) and *Nitrospira* (Reveillaud et al., 2014), but because *Vibrio* is abundant in both host and environmental habitats, distinguishing established host associations from incidental or ephemeral colonization from surrounding habitats is difficult.

Because vibrios are diverse in their habitat preferences and potentially act as socially cohesive units, large-scale analysis of *Vibrio* community structure across habitats may provide important insights into its ecology. Analyses of this type have historically involved culturing isolates from target habitats and sequencing multiple loci in order to gain sufficient taxonomic resolution within a sample, requiring the use of *Vibrio*-specific primers (Preheim et al., 2011a; Szabo et al., 2013), and making-large scale, non-targeted, multi-habitat analyses challenging and costly. More recently, oligotyping rRNA gene amplicon sequences affords extremely high resolution analysis of community structure by selecting a subset of highly informative nucleotide sites within single loci of 16S rRNA gene hypervariable regions alone (Eren et al., 2013a, 2014; Reveillaud et al., 2014). At the same time, repositories of 16S rRNA gene sequences have grown in size and scope. The Visualization and Analysis of Microbial Population Structures (VAMPS) database is one such repository that contains over 1000 datasets representing hundred of millions of publically available 16S rRNA gene sequences (Huse et al., 2014).

The aim of the present study was to use oligotyping to explore the distribution of *Vibrio* communities in a range of substrate-associated (both biotic and abiotic) and free-living aquatic environments. We used this method to test the hypothesis that distinct *Vibrio* communities occur in different habitats, and are characterized by clear distinctions between host habitats and their surrounding water.

MATERIALS AND METHODS

SEQUENCE COLLECTION

The VAMPS database houses 16S rRNA gene amplicon sequence data projects from a wide variety of environmental and host-associated habitats. We identified seven existing projects to target for analyses of *Vibrio* diversity, representing free-living and host (abiotic and biotic) substrate associated habitats (**Table 1**). We chose projects with the occurrence of at least three samples with >300 sequences identified as *Vibrio* by the Global Alignment for Sequence Taxonomy (GAST) pipeline (Huse et al., 2008) in the VAMPS database. In rare cases, a sample was

Table 1 | Overview of projects used from the VAMPS database with their original citation.

VAMPS project	Habitat	Sample number	Mean *Vibrio* relative abundance	Geographic location	Salinity	Citation or SRA BioProject accession number
ICM_PML_Bv6	Seawater	3	0.31 (SE 0.13)	English Channel	Marine	Gilbert et al., 2012
LAZ_MHB_Bv6	Seawater	14	0.0017 (SE 0.00021)	Northwestern Atlantic	Marine	SRP049014
LAZ_NMS_Bv6	Saltmarsh	11	0.173 (SE 0.019)	New England, USA	Mixed	SRP059013
SLM_NIH_Bv6	PAH spiked sand	11	0.012 (SE 0.002)	Gulf of Mexico	Mixed	Kappell et al., 2014
LAZ_SEA_Bv6	Seawater - associated with plastic	32	0.0036 (SE 0.00058)	Northwestern Atlantic	Marine	SRP026054
LAZ_SEA_Bv6	Plastic-associated	27	0.0032 (SE 0.0005)	Northwestern Atlantic	Marine	SRP026054
JCR_SPO_Bv6	Seawater - associated with sponge	11	0.055 (SE 0.023)	Northeastern Atlantic	Marine	Reveillaud et al., 2014
JCR_SPO_Bv6	Sponge-associated	49	0.09 (SE 0.014)	Northeastern Atlantic	Marine	Reveillaud et al., 2014
VTS_MIC_Bv6	Aquarium water - associated with fish	31	0.016 (SE 0.0037)	MBL, Woods Hole, USA	Mixed	SRP047374 (but see Supplementary Data Sheet 1)
VTS_MIC_Bv6	Fish-associated	20	0.3 (SE 0.039)	MBL, Woods Hole, USA	Mixed	SRP047374 (but see Supplementary Data Sheet 1)

The mean Vibrio relative abundance across all samples in a given project is shown with standard error. For sequences first published by this study the accession numbers for that project's NCBI Sequence Read Archive (SRA) BioProject is given.

included that had less than 300 sequences in order to increase the number of samples for habitats with low sample numbers, or when water associated with a specific substrate was of interest but possessed low *Vibrio* sequence representation. All projects were sequenced at the Marine Biological Laboratory (MBL) Keck sequencing facility on an Illumina HiSeq 1000, and employed identical protocols in the generation and sequencing of 16S rRNA gene sequences, including the same primer cocktails to target the V6 hypervariable region, as described elsewhere (Eren et al., 2013b). Each project also followed the same standard MBL sequence analysis pipeline, where only perfectly overlapping paired-ends reads with zero mismatches passed quality filters (Eren et al., 2013b), and taxonomic assignment was done using the GAST pipeline (Huse et al., 2008). Both quality filtering and taxonomic assignments had already been made for all sequences across all projects as part of standard VAMPS protocols, and we were therefore able to directly download sequences using VAMPS's "data export -> TaxBySeq" feature, using the query "Bacteria; Proteobacteria; Gammaproteobacteria; Vibrionales; Vibrionaceae; *Vibrio*."

Although sequence generation was identical for all projects, sample collection varied. Saltmarsh water sample collection (LAZ_NMS_Bv6, this study) employed automated collection via the Phytoplankton Sampler (PPS) (McLane Research Laboratories, Inc., East Falmouth, MA) that filters 500 mL of water through a 0.65 μm flat filter (EMD Millipore Durapore PVDP hydrophilic membrane filters) (Billerica, MA) twice a day, and stores the filter in RNAlater (Qiagen, Valencia, CA) buffer. As part of a broader project to understand microbial populations in coastal environments, we deployed PPS samplers in tidal creeks (Mill and Salt Ponds, Nauset Marsh System, MA) that receive daily tidal fluxes from the Atlantic Ocean off Cape Cod, MA. DNA extraction and purification of filters used a modified salt precipitation method with bead-beating (Gentra Puregene, Qiagen, Valencia, CA). The Rhode Island Department of Environmental Management (RIDEM) collected seawater samples from the northeastern reach of Narragansett Bay called Mount Hope Bay, MA (LAZ_MHB_Bv6) as part of their monthly water quality survey for shellfish safety. Samples collected manually from surface waters in sterile 1 L polyethylene terephthalate (PET) bottles at 17 stations throughout the 36 km^2 bay were subsequently filtered through 0.22 μm polyethersulphone membrane Sterivex filters (Millipore, Billerica, MA) followed by DNA extraction as above. Collection details for other samples and metadata are found in respective publications (**Table 1**).

Samples from fish and fish tanks (VTS_MIC_Bv6) were collected as part of an experiment to understand the role of salinity and external microbiota on fish microbiomes (Schmidt et al., Submitted) (Supplementary Data Sheet 1). We acclimated ∼1 inch Black Molly fish (*Poecilia sphenops*) to four salinity levels (salinities 0, 5, 18, and 30) over 30 days using nanopure water and Instant Ocean® (Blacksburg, VA) salt mix, then maintained each fish at target salinity for 12 days. Each salinity treatment contained four independent tanks, each with two fish. For our analyses here, we grouped salinities 0 and 5 (FreshwaterFish) and salinities 18 and 30 (MarineFish). After 12 days we euthanized fish in 1 mg/mL MS-222 and homogenized the entire fish. We then extracted microbial gDNA from the homogenate using a modified Gentra Puregene Yeast/Bac (Qiagen, Valencia, CA) extraction protocol (Supplementary Data Sheet 1). We collected microbial communities from tank water using sterile 1 L PET bottles, and extracted gDNA according to protocols outlined above for LAZ_MHB_Bv6 samples.

OLIGOTYPE GENERATION AND ANALYSIS

Oligotyping is a supervised method that allows the identification of closely related but distinct bacterial taxa in high-throughput sequencing datasets of marker genes. This novel bioinformatics approach is capable of uncovering ecological patterns of microbial communities at finer scales than previously possible with *de novo* approaches (Eren et al., 2013a). Oligotying exploits the fact that some positions within a DNA marker sequence are more ecologically informative than others. The method identifies highly variable locations using Shannon entropy (that is, "entropy components"), and uses only these positions to discriminate ecological units, so called oligotypes. This process reduces the impact of noise caused by sequencing error by relying on only a small number of nucleotide positions, discarding the redundant parts of reads for the identification of oligotypes. The open-source pipeline for oligotyping is available from http://oligotyping.org. This method has been used previously to identify *Gardnerella* distributions in vaginal samples (Eren et al., 2011), *Nitrospira* specificity in sponges (Reveillaud et al., 2014), and *Blautia* specificity in animal hosts (Eren et al., 2014).

In order to get the best possible insights into *Vibrio* oligotype distributions across habitat types, we grouped samples into three broad analysis groupings; substrate (host) associated habitats only, substrate habitats along with their surrounding water samples, and environmental and substrate associated habitats (**Table 2**). First, we analyzed only substrate-associated samples (fish, sponge and our abiotic substrate—plastic marine debris). We subsampled these datasets to the median *Vibrio* sequence count of 30,000 prior to analysis in order to minimize the range in initial sequencing depth between samples, which can otherwise reduce the entropy value of discriminating points in datasets with much lower sequence counts. We then processed them through the oligotyping pipeline, as described in Eren et al. (2013a). To establish entropy components that fully decomposed our sequence data, we started with the strongest two components, then manually chose the next component that best removed remaining entropy in the resulting oligotypes, and re-ran the analysis. This iterative process yielded a final 12 entropy points that fully decomposed our sequences. We allowed for oligotypes to occur in only a single sample (-s 1) but discarded them if they did not represent at least 0.5% of the relative abundance of that sample (–a 0.5). Not all samples contained 30,000 sequences, and sequencing depth ranged from 83 to 30,000 after rarefaction (**Table 2**).

Global alignment prior to oligotyping for short Illumina reads is unnecessary, as positional shifts in sequencing reads due to natural indels will produce entropy peaks at the position of insertion or deletion (and subsequent positions), and the decomposition of the dataset based on any of these peaks will eventually result in the same oligotypes as if they would have been previously aligned.

Table 2 | Description of samples for each analysis grouping.

Analysis grouping	Analysis grouping/ Figure	Number of samples	Subsampled sequence depth range	Oligotypes before (after) quality filtering	Percentage of reads represented by top 5% (10%) oligotypes	Components (position in alignment)
All substrates	1	104	83–30,000	882 (74)	76 (94)	13, 15, 20, 21, 22, 23, 25, 31, 32, 45, 50, 55
Plastics and surrounding seawater	2C	71	83–13,359	415 (71)	90 (96)	13, 15, 20, 21, 22, 23, 25, 31, 32, 45, 50, 55
Sponges and surrounding seawater	2B	58	1822–10,000	681 (45)	71 (90)	13, 15, 20, 21, 22, 23, 25, 31, 32, 45, 50, 55
Fish and surrounding water	2A	51	636–85,000	604 (21)	80 (97)	13, 15, 20, 21, 22, 23, 25, 31, 32, 45, 50, 55
Mixed habitat	4	179	83–20,000	1452 (99)	65 (90)	13, 15, 20, 21, 22, 23, 25, 31, 32, 37, 45, 50, 55

Some samples were included in more than a single grouping.

Furthermore, V6 primers target a hypervariable region with few insertions or deletions, and Illumina technology does not have indel error issues. We therefore did not create an alignment prior to entropy analyses. Instead, we aligned sequences at their 3′ end and padded any length discrepancies with gaps at the 5′ end prior to entropy analysis, as detailed in Eren et al. (2013a). We do note that a single Oligotype, Oligotype 10, was not fully decomposed (Supplementary Data Sheet 1). We note that this oligotype varies widely from all other *Vibrio* sequences in this study, and would require an additional 4 entropy components to fully decompose. Furthermore, it occurred in high abundance only in the Sand-PAH habitat. We make no conclusions about this oligotype across any habitat.

Next, we examined each substrate or host sample alongside its respective water sample. For fish and plastics, water samples were directly associated with the host, and collected at the same time and place as host material. Sponge and corresponding seawater were collected simultaneously, although not all sponge samples have a corresponding water samples (see Reveillaud et al., 2014). For each analysis, we subsampled *Vibrio* sequences to the median sequencing depth. The same 12 entropy components and oligotyping parameters as above fully decomposed all but one oligotype, and were therefore used again in this analysis.

We then analyzed oligotype distributions across the broadest range of samples and projects included in this study in a single analysis using the same methods and 12 entropy components, with one additional component added to fully resolve novel oligotypes from additional samples (13 components total). The added samples included water samples from saltmarshes (Saltmarsh), seawater from a large coastal bay (Seawater), open ocean seawater (Seawater), and sand samples from oiled beaches in the Gulf of Mexico inoculated with Polycyclic Aromatic Hydrocarbons (PAHs) (Sand-PAH) (**Tables 1**, **2**). As above, samples were subsampled down to the median value of 20,000. An interactive html file of the results from this oligotyping analysis grouping (Mixed Habitat) is included in the Supplementary Material under "html_files/html.index" (Supplementary Data Sheet 2).

Finally, we grouped the representative sequences from the 10 most abundant oligotypes (30 total) from each analysis grouping. Since most oligotypes were abundant across multiple projects, this list collapsed into 17 unique oligotypes across all three analysis runs. These 17 oligotypes were assigned identifiers ("Oligotype1" through "Oligotype17") that remained consistent across all three runs (**Table 3** and Table S1).

VISUALIZATIONS AND STATISTICAL ANALYSES OF OLIGOTYPE DISTRIBUTIONS

To visualize *Vibrio* community similarity between samples and habitats we constructed Nonmetric Multidimensional Scaling (NMDS) plots as part of our oligotyping pipelines. We included the covariance ellipsoids calculated as part of the oligotyping pipeline on these plots to visualize the spread of a given habitat's community variance. Importantly, covariance ellipsoids delineate the total high-dimensional space, not only the two axes shown in the NMDS plot. Statistical analyses of *Vibrio* oligotype distributions between and within habitat types followed a 4-step analysis using the ecological statistics packages PrimerE v.6 and the R package Vegan (Oksanen et al., 2012). First, we normalized an oligotype matrix, which consisted of samples across rows and oligotypes down columns (Supplementary Data Sheet 2), by percent per sample (i.e., to 100% total for each sample) and calculated pairwise Bray-Curtis similarities. Second, we assigned each sample to a habitat "factor" based on where it was collected (e.g., on plastics, sponges or seawater) and tested the null hypothesis that there were no community differences between habitat types using Analysis of Similarity (ANOSIM) permutation tests. This test builds a random distribution of oligotype abundances using 9999 permutations then assesses the likelihood that observed oligotype distributions across *a priori* assigned habitat factors occurred by chance. We then conducted pairwise ANOSIM tests to determine whether significant differences occurred between individual habitat factors. Third, when statistical groupings did occur (as they did in most cases), we identified the oligotypes that contributed most to the formation of these groupings using Similarity Percentages (SIMPER). SIMPER decomposes average Bray-Curtis similarities between all pairwise habitat comparisons into percentage contributions of each oligotype.

Table 3 | Summary of the 10 most abundant oligotypes from each of three oligotyping analysis groupings.

	MEGABLAST results		Percentage of each isolation source category for MEGABLAST hits						
Oligotype ID	**Number of 100% hits to nr database**	**Top species level hits (number of hits to that species)**	**1**	**2**	**3**	**4**	**5**	**No data**	**Habitats with Similarity Percentage (SIMPER) results > 10%**
Oligotype 1	22	*V. alfacsensis* (6) *V. sinaloensis* (2)	45.5	9.1	40.9	0.0	0.0	4.5	Sponge/MarineFish/MarineWater
Oligotype 2	171	*V. metschnikovii* (3) *V. neptunius* (11)	42.9	15.3	12.4	11.2	0.0	18.2	FreshwaterFish/FreshWater
Oligotype 3	780	*V. scophthalmi* (4) *V. ichthyoenteri* (31) *V. anguillarum* (37)	55.7	11.4	6.9	0.1	0.0	25.9	MarineFish/MarineWater/Seawater/Saltmarsh/Plastic
Oligotype 4	39	*V. cholerae* (9) *V. vulnificus* (17) *V. mimicus* (5)	2.6	0.0	0.0	0.0	30.8	66.7	FreshwaterFish/FreshWater/Sponge/MarineFish
Oligotype 5	1000*	*V. splendidus* (52) *V. mediterranei* (35) *V. gigantis* (39)	45.3	21.4	8.1	0.1	0.0	25.1	Sand-PAH/Seawater/Saltmarsh/Plastic
Oligotype 6	23	*V. ichthyoenteri* (1) *V. ordalii* (1)	70.8	8.3	0.0	0.0	0.0	20.8	Sponge/MarineFish/MarineWater
Oligotype 7	46	*V. ponticus* (12) *V. nigripulchritudo* (9)	30.4	8.7	4.3	0.0	0.0	56.5	Seawater
Oligotype 8	221	*V. cholerae* (184)	1.4	1.4	0.5	3.6	11.8	81.4	FreshwaterFish/FreshWater
Oligotype 9	1	*V. vulnificus* (1)	0.0	0.0	0.0	0.0	100	0.0	
Oligotype 10	18	*V. alginolyticus* (1)	11.1	0.0	0.0	89.0	0.0	0.0	Sand-PAH
Oligotype 11	1	None	0.0	0.0	0.0	0.0	0.0	100	
Oligotype 12	113	*V. coralliilyticus* (2)	27.4	39.8	0.0	0.0	0.0	32.7	Seawater/Plastic
Oligotype 13	66	*V. azureus* (13) *V. harveyi* (2)	54.5	6.1	0.0	11.6	5.5	22.7	Plastic
Oligotype 14	0	*V. azureus* (5) *V. owensii* (2)	0.0	0.0	0.0	0.0	0.0	100.0	
Oligotype 15	245	*V. vulnificus* (4) *V. shilonii* (3)	94.5	2.4	0.4	0.0	0.0	3.4	
Oligotype 16	140	*V. kanaloae* (3) *V. splendidus* (2)	42.1	14.3	11.4	2.1	0.0	30.0	
Oligotype 17	6	*V. splendidus* (3)	84.0	0.0	0.0	0.0	0.0	16.0	

Overlap in the most abundant sequences between groupings reduced the total number to 17. Oligotype names were assigned arbitrarily, but are consistent across all groupings. The species assignments given by reports from 100% MEGABLAST hits, and the number of hits to each species, is shown. The proportion of MEGABLAST hits isolated from each of the five habitat categories are also shown. Categories are; 1. Marine Host, 2. Seawater, 3. Other Marine, 4. Terrestrial or Human, and 5. Freshwater (see Materials and Methods). (*) Indicates maximum requested hits.

To gain insight into the taxonomy and ecology of our oligotype sequences, we isolated the representative sequence from the 17 most abundant oligotypes outlined above. We then used MEGABLAST to query these sequences against National Center for Biotechnology Information's (NCBI) nr database in June 2014 (nr = non-redundant amalgamation of GenBank, RefSeq, EMBL, DDBJ and PDB databases). We kept only 100% matches across the entire 60 bp query, and extracted the "isolation source" and "host" feature using Geneious (v. 6.1) annotation tables. We also extracted the most abundant 2 or 3 taxonomies from perfect hits, not including "uncultured bacterium" (**Table 3**). We created a PhyML tree of existing full-length *Vibrio* 16S rRNA gene sequences downloaded from type strains in the SILVA ARB v5.1 database and then added our oligotype sequences to this tree using the Maximum Parsimony feature in ARB across the V6 region only (using a V6 "filter" in ARB).

Our rationale for building this tree was not to reconstruct phylogenetic relationships between *Vibrio* species but rather to make some inference about the habitats from which closely related *Vibrio* 16S rRNA gene sequences have been isolated. To this end, we binned the isolation source and host annotations of both our oligotype MEGABLAST hits, and our ARB isolates, into five

habitat categories. These were "Marine Host," "Seawater," "Other Marine," "Terrestrial," and "Freshwater." Marine host included all sequences isolated from the skin or innards of a host in seawater (e.g., tunicates, fish, crustaceans, sponges and sea cucumbers). "Other Marine" included sediment, biofilms and algae, or other marine plant associated sources. "Terrestrial" included any sample taken from the terrestrial environment, or from a terrestrial host (e.g., humans, birds, and plants), and freshwater included hits isolated from a freshwater environment, or freshwater host (e.g., freshwater shrimp). We color-coded the proportion of each category and displayed them at the terminal nodes of each oligotype sequence in our PhyML reference tree. Lastly, to visualize the isolation source of the ARB isolate reference sequences, we color-coded the nodes of the tree according to the isolation source of each reference sequence at the terminal node of that branch. We used the online software Interactive Tree of Life (iTOL) (Letunic and Bork, 2011) to visualize the phylogenetic tree. The proportion of NCBI hits for each oligotype that fit into each category also appear in **Table 3**.

"WITHIN-HABITAT VARIANCE" OF *VIBRIO* COMMUNITIES

To determine differences in habitat specificity, we calculated the median *Vibrio* community variance across all datasets within the same habitat (its "within-habitat variance"), and compared that across habitats. This allowed us to determine if some habitats contained a specific *Vibrio* community, or if *Vibrio* communities varied widely even within the same habitat. To calculate the median variance of all datasets in a habitat, we normalized our oligotype abundance matrixes by the maximum value of each oligotype, then calculated Bray-Curtis community similarity between all pairwise comparisons using vegdist{vegan} function in R (Oksanen et al., 2012). We then calculated each habitat's multidimensional "centroid" using the median value of each sample within a habitat across all principal components. The distance of each sample to its habitat centroid was calculated across all principal components. The variance around the median value of sample-centroid distances was then compared across habitats in a standard ANOVA, followed by pairwise Tukey's Honestly Significant Difference (HSD) tests. This entire process, from centroid calculation to HSD tests was implemented using the betadisper{vegan} function in R. To visualize our results, we plotted each sample along their first two principal components, and plotted the multidimensional centroid. We then drew covariance ellipsoids around each habitat to illustrate the median distance for all samples in a habitat around its centroid. Median sample-centroid distances for each habitat were also plotted to better visualize the within-habitat variance.

RESULTS

OLIGOTYPE DISTRIBUTION ACROSS BIOTIC AND ABIOTIC SUBSTRATES

Oligotyping analysis of substrate associated-habitats (fish, sponges, plastics) yielded 74 unique oligotypes across 104 samples from 1,543,415 initial sequences. The minimum relative abundance threshold removed 808 rare oligotypes. The most abundant 5 oligotypes represented 76% of the reads, with the top 10 representing 94% (**Table 2**). Oligotypes that were abundant in at least one sample (>1% relative abundance) were always found across all three substrate types, meaning abundant oligotypes were ubiquitous across all host-associated habitats. Oligotype richness varied across host/substrate type. Of the 74 total oligotypes, 24.1 ± SE 2.2 were found in FreshwaterFish, 27.0 ± SE0.71 in MarineFish, 31 ± SE 0.9 in Sponge and only 18 ± SE1.5 for Plastic samples.

Pairwise Analysis of Similarity (ANOSIM) tests showed significant groupings in oligotype communities according to habitat, except between high salinity fish (MarineFish) and sponges, whose communities could not be significantly distinguished. This ANOSIM result is also visible in our NMDS analyses which shows clear overlap between both sponges (DeepSponge and ShallowSponge) and high salinity fish (MarineFish), yet separation from plastic and low salinity fish (FreshwaterFish) habitats (**Figure 1**). An ANOSIM test comparing marine biotic substrates (MarineFish and Sponges pooled together) revealed a significant grouping that excluded Plastic, an abiotic substrate. Similarity Percentages (SIMPER) analysis corroborated these results by illustrating the strong contribution of Oligotypes 1, 4, and 6 to both MarineFish and Sponge within-habitat similarity, and distinguished those habitats from FreshwaterFish and Plastics. Oligotypes 2, 4, and 8 contributed to both within habitat similarity, and between habitat differences for FreshwaterFish samples. Plastics were dominated by Oligotype 5, which also distinguished it from other habitat types, including biotic substrates (**Tables 4A,B**).

OLIGOTYPE DISTRIBUTIONS BETWEEN SUBSTRATES AND THEIR SURROUNDING WATER

Isolating individual biotic (hosts) and abiotic substrates along with their surrounding water allowed for direct comparisons of attached vs. free-living *Vibrio* communities. Fish hosts (FreshwaterFish and MarineFish) on average showed a nearly 20-fold enrichment of total *Vibrio* relative abundance compared to their surrounding water (30% ± SE 2.9 in fish vs. 1.6% ± SE 0.37 in water, **Table 1**), but samples from sponges or plastic showed no significant enrichment [although a non-significant trend of enrichment was evident for Sponge habitats (**Table 1**)]. Despite this enrichment in fish hosts, we could not differentiate *Vibrio* community structure in fish microbiome samples and their surrounding environment with ANOSIM analyses, so long as comparisons were made within the same salinity category (Marine and Freshwater).

Interestingly, although *Vibrio* communities between fish and their surrounding water at a given salinity were statistically indistinguishable, communities between fresh and marine salinity environments showed dramatic differences in both community structure and relative abundance of total *Vibrio* (**Figure 2**, **Middle**). Oligotypes 2, 4, and 8 dominated both FreshwaterFish and FreshWater (cumulative abundance in FreshwaterFish/FreshWater = 87.2%/71.9%), while Oligotypes 1, 3, and 6 dominated MarineFish and MarineWater (cumulative abundance in MarineFish/MarineWater = 60%/53% (**Figure 3**). Experimental aquaria without fish at low salinity (water only) showed a strong dominance of Oligotype 2, 4, and 8 (cumulative

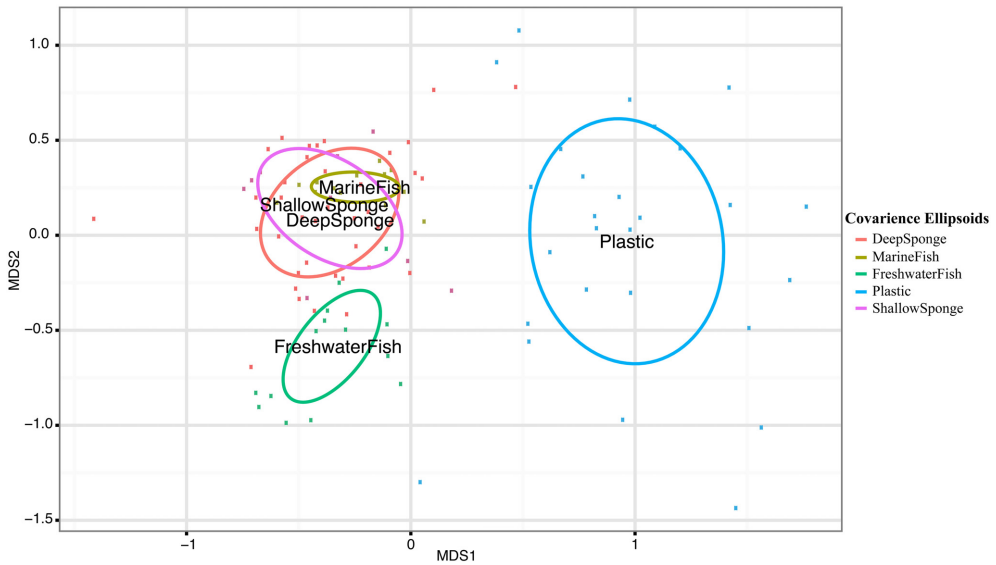


FIGURE 1 | Nonmetric Multidimensional Scaling (NMDS) plot of host-associated *Vibrio* communities based on oligotype distributions. Labels are located at the center of covariance ellipsoids around each host type. Pairwise ANOSIM permutation tests reveal all host habitats can be significantly differentiated except MarineFish and Sponge communities (Table S1).

Table 4 | (A,B) SIMPER analysis output for the "All substrate" analysis grouping.

(A) Within habitat markers							
FreshwaterFish	**Cont (%)**	**MarineFish**	**Cont (%)**	**Sponge**	**Cont (%)**	**Plastic**	**Cont (%)**
Oligo2	72.85	Oligo1	32.85	Oligo1	37.45	Oligo5	69.9
Oligo4	11.61	Oligo6	20.49	Oligo6	18.67	Oligo3	11.07
Oligo8	6.85	Oligo3	15.16	Oligo4	15.73	Oligo15	5.53
		Oligo4	12.85	Oligo2	6.36		
		Oligo7	7.31	Oligo8	5.43		
		Oligo5	4.98	Oligo7	4.54		
				Oligo3	2.7		
(B) Between habitat markers							
FreshwaterFish and MarineFish Average dissimilarity = 83.6		**FreshwaterFish and Sponge Average dissimilarity = 76.16**		**FreshwaterFish and Plastic Average dissimilarity = 90.26**			
Oligo2	33.64	Oligo2	33.5	Oligo2	30.08		
Oligo1	15.51	Oligo1	14.93	Oligo5	24.16		
Oligo6	11.57	Oligo8	12.8	Oligo8	8.38		
Oligo4	11.1	Oligo4	11.67	Oligo4	8.02		
Oligo8	8.72	Oligo6	9.46	Oligo15	5.21		
Cumulative	80.54	Cumulative	82.36	Cumulative	75.85		
MarineFish and Sponge Average dissimilarity = 63.21		**MarineFish and Plastic Average dissimilarity = 85.02**		**Sponge and Plastic Average dissimilarity = 89.48**			
Oligo1	19.77	Oligo5	23.41	Oligo5	23.33		
Oligo6	16.14	Oligo1	16.07	Oligo1	13.31		
Oligo4	15.01	Oligo6	12	Oligo6	8.46		
Oligo8	8.77	Oligo4	9.34	Oligo4	8.18		
Oligo2	8.45	Oligo7	7.04	Oligo2	7.64		
Cumulative	68.14	Cumulative	67.86	Cumulative	60.92		

A: The percent contribution of each oligotype to within-habitat Bray-Curtis similarity is shown (Cont%). B: The percent contribution of each oligotype to Bray-Curtis dissimilarities between two habitats is shown, along with average Bray-Curtis dissimilarities.

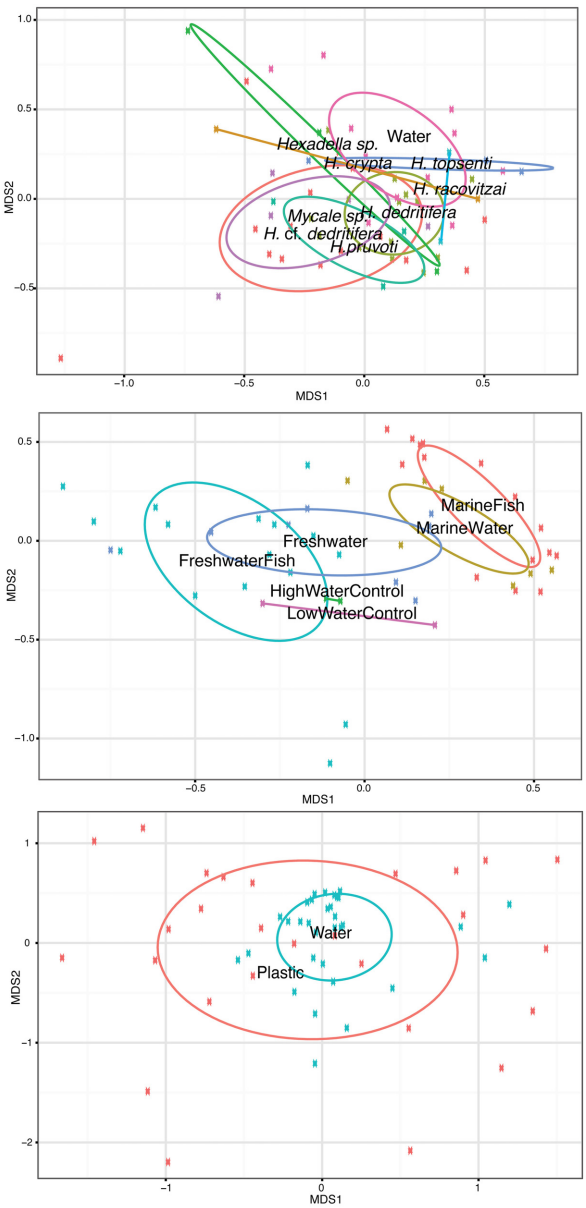


FIGURE 2 | Samples from the three host-associated studies included in this meta-analysis separated out into individual NMDS plots alongside water samples collected as part of the same study. Top: Sponges are labeled according to species. Water was collected next to *Mycale* sp. and *Hexadella* cf. *dedritifera* samples only. **Middle:** Fish water was sampled directly from aquaria in which fish were housed, in addition to "control" water that was sampled from aquaria at an identical salinity, but without any fish. **Bottom:** Ocean surface water was collected at the same time and location as corresponding plastics. ANOSIM permutation tests show that Fish, Sponge and Plastic associated *Vibrio* communities cannot be statistically separated out from their surrounding water environment. Labels for each habitat are within their respective covariance ellipsoid.

abundance = 67%), showing strong similarities to those aquaria that did house fish. However, marine aquaria without fish showed a strong dominance of Oligotype 2, inconsistent with marine aquaria that did contain fish. Interestingly, Oligotype 4 was found across all salinities, in water, fish and control samples, at greater than 10% relative abundance.

Sponge associated *Vibrio* communities were also statistically indistinguishable from their surrounding water (ANOSIM $P = 0.31$), although this may have been in part due to the sample size difference between sponge and associated seawater samples (Sponge = 49, Associated seawater = 11), and the high variability of particular oligotypes in some Sponge samples. Sponges showed significantly smaller relative abundance of Oligotype 7 and 5, and significantly increased abundance of Oligotypes 8 and 2 (Pairwise T-tests $P < 0.01$ in all cases). Sponge associated *Vibrio* communities also showed no clear groupings according to species (**Figure 2**, **Top**).

Lastly, *Vibrio* communities from plastic substrates overlapped completely with seawater communities collected alongside them (**Figure 2**, **Bottom**), and Oligotypes 3, 5, and 12, dominated both Plastics and associated seawater. We found no oligotype to be significantly enriched on plastic samples compared to their surrounding water, nor were there significant increases in total *Vibrio* on plastic substrates. In several cases, we found a single oligotype that did not occur in the surrounding water but dominated in relative abundance on an individual plastic substrate. This pattern was particularly apparent with Oligotypes 13, 2, and 7, which reached extremely high relative abundances on multiple occasions (e.g., Oligotype 2 at 91% relative abundance in the "10/09-Plastic" sample) (**Figure 4**).

OLIGOTYPE DISTRIBUTIONS ACROSS BROAD HABITAT TYPES

In order to gain as broad a view as possible of *Vibrio* oligotype distributions across habitats, we included 179 samples from 7 environmental and host-associated habitats spanning a wide range of environmental and geographical gradients (**Figure 5**). This analysis yielded 99 oligotypes, of which the top 5 represented 65% of all reads, while the top 10 represented 90%. We observed the top 10 oligotypes from this analysis at high abundance in previous analyses (as determined by identical representative sequences), except Oligotype 10, which was novel to Sand-PAH mesocosms and is a highly divergent oligotype which could not be fully resolved. Sand-PAH samples were taken from beach sand communities near the Deepwater Horizon oil spill, and were likely enriched for PAH-associated species (Kappell et al., 2014).

All oligotypes that were highly abundant in a single sample (>10% relative *Vibrio* abundance) occurred across all other habitat types. Abundant oligotypes were therefore also likely to be common across a wide variety of habitat types (**Figure 6**). Oligotype 5 in particular was found to be both highly abundant and frequent, occurring in all 179 samples analyzed across all habitats (**Figure 6**). Several oligotypes did not follow this general pattern, and despite a relative ubiquity, they maintained at low mean relative abundances across all the samples in which they occurred (e.g., Oligotypes 1, 2, 13, and 14, **Figure 6**). Comparing the mean relative abundance of individual oligotypes between marine hosts (Sponge, MarineFish) and marine environments (Seawater, Saltmarsh, Sand-PAH) revealed that 10 of the top 17 oligotypes (**Table 3**) were significantly different between these habitat categories at the bonferroni-corrected alpha level of 0.0029. SIMPER revealed the strong influence of Oligotype 3 and

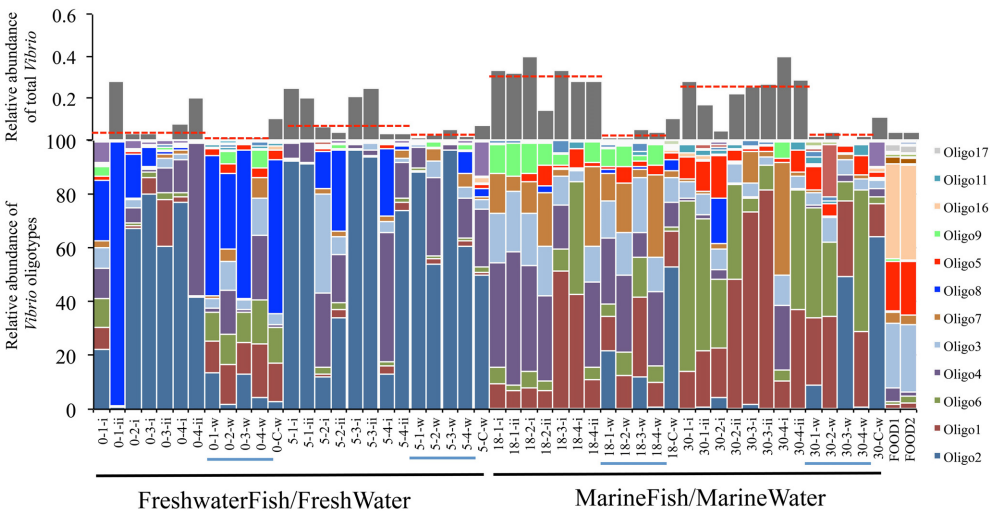


FIGURE 3 | Oligotype distribution for FreshwaterFish and MarineFish tissue (FreshwaterFish: 0-x-i and 5-x-ii, MarineFish: 18-x-i and 18-x-ii), associated water samples (x-x-W, blue lines), and water samples from aquaria containing no fish (x-C-W). The relative abundance of each oligotype within the total *Vibrio* diversity for each sample is shown in stacked bar graphs (bottom), and the proportion of the total *Vibrio* (relative abundance) within all bacterial diversity for each sample is shown with gray bars (top, red-dashed lines are median values for each host-habitat). A clear division between low and high salinity samples is seen, despite considerable variation within salinities. Significant differences in median total *Vibrio* relative abundance exists between FreshwaterFish and MarineFish samples, and between MarineFish samples and their surrounding water. The *Vibrio* community of fish food used during experimental period is also shown ("FOODX").

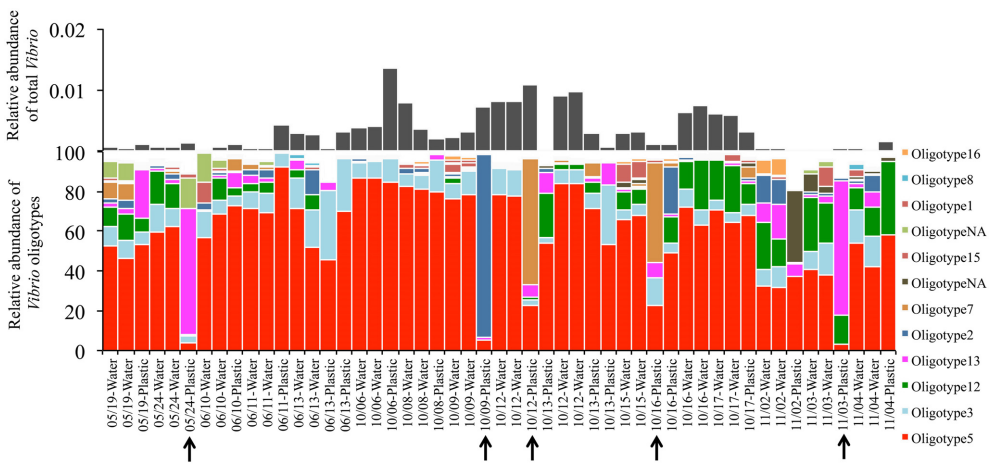


FIGURE 4 | Oligotype distributions for Plastic samples and their surrounding water. Sample labels indicate the date the samples were collected and the sample type. Water and plastic samples collected on the same date are associated with one another. All water was collected at the surface. Black arrows indicate plastic samples that contain a single oligotype at greater than 50% relative abundance. "OligotypeNA" represents an oligotype that was not among the top 10 most abundant oligotypes from any three of our oligotype groupings.

5 in Saltmarsh and non-host associated Seawater communities, and Oligotype 5 in Sand-PAH mesocosms. Oligotype 5 was often extremely abundant in seawater samples, including those from a *Vibrio* bloom (from project ICM_PML_Bv6, **Table 1**), and on plastic samples (**Figure 4**). ANOSIM analyses revealed that across all habitat pairwise comparisons only Sponge and MarineFish, Seawater and Saltmarsh, Sand-PAH and Plastic, and Sand-PAH and Seawater could not be significantly differentiated from one another (Table S1).

Analysis of "within-habitat similarity" (a measure of sample dispersion within a habitat) showed significant differences between habitats. *Post-hoc* Tukeys pairwise tests revealed Sand-PAH mesocosms (median distance to centroid = 0.047) and Saltmarsh (median distance to centroid = 0.046) contained significantly less variance than Seawater (median distance to centroid = 0.388) and Sponge (median distance to centroid = 0.49) habitats. We note here however, that these results do not control for the larger geographic area over which Seawater and Sponge

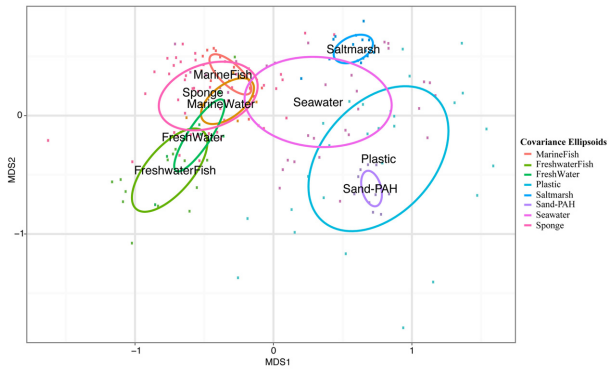


FIGURE 5 | NMDS plot with covariance ellipsoids for both host-associated and environmental samples. Sample names refer to habitat type and VAMPS project listed in **Table 1**.

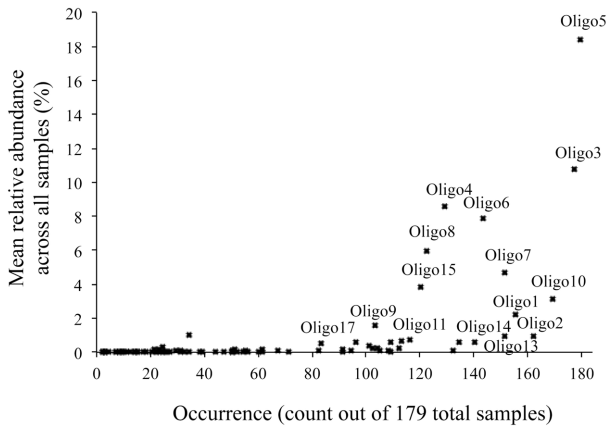


FIGURE 6 | Commonness and abundance plot of all oligotypes that were part of the mixed habitat sample grouping analysis (Table 2). The occurrence (presence/absence) of each of the 99 oligotypes across all 179 samples is plotted along the x-axis while its mean relative abundance across all samples is plotted on the y-axis. Samples that are both common and abundant are found in the top right, while those that are common, but rare are in the bottom right. Both rare and uncommon are found in the bottom left. The top 17 most abundant oligotypes from **Table 3** are also labeled.

samples were collected as compared to Sand-PAH and Saltmarsh samples. All other pairwise comparisons were insignificant at the 0.05 alpha level after multiple comparison adjustments.

PHYLOGENETIC AND METADATA ANALYSIS OF ABUNDANT OLIGOTYPES

Our phylogenetic analysis of ARB reference sequences revealed no 16S rRNA gene phylogeny-habitat relationship. This is evidenced by the spread of reference sequences from the same isolation source across the phylogenetic tree (**Figure 7**). Our 17 abundant oligotypes were also not monophyletic according to the habitat in which they were most abundant (e.g., sponges or fish). For example, oligotype sequences dominant in FreshwaterFish and Freshwater (Oligotype 2, 4, and 8) were found to branch in different parts of our phylogenetic tree (**Figure 7**).

MEGABLAST queries of abundant oligotypes returned on average 120 perfect matches from the NCBI nr database, although variance in this number was high (ranging from 0 to 1000). No oligotype with more than two perfect NCBI hits came from only one source, and any oligotype with more than 50 perfect NCBI hits was isolated from at least three different sources. Oligotype 5, which was the most abundant across our habitats (**Figure 6**), also had the maximum number of allowable perfect hits (1000). Which habitat a query sequence came from was an extremely poor predictor for the isolation source of its NCBI hits, although Oligotypes 4 and 8, which contributed to FreshwaterFish had previously been isolated from freshwater environments (**Table 3**), and both matched *V. cholerae,* found in brackish water, while abundant oligotypes in marine organisms (e.g., 1 and 6), did not return any previous isolations from freshwater environments. Surprisingly, Oligotype 2, which contributed most to freshwater environments, had no perfect hits from freshwater sources. The majority of isolation sources from all our NCBI hits (76%) were marine hosts, although we note the potential for database biases.

DISCUSSION

Our results represent the first attempt to use subtle nucleotide variation at a single, 60 bp gene marker to make sense of community level patterns across habitats within the genus *Vibrio*. Although our results are not the first to explore *Vibrio* community patterns across diverse habitats (Hunt et al., 2008; Preheim et al., 2011a; Szabo et al., 2013), we use preexisting sequence data to extend conclusions made by previous authors to a broader survey of unexplored habitats and provide novel insights into substrate-associated communities and their surrounding water.

The breadth and scope of our analyses, 211 samples representing seven unique habitats, revealed or supported several interesting patterns suggestive of two broad hypotheses regarding different aspects of *Vibrio* ecology and life history. First, *Vibrio* contains many generalist taxa, each adapted to a wide range of animal hosts. Second, our data suggest even these "host-adapted" vibrios occur as members of free-living communities facilitating long distance dispersal to disparate hosts. We suggest both of these characteristics, combined with previous understanding of rapid growth rates (McDonough et al., 2013; Skorupski and Taylor, 2013) fit the description of an "r-strategist" life history.

Several patterns within our data support these suggestions. Fish acclimated to marine salinities share highly similar *Vibrio* communities with geographically and phylogenetically distinct sponges (**Figure 1**, **Table 4**). Both marine acclimated fish and sponge communities were typified by a strong dominance of Oligotypes 1, 4, and 6, all of which were above 15% mean relative abundance in both "Sponge" and "MarineFish" samples. These oligotypes contributed at least 12% of within group Bray-Curtis similarities for each habitat (**Table 4**). Furthermore, ANOSIM analysis at the community level (i.e., using all oligotypes in each community) found these habitats could not be significantly separated, a conclusion supported by their overlapping distribution in 2D projections (**Figures 1**, **5**). Although these communities do not appear to be host specific, they do appear to be specific to biotic hosts. Plastic communities, on an abiotic substrate, were

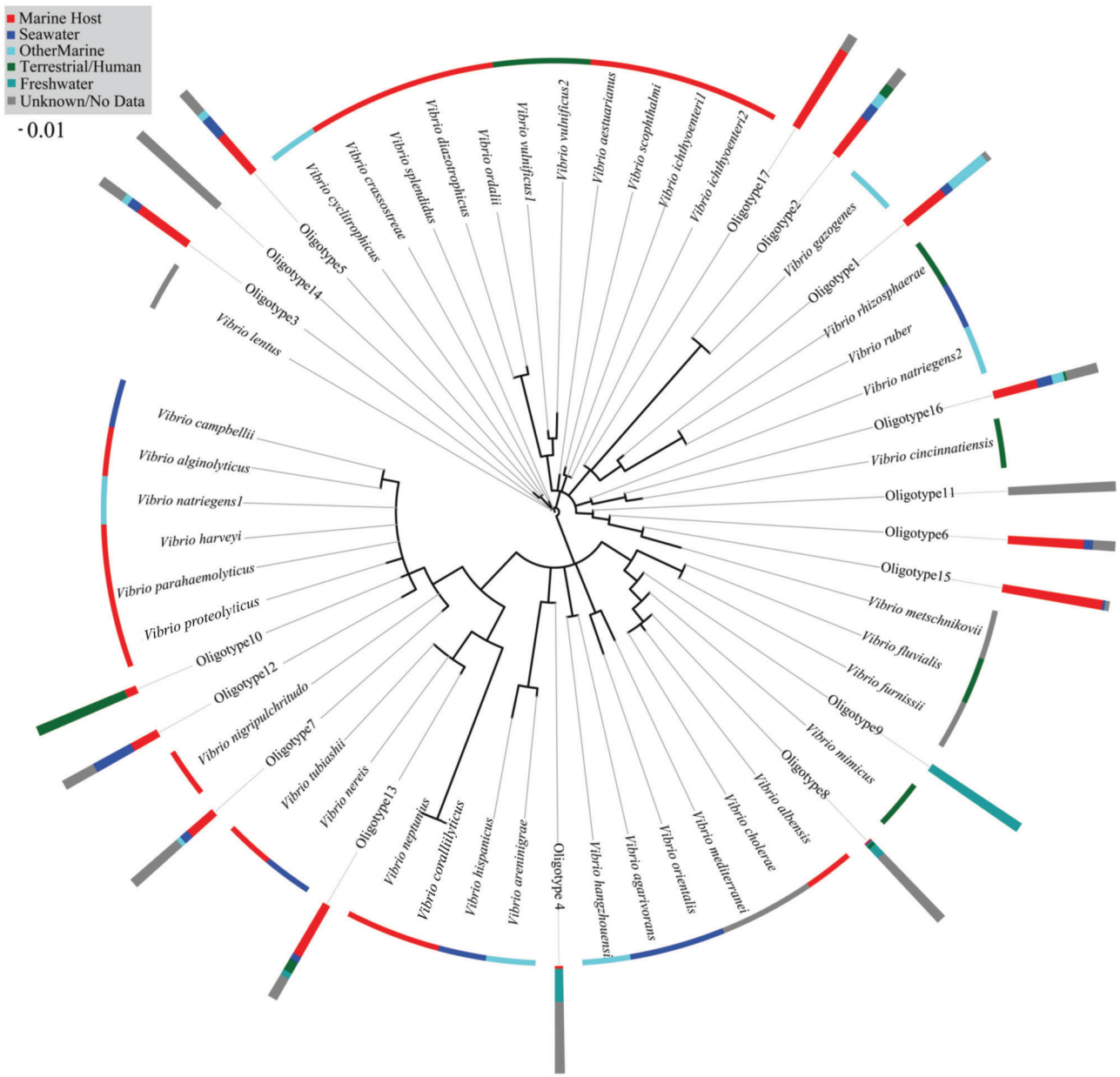


FIGURE 7 | PhyML phylogeny of full-length 16S rRNA gene sequences from type strain isolates in the SILVA ARB database, with the 17 most abundant oligotypes found in this study added by Maximum Parsimony over the 60 bp region. The isolation source for each type strain is color coded at its terminal node. The proportion of MEGABLAST hits that fell into each isolation source category is coded for the 17 oligotype sequences added to the tree, and shown in bar graphs at each oligotype node.

typified by Oligotypes 5, 3, and 12, and while oligotypes that typified marine fish and sponges sometimes occurred on plastics, they were always in low abundance. Furthermore, pairwise ANOSIM tests showed that although Sponges and MarineFish could not be significantly separated both could be separated from Plastic. We confirmed this result with significant ANOSIM groupings of MarineFish and Sponge datasets pooled together, at the exclusion of Plastic samples (data not shown). We found groupings were again characterized by high abundance of Oligotypes 1, 4, and 6 on biotic substrates, and Oligotype 5, 3, and 12 on Plastic. This suggests that vibrios behave differently with respect to adaptation to and colonization of biotic vs. abiotic substrates. Furthermore, the finding that *Vibrio* oligotypes associated with plastics overlap with those in the surrounding seawater suggests that recruitment may take place far from the origins of the plastic marine debris itself, typically thought to be land-based.

Explaining the similarity between *Vibrio* communities in fish and sponges is challenging, but overlap in habitat geography during sampling, or overlap of laboratory sample preparation

in space or time can be ruled out. Sponges were collected by SCUBA and Remote Operated Vehicle (ROV) deep-sea dives from the Northeast Atlantic and Mediterranean Sea (Reveillaud et al., 2014) from 1981 to 2011, while all fish were collected from experimental aquaria filled with sterilized water and Instant Ocean® salt mix in a Woods Hole, MA laboratory in 2013 (Schmidt et al., submitted; Supplementary Data Sheet 1). Furthermore, although sequencing was conducted on the same Illumina HiSeq instrument, run dates were several months apart, and sample storage occurred in different freezers, making contamination between projects unlikely. In addition to community level patterns outlined above, our MEGABLAST analysis of the representative sequences from oligotypes most representative of our sponge and fish samples revealed that each was previously isolated from a variety of marine hosts including crabs, jellyfish, sea squirts, corals, fish, clams, and sea cucumbers (**Table 3**, Table S2). Lastly, a previous comparison of two phylogenetically distinct hosts (mussels and crabs) also found significant overlap in *Vibrio* communities (Preheim et al., 2011a), although we note in this case hosts were not geographically or temporally distinct.

Such a large degree of host plasticity does not appear to be a ubiquitous feature among all bacterial taxa, and *Vibrio* clearly differs from other taxa at the community level. A previous study on sponge-microbe associations demonstrated host-specificity within the genus *Nitrospira* (Reveillaud et al., 2014). This study used the same oligotyping pipeline of V6 sequences, from the same samples analyzed here, and the authors showed strong host specificity of *Nitrospira* oligotypes to sponges at the species level (see Figure 4 in Reveillaud et al., 2014). They found closely related sponge species had differential enrichment preferences for closely related *Nitrospira* phylogenetic lineages across varying bathymetric and geographic areas. Our oligotyping analysis, focusing on *Vibrio,* highlighted the lack of sponge-specific patterns within this genus (**Figure 2**, **Top**), and is therefore in stark contrast to the patterns illustrated for *Nitrospira*.

Further supporting our suggestion that taxa within the genus *Vibrio* are generalist, long distance dispersing, r-strategists is the commonality of some oligotypes across the broad range of host associated and environmental habitat types in this study (**Figure 7**). Although *Vibrio* does not form spores (Madigan et al., 2009), it is known to enter a "viable but uncultivable" state under stressful or nutrient limiting conditions (Ramaiah et al., 2002). Research with the squid symbiont *Vibrio fischeri* found the bacteria quickly became uncultivable and non-luminescent in nutrient poor water outside of its host, but retained the ability to colonize, and luminesce, given re-entry to a suitable host (Lee and Ruby, 1995). Our results demonstrate that Oligotypes 1, 6, 4, 8, and 9 were all significantly enriched in marine hosts compared to marine environmental samples, with as much as a 168-fold enrichment, yet they all occurred at low abundance in open-ocean, host-independent samples. It is possible the rare occurrence of these "host-associated" oligotypes in seawater samples represent taxa that have entered viable but uncultivable states, giving them the ability to disperse long distances in nutrient poor waters between opportunistic colonization of a wide variety of marine hosts. Conversely, Oligotype 5 was found at an average relative abundance in environmental samples (Seawater, Saltmarsh) and an abiotic substrate (Plastic) of >40%, while averaging only 2.8% in hosts. This oligotype was found across all 179 samples analyzed as part of this study (**Figure 6**), and was widely represented in our MEGABLAST results (**Table 3**). Analysis of *Vibrio* sequences recovered from an earlier study of plastic marine debris samples (Zettler et al., 2013) also detected Oligotype 5 (data not shown) despite employing a different sequencing platform. Together, these data suggest a widely distributed, predominantly "non-host associated" *Vibrio* found in hosts only through chance or ephemeral colonization.

SALINITY DRIVES *VIBRIO* STRUCTURE IN WATER AND HOST COMMUNITIES

Salinity is a known driver of *Vibrio* community structure and most vibrios are thought to occur in brackish or marine environments (Takemura et al., 2014). Schmidt et al. (submitted) (Supplementary Data Sheet 1) experimentally manipulated salt concentrations in aquaria containing a euryhaline (salt tolerant) fish, and characterized the resulting bacterial community. They found that communities in both the fish and water changed across the salt gradient, but that they did not change concurrently, resulting in drastically different communities in fish and tank water. Interestingly, fish/water differences at high salinities (18 and 30 ppt) were in part driven by high *Vibrio* relative abundance in fish, vs. its relative rarity in tank water (**Figure 3**). *Vibrio* also partly drove differences in fish microbiomes across the salinity gradient, with a nearly 10-fold increase in total *Vibrio* relative abundance from 0 to 30 ppt acclimated fish. This study did not, however, resolve bacteria below the genus level, and could not make conclusions about variation within a genus across salinities. The study therefore did not assess if increases in total *Vibrio* relative abundance up the gradient were due to the same taxa becoming more abundant, or to the addition of novel taxa at higher salinities. Nor were they able to assess if *Vibrio* inside fish were the same as those found in the tank water.

The fine scale resolution provided by our oligotyping analyses allowed us to answer these questions, and we show that *Vibrio* community structure between water and fish are broadly consistent, with both habitats sharing similar occurrence and relative abundances of particular *Vibrio* oligotypes (**Figure 3**), despite an overall enrichment of *Vibrio* relative abundance in fish vs. water. We also show that FreshWater (tank water community) and FreshwaterFish (fish microbiome community) cannot be significantly distinguished with ANOSIM tests, and SIMPER analyses find the same oligotypes (2, 4, and 8) are representative of both FreshWater and FreshwaterFish (**Table 4**). The same is true for comparisons between high salt acclimated fish (MarineFish) and their water (MarineWater), which are both characterized by Oligotypes 1, 6, 3, and 4. This trend for both salinities is evident from 2D projections of community structure (**Figure 2**, **Middle**), which show overlapping ellipsoids of fish and water habitats (although some separation is evident). Despite an overall shift in community structure across the salinity gradient, Oligotype 4 remains at high abundance in both fresh and marine samples. Oligotype 4 was highly enriched in all host-associated samples (Sponges, FreshwaterFish, MarineFish), and

extremely rare in environmental water samples not associated with a host collection (i.e., completely independent of hosts). Furthermore, MEGABLAST results from this oligotype show its previous isolation from both marine and freshwater hosts, but never from seawater (**Table 4**). Together these results are suggestive of host-associated, potentially salt-tolerant *Vibrio* taxa.

ADVANTAGES AND LIMITATIONS OF OLIGOTYPING ANALYSIS FOR *VIBRIO*

By separating regions of a marker gene that contain biologically meaningful variation from stochastic error, oligotyping allows single nucleotide differences across short marker genes to identify potentially ecologically meaningful patterns with extremely small amounts of information (Eren et al., 2013a). This technique provides substantial benefits, avoiding the need for lengthy and expensive culturing and sequencing protocols, and allows researchers to tap into massive existing databases to ask novel ecological questions at high-throughput levels across global scales. We note, however, that some serious limitations do exist for these type of data. 16S rRNA gene sequences can be nearly identical across multiple *Vibrio* species (Gomez-Gil et al., 2004), and even contain variance between copies of 16S rRNA genes within a single genome, making its use as a phylogenetic tool difficult or impossible. In addition, because our analyses use only 60 bp of DNA sequence at a single marker gene, our data are insufficient for any phylogenetic inference, and we cannot deduce relatedness between individual oligotypes. This provides major limitations in our ability to address some hypotheses about the evolutionary history of vibrios adapted to specific habitats. We could not investigate, for example, if the abundant oligotypes of fish and sponges (Oligotypes 1, 4, and 6) share common ancestry, which would suggest speciation within the genus after association with the host. We also cannot confidently tie our conclusions to previous observations about particular *Vibrio* species, such as *V. splendidus*' potentially recent adaption to particulate adhesion (Hunt et al., 2008), or *V. cholerae's* affinity for freshwater (Skorupski and Taylor, 2013), since we cannot make taxonomic assignments to any of our oligotypes. High-resolution taxonomic assignment of *Vibrio* has been a significant challenge, necessitating the use of genomic analyses including DNA-DNA hybridization, multilocus sequence analysis (MLSA), and genome sequencing for species- or strain-level identification (Thompson et al., 2005). Analyses of multiple loci, or entire genome sequences, are therefore required to make any phylogenetic inference (Thompson et al., 2005; Preheim et al., 2011b). However, this study shows the utility of oligotyping as an easily adaptable, high-throughput and unbiased method for large-scale analyses of data from publically available sequence data repositories, and suggests its wide application could greatly extend the range of possibilities to explore microbial ecology studies of particular genera.

CONCLUSIONS

Our analysis combines a novel bioinformatics technique with large quantities of *Vibrio* 16S rRNA gene sequence data to reveal patterns of *Vibrio* ecology across a wide range of environmental, host, and abiotic substrate-associated habitats. Despite the drawbacks for phylogenetic and taxonomic inference of using a single, short rRNA gene sequence, our analyses show strong convergence between host-associated communities, despite wide geographic and phylogenetic distance between them. We also show a surprising overlap, and a lack of significant divisions, between *Vibrio* communities in hosts and those found in their surrounding aquatic environments. Our results further support that *Vibrio*, as a genus, is largely populated by generalist r-strategist species, capable of long distance dispersal, a wide breadth of growth requirements, and rapid growth rates (Szabo et al., 2013).

ACKNOWLEDGMENTS

We would like to thank the VAMPS bioinformatics team, principally Andy Voorhis and Anna Shipunova. We also thank Joseph Migliore with the Rhode Island Department of Environmental Management for assistance collecting samples from Mt. Hope Bay. Open ocean plastic samples were collected by students and staff at the SEA Education Association/SEA Semester in Woods Hole. This work was supported by an NSF Collaborative grant to Erik Zettler (OCE-1155379), Tracy J. Mincer (OCE-1155671) and Linda A. Amaral-Zettler (OCE-1155571), NSF TUES grant to Erik Zettler and Linda A. Amaral-Zettler (DUE-1043468). Additional support came from the Woods Hole Center for Oceans and Human Health from the National Institutes of Health and National Science Foundation (NIH/NIEHS 1 P50 ES012742-01 and NSF/OCE 0430724-J: Linda A. Amaral-Zettler and Leslie Murphy) and an NSF/OCE-1128039 award (Linda A. Amaral-Zettler and Leslie Murphy). Victor Schmidt was supported during this work by an NSF IGERT fellowship (DGE 0966060, Dr. David Rand, PI).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2014.00563/abstract

REFERENCES

Aiso, K., Simidu, U., and Hasuo, K. (1968). Microflora in the digestive tract of inshore fish in Japan. *Microbiology* 52, 361–364.

Andrews, J., and Harris, R. (1986). "r- and K-selection and microbial ecology," in *Advances in Microbial Ecology*, ed K. C. Marshall (New York, NY: Springer), 99–147.

Austin, B., and Austin, D. (2007). *Bacterial Fish Pathogens; Diseases of Farmed and Wild Fish, 4th Edn*. Chichester: Praxis Publishing.

Chin, C., Sorenson, J., Harris, J. B., Robins, W. P., Charles, R. C., Jean-Charles, R. R., et al. (2011). The origin of the Haitian cholera outbreak strain. *N. Engl. J. Med.* 364, 33–42. doi: 10.1056/NEJMoa1012928

Cordero, O. X., Wildschutte, H., Kirkup, B., Proehl, S., Ngo, L., Hussain, F., et al. (2012). Ecological populations of bacteria act as socially cohesive units of antibiotic production and resistance. *Science* 337, 1228–1231. doi: 10.1126/science.1219385

Eren, M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., et al. (2013a). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* 4, 1111–1119. doi: 10.1111/2041-210X.12114

Eren, M., Sogin, M. L., Morrison, H. G., Vineis, J. H., Fisher, J. C., Newton, R. J., et al. (2014). A single genus in the gut microbiome reflects host preference and specificity. *ISME J.* doi: 10.1038/ismej.2014.97. [Epub ahead of print].

Eren, M., Vineis, J. H., Morrison, H. G., and Sogin, M. L. (2013b). A filtering method to generate high quality short reads using Illumina paired-end technology. *PLoS ONE* 8:e66643. doi: 10.1371/journal.pone.0066643

Eren, M., Zozaya, M., Taylor, C. M., Dowd, S. E., Martin, D. H., and Ferris, M. J. (2011). Exploring the Diversity of *Gardnerella vaginalis* in the genitourinary

tract microbiota of monogamous couples through subtle nucleotide variation. *PLoS ONE* 6:e26732. doi: 10.1371/journal.pone.0026732

FAO. (2012). *The State of World Fisheries and Aquaculture*. Rome: Food and Agriculture Organization of the United Nations.

Gilbert, J. A., Steele, J. A., Caporaso, J. G., Steinbrück, L., Reeder, J., Temperton, B., et al. (2012). Defining seasonal marine microbial community dynamics. *ISME J.* 6, 298–308. doi: 10.1038/ismej.2011.107

Gomez-Gil, B., Soto-Rodríguez, S., García-Gasca, A., Roque, A., Vazquez-Juarez, R., Thompson, F. L., et al. (2004). Molecular identification of *Vibrio harveyi*-related isolates associated with diseased aquatic organisms. *Microbiology* 150, 1769–1777. doi: 10.1099/mic.0.26797-0

Heidelberg, J. F., Eisen, J., Nelson, W. C., Clayton, R., Gwinn, M. L., Dodson, R. J., et al. (2000). DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* 406, 477–483. doi: 10.1038/35020000

Hunt, D. E., David, L. A., Gevers, D., Preheim, S. P., Alm, E. J., and Polz, M. F. (2008). Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* 320, 1081–1085. doi: 10.1126/science.1157890

Huse, S. M., Dethlefsen, L., Huber, J., Mark Welch, D., Welch, D. M., Relman, D., et al. (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet.* 4:e1000255. doi: 10.1371/journal.pgen.1000255

Huse, S. M., Mark Welch, D. B., Voorhis, A., Shipunova, A., Morrison, H. G., Eren, A. M., et al. (2014). VAMPS: a website for visualization and analysis of microbial population structures. *BMC Bioinformatics* 15:41. doi: 10.1186/1471-2105-15-41

Kappell, A. D., Wei, Y., Newton, R. J., Van Nostrand, J. D., Zhou, J., McLellan, S. L., et al. (2014). The polycyclic aromatic hydrocarbon degradation potential of Gulf of Mexico native coastal microbial communities after the Deepwater Horizon oil spill. *Front. Microbiol.* 5:205. doi: 10.3389/fmicb.2014.00205

Lee, K., and Ruby, E. G. (1995). Symbiotic role of the viable but nonculturable state of *Vibrio fischeri* in Hawaiian coastal seawater. *Appl. Environ. Microbiol.* 61, 278–283.

Letunic, I., and Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* 39, W475–W478. doi: 10.1093/nar/gkr201

Liston, J. (1956). Quantitative variations in the bacterial flora of flatfish. *J. Gen. Microbiol.* 15, 305–314.

Madigan, M., Martinko, J., Dunlap, P., and Clark, D. (2009). *Brock Biology of Microorganisms*, *12th Edn*. San Francisco, CA: Pearson Benjamen Cummings.

McDonough, E., Bradley, E., and Camilli, A. (2013). "Regulating the transition of *Vibrio cholerae* out of the host," in *Regulation of Bacterial Virulence*, eds E. Vasil and A. Darwin (Washington, DC: ASM Press), 566–578.

Mutreja, A., Kim, D. W., Thomson, N. R., Connor, T. R., Lee, J. H., Kariuki, S., et al. (2011). Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* 477, 462–465. doi: 10.1038/nature10392

Oksanen, J., Blanchet, G., Kindt, R., Legendre, P., O'Hara, R., Simpson, G., et al. (2012). *Vegan: Community Ecology Package*. R package version 1.17-3.

Preheim, S. P., Boucher, Y., Wildschutte, H., David, L., Veneziano, D., Alm, E. J., et al. (2011a). Metapopulation structure of Vibrionaceae among coastal marine invertebrates. *Environ. Microbiol.* 13, 265–275. doi: 10.1111/j.1462-2920.2010.02328.x

Preheim, S. P., Timberlake, S., and Polz, M. F. (2011b). Merging taxonomy with ecological population prediction in a case study of Vibrionaceae. *Appl. Environ. Microbiol.* 77, 7195–7206. doi: 10.1128/AEM.00665-11

Ramaiah, N., Ravel, J., Straube, W. L., Hill, R. T., and Colwell, R. R. (2002). Entry of *Vibrio harveyi* and *Vibrio fischeri* into the viable but nonculturable state. *J. Appl. Microbiol.* 93, 108–116. doi: 10.1046/j.1365-2672.2002.01666.x

Reveillaud, J., Maignien, L., Murat, E., A., Huber, J., Apprill, A., Sogin, M. L., et al. (2014). Host-specificity among abundant and rare taxa in the sponge microbiome. *ISME J.* 8, 1198–1209. doi: 10.1038/ismej.2013.227

Ruby, E., and Nealson, K. (1976). Symbiotic association of *Photobacterium fischeri* with the marine luminous fish *Monocentris japonica*: a model of symbiosis based on bacterial studies. *Biol. Bull.* 151, 547–586.

Skorupski, J., and Taylor, R. (2013). "Toxin and virulence regulation in *Vibrio cholerae*," in *Regulation of Bacterial Virulence*, eds M. Vasil and A. Darwin (Washington, DC: ASM Press), 241–262.

Szabo, G., Preheim, S. P., Kauffman, K. M., David, L., Shapiro, J., Alm, E. J., et al. (2013). Reproducibility of Vibrionaceae population structure in coastal bacterioplankton. *ISME J.* 7, 509–519. doi: 10.1038/ismej.2012.134

Takemura, A. F., Chien, D. M., and Polz, M. F. (2014). Associations and dynamics of Vibrionaceae in the environment, from the genus to the population level. *Front. Microbiol.* 5:38. doi: 10.3389/fmicb.2014.00038

Thompson, F. L., Gevers, D., Thompson, C. C., Dawyndt, P., Hoste, B., Munn, C. B., et al. (2005). Phylogeny and molecular identification of Vibrios on the basis of multilocus sequence analysis phylogeny. *Appl. Environ. Microbiol.* 71, 5107–5115. doi: 10.1128/AEM.71.9.5107

Zettler, E. R., Mincer, T. J., and Amaral-Zettler, L. A. (2013). Life in the "Plastisphere": microbial communities on plastic marine debris. *Environ. Sci. Technol.* 47, 7137–7146. doi: 10.1021/es401288x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 August 2014; accepted: 07 October 2014; published online: 13 November 2014.

Citation: Schmidt VT, Reveillaud J, Zettler E, Mincer TJ, Murphy L and Amaral-Zettler LA (2014) Oligotyping reveals community level habitat selection within the genus Vibrio. Front. Microbiol. ***5****:563. doi: 10.3389/fmicb.2014.00563*

This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology.

frontiers in
MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 16 February 2015
doi: 10.3389/fmicb.2015.00049

Oligotyping reveals stronger relationship of organic soil bacterial community structure with N-amendments and soil chemistry in comparison to that of mineral soil at Harvard Forest, MA, USA

Swathi A. Turlapati[1,2], Rakesh Minocha[2]*, Stephanie Long[2], Jordan Ramsdell[3] and Subhash C. Minocha[1]

[1] *Department of Biological Sciences, University of New Hampshire, Durham, NH, USA*
[2] *Northern Research Station, United States Department of Agriculture Forest Service, Durham, NH, USA*
[3] *Hubbard Center for Genome Studies, University of New Hampshire, Durham, NH, USA*

Edited by:
A. Murat Eren, Marine Biological Laboratory, USA

Reviewed by:
Sandra L. McLellan, University of Wisconsin–Milwaukee, USA
Tom O. Delmont, Marine Biological Laboratory, USA

***Correspondence:**
Rakesh Minocha, Northern Research Station, United States Department of Agriculture Forest Service, 271 Mast Road, Durham, NH 03824, USA
e-mail: rminocha@unh.edu; rminocha@fs.fed.us

The impact of chronic nitrogen amendments on bacterial communities was evaluated at Harvard Forest, Petersham, MA, USA. Thirty soil samples (3 treatments × 2 soil horizons × 5 subplots) were collected in 2009 from untreated (control), low nitrogen-amended (LN; 50 kg NH_4NO_3 ha^{-1} yr^{-1}) and high nitrogen-amended (HN; 150 kg NH_4NO_3 ha^{-1} yr^{-1}) plots. PCR-amplified partial 16S rRNA gene sequences made from soil DNA were subjected to pyrosequencing (Turlapati et al., 2013) and analyses using oligotyping. The parameters M (the minimum count of the most abundant unique sequence in an oligotype) and s (the minimum number of samples in which an oligotype is expected to be present) had to be optimized for forest soils because of high diversity and the presence of rare organisms. Comparative analyses of the pyrosequencing data by oligotyping and operational taxonomic unit clustering tools indicated that the former yields more refined units of taxonomy with sequence similarity of ≥99.5%. Sequences affiliated with four new phyla and 73 genera were identified in the present study as compared to 27 genera reported earlier from the same data (Turlapati et al., 2013). Significant rearrangements in the bacterial community structure were observed with N-amendments revealing the presence of additional genera in N-amended plots with the absence of some that were present in the control plots. Permutational MANOVA analyses indicated significant variation associated with soil horizon and N treatment for a majority of the phyla. In most cases soil horizon partitioned more variation relative to treatment and treatment effects were more evident for the organic (Org) horizon. Mantel test results for Org soil showed significant positive correlations between bacterial communities and most soil parameters including NH_4 and NO_3. In mineral soil, correlations were seen only with pH, NH_4, and NO_3. Regardless of the pipeline used, a major hindrance for such a study remains to be the lack of reference databases for forest soils.

Keywords: bacterial community, forest soils, microbiome, oligotypes, pyrosequencing, QIIME software, OTUs, entropy

INTRODUCTION

Soils harbor an immense diversity of bacteria (Torsvik et al., 2002; Trevors, 2010 and references therein), most of it is hidden from experimental analyses (Wall et al., 2010). A vast majority of soil microbes are recalcitrant to culture methods thus increasing the complexity of any study to expose this concealed diversity (Sait et al., 2002; Nunes da Rocha et al., 2009; Vartoukian et al., 2010; Lombard et al., 2011). Recently developed molecular tools have enabled us to analyze the expanse of variation in bacterial populations using culture-independent methods such as polymerase chain reaction (PCR) amplification of partial or full-length genes of 16S rRNA, and their in-depth sequencing (Janssen, 2006; Větrovský and Baldrian, 2013). Next generation sequencing approaches (e.g., pyrosequencing and Illumina technology) generate data that are several orders of magnitude superior than traditional sequencing methods; still they have limited ability for the assessment of the total microbial diversity in a soil sample, albeit the taxonomic richness (Margulies et al., 2005; Huse et al., 2009; Gloor et al., 2010). This is primarily due to the lack of reference genome libraries even for the dominant bacterial species in forest soil ecosystems (Howe et al., 2014).

In nature, microbes are vital contributors to biogeochemical transformations and hence examining their response to anthropomorphic activities over long periods is critical for

Abbreviations: Con, control; LN, low nitrogen; HF, Harvard Forest; HN, high nitrogen; Min, mineral; Org, organic; CT, confidence threshold.

understanding ecological processes (Falkowski et al., 2008). Several studies have been conducted to reveal the extent of diversity of the soil bacterial and fungal communities being influenced by various factors including aboveground plant populations (Carney and Matson, 2006; Uroz et al., 2010; McGuire et al., 2012), soil type (Roesch et al., 2007; Lauber et al., 2008), soil pH (Fierer and Jackson, 2006; Lauber et al., 2009; Rousk et al., 2010), and differences in geographic location (Fulthorpe et al., 2008; Langenfeld et al., 2013; Bischoff et al., 2014). It is evident that soil microbial communities are influenced by numerous human activities, particularly land management practices, including long term nitrogen (N) fertilization of both agricultural and forest soils (Wu et al., 2008, 2011; Hallin et al., 2009; Ramirez et al., 2010; Fierer et al., 2012; Sridevi et al., 2012; Coolon et al., 2013; Turlapati et al., 2013). Wagg et al. (2014) have recently stressed the importance of changes in soil microbial diversity on nutrient cycling at several sites.

During the 1990s, N added to the atmosphere through human activity was much higher (160 Tg Y^{-1}) than through natural biological fixation processes (110 Tg Y^{-1}) (Gruber and Galloway, 2008). These authors suggested that increased N has multiple effects on cycling of other elements in the environment including carbon (C), leading to global warming. Acidification due to N saturation causes forest soils to become deficient in important labile pools of nutrients, particularly Ca^{2+} and Mg^{2+} (Currie et al., 1999). Calcium deficiency in the soil is known to predispose plants to disease and pathogen infection, thus contributing to decline in forest productivity as seen in red spruce and sugar maple stands in the Northeastern US (Shortle and Smith, 1988; Long et al., 1997; Minocha et al., 2010; Schaberg et al., 2011). These aboveground changes are often accompanied by belowground changes in soil chemistry that alter the microbiome, which in turn impacts the biogeochemical cycling of essential nutrients (N, C, and P). Recent studies have reported reduced microbial biomass and activity with N fertilization of forest soils (Treseder, 2008; Janssens et al., 2010). Other reports have indicated either an increase (Cusack et al., 2010), or a neutral response (Zhao et al., 2013) in microbial biomass with N fertilization. The variability of these findings suggests that the effects of N addition on microbial populations may be site-specific.

At the HF Long-Term Ecological Research site located in Petersham, MA, USA (HF)[1], experimental plots were set up in 1989 to study the long-term effects of N addition on above- and belowground communities (Magill et al., 2004). Past studies from this site have shown negative shifts in the ratio of fungal: bacterial biomass, microbial biomass C, and substrate-induced respiration rates in response to N additions (Bowden et al., 2004; Frey et al., 2004; Wallenstein et al., 2006; Ramirez et al., 2012). Soil from N treatment plots at HF were reported to accumulate more C due to a decrease in decomposition rate (Frey et al., 2014). Restriction fragment length polymorphism (RFLP) and PCR profiles for DNA extracted from N-treated soil samples exhibited altered functional N-cycle gene composition (Compton et al., 2004). Using pyrosequencing of the PCR-amplified 16S rRNA genes from the soil DNA, our group showed that N addition caused profound rearrangements in the structure of bacterial communities at the HF site (Turlapati et al., 2013). Major changes were recorded in the community structure of *Acidobacteria*, α and β subclasses of *Proteobacteria, and Verrucomicrobia* were observed. These conclusions were derived using the UCLUST tool in Quantitative Insights into Microbial Ecology (QIIME) toolkit using the latest version available (1.4.0) at that time (Caporaso et al., 2010b; Edgar, 2010) to cluster the sequences into operational taxonomic units (OTUs) at 97% sequence identity. It was observed that 2% of the total OTUs in this dataset contained $\geq$50% of the total sequences; on the other hand, up to 80% of total OTUs were highly diverse and contained $\sim$10% of the total sequences.

It has been suggested that there is substantial phylogenetic diversity in marine and soil environments attributable to the occurrence of rare bacterial populations (Sogin et al., 2006; Lynch et al., 2012). Although the diversity of HF soil microbes could be estimated to some extent in our previous study, it was not possible to examine the sequence diversity within each abundant OTU since classification was assigned only to the OTU representative sequences (Turlapati et al., 2013). This is important because OTU clustering (often done at 97% sequence identity) is less powerful in identifying phylogenetically distinct organisms that differ by a small number of nucleotides (Eren et al., 2013). Oligotyping is a recently developed computational tool that allows users to choose entropy components ('supervised tool') that have high variability in order to resolve underlying diversity among sequences within each OTU or taxonomic group (Eren et al., 2013).

With the aim of analyzing the effects of prolonged N treatment on individual bacterial groups, and identifying additional genera/families whose presence and/or abundance may be correlated with alterations in soil factors, we subjected our pyrosequencing dataset (Turlapati et al., 2013) to the oligotyping pipeline and taxonomy was assigned using the recently updated RDP database (Cole et al., 2013). The specific objectives of the study were to: (1) demonstrate the applicability of the oligotyping pipeline for forest soil datasets; (2) study the effects of N-amendment on individual bacterial taxa and compare these with previous findings based on OTU clustering; and (3) evaluate the effects of soil chemistry on bacterial communities of Org and Min horizons.

MATERIALS AND METHODS

SITE DESCRIPTION AND SOIL SAMPLE COLLECTION

The study site is a mixed hardwood stand naturally regenerated after being clear-cut in 1945 and is located on Prospect Hill at the HF[2], Petersham, MA, USA. The stand is comprised of predominantly red oak (*Quercus rubra L*) and black oak *(Q. vetulina* Lam.), mixed with red maple (*Acer rubrum* L.), American beech (*Fagus grandifolia* Ehrh.) and black birch (*Betula lenta* L.). Soil at this site is mostly stony to sandy loam formed from glacial till. For more details on site description including vegetation, climate, site topography, and N amendments refer to Aber et al. (1993) and Magill et al. (2004).

As described in our previous report (Turlapati et al., 2013), three 30 m × 30 m treatment plots (further subdivided into 36 subplots; each measuring 5 m × 5 m) were used for sample collection.

[1]http://harvardforest.fas.harvard.edu/research/LTER

[2]http://harvardforest.fas.harvard.edu/

These plots were established in May 1988 as part of a long-term study on chronic N effects on ecosystem function. One plot served as a Con and 2 other plots were treated with NH_4NO_3; low N (LN; treated with 50 kg N ha^{-1} yr^{-1}), and high N (HN; treated with 150 kg N ha^{-1} yr^{-1}). Ammonium nitrate solution was applied by a backpack sprayer yearly in six equal doses at 4-week intervals from May to September. Soil samples were collected from five randomly selected subplots within each treatment plot in September 2009 using a soil corer (7.5 cm diameter). The upper Org layer (Org, average 8 cm) was separated from the lower Min layer. In a few cases where the Org horizon was 10–12 cm deep, deeper coring was needed to get to the Min soil. Thirty samples (5 cores per plot × 2 horizons × 3 treatments) were collected in polyethylene bags and brought to the laboratory on ice. Samples were sieved (2 mm pore size) to remove roots, debris and stones, and then stored at −20°C for further use.

SOIL CHEMICAL ANALYSES

Air-dried soil samples (20–40 g) were sent to the Soil Testing Service Laboratory at the University of Maine, Orono, ME, USA[3] for analyses. Nitrate and NH_4 -N were extracted in potassium chloride and determined colorimetrically by Ion Analyzer in 2012. The rest of the analyses were carried out in 2010 as described in Turlapati et al. (2013). The methods for the extraction of polyamines and amino acids were described in our previous publication (Frey et al., 2014).

DNA ISOLATION, PCR, PYROSEQUENCING, AND DATA QUALITY FILTERING

As previously described in Turlapati et al. (2013), PowerSoil® DNA isolation kit (MO-BIO Laboratories, Carlsbad, CA, USA) was used to isolate genomic DNA from 0.5 g of soil samples. Universal primers (F968 5′AA CGC GAA GAACCT TAC3′ and R1401-1a 5′CGG TGT GTA CAA GGC CCG GGA ACG3′) as described in Brons and van Elsas (2008) with 30 barcodes (10 bp, one for each soil sample) were used for PCR to generate ∼433-bp amplicons corresponding to the V6–V8 hypervariable region of the bacterial 16S rRNA encoding gene. The PCR amplifications were conducted in triplicate using Phusion® Taq Master Mix (New England Biolabs, Ipswich, MA, USA) with 50 ng of template DNA in a final volume of 50 μL. The reactions were performed in a PTC-100® Programmable Thermal Cycler (MJ Research, Inc., Waltham, MA, USA) with the following conditions: an initial denaturation at 95°C for 5 min, followed by 20 cycles of denaturation at 95°C for 30 s, annealing at 61°C for 30 s, and extension at 72°C for 45 s, with a final extension at 72°C for 10 min. The triplicate reaction products (amplicons) from each soil sample were pooled for sequencing. DNA purification kit (Zymo Research, Irvine, CA, USA) was used to purify the pooled PCR products which were then subjected to further cleaning via the Agencourt® AMPure® XP Bead Purification method (Agencourt Bioscience Corporation, Beverly, MA, USA) to remove fragments <100 bp. The quality of the PCR products were evaluated in an Agilent 2100 Bioanalyzer using the DNA 1000 LabChip (Agilent Technologies, Palo Alto, CA, USA). The 30 bar-coded samples were pooled in equimolar quantities (Margulies et al., 2005) in order to process for sequencing (Roche 454 GS-FLX Titanium System) at the University of Illinois, USA[4] in a full picotiter plate.

OLIGOTYPING ANALYSIS

The forward primer (549,500 sequences) pyrosequencing data were quality filtered in QIIME (version 1.4.0) with default settings for most steps as described in Caporaso et al. (2010b). Chimeric sequences were also removed using Chimera Slayer in QIIME (Table S1). After removal of low quality and chimeric sequences, the remaining data were used as an input taxonomic assignment using the Ribosome Database Project (RDP) classifier version 2.7[5]. In the first step, IDs of the sequences corresponding to each phylum were extracted individually from the dataset using CT ≥0.8 with Perl script (Supplementary Material – 1). The sequences corresponding only to these IDs (selected at CT ≥0.8) were then extracted out by using the command filter.fasta.py (available in QIIME) and were further processed for oligotyping analyses. Selected sequences corresponding to a phylum or a subgroup/class were aligned using PyNAST (Caporaso et al., 2010a). PyNAST enables the alignment of sequences against a template database such as Greengenes (McDonald et al., 2012) in QIIME. Before beginning oligotyping, the uninformative gaps in the alignment were removed along with 10–15 nucleotides at the end of each read, which were trimmed to attain reads of similar lengths.

Oligotyping was conducted individually for each phylum. The only exceptions were *Proteobacteria* and *Acidobacteria* where oligotyping had to be conducted individually for each class or subgroup because among some of the subgroups there were several-fold differences in the number of sequences (e.g., within *Acidobacteria*, subgroup *Gp*1 accounted for 151,462 sequences and *Gp*4 had only 715 sequences); these required different *M* values (the minimum substantive abundance of the most abundant unique sequence in an oligotype) for analyses. To begin with all sequences were 430 bp in size; after alignment and filtering the gaps, the sizes of the aligned reads were different for different phyla (Figure S1).

Entropy and initial oligotyping analyses were conducted according to Eren et al. (2013). The oligotyping method utilizes Shannon (1948) entropy for detecting the amount of diversity associated with each nucleotide position and provides a way of identifying positions associated with greater variability. The entropy peaks for nucleotide positions ranged from 0 to slightly >2 for most of the datasets under consideration with the exception of the phylum *Nitrospira* (Figure S1). We observed that a higher number of peaks with entropy values ≥1.0 resulted in greater bacterial diversity. In the first step, oligotypes were generated by using the first position with the highest entropy value. To decompose these oligotypes further, supervised oligotyping steps (user-defined nucleotide selection for decomposing entropy) were used. On the average 50 positions were chosen to decompose entropy for most datasets. Oligotyping was continued until no peaks were left unresolved that could further decompose the oligotype. Peaks with entropy values of <0.2 often did

[3]http://anlab.umesci.maine.edu

[4]http://www.biotech.illinois.edu/htdna

[5]http://sourceforge.net/projects/rdp-classifier/

not yield additional oligotypes and were considered background noise. Details for the above steps are described in Eren et al. (2013).

For supervised oligotyping, *M* values had to be modified for forest soil samples from those suggested for other ecosystems by Eren et al. (2013). Since the total number of sequences varied across different bacterial phyla, the *M* values varied accordingly for each analysis as shown in Table S2. This value was usually lower (2–5) for relatively smaller datasets in the range of 200–5000 reads. In order to correct for technical errors due to pyrosequencing, the value for parameter 's' ('s' is the minimum number of samples in which an oligotype is expected to be present) was set to two with the assumption that any sequence that occurs in two biological samples represents an element of a microbiome and is not an error. Huse et al. (2010) found that even with deep sequencing, OTUs with one sequence rarely occurred in other replicates and the chance of a spurious OTU occurring in two environmental samples is not realistic which supports our assumption. To capture the rare biosphere in soil samples, no values were assigned to parameter 'a' (the minimum percent abundance of an oligotype in at least one sample) and parameter 'A' (the minimum actual abundance of an oligotype across all samples). Oligotype representative sequences (the most abundance sequence within an oligotype) were classified at CT $\geq$0.8 (a generally accepted threshold of 80% for assigning taxonomy) using the RDP classifier tool version 2.7[6]. For determining percent alignment scores, the sequences corresponding to individual oligotypes were extracted from the oligo-representatives directory and aligned using the ClustalW2 tool[7].

In one small comparative study with a subset of data, oligotype representative sequences were classified at CT $\geq$0.5 as well as $\geq$0.8 using the RDP classifier to determine if additional genera could be identified using a lower CT value.

The oligotype representative sequences have been deposited in the NCBI short read archive. The accession numbers are presented in Table S1.

OTU CLUSTERING AND TAXONOMIC ANALYSES

To compare the oligotyping method with OTU clustering, quality-filtered reads of four phyla (*Actinobacteria, Bacteroidetes, Firmicutes,* and *Proteobacteria*) were selected and clustered into OTUs using UCLUST (Edgar, 2010) set at a 97% identity threshold in QIIME (version 1.8.0). The OTU representative sequences were picked in QIIME based on the most abundant sequence in each OTU. Similarly, the oligotyping method also assumes that the most abundant unique sequence is the oligotype representative sequence. For each dataset, sequences were filtered for minimum abundance (*n* size) for each OTU using the same value that was used for *M* in the corresponding oligotyping analyses. In addition, in order to match the parameter set for oligotyping ($s = 2$), OTUs that were not present in at least two samples were removed from the OTU table using python scripts (Supplementary Material – 2). The OTU representative sequences were classified using RDP version 2.2 (currently used version) in QIIME.

In order to understand the reason for the difference in the genera identified by the two methods, we assigned taxonomy to OTU and oligo representative sequences using RDP in QIIME pipeline version 2.2 and online RDP classifier version 2.2.

STATISTICAL ANALYSES

SYSTAT (version 10.2, Systat Software, Inc., San Jose, CA, USA) was used for standard statistical tests, including paired *t*-tests and two-way analysis of variance (ANOVA), on the soil NH_4 and NO_3 data. Non-metric dimensional scaling (NMS) analyses were conducted using PC-ORD (version 6.03, MJM Software Design, Gleneden Beach, OR, USA). To normalize the data, digit one was added to all data before log10 transformation. Briefly, following settings were used for NMS: number of axes = 4, maximum number of iterations = 500; stability criterion (the standard deviation in stress over the last 10 iterations) = 10^{-6}; number of runs with real data = 100; and the number of runs with randomized data = 250. Random numbers were chosen as a source of starting ordinations. The tie handling was done by penalizing unequal ordination distance (Kruskal's secondary approach). The following were chosen as output options: varimax, randomization test, plot stress vs. iterations and calculate scores for OTUs by weighted averages. Dimensionality of solutions was selected for these analyses based on the assessment using a graph of stress as a function of dimensionality (scree plot). A Monte Carlo test was used to examine the stress and the strength of NMS results. Two-way permutational MANOVA was conducted using the Bray-Curtis distances to evaluate the effect of the horizons and the treatment, and the interaction between them. Mantel tests were conducted to evaluate the significance of correlations among Bray-Curtis distance scores and soil chemistry and soil Org N metabolites.

RESULTS

SOIL CHEMISTRY

While NH_4 concentration in the Org soil was significantly higher than that in the Min soil for all treatments (Table S3), there was no difference in NO_3 levels between the two horizons. Long-term N treatment did not significantly alter either NH_4 or NO_3 concentrations of either soil horizon (Table S3). Other details on soil analyses are described in Frey et al. (2004) and Turlapati et al. (2013).

PARAMETERS OF OLIGOTYPING ANALYSES

Soils are highly diverse and harbor an abundance of rare microbes. Rare microbes are more prone to primer-PCR amplification and sequencing biases thus making it harder to identify such individuals. In addition, it is well known that soil replicates have high microsite variability in chemistry as well as bacterial populations. This combination of rarity and microsite variability perhaps is the reason that the same microbes are not present in all replicate samples, and why the guidelines suggested for other biomes in Eren et al. (2013) did not work with the HF soil samples. Thus for analysis of these soils, the suggested guidelines had to be modified (personal communication with Dr. A. Murat Eren, Marine Biological Laboratories, Woods Hole, MA, USA).

[6]http://sourceforge.net/projects/rdp-classifier/files/rdp-classifier/rdp_classifier_2.7.zip/download

[7] http://www.ebi.ac.uk/Tools/msa/clustalw2/

M is the value that is used to filter out potential noise in a sample. For this reason it is generally kept at a reasonably high number. It is likely that with forest soils, sequences representing rare taxa may be filtered out as noise due to their low abundance at high *M* values. In addition, high *M* values in such cases filtered out more than 50% of all sequences. The more diverse the phyla are (in terms of number of entropy components) more are the sequences that are filtered out with high *M* values (Table S2, e.g., see *Bacteroidetes vs. Gp10*). With the goal of retaining maximum sequences and diversity in terms of the number of oligotypes, several *M* values were tested for most phyla before final analyses. We tested different *M* values *using* α- and β-*Proteobacteria*, two well-known bacterial classes that are highly diverse and varied in the size of data for our soil samples with 38,858 and 3,340 sequences, respectively. The *s* value was set at two for both analyses. Lowering *M* values resulted in the identification of more genera at CT ≥0.8 in both classes (**Table 1**). For example at CT of ≥0.8, the total number of genera identified with an $M = 75$ for α-*Proteobacteria* was 5, but with $M = 15$, 11 genera were identified. Similarly, for β-*Proteobacteria*, $M = 25$ identified only two genera while $M = 3$ identified eight genera (**Table 1**) In α-*Proteobacteria*, genera such as *Labrys* and *Acidisoma* were identified with $M = 15$; they would be missed at higher *M* values. Similar results were obtained within β-*Proteobacteria* (**Table 1**). Therefore, final data analyses in this study were conducted using an *M* value that was based on two criteria, namely, the retention of maximum sequences, and maximum diversity (in terms of number of oligotypes) with taxonomic assignment at CT values no less than 0.8. Data presented here show that using high *M* values for groups that have low abundance would not identify genera with high confidence limits [e.g., at $M = 2$ *Paenibacillus* (*Firmicutes*) at CT = 1 and *Mucilaginibacter* (*Bacteroidetes*) at CT of 0.99–1 were identified].

ALIGNMENT OF SEQUENCES WITHIN OLIGOTYPES

We compared the sequence identities within OTUs clustered at ≥97% similarity (Edgar, 2010) in QIIME with those in oligotypes that are generated by manually selecting components of high entropy values. The results show that the sequence identities within an oligotype ranged between 99.5 and 100%. In order to reduce variation among sequence identities within an oligotype, all peaks with entropy values greater than 0.6 were resolved. The oligotyping process left very few unresolved peaks of relatively low entropy (<0.2) values within oligotypes that often had <100 sequences. These low entropy peaks were considered as the background noise (**Table 2**). However, even when the oligotyping analysis appeared not fully resolved, the range of % identities

Table 1 | Effect of varying *M* values on percent of reads retained, total number of oligotypes, and genera identified at 0.8 CT with RDP database.

Class	*M* value	% of reads retained	Number of oligotypes	Genera identified at 0.8 CT in RDP	Total number of genera
α-*Proteobacteria*	75	35	73	*Bradyrhizobium, Rhodomicrobium, Rhizomicrobium, Methylocella, Methylosinus*	5
"	50	44	118	*Bradyrhizobium, Rhodomicrobium, Rhizomicrobium, Methylocella, Methylosinus, Hyphomicrobium, Phenylobacterium*	7
"	30	55	211	*Bradyrhizobium, Rhodomicrobium, Rhizomicrobium, Methylocella, Methylosinus, Hyphomicrobium, Phenylobacterium, Acidisphaera, Bauldia*	9
"	15*#	67	389	*Bradyrhizobium, Rhodomicrobium, Rhizomicrobium, Methylocella, Methylosinus, Hyphomicrobium, Phenylobacterium, Acidisphaera, Bauldia, Labrys, Acidisoma*	11
β-*Proteobacteria*	25	64	27	*Burkholderia, Herbaspirillum*	2
"	20	65	29	*Burkholderia, Herbaspirillum*	2
"	15	69	37	*Burkholderia, Herbaspirillum, Comamonas, Herminiimonas*	4
"	10	74	48	*Burkholderia, Herbaspirillum, Comamonas, Herminiimonas, Collimonas*	5
"	5$	81	83	*Burkholderia, Herbaspirillum, Comamonas, Herminiimonas, Collimonas, Nitrosospira, Variovorax*	7
"	3*$	86	123	*Burkholderia, Herbaspirillum, Comamonas, Herminiimonas, Collimonas, Nitrosospira, Variovorax, Aquabacterium*	8
"	2	88	157	*Burkholderia, Herbaspirillum, Comamonas, Herminiimonas, Collimonas, Nitrosospira, Variovorax, Aquabacterium*	8

The datasets used for illustrating this point are α-Proteobacteria (38,858 sequences) and β-Proteobacteria (3,440 sequences); s value used for these analyses was 2.
**Indicates final M values used for the rest of the analyses. # Indicates that below an M value of 15 a machine generated error was encountered due to exceedingly large numbers of oligotypes. $ Indicates that Nitrosospira, Variovorax, and Aquabacterium were identified at CT of (0.98–1).*

was still within 99.76–100%. **Table 2** shows two such examples: in one case there are two peaks with entropy values of <0.25, and in another case there is only one large peak with an entropy value at 0.65. In such cases, the small peaks were deliberately disregarded because their further decomposition did not result in additional oligotypes. The sequence identities within OTUs clustered using the same dataset ranged from 97 to 100% (results not shown).

TAXONOMIC ASSIGNMENTS USING DIFFERENT TOOLS

Major motivation for examining the use of oligotyping as a tool was to reveal concealed microbial population diversity that could not be seen using the OTU clustering approach (Turlapati et al., 2013) with the additional focus on the effects of N treatment on the distribution of these taxa. Comparison of taxonomic assignments between OTUs described earlier using QIIME 1.4.2 and the data presented in this report show more genera were identified using the oligotyping analysis (**Table 3**). Although the same data set was used for both analyses, genus level information for phyla such as *Acidobacteria*, *Bacteroidetes,* and *Chloroflexi* was generated in the present study. It should be pointed out that the use of an updated version (2.7) of RDP also revealed four previously unidentified phyla (AD3, *Cyanobacteria,* TM6 *and* WPS-2; Table S2).

Whereas OTU clustering resulted in 2% of the OTUs containing ∼50% of the total sequences (Turlapati et al., 2013), for oligotyping, 2% of the oligotypes contained ∼38% of the total number of sequences (results not shown). Another major difference was that 80% of the OTUs contained 10% of sequences in comparison to 20% sequences in 80% of oligotypes. A comparison of two databases (Greengenes used in QIIME vs. RDP online database) for classifying OTUs as well as oligotypes representative sequences resulted in a significantly lower number of taxa identified from both OTU and oligotype datasets with Greengenes as compared to RDP (**Table 4**). More genera were discernible when CT ≥0.5 was used (vs. ≥0.8) with oligotype representative sequences, (Table S4). One such example is the genus *Terriglobus* (*Acidobacteria-Gp*1), which was identified only at CT ≥0.5 (Figure S2).

BACTERIAL DIVERSITY ANALYSES

Details on the various steps of oligotyping analysis and the number of oligotypes identified for each phylum are given in Table S2. In general, no direct correlation was observed between the numbers of sequences and the oligotypes. In all, sequences affiliated with 73 known genera were identified from these soils as compared to 27 genera observed in our previous study (**Table 3**). Oligotyping revealed that sequences corresponding to some genera were present in control but absent in N-treated soils. In other instances, those present in N-amended plots were not found in control plots.

BACTERIAL COMMUNITIES AND TREATMENT RELATIONSHIPS

Non-metric multidimensional scaling (NMS) followed by permutational MANOVA (Permanova) with ordination scores, i.e., the Bray-Curtis distance, obtained from normalized total oligotype data revealed significant differences among bacterial communities based on treatment and soil horizon (**Figure 1**; Table S5). In general, all five replicate soil samples from within the same treatment plot clustered together and displayed stronger similarities among their oligotypes as compared to those from other treatment plots. For the two largest phyla, *Acidobacteria* and *Proteobacteria*, NMS and Permanova analyses were conducted for individual subgroups and classes, respectively (Figure S3; Table S5). With the exception of *Bacteroidetes* and subgroup *Gp*10 in *Acidobacteria,* the bacterial communities of the two horizons were significantly different. Additionally, within each soil horizon significant treatment effects on the structure of the bacterial community were observed for all phyla except for *Chloroflexi*, *Firmicutes*, TM7, *Bacteroidetes*, and subgroups *Gp*6 and *Gp*10 of *Acidobacteria* (**Figure 1** and Table S5; Figure S3).

The highest variation in partitioning was observed in the phylum *Acidobacteria,* where horizon explained 49.0% ($P \leq 0.0002$) of the variation among samples, and treatment accounted for about 21% ($P \leq 0.0004$) of the variation (**Figure 1A**; Table S5). Most of this variation was due to subgroups *Gp*1, *Gp*2, and *Gp*3 (Figure S3). In the second most diverse phylum, *Proteobacteria,* horizon explained 36% ($P \leq 0.0002$) of variation among the samples and the treatment accounted for 16.5% ($P \leq 0.0006$) of the variation;

Table 2 | CLUSTALW percent identity alignment scores for the sequences within each oligotype, taxonomic affiliation of the oligotype, total number of unresolved peaks, and the entropy values associated with the nucleotide components of unresolved peaks.

Oligo ID	Taxonomic affiliation	Number of sequences within oligotype (number of 100% identical sequences)	% identity among the sequences	Number of unresolved peaks	Entropy associated with the nucleotide components of unresolved peaks
00054	*Acidobacteria: Gp1*	498 (462)	99.53–100	Background noise*	<0.10
00009	*Acidobacteria: Gp3*	469 (436)	99.30–100	Background noise*	<0.15
00008	*α-Proteobacteria*	276 (255)	99.53–100	Background noise*	<0.10
00012	*β-Proteobacteria*	52	100	none	–
00028	*β-Proteobacteria*	23 (22)	99.76–100	2	0.25
00040	*β-Proteobacteria*	14 (13)	99.76–100	2	0.37
00165	*Verrucomicrobia*	54 (43)	99.54–100	1	0.65

**Background noise is defined as the situation when the unresolved peaks with entropy values associated with the nucleotide components fall within the range of 0.1–0.2. This often happened with oligotypes having >100 sequences.*

Table 3 | Comparison of genera identified by QIIME (version 1.4.0) UCLUST clustering [method used in our previous study (Turlapati et al., 2013)] with those identified in the present study using oligotyping.

Phylum/class	Genera identified in Turlapati et al. (2013) using QIIME UCLUST clustering method	Total number	Genera identified in present study using oligotyping	Total number
Acidobacteria	–	0	*Acidobacterium, Bryobacter, Edaphobacter, Granulicella*	4
α-Proteo	*Bradyrhizobium, Methylocella, Phenylobacterium, Rhodomicrobium*	4	*Acidisoma, Acidisphaera, Bauldia, Bradyrhizobium, Hyphomicrobium, Labrys, Methylocella, Methylosinus, Phenylobacterium, Rhizomicrobium, Rhodomicrobium*	11
β-Proteo	*Burkholderia*	1	*Aquabacterium, Burkholderia, Collimonas, Comamonas, Herbaspirillum, Herminiimonas, Nitrosospira, Variovorax*	8
δ-Proteo	*Byssovorax**	1	–	0
γ-Proteo	*Aquicella, Dyella, Legionella, Pseudomonas, Serratia, Stenotrophomonas*	6	*Aquicella, Coxiella, Dyella, Legionella, Nevskia, Pseudomonas, Rhodanobacter, Rudaea, Serratia, Stenotrophomonas, Yersinia*	11
Actinobacteria	*Actinospica, Actinocorallia*, Catenulispora, Conexibacter, Kitasatospora, Mycobacterium, Nocardia,*	7	*Aciditerrimonas, Actinospica, Catenulispora, Conexibacter, Kitasatospora, Microbacterium, Mycetocola, Mycobacterium, Nocardia, Solirubrobacter, Streptacidiphilus*	11
Verrucomicrobia	*Opitutus*	1	*Alterococcus, Opitutus*	2
Chlamydiae	*Neochlamydia, Parachlamydia, Rhabdochlamydia*	3	*Neochlamydia, Parachlamydia, Simkania*	3
Chloroflexi	–	0	*Ktedonobacter, Thermosporothrix*	2
Elusimicrobia	–	0	*Elusimicrobium*	1
Firmicutes	*Bacillus, Paenibacillus*	2	*Ammoniphilus, Bacillus, Brochothrix, Clostridium XI, Clostridium sensu stricto, Cohnella, Lactococcus, Lysinibacillus, Paenibacillus, Solibacillus, Sporosarcina, Viridibacillus*	12
Gemmatimonadetes	*Gemmatimonas*	1	*Gemmatimonas*	1
Bacteroidetes	–	0	*Flavobacterium, Mucilaginibacter, Niabella, Pedobacter, Sphingobacterium, Terrimonas*	6
Nitrospira	*Nitrospira*	1	*Nitrospira*	1
Total genera		**27**		**73**

*Genera were identified at CT ≥0.8 using the online RDP classifier. *The genera Byssovorax and Actinocorallia were identified by OTU clustering.*

a major part of this variation was seen in *α*-and *γ*-*Proteobacteria* (Figure S3). Often there was an overlap between the LN- and HN-amended soil communities. Treatment effects were generally more pronounced in the Org soil horizon.

SOIL CHARACTERISTICS AND BACTERIAL COMMUNITIES

Mantel test results on soil chemistry and the bacterial community revealed strong correlations for the Org soil. Pooled oligotyping data from all phyla found in the Org soil horizon showed a stronger positive correlation between the entire bacterial community and soil pH, Ca, P, K, Zn, Mg, NH_4, NO_3, and total C (**Table 5**) as compared to Min soil. When data for each phylum were analyzed separately, significant correlations were observed for *Acidobacteria*, *Proteobacteria*, *Actinobacteria*, *Verrucomicrobia*, WPS-2, and AD3. Exceptions included *Proteobacteria* which showed no correlation with Ca and P, nor did *Chlamydiae* with NH_4 and total C. The remaining phyla had correlations with fewer elements, specifics of which varied (**Table 5**).

Oligotyping data pooled for all analyzed phyla from the Min horizon community showed strong positive correlations only with soil pH, acidity, NH_4, and NO_3. With few exceptions, this was

Table 4 | The effect of different databases on classification.

Database and tools comparisons for classification of oligo representatives (ORs) and operating taxonomic units OTUs				
Method	**Oligotyping**		**OTU clustering**	
CT	**0.8**		**0.8**	
Tool	**RDP classifier online**	**RDP in QIIME**	**RDP classifier online**	**RDP in QIIME**
Database	**RDP**	**Greengenes**	**RDP**	**Greengenes**
Actinobacteria	11	6	13	6
Bacteroidetes	6	2	3	1
Firmicutes	12	15	14	13
α-Proteo	11	6	14	11
β-Proteo	8	3	6	2
δ-Proteo	0	1	2	1
γ-Proteo	11	8	11	7
Total	**59**	**41**	**63**	**41**

Oligotyping representative as well as OTU representative sequences were classified using the online RDP classifier version 2.2 (RDP database) and RDP version 2.2 within QIIME version 1.8.0 (Greengenes database). The OTU clustering was performed using UCLUST in QIIME with the identical parameters (s and M) that were set up for oligotyping. All genera identified at ≥0.8 CT.

true for analyses of each individual phylum. There was a positive correlation of Al with *Acidobacteria* and WPS-2, and Ca with *Proteobacteria* (**Table 5**).

CHANGES IN COMMUNITY STRUCTURE OF GENERA ASSOCIATED WITH SOIL HORIZON AND N-AMENDMENTS

Across all three treatments, oligotypes corresponding to five genera (*Ammoniphilus, Clostridium X1, Solibacillus, Sporosarcina,* and *Viridibacillus*) were found in Min soils but were absent in Org soils, and one genus (*Conexibacter*) was identified in Org soils but not in Min soils (**Table 6**; **Figures 2A,C**). Overall, oligotypes corresponding to five genera (*Aquabacterium, Nitrosospira, Yersinia, Legionella,* and *Niabella*) appeared exclusively in N-treated soils (in both horizons combined), while those representing eight genera (*Comamonas, Microbacterium, Mycetocola, Brochothrix, Flavobacterium, Pedobacter, Sphingobacterium,* and *Terrimonas*) were present in the control but absent from the N-treated soils (**Table 6**; **Figures 2B,C**). A comparison of the presence of oligotypes specific to each genus revealed that some oligotypes were unique to Org soils while others were exclusively present in Min soils (Table S6; **Figure 2**). Additionally some oligotypes were present only in the N-amended soils and not in the control plots. Although Figures and Tables show all identified genera, only those that exhibited significant differences with treatment or soil horizon are discussed in this section.

TAXONOMY BY PHYLUM

Acidobacteria, which was the most abundant phylum in the HF soils, constituted ∼50% of total sequences and 22% of total oligotypes (Table S2). Subgroups *Gp*1, *Gp*2, and *Gp*3 accounted for most of the *Acidobacteria* sequences and oligotypes. Soil samples from control plots had three times more sequences affiliated with *Gp*1 as compared to *Gp*3. However, based on the total number of oligotypes, the *Gp*3 subgroup was more diverse than *Gp*1 in both soil horizons (Figures S4A,B). *Gp*13 was represented by a large number of oligotypes each containing small sets of sequences. Four new genera were identified in the phylum *Acidobacteria* since our last report with the same soil samples (**Table 3**). All of the genera identified in the Org horizon were represented by a significantly larger number of sequences as compared to those identified in the Min soil horizon for all three treatments (Figure S5A). However, the number of oligotypes did not vary greatly between soil horizons. Sequences of the genus *Edaphobacter* were significantly higher in HN vs. the control while *Acidobacterium*, *Granulicella*, and *Bryobacter* showed a reverse trend in the Org horizon (Figure S5A). Oligotypes for *Edaphobacter* were more abundant in N treatment plots relative to control in Min soil of both (**Figure 2C**).

Proteobacteria was the second most abundant phylum and was highly diverse in these soils. This phylum comprised ∼20% of the total sequences and total oligotypes (Table S2). Classes *α-Proteobacteria* and *γ-Proteobacteria* together accounted for >70% of this phylum's sequences and oligotypes in each horizon (Figures S6A,B). Regardless of the number of sequences, oligotyping data revealed that among all other classes and phyla, *α-Proteobacteria* were the most diverse in both soil horizons. In general, LN treatment was positively correlated with sequences and oligotype numbers. Altogether, 30 genera were found in *Proteobacteria* (Figures S5B–D, **Table 3**): 11 each in *α-* and *γ-Proteobacteria* (Figures S5B–D), and eight in *β-Proteobacteria* (Figure S5C). Fourteen out of these 30 genera were present in relatively high abundance (sequences ≥50). Shifts in community structure were observed among most genera in response to N treatments (Figures S6A,B). An increase in the number of sequences was observed for several genera in soil treated with LN (Figures S5B,D).

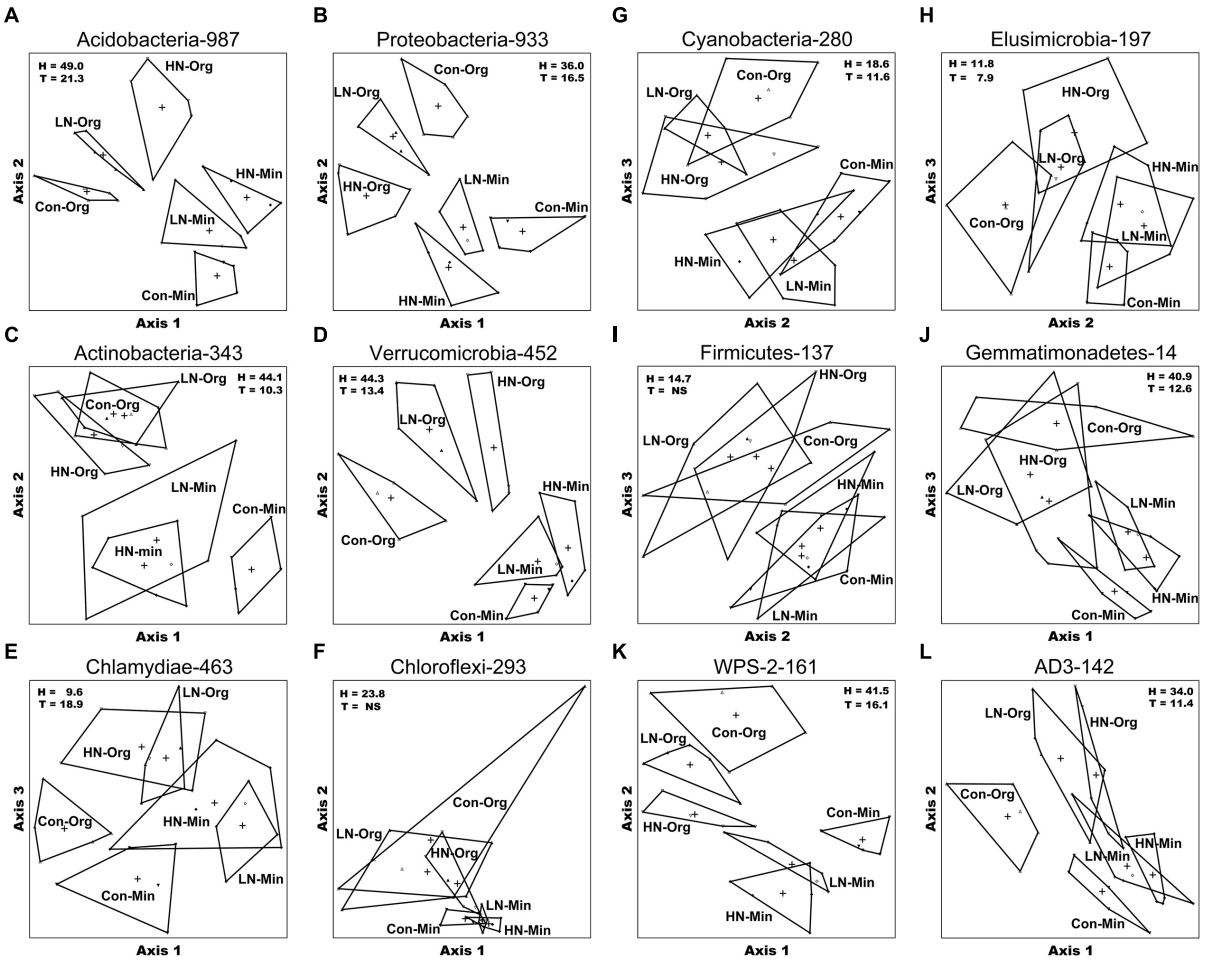


FIGURE 1 | Non-metric dimensional scaling (NMS) ordination for oligotypes of 30 soil samples for 12 of the 16 identified bacterial phyla. Letter symbols refer to the NMS analyses for the following phyla: **(A)** Acidobacteria; **(B)** Proteobacteria; **(C)** Actinobacteria; **(D)** Verrucomicrobia; **(E)** Chlamydiae; **(F)** Chloroflexi; **(G)** Cyanobacteria; **(H)** Elusimicrobia; **(I)** Firmicutes; **(J)** Gemmatimonadetes; **(K)** WPS-2; and **(L)** AD3. Each soil type is represented by 5 replicates and a centroid, which is indicated by a single symbol and treatment-soil horizon name. The number next to the phylum name represents the total number of oligotypes identified within each phylum. H = % of variation partitioned by horizon and T = % of variation partitioned by treatment.

Actinobacteria constituted only 7% of total sequences and 7.6% of oligotypes (Table S2). Eleven genera were identified from this phylum relative to the prior findings of seven (by OTU clustering) for the same data set (**Table 3**). The absence of certain genera in N-amended soils was observed in a few cases (Figure S5E).

Verrucomicrobia was the third most abundant phylum with ∼10% of sequences and ∼10% of oligotypes in these soils (Table S2). Only two genera (**Table 3**) from this phylum were identified. Genus *Opitutus* was present in tenfold higher numbers than *Alterococcus* (Figure S5F). The number of sequences for *Opitutus* was lower in N-treated soils than in Control. For both *Actinobacteria* and *Verrucomicrobia*, horizon and treatment explained 44 and 10–13% of the total variation, respectively (**Figures 1C,D**).

Chlamydiae constituted 2.5% of the total sequences and ∼10% of oligotypes (Table S2). Three genera were identified in this phylum (**Table 3**); the relative abundance of oligotypes corresponding to the genus *Parachlamydia* were highest in N treatment plots relative to control in Org soil (**Figure 2B**; Figure S5G).

Chloroflexi constituted 2.7% of the total sequences and 6.5 % of oligotypes (Table S2), with only two genera (*Ktedonobacter* and *Thermosporothrix* – Figure S5H; **Table 3**), which were absent in the control treatment in the Org horizon but present in the Min soils (**Figure 2A**).

Although *Firmicutes* constituted only about 0.3% of total sequences, with 3.0% oligotypes, 12 genera were identified in this phylum (Table S2; Figure S5J). Most of these were present in very low numbers and were more prevalent in Min soils (e.g., *Bacillus* and *Paenibacillus*); all five of the genera seen in the Min soil horizon were absent in the Org soil horizon (**Table 6**; **Figure 2**); there was little effect of N treatments (**Figure 1I**; **Table 5**).

Together the phyla TM7, *Gemmatimonadetes*, and *Nitrospira* comprised ∼1% of the total sequences and 2.4% of oligotypes

Table 5 | Relationship between soil chemistry and Bray-Curtis (Sorenson) distance measures of the normalized oligotypes data (Mantel test) conducted using PC-ORD software (version 6).

Soil Chemistry	pH	Ca	P	Mn	K	Zn	Mg	Acidity	Al	NH_4	NO_3	Total C
Organic soil horizon												
All Bacteria	**0.001***	**0.007***	**0.007***	–	**0.004***	**0.001**	**0.001***	–	–	**0.015***	**0.002***	**0.002***
Acidobacteria	0.011*	0.028*	0.026*	–	0.012*	0.003*	0.002*	–	–	0.021*	0.001*	0.010*
Proteobacteria	0.027*	–	–	–	0.014*	0.001*	0.004*	–	–	0.041*	0.001*	0.037*
Actinobacteria	0.001*	0.006*	0.024*	–	–	0.021*	–	–	–	0.008*	0.024*	0.002*
Verrucomicrobia	0.001*	0.002*	0.012*	–	0.047*	0.004*	0.014*	–	–	0.003*	0.002*	0.006*
Chlamydiae	–	–	–	–	0.009*	0.011*	–	–	–	–	0.001*	–
Chloroflexi	–	–	0.001*	0.019*	0.015*	0.001*	0.006*	0.033*	–	0.013*	–	0.006*
Cyanobacteria	–	–	–	–	0.010*	0.009*	0.014*	–	–	–	0.001*	0.049*
Elusimicrobia	–	0.029*	–	–	–	–	–	–	–	0.011*	0.001*	–
Firmicutes	–	–	–	–	0.026*	0.017*	–	0.041*	–	0.021*	–	0.014*
Gemmatimonadetes	0.016*	0.008*	–	–	–	–	–	–	–	–	–	–
TM7	–	–	–	–	–	–	–	–	–	–	–	–
WPS-2	0.001*	0.023*	0.013*	–	0.003*	0.001*	0.010*	–	–	0.005*	0.002*	0.002*
AD3	0.008*	0.037*	0.002*	–	0.001*	0.001*	0.001*	0.035*	–	0.014*	0.001*	0.003*
Bacteroidetes	0.033*	–	0.017*	–	0.043*	–	–	0.016*	–	–	–	0.010*
Mineral soil horizon												
All Bacteria	**0.047***	–	–	–	–	–	–	**0.020***	–	**0.017***	**0.005***	–
Acidobacteria	0.005*	–	–	–	–	–	–	0.009*	0.035*	0.037*	0.008*	–
Proteobacteria	–	0.009*	–	–	–	–	–	–	–	0.013*	0.005*	–
Actinobacteria	–	–	–	–	–	–	–	–	–	0.013*	–	–
Verrucomicrobia	0.029*	–	–	–	–	–	–	0.013*	–	0.031*	0.032*	–
Chlamydiae	–	–	–	–	–	0.040*	–	–	–	0.032*	0.001*	–
Chloroflexi	–	–	–	–	–	–	–	–	–	–	0.011*	–
Cyanobacteria	–	–	–	–	–	0.023*	–	–	–	–	0.002*	–
Elusimicrobia	–	0.041*	–	–	–	0.031*	–	–	–	0.050*	0.044*	–
Firmicutes	–	–	0.014*	–	–	0.024*	–	–	–	–	–	–
Gemmatimonadetes	0.016*	–	–	–	–	–	0.035*	0.041*	–	–	–	–
TM7	–	–	–	–	–	–	–	–	–	–	–	–
WPS-2	0.003*	–	–	–	–	–	0.026*	0.005*	0.009*	0.007*	0.010*	–
AD3	–	–	–	–	–	–	0.023*	–	–	–	–	–
Bacteroidetes	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

All 15 samples from each soil horizon were pooled for these analyses. The iterations were set to 5000. Asterisks Indicates significant correlations ($P \leq 0.05$). –Indicates non-significance. NA denotes that analyses could not be performed due to insufficient data.

(Table S2). From these three phyla, only two genera were identified; *Gemmatimonas* in *Gemmatimonadetes* and *Nitrospira* in *Nitrospira* (Figures S5L,M). The number of sequences representing the genus *Nitrospira* was greater in Min vs. the Org soil horizon and in LN treatment vs. the control and HN treatment. While both treatment and horizon explained significant variation found in *Gemmatimonadetes* (**Figure 1J**; Table S5) only horizon-specific effects were seen for TM7 (Figure S3L). NMS generated a four dimensional solution for TM7 (Table S5). Because of a small dataset, no further analyses were conducted on *Nitrospira*.

The phyla TM6, *Cyanobacteria*, *Elusimicrobia*, WPS-2, and AD3 were not identified in our previous report with the same data. TM6 was represented by only 36 sequences, most of which were filtered out in the initial oligotyping run and, therefore, were not analyzed further (Table S2). The other five phyla together constituted a relatively small portion of the total sequences (∼5.2%, with numbers ranging from 142 to 280) and 17.2% of the total oligotypes (Table S2). Among these phyla, WPS-2 was the most affected by treatment as well as by horizon (**Figure 1K**). Only the genus *Elusimicrobium* was identified in *Elusimicrobia* (Figure S5I).

Table 6 | Horizon specific genera (shown with an *) and genera that appeared or disappeared with N-amendments.

Phylum or class	Con-Org	LN-Org	HN-Org	Con-Min	LN-Min	HN-Min
Actinobacteria	*Conexibacter** *Microbacterium* *Mycetocola*	–	*Conexibacter**	*Microbacterium* *Mycetocola*	–	–
Firmicutes	*Brochothrix*	–	–	*Ammoniphilus** *Brochothrix* *Clostridium XI** *Sporosarcina**	*Ammoniphilus** *Clostridium XI** *Solibacillus** *Viridibacillus**	*Clostridium XI** *Viridibacillus**
β-Proteobacteria	*Comamonas*	*Aquabacterium*	*Aquabacterium* *Nitrosospira*	*Comamonas*	*Aquabacterium*	*Aquabacterium* *Nitrosospira*
γ-Proteobacteria	–	*Legionella,* *Yersinia*	*Legionella*	–	*Yersinia*	*Legionella*
Bacteroidetes	*Flavobacterium* *Pedobacter* *Sphingobacterium* *Terrimonas*	*Niabella*	–	*Flavobacterium* *Pedobacter* *Sphingobacterium* *Terrimonas*	*Niabella*	–

Bacteroidetes contained only 259 sequences but 36 oligotypes, and six distinct genera were identified in this phylum; four of which were absent from N-amended soils (**Table 6**; Figure S5K). NMS analyses yielded only one dimension solution for this phylum, and thus no figure was generated (Table S5).

DISCUSSION

The primary objective for re-analyses of the pyrosequencing data from the previous study (Turlapati et al., 2013) was to determine additional diversity by applying more efficient and reliable bioinformatics tools. The results enabled us to identify a total of 4534 oligotypes belonging to 15 different bacterial phyla with 73 genera. The same dataset had previously resulted in 6936 OTUs belonging to 11 different bacterial phyla with only 27 identifiable genera. There are two main explanations for this apparent discrepancy: the first being the updating of tools and databases used for classification to newer versions since our previous report; and the second (somewhat obscure) is the difference between the classifiers and databases currently being used for the two clustering methods. Whereas the RDP classifier version 2.2 in QIIME pipeline uses the Greengenes database for OTU- rep classification, the RDP online classifier version 2.7 used for oligotyping has its own built-in RDP database. Although oligotyping and OTU clustering identify similar numbers of taxa (in a comparative study using four phyla constituting ∼28% of total sequences, **Table 4**), the former has greater resolving ability for classifying nearly identical, closely related organisms provided a good reference genomic library is available to compare against. Using entropy analysis, oligotyping simultaneously clusters multiple sequences based on similarity/identity of each nucleotide along the entire length of the read for all sequences within a given group. However, OTU groups sequences at 97% similarity with a representative sequence. The 3% difference in nucleotides may occur anywhere along the entire length of the read, and that location can vary from sequence to sequence within the group. That is to say, 2 sequences within an OTU can differ from the representative sequence at different nucleotide positions. Additionally, once a sequence has been selected and grouped within one OTU it cannot be assigned to another even if it has greater similarity with the representative sequence of the second OTU. It is this difference in the way sequences are clustered by these methods that makes oligotyping more powerful in grouping closely related organisms. While a relationship between oligotypes and most soil chemistry parameters was observed only a few such relationships were observed for OTU data by the earlier report (Turlapati et al., 2013).

A major limitation of any study involving a soil microbiome (including the present one) is the lack of reference genomes (even for dominant taxa; Howe et al., 2014). Metagenomic outputs of most current high-throughput sequencing technologies (e.g., Illumina) often result in a mixture of multiple genomes most of which do not cover a complete genome of the organisms of interest since complete reference genomes of known organisms are lacking (Simon and Daniel, 2011; Teeling and Glöckner, 2012). Therefore, many studies still rely on universally occurring DNA sequences (either partial or complete), e.g., the 16S rRNA genes. Single cell genomics offers a powerful technique for characterizing the genome of a single organism (Zaremba-Niedzwiedzka et al., 2013; Macaulay and Voet, 2014); however, this is still an emerging technology and is difficult to apply to the microbiome of complex systems like forest soils. In the absence of genome-specific sequence libraries from forest soils, it is difficult to assign the terminal taxonomic identity (e.g., at species level) even to 16S rRNA genes or any other gene sequences. Thus, amplicon-sequencing (although known to be biased against rare organisms) remains a realistic approach to estimate the diversity of large microbial populations in complex environments.

Oligotyping enables the detection and classification of distinct subpopulations within a genus, or even within a single species

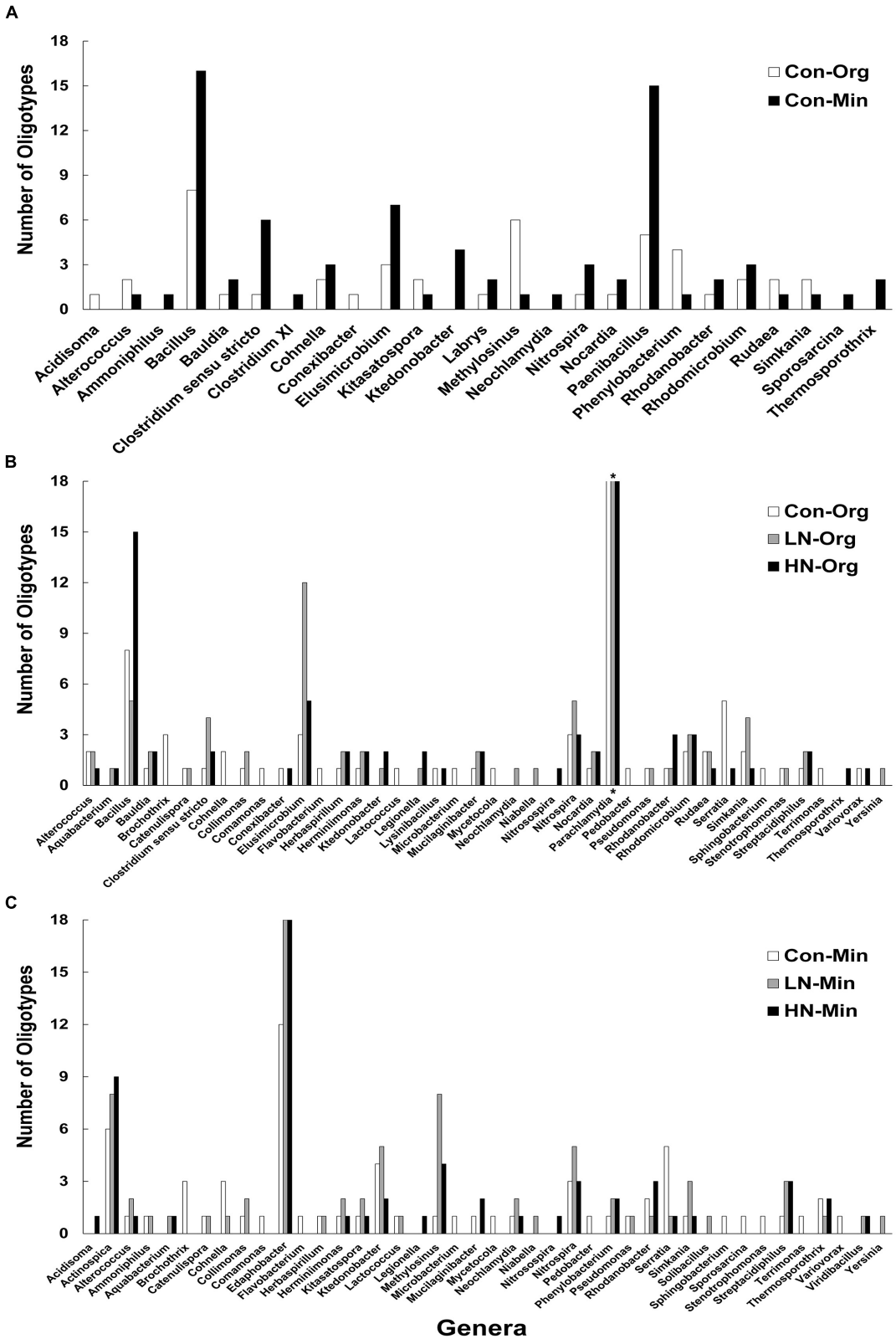


FIGURE 2 | Comparison of numbers of oligotypes for the identified genera in soil samples: (A) control organic vs. mineral soil; (B) control vs. N-amended organic soil; and (C) control vs. N-amended mineral soil. The genera presented here were selected based on ≥50% change relative to control. Asterisk in **(B)** denotes the presence of >18 oligotypes for con (24), LN (44), and HN (39) for Parachlamydia.

as was shown for *Gardnerella vaginalis* in humans (Eren et al., 2011). With forest soil microbiomes, although taxonomy at the species level could not be assigned because of the lack of reference genome data, oligotyping did enable us to detect subpopulations within a genus. Furthermore, the distribution of many of these subpopulations often varied between the soil horizons and among long-term treatments with N fertilizer (Table S6).

The results presented here demonstrate the applicability of oligotyping to complex microbiomes of forest soils. However, some adjustments may be necessary to the stringency of parameters that Eren et al. (2013) had suggested. For example, in order to assess the diversity of closely related bacterial populations in an ecosystem by oligotyping of 16S rRNA gene sequences, Eren et al. (2013) emphasized the importance of at least four critical parameters (namely 's,' 'a,' 'A' and 'M') that minimize the impact of sequencing errors in determining the outcome of results. They further summarized that s and M are critical components used to reduce the noise in such analyses. Soil samples have greater microsite variability, bacterial diversity and the occurrence of rare organisms as compared to other microbiomes (e.g., human body and marine waters). All of the data in the present study were analyzed using *M* values based on two criteria: namely, retaining maximum sequences and the diversity in terms of the number of oligotypes with an $s = 2$ (due to high microsite variability among five replicates). In the present study, high diversity was evident from a large number of entropy peaks with high values for components for most phylum level datasets (Figure S1). Therefore, oligotyping of these soil samples required several rounds of supervised (user-defined component selection) analyses before all of the entropy peaks could be decomposed. Lowering the *M* values from those suggested by Eren et al. (2013), especially for relatively larger datasets led to the identification of more genera at CT $\geq$0.8 (**Table 1**). This suggests that many organisms are probably present in low abundance (constituting the rare microbiome) in HF soil. Even with much smaller datasets, where *M* values of two or three were used, genera were identified at CT $\geq$0.8 (e.g., genus *Aquabacterium* of class *β-Proteobacteria* – **Table 1**). Huse et al. (2010) suggested that if a sequence occurs in two separate environmental samples (i.e., $s = 2$), then the chance of it being noise or a technical error is almost zero and thus should be considered as a sequence affiliated with a rare organism. Using this same dataset, Turlapati et al. (2013) earlier reported 4093 singleton OTUs among a total of 11,029 (37%) with $s = 1$ (default). Therefore, the present classification with oligotyping with $s = 2$ should be more reliable as compared to OTU clustering in eliminating noise.

The primers used in the present study specifically target the V6–V8 region of 16S rRNA genes and were chosen due to the high sequence variability associated with this region (Brons and van Elsas, 2008). The poor ability of RDP to assign taxonomy to V6 reads at CT $\geq$0.8 as compared to CT $\geq$0.5 has been reported by Claesson et al. (2009). Similarly, in this present study, a greater number of genera were identified at CT $\geq$0.5 vs. CT $\geq$0.8 (Table S5). For example, genus *Terriglobus* (*Acidobacteria, Gp1*) could only be identified at CT $\geq$0.5; sequences for this genus were found in all 30 samples and were significantly higher in HN-Org soils as compared to control. These observations suggest that the standard CT value of $\geq$0.8 at the genus level may need to be adjusted when working with the V6-V8 hypervariable regions of the 16S rRNA gene especially in ecosystems for which reference genome libraries are lacking.

Available analytical tools and public databases, such as RDP, are constantly being updated to meet increasing demand for taxonomic classification arising from high throughput outputs created by next generation sequencing platforms (Cole et al., 2009, 2013). Mclean et al. (2013) reported 31 candidate phyla including recently identified TM6 in the bacterial population of a hospital sink. In the present study, A total of 16 phyla were identified as compared to 11 in our previous report which used the same dataset in the QIIME pipeline (Turlapati et al., 2013). The RDP classifier assigned phyla names such as AD3, *Elusimicrobia, Cyanobacteria*, TM6, WPS-2 to the sequences that were termed unclassified in our previous study. No genera were identified within these phyla with the exception of *Elusimicrobium*. Most importantly, the overall unclassified sequences previously constituting 15–20% of the total were reduced to 0.5% in the current analysis.

Although *Acidobacteria* constituted >50% of total sequences, only four genera were identified in this phylum. The availability of reference genomes would be useful in further classifying this phylum; however, to date only eight genera have been taxonomically described in this phylum (Männistö et al., 2011 and references therein). Naether et al. (2012) reported that within *Acidobacteria Gp*1, *Gp*2, and *Gp*3 organisms favor nutrient-limited soils as compared to other subgroups. The dominance of these three subgroups of *Acidobacteria* in both soil horizons at HF suggests that these soils are perhaps nutrient limited. VanInsberghe et al. (2013) reported that in comparison with *Proteobacteria, Acidobacteria* are more prevalent in soils with low resource availability; our results are in agreement with this report and further reinforce the conclusion that HF soils are nutrient poor. Differences observed between bacterial communities in the Org and Min soil are clearly attributable to the differences in the soil chemistry of the two horizons (Table S3). Although HF soils may be nutrient limited, bacteria in the Org horizon are perhaps adapted to relatively nutrient-rich environment compared to those in the Min horizon. Our results demonstrate that with few exceptions, the Org soil communities were more impacted by N-treatment as compared to the Min soil communities. Compared to Min soil horizon, bacteria in the Org soils demonstrated stronger relationships with most of the soil chemistry parameters (**Table 5**). Fierer et al. (2012) also reported greater phylogenetic shifts in microbial communities that prefer a nutrient rich environment following N fertilization.

Naether et al. (2012) also reported correlations between edaphic factors such as pH, C, N, C/N ratio, and P with corresponding OTUs (16S clones) and terminal-RFLP (T-RFLP) for most of the subgroups of *Acidobacteria* found in soil from 30 forested and 27 grassland sites. They found either positive or negative correlations of different OTUs or T-RFLPs within respective subgroups of *Acidobacteria* over a wide pH and nutrient range. Another study involving 87 soil types with pH values ranging from 3.5 to 8.5 reported an overall inverse relationship between soil pH and the relative abundance of *Acidobacteria* (Jones et al., 2009). However, a closer look at data shows that within a narrow range of pH from 3.5 to 4.5, this inverse relationship is not held. The pH of

HF soil ranged from 3.8 to 4.4 for Org and 4.3 to 4.8 for Min. At HF, a positive correlation between the subgroups of *Acidobacteria* and pH in this narrow range for each soil horizon indicates that for optimal growth, this group prefers the higher end of this narrow range at HF (Table S3). Our results are in agreement with those of Sait et al. (2006) who found that subgroup *Gp*1 ideally requires moderately acidic conditions (pH 4.0–5.5).

Despite the existence of significant effects on aboveground foliar Org N metabolites (polyamines and amino acids; Table S3), tree physiology and productivity (Minocha et al., 2000; Bauer et al., 2004; Magill et al., 2004; Frey et al., 2014), and changes in soil microbial diversity at the HF, the lack of any lasting effects of N-amendment on soil NH_4 and NO_3 concentrations is interesting and apparently contradictory. We speculate that this is due to the combined effects of the fast uptake of the fertilizer (applied only during the growing season) by the macroflora, its rapid conversion into other inorganic N and Org N metabolites (e.g., polyamines and amino acids), and leaching of a significant amount of applied N.

Polyamines are present in all living organisms. They are required for growth and are also involved in stress responses (Minocha et al., 2014). Polyamines and specific-amino acids (e.g., glutamine and arginine) are known to be major N storage metabolites in plants, especially under excess N conditions. Concentrations of these compounds were found to be high in the foliage of trees growing in the N-treated plots at the HF (Minocha et al., 2000; Bauer et al., 2004). Changes in concentrations of polyamines and amino acids were observed in the same soils used for the present study (Frey et al., 2014). These findings suggest that the effects of N-addition on shifts in bacterial community structure may have resulted partially through effects of N-amendments on the growth of the aboveground plant community and vice versa. Also, it can be hypothesized that changes observed in the microbiome are due to the preference of certain microbial taxa for excess N that was present immediately following N application. Then, over the longer term of repeated N applications, they were stabilized and became a major component of the microbiome during the phase when soil inorganic N reverted back to original levels. This argument is supported by the observation that some of the functionally important genera (e.g., *Nitrosospira* of *phylum Proteobacteria*, which include well-known NH_3 oxidizers/nitrifiers) appeared mostly in N-treated soils. Using the amplification of a functional N-transformation gene *amo*A, He et al. (2007) observed an increased abundance of *Nitrosospira* sequences in response to N-treatments at a Chinese Agricultural Experimental Station. In our study the sequences and oligotypes corresponding to *Nitrospira* were higher in LN-treated Min soil. However, using 16S markers, Wertz et al. (2012) observed no change in the abundance of sequences for *Nitrospira* (another potential nitrifier group within the phylum *Nitrospira*) in response to long term N-fertilization at five forested sites in British Columbia, Canada.

CONCLUSION

A total of 46 previously unidentified genera were recognized by oligotyping vs. OTU clustering analysis of PCR-amplified partial 16S rDNA sequences from HF soil DNA. Because of the lack of a reference genome database for forest soils, both clustering approaches yield limited information at the genus and species level; however, oligotyping enables reliable classification of closely related organisms because of the high stringency of this tool. This analytical approach further revealed strong correlations between soil chemistry and oligotypes; no such correlations were discernible with the OTU clustering approach. Based on the fact that we could identify several genera at CT $\geq$0.98 using a relatively lower *M* value, we suggest that lowering *M* values may be appropriate for the complex microbiomes such as forest soils that are comprised of an enormous diversity of bacteria that are often present in low abundance. As suggested by Mantel test results, bacterial communities in the Org soil at HF have high preference for a nutrient rich environment and the communities found in the Min soil are better adapted to nutrient poor conditions. Overall, effects of N-treatment on the microbiome were more evident in the Org soil than the Min soil horizon, perhaps due to the fact that N utilization requires an abundance of C, which three times higher in the Org as compared to Min soil.

ACKNOWLEDGMENTS

The authors are grateful to Kenneth R. Dudzik for his technical assistance. They also extend their thanks to Dr. Murat Eren for providing guidance in using the oligotyping pipeline. Partial funding was provided by the New Hampshire Agricultural Experiment Station. This is Scientific Contribution Number 2568.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2015.00049/abstract

REFERENCES

Aber, J. D., Alison, M., Boone, R., Melillo, J. M., and Steudler, P. (1993). Plant and soil responses to chronic nitrogen additions at the Harvard Forest, Massachusetts. *Ecol. Appl.* 3, 156–166. doi: 10.2307/1941798

Bauer, G. A., Bazzaz, F. A., Minocha, R., Long, S., Magill, A., Aber, J., et al. (2004). Effects of chronic N additions on tissue chemistry, photosynthetic capacity, and carbon sequestration potential of a red pine (*Pinus resinosa Ait.*) stand in the NE United States. *For. Ecol. Manage.* 196, 173–186. doi: 10.1016/j.foreco.2004.03.032

Bischoff, J., Mangelsdorf, K., Schwamborn, G., and Wagner, D. (2014). Impact of lake-level and climate changes on microbial communities in a terrestrial permafrost sequence of the El'gygytgyn Crater, far East Russian Arctic. *Permafrost Periglacial Process.* 25, 107–116. doi: 10.1002/ppp.1807

Bowden, R. D., Davidson, E., Savage, K., Arabia, C., and Steudler, P. (2004). Chronic nitrogen additions reduce total soil respiration and microbial respiration in temperate forest soils at the Harvard Forest. *For. Ecol. Manage.* 196, 43–56. doi: 10.1016/j.foreco.2004.03.011

Brons, J. K., and van Elsas, J. D. (2008). Analysis of bacterial communities in soil by use of denaturing gradient gel electrophoresis and clone libraries, as influenced by different reverse primers. *Appl. Environ. Microbiol.* 74, 2717–2727. doi: 10.1128/AEM.02195-07

Caporaso, J. G., Bittinger, K., Bushman, F. D., Desantis, T. Z., Andersen, G. L., and Knight, R. (2010a). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* 26, 266–267. doi: 10.1093/bioinformatics/btp636

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010b). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Carney, K. M., and Matson, P. A. (2006). The influence of tropical plant diversity and composition on soil microbial communities. *Microb. Ecol.* 52, 226–238. doi: 10.1007/s00248-006-9115-z

Claesson, M. J., O'sullivan, O., Wang, Q., Nikkilä, J., Marchesi, J. R., Smidt, H., et al. (2009). Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PLoS ONE* 4:e6669. doi: 10.1371/journal.pone.0006669

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., et al. (2009). The ribosomal database project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* 37, D141–D145. doi: 10.1093/nar/gkn879

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., Mcgarrell, D. M., Sun, Y., et al. (2013). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 42, D633–D642. doi: 10.1093/nar/gkt1244

Compton, J. E., Watrud, L. S., Arlene Porteous, L., and Degrood, S. (2004). Response of soil microbial biomass and community composition to chronic nitrogen additions at Harvard forest. *For. Ecol. Manage.* 196, 143–158. doi: 10.1016/j.foreco.2004.03.017

Coolon, J. D., Jones, K. L., Todd, T. C., Blair, J. M., and Herman, M. A. (2013). Long-term nitrogen amendment alters the diversity and assemblage of soil bacterial communities in Tallgrass Prairie. *PLoS ONE* 8:e67884. doi: 10.1371/journal.pone.0067884

Currie, W. S., Aber, J. D., and Driscoll, C. T. (1999). Leaching of nutrient cations from the forest floor: effects of nitrogen saturation in two long-term manipulations. *Can. J. For. Res.* 29, 609–620. doi: 10.1139/x99-033

Cusack, D. F., Silver, W. L., Torn, M. S., Burton, S. D., and Firestone, M. K. (2010). Changes in microbial community characteristics and soil organic matter with nitrogen additions in two tropical forests. *Ecology* 92, 621–632. doi: 10.1890/10-0459.1

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., et al. (2013). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* 4, 1111–1119. doi: 10.1111/2041-210X.12114

Eren, A. M., Zozaya, M., Taylor, C. M., Dowd, S. E., Martin, D. H., and Ferris, M. J. (2011). Exploring the diversity of *Gardnerella vaginalis* in the genitourinary tract microbiota of monogamous couples through subtle nucleotide variation. *PLoS ONE* 6:e26732. doi: 10.1371/journal.pone.0026732

Falkowski, P. G., Fenchel, T., and Delong, E. F. (2008). The microbial engines that drive earth's biogeochemical cycles. *Science* 320, 1034–1039. doi: 10.1126/science.1153213

Fierer, N., and Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. *Proc. Natl. Acad. Sci. U.S.A.* 103, 626–631. doi: 10.1073/pnas.0507535103

Fierer, N., Lauber, C. L., Ramirez, K. S., Zaneveld, J., Bradford, M. A., and Knight, R. (2012). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *ISME J.* 6, 1007–1017. doi: 10.1038/ismej.2011.159

Frey, S. D., Knorr, M., Parrent, J. L., and Simpson, R. T. (2004). Chronic nitrogen enrichment affects the structure and function of the soil microbial community in temperate hardwood and pine forests. *For. Ecol. Manage.* 196, 159–171. doi: 10.1016/j.foreco.2004.03.018

Frey, S. D., Ollinger, S., Nadelhoffer, K., Bowden, R., Brzostek, E., Burton, A., et al. (2014). Chronic nitrogen additions suppress decomposition and sequester soil carbon in temperate forests. *Biogeochemistry* 121, 305–316. doi: 10.1007/s10533-10014-10004-10530

Fulthorpe, R. R., Roesch, L. F. W., Riva, A., and Triplett, E. W. (2008). Distantly sampled soils carry few species in common. *ISME J.* 2, 901–910. doi: 10.1038/ismej.2008.55

Gloor, G. B., Hummelen, R., Macklaim, J. M., Dickson, R. J., Fernandes, A. D., Macphee, R., et al. (2010). Microbiome profiling by illumina sequencing of combinatorial sequence-tagged PCR products. *PLoS ONE* 5:e15406. doi: 10.1371/journal.pone.0015406

Gruber, N., and Galloway, J. N. (2008). An earth-system perspective of the global nitrogen cycle. *Nature* 451, 293–296. doi: 10.1038/nature06592

Hallin, S., Jones, C. M., Schloter, M., and Philippot, L. (2009). Relationship between N-cycling communities and ecosystem functioning in a 50-year-old fertilization experiment. *ISME J.* 3, 597–605. doi: 10.1038/ismej.2008.128

He, J.-Z., Shen, J.-P., Zhang, L.-M., Zhu, Y.-G., Zheng, Y.-M., Xu, M.-G., et al. (2007). Quantitative analyses of the abundance and composition of ammonia-oxidizing bacteria and ammonia-oxidizing archaea of a Chinese upland red soil under long-term fertilization practices. *Environ. Microbiol.* 9, 2364–2374. doi: 10.1111/j.1462-2920.2007.01358.x

Howe, A. C., Jansson, J. K., Malfatti, S. A., Tringe, S. G., Tiedje, J. M., and Brown, C. T. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl. Acad. Sci. U.S.A.* 111, 4904–4909. doi: 10.1073/pnas.1402564111

Huse, S. M., Dethlefsen, L., Huber, J. A., Welch, D. M., Relman, D. A., and Sogin, M. L. (2009). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet.* 4:e1000255. doi: 10.1371/journal.pgen.1000255

Huse, S. M., Welch, D. M., Morrison, H. G., and Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.* 12, 1889–1898. doi: 10.1111/j.1462-2920.2010.02193.x

Janssen, P. H. (2006). Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Appl. Environ. Microbiol.* 72, 1719–1728. doi: 10.1128/AEM.72.3.1719-1728.2006

Janssens, I. A., Dieleman, W., Luyssaert, S., Subke, J. A., Reichstein, M., Ceulemans, R., et al. (2010). Reduction of forest soil respiration in response to nitrogen deposition. *Nat. Geosci.* 3, 315–322. doi: 10.1038/ngeo844

Jones, R. T., Robeson, M. S., Lauber, C. L., Hamady, M., Knight, R., and Fierer, N. (2009). A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. *ISME J.* 3, 442–453. doi: 10.1038/ismej.2008.127

Langenfeld, A., Prado, S., Nay, B., Cruaud, C., Lacoste, S., Bury, E., et al. (2013). Geographic locality greatly influences fungal endophyte communities in *Cephalotaxus harringtonia*. *Fungal Biol.* 117, 124–136. doi: 10.1016/j.funbio.2012.12.005

Lauber, C. L., Hamady, M., Knight, R., and Fierer, N. (2009). Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl. Environ. Microbiol.* 75, 5111–5120. doi: 10.1128/AEM.00335-09

Lauber, C. L., Strickland, M. S., Bradford, M. A., and Fierer, N. (2008). The influence of soil properties on the structure of bacterial and fungal communities across land-use types. *Soil Biol. Biochem.* 40, 2407–2415. doi: 10.1016/j.soilbio.2008.05.021

Lombard, N., Prestat, E., Van Elsas, J. D., and Simonet, P. (2011). Soil-specific limitations for access and analysis of soil microbial communities by metagenomics. *FEMS Microbiol. Ecol.* 78, 31–49. doi: 10.1111/j.1574-6941.2011.01140.x

Long, R. P., Horsley, S. B., and Lilja, P. R. (1997). Impact of forest liming on growth and crown vigor of sugar maple and associated hardwoods. *Can. J. For. Res.* 27, 1560–1573. doi: 10.1139/x97-074

Lynch, M. D. J., Bartram, A. K., and Neufeld, J. D. (2012). Targeted recovery of novel phylogenetic diversity from next-generation sequence data. *ISME J.* 6, 2067–2077. doi: 10.1038/ismej.2012.50

Macaulay, I. C., and Voet, T. (2014). Single cell genomics: advances and future perspectives. *PLoS Genet.* 10:e1004126. doi: 10.1371/journal.pgen.1004126

Magill, A. H., Aber, J. D., Currie, W. S., Nadelhoffer, K. J., Martin, M. E., Mcdowell, W. H., et al. (2004). Ecosystem response to 15 years of chronic nitrogen additions at the Harvard Forest LTER, Massachusetts, USA. *For. Ecol. Manage.* 196, 7–28. doi: 10.1016/j.foreco.2004.03.033

Männistö, M. K., Rawat, S., Starovoytov, V., and Häggblom, M. M. (2011). Terriglobus saanensis sp. nov., an *acidobacterium* isolated from tundra soil. *Int. J. Syst. Evol. Microbiol.* 61, 1823–1828. doi: 10.1099/ijs.0.026005-0

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380. doi: 10.1038/nature03959

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., Desantis, T. Z., Probst, A., et al. (2012). An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* 6, 610–618. doi: 10.1038/ismej.2011.139

McGuire, K., Fierer, N., Bateman, C., Treseder, K., and Turner, B. (2012). Fungal community composition in neotropical rain forests: the influence of tree diversity and precipitation. *Microb. Ecol.* 63, 804–812. doi: 10.1007/s00248-011-9973-x

Mclean, J. S., Lombardo, M.-J., Badger, J. H., Edlund, A., Novotny, M., Yee-Greenbaum, J., et al. (2013). Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *Proc. Natl. Acad. Sci. U.S.A.* 110, E2390–E2399. doi: 10.1073/pnas.1219809110

Minocha, R., Long, S. L., Magill, A., Aber, J., and Mcdowell, W. (2000). Foliar free polyamine and inorganic ion content in relation to soil and soil solution chemistry in two fertilized forest stands at the Harvard Forest, Massachusetts. *Plant Soil* 222, 119–137. doi: 10.1023/A:1004775829678

Minocha, R., Long, S., Thangavel, P., Minocha, S. C., Eagar, C., and Driscoll, C. T. (2010). Elevation dependent sensitivity of northern hardwoods to Ca addition at Hubbard Brook experimental forest, NH USA. *For. Ecol. Manage.* 260, 2115–2125. doi: 10.1016/j.foreco.2010.09.002

Minocha, R., Majumdar, R., and Minocha, S. C. (2014). Polyamines and abiotic stress in plants: A complex relationship. *Front. Plant Sci.* 5:175. doi: 10.3389/fpls.2014.00175

Naether, A., Foesel, B. U., Naegele, V., Wust, P. K., Weinert, J., Bonkowski, M., et al. (2012). Environmental factors affect Acidobacterial communities below the subgroup level in grassland and forest soils. *Appl. Environ. Microbiol.* 78, 7398–7406. doi: 10.1128/AEM.01325-12

Nunes da Rocha, U., Van Overbeek, L., and Van Elsas, J. D. (2009). Exploration of hitherto-uncultured bacteria from the rhizosphere. *FEMS Microbiol. Ecol.* 69, 313–328. doi: 10.1111/j.1574-6941.2009.00702.x

Ramirez, K. S., Craine, J. M., and Fierer, N. (2012). Consistent effects of nitrogen amendments on soil microbial communities and processes across biomes. *Glob. Chang. Biol.* 18, 1918–1927. doi: 10.1111/j.1365-2486.2012.02639.x

Ramirez, K. S., Lauber, C. L., Knight, R., Bradford, M. A., and Fierer, N. (2010). Consistent effects of nitrogen fertilization on soil bacterial communities in contrasting systems. *Ecology* 91, 3463–3470. doi: 10.1890/10-0426.1

Roesch, L. F. W., Fulthorpe, R. R., Riva, A., Casella, G., Hadwin, A. K. M., Kent, A. D., et al. (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J.* 1, 283–290.

Rousk, J., Baath, E., Brookes, P. C., Lauber, C. L., Lozupone, C., Caporaso, J. G., et al. (2010). Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME J.* 4, 1340–1351. doi: 10.1038/ismej.2010.58

Sait, M., Davis, K. E., and Janssen, P. H. (2006). Effect of pH on isolation and distribution of members of subdivision 1 of the phylum *Acidobacteria* occurring in soil. *Appl. Environ. Microbiol.* 72, 1852–1857. doi: 10.1128/AEM.72.3.1852-1857.2006

Sait, M., Hugenholtz, P., and Janssen, P. H. (2002). Cultivation of globally distributed soil bacteria from phylogenetic lineages previously only detected in cultivation-independent surveys. *Environ. Microbiol.* 4, 654–666. doi: 10.1046/j.1462-2920.2002.00352.x

Schaberg, P., Minocha, R., Long, S., Halman, J., Hawley, G., and Eagar, C. (2011). Calcium addition at the Hubbard Brook experimental forest increases the capacity for stress tolerance and carbon capture in red spruce (*Picea rubens*) trees during the cold season. *Trees* 25, 1053–1061. doi: 10.1007/s00468-011-0580-8

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Shortle, W. C., and Smith, K. T. (1988). Aluminum-induced calcium deficiency syndrome in declining red spruce. *Science* 240, 1017–1018. doi: 10.1126/science.240.4855.1017

Simon, C., and Daniel, R. (2011). Metagenomic analyses: past and future trends. *Appl. Environ. Microbiol.* 77, 1153–1161. doi: 10.1128/AEM.02345-10

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., et al. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. U.S.A.* 103, 12115–12120. doi: 10.1073/pnas.0605127103

Sridevi, G., Minocha, R., Turlapati, S. A., Goldfarb, K. C., Brodie, E. L., Tisa, L. S., et al. (2012). Soil bacterial communities of a calcium-supplemented and a reference watershed at the Hubbard Brook Experimental Forest (HBEF), New Hampshire, USA. *FEMS Microbiol. Ecol.* 79, 728–740. doi: 10.1111/j.1574-6941.2011.01258.x

Teeling, H., and Glöckner, F. O. (2012). Current opportunities and challenges in microbial metagenome analysis—a bioinformatic perspective. *Brief. Bioinform.* 13, 728–742. doi: 10.1093/bib/bbs039

Torsvik, V., Øvreås, L., and Thingstad, T. F. (2002). Prokaryotic diversity–magnitude, dynamics, and controlling factors. *Science* 296, 1064–1066. doi: 10.1126/science.1071698

Treseder, K. K. (2008). Nitrogen additions and microbial biomass: a meta-analysis of ecosystem studies. *Ecol. Lett.* 11, 1111–1120. doi: 10.1111/j.1461-0248.2008.01230.x

Trevors, J. T. (2010). One gram of soil: a microbial biochemical gene library. *Antonie Van Leeuwenhoek* 97, 99–106. doi: 10.1007/s10482-009-9397-5

Turlapati, S. A., Minocha, R., Bhiravarasa, P. S., Tisa, L. S., Thomas, W. K., and Minocha, S. C. (2013). Chronic N-amended soils exhibit an altered bacterial community structure in Harvard Forest, MA, USA. *FEMS Microbiol. Ecol.* 83, 478–493. doi: 10.1111/1574-6941.12009

Uroz, S., Buée, M., Murat, C., Frey-Klett, P., and Martin, F. (2010). Pyrosequencing reveals a contrasted bacterial diversity between oak rhizosphere and surrounding soil. *Environ. Microbiol. Rep.* 2, 281–288. doi: 10.1111/j.1758-2229.2009.00117.x

VanInsberghe, D., Hartmann, M., Stewart, G. R., and Mohn, W. W. (2013). Isolation of a substantial proportion of forest soil bacterial communities detected via pyrotag sequencing. *Appl. Environ. Microbiol.* 79, 2096–2098. doi: 10.1128/AEM.03112-12

Vartoukian, S. R., Palmer, R. M., and Wade, W. G. (2010). Strategies for culture of 'unculturable' bacteria. *FEMS Microbiol. Lett.* 309, 1–7. doi: 10.1111/j.1574-6968.2010.02000.x.

Větrovský, T., and Baldrian, P. (2013). The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS ONE* 8:e57923. doi: 10.1371/journal.pone.0057923

Wagg, C., Bender, S. F., Widmer, F., and Van Der Heijden, M. G. A. (2014). Soil biodiversity and soil community composition determine ecosystem multifunctionality. *Proc. Natl. Acad. Sci. U.S.A.* 111, 5266–5270. doi: 10.1073/pnas.1320054111

Wall, D. H., Bardgett, R. D., and Kelly, E. (2010). Biodiversity in the dark. *Nat. Geosci.* 3, 297–298. doi: 10.1038/ngeo860

Wallenstein, M. D., Mcnulty, S., Fernandez, I. J., Boggs, J., and Schlesinger, W. H. (2006). Nitrogen fertilization decreases forest soil fungal and bacterial biomass in three long-term experiments. *For. Ecol. Manage.* 222, 459–468. doi: 10.1016/j.foreco.2005.11.002

Wertz, S., Leigh, A. K., and Grayston, S. J. (2012). Effects of long-term fertilization of forest soils on potential nitrification and on the abundance and community structure of ammonia oxidizers and nitrite oxidizers. *FEMS Microbiol. Ecol.* 79, 142–154. doi: 10.1111/j.1574-6941.2011.01204.x

Wu, M., Qin, H., Chen, Z., Wu, J., and Wei, W. (2011). Effect of long-term fertilization on bacterial composition in rice paddy soil. *Biol. Fertil. Soils* 47, 397–405. doi: 10.1007/s00374-010-0535-z

Wu, T., Chellemi, D. O., Graham, J. H., Martin, K. J., and Rosskopf, E. N. (2008). Comparison of soil bacterial communities under diverse agricultural land management and crop production practices. *Microb. Ecol.* 55, 293–310. doi: 10.1007/s00248-007-9276-4

Zaremba-Niedzwiedzka, K., Viklund, J., Zhao, W., Ast, J., Sczyrba, A., Woyke, T., et al. (2013). Single-cell genomics reveal low recombination frequencies in freshwater bacteria of the SAR11 clade. *Genome Biol.* 14, R130. doi: 10.1186/gb-2013-14-11-r130

Zhao, J., Wan, S., Fu, S., Wang, X., Wang, M., Liang, C., et al. (2013). Effects of understory removal and nitrogen fertilization on soil microbial communities in *Eucalyptus plantations*. *For. Ecol. Manage.* 310, 80–86. doi: 10.1016/j.foreco.2013.08.013

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 21 August 2014; accepted: 14 January 2015; published online: 16 February 2015.

Citation: Turlapati SA, Minocha R, Long S, Ramsdell J and Minocha SC (2015) Oligotyping reveals stronger relationship of organic soil bacterial community structure with N-amendments and soil chemistry in comparison to that of mineral soil at Harvard Forest, MA, USA. Front. Microbiol. ***6****:49. doi: 10.3389/fmicb.2015.00049*

This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology.

frontiers in MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 14 October 2014
doi: 10.3389/fmicb.2014.00526

Oligotyping reveals differences between gut microbiomes of free-ranging sympatric Namibian carnivores (*Acinonyx jubatus, Canis mesomelas*) on a bacterial species-like level

Sebastian Menke[1,2], Wasimuddin[3], Matthias Meier[1], Jörg Melzheimer[2], John K. E. Mfune[4], Sonja Heinrich[2], Susanne Thalwitzer[2], Bettina Wachter[2] and Simone Sommer[1,5]*

[1] Evolutionary Genetics, Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany
[2] Evolutionary Ecology, Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany
[3] Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, Brno, Czech Republic
[4] Department of Biological Sciences, University of Namibia, Windhoek, Namibia
[5] Institute of Experimental Ecology, University of Ulm, Ulm, Germany

Edited by:
Lois Maignien, University of Western Brittany, France

Reviewed by:
Spyridon Ntougias, Democritus University of Thrace, Greece
Thulani Peter Makhalanyane, University of Pretoria, South Africa
A. Murat Eren, Marine Biological Laboratory, USA

***Correspondence:**
Simone Sommer, Institute of Experimental Ecology, University of Ulm, Albert-Einstein Allee 11, Ulm 89069, Germany
e-mail: simone.sommer@uni-ulm.de

Recent gut microbiome studies in model organisms emphasize the effects of intrinsic and extrinsic factors on the variation of the bacterial composition and its impact on the overall health status of the host. Species occurring in the same habitat might share a similar microbiome, especially if they overlap in ecological and behavioral traits. So far, the natural variation in microbiomes of free-ranging wildlife species has not been thoroughly investigated. The few existing studies exploring microbiomes through 16S rRNA gene reads clustered sequencing reads into operational taxonomic units (OTUs) based on a similarity threshold (e.g., 97%). This approach, in combination with the low resolution of target databases, generally limits the level of taxonomic assignments to the genus level. However, distinguishing natural variation of microbiomes in healthy individuals from "abnormal" microbial compositions that affect host health requires knowledge of the "normal" microbial flora at a high taxonomic resolution. This gap can now be addressed using the recently published oligotyping approach, which can resolve closely related organisms into distinct oligotypes by utilizing subtle nucleotide variation. Here, we used Illumina MiSeq to sequence amplicons generated from the V4 region of the 16S rRNA gene to investigate the gut microbiome of two free-ranging sympatric Namibian carnivore species, the cheetah (*Acinonyx jubatus*) and the black-backed jackal (*Canis mesomelas*). Bacterial phyla with proportions >0.2% were identical for both species and included Firmicutes, Fusobacteria, Bacteroidetes, Proteobacteria and Actinobacteria. At a finer taxonomic resolution, black-backed jackals exhibited 69 bacterial taxa with proportions ≥0.1%, whereas cheetahs had only 42. Finally, oligotyping revealed that shared bacterial taxa consisted of distinct oligotype profiles. Thus, in contrast to 3% OTUs, oligotyping can detect fine-scale taxonomic differences between microbiomes.

Keywords: gut microbiome, bacteria, oligotyping, carnivores, cheetah (*Acinonyx jubatus*), black-backed jackal (*Canis mesomelas*), Namibia

INTRODUCTION

Gut-associated bacterial communities and their mammalian hosts are highly dependent on each other. Recent investigations have applied metagenomic approaches to increase our understanding of the factors that shape these host-gut bacterial relationships (Kau et al., 2011; Muegge et al., 2011; Schloissnig et al., 2012). The interpretation of these results is, however, challenging because several factors such as host-bacteria co-evolution (Ley et al., 2008; Ochman et al., 2010; Yeoman et al., 2011), host genotype (Benson et al., 2010; Spor et al., 2011; Bolnick et al., 2014), life history traits and behavior (Ezenwa et al., 2012), social organization (Koch and Schmid-Hempel, 2011), health status, diet (Turnbaugh et al., 2009) and the environment itself (Coolon et al., 2010; Nelson et al., 2013) are simultaneously involved in shaping the gut microbiome. In contrast, variations in the gut microbiome affect the host by causing, for example, obesity (Turnbaugh et al., 2006) and changes in exploratory behavior or anxiety (Bercik et al., 2011; Bravo et al., 2011), all of which may affect the overall health status of the host (Sekirov et al., 2010; Hooper et al., 2012).

Previous studies on gut microbiomes have focused largely on the variation exhibited in humans or laboratory organisms. Only recently the interest has grown to study host-gut bacterial associations also in wildlife species (Schwab et al., 2011; Amato et al., 2013; Nelson et al., 2013; Delsuc et al., 2014). Such studies are facing many challenges, but offer, when sample sizes are large, important insight in the "normal" variation of the gut microbiome of free-ranging species. Only under natural conditions we can also detect how changes in the mentioned factors

affect bacterial communities and thus host nutrition and health (McKenna et al., 2008; Amato, 2013). In addition, such data may provide reference information on habitat quality which has many implications for species conservation.

Here, we present for the first time a comparison of the gut microbiomes of two sympatric free-ranging mammalian carnivorous species, represented by a felid (cheetah, *Acinonyx jubatus*) and a canid (black-backed jackal, *Canis mesomelas*). We investigated their gut bacterial diversity by applying high-throughput sequencing to characterize the V4 hyper-variable region of the 16S rRNA gene. Differences in bacterial composition between feline and canine species have been shown previously in domestic animals (Handl et al., 2011), but microbiomes in cheetahs have only been investigated in a few zoo individuals (Ley et al., 2008; Becker et al., 2014) and so far no study has investigated microbiomes in black-backed jackals. Thus, with our study we aim to provide the microbiomes of free-ranging cheetahs and black-backed jackals and to investigate hypotheses on bacterial diversity derived from known species characteristics.

We hypothesized that diet, foraging behavior, social system and home range size, four species characteristics that differ between the species, should also lead to differences in their gut-microbial diversities. The carnivorous diet of cheetahs (Eaton, 1974; Wachter et al., 2006) is likely to be associated with a lower microbial diversity than the omnivorous diet of black-backed jackals (Goldenberg et al., 2010), because the digestive requirements for a carnivorous species can be expected to be lower (Ley et al., 2008). Accordingly, cheetahs should harbor a lower gut-microbial community compared to black-backed jackals. Moreover, cheetahs feed only on freshly killed prey animals (Caro, 1994), not on carcasses as black-backed jackals occasionally do (Walton and Joly, 2003). An intake of a more diverse bacterial community due to scavenging can be expected and accordingly a lower microbial diversity in cheetahs than in black-backed jackals. Also, the intraspecific contact rate of mainly solitary cheetahs (Caro, 1994) is likely to be lower than the one of the group living black-backed jackals (Walton and Joly, 2003), resulting in a lower bacterial transmission and therefore an expected lower microbial diversity in cheetahs than in black-backed jackals. In contrast, the larger home range sizes of cheetahs (Marker et al., 2008) compared to black-backed jackals (Jenner et al., 2011; Kamler et al., 2012) are likely to result in cheetahs encountering a larger variety of environmental bacteria than black-backed jackals and therefore are expected to exhibit a higher microbial diversity. If the home range size has the stronger influence in shaping the microbiome of a host species, we expect the cheetah to exhibit a higher microbial diversity, but if diet, foraging behavior and social system have the stronger influence, we expect the black-backed jackal to exhibit a higher microbial diversity.

The taxonomic resolution of bacterial communities based on high-throughput sequencing of 16S rRNA gene amplicons is limited. The short fragment sizes, the single locus approach and the limitations in resolution and richness of most current databases hinders a taxonomic assignment of sequencing reads better than family or genus level. *De novo* clustering of reads into operational taxonomic units (OTUs) based on a similarity threshold (e.g., 97%, Caporaso et al., 2012) aims to increase the resolution. Nevertheless, multiple 3% OTUs can be assigned to a single genus and thus still contain unexplained diversity. Differences in bacterial communities between species, however, are manifested on a bacterial species or strain level (Suchodolski, 2011) and are therefore not detectable with the conventional OTU approach. Recently, this problem was addressed by Eren et al. (2013) who developed an "oligotyping" approach which reveals differences between bacterial communities on a low level of taxonomic discrimination by targeting subtle nucleotide variation. First publications that successfully applied oligotyping showed the potential of this method by having tracked human fecal *Lachnospiraceae* in sewage (McLellan et al., 2013) or having described the diversity of a single bacterial species in the genitourinary tract of monogamous sexual partners (Eren et al., 2011). If a comprehensive database is available, it is even possible to assign species level taxonomy to oligotypes (Eren et al., 2014).

In this study, we aim to investigate whether the diversity and proportion of bacterial taxa differ between the cheetah and the black-backed jackal and whether oligotype profiles within shared bacterial taxa are the same or not. To our knowledge, this is the first study that applies the new oligotyping approach in conjunction with the common OTU approach to describe the gut microbiome of sympatric carnivores using high-throughput sequencing of the 16S rRNA gene. This study contributes to our understanding on the extent to which host characteristics contribute to the variability of bacterial communities, from the bacterial phylum level down to a bacterial species-like level in oligotype profiles of shared bacterial taxa.

MATERIALS AND METHODS

SAMPLE COLLECTION AND DNA EXTRACTION

We used fecal samples collected in central Namibia by the cheetah research project (CRP) and the black-backed jackal project (BBJP) of the Leibniz Institute for Zoo and Wildlife Research (IZW) in Berlin, Germany. Amplification of bacterial DNA was possible in 68 samples of clinically healthy free-ranging cheetahs and 50 samples of clinically healthy free-ranging black-backed jackals. Fecal samples from cheetahs were collected from the rectum of immobilized animals during health monitoring, whereas fecal samples from black-backed jackals were collected from the rectum of individuals that were dissected after being shot by local farmers or hunters as problem animals. The CRP and the BBJP hold research permits from the Namibian Ministry of Environment and Tourism (MET) and all work has been carried out in accordance to the relevant regulatory standards. Samples were kept cool in a car freezer for transport to the research station, deep frozen in liquid nitrogen and transported to the IZW, where they were stored at −80°C in deep freezers.

We applied a combined approach of mechanic disruption and enzymatic lysis of bacterial cells. Approximately 200 mg of thawed feces were filled into a 2 ml lysis tube (Precellys SK-38) to which 1.4 ml buffer ASL (QIAamp Mini Stool Kit) was added. A precellys homogenizer was used to homogenize individual samples (2 × 5200 rpm for 20 s with 10 s pause). After the centrifugation of the fecal suspension, 1.2 ml of the supernatant was used to proceed with the isolation as recommended by the QIAamp Mini Stool Kit protocol (Qiagen, Hilden, Germany). This kit contains an Inhibitex tablet that absorbs PCR inhibitors which often cause problems when amplifying DNA from fecal isolates. All handling

material (sterile scraper and plate, gloves etc.) was exchanged after each single preparation and the workbench was sterilized before the next extraction.

16S rDNA LIBRARY PREPARATION AND SEQUENCING

16S rDNA libraries for cheetahs and black-backed jackals were prepared independently but following the same protocol. We used the approach and the chemistry of Fluidigm (Access Array™ System for Illumina Sequencing Systems, ©Fluidigm Corporation) in which PCR and barcoding occur simultaneously. The primers 515F (5′-GTGCCAGCMGCCGCGGTAA-3′) and 806R (5′-GGACTACHVGGGTWTCTAAT-3′) which target a 291 bp-fragment of the hypervariable V4 region of the 16S rRNA gene were used for amplification (Caporaso et al., 2012; Kuczynski et al., 2012). These primers had to be modified according to the Fluidigm protocol and thus were tagged with sequences (CS1 forward tag and CS2 reverse tag) which were complementary to the respective forward or reverse access array barcode primers for Illumina. Final concentrations for the 10 μl target specific 4-primer amplicon tagging reaction were 10 ng/μl DNA, 1X FastStart PCR grade nucleotide mix buffer without $MgCl_2$ (Roche), 4.5 mM $MgCl_2$ (Roche), 200 μM of each PCR grade nucleotide (Roche), 0.05 U/μl FastStart high fidelity enzyme blend (Roche), 1X access array loading reagent (Fluidigm), 400 nM access array barcode primers for Illumina (Fluidigm), 5% DMSO (Roche), 2.4% PCR certified water and 50 nM target specific primers (TS-515F and TS-806R). In a standard PCR machine we ran the samples as described in the manufacture's protocol (Access Array®, Fluidigm 2012, San Francisco, USA). All individually barcoded samples were subsequently purified using SPRI Based Size Selection (Beckman Coulter Genomics, Brea, CA) with a 1:1 ratio of amplicons to beads and quantified with the Quant-iT™ PicoGreen® kit (Invitrogen/Life Technologies, Green Islands, NY). We then pooled all samples with an equal amount of 15 ng of DNA and diluted the pool down to 8 nM in hybridization buffer. Finally, the libraries were sequenced in two different paired-end runs on Illumina® MiSeq.

BIOINFORMATICS

We applied the same basic bioinfomatic pipeline to all demultiplexed reads from 68 cheetahs and 50 black-backed jackals. Initially, paired-end reads were merged using FLASH (Magoè and Salzberg, 2011) and primers were cut with the software cutadapt (Martin, 2011). Then, we performed a quality filtering (Q30) and converted fastq-files to fasta-files using the FASTX-Toolkit (FASTX-Toolkit)[1]. Subsequently, all individual fasta-files were merged into a single file and used as a starting point for downstream analyses in the "Quantitative Insights Into Microbial Ecology" (QIIME) software package (Caporaso et al., 2010b). Reads were checked for chimera using the UCHIME algorithm implemented in USEARCH 6.1. Afterwards, reads were pre-clustered at 60% identity against the reference data base using PyNast (Caporaso et al., 2010a). Any reads that failed to hit were discarded. For designation of OTUs, we followed the generally accepted similarity threshold of 97% (Muegge et al., 2011; Caporaso et al., 2012; Bermingham et al., 2013) and applied an open-reference OTU-picking approach using the USEARCH algorithm (Edgar, 2010; Edgar et al., 2011). Thus, besides OTUs which consisted of reads that were clustered against the Greengenes database (version 13.5, http://greengenes.lbl.gov), the remaining reads were clustered into OTUs *de novo* because they did not hit the reference sequence collection. Subsequently, singletons were removed and taxonomy was assigned using the ribosomal database project (RDP) classifier with a minimum confidence to record assignment set at 0.8 (Wang et al., 2007). Finally, reads were cleaned of any non-bacterial ribosomal reads. Alpha diversity for cheetahs and black-backed jackals was calculated on sub-samples of 8000 reads per individual to eliminate the differences in sequencing effort between species and individuals. We calculated (1) the OTU abundance, (2) the Shannon index, which is widely used to calculate diversity based on the number of different data categories and their respective abundance in a data set (Shannon and Weaver, 1949; Spellerberg and Fedor, 2003), and (3) phylogenetic diversity (PD), which is the sum of the branch lengths for all taxa that are part of a given sample (Faith, 1992). We compared alpha diversity measures between cheetahs and black-backed jackals using the Wilcoxon rank sum test. To estimate to which extent the total alpha diversity of an individual was sampled, we plotted the accumulation of Shannon index and PD against sampling effort (number of reads) for each individual (Supplementary Figure 1). Because some bacterial taxa were only present in cheetahs and others only in black-backed jackals we tested whether the proportions of taxa differed significantly between the two species using a Kolmogorov-Smirnov (K-S) test with 1000 bootstraps to calculate the *p*-value ["ks.boot" function in R package "Matching" (Sekhon, 2011)]. The K-S test only compares the similarity in sample diversity but does not account for community composition. Therefore, we calculated beta diversity on a subset (8000 reads) of each cheetah and black-backed jackal microbiome using the unweighted UniFrac metric (Lozupone and Knight, 2005; Lozupone et al., 2011) and applied a PERMANOVA approach ("adonis" in R package "vegan"). We tested the significance of the differences in community composition with a permutation test with 1000 permutations. In addition, we calculated the mean Bray-Curtis distance in cheetahs and black-backed jackals separately (Bray and Curtis, 1957).

We applied oligotyping on all reads which were assigned to a bacterial taxon that was shared between cheetahs and black-backed jackals (*Bacteroides, Blautia, Clostridium, Collinsella, Dorea, Enterococcus*, [*Eubacterium*], *Lactobacillus*, *Megamonas*, *Parabacteroides*, *Peptococcus*, *Peptostreptococcus*, *Phascolarctobacterium*, [*Prevotella*], *Ruminococcus*, [*Ruminococcus*], *Slackia*, *SMB53*, *Streptococcus*, *Sutterella*). Because most reads assigned to *Enterobacteriaceae* and *Fusobacteriaceae* were not resolved any better, they were extracted from the family level. To compare oligotype profiles between cheetahs and black-backed jackals, we pooled reads of shared bacterial taxa according to host species for further analyses. Taxa in squared brackets such as [*Ruminococcus*] are recommended groupings by Greengenes database managers based on whole genome phylogeny. However, they are not

[1]Available online at: http://hannonlab.cshl.edu/fastx_toolkit/index.html by Hannon Lab.

officially recognized groupings according to Bergey's manual of determinative bacteriology (Bergey et al., 1975) based on physiochemical and morphological traits. Reads were aligned against the Greengenes "core_set_aligned.fasta.imputed" alignment using PyNAST (Caporaso et al., 2010a). We stripped common gaps from each alignment and rarefied both samples (cheetah and black-backed jackal) within each alignment depending on the maximum shared sequence abundance. Subsequently, we conducted the entropy analysis which is based on the Shannon entropy for each base position in the alignment. Oligotyping (version 0.96; http://oligotyping.org) was performed using as many highly variable base positions as necessary to resolve all oligotypes in a bacterial taxon. If a taxon had 200,000–50,000 reads per sample (cheetah and black-backed jackal), an oligotype was considered true if it occurred in more than 1% of all reads ($a = 1$), if the most abundant unique sequence occurred at a minimum of 50 reads ($M = 50$) and if the minimum actual abundance of an oligotype in both samples was more than 500 reads ($A = 500$). For taxa that had less reads per sample (50,000-30,000; 30,000-10,000; 10,000-750) we downsized the M-value (25; 10; 5) and the A-value (250; 50; 10), respectively. In addition, we applied oligotyping also on the individual level to all bacterial taxa for which oligotype profiles were ≥75% dissimilar in the between-species comparison. We excluded *Lactobacillus*, *Megamonas*, and *Parabacteroides* from this analysis due to the limited number of host individuals which contributed to this bacterial taxon. Because sequencing depth differed between host species and proportions of bacterial taxa differed naturally between individuals, parameters for oligotyping on the individual level differed from the between-species comparison. In order to treat all samples equally, we only applied the minimum percent abundance parameter ($a = 5\%$, $A = 0$, $M = 0$).

We tested whether the number of oligotypes differed significantly between cheetahs and black-backed jackals by comparing the numbers of oligotypes found for each species and for each taxon against each other, again using the ks.boot function (R-package "Matching"; 1000 bootstraps). Furthermore, we measured the dissimilarity in oligotype profiles as the average proportion of reads that differed between cheetahs and black-backed jackals across oligotypes for each bacterial taxon. We then tested whether these average proportions differed between cheetahs and black-backed jackals using a permutation test where we compared observed average proportions to those obtained by randomly assigning each sequence to either one or the other carnivore species with the same probability (0.5). For visualization of heatmap, alpha diversity measures and oligotype barplots we used the packages "phyloseq" (McMurdie and Holmes, 2013) and "ggPlot2" (Wickham, 2009) in R. The principal coordinate analysis (PCoA) plot was produced in QIIME (Caporaso et al., 2010b). All statistical analyses were conducted in R 3.0.2 (R Core Team, 2013). Sequencing data is deposited at the Sequence Read Archive (SRA) under the accession number SRP044660.

RESULTS

Initially, our next generation sequencing approach of the hypervariable V4 region of the 16S rRNA gene provided 6,117,462 reads for 68 cheetahs and 1,757,276 for 50 black-backed jackals. The number of reads was higher in cheetahs because black-backed jackals were sequenced together with other projects in one Illumina MiSeq run which decreased the number of reads per sample. To account for this bias in number of reads between the two species, all diversity estimates were calculated based on a subsampling of 8000 reads per individual. After all initial quality filtering steps were applied to prepare reads for further analyses in QIIME, we proceeded with 5,339,319 reads (87.3%) for cheetahs and 1,344,632 reads (76.5%) for black-backed jackals with an average read length of 252 bp. The open-reference OTU picking resulted in 4033 OTUs which consist of reads that were clustered against the Greengenes database. In addition, 16,271 OTUs were picked *de novo* because associated reads did not hit the reference sequence collection. Rarefaction analyses based on the Shannon index and PD revealed that the sequencing effort was sufficient to describe and compare bacterial communities within and between the two species (Supplementary Figure 1).

Some reads could not be assigned to a phylum and thus remained on the kingdom level of bacteria (0.5% reads of cheetahs and 0.3% reads of black-backed jackals). Basically, cheetahs and black-backed jackals had the same most abundant (>0.2%) bacterial phyla (**Figure 1**). Cyanobacteria and Tenericutes were only represented in the black-backed jackal with proportions ≥0.1% (0.2 and 0.1%, respectively). Differences in proportions of bacterial phyla were pronounced between the species for Actinobacteria (15.5% cheetah vs. 3.8% black-backed jackals), Bacteroidetes (5.8% cheetah vs. 26.1% black-backed jackals) and Firmicutes (56.2% cheetah vs. 40.5% black-backed jackals). Cheetahs and black-backed jackals had these phyla in common with domestic cats, dogs, and other carnivores (**Table 1**).

Bacterial reads of both species were assigned to 74 taxa with some being present in either one or both species with a proportion ≥0.1% at the finest resolution (Supplementary Table 1). Out of these, the black-backed jackal was represented in 68 taxa, whereas the cheetah was represented in only 42 taxa. Overall, the two species shared 37 taxonomic assignments (Supplementary Table 1). On the genus level, *Clostridium* (24.5%), *Collinsella* (12.2%), and *Blautia* (8.9%) were the taxa with the highest proportions in cheetahs, whereas in black-backed jackals *Bacteroides* (15.1%), *Clostridium* (9.2%), and *Fusobacterium* (8.4%) had the highest proportions.

To determine whether cheetahs and black-backed jackals can be distinguished from each other based on the diversity of their bacterial communities, we calculated alpha diversity using OTU abundance, Shannon index and PD (**Figure 2**). OTU abundance was higher in black-backed jackals than in cheetahs (Wilcoxon rank sum test: $W = 111.5$, $p < 0.001$), black-backed jackals had a more diverse bacterial community than cheetahs as revealed by the Shannon index ($W = 141$, $p < 0.001$) and bacterial communities were ecologically more diverse when incorporating information on bacterial phylogeny ($W = 38$, $p < 0.001$). Also, the proportions of bacterial taxa between cheetahs and black-backed jackals were significantly different (ks.boot test (1000 permutations): $D = 0.35$, $p < 0.01$). Beta diversity calculated using the UniFrac metric, which also incorporates bacterial taxonomy, revealed a strong discrimination between the two

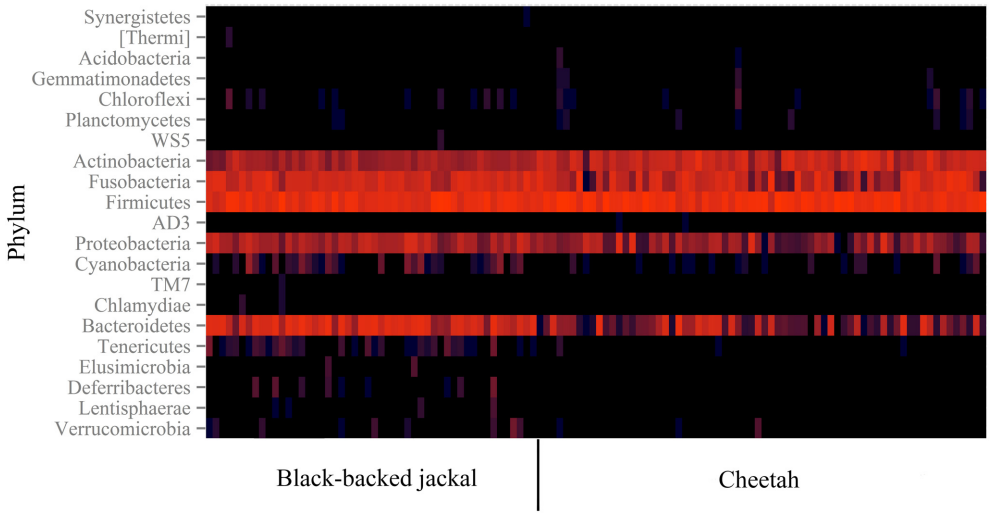


FIGURE 1 | Heatmap of 16S rRNA gene reads assigned to taxonomy on the bacterial phylum level. The color encodes the abundance of OTUs (log-scale of base 4) allocated to a specific phylum. Each column represents data of an individual black-backed jackal (left) or cheetah (right) based on a sub-sampling of 8000 reads per individual. The two species are similar in their phylum profiles and share all phyla with proportions above 0.2%.

Table 1 | Proportions (%) of dominant phyla present in domestic cats (cat 1: Tun et al., 2012; cat 2: Handl et al., 2011), free-ranging Iberian lynx (*Lynx pardinus*, Alcaide et al., 2012) and captive cheetahs (Becker et al., 2014), domestic dogs (K9BP dog and K9C dog: Swanson et al., 2010; dog: Handl et al., 2011) and free-ranging wolf (*Canis lupus*, Zhang and Chen, 2010) and profiles for cheetah (red) and black-backed jackal (red) of this study.

Phylum	Cat 1	Cat 2	Lynx	Captive cheetahs	Cheetah	K9BP dog	K9C dog	Dog	Wolf	Black-backed jackal
Bacteroidetes/chlorobi group	67.54	NA	NA	NA	NA	37.67	36.75	NA	NA	NA
Firmicutes	12.98	92.10	43.25	94.70	56.20	34.72	30.52	95.36	60.00	40.50
Proteobacteria	5.85	0.00	4.27	0.40	4.20	13.08	15.26	0.00	9.20	6.90
Fusobacteria	0.68	0.04	10.45	0.60	18.10	7.13	8.64	0.30	9.20	21.80
Bacteroidetes	8.68	0.45	39.43	0.00	5.80	3.14	4.47	2.25	16.90	26.10
Actinobacteria	1.16	7.31	1.78	4.30	15.50	1.01	1.00	1.81	4.60	3.80

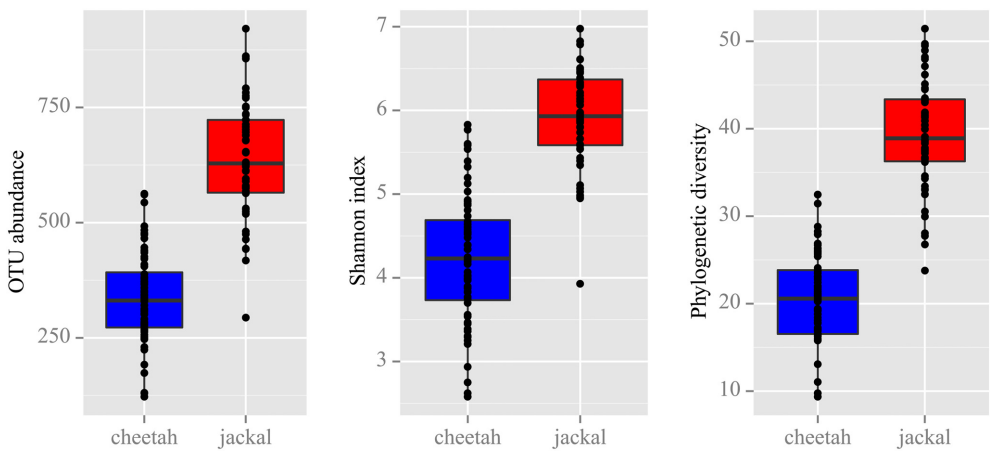


FIGURE 2 | Alpha diversity measures based on the rarefaction of 8000 reads revealed that black-backed jackals are significantly more diverse than cheetahs for OTU abundance, Shannon index and phylogenetic diversity (Wilcoxon rank sum test: all $p < 0.001$).

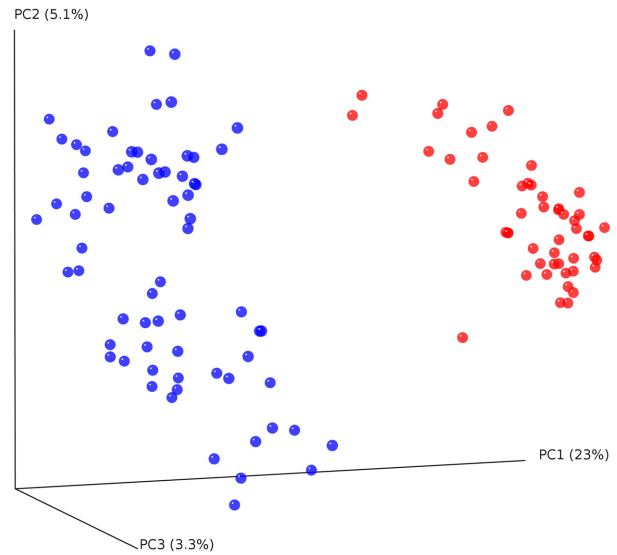


FIGURE 3 | Principal coordinate analysis (PCoA) plot of cheetahs (blue) and black-backed jackals (red) based on unweighted UniFrac metric. The two host species differ significantly in their bacterial communities [PERMANOVA: $R^2 = 0.74$, $F_{(1,\ 116)} = 337,02$, $p < 0.001$].

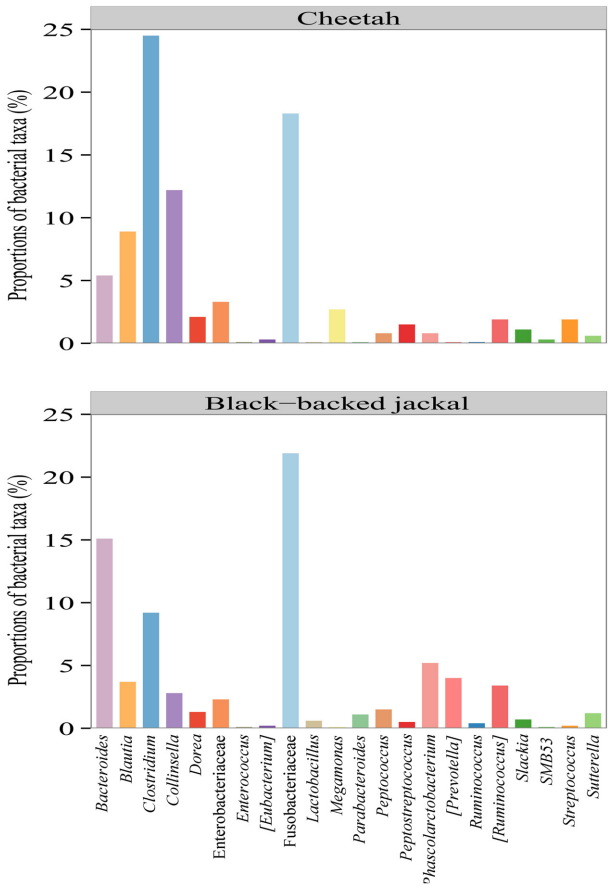


FIGURE 4 | Proportions of bacterial taxa which were present in both the cheetah and the black-backed jackal with proportions ≥0.1% (pooled on host species level) at the finest taxonomic resolution. Oligotyping was performed on all 20 bacterial genera and the two bacterial families to resolve the within taxon variation of bacteria between host species on a bacterial species-like level.

species [**Figure 3**; PERMANOVA: $R^2 = 0.74$, $F_{(1,\ 116)} = 337,02$, $p < 0.001$]. The three axes of the three-dimensional PCoA plot based on UniFrac distance measures explained more than 30% of the variation in the data set. The Bray-Curtis distance matrix revealed that cheetahs were less similar among each other than black-backed jackals (mean Bray-Curtis similarity indices for cheetahs = 34, for black-backed jackals = 40).

Oligotyping of bacterial reads extracted from 20 shared genera and two shared families (**Figure 4**) revealed differences in representative oligotypes and oligotype diversity between cheetahs and black-backed jackals (**Table 2**, Supplementary Table 2). *Collinsella* and *Lactobacillus* were the genera with the lowest and highest number of oligotypes, respectively, in both species. In general, black-backed jackals had a higher number of oligotypes for 12 out of 22 taxa, particularly in *Blautia* and *Megamonas*. Only in the genera [*Eubacterium*] and *Phascolarctobacterium* cheetahs carried a higher number of oligotypes. In *Clostridium*, *Enterococcus, Enterobacteriaceae*, *Fusobacteriaceae, Peptococcus*, *Peptostreptococcus*, [*Ruminococcus*], and *Sutterella* the number of oligotypes were identical in both species. Within each genus and within the two families of bacteria the proportions of shared oligotypes varied substantially between cheetahs and black-backed jackals, and some oligotypes were exclusively found in one species (Supplementary Figure 2). In 60% of cases the oligotype with the highest proportion was different for both species (Supplementary Figure 2). Overall, the distribution of number of oligotypes per taxon did not differ significantly between cheetahs and black-backed jackals [K-S test (1000 permutations): $D = 0.18$, $p = 0.69$]. The observed differences in proportions of oligotypes between the two species strongly varied between bacterial taxa (**Figure 5**). The genus *Slackia* exhibited the highest oligotype differences, whereas the family *Enterobacteriaceae* only showed a weak differentiation between the host species. Observed differences differed strongly from a random assignment of oligotypes to cheetah and black-backed jackal within each bacterial taxon (**Figure 5**; randomization test: $p < 0.001$). When we applied the oligotyping approach on the level of cheetah and black-backed jackal individuals, results were in line with the species-level comparison and individuals exhibited strong differences in oligotype profiles according to species identity (**Figure 6**). Nevertheless, results from oligotyping on the level of species and individuals are only comparable in a qualitative but not in a quantitative way due to differences in the data sets and parameters which were used for oligotyping.

DISCUSSION

Gut-associated microbial communities of two free-ranging Namibian carnivore species vary increasingly from the bacterial phylum level down to the bacterial species-like level of oligotypes. In general, most bacterial phyla to which reads were assigned were shared among both species, whereas the phyla Cyanobacteria and

Table 2 | Oligotyping results for the cheetah and the black-backed jackal for shared bacterial genera and families.

Taxon	Initial reads/ after filter	Reads after filter cheetah/jackal	OT in cheetah	OT in jackal	Number of shared OT	Base positions required to resolve OT
GENUS						
Bacteroides	400,000/285,663	154,366/131,297	16	17	16	30
Blautia	99,000/81,059	41,554/39,505	6	17	6	21
Clostridium	246,200/219,461	109,061/110,400	12	12	12	9
Collinsella	75,800/73,271	36,696/36,575	5	6	5	5
Dorea	35,000/29,436	14,862/14,574	10	12	7	15
Enterococcus	3100/2606	1295/1311	10	10	10	16
[Eubacterium]	6000/5418	2755/2663	11	9	7	13
Lactobacillus	12,000/9719	5151/4568	17	19	12	20
Megamonas	1500/1231	633/598	4	10	4	14
Parabacteroides	8600/6539	3363/3176	13	16	12	35
Peptococcus	39,000/36,792	18,237/18,555	9	9	9	5
Peptostreptococcus	13,200/12,508	6233/6275	10	10	10	8
Phascolarctobacterium	74,000/64,533	31,905/32,628	10	8	8	17
[Prevotella]	6300/5157	2682/2475	14	15	14	15
Ruminococcus	10,700/9255	4700/4485	6	9	6	25
[Ruminococcus]	89,400/78,556	40,974/37,582	10	10	10	10
SMB53	3300/2756	1385/1371	11	12	11	14
Slackia	18,000/16,058	8458/7627	9	10	4	15
Streptococcus	5400/5168	2605/2563	7	11	6	7
Sutterella	31,800/29,494	14,810/14,684	8	8	8	9
FAMILY						
Enterobacteriaceae	31,859/31,642	24,430/24,264	7	7	7	4
Fusobacteriaceae	581,800/497,644	263,944/233,700	15	15	15	16

Reads were extracted from the respective bacterial taxon to which they were assigned to and sub-sampled according to the maximum number of shared reads. After filtering, the remaining reads were used for oligotyping with the minimum number of base positions required to resolve all oligotypes (OT) in both samples.

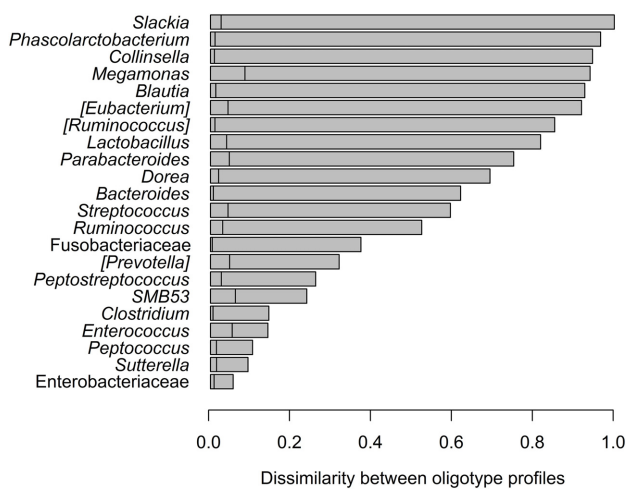


FIGURE 5 | Dissimilarity in oligotype profiles for each bacteria taxon shared between cheetahs and black-backed jackals. Observed dissimilarities are ranked from highest (top) to lowest differences (bottom). The vertical lines in each horizontal bar represent the 95% quantile above which differences between the observed and randomized dissimilarities become statistically significant. All observed dissimilarities were higher than randomized ones ($p < 0.001$).

Tenericutes were only present in black-backed jackals with proportions $\geq$0.1%. The phylum Tenericutes, especially the genus *Mollicutes*, comprises many parasitic bacteria (e.g., *Mycoplasma canis*) which can cause urogenital tract diseases (Chalker, 2005; Waltzek et al., 2012). Cyanobacteria are present in many terrestrial habitats but they are also very abundant in aquatic habitats (Stanier and Bazine, 1977). Thus, black-backed jackals might carry a higher proportion of Cyanobacteria because they also feed on amphibians (Walton and Joly, 2003). In addition, the proportions of shared bacterial genera were different between the two species. Black-backed jackals had higher proportions of *Bacteroides* and [*Prevotella*] than cheetahs. These two genera are known to be influenced by the diet (David et al., 2013). *Bacteroides* is associated with animal protein, several amino acids and saturated fats, whereas *Prevotella* is associated with hemicellulose and simple carbohydrates (Wu et al., 2011). Thus, the omnivorous diet of black-backed jackals requires a higher proportion of *Prevotella* to digest also plant material. In contrast, cheetahs would be expected to harbor a higher proportion of *Bacteroides* because of their strict carnivorous diet, yet this was not the case in our study. Furthermore, the proportions of the genera *Blautia, Clostridium, Megamonas*, and *Peptostreptococcus* increased when hosts fed on a diet with high fat contents compared to a diet with low fat contents (Bermingham et al., 2013).

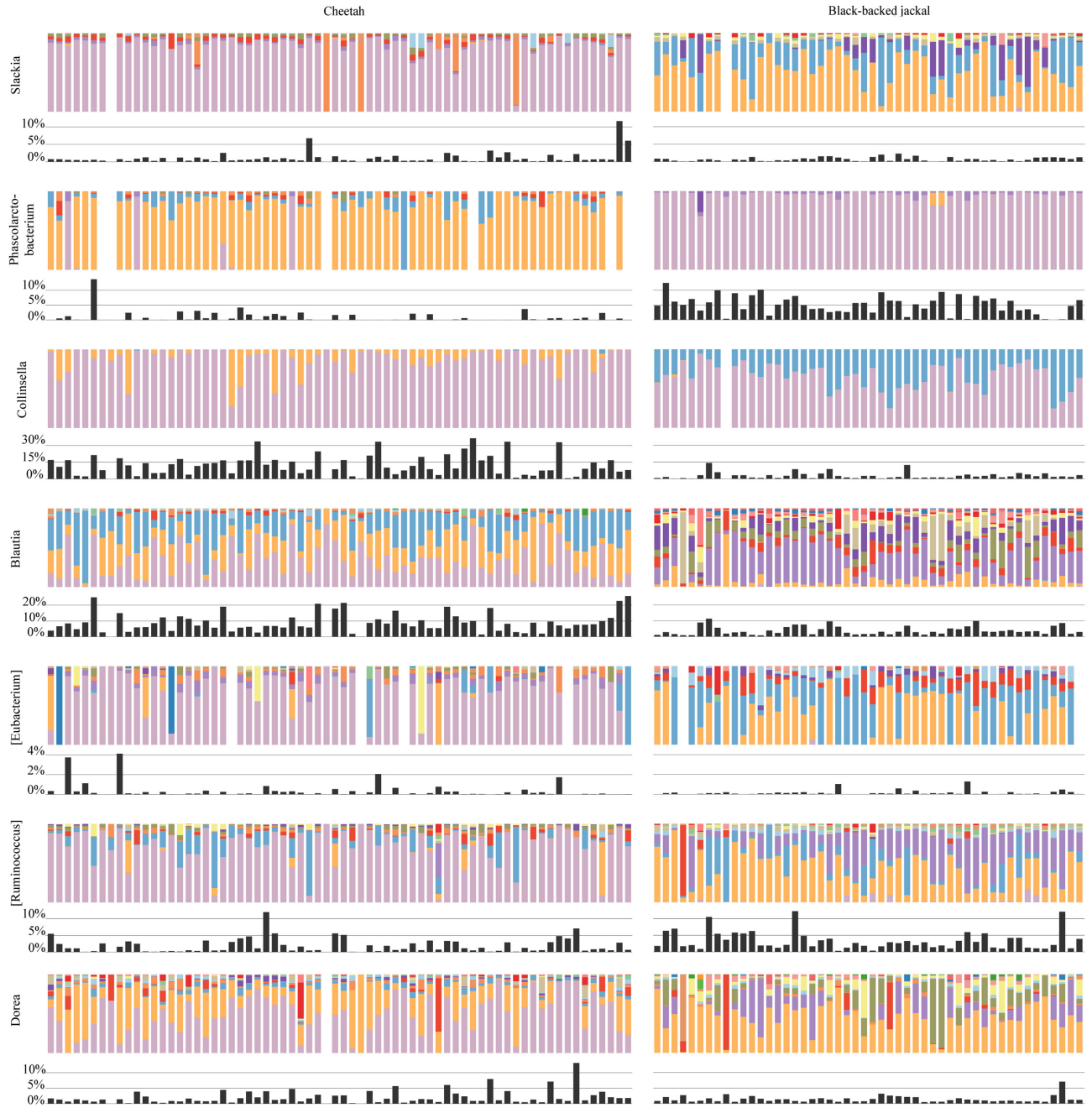


FIGURE 6 | Oligotype profiles for each cheetah and black-backed jackal individual for shared bacterial taxa which were ≥75% dissimilar in the species level comparison. *Lactobacillus*, *Megamonas*, and *Parabacteroides* are excluded from this analysis due to the limited number of host individuals which contributed to this bacterial taxon. Colors of bars within but not between bacterial taxa reflect the same oligotype and the respective proportion present in each individual. Black bars show the proportions of reads that were assigned to each bacterial taxon for every individual.

This relationship was reversed for *Lactobacillus* spp. which are usually seen as a beneficial group of microbes supporting nutrient acquisition in herbivores (Famularo et al., 2005). In the present study, *Lactobacillus* had a higher proportion in the omnivorous black-backed jackal, whereas the other mentioned genera had higher proportions in the cheetah. Thus, our findings suggest that a strictly carnivorous diet leads to a higher fat intake than an omnivorous diet.

Microbiomes of cheetahs and black-backed jackals share some characteristics with microbiomes of domestic cats and dogs (Swanson et al., 2010; Handl et al., 2011; Tun et al., 2012) and other mammals (Ley et al., 2008; Zhang and Chen, 2010;

Alcaide et al., 2012). The microbiomes differ mainly in proportions of phyla rather than differences in diversity (**Table 1**). In one of the first studies on gut-microbial communities using 454 next-generation sequencing in domestic cats and dogs, the phylum with the highest proportions was Firmicutes followed by Actinobacteria in cats and Bacteroidetes in dogs (Handl et al., 2011). Black-backed jackals harbor the same taxa in high proportions as dogs, whereas for cheetahs the second most abundant phylum was Fusobacteria followed closely by Actinobacteria. When looking at a higher taxonomic resolution, cheetahs and black-backed jackals shared the same genera (*Slackia* and *Collinsella*) within the phylum Actinobacteria which was also true for domestic dogs but not for domestic cats that carried *Eggerthella* and *Olsenella* (Handl et al., 2011). To our knowledge, only two studies focused on the microbiomes of free-ranging wildlife species belonging to the feline (Alcaide et al., 2012) and canine family (Zhang and Chen, 2010). These studies investigated the gut-bacterial communities of free-ranging Iberian lynx (*Lynx pardinus*) and free-ranging wolf (*Canis lupus*), respectively. Cheetahs and Iberian lynx both have high proportions of the phylum Firmicutes (56.20 vs. 43.25%) and similar proportions for the phylum Proteobacteria (4.20 vs. 4.27%). However, differences were quite pronounced for Fusobacteria (18.10 vs. 10.45%), *Bacteroides* (5.80 vs. 39.43%) and Actinobacteria (15.50 vs. 1.78%). Gut-bacterial communities of two captive cheetahs analyzed with 16S rRNA gene clone libraries (Becker et al., 2014) were very different from bacterial communities of free-ranging cheetahs in our study. Captive cheetahs had high proportions of Firmicutes (94.7 vs. 56.2% in free-ranging cheetahs) and low proportions of, e.g., Fusobacteria (0.6 vs. 18.1%). In contrast, black-backed jackals and free-ranging wolves exhibited similar ranks for bacterial phyla. Nevertheless, proportions also differed between Firmicutes (40.50 vs. 60.00%), Fusobacteria (21.80 vs. 9.20%) and *Bacteroides* (26.10 vs. 16.90%). Although these findings might be biased to some extent by the varying extraction and sequencing methods and differences in samples sizes, they reveal large variation in bacterial proportions already at the phylum level.

Comparisons between cheetahs and black-backed jackals using the alpha diversity measures OTU abundance, Shannon index and PD revealed that black-backed jackals had higher alpha diversities for all measures. Also, beta diversity measures based on UniFrac and Bray-Curtis distance clearly discriminated bacterial diversity according to host species. The microbial community in the black-backed jackal needs to achieve digestion of prey items from various sources ranging from meat to plant material, whereas bacteria in cheetahs are confronted with a much more restricted diet. In addition, the social system of black-backed jackals and their foraging behavior favors the exchange of bacteria via contact with conspecifics and intake of bacteria from carcasses (Walton and Joly, 2003; VanderWaal et al., 2013). The fact that cheetahs use a larger territory promoting microbial intake from a wide range of habitat types seems to be of minor impact compared to the factors that drive bacterial diversification in black-backed jackals. Although evidence exists that individuals exhibit different microbiomes in geographically distant habitats, differences have only been demonstrated for within-species comparisons (Fallani et al., 2010; Linnenbrink et al., 2013). When looking at between-species differences, environmental factors are difficult to distinguish from other factors such as host phylogeny, behavior or diet (Ley et al., 2008; Phillips et al., 2012).

Assignment of taxonomy to bacterial OTUs is a common approach to investigate bacterial taxa present in samples of interest. Nevertheless, due to the restrictions in resolution and richness of bacterial databases, assignments are rarely better than genus level. To resolve bacteria within the same genus between host species requires a higher taxonomic resolution. We have achieved this by oligotyping reads extracted from shared bacterial taxa in cheetahs and black-backed jackals. Thereby, we revealed a strong association and differentiation of oligotypes according to host species which was not explained by the OTU-clustering approach. Some genera exhibited strong differences in oligotype profiles, whereas others were similar with changes only in proportions of oligotypes. These differences in oligotypes might be due to a co-evolutionary fine-tuning of some genera according to the digestive requirements of the host (Ley et al., 2008). Alternatively, they may reflect the bacterial "speciation" in an enclosed host system in which almost no genetic exchange exists with the respective bacteria in another host species. Although oligotypes are different between the two carnivores, a functional redundancy might enable them to possess similar microbial genes and metabolic pathways (Suchodolski et al., 2009; Muegge et al., 2011). In most taxa one oligotype accounted for the majority of bacterial reads, which demonstrates that the diversity of genera might be more important for the digestive requirements of a host than the bacterial diversity within a genus.

CONCLUSION

Oligotyping in our study revealed gut microbiome differences between the cheetah and the black-backed jackal at a high taxonomic resolution. As a technique, oligotyping encompasses the limitation of the OTU approach by decomposing bacterial OTUs or taxa by minimizing the entropy within a group of reads down to a single base. Thus, resolved oligotypes act as proxies for bacterial species and thereby increase the amount of in-depth information that can be extracted from short sequencing reads of the 16S rRNA gene. By applying this new approach in conjunction with the OTU clustering approach we described similarities and differences between these two carnivore species from the bacterial phylum down to the species-like level of oligotypes. Cheetahs exhibited a lower bacterial diversity than black-backed jackals indicating that the size of home ranges is less important in shaping the microbiome of a host than the respective diet, foraging behavior and social system.

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: Sebastian Menke, Simone Sommer. Performed the experiments: Sebastian Menke, Matthias Meier, Wasimuddin. Field logistic and sample collection: Sebastian Menke, Matthias Meier, Jörg Melzheimer, Sonja Heinrich, Susanne Thalwitzer, Bettina Wachter, John K. E. Mfune. Analyzed the genomic data: Sebastian Menke. Statistical analysis: Sebastian Menke. Wrote the paper: Sebastian Menke drafted the manuscript, Simone Sommer, Bettina Wachter, Wasimuddin

critically reviewed the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We would like to thank the German Research Foundation (DFG; SO 428/10-1), the Messerli Foundation in Switzerland and the Leibniz Institute for Zoo and Wildlife Research (IZW) in Germany for funding. Wasimuddin's visit in Simone Sommer's lab was partly supported by the Erasmus Lifelong Learning Programme and the Czech NextGen Project (CZ.1.07/2.3./20.0303). We also thank the Namibian Ministry of Environment and Tourism for permission to conduct the research (CRP: Permits 525/2002-1689/2012; BBJP: Permits 1618/2011 and 1723/2012), A Schmidt for technical and laboratorial assistance, M Allgaier (Berlin Center for Genomics in Biodiversity Research, BeGenDiv) for advice on the sequencing strategy, A Courtiol for statistical assistance, I Heckman for programming helpful python scripts, O Aschenborn for sharing his veterinary experience, B Wasiolka for valuable assistance in the field and M Gillingham for fruitful discussions. We specially thank the Namibian farmers and predator controllers for their immense support and collaboration, without them, this project would not have been possible.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2014.00526/abstract

REFERENCES

Alcaide, M., Messina, E., Richter, M., Bargiela, R., Peplies, J., Huws, S. A., et al. (2012). Gene sets for utilization of primary and secondary nutrition supplies in the distal gut of endangered Iberian lynx. *PLoS ONE* 7:e51521. doi: 10.1371/journal.pone.0051521

Amato, K. R. (2013). Co-evolution in context: the importance of studying gut microbiomes in wild animals. *Microbiome Sci. Med.* 1, 10–29. doi: 10.2478/micsm-2013-0002

Amato, K. R., Yeoman, C. J., Kent, A., Righini, N., Carbonero, F., Estrada, A., et al. (2013). Habitat degradation impacts black howler monkey (*Alouatta pigra*) gastrointestinal microbiomes. *ISME J.* 7, 1344–1353. doi: 10.1038/ismej.2013.16

Becker, A. A., Hesta, M., Hollants, J., Janssens, G. P., and Huys, G. (2014). Phylogenetic analysis of faecal microbiota from captive cheetahs reveals underrepresentation of Bacteroidetes and Bifidobacteriaceae. *BMC Microbiol.* 14:43. doi: 10.1186/1471-2180-14-43

Benson, A. K., Kelly, S. A., Legge, R., Ma, F., Low, S. J., Kim, J., et al. (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc. Natl. Acad. Sci. U.S.A.* 107, 18933–18938. doi: 10.1073/pnas.1007028107

Bercik, P., Denou, E., Collins, J., Jackson, W., Lu, J., Jury, J., et al. (2011). The intestinal microbiota affect central levels of brain-derived neurotropic factor and behavior in mice. *Gastroenterology* 141, 599–609, 609.e1–3. doi: 10.1053/j.gastro.2011.04.052

Bergey, D. H., Buchanan, R. E., and Gibbons, N. E. (1975). *Bergey's Manual of Determinative Bacteriology*. Baltimore, Williams and Wilkins Co.

Bermingham, E. N., Young, W., Kittelmann, S., Kerr, K. R., Swanson, K. S., Roy, N. C., et al. (2013). Dietary format alters fecal bacterial populations in the domestic cat (*Felis catus*). *Microbiologyopen* 2, 173–181. doi: 10.1002/mbo3.60

Bolnick, D. I., Snowberg, L. K., Hirsch, P. E., Lauber, C. L., Org, E., Parks, B., et al. (2014). Individual diet has sex-dependent effects on vertebrate gut microbiota. *Nat. Commun.* 5, 1–13. doi: 10.1038/ncomms5500

Bravo, J. A., Forsythe, P., Chew, M. V., Escaravage, E., Savignac, H. M., Dinan, T. G., et al. (2011). Ingestion of *Lactobacillus* strain regulates emotional behavior and central GABA receptor expression in a mouse via the vagus nerve. *Proc. Natl. Acad. Sci. U.S.A.* 108, 16050–16055. doi: 10.1073/pnas.1102999108

Bray, J. R., and Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* 27, 325. doi: 10.2307/1942268

Caporaso, J. G., Bittinger, K., Bushman, F. D., DeSantis, T. Z., Andersen, G. L., and Knight, R. (2010a). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* 26, 266–267. doi: 10.1093/bioinformatics/btp636

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010b). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* 6, 1621–1624. doi: 10.1038/ismej.2012.8

Caro, T. M. (1994). *Cheetahs of the Serengeti Plains: Group Living in an Asocial Species*. Chicago: University of Chicago Press.

Chalker, V. J. (2005). Canine mycoplasmas. *Res. Vet. Sci.* 79, 1–8. doi: 10.1016/j.rvsc.2004.10.002

Coolon, J. D., Jones, K. L., Narayanan, S., and Wisely, S. M. (2010). Microbial ecological response of the intestinal flora of *Peromyscus maniculatus* and *P. leucopus* to heavy metal contamination. *Mol. Ecol.* 19, 67–80. doi: 10.1111/j.1365-294X.2009.04485.x

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., et al. (2013). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505, 559–563. doi: 10.1038/nature12820

Delsuc, F., Metcalf, J. L., Wegener Parfrey, L., Song, S. J., González, A., and Knight, R. (2014). Convergence of gut microbiomes in myrmecophagous mammals. *Mol. Ecol.* 23, 1301–1317. doi: 10.1111/mec.12501

Eaton, T. L. (1974). *The Cheetah—the Biology, Ecology, and Behavior of an Endangered Species*. New York, NY: Van Nostrand Reinhold Company.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27, 2194–2200. doi: 10.1093/bioinformatics/btr381

Eren, A. M., Borisy, G. G., Huse, S. M., and Mark Welch, J. L. (2014). Oligotyping analysis of the human oral microbiome. *Proc. Natl. Acad. Sci. U.S.A.* 111, E2875–E2884. doi: 10.1073/pnas.1409644111

Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., et al. (2013). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* 4, 1111–1119. doi: 10.1111/2041-210X.12114

Eren, A. M., Zozaya, M., Taylor, C. M., Dowd, S. E., Martin, D. H., and Ferris, M. J. (2011). Exploring the diversity of *Gardnerella vaginalis* in the genitourinary tract microbiota of monogamous couples through subtle nucleotide variation. *PLoS ONE* 6:e26732. doi: 10.1371/journal.pone.0026732

Ezenwa, V. O., Gerardo, N. M., Inouye, D. W., Medina, M., and Xavier, J. B. (2012). Animal behavior and the microbiome. *Science* 338, 198–199. doi: 10.1126/science.1227412

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* 61, 1–10. doi: 10.1016/0006-3207(92)91201-3

Fallani, M., Young, D., Scott, J., Norin, E., Amarri, S., Adam, R., et al. (2010). Intestinal microbiota of 6-week-old infants across Europe: geographic influence beyond delivery mode, breast-feeding, and antibiotics. *J. Pediatr. Gastroenterol. Nutr.* 51, 77–84. doi: 10.1097/MPG.0b013e3181d1b11e

Famularo, G., De Simone, C., Pandey, V., Sahu, A. R., and Minisola, G. (2005). Probiotic lactobacilli: an innovative tool to correct the malabsorption syndrome of vegetarians? *Med. Hypotheses* 65, 1132–1135. doi: 10.1016/j.mehy.2004.09.030

Goldenberg, M., Goldenberg, F., Funk, S. M., Millesi, E., and Henschel, J. (2010). Diet composition of black-backed jackals, *Canis mesomelas* in the Namib desert. *Folia Zool.* 59, 93–101.

Handl, S., Dowd, S. E., Garcia-Mazcorro, J. F., Steiner, J. M., and Suchodolski, J. S. (2011). Massive parallel 16S rRNA gene pyrosequencing reveals highly diverse fecal bacterial and fungal communities in healthy dogs and cats. *FEMS Microbiol. Ecol.* 76, 301–310. doi: 10.1111/j.1574-6941.2011.01058.x

Hooper, L. V., Littman, D. R., and Macpherson, A. J. (2012). Interactions between the microbiota and the immune system. *Science* 336, 1268–1273. doi: 10.1126/science.1223490

Jenner, N., Groombridge, J., and Funk, S. (2011). Commuting, territoriality and variation in group and territory size in a black-backed jackal population reliant on a clumped, abundant food resource in Namibia. *J. Zool.* 248, 231–238. doi: 10.1111/j.1469-7998.2011.00811.x

Kamler, J. F., Stenkewitz, U., Klare, U., Jacobsen, N. F., and Macdonald, D. W. (2012). Resource partitioning among cape foxes, bat-eared foxes, and black-backed jackals in South Africa. *J. Wildl. Manag.* 76, 1241–1253. doi: 10.1002/jwmg.354

Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., and Gordon, J. I. (2011). Human nutrition, the gut microbiome and the immune system. *Nature* 474, 327–336. doi: 10.1038/nature10213

Koch, H., and Schmid-Hempel, P. (2011). Socially transmitted gut microbiota protect bumble bees against an intestinal parasite. *Proc. Natl. Acad. Sci. U.S.A.* 108, 19288–19292. doi: 10.1073/pnas.1110474108

Kuczynski, J., Lauber, C. L., Walters, W. A., Parfrey, L. W., Clemente, J. C., Gevers, D., et al. (2012). Experimental and analytical tools for studying the human microbiome. *Nat. Rev. Genet.* 13, 47–58. doi: 10.1038/nrg3129

Ley, R. E., Hamady, M., Lozupone, C., Turnbaugh, P. J., Ramey, R. R., Bircher, J. S., et al. (2008). Evolution of mammals and their gut microbes. *Science* 320, 1647–1651. doi: 10.1126/science.1155725

Linnenbrink, M., Wang, J., Hardouin, E. A., Künzel, S., Metzler, D., and Baines, J. F. (2013). The role of biogeography in shaping diversity of the intestinal microbiota in house mice. *Mol. Ecol.* 22, 1904–1916. doi: 10.1111/mec.12206

Lozupone, C., and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* 71, 8228–8235. doi: 10.1128/AEM.71.12.8228-8235.2005

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., and Knight, R. (2011). UniFrac: an effective distance metric for microbial community comparison. *ISME J.* 5, 169–172. doi: 10.1038/ismej.2010.133

Magoè, T., and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27, 2957–2963. doi: 10.1093/bioinformatics/btr507

Marker, L. L., Dickman, A. J., Mills, M. G., Jeo, R. M., and Macdonald, D. W. (2008). Spatial ecology of cheetahs on north-central Namibian farmlands. *J. Zool.* 274, 226–238. doi: 10.1111/j.1469-7998.2007.00375.x

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* 17, 10–12. doi: 10.14806/ej.17.1.200

McKenna, P., Hoffmann, C., Minkah, N., Aye, P. P., Lackner, A., Liu, Z., et al. (2008). The macaque gut microbiome in health, lentiviral infection, and chronic enterocolitis. *PLoS Pathog* 4:e20. doi: 10.1371/journal.ppat.0040020

McLellan, S. L., Newton, R. J., Vandewalle, J. L., Shanks, O. C., Huse, S. M., Eren, A. M., et al. (2013). Sewage reflects the distribution of human faecal *Lachnospiraceae*. *Environ. Microbiol.* 15, 2213–2227. doi: 10.1111/1462-2920.12092

McMurdie, P. J., and Holmes, S. (2013). *Package "phyloseq."* Available online at: http://bioconductor.fhcrc.org/packages/2.13/bioc/manuals/phyloseq/man/phyloseq.pdf [Accessed November 26, 2013].

Muegge, B. D., Kuczynski, J., Knights, D., Clemente, J. C., Gonzalez, A., Fontana, L., et al. (2011). Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* 332, 970–974. doi: 10.1126/science.1198719

Nelson, T. M., Rogers, T. L., Carlini, A. R., and Brown, M. V. (2013). Diet and phylogeny shape the gut microbiota of Antarctic seals: a comparison of wild and captive animals. *Environ. Microbiol.* 15, 1132–1145. doi: 10.1111/1462-2920.12022

Ochman, H., Worobey, M., Kuo, C.-H., Ndjango, J.-B. N., Peeters, M., Hahn, B. H., et al. (2010). Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLoS Biol.* 8:e1000546. doi: 10.1371/journal.pbio.1000546

Phillips, C. D., Phelan, G., Dowd, S. E., McDonough, M. M., Ferguson, A. W., Delton Hanson, J., et al. (2012). Microbiome analysis among bats describes influences of host phylogeny, life history, physiology and geography. *Mol. Ecol.* 21, 2617–2627. doi: 10.1111/j.1365-294X.2012.05568.x

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R foundation for statistical computing. Available online at: http://www.R-project.org

Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., et al. (2012). Genomic variation landscape of the human gut microbiome. *Nature* 493, 45–50. doi: 10.1038/nature11711

Schwab, C., Cristescu, B., Northrup, J. M., Stenhouse, G. B., and Gänzle, M. (2011). Diet and environment shape fecal bacterial microbiota composition and enteric pathogen load of grizzly bears. *PLoS ONE* 6:e27905. doi: 10.1371/journal.pone.0027905

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *J. Stat. Softw.* 42, 1–52.

Sekirov, I., Russell, S. L., Antunes, L. C. M., and Finlay, B. B. (2010). Gut microbiota in health and disease. *Physiol. Rev.* 90, 859–904. doi: 10.1152/physrev.00045.2009

Shannon, C. E., and Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana: The University of Illinois Press.

Spellerberg, I. F., and Fedor, P. J. (2003). A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the "Shannon–Wiener" Index. *Glob. Ecol. Biogeogr.* 12, 177–179. doi: 10.1046/j.1466-822X.2003.00015.x

Spor, A., Koren, O., and Ley, R. (2011). Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat. Rev. Microbiol.* 9, 279–290. doi: 10.1038/nrmicro2540

Stanier, R. Y., and Bazine, G. C. (1977). Phototrophic prokaryotes: the cyanobacteria. *Annu. Rev. Microbiol.* 31, 225–274. doi: 10.1146/annurev.mi.31.100177.001301

Suchodolski, J. (2011). Intestinal microbiota of dogs and cats: a bigger world than we thought. *Vet. Clin. North Am. Small Anim. Pract.* 41, 261–272. doi: 10.1016/j.cvsm.2010.12.006

Suchodolski, J. S., Dowd, S. E., Westermarck, E., Steiner, J. M., Wolcott, R. D., Spillmann, T., et al. (2009). The effect of the macrolide antibiotic tylosin on microbial diversity in the canine small intestine as demonstrated by massive parallel 16S rRNA gene sequencing. *BMC Microbiol.* 9:210. doi: 10.1186/1471-2180-9-210

Swanson, K. S., Dowd, S. E., Suchodolski, J. S., Middelbos, I. S., Vester, B. M., Barry, K. A., et al. (2010). Phylogenetic and gene-centric metagenomics of the canine intestinal microbiome reveals similarities with humans and mice. *ISME J.* 5, 639–649. doi: 10.1038/ismej.2010.162

Tun, H. M., Brar, M. S., Khin, N., Jun, L., Hui, R. K.-H., Dowd, S. E., et al. (2012). Gene-centric metagenomics analysis of feline intestinal microbiome using 454 junior pyrosequencing. *J. Microbiol. Methods* 88, 369–376. doi: 10.1016/j.mimet.2012.01.001

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027–1131. doi: 10.1038/nature05414

Turnbaugh, P. J., Ridaura, V. K., Faith, J. J., Rey, F. E., Knight, R., and Gordon, J. I. (2009). The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci. Transl. Med.* 1, 6ra14. doi: 10.1126/scitranslmed.3000322

VanderWaal, K. L., Atwill, E. R., Isbell, L. A., and McCowan, B. (2013). Linking social and pathogen transmission networks using microbial genetics in giraffe (*Giraffa camelopardalis*). *J. Anim. Ecol.* 83, 406–414. doi: 10.1111/1365-2656.12137

Wachter, B., Jauernig, O., and Breitenmoser, U. (2006). Determination of prey hair in faeces of free-ranging Namibian cheetahs with a simple method. *Cat. News* 44, 8–9.

Walton, L. R., and Joly, D. O. (2003). *Canis mesomelas*. *Mamm. Species* 715, 1–9. doi: 10.1644/715

Waltzek, T. B., Cortés-Hinojosa, G., Wellehan, J. F. X. Jr., and Gray, G. C. (2012). Marine mammal zoonoses: a review of disease manifestations. *Zoonoses Public Health* 59, 521–535. doi: 10.1111/j.1863-2378.2012.01492.x

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/AEM.00062-07

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer.

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334, 105–108. doi: 10.1126/science.1208344

Yeoman, C. J., Chia, N., Yildirim, S., Miller, M. E. B., Kent, A., Stumpf, R., et al. (2011). Towards an evolutionary model of animal-associated microbiomes. *Entropy* 13, 570–594. doi: 10.3390/e13030570

Zhang, H., and Chen, L. (2010). Phylogenetic analysis of 16S rRNA gene sequences reveals distal gut bacterial diversity in wild wolves (*Canis lupus*). *Mol. Biol. Rep.* 37, 4013–4022. doi: 10.1007/s11033-010-0060-z

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 10 June 2014; paper pending published: 09 July 2014; accepted: 21 September 2014; published online: 14 October 2014.

Citation: Menke S, Wasimuddin, Meier M, Melzheimer J, Mfune JKE, Heinrich S, Thalwitzer S, Wachter B and Sommer S (2014) Oligotyping reveals differences between gut microbiomes of free-ranging sympatric Namibian carnivores (Acinonyx jubatus, Canis mesomelas) on a bacterial species-like level. Front. Microbiol. ***5****:526. doi: 10.3389/fmicb.2014.00526*

This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology.

frontiers in MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 07 November 2014
doi: 10.3389/fmicb.2014.00568

Dynamics of tongue microbial communities with single-nucleotide resolution using oligotyping

Jessica L. Mark Welch[1,2]*, Daniel R. Utter[1,2], Blair J. Rossetti[2], David B. Mark Welch[1], A. Murat Eren[1] and Gary G. Borisy[1,2]

[1] *Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA, USA*
[2] *Department of Microbiology, The Forsyth Institute, Cambridge, MA, USA*

Edited by:
Angel Angelov, Technische Universität München, Germany

Reviewed by:
Peter Bergholz, North Dakota State University, USA
Thomas Jefferson Sharpton, Oregon State University, USA

****Correspondence:***
Jessica L. Mark Welch, Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, 7 MBL Street, Woods Hole, MA 02543, USA
e-mail: jmarkwelch@mbl.edu

The human mouth is an excellent system to study the dynamics of microbial communities and their interactions with their host. We employed oligotyping to analyze, with single-nucleotide resolution, oral microbial 16S ribosomal RNA (rRNA) gene sequence data from a time course sampled from the tongue of two individuals, and we interpret our results in the context of oligotypes that we previously identified in the oral data from the Human Microbiome Project. Our previous work established that many of these oligotypes had dramatically different distributions between individuals and across oral habitats, suggesting that they represented functionally different organisms. Here we demonstrate the presence of a consistent tongue microbiome but with rapidly fluctuating proportions of the characteristic taxa. In some cases closely related oligotypes representing strains or variants within a single species displayed fluctuating relative abundances over time, while in other cases an initially dominant oligotype was replaced by another oligotype of the same species. We use this high temporal and taxonomic level of resolution to detect correlated changes in oligotype abundance that could indicate which taxa likely interact synergistically or occupy similar habitats, and which likely interact antagonistically or prefer distinct habitats. For example, we found a strong correlation in abundance over time between two oligotypes from different families of Gamma Proteobacteria, suggesting a close functional or ecological relationship between them. In summary, the tongue is colonized by a microbial community of moderate complexity whose proportional abundance fluctuates widely on time scales of days. The drivers and functional consequences of these community dynamics are not known, but we expect they will prove tractable to future, targeted studies employing taxonomically resolved analysis of high-throughput sequencing data sampled at appropriate temporal intervals and spatial scales.

Keywords: human microbiome, oral microbiota, 16S ribosomal RNA, *Haemophilus*, *Neisseria*, *Streptococcus*, *Veillonella*

INTRODUCTION

Understanding microbial community dynamics requires knowledge of the time scale over which microbial communities adapt and change. Studies using rRNA gene-based approaches to investigate microbial communities sampled at intervals of weeks to months found that these communities correlated to environmental conditions (Fuhrman et al., 2006; Dethlefsen et al., 2008; Gilbert et al., 2012; Chow et al., 2013). Indications that changes of interest may occur over shorter time scales led to studies that sampled at daily intervals in a marine system and in the human microbiome (Dethlefsen and Relman, 2011; Caporaso et al., 2011; Koenig et al., 2011; Gajer et al., 2012; Martínez et al., 2013; Needham et al., 2013; David et al., 2014). These studies established that microbial communities are resilient, with episodic shifts in community composition followed by reversion to previous states. Remarkably, within that overall stability, dramatic fluctuations in community composition could occur on time scales of the order of days.

Our understanding of microbial community dynamics at the species level has heretofore been hindered by the use of analysis methods that cluster sequences into operational taxonomic units (OTUs) based on arbitrary similarity thresholds. Such methods have the twin drawbacks that they generate heterogeneous groupings of limited biological relevance and that they do not make full use of available sequence information that would allow finer taxonomic resolution. Many described microbial species differ by only 1 or 2% in rRNA gene sequence, yet standard analysis methods lump them together by clustering sequences that are more than 97% identical. A recently developed computational method called oligotyping (Eren et al., 2013) removes this hindrance. Oligotyping uses a calculation of Shannon entropy to identify nucleotide positions of high variation (i.e., high information content) in a dataset, and employs only these positions to partition the sequence dataset into groups called oligotypes. It exploits all available informative data, reduces the effect of noise, and generates homogeneous groupings in the sense that

nearly every read assigned to an oligotype, if classified individually by BLAST, would have the same taxonomic annotation (Eren et al., 2014). Oligotyping allows the analysis of high-throughput sequencing datasets with single-nucleotide resolution. A different approach that also achieves single-nucleotide resolution has recently been reported (Tikhonov et al., 2014).

Application of oligotyping to the human oral microbiota presents an opportunity to analyze a tractable microbial community with a level of taxonomic resolution that permits differentiation among important species and, in favorable cases, analysis of within-species dynamics. The human mouth is an excellent test bed for microbiome analysis for several reasons: it is home to a well-studied microbial community for which a highly curated Human Oral Microbiome Database (HOMD) (www.homd.org) has been established (Dewhirst et al., 2010); a high proportion of the human oral microbiota have been cultured (65%); fully sequenced genomes are available for many (50%) of the oral microbiota; and, importantly, a foundation for defining the healthy human oral microbiome has been laid by the Human Microbiome Project (HMP) (http://commonfund.nih.gov/hmp/index.aspx) which sampled nine oral sites from over 200 healthy individuals and generated millions of sequences (Human Microbiome Project Consortium, 2012).

The oral microbiota may be deconstructed into overlapping but distinct communities. For example, the human tongue is the substrate for an abundant microbiota different in composition from the microbiota on the teeth and on the mucosal surfaces of the mouth, as first indicated by distinctive distribution of a few taxa in DNA hybridization and early sequencing studies (Mager et al., 2003; Aas et al., 2005; Socransky and Haffajee, 2005; Zaura et al., 2009). Analysis of the HMP data confirmed the finding of broad differences in the microbiome of the tongue dorsum as compared to plaque and to the surfaces of the gums, cheek and hard palate (Segata et al., 2012).

The application of oligotyping to the HMP data for the oral microbiome (Eren et al., 2014), in combination with habitat analysis of oligotype distribution across nine oral sites, identified a level of ecological and functional biodiversity in the oral microbiome not previously recognized. We identified oral site-specialists, established correlations between sites within individual mouths, and revealed predominance of certain oligotypes within individuals that would not have been seen with OTU clustering. Some oligotypes differing by a few nucleotides or even as little as a single nucleotide showed strikingly different distributions across oral sites or among individuals, suggesting that even single-nucleotide differences in the 16S rRNA gene can act as markers for underlying, biologically significant differences elsewhere in the genome.

The HMP data provided an invaluable baseline for assessing variation in the microbiome across a range of individuals whose health status was carefully documented. However, this baseline represents a single "snapshot" in time from each of the sampled individuals, meaning that the significance of some distributional patterns of oligotypes remained unclear. Some very closely related oligotypes, for example representing different species of *Streptococcus*, were detected in the tongue of every individual, but in widely different proportions in different individuals; were these proportions a stable characteristic of an individual's microbiota or did they change over time and over what time scales? Other closely related oligotypes apparently represented different strains within a single species. For example, in the *Neisseria flavescens/subflava* group, one or another of these oligotypes would dominate the tongue community in an individual, making up 90% or more of the reads of that taxon. Is one oligotype stably dominant in each individual, or does the dominance relationship fluctuate?

A time-resolved high-throughput sequencing dataset from the tongue of two individuals (Caporaso et al., 2011) provided an ideal opportunity to test the stability of these distributions over time as well as to generate a more precise and unified description of the characteristic microbiota of the tongue. We carried out oligotyping on this dataset and compared the resulting oligotypes to those detected in HMP data. Oligotyping, similar to other *de novo* partitioning approaches, creates units that are dataset-specific and not inherently comparable across datasets. We overcame this limitation by making taxonomic assessments for each oligotype by reference to the HOMD. This association of oligotypes from separate datasets allowed us to apply the insights gained from a large time-series study of two individuals to the analysis of a large cross-sectional study with many individuals. It also provided resolution sufficient to discriminate very closely-related taxa, so that for the first time we can describe with species-level or near-species level precision the overall composition and temporal dynamics of the tongue microbial community.

METHODS

SAMPLE COLLECTION

This is a re-analysis of existing sequence data; procedures for informed consent, institutional review board approval, and sample collection and sequencing are described in the original publications (Caporaso et al., 2011; The Human Microbiome Project Consortium, 2012; Aagaard et al., 2013).

PREPARING THE SEQUENCE DATA

The study by Caporaso et al. (2011) describes in detail the sample collection, sequencing, and quality filtering of reads used in this study. Briefly, one male and one female adult were sampled approximately daily over 15 months (male) and 6 months (female). The V4 region of the 16S rRNA gene was amplified from tongue samples and amplicons were sequenced using the Illumina HiSeq platform (Illumina, Inc., San Diego, CA, USA). We obtained the quality-filtered data from MG-RAST (http://metagenomics.anl.gov/) using sample accession IDs 4457768.3 through 4459735.3. To eliminate the artificial length variation among reads introduced by the original quality trimming, we re-trimmed each read to 130 nucleotides, and removed the reads that were shorter. For each sample with >20,000 reads we randomly subsampled to 20,000 reads to minimize the sampling bias in our results. The resulting dataset contained 508 samples and a total of 7,538,132 sequencing reads. We used GAST (Huse et al., 2008) to assign taxonomy at the family level to each read in the dataset.

OLIGOTYPING

We used oligotyping pipeline version 1.0 available from http://oligotyping.org (Eren et al., 2013) on each taxonomic family separately. For each family, we used the auto component command (-c) to select the two nucleotide positions with the highest Shannon entropy, partioning each family into up to eight groups. Groups were further divided by manually adding additional nucleotide positions (using the -C parameter) based on the recalculated Shannon entropy and on the absolute and relative abundance distribution among samples of the unique sequences within a grouping. No more than 5 nucleotide positions were added in a single iteration. The minimum substantive abundance threshold for an oligotype (-M) was set to 500 reads. Upon completion of the oligotyping analysis for each family, we concatenated the resulting observation matrices to generate a single observation matrix reporting counts (i.e. number of reads assigned to each oligotype in each sample). We also converted counts to percent abundances within each sample and used these normalized relative abundances for all analyses except the cross-correlation analysis which was performed on the count data. We assigned taxonomic values to each oligotype by a BLAST search using NCBI executables (--blast-ref-db) against the HOMD RefSeq v.12.0 obtained from www.homd.org. Each oligotype was assigned the taxonomy of the closest match(es) in HOMD except for the one oligotype that had no match within 90% of any sequence in HOMD.

CROSS-CORRELATION AND AUTO-CORRELATION ANALYSIS

We carried out cross-correlation analysis using Matlab R2014a (version 8.3) using the counts matrix for each oligotype (the number of reads assigned to each oligotype in each sample) and using the percent matrix (the counts normalized within each sample). Results using Pearson cross-correlation are shown (Matlab function `corr`); we also carried out the same analysis using Spearman and Kendall with comparable results. Significance (*p*-value) was calculated using the `corr` function which employs a Student's t distribution for a transformation of the correlation. We used the Bonferroni correction for multiple tests by multiplying significance estimates by $315^2 \sim= 10^5$. Auto-correlation analysis was carried out using the Matlab function `xcorr` on percent-normalized data for the entire time course for each subject and for subsets of the male time course, and in each case was evaluated over a window of plus or minus 21 days. Potential periodicity of oligotype abundance was analyzed with Fourier transforms using the Matlab functions `fft` and `periodogram`. For this analysis, linear interpolation was used to estimate the relative abundance of oligotypes on days without sequencing data.

ANALYSIS OF V3-V5 READS FROM HMP DATA FOR MULTIPLE TIME POINTS

We used the HMP 16S sequence data from the V3-V5 region. Quality filtering and trimming, chimera removal, and taxonomic assignment of reads were previously performed (The Human Microbiome Project Consortium, 2012) using mothur (Schloss et al., 2011) and the reads were uploaded into a MySQL database. From this data we selected subjects from whom two tongue dorsum samples were available with at least 600 reads from each sample. We counted the number of reads assigned to each genus in each sample, and clustered this abundance data using the Morisita-Horn dissimilarity index.

BLAST SEARCHES OF MICROBIAL GENOMES

We conducted BLAST searches at HOMD (www.homd.org) using blastn against all oral microbial genomic DNAs annotated by HOMD, and at NCBI (www.ncbi.nlm.nih.gov) using megablast against all completed microbial genomes and against draft genomes of *Haemophilus* and *Neisseriaceae*.

RESULTS

OLIGOTYPING RESULTS

We used oligotyping to re-analyze time series data sampled from the tongues of two individuals at up to 396 time points (Caporaso et al., 2011). We oligotyped each of the 17 most abundant bacterial families, selecting sets of sequence reads based on their family-level taxonomic assignment using GAST (Huse et al., 2008). These 17 families together represented 99% of reads in the combined tongue data set, and this family-level oligotyping achieved a similarly comprehensive result to the phylum-level oligotyping of HMP data as previously described (Eren et al., 2014) but with lower complexity in the supervision process. The number of nucleotides we used to define oligotypes in the time series data set ranged from 3 (for Actinomycetaceae and Bacillales) to 24 (for Neisseriaceae). We partitioned the data into 315 oligotypes (Table S1) and assigned taxonomic identification to each by BLAST search of the representative sequence for each oligotype against the Human Oral Microbiome Database (HOMD, Dewhirst et al., 2010). Oligotyping of 16S rRNA gene tag sequence data from the tongue dorsum as well as eight other oral sites for 148 individuals sequenced in the V3-V5 region, and 77 individuals sequenced in the V1-V3 region, was previously described (Eren et al., 2014). Results from that study provide the foundation for the current study.

PHASE TRANSITION OF A MICROBIAL COMMUNITY

With the single-nucleotide resolution achievable by oligotyping, strains or variants within a taxon that differ in their rRNA sequence are in principle detectable and their population dynamics open to analysis. We previously found, for example, a case of closely-related oligotypes within the genus *Neisseria*, in which the *Neisseria* population on the tongue of each subject was dramatically different from the mean abundance of the oligotypes across all sampled subjects. Remarkably, the *Neisseria* population on an individual tongue was generally dominated by one or another of these oligotypes (**Figure 1** and Eren et al., 2014). To understand the cause of this distributional pattern, it is important to know whether the differences between individuals are stable, or whether populations within individuals change over time.

When only a single sample is analyzed from each of many individuals, as in **Figure 1**, it is impossible to assess whether populations within an individual are stable or dynamic. Most individuals possessed a population in which a single oligotype made up at least 90% of the *Neisseria* reads, but the oligotype varied from individual to individual, suggesting the possibility of multiple stable states, each dominated by a single *Neisseria* oligotype.

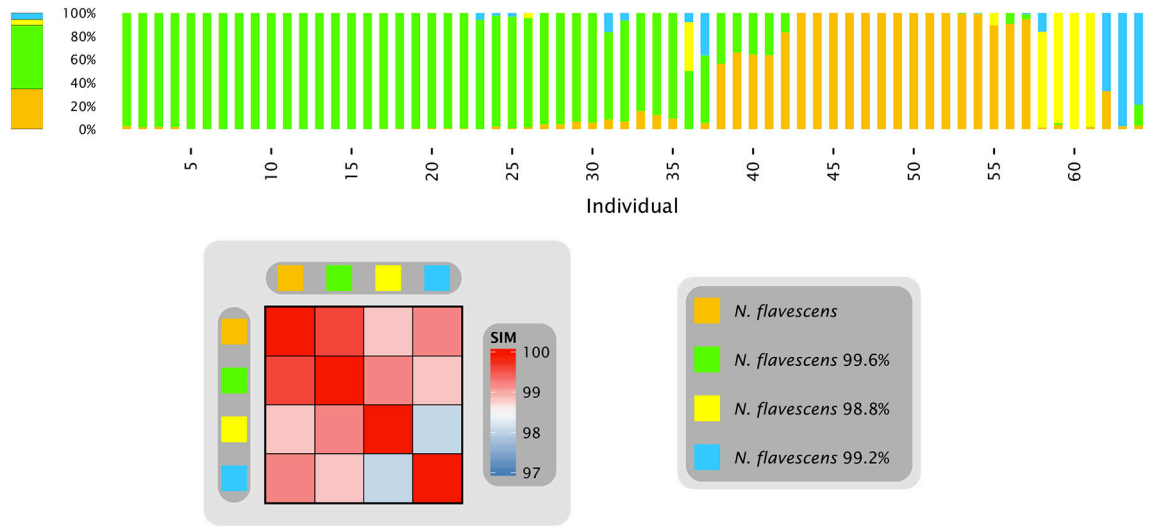


FIGURE 1 | *Neisseria* oligotypes in the tongue. Human Microbiome Project data from the tongue, sequenced in the V1-V3 region. The four oligotypes with at least 0.5% mean abundance across all sampled individuals are shown with their percent identity to *N. flavescens,* the closest match in HOMD. Colored bars represent the relative abundance of these four oligotypes in samples from an individual tongue; the 64 samples shown are those in which these *Neisseria* oligotypes represented a total of at least 10 reads. The bar on the left represents the relative abundance of each of these four oligotypes averaged across all 64 tongue samples. The heat map shows the similarity of each pair of oligotypes. Data are from Eren et al. (2014).

However, some individuals had populations lacking a dominant oligotype. Did the more mixed populations represent short-lived transitions between the stable states in individuals who were by chance sampled during the transition? Alternatively, did certain individuals stably maintain a mixed *Neisseria* population?

Oligotyping of a time series from the tongue of two individuals (**Figure 2**) answered some of these questions and provided plausible explanations for the observed distributions. Most of the *Neisseria* in both subjects consisted of three major oligotypes, shown in **Figure 2** as light blue (*Neisseria* A), dark blue (*Neisseria* B), and green (*Neisseria* C) with a small amount of a fourth oligotype shown as magenta (*Neisseria* D). The *Neisseria* population in both subjects was initially dominated by type C, which was the only *Neisseria* oligotype detectable in the first three samples from the female and two samples from the male. The additional types A and B then became detectable in both individuals, increasing rapidly as a proportion of the total *Neisseria* (**Figure 2**). In the female, type A was initially the more abundant of these two, but rapidly faded in abundance relative to type B, which became the dominant *Neisseria* in the female after approximately day 35. In the male, by contrast, type B increased and then decreased in relative abundance several times before diminishing in proportion until its abundance was negligible and the population was dominated by type A after approximately day 100. These dynamics display two main characteristics which, taken together, may be termed a phase transition. The major behavior is one of stability. For most of the time, the oligotype distribution within an individual was essentially invariant, irrespective of whether the dominant oligotype in the individual was type A or type B. The second property was of abrupt transition to an alternate oligotype. The time series data showed several instances in which a community initially dominated by one oligotype became transiently mixed and then transitioned to a state where one oligotype was dominant. These properties suggest that the evenly mixed populations of *Neisseria* on the tongue found in some individuals in the HMP data are transient states. Occasional replacement of the dominant oligotype argues against strong founder effects and priority effects for this taxon in the tongue microbiota. Throughout these transitions the fourth oligotype, type D, did not participate in the apparently competitive or exclusionary dynamics of types A and B, but persisted in relatively stable proportion in the community, likely demonstrating a subdivision of functional/ecological roles even among these very closely related taxa.

DIFFERENCES AMONG INDIVIDUALS ARE COMPARABLE TO FLUCTUATIONS OVER TIME

The stable dominance of one oligotype of *Neisseria* in each individual, relative to the other *Neisseria* oligotypes, occurred in a context of rapid fluctuation in the abundance of *Neisseria* and all other taxa as a proportion of the total community. The overall behavior of the system was a dynamic equilibrium with rapid fluctuations in relative abundance but without long-term directionality, as shown in **Figures 3**, **4**. **Figure 3** shows the relative proportions of the five most abundant *Streptococcus* oligotypes over time in each individual. The most abundant *Streptococcus* oligotype overall, labeled *Streptococcus* A in the figure, is identical to *S. mitis*, *S. oralis,* and *S. infantis* in the V4 region; this oligotype ranged in abundance from 9 to 75% of the *Streptococcus* in the female subject and from 10 to 92% of the *Streptococcus* in the male (**Figure 3A**). Relative abundance of taxa not only ranged widely but also changed quickly as seen, for example, in samples 269 and 270 from the male subject, in which the relative abundance of *S. mitis/oralis/infantis* dropped from 78 to 10% of the *Streptococcus* in the sample over the course of a

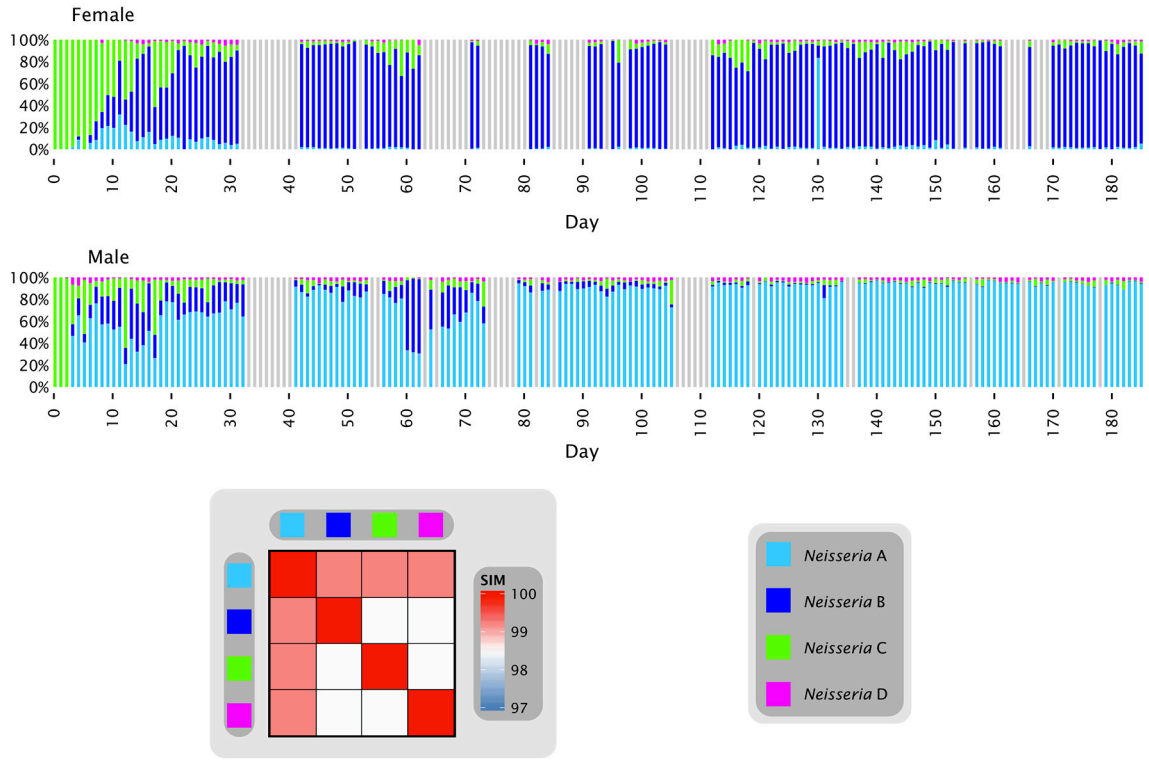


FIGURE 2 | *Neisseria* oligotypes in time series data from the tongue. Oligotypes with at least 0.5% mean abundance in at least one of the two individuals are shown. Colored bars represent the relative abundance of each of the four oligotypes in a single sample; data shown is for days 0 to 185 and gray bars represent days for which data is unavailable. The heat map shows the similarity of each pair of oligotypes. *Neisseria* A (oligo_001 in Table S1) is 100% identical to the HOMD reference sequence of *N. subflava*; *Neisseria* B (oligo_011) and *Neisseria* D (oligo_024) are each 99.2% identical to the *N. subflava* reference sequence; and *Neisseria* C (oligo_017) is identical to the reference sequences for *N. mucosa*, *N. flava*, *N. pharyngis*, *N. oralis*, and *Neisseria* spp. Human Oral Taxon (HOT) 015 and 018. All are within 98% identity of one another.

single day (Table S1). For comparison, the corresponding oligotypes identical to *S. mitis*, *S. oralis*, and *S. infantis* sampled from the tongue dorsum of multiple individuals from the HMP together ranged from 1 to approximately 90% of the *Streptococcus* genus on the tongue (**Figure 3B** and Eren et al., 2014). Thus, a substantial fraction of the range of variability observed across individuals was also observed within a single individual over time.

The proportions of *Neisseria* and *Streptococcus* can be seen in the context of other major tongue dorsum oligotypes in **Figure 4**. The major oligotypes shown in the figure each ranged from double-digit abundance to near-absence in samples over the course of the time series. The wide fluctuations in sample composition within an individual raised the question of the significance of differences between individuals compared to the variation that exists within an individual over time. OTU-level analysis of the tongue dorsum time course data showed that between-subject UniFrac distances were greater than within-subject distances (Caporaso et al., 2011), and likewise OTU-level analysis of HMP data showed between-subject differences within a body site greater than within-subject differences (Human Microbiome Project Consortium, 2012). Such inter-individual differences are also reflected in our oligotyping analysis, in the form of sometimes widely differential mean abundances of oligotypes between the two individuals, such as a greater abundance of *Neisseria* in the male and a greater abundance of several *Streptococcus* oligotypes in the female (**Figures 3**, **4**, and Table S1). These differences are concrete examples of the underlying taxon composition that leads to higher community dissimilarity scores between than within individuals. However, we wondered whether the summary statistic of average community dissimilarity was obscuring the magnitude of the shifts in community composition within individuals over time and giving a misleading impression about the relative importance of differences between and within individuals.

For a quantitative comparison of variation within and between subjects in these different studies, standard beta-diversity indices are not calculable because the studies analyzed different regions of the 16S rRNA gene and employed different amplification and sequencing protocols. However, some of the HMP subjects were sampled on more than one visit, affording the opportunity to assess the variability over time in these individuals compared to variation between subjects measured using the same study protocol. To make this comparison we identified 104 subjects for whom tongue dorsum samples from the V3-V5 region were available from two different visits, which were separated by 30–359 days (Aagaard et al., 2013). Reads in these samples had previously been trimmed and classified to genus using standard HMP pipelines (The Human Microbiome Project Consortium, 2012). Using data

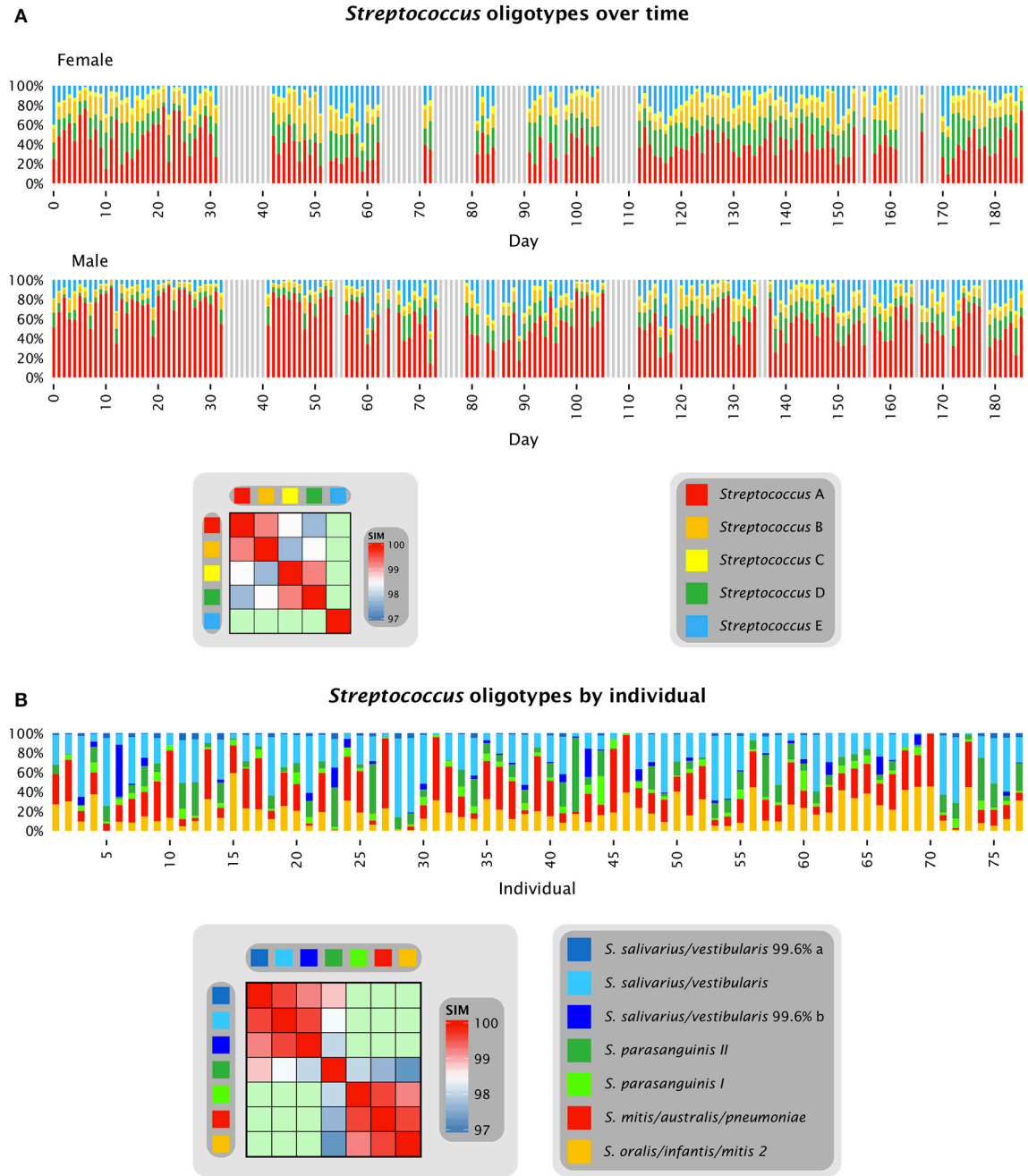


FIGURE 3 | *Streptococcus* oligotypes in the tongue. (A) Relative abundance of *Streptococcus* oligotypes in time series data from the tongue. Oligotypes with at least 0.5% mean abundance in at least one of the two individuals are shown. Colored bars represent the relative abundance of each of the five oligotypes in a single sample; data shown is for days 0 to 185 and gray bars represent days for which no data is available. The heat map shows the similarity of each pair of oligotypes. *Streptococcus* A (oligo_003 in Table S1) is identical to the reference sequences for 6 species in HOMD including *S. mitis*, *S. mitis* biovar 2, *S. infantis*, *S. oralis*, and *Streptococcus spp.* HOT 070 and 071; *Streptococcus* B (oligo_008) is identical to the reference sequences for 7 species in HOMD including *S. parasanguinis I*, *S. parasanguinis II*, *S. australis*, and *Streptococcus spp.* HOT 057, 065, 066, and 067; *Streptococcus* C (oligo_012) is identical to the reference sequences for *S. peroris* and *Streptococcus spp.* HOT 068 and 074; *Streptococcus* D (oligo_027) is identical to the reference sequences for *S. cristatus, S. gordonii, S. sinensis, S. oligofermentans,* and *Streptococcus spp.* HOT 056 and 069; and *Streptococcus* E (oligo_006) is identical to the reference sequences for *S. salivarius* and *S. vestibularis*. *Streptococcus* A, B, C, and D are all within 97% identity of one another as shown by the heat map. **(B)** Relative abundance of *Streptococcus* oligotypes in HMP data from the tongue, sequenced in the V1-V3 region. Oligotypes with at least 0.5% mean abundance across all sampled individuals are shown, and are assigned the name of the closest match in HOMD; where the closest match is not 100% identical, the percent identity is shown. In addition to the taxa listed in the key, the oligotype identified as *S. oralis/infantis/mitis biovar 2* is also identical to *Streptococcus spp.* HOT 055, 058, 061, and 070 and the oligotype identified as *S. mitis/australis/pneumoniae* is also identical to *Streptococcus spp.* HOT 070, 071, and 074. Data for **(B)** are from Eren et al. (2014).

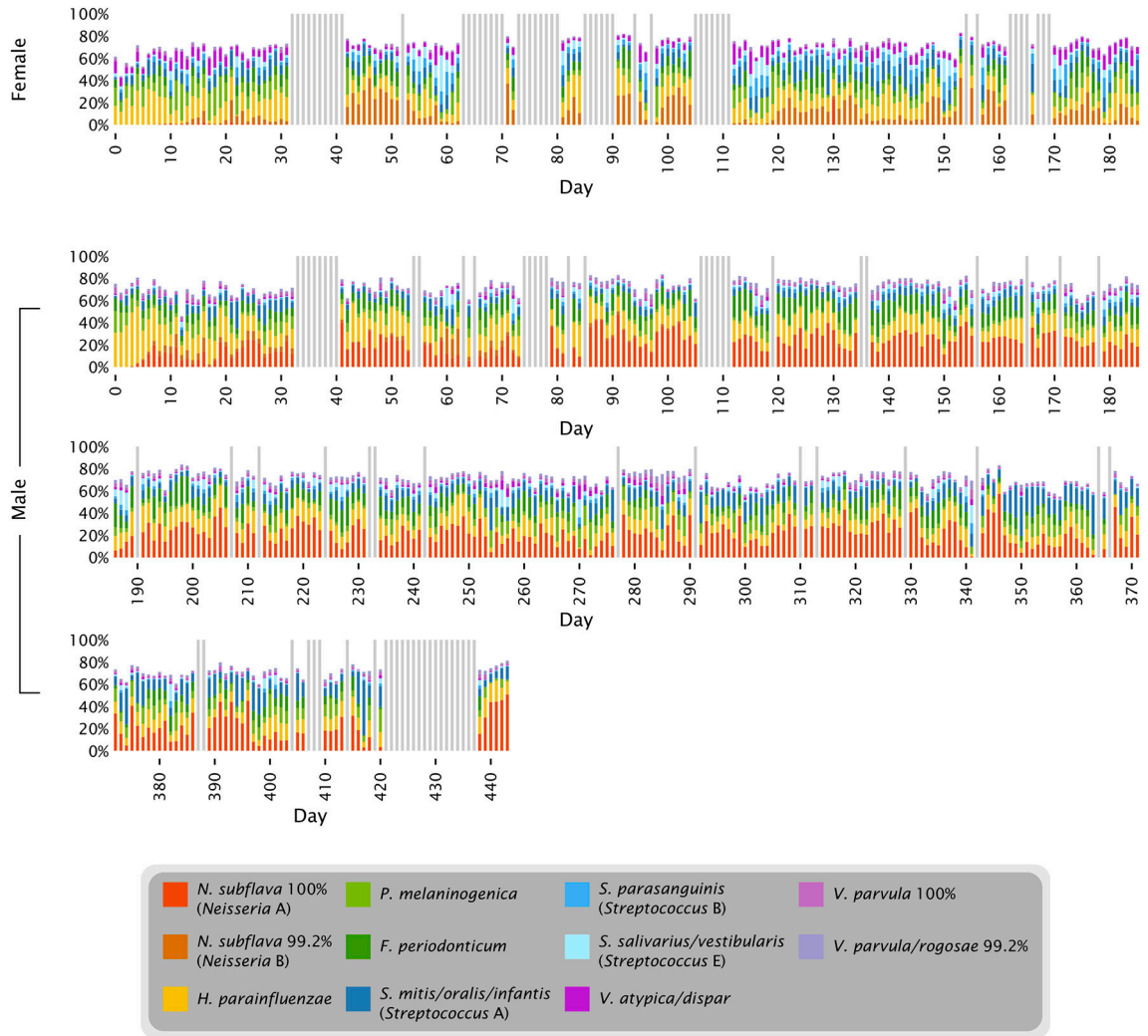


FIGURE 4 | Time series of abundant oligotypes. The 11 oligotypes shown include the 8 most abundant in each subject; 5 oligotypes are in the top 8 in both subjects. Colored bars represent the abundance of each oligotype in each sample; gray bars represent days for which no data is available.

on the number of reads classified into each genus for each of these samples, we carried out a cluster analysis using the Morisita-Horn dissimilarity index. **Figure 5** shows the resulting clusters. For each of the 104 subjects, the two samples from different time points are connected by an arc. As can be seen in the figure, for some subjects the two samples from different time points cluster tightly together (short arcs), but for many subjects the two samples are located in different clusters (long arcs). These clusters can be related back to the taxon composition of each sample; for example, the cluster colored in light blue consists of samples that are more than 50% *Streptococcus* while samples in the cluster shown in red have a high proportion of *Fusobacterium*. This analysis supports the conclusion that many of the apparent microbiome differences between individuals seen in the HMP data are a result not of stable differences from person to person, but of the limited information that results from "snapshot" sampling a continuously changing system (an individual tongue) at a single time point.

CORRELATED ABUNDANCE BETWEEN MEMBERS OF DIFFERENT GENERA

The time series abundance data permit an assessment of the degree of correlation or anti-correlation in the abundance of individual oligotypes. Such an assessment would provide a basis for inferring significant biological associations of taxa. Remarkably, the data showed strong correlations between pairs of oligotypes both within a taxon and across taxa (Figure S1).

The strongest correlation was between two oligotypes that are among the 10 most abundant in the dataset and whose best match in HOMD is to the same taxon, *Veillonella parvula*. One, oligo_007, is identical to the *V. parvula* reference sequence and the other, oligo_009, differs from it by a single nucleotide (**Figure 6A**). The strength of their correlation suggests either that they are in an extraordinarily close symbiosis or that they represent two distinct rRNA genes present in the same cell. One advantage of the oral microbiome as a study subject is the presence of sequenced genomes for a high fraction of oral microbial

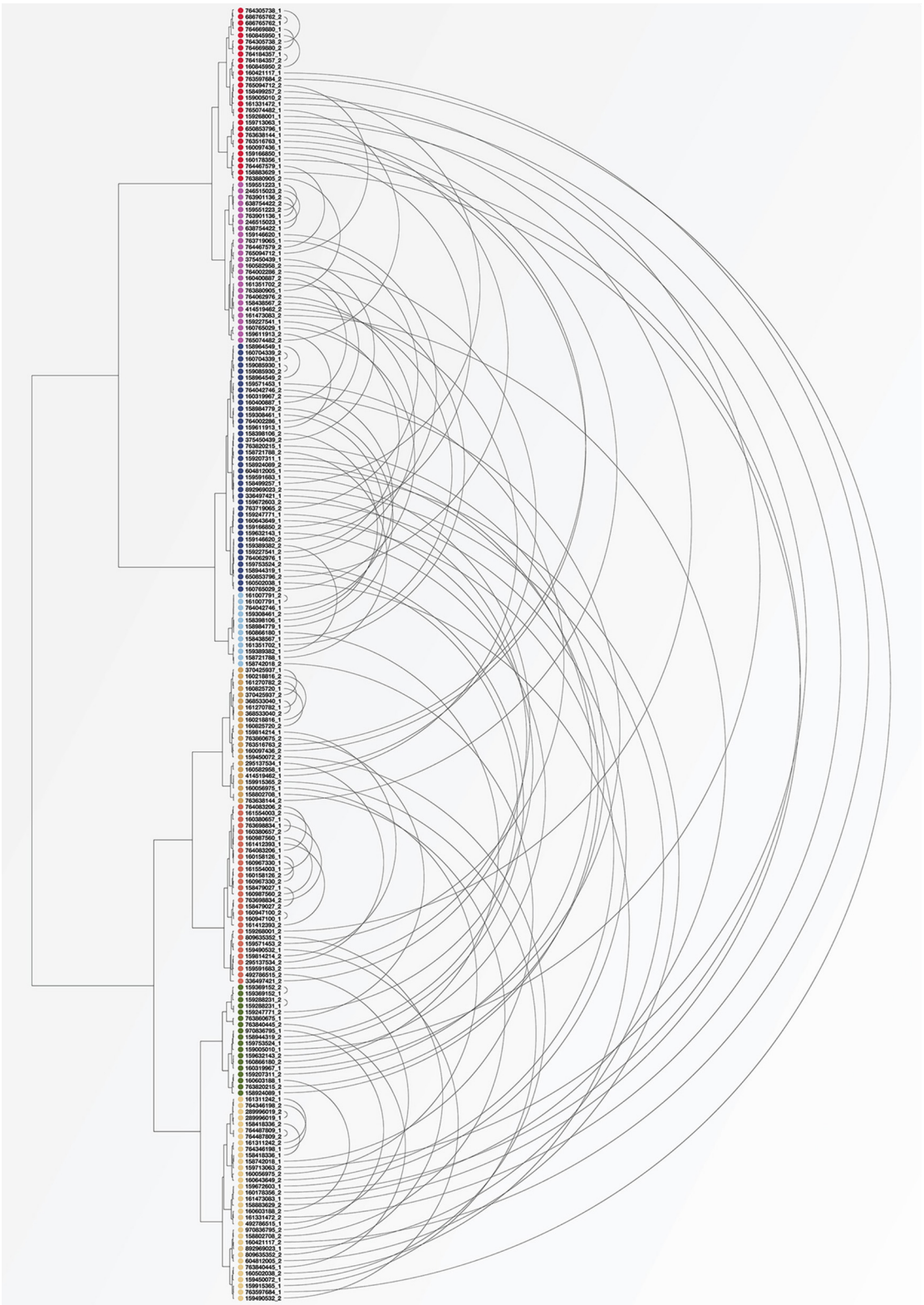


FIGURE 5 | Cluster analysis of individuals sampled by HMP at two visits. Each dot represents a tongue dorsum sample from one of 104 individuals sampled at two visits at least 30 days apart. Samples were classified to genus and clustered using the Morisita-Horn dissimilarity index. Arcs connect the two samples from each subject. Short arcs indicate subjects whose community composition was similar at the two visits; long arcs indicate subjects whose second sample was substantially different in composition from the first.

taxa (Dewhirst et al., 2010), allowing a direct test of the possibility that any two given rRNA genes are present in a single organism. We carried out a BLAST search of genomic DNA using the HOMD web site (HOMD.org) and found both *V. parvula* oligotypes in the sequenced genome of *V. parvula* DSM 2008/ATCC 10790. We conclude that the two tightly correlated *V. parvula* oligotypes represent two sequences found in the same organism.

In contrast to the *V. parvula* oligotypes, another strongly-correlated pair of oligotypes (oligo_024 and oligo_030) represent species in different taxonomic families: one member of the pair differs by a single nucleotide from the *Haemophilus parainfluenzae* reference and the other differs by a single nucleotide from the *Neisseria subflava* reference (**Figure 6B**). Partially or completely sequenced genomes are available for *N. subflava* as well as the related taxa *N. flavescens* and *N. mucosa*, and for *H. parainfluenzae* as well as the related *H. influenzae* and *H. haemolyticus*, among others. BLAST searches revealed that the *H. parainfluenzae* oligotype oligo_030 is no more than 87% identical to any region of any sequenced *Neisseria* genome, while the reverse is true for the *N. subflava* oligotype oligo_024: it is no more than 87% identical to any sequenced *Haemophilus* genome. We conclude that these two oligotypes reside in different organisms, and their strong correlation reflects either a close symbiotic interaction between them, or strong specialization of both organisms to the same micro-habitat.

The abundance traces of the two pairs of oligotypes shown in **Figures 6A,B** are nearly identical to those obtained by other investigators who analyzed the same dataset using an entirely different method aimed at identifying biologically meaningful units with single-nucleotide resolution (Tikhonov et al., 2014). This similarity supports the general validity of both methods. However, we reach opposite conclusions concerning which of these pairs is made up of sequences present in the same genome and which are in different genomes. We conclude based on whole-genome sequences that the two *Veillonella* sequences are in the same genome and that the *Haemophilus* and *Neisseria* sequences are in different cells. Tikhonov et al. confined their analysis to the sequences *per se*. Based on autocorrelation coefficients, they concluded that the two sequences which we identify as *Veillonella* are at least partially contributed by different cells and that the sequences we identify as *Haemophilus* and *Neisseria* likely originate from the same cells. We believe our conclusions benefit from the genome-mining and cross-referencing to HOMD, but future work is necessary to determine which conclusion is correct.

The most abundant oligotypes in the tongue time series dataset are not strongly correlated with one another. For example, the two most abundant oligotypes in the dataset, which are identical to the HOMD reference sequences for *N. subflava* and *H. parainfluenzae*, each make up more than 10% of the entire dataset and have abundance distributions that are weakly correlated with each other (**Figure 6C** and Figure S1). The weak correlation of these highly abundant oligotypes contrasts with the tight correlation of their lower-abundance variants discussed above and suggests differences in the underlying biology of the high- and low-abundance types. Possibly, the high-abundance oligotypes represent generalist organisms that do not require specialized habitat or tight taxon-taxon associations. Alternatively, the more abundant oligotypes may encompass a heterogeneous collection of organisms with identical V4 regions of the 16S rRNA gene but with distinctive habitat requirements.

Additional, moderately positive correlations exist among pairs of oligotypes from widely different taxa such as *Streptococcus*, *Haemophilus*, and *Alloprevotella* (Figure S1) and likely result from a preference for similar habitats or environmental conditions. In contrast, the fourth and fifth most abundant oligotypes overall, whose sequences are identical to the *Fusobacterium periodonticum* and *Prevotella melaninogenica* reference sequences, are moderately anticorrelated (**Figure 6D**); this anticorrelation could result from an active antagonism between two taxa or from a preference for incompatible microhabitats. In sum, correlation analysis of the time series data provides strong indications of possible functional or habitat associations among diverse taxa.

ARE THE FLUCTUATIONS IN OLIGOTYPE ABUNDANCE PERIODIC?

Casual inspection of the time series data gives the impression that the oligotype fluctuations could be periodic. One possible hypothesis for periodic variation in the composition of the tongue microbiome is a periodic variation in host activity such as might occur over weekends as opposed to the workweek. We tested for reproducible periodicity in the data by carrying out auto-correlation and Fourier transform analysis for each oligotype. Auto-correlations were evaluated over a window of plus or minus 21 days. Consistent with the observation of rapid fluctuations, the auto-correlation signal was strongest for a one-day time lag, which agrees with the results of Tikhonov et al. (2014). However, no consistent signal was observed for any of the abundant oligotypes either with auto-correlation or with Fourier analysis that would suggest a weekly or other periodicity. A few minor oligotypes showed a weak signal corresponding to weekly periodicity but the signal was not of sufficient magnitude to admit of a strong conclusion. Proper evaluation of such a possibility will require a directed investigation.

A CHARACTERISTIC TONGUE MICROBIOTA

Oligotyping three datasets from the tongue (one time course and two broad samplings of individuals) showed that a limited number of species-level or near-species-level taxa consistently make up the majority of the microbiota on the tongue. Detailed taxonomic comparison of oligotypes across these datasets is not straightforward, because different regions of the 16S rRNA gene were sequenced in each case: V4 for the time course data and V1-V3 and V3-V5 for the HMP data. Nonetheless, taxonomic assessments can be made by comparing each sequence to a curated reference database, the HOMD, and using the matching reference sequence(s) as an estimate of the taxonomy of the oligotype. Twenty such reference sequences, or groups of closely related reference sequences, collectively account for 91–93% of the reads from each dataset (**Figure 7**). Eighteen of these 20 were detected in every sample or nearly every sample from both individuals in the time series (Table S1). Thus, while the temporal core microbiome in this dataset is composed of only a small fraction of the taxa that are detectable (Caporaso et al., 2011), this temporal core nonetheless constitutes the majority of the organisms on the tongue.

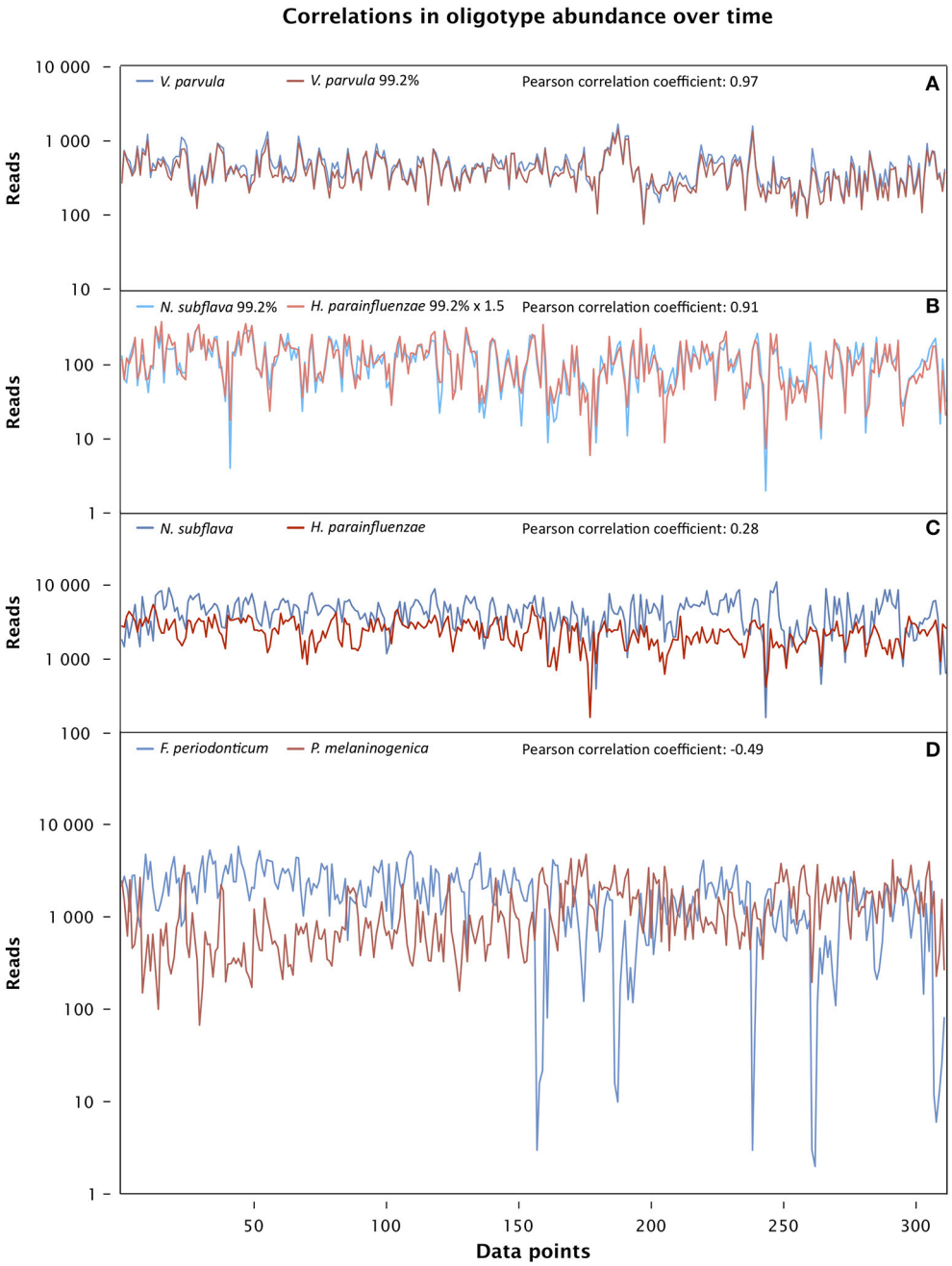


FIGURE 6 | Time series correlation analysis. The abundance of each oligotype, measured in reads, is plotted for each sample from day 66 through 420 in the male subject, and the Pearson correlation coefficient for each pair of oligotypes over this data range is shown above the plot. **(A)** Two strongly correlated oligotypes matching the taxon *Veillonella parvula* (oligo_007 and oligo_009 in Table S1) ($p << 0.0001$). **(B)** Strongly correlated oligotypes 99.2% identical to *Neisseria subflava* (oligo_024) and *Haemophilus parainfluenzae* (oligo_030) ($p << 0.0001$). **(C)** Two oligotypes identical to *Neisseria subflava* (oligo_001) and *Haemophilus parainfluenzae* (oligo_002) are weakly correlated (Pearson, $p = 0.18$ after Bonferroni correction; Spearman, $p < 0.01$ after Bonferroni correction). **(D)** Oligotypes identical to *Fusobacterium periodonticum* (oligo_004) and *Prevotella melaninongenica* (oligo_005), showing moderate anti-correlation ($p << 0.0001$).

In contrast to these similarities, there are also differences among the abundant taxa present in the time series compared to the HMP data. Several taxa are relatively depauperate in the time series data set compared to HMP, including *Actinomyces spp.*, *Leptotrichia spp.*, and *Porphyromonas sp.* HOT 279; these differences may reflect true characteristics of the microbiomes of the sampled individuals, or may result from primer bias or other technical differences in experimental procedures. The genus *Rothia* is represented in all three datasets but the oligotypes representing this genus do not match the same

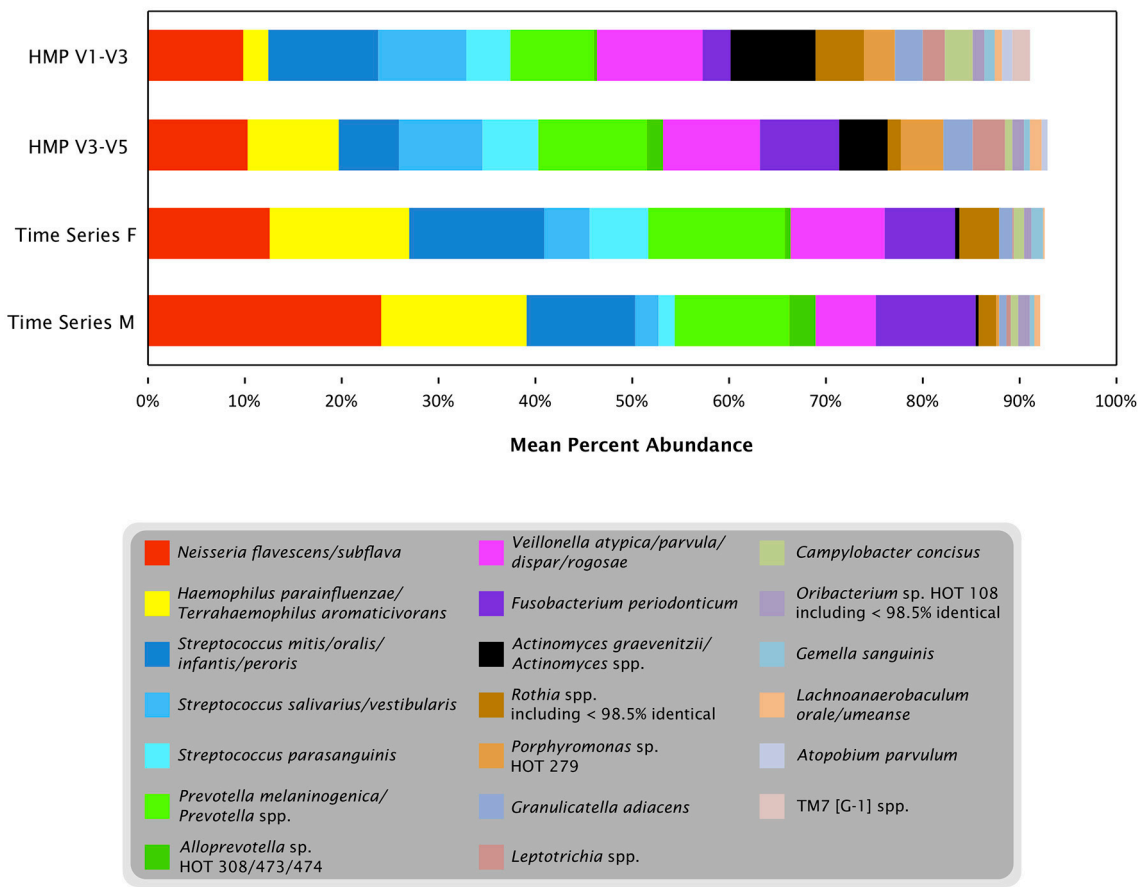


FIGURE 7 | The characteristic microbiota of the tongue. Colored bars represent the mean abundance of the taxa that consistently occur in the time series and HMP datasets. For each of the taxa shown, the HOMD reference sequence for the taxon is at least 98.5% identical to one or more oligotypes that make up at least 0.5% of the time series tongue data (Female, Male) and/or the V3-V5 and V1-V3 tongue dorsum data from the HMP. Exceptions to the 98.5% identity criterion (*Rothia* spp. and *Oribacterium* sp. HOT 108) indicate cases where a precisely matched reference sequence may not yet be represented in HOMD.

species consistently across datasets. This inconsistency may be explained by technical or biological differences, but an alternative possibility is that sequences from this genus represent one or a few taxa that are consistently present across datasets but have 16S rRNA gene sequences divergent from the reference sequences currently represented in HOMD. For such a taxon the closest match in the V1-V3 region may be to one reference sequence; in the V4 region its closest match may be to a different reference sequence, and these differences in taxonomic assignment in the different regions may obscure the consistency with which the identical taxon is present across datasets.

DISCUSSION

MICROBIAL COMMUNITY DYNAMICS FOLLOWED USING OLIGOTYPING

Understanding the forces that shape microbial communities in the human microbiome requires following dynamic changes in these communities over time. Rapid decreases in the cost of DNA sequencing have made it possible to generate the large amounts of data required for studies of dynamics, but analysis methods limited to the genus or OTU level have limited the opportunities for analyzing the dynamics within a single species or between closely related species. This study provides an example of the single-nucleotide taxonomic resolution of oligotyping which, in turn, enables analysis of microbial dynamics and associations that would otherwise not be possible if taxa were lumped into heterogeneous groups.

PHASE TRANSITIONS OF OLIGOTYPES

Our observation of changing relative abundance of *Neisseria* oligotypes on the tongues of two different individuals showed that in these instances, replacement of an initially dominant oligotype occurred over a time scale of days, and the newly dominant type remained dominant for the rest of the months-long sampling period. Thus the period of transition was relatively abrupt

in comparison to the duration of the subsequent dominant phase. The causes both of the replacement, and of the stable dominance, remain uncertain. After the first few days of sampling the two oligotypes that became dominant were different in the two individuals but were detected in nearly every sample from each of them. It is possible that these two oligotypes newly invaded the tongue habitat of these individuals near the beginning of the time course and, once present, proliferated in what was for them a favorable environment. Alternatively, it is possible that they were present but simply below the detection limit for the first few days, and their sudden proliferation was caused by changes in the oral environment or the surrounding microbiota, changes perhaps occasioned by the daily sampling itself. In both individuals the oligotype that was not dominant nevertheless persisted in low abundance, showing that (unsurprisingly for the oral environment) dispersal is not the limiting factor regulating the abundance of these taxa in a given mouth. The dynamics displayed by these oligotypes are similar to the behavior of some closely-related 97% OTUs in a time series of gut and saliva samples from two individuals (David et al., 2014), in which rapid transitions are followed by extended periods of stable dominance of one of the OTUs. A similar pattern was also observed in the within-species dynamics of *Staphylococcus epidermidis* in a time series from the gut microbiome of a premature infant (Sharon et al., 2013) in which the changes in strain abundance were at least partially attributable to the dynamics of infecting bacteriophage. The extended dominance periods we observe are difficult to explain as a consequence of phage-driven dynamics, however, unless one invokes development of host resistance or changes in phage infectivity (Sharon et al., 2013) or the presence of multiple strains that have different virus sensitivities and that are succeeding one another, but which are indistinguishable in 16S rRNA gene sequence (Fuhrman, 2009) and thus undetectable with this data.

IMPLICATIONS OF HIGH VARIABILITY IN TAXON RELATIVE ABUNDANCE OVER TIME

The high variability and rapid change in microbial communities in the time series data set were noted by Caporaso et al. (2011) as well as the contribution of blooms of particular genus-level taxa to the dissimilarity of the overall community over time. Our oligotyping results extend these findings to the species- or near-species level, as shown in the example of *Streptococcus* in which dramatic changes occur in the relative, as well as the absolute, abundance of each oligotype as a proportion of the genus abundance over time. From our analysis of the HMP data for the *Streptococcus* community of many individuals at a single time point, it was evident that a number of major *Streptococcus* taxa were present in every individual; however it was not possible to determine whether their abundances fluctuated over time or whether communities in some individuals were strongly and continually biased in favor of one or another taxon. Our results with the time-series data for the tongue dorsum suggest that a substantial portion of the variation in taxon abundance occurring between individuals in the HMP data can be explained by the temporal variation of abundance within individuals.

This high variability has implications for the fine-scale spatial and metabolic structure of the tongue flora. Given our observation of a consistent, characteristic tongue dorsum microbiota over time and across individuals, one could hypothesize that these taxa comprise a tightly integrated community with finely tuned metabolic interactions with one another and with cells of different microbial species intimately intermingled at micron scales in a relatively constant stoichiometry. The high overall variability in relative abundance among these taxa, however, argues against such a hypothesis. Rather, the microbiota likely constitute a number of distinct assemblages occupying different spatial positions, preferring different environments, or succeeding one another over time. Certain subsets of the assemblage that show correlated distribution, such as the oligotypes identified with *H. parainfluenzae* and members of the *N. subflava* group, may constitute a functional unit. Other anti-correlated subsets, however, such as the oligotypes identified with *F. periodonticum* and *P. melaninogenica*, may reflect that the corresponding taxa interact in an antagonistic fashion or that they prefer different environmental conditions.

The reasons underlying the large fluctuations in relative abundance across taxa are an interesting question for further study. Disturbances caused by oral hygiene procedures and ingestion of food or liquids occur with higher frequency than the observed community fluctuations and are unlikely to be the sole driver of these fluctuations. For an assemblage residing on a shedding epithelial surface, the sporadic availability of new surfaces for colonization may give a temporary advantage to taxa that are more effective initial colonizers or may be, by chance, spatially well-positioned to colonize new habitat. Alternate explanations include changes in activity of the host immune system, diurnal physiological changes, the dynamics of bacteriophage populations, competition, or stochastic variation. Sporadic changes in host behavior may also be responsible.

THE USE OF HOMD TO CONNECT ORAL OLIGOTYPE DATASETS

Short regions of the rRNA gene have limitations for high-resolution identification and differentiation of microbes. Potential confusion arises when taxa of interest are differentiated by only one or a few nucleotides in the sequenced region, but these limitations can be mitigated by making use of taxonomic information to relate distinct datasets to one another.

An example in the data shown here is the time series oligotype labeled *Streptococcus* D (**Figure 3A**). This oligotype is identical in the V4 region to the HOMD reference sequence for *S. gordonii* but is also only a single nucleotide different from the HOMD reference sequence for *Streptococcus parasanguinis*. Additional information about the likely taxonomy of this oligotype comes from the HMP datasets from V1-V3 and V3-V5; neither of these datasets shows a significant contribution of *S. gordonii* to the tongue microbiota, while both show a substantial contribution of *S. parasanguinis* (**Figure 3B**). Evaluation of the time series data in the context of the HMP data therefore suggests that *Streptococcus* D is more accurately identified as a variant of *S. parasanguinis*. Similar considerations apply to the *Neisseria*

oligotypes (**Figure 1**). The species *N. flavescens*, *N. subflava*, and *N. flava* form a phylogenetically distinct group according to whole-genome sequence data (Bennett et al., 2013) and are shown by HMP data to be important in the tongue microbiota (Eren et al., 2014). In the V4 region there are only 1 or 2 nucleotide differences between these taxa, leading to ambiguity in identification of the time series oligotypes in the absence of additional information. This information can be found in HMP data: using V1-V3 sequences, the abundant oligotypes of this group are unequivocally identified as most similar to *N. flavescens*, which differs from *N. flava* and *N. subflava* by 11 nucleotides in the oligotyped region of V3. These two examples demonstrate the power of a well-curated database and applying multiple lines of evidence to the identification of taxa.

THE CORE TONGUE MICROBIOME

With the species-level description of its consistent core microbiome that we present here, the tongue becomes one of only a small number of habitats for which a numerically abundant core microbiome has been described at the species level. Our results support the conclusions of Kraal et al. (2014) who analyzed whole-genome shotgun samples from the HMP and concluded that the species *Veillonella dispar* was abundant in every tongue microbiome sampled and that three other species (*S. parasanguinis*, *S. salivarius*, and *P. melaninogenica*) each were abundant in at least 87% of tongues sampled. Given the close similarity of the microbiomes of the tongue and of saliva (Mager et al., 2003; Eren et al., 2014) it is not surprising that the set of genus-level taxa detected in all or nearly all saliva samples by Stahringer et al. (2012) is also concordant with our set of core taxa, as is the set of genera found in all saliva samples by Lazarevic et al. (2010). The presence of a consistent core tongue microbiota argues against the idea that many functions in the overall oral microbial community can be carried out by any one of a number of interchangeable taxa, and argues instead for the presence of niche specialists whose role is not readily filled by alternative taxa (Fuhrman, 2009). The relative simplicity of the core tongue microbiota contrasts with the hundreds of taxa that are described from the mouth as a whole (Dewhirst et al., 2010), many of which are specialized to a subset of habitats within the mouth (Eren et al., 2014). It may be a general characteristic of microbial ecosystems to appear enormously complicated when considered at spatial scales that lump together disparate habitats, but to resolve into more tractable communities when the habitat is accurately and narrowly defined.

MAKING FULL USE OF THE INFORMATION IN HIGH-THROUGHPUT SEQUENCING DATA SETS

There is a growing recognition that high-throughput sequencing data contains information that is not fully expressed by partitioning the data into conventional OTUs. Some form of partitioning is necessary because both neutral variation in natural populations and sequencing errors create a profusion of sequence variants without underlying biological meaning. However, OTUs that are defined purely by a threshold of sequence similarity are phylogenetically and ecologically heterogeneous and inconsistent (Prosser et al., 2007; Schloss and Westcott, 2011; Koeppel and Wu, 2013; Schmidt et al., 2014). Alternative approaches make use of the fact that the noise arising from neutral variation and sequencing errors is randomly distributed with respect to ecology. For example, an approach termed "distribution-based clustering" employs information about the distribution of sequences among habitats or samples to differentiate noise from meaningful variation and thus inform the definition of taxonomic units (Preheim et al., 2013). In a "denoising" approach (Tikhonov et al., 2014), sequencing error and temporal cross-correlation were analytically distinct but temporal cross-correlation analysis was used to determine which unique sequences were "real," i.e., not attributable to noise.

Oligotyping is an information theory-based approach that employs Shannon entropy to identify nucleotide positions of high variation within a dataset (Eren et al., 2013), thereby distinguishing meaningful variation from sequencing errors (Huse et al., 2007; Minoche et al., 2011). Like the cross-correlation approaches, the Shannon entropy method has the capacity to discriminate among closely related taxa at the sub-species level. However, unlike these other approaches, the Shannon entropy method partitions the data into oligotypes independent of cross-sample correlations. This independence means that habitat or temporal correlation analysis can be employed at a later stage in data analysis, providing an independent way of assessing the biological meaning and distinctiveness of sequence variants.

For the human oral microbiome, the presence of a highly curated database and a large number of sequenced genomes provides an additional layer of analytic power. Sequence differences that rise above the level of noise, as identified by oligotyping or cross-correlation, can be associated with known taxa via the HOMD, allowing the comparison of data across datasets even when different regions of the 16S rRNA gene were employed for sequencing. Distinguishing whether oligotypes represent different 16S rRNA genes within a single organism or are tags for different organisms is enabled by access to full genomes. This cross-dataset analysis and genome-mining capability greatly expands the usefulness of datasets. In summary, we have used high-resolution taxonomic analysis of high-throughput time series data to provide insight into the microbial population dynamics of the tongue. Our results have revealed phase transitions of closely related taxa and unanticipated associations of taxa from different genera. We expect that our approach will permit future, targeted analyses of specific microbial interactions and dynamics.

AUTHOR CONTRIBUTIONS

Jessica L. Mark Welch, A. Murat Eren, and Gary G. Borisy conceived and designed the work; Jessica L. Mark Welch, Daniel R. Utter, Blair J. Rossetti, David B. Mark Welch, and A. Murat Eren analyzed data; Jessica L. Mark Welch and Gary G. Borisy wrote the paper; and all authors reviewed, edited, and approved the final manuscript.

ACKNOWLEDGMENTS

We thank two reviewers for their careful reading and comments, which greatly improved the manuscript. Supported by

National Institutes of Health (NIH) National Institute of Dental and Craniofacial Research Grant DE022586 (to Gary G. Borisy). Daniel R. Utter was supported by the Woods Hole Partnership Education Program; A. Murat Eren was supported by a G. Unger Vetlesen Foundation grant to the Marine Biological Laboratory; David B. Mark Welch was supported by NSF DBI-1262592.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2014.00568/abstract

REFERENCES

Aagaard, K., Petrosino, J., Keitel, W., Watson, M., Katancik, J., Garcia, N., et al. (2013). The Human Microbiome Project strategy for comprehensive sampling of the human microbiome and why it matters. *FASEB J.* 27, 1012–1022. doi: 10.1096/fj.12-220806

Aas, J. A., Paster, B. J., Stokes, L. N., Olsen, I., and Dewhirst, F. E. (2005). Defining the normal bacterial flora of the oral cavity. *J. Clin. Microbiol.* 43, 5721–5732. doi: 10.1128/JCM.43.11.5721

Bennett, J. S., Jolley, K. A., and Maiden, M. C. J. (2013). Genome sequence analyses show that Neisseria oralis is the same species as "Neisseria mucosa var. *heidelbergensis*." *Int. J. Syst. Evol. Microbiol.* 63, 3920–3926. doi: 10.1099/ijs.0.052431-0

Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., et al. (2011). Moving pictures of the human microbiome. *Genome Biol.* 12:R50. doi: 10.1186/gb-2011-12-5-r50

Chow, C.-E. T., Sachdeva, R., Cram, J. A., Steele, J. A., Needham, D. M., Patel, A., et al. (2013). Temporal variability and coherence of euphotic zone bacterial communities over a decade in the Southern California Bight. *ISME J.* 7, 2259–2273. doi: 10.1038/ismej.2013.122

David, L. A., Materna, A. C., Friedman, J., Campos-Baptista, M. I., Blackburn, M. C., Perrotta, A., et al. (2014). Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* 15:R89. doi: 10.1186/gb-2014-15-7-r89

Dethlefsen, L., Huse, S., Sogin, M. L., and Relman, D. A. (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol.* 6:e280. doi: 10.1371/journal.pbio.0060280

Dethlefsen, L., and Relman, D. A. (2011). Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl. 1), 4554–4561. doi: 10.1073/pnas.1000087107

Dewhirst, F. E., Chen, T., Izard, J., Paster, B. J., Tanner, A. C. R., Yu, W.-H., et al. (2010). The human oral microbiome. *J. Bacteriol.* 192, 5002–5017. doi: 10.1128/JB.00542-10

Eren, A. M., Borisy, G. G., Huse, S. M., and Mark Welch, J. L. (2014). Oligotyping analysis of the human oral microbiome. *Proc. Natl. Acad. Sci. U.S.A.* 111, E2875–E2884. doi: 10.1073/pnas.1409644111

Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., et al. (2013). Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* 4, 1111–1119. doi: 10.1111/2041-210X.12114

Fuhrman, J. A., Hewson, I., Schwalbach, M. S., Steele, J. A., Brown, M. V., and Naeem, S. (2006). Annually reoccurring bacterial communities are predictable from ocean conditions. *Proc. Natl. Acad. Sci. U.S.A.* 103, 13104–13109. doi: 10.1073/pnas.0602399103

Fuhrman, J. A. (2009). Microbial community structure and its functional implications. *Nature* 459, 193–199. doi: 10.1038/nature08058

Gajer, P., Brotman, R. M., Bai, G., Sakamoto, J., Schütte, U. M. E., Zhong, X., et al. (2012). Temporal dynamics of the human vaginal microbiota. *Sci. Transl. Med.* 4, 132ra52. doi: 10.1126/scitranslmed.3003605

Gilbert, J. A., Steele, J. A., Caporaso, J. G., Steinbrück, L., Reeder, J., Temperton, B., et al. (2012). Defining seasonal marine microbial community dynamics. *ISME J.* 6, 298–308. doi: 10.1038/ismej.2011.107

Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Huse, S. M., Dethlefsen, L., Huber, J. A., Mark Welch, D., Relman, D. A., and Sogin, M. L. (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet.* 4:e1000255. doi: 10.1371/journal.pgen.1000255

Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., and Mark Welch, D. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* 8:R143. doi: 10.1186/gb-2007-8-7-r143

Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., et al. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl.), 4578–4585. doi: 10.1073/pnas.1000081107

Koeppel, A. F., and Wu, M. (2013). Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. *Nucleic Acids Res.* 41, 5175–5188. doi: 10.1093/nar/gkt241

Kraal, L., Abubucker, S., Kota, K., Fischbach, M. A., and Mitreva, M. (2014). The prevalence of species and strains in the human microbiome: a resource for experimental efforts. *PLoS ONE* 9:e97279. doi: 10.1371/journal.pone.0097279

Lazarevic, V., Whiteson, K., Hernandez, D., François, P., and Schrenzel, J. (2010). Study of inter- and intra-individual variations in the salivary microbiota. *BMC Genomics* 11:523. doi: 10.1186/1471-2164-11-523

Mager, D. L., Ximenez-Fyvie, L. A., Haffajee, A. D., and Socransky, S. S. (2003). Distribution of selected bacterial species on intraoral surfaces. *J. Clin. Periodontol.* 30, 644–654. doi: 10.1034/j.1600-051X.2003.00376.x

Martínez, I., Muller, C. E., and Walter, J. (2013). Long-term temporal analysis of the human fecal microbiota revealed a stable core of dominant bacterial species. *PLoS ONE* 8:e69621. doi: 10.1371/journal.pone.0069621

Minoche, A. E., Dohm, J. C., and Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol.* 12:R112. doi: 10.1186/gb-2011-12-11-r112

Needham, D. M., Chow, C.-E. T., Cram, J. A., Sachdeva, R., Parada, A., and Fuhrman, J. A. (2013). Short-term observations of marine bacterial and viral communities: patterns, connections and resilience. *ISME J.* 7, 1274–1285. doi: 10.1038/ismej.2013.19

Preheim, S. P., Perrotta, A. R., Martin-Platero, A. M., Gupta, A., and Alm, E. J. (2013). Distribution-based clustering: using ecology to refine the operational taxonomic unit. *Appl. Environ. Microbiol.* 79, 6593–6603. doi: 10.1128/AEM.00342-13

Prosser, J. I., Bohannan, B. J. M., Curtis, T. P., Ellis, R. J., Firestone, M. K., Freckleton, R. P., et al. (2007). The role of ecological theory in microbial ecology. *Nat. Rev. Microbiol.* 5, 384–392. doi: 10.1038/nrmicro1643

Schloss, P. D., Gevers, D., and Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE* 6:e27310. doi: 10.1371/journal.pone.0027310

Schloss, P. D., and Westcott, S. L. (2011). Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl. Environ. Microbiol.* 77, 3219–3226. doi: 10.1128/AEM.02810-10

Schmidt, T. S. B., Matias Rodrigues, J. F., and von Mering, C. (2014). Ecological consistency of SSU rRNA-based operational taxonomic units at a global scale. *PLoS Comput. Biol.* 10:e1003594. doi: 10.1371/journal.pcbi.1003594

Segata, N., Haake, S. K., Mannon, P., Lemon, K. P., Waldron, L., Gevers, D., et al. (2012). Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol.* 13:R42. doi: 10.1186/gb-2012-13-6-r42

Sharon, I., Morowitz, M. J., Thomas, B. C., Costello, E. K., Relman, D. A., and Banfield, J. F. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.* 23, 111–120. doi: 10.1101/gr.142315.112

Socransky, S. S., and Haffajee, A. D. (2005). Periodontal microbial ecology. *Periodontol. 2000* 38, 135–187. doi: 10.1111/j.1600-0757.2005.00107.x

Stahringer, S. S., Clemente, J. C., Corley, R. P., Hewitt, J., Knights, D., Walters, W., et al. (2012). Nurture trumps nature in a longitudinal survey of salivary bacterial communities in twins from early adolescence to early adulthood. *Genome Res.* 22, 2146–2152. doi: 10.1101/gr.140608.112

The Human Microbiome Project Consortium. (2012). A framework for human microbiome research. *Nature* 486, 215–221. doi: 10.1038/nature 11209

Tikhonov, M., Leach, R. W., and Wingreen, N. S. (2014). Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *ISME J.* doi: 10.1038/ismej.2014.117. [Epub ahead of print].

Zaura, E., Keijser, B. J. F., Huse, S. M., and Crielaard, W. (2009). Defining the healthy "core microbiome" of oral microbial communities. *BMC Microbiol.* 9:259. doi: 10.1186/1471-2180-9-259

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 07 August 2014; accepted: 08 October 2014; published online: 07 November 2014.

Citation: Mark Welch JL, Utter DR, Rossetti BJ, Mark Welch DB, Eren AM and Borisy GG (2014) Dynamics of tongue microbial communities with single-nucleotide resolution using oligotyping. Front. Microbiol. ***5****:568. doi: 10.3389/fmicb.2014.00568*

This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology.

frontiers in
MICROBIOLOGY

METHODS ARTICLE
published: 14 November 2014
doi: 10.3389/fmicb.2014.00601

The R package *otu2ot* for implementing the entropy decomposition of nucleotide variation in sequence data

Alban Ramette[1]* and Pier Luigi Buttigieg[2,3]

[1] HGF-MPG Group for Deep Sea Ecology and Technology, Max Planck Institute for Marine Microbiology, Bremen, Germany
[2] Organic Geochemistry Department, MARUM - Center for Marine Environmental Sciences, Bremen, Germany
[3] HGF-MPG Group for Deep Sea Ecology and Technology, Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, Bremerhaven, Germany

Edited by:
A. Murat Eren, Marine Biological Laboratory, USA

Reviewed by:
Lois Maignien, University of Western Brittany, France
Christopher Quince, University of Warwick, UK

***Correspondence:**
Alban Ramette, Institute of Social and Preventive Medicine (ISPM), University of Bern, Finkenhubelweg 11, 3012 Bern, Switzerland
e-mail: ramette@ispm.unibe.ch

Oligotyping is a novel, supervised computational method that classifies closely related sequences into "oligotypes" (OTs) based on subtle nucleotide variation (Eren et al., 2013). Its application to microbial datasets has helped reveal ecological patterns which are often hidden by the way sequence data are currently clustered to define operational taxonomic units (OTUs). Here, we implemented the OT entropy decomposition procedure and its unsupervised version, Minimal Entropy Decomposition (MED; Eren et al., 2014c), in the statistical programming language and environment, R. The aim of this implementation is to facilitate the integration of computational routines, interactive statistical analyses, and visualization into a single framework. In addition, two complementary approaches are implemented: (1) An analytical method (the broken stick model) is proposed to help identify OTs of low abundance that could be generated by chance alone and (2) a one-pass profiling (OP) method, to efficiently identify those OTUs whose subsequent oligotyping would be most promising to be undertaken. These enhancements are especially useful for large datasets, where a manual screening of entropy analysis results and the creation of a full set of OTs may not be feasible. The package and procedures are illustrated by several tutorials and examples.

Keywords: oligotyping, minimum entropy decomposition, one-pass decomposition, diversity, next generation sequencing

INTRODUCTION

Eren et al. (2013) implemented a technique called *oligotyping* to help identify highly variable nucleotide positions of 16S rRNA gene sequences by calculating their Shannon entropy values. Subtle variations are used to iteratively classify the sequences into oligotypes (OTs), which may offer an interesting way to resolve ecologically meaningful differences between closely related organisms. In some cases, especially when processing data generated from sequencing methods prone to insertions or deletions (e.g. 454 Massively Parallel Tag Sequencing), sequence alignment must be performed prior to oligotyping to ensure meaningful classification (see the example below). The oligotyping procedure is straightforward: Sequences are assigned to the same taxonomic group or clustered together in one OTU before oligotyping analysis performs a systematic identification of nucleotide positions that represent information-rich variations across the group or OTU. The variation at these positions is then used to bin the sequences into OTs. If sample information is available for each sequence originating from one OTU, a sample-by-OT table is then produced, which can be subjected to traditional multivariate analyses (e.g., Legendre and Legendre, 1998; Ramette, 2007; Buttigieg and Ramette, in press).

Depending on the degree of variability in a sequenced region, the identity threshold between different OTs may be as low as 0.2%, i.e., about an order of magnitude lower than the 3% identity threshold that is currently being used to define OTUs. Consequently, the marginal diversity space left unexplored by coarse-grained methods requires attention and its significance needs to be assessed in its evolutionary and environmental context. Indeed, the subtle nucleotide variation detected by oligotyping among 16S ribosomal RNA gene amplicon reads has revealed ecologically meaningful microdiversity patterns hidden in sequence datasets. For instance, the technique has successfully identified subtle nucleotide variations that were associated with distinct environments, hosts, body location, or epidemiological states in human oral (Eren et al., 2014a), gut (Eren et al., 2014b), and bacterial vaginosis (Eren et al., 2011) microbiomes, but also in wastewater communities (McLellan et al., 2013), or among spatially structured communities in Arctic deep-sea sediments (Buttigieg and Ramette, submitted).

In addition to its ecological applications, the procedure is also computationally interesting because it identifies a relatively small subset of nucleotide positions in a set of sequences associated with high entropy values, thus reducing subsequent computational effort. However, the original oligotyping procedure is supervised: it relies on user input to decide how many components (i.e., positions with high entropy values) and which entropy threshold to be considered for further rounds of oligotyping. The supervised method may work when dealing with a few, well-targeted OTUs, but if we are to cope with very large datasets, as commonly encountered in environmental and clinical microbiology, a more scalable, automatic procedure is required. Recently, Eren

and colleagues proposed a computationally efficient procedure to partition marker gene datasets in an unsupervised fashion, which they termed *Minimum Entropy Decomposition* (MED; http://oligotyping.org/MED/; Eren et al., 2014c). This approach iteratively partitions large sets of sequences by repeating the oligotyping procedure until no more high entropy nucleotide positions are identified in any of the partitions of those sequences.

With regard to their implementation, the original oligotyping and MED software scripts are written in Python to efficiently handle the FASTA sequences, Shannon entropy calculations, and navigation across numerous directories that are created during the successive rounds of OT generation. The following Python modules need to be manually installed: *Matplotlib* (http://matplotlib.sourceforge.net/), *BioPython* (http://biopython.org/wiki/Biopython), *SciPy* (http://www.scipy.org/), *PyCogent* (http://pycogent.org/), and *Django* (https://www.djangoproject.com/), to generate user-friendly HTML outputs. The final stage of data visualization and further ecological analysis of sample-by-OT patterns rely on using the R language (R Core Team, 2014) and its libraries. Several R scripts are used to reduce the dimensionality of large datasets, calculate dissimilarity matrices, or to visualize data (e.g., using the functions *heatmap* and *barplot*). The oligotyping and MED scripts also have some dependencies such as NCBI executable (especially *blastn*) to match the most interesting OT sequences directly to their closest relatives in local or publicly available sequence databases.

Here, the R package *otu2ot*, which stands for "OTU to OT" is described and examples as well as tutorials are provided to illustrate the library's installation and functioning. The oligotyping and MED routines are implemented solely using R scripts in order to facilitate the integration of computational routines, interactive statistical analyses, and visualization into one common framework. Additional methods are also presented such as the broken stick model procedure to help identify OTs of low abundance that could be generated by chance alone. Further, a one-pass entropy profiling approach is compared to MED, as a method to efficiently identify those OTUs whose decomposition into OTs would be most promising. This latter method is especially useful for large datasets, where a complete decomposition to OTs may not be computationally feasible.

METHODS

R IMPLEMENTATION AND DEPENDENCIES

R (http://www.R-project.org/) is a widely used language and environment for statistical computation and graphics. The core of R is an interpreted computer language which allows branching and looping as well as modular programming using functions. Although most of the user-visible functions in R are written in the R language itself, procedures written in the C, C++, or FORTRAN languages, can be easily called to further improve computational efficiency.

To develop *otu2ot*, R version 3.1.0 was used within RStudio (version 0.98.953; http://www.rstudio.com/). Within *otu2ot*, the R library *seqinR* (Charif and Lobry, 2007) is called to efficiently import FASTA sequences. The optional libraries *FactoMineR* (Husson et al., 2014) and *vegan* (Oksanen et al., 2013) may also be used to calculate specific coefficients and to perform multivariate analysis of community data, respectively, but are not mandatory to perform the oligotyping or MED procedures. The package can be easily installed as described in the tutorials (Supporting Information). An active repository is available at: https://github.com/aramette/otu2ot.

EXPECTED INPUT DATA FORMAT

The *otu2ot* library expects input FASTA files to have a specific format, identical to that required by the original oligotyping pipeline, as described at: http://oligotyping.org/.

All of the (aligned) sequences from an OTU of interest have to be present in a single multi-FASTA file, and all reads must have the following format:

```
>[SampleName]_[ReadId]
GTTGAAAAAGTTAGTGGTGAAATCCCAGA
```

where "[SampleName]" refers to the name of the sample from which the sequences originated from and "[ReadId]" refers to a unique sequence identifier.

DIFFERENCES TO ORIGINAL OLIGOTYPING AND MED IMPLEMENTATIONS

In its current version (1.4), *otu2ot* does not implement two optional features found in the original procedure: (1) the selection of several components in the MED procedure, and (2) the subsequent BLAST analysis of the most abundant unique OT sequences against NCBI's *nr* database. This latter option may be readily integrated using additional R libraries such as BoSSA (http://cran.r-project.org/web/packages/BoSSA/) at a later stage. Other features are implemented, however, namely the broken stick model (BSM) and a one-pass (OP) procedure, as follows.

The BSM is implemented to help identify which OTs have a read abundance greater than one would expect by chance. Following the decomposition of an OTU into OTs, only those OTs which satisfy this condition are further considered for community analysis. The original BSM idea originates from niche theory (MacArthur, 1957), where the sub-division of niche space among species is thought to be analogous to randomly breaking a stick into p pieces. When applied to oligotyping data, the procedure is as follows: The total number of sequences clustered into one OTU is randomly split into p subsets (i.e., "pieces" of the broken stick) where p is defined by the number of OTs detected. The pieces are then sorted by decreasing size. By repeating these two steps many times and averaging the results over all executions the BSM generates the OT abundances which would occur by chance alone, that is, the distribution of OT abundances if there was no structure in the data. The R script used in our implementation uses a simple formula that provides the expected abundance values for a given partition under the BSM (Legendre and Legendre, 1998):

$$b_k = \frac{1}{p}\sum_{i=k}^{p}\frac{1}{i}$$

where p is the number of pieces (i.e., the number of OTs) and b_k is the expected abundance of the kth OT under the BSM.

One may then choose to limit their analyses to those OTs whose abundances are larger than those generated by the BSM. This procedure allows the use of a null abundance model to focus

on OTs whose abundances are likely to be non-random, instead of relying on an arbitrary choice of minimum OT abundance or on external knowledge to allow a given OT to be further considered for downstream analyses. This approach may thus help lessen the subjectivity which threatens reproducibility and consistency in defining what a minimum OT abundance should be. The BSM has been advocated as appropriate to describe the right-hand side of the rank frequency curve, i.e., the distribution of the rare species (Frontier, 1985), so it may be useful for OT abundance distributions, which are conceptually similar. In addition, the same approach is often applied to the solution of a principal component analysis in order to suggest the minimum number of principal axes needed to satisfactorily represent a data matrix (Legendre and Legendre, 1998). However, when considering results from oligotyping procedures, it is important to note that other models of species distribution exist and should also be evaluated (e.g., the log series, log normal, or neutral model): it would be hasty to favor the BSM over any alternatives at this stage. Future research using, for instance, simulated datasets with known amounts of sequencing error or rare sequences could be used to validate the application of the BSM approach to OT abundance modeling.

A one-pass (OP) procedure is also proposed to rapidly assess the amount of microdiversity present in a set of sequences. The procedure is similar to oligotyping, but it only performs one round of entropy calculations. When an entropy profile is obtained, only the nucleotide positions with Shannon entropy values greater than a chosen threshold are concatenated, and these concatenated sequences are then used to classify the sequences into OTs. Here we determined how OP compares to MED, in terms of computational speed and in its ability to capture ecological information such as variance in community composition, community patterns, or presence of rare types (e.g., singletons). We also evaluated whether OP can be used as a first screen across a large number of OTUs, before using the more computationally-demanding MED procedure to analyze microbial diversity on targeted OTUs.

FUTURE DEVELOPMENTS

The motivation behind the creation of *otu2ot* is twofold. First, it provides a more transparent, single-language implementation of the scripts used for oligotyping and MED, in order to promote more development of these tools and approaches. At this stage, less emphasis has been given to the improvement of computational performance, but this could be obtained by code optimization and interfacing with C or C++. This should be addressed when the phase of prototyping methods such as oligotyping, MED, or OP is over and large datasets need to be efficiently analyzed. In that respect, R is receiving much attention and is being actively developed to support very efficient parallel computing solutions, large memory data handling, and seamless interfacing with compiled code (e.g., http://cran.r-project.org/web/views/HighPerformanceComputing.html). R also has a large and growing user base among data scientists and ecologists, who continually submit new packages which can be integrated with *otu2ot*, further motivating development in this language. To improve interactive data exploration and visualization, developers have contributed R libraries such as *shiny* (http://shiny.rstudio.com/), which may readily turn a set of R functions into interactive web interfaces. Beyond R and interfacing with C or C++, other high-level languages may also be used to efficiently implement oligotyping, MED, or OP, at least for the entropy decomposition steps. These may be worth comparing to this R implementation in the future. For instance, the Julia language (http://julialang.org/) is a high-performance, dynamic programming language with a syntax that would be familiar to R users, and which seems to improve computing speed by several orders of magnitude when compared to a range of functions implemented in R.

EXAMPLE DATASETS

The original data available (e.g., "mock" dataset) on the website (http://oligotyping.org/) were used to create the R functions and ensure that results were concordant with those of the original

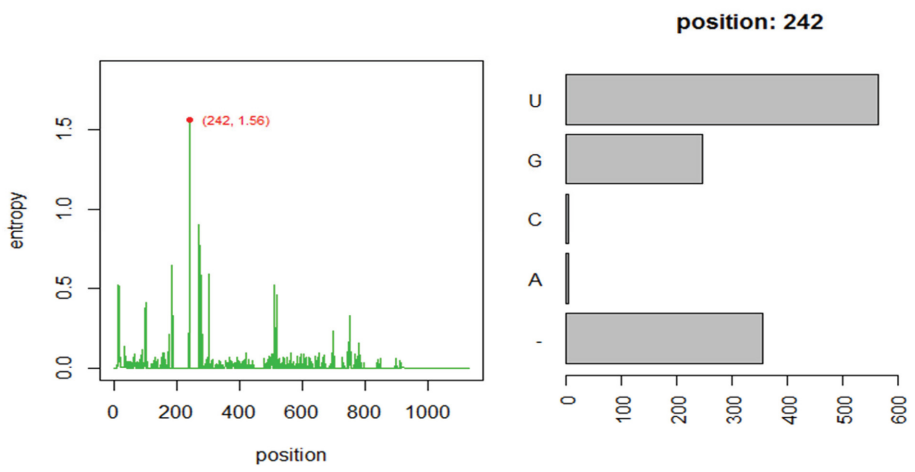


FIGURE 1 | Entropy profile of file "HGB_0013_GXJPMPL01A3OQX.fasta" and further nucleotide composition of the position of higher Shannon entropy (position 242).

implementation. These data are also included in the *otu2ot* package. The dataset used in Buttigieg and Ramette's (submitted) application of oligotyping was also used here. It corresponds to a set of sequence-abundant OTUs (abundance greater than or equal to 100 reads), derived from sequencing of sediment samples from the Hausgarten Long-Term Ecological Research station (Eastern Fram Strait, Arctic sea), which were clustered at the 97% sequence identity level of the 16S rRNA gene. The sequence data were produced by 454 Massively Parallel Tag Sequencing, and sequence alignment was performed to account for insertions and deletions. The analysis of the full OTU dataset from this site was previously published (Jacob et al., 2013). All relevant datasets are provided as Supplementary Information.

In the following section, a few plots and results are provided to illustrate how to use the *otu2ot* package and its functions. Here, we compare the results and performance of different methods and less emphasis is given to the ecological interpretation of the resulting OT tables, which can be found elsewhere (Buttigieg and Ramette, submitted). It should be noted that our examples began at an OTU-level resolution and further explored the extent of OT microdiversity within OTUs. It is equally interesting to choose a coarser taxonomic level (e.g., Phylum, Class) where more robust membership is expected, and then perform oligotyping methods. This would alleviate issues originating from splitting sequences into different OTUs as a result of the OTU clustering step.

EXAMPLES OF APPLICATION

MED ANALYSIS OF ONE OTU DATASET

Using one abundant OTU (1175 sequences, 1133 positions) whose sequences are provided in file HGB_0013_GXJPMPL01A3OQX.fasta, we generated a Shannon entropy profile of the alignment and a nucleotide composition

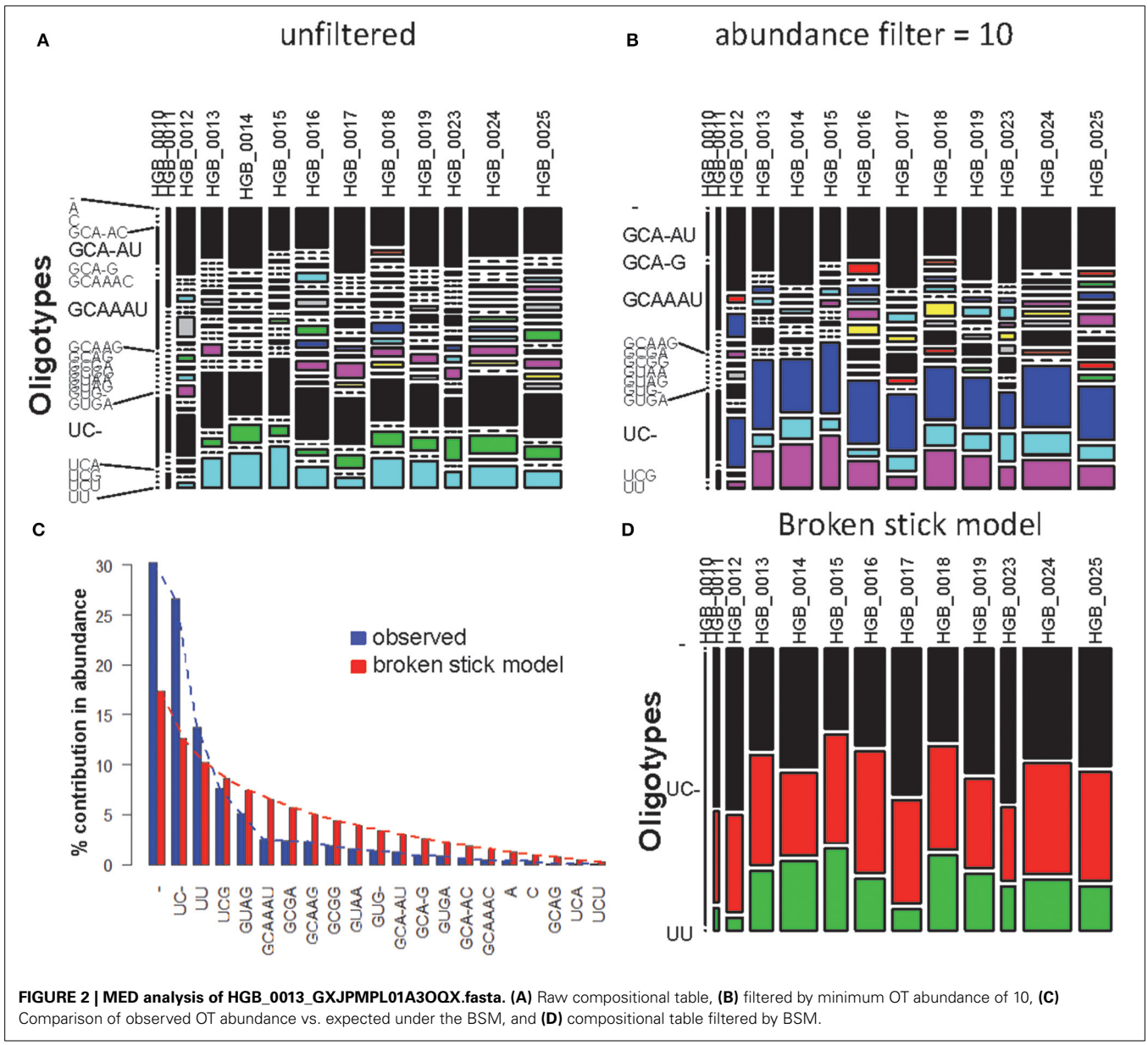


FIGURE 2 | MED analysis of HGB_0013_GXJPMPL01A3OQX.fasta. (A) Raw compositional table, **(B)** filtered by minimum OT abundance of 10, **(C)** Comparison of observed OT abundance vs. expected under the BSM, and **(D)** compositional table filtered by BSM.

profile of the position with the highest Shannon entropy (position 242). Note that alignment gaps (−) are also considered as informative in these calculations (**Figure 1**; Tutorial 1). By using the sample information in each sequence header, a raw sample-by-OT compositional table was generated (**Figure 2A**), which can be filtered by minimum OT abundance in the table (**Figure 2B**) or further filtered by applying the broken stick model (BSM) rule (**Figures 2C,D**).

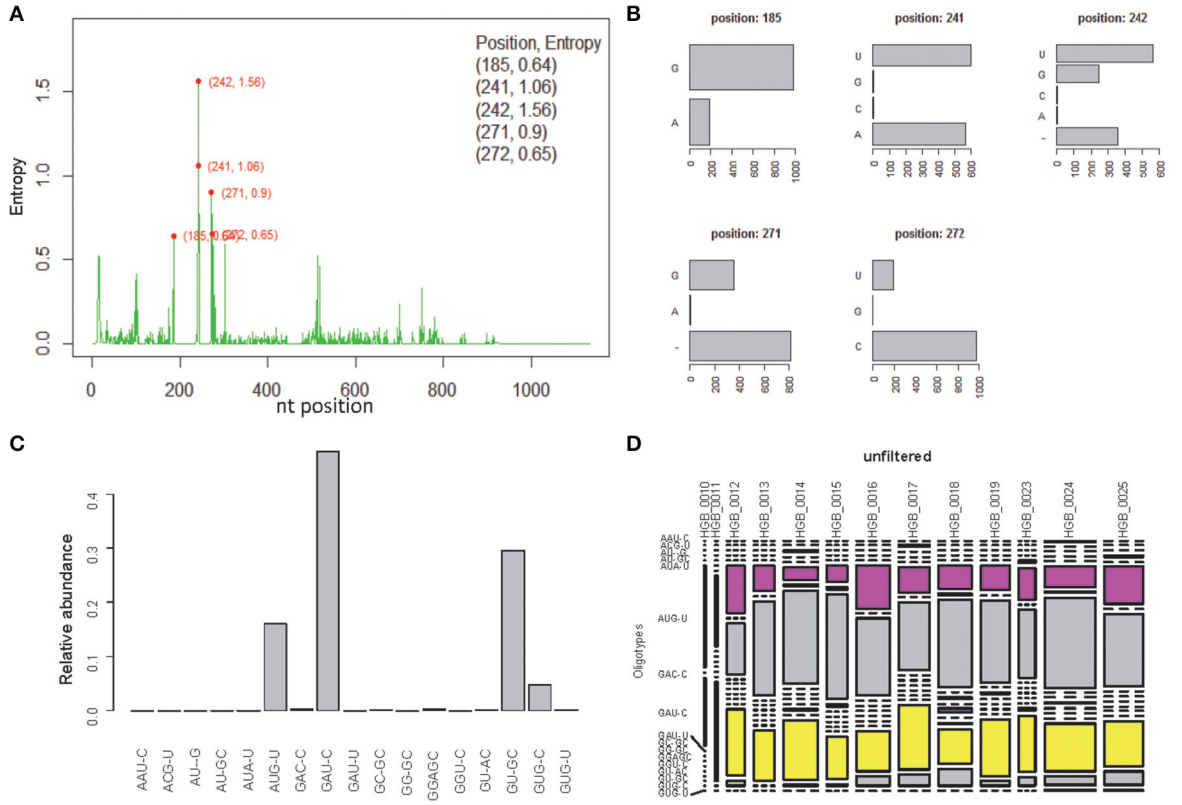


FIGURE 3 | One-Pass (OP) analysis of HGB_0013_GXJPMPL01A3OQX.fasta. (A) Shannon entropy profile, **(B)** nucleotide composition of the 5 high-entropy positions, **(C)** Relative abundance of each OT obtained by OP, **(D)** raw compositional table.

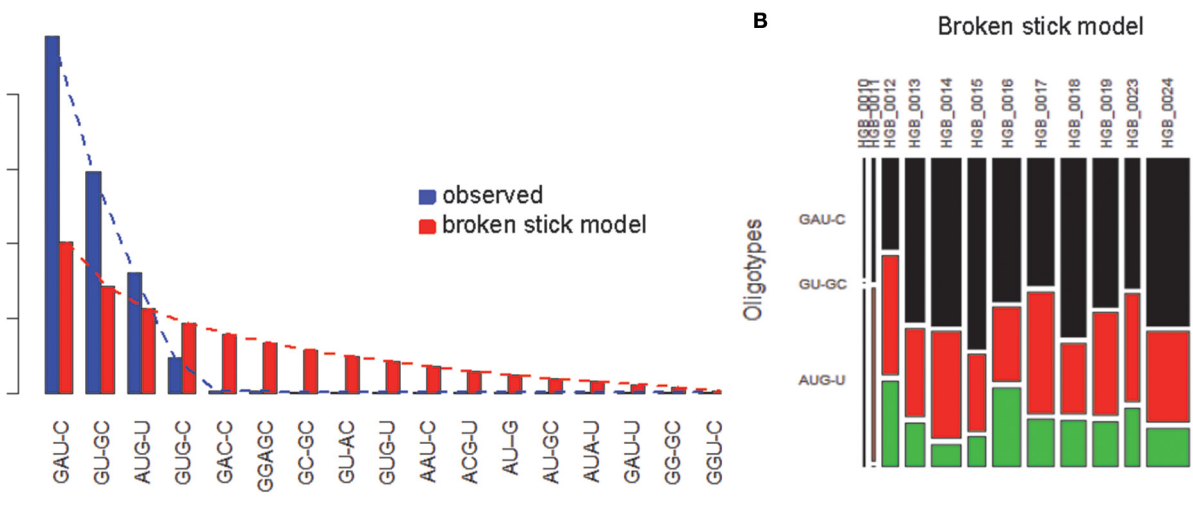


FIGURE 4 | BSM filtering applied to the OT table generated from HGB_0013_GXJPMPL01A3OQX.fasta by OP. (A) broken-stick model evaluation, **(B)** BSM filtered compositional table.

Table 1 | Correlation between OT abundance values obtained by MED (rows) and those obtained by OP (columns).

		*OP:GAU-C	*OP:GU-GC	*OP:AUG-U	*OP:GUG-C	*OP:GAC-C	*OP:GGAGC	OP:GC-GC	*OP:GU-AC	*OP:GUG-U	*OP:AAU-C	OP:ACG-U	OP:AU–G	*OP:AU-GC	OP:AUA-U	OP:GAU-U	*OP:GG-GC	OP:GGU-C
	Counts	**562**	**348**	**188**	**56**	**4**	**3**	**2**	**2**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
*MED.-	**355**	**0.864**	**0.999**	0.624	0.556	0.357	−0.027	0.098	−0.117	0.527	0.512	0.284	0.326	0.201	−0.131	−0.193	0.512	0.326
*MED.UC-	**313**	**0.966**	**0.871**	0.716	0.67	0.49	0.108	0.095	−0.148	0.613	0.612	0.108	0.13	0.218	−0.309	−0.002	0.612	0.13
*MED.UU	**161**	**0.913**	0.658	0.356	0.418	0.773	0.3	0.314	0.057	0.34	0.404	−0.222	0.439	0.056	−0.222	0.195	0.404	0.439
*MED.UCG	**89**	**0.891**	**0.829**	0.435	0.631	0.591	0.229	0.216	0.216	0.622	0.669	0.064	0.284	0.174	0.064	−0.157	0.669	0.284
MED.GUAG	**60**	0.71	0.792	0.752	0.763	0.08	0.125	0.118	0.051	0.584	0.576	0.396	−0.236	0.215	−0.056	−0.326	0.576	−0.236
MED.GCAAAU	**29**	−0.065	0.044	0.609	0.186	−0.346	−0.035	−0.052	−0.165	0.061	−0.188	−0.188	−0.188	0.271	−0.188	−0.188	−0.188	−0.188
*MED.GCGA	**28**	0.494	0.474	**0.834**	0.521	−0.044	−0.265	−0.3	−0.3	0.698	0.227	−0.142	−0.019	0.719	−0.142	−0.019	0.227	−0.019
MED.GCAAG	**27**	0.315	0.481	0.281	0.355	−0.037	0.192	0.438	0.284	−0.178	−0.225	0.609	−0.016	−0.016	0.192	−0.016	−0.225	−0.016
MED.GCGG	**22**	0.44	0.276	0.462	0.725	0.311	0.676	0.34	0.528	0.058	0.294	0.167	−0.216	−0.216	0.039	−0.216	0.294	−0.216
MED.GUAA	**18**	0.574	0.536	0.684	0.773	0.191	0.366	0.374	0.374	0.541	0.593	−0.087	−0.314	0.14	0.14	−0.314	0.593	−0.314
*MED.GUG-	**16**	0.659	0.631	0.652	**0.912**	0.22	0.324	0.208	0.073	0.749	0.507	0.324	−0.225	0.507	−0.225	−0.225	0.507	−0.225
MED.GCA-AU	**15**	0.11	0.021	0.714	0.448	−0.256	0.143	0.087	−0.039	0.087	−0.196	−0.196	−0.196	0.313	−0.196	−0.196	−0.196	−0.196
*MED.GCA-G	**10**	0.566	0.52	0.526	0.657	0.286	0.059	−0.102	−0.102	**0.849**	0.317	0.059	0.059	**0.833**	−0.198	0.059	0.317	0.059
*MED.GUGA	**10**	0.154	0.208	0.627	0.509	−0.309	0.059	0.278	0.088	0.469	−0.198	−0.198	−0.198	**0.833**	0.059	−0.198	−0.198	−0.198
MED.GCA-AC	**6**	0.355	0.343	0.531	0.266	0.03	−0.158	−0.234	−0.234	0.272	−0.158	−0.158	0.527	0.527	−0.158	−0.158	−0.158	0.527
MED.GCAAAC	**5**	0.49	0.245	0.497	0.245	0.158	−0.228	−0.337	−0.337	0.539	0.365	−0.228	−0.228	0.365	−0.228	0.365	0.365	−0.228
*MED.A	**4**	0.129	−0.103	0.012	0.488	0.359	**0.946**	0.619	**0.879**	−0.16	−0.108	−0.108	−0.108	−0.108	0.243	−0.108	−0.108	−0.108
*MED.C	**4**	0.632	0.341	−0.073	0.23	**1**	0.433	0.178	0.178	0.178	0.433	−0.192	0.433	−0.192	−0.192	0.433	0.433	0.433
*MED.GCAG	**1**	0.172	0.195	0.51	0.463	−0.192	−0.083	−0.123	−0.123	0.677	−0.083	−0.083	−0.083	**1**	−0.083	−0.083	−0.083	−0.083
MED.UCA	**1**	−0.282	−0.037	0.116	−0.16	−0.192	−0.083	−0.123	−0.123	−0.123	−0.083	−0.083	−0.083	−0.083	−0.083	−0.083	−0.083	−0.083
*MED.UCU	**1**	0.602	0.512	0.247	0.324	0.433	−0.083	−0.123	−0.123	0.677	**1**	−0.083	−0.083	−0.083	−0.083	−0.083	**1**	−0.083

Pearson correlation coefficients > 0.6 (absolute values) are indicated in gray. Bold, underlined values are those >0.8. OT names with an asterisk (*) are those which are associated with a correlation coefficient > 0.8 at least once, indicating a good match between the two methods.

ONE-PASS (OP) APPROACH

OP analysis of the same alignment file indicated 5 positions associated with high Shannon entropy values (**Figures 3A,B**; Tutorial 2). Further concatenation and binning of the sequence data led to 4 dominant OTs (**Figure 3C**) out of the 17 OTs generated by OP. Most of the rarer OTs were, in fact, singletons (**Figure 3D**). Subsequent BSM filtering (**Figure 4A**) led to a compositional table (**Figure 4B**) very similar to the one obtained by MED followed by BSM filtering (**Figure 2D**). Despite those similar plots, a number of differences may be observed which require careful investigation to fully compare the results produced by OP and MED (Tutorial 3).

As expected, the OP table has fewer columns (corresponding to 17 OTs) than the MED table (21 OTs). MED splits the initial number of sequences (1175) to greater extent, but OP displays more singleton OTs. When OT abundances were correlated across tables, high correlation values were mostly obtained among abundant OTs (**Table 1**), particularly for OT abundances >50 sequences. This may explain why community patterns that are extracted by multivariate techniques, many of which focus on the most abundant types, were found to be very similar overall. Because OP does not decompose the sequence pool to the same extent as MED, many OTs obtained by MED (11 out of 21), including some rather abundant ones, did not correlate with any OTs obtained by OP.

When both OT tables were rarefied according to the BSM, both were left with only 3 OTs, corresponding to 70.5 and 93.4% of all sequences for MED and OP, respectively, and which led to very similar abundance profiles (**Table 2**). Interestingly, the total community variance still present in each dataset was very different with nearly twice as much for OP (953) as for MED (472) (**Table 3**). Despite this notable difference, common statistical procedures based on dissimilarity indices failed to distinguish between these OT tables (Tutorial 3). Further, correlation coefficients between the raw tables or between distance matrices, calculated using differential weighting of double absences, led to the same conclusion: there were highly significant and strong correlations between the results obtained with the two approaches. OP generated more singleton OTs than MED, which may be observed when an asymmetric (Bray-Curtis) vs. symmetric (Euclidean) dissimilarity coefficient is used (**Figures 5A,B**, respectively). If OP is to be used to speed up the computation in lieu of MED, the best strategy would be to always use a filtering of the raw tables to avoid the increased generation of singleton OTs by OP.

One key parameter for ecological comparison and interpretation is the direct correlation among compositional tables (e.g., Gobet et al., 2010), as this ultimately determines the amount of change in community composition. Neither the RV coefficient nor the Mantel test was sensitive enough to capture the fine differences in the highly comparable compositional tables produced by each method (**Table 3**). However, Procrustes correlation analysis of the correspondence analysis (CA) results was found to be the most sensitive approach (Tutorial 3).

Table 2 | Sample-by-OT tables produced by MED and OP after applying the BSM procedure.

	MED			OP		
	-	UC-	UU	GU-GC	GAU-C	AUG-U
HGB_0010	0	2	0	0	2	3
HGB_0011	7	4	1	7	5	0
HGB_0012	25	15	2	25	19	18
HGB_0013	22	24	13	22	41	11
HGB_0014	43	30	25	42	66	9
HGB_0015	18	24	18	18	45	7
HGB_0016	28	35	15	28	54	30
HGB_0017	41	29	6	41	43	16
HGB_0018	26	29	21	24	61	16
HGB_0019	35	25	16	34	49	15
HGB_0023	21	10	6	20	24	11
HGB_0024	52	52	24	51	95	22
HGB_0025	37	34	14	36	58	30

COMPARISON OF OP AND MED FOR SEVERAL OTU DATASETS

A set of 269 OTU FASTA alignments coming from the same study as HGB_0013_GXJPMPL01A3OQX.fasta was submitted to both MED and OP to systematically compare their output (Tutorial 4). MED took about 10 times longer to complete than OP on

Table 3 | Summary of the comparison between 1) OP vs. MED and 2) using the raw compositional table or a compositional table filtered by applying the BSM procedure.

Type of data	Raw abundance		BSM	
Method	OP	MED	OP	MED
Table name in the tutorials	TOP0	TM0	TOP_BSM	TM_BSM
Total number of OTs	17	21	3	3
Number of singleton OTs (%)	8 (47%)	3 (14%)	0 (0%)	0 (0%)
Total variance	974.2	543.0	953.3 (97.9%)[$]	472.1 (86.9%)[$]
RV Coefficient	rv: 0.9848*		rv: 0.9824*	
Mantel test: Bray-Curtis, Euclidean index	r: 0.994*, r: 0.981*		r: 0.987*, r: 0.975*	
Correlation of CA ordination plots (Procrustes rotation)	r: 0.787*		r: 0.879*	
Number of OTs highly correlated (>0.8) to OTs produced with the other approach (% of the total number of OTs) (see **Table 1**)	11 (64.7%)	12 (57.1%)	2 (66.7%)	3 (100%)

** $P < 0.01$.*

[$] *percentage referring to the variance in the corresponding raw abundance table.*

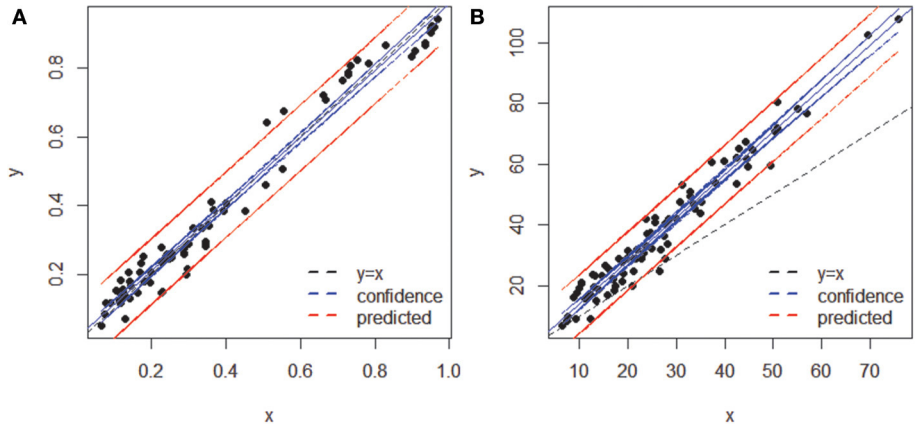


FIGURE 5 | Comparison of sample dissimilarities obtained by MED (y axis) and of those obtained by OP (x axis) (A) using an asymmetric (Bray-Curtis) dissimilarity coefficient, (B) using a symmetric (Euclidean) coefficient. Notice the departure from the 1:1 line due to the inclusion of rare OTs in (B). The blue and red lines represent confidence (95%) and prediction lines, respectively, of the linear regression models.

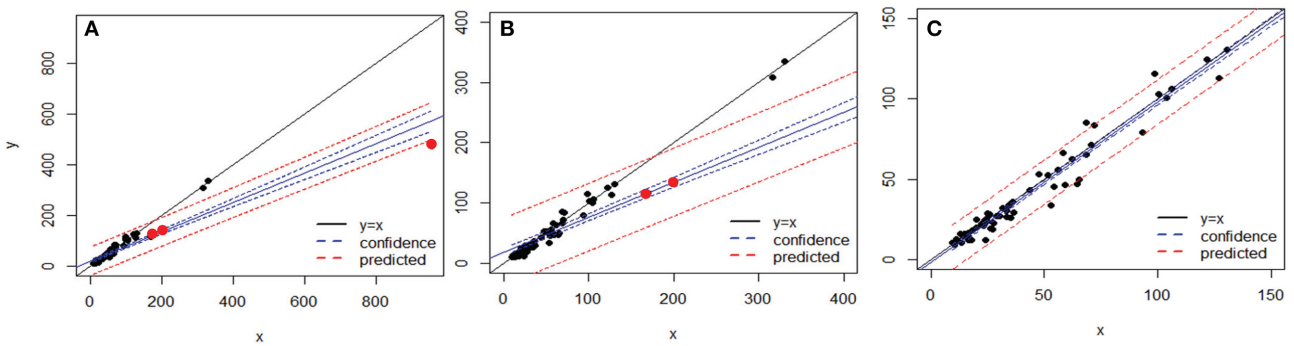


FIGURE 6 | Comparison of variances obtained by MED (y axis) vs. OP (x axis) across 67 sequence alignments (OTUs), both after BSM filtering. **(A)** whole dataset, **(B)** after rescaling to variances below 400 on each axis, and **(C)** after removing the three points that made the y = x line deviate (as red dots in A and B). The blue and red lines represent linear confidence (95%) and prediction lines, respectively.

the same data (about 10 min and less than a min, respectively, on a desktop computer [3.40 GHz, 8 GB RAM, 64-bit Windows 7 OS], when the plotting option was disabled). A total of 217 datasets had Shannon entropy >0.6. The RV coefficients comparing the correlation between the raw OT tables generated by the two approaches ranged from 0.78 to 1.0 (mean of 0.97) and were highly significant. Using CA as a finer approach to detect subtle changes in community composition (see above), 198 OT tables could be represented by a 2D solution and 19 OT tables produced a one-dimensional solution. The former were then used to compare ordination of the samples under the two approaches and 76% of them were found to display significantly related ordination plots, with Procrustes correlation coefficients ranging from 0.54 to 1.0 (mean 0.86).

After applying BSM filtering to MED- and OP-generated tables, only 79 and 123 datasets still contained OTs, respectively, with 67 datasets in common to both techniques. The comparison of the variance in each dataset across the 67 sequence alignments identified three datasets which were mainly responsible for the departure from an exact match between the variances obtained by the two methods for each dataset analyzed (**Figure 6**). Removing those three datasets, in which OP identified generally higher variance than MED (Tutorial 4), led to a near 1:1 correspondence between the variance obtained by MED and by OP (**Figure 6C**).

To better explore the nature of this discrepancy, the three outlier datasets were further compared to the rest of the datasets in terms of maximum entropy level and number of components in the initial sequence alignments; however, these three datasets did not show any particularly extreme behavior (**Figure 7**). Likewise, no obvious relationship could be found between the variance in an OT table and either the maximum entropy or number of components found in the initial sequence alignments (Tutorial 4).

The RV coefficients, ranging from 0.79 to 1.0, were very similar to those reported for the methodological comparison based on the raw data. When CA was applied, 25 (48%) out of 57 remaining datasets had a valid 2D representation, from which 19 (i.e., 76%) were significantly correlated across methods with Procrustes coefficients ranging from 0.57 to 1.0 (mean of 0.77). Only seven out of 25 had a Procrustes correlation coefficient > 0.8 (Tutorial 4), thus indicating that few datasets had strong agreement between the CA solutions produced by MED and those produced by OP.

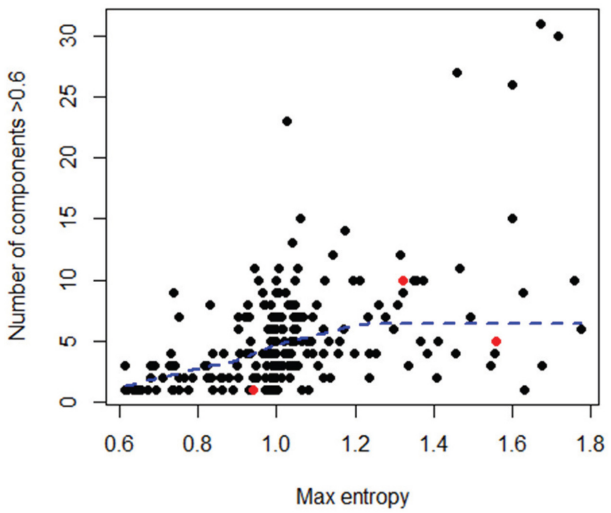


FIGURE 7 | Comparison of number of components (i.e., high entropy positions) and maximum entropy level for all datasets, considering only the initial sequences in each dataset. In red are the 3 outlier datasets shown in **Figure 6**. A LOWESS local fitting line (dotted line) was used to describe the complex shape of the relationship.

CONCLUSIONS

The initial choice of file "HGB_0013_GXJPMPL01A3OQX.fasta," which was randomly done, was to some extent unfortunate because that dataset belongs to one of the outlier datasets identified above. When all datasets were used to allow for a more robust methodological comparison, OP seemed to offer a good approach to first screen a large number of sequence datasets (i.e., OTUs), which may then be submitted to MED for more in-depth, and more computationally-demanding, analysis of the existing microdiversity. As demonstrated here, however, the OT tables produced by OP and MED might sometimes not necessarily capture the same ecological information, and this was particularly notable when investigating the fine correspondences between OT abundance and sample mapping in ordination space.

ACKNOWLEDGMENTS

We are grateful to Christian Quast for his assistance in preparing data exports from the SILVAngs pipeline. This is a contribution to the Micro B3 project, funded by the European Union's Seventh Framework Programme (Joint Call OCEAN. 2011-2: Marine microbial diversity—new insights into marine ecosystems functioning and its biotechnological potential) under the grant agreement no 287589. Alban Ramette is funded by the Max Planck Society.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2014.00601/abstract

REFERENCES

Buttigieg, P. L., and Ramette, A. (in press). A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. *FEMS Microbiol. Ecol.* doi: 10.1111/1574-6941.12437

Charif, D., and Lobry, J. R. (2007). "SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis," in *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations*, eds U. Bastolla, M. Porto, E. Roman, and M. Vendruscolo, (New York, NY: Springer Verlag), 207–232.

Eren, A. M., Borisy, G. G., Huse, S. M., and Welch, J. L. M. (2014a). Oligotyping analysis of the human oral microbiome. *Proc. Natl. Acad. Sci. U.S.A.* 111, E2875–E2884. doi: 10.1073/pnas.1409644111

Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., et al. (2013). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* 4, 1111–1119. doi: 10.1111/2041-210X.12114

Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H., and Sogin, M. L. (2014c). Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J.* doi: 10.1038/ismej.2014.195. [Epub ahead of print].

Eren, A. M., Sogin, M. L., Morrison, H. G., Vineis, J. H., Fisher, J. C., Newton, R. J., et al. (2014b). A single genus in the gut microbiome reflects host preference and specificity. *ISME J.* doi: 10.1038/ismej.2014.97. [Epub ahead of print].

Eren, A. M., Zozaya, M., Taylor, C. M., Dowd, S. E., Martin, D. H., and Ferris, M. J. (2011). Exploring the diversity of *Gardnerella vaginalis* in the genitourinary tract microbiota of monogamous couples through subtle nucleotide variation. *PLoS ONE* 6:e26732. doi: 10.1371/journal.pone.0026732

Frontier, S. (1985). Diversity and structure of aquatic ecosystems. *Oceanogr. Mar. Biol. Annu. Rev.* 23, 253–312.

Gobet, A., Quince, C., and Ramette, A. (2010). Multivariate Cutoff Level Analysis (MultiCoLA) of large community data sets. *Nucleic Acids Res.* 38, e155. doi: 10.1093/nar/gkq545

Husson, F., Josse, J., et al. (2014). *FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R. R package version 1.26*. Available online at: http://CRAN.R-project.org/package=FactoMineR.

Jacob, M., Soltwedel, T., et al. (2013). Biogeography of deep-sea benthic bacteria at regional scale (LTER HAUSGARTEN, Fram Strait, Arctic). *PLoS ONE* 8:e72779. doi: 10.1371/journal.pone.0072779

Legendre, P., and Legendre, L. (1998). *Numerical Ecology.* Amsterdam, The Netherlands: Elsevier Science B.V.

MacArthur, R. (1957). On the relative abundance of bird species. *Proc. Natl. Acad. Sci. U.S.A.* 43, 293–295 doi: 10.1073/pnas.43.3.293

McLellan, S. L., Newton, R. J., et al. (2013). Sewage reflects the distribution of human faecal Lachnospiraceae. *Environ. Microbiol.* 15, 2213–2227. doi: 10.1111/1462-2920.12092

Oksanen, J., Kindt, R., et al. (2013). *Vegan: Community Ecology Package. R Package Version, 2.0-10.*

Ramette, A. (2007). Multivariate analyses in microbial ecology. *FEMS Microbiol. Ecol.* 62, 142–160. doi: 10.1111/j.1574-6941.2007.00375.x

R Core Team (2014). *R: a Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Avaliable online at: http://www.R-project.org/.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 04 August 2014; accepted: 23 October 2014; published online: 14 November 2014.

*Citation: Ramette A and Buttigieg PL (2014) The R package otu2ot for implementing the entropy decomposition of nucleotide variation in sequence data. Front. Microbiol. **5**:601. doi: 10.3389/fmicb.2014.00601*

This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on