Taxonomic annotation errors incorrectly assign the family Pseudoalteromonadaceae to the order Vibrionales in Greengenes: Implications for microbial community assessments

Keri Ann Lydon Correspond., Erin K Lipp Correspond.

1 Department of Environmental Health Science, University of Georgia, Athens, GA, United States

Corresponding Authors: Keri Ann Lydon, Erin K Lipp
Email address: keri.lydon@gmail.com, elipp@uga.edu

Next-generation sequencing has provided powerful tools to conduct microbial ecology studies. Analysis of community composition relies on annotated databases of curated sequences to provide taxonomic assignments; however, these databases occasionally have errors with implications for downstream analyses. Systemic taxonomic errors were discovered in Greengenes database (v13_5 and 13_8) related to orders Vibrionales and Alteromonadales. These orders have family level annotations that were erroneous at least one taxonomic level, e.g., 100% of sequences assigned to the Pseudoalteromonadaceae family were placed improperly in Vibrionales (rather than Alteromonadales) and >20% of these sequences were indeed Vibrio spp. but were improperly assigned to the Pseudoalteromonadaceae family (rather than to Vibrionaceae). Use of this database is common; we identified 68 peer-reviewed papers since 2013 that likely included erroneous annotations specifically associated with Vibrionales and Pseudoalteromonadaceae, with 20 explicitly stating the incorrect taxonomy. Erroneous assignments using these specific versions of Greengenes can lead to incorrect conclusions, especially in marine systems where these taxa are commonly encountered as conditionally rare organisms and potential pathogens.
Taxonomic annotation errors incorrectly assign the family Pseudoalteromonadaceae to the order Vibrionales in Greengenes: Implications for microbial community assessments

Keri Ann Lydon1* and Erin K. Lipp1*

1 University of Georgia, Department of Environmental Health Science, Athens, GA

*Co-Corresponding authors: 150 East Green St., Athens, GA, 30602; keri.lydon@gmail.com; elipp@uga.edu; (706) 583-8138; @SciKeri

Running title: Greengenes errors in Vibrionales
Abstract

Next-generation sequencing has provided powerful tools to conduct microbial ecology studies. Analysis of community composition relies on annotated databases of curated sequences to provide taxonomic assignments; however, these databases occasionally have errors with implications for downstream analyses. Systemic taxonomic errors were discovered in Greengenes database (v13_5 and 13_8) related to orders Vibrionales and Alteromonadales. These orders have family level annotations that were erroneous at least one taxonomic level, e.g., 100% of sequences assigned to the Pseudoalteromonadaceae family were placed improperly in Vibrionales (rather than Alteromonadales) and >20% of these sequences were indeed *Vibrio* spp. but were improperly assigned to the Pseudoalteromonadaceae family (rather than to Vibrionaceae). Use of this database is common; we identified 68 peer-reviewed papers since 2013 that likely included erroneous annotations specifically associated with Vibrionales and Pseudoalteromonadaceae, with 20 explicitly stating the incorrect taxonomy. Erroneous assignments using these specific versions of Greengenes can lead to incorrect conclusions, especially in marine systems where these taxa are commonly encountered as conditionally rare organisms and potential pathogens.
Introduction

Analysis of 16S rRNA gene sequences has dramatically changed the way microbiologists understand the ecology of whole bacterial communities in an ecosystem. We are now able to sequence millions of reads of this gene in mixed samples to understand changes and dynamics in microbial composition. To analyze these data, sequence reads are compared against a curated ribosomal sequence database with known taxonomic identities. Frequently used databases include Greengenes (DeSantis et al., 2006; McDonald et al., 2012; http://greengenes.secondgenome.com), SILVA (Pruesse et al., 2007), the ribosomal database project (RDP) (Cole et al., 2014), or NCBI (Federhen, 2012). Greengenes is highly cited and was first introduced in 2006 for assigning taxonomies to Archaea and Bacteria using an automated de novo tree-based approach (DeSantis et al., 2006; McDonald et al., 2012). One of the reasons for its popularity is its ease of use given that it has been assimilated into 16S rRNA gene sequence analysis pipelines such as QIIME (Caporaso et al., 2010). Although Greengenes is one of the smallest databases, it has been suggested as the preferred database for classification of taxonomy because of its capacity to assign taxonomy to great depth (e.g., species level identification) (Werner et al., 2012).

In a recent analysis of 16S rRNA gene sequences using QIIME version 1.9.1 (Caporaso et al., 2010) in our laboratory, we observed that the Greengenes taxonomy for the orders Vibrionales and Alteromonadales appeared to have errors when using version 13_8. Further investigation of Greengenes versions revealed that these errors first appeared in version 13_5, but were not present in earlier versions (Accessed 26 March 2018; ftp://ftp.microbio.me/greengenes_release/).
The taxonomic orders Vibrionales and Alteromonadales include heterotrophic Gram-negative bacteria within the class Gammaproteobacteria that are important members in a number of environments, especially in marine systems. The genus *Vibrio* in Vibrionales includes opportunistic, ubiquitous, and conditionally rare taxa that are studied for their contributions to nutrient cycling in the marine environment (Thompson and Polz, 2006; Takemura et al., 2014; Vezzulli et al., 2016), and more notably pathogenicity in both invertebrates and vertebrates (e.g., Austin and Zhang, 2006). Furthermore, the family Vibrionaceae includes several common human pathogens (e.g., Newton et al., 2012; U.S. Centers for Disease Control (CDC), 2016).

Pseudoalteromonadaceae, which belong to the Alteromonadales order, are also opportunistic and conditionally rare in the marine environment, where they have been associated with disease (Choudhry et al., 2015; Beurmann et al., 2015) and as biofilm forming bacteria associated with the surfaces of eukaryotic organism (Davis et al., 1989; Holmström and Kjelleberg, 1999). They occupy a similar niche as Vibrionaceae, yet *Pseudoalteromonas* sp. (within the family Pseudoalteromonadaceae) are known antagonists of *Vibrio* spp. (Morya et al., 2014; Richards et al., 2017)

Edgar (2018a) published a non-peer reviewed pre-print (25 March 2018) that found the annotation error rate in Greengenes version 13_5 to be approximately 15%, including mismatches for the families Vibrionaceae and Pseudoalteromonadaceae between SILVA version 128 and Greengenes 13_5 (amended to a 17% error rate in a revised version [Edgar et al., 2018b]). Although the analyses reported by Edgar (2018b) suggest broader issues with taxonomic mismatches (also recognized by DeSantis et al. (2006, with the original release of Greengenes) among others (e.g., Beiko, 2016; Kozlov et al., 2016; Balvočiūtė and Huson, 2017)), we specifically assessed the extent of the misclassification associated with order
Vibrionales and family Pseudoalteromonadaceae, given the importance of these taxa in marine systems. Furthermore, we highlight that these errors do not appear to be due to real differences in inferred taxonomy between databases but are rather an example of a specific assignment error in Greengenes curation.

**Methods**

**Taxonomic Comparison of Pseudoalteromonadaceae Across Curated Databases.** We searched for the bacterial family Pseudoalteromonadaceae within the representative sequences and taxonomy for Greengenes version 13_8 provided within QIIME version 1.9.1 and found that 100% of Pseudoalteromonadaceae sequences (n=164) were misclassified within the Vibrionales order when they properly belong to the Alteromonadales order (Supplementary Data File 1). This was also observed in an earlier version of Greengenes (13_5). Subsequently, we checked taxonomies against NCBI (16 rRNA database), RDP (Type Strains), and SILVA (ref NR) using the web interfaces of BLAST (Morgulis et al., 2008; https://blast.ncbi.nlm.nih.gov), SeqMatch (Cole et al., 2014; http://rdp.cme.msu.edu/seqmatch/seqmatch_intro.jsp), and SINA (Pruesse et al., 2012; https://www.arb-silva.de/aligner/), respectively, for all Pseudoalteromonadaceae sequences from the Greengenes representative sequences in version 13_8 to determine if there were additional mismatches at other levels of taxonomy.

**Phylogenetic Tree of Pseudoalteromonadaceae in Greengenes.** The 16S rRNA gene representative sequences assigned to Pseudoalteromonadaceae in Greengenes (n=164) were used to create a phylogenetic tree in Unipro UGENE v.1.26.3 (Okonechnikov et al., 2012). Reads were first aligned using MUSCLE (Edgar, 2004), before subsequent tree building using PhyML.
maximum likelihood with aLRT SH-like branch support (Guindon et al., 2010). Results were visualized in iTOL (version 4.0.3; Letunic and Bork, 2007), including the addition of color strips which indicate the assigned taxonomy results generated in this study from NCBI, RDP, SILVA, Greengenes queries.

**Literature Search for Taxonomic Errors.** We attempted to examine how widespread such a taxonomic annotation error might have been propagated in the microbial ecology literature. We conducted a systematic literature search within Google Scholar using the terms "Greengenes + Vibrionales + Pseudoalteromonadaceae" or "Greengenes + Vibrionales" for the years 2013 – 2018 (Greengenes 13_5 and 13_8 were released in 2013), including only published peer-reviewed papers accessible in English and which used a next-generation amplicon sequencing approach (i.e., excluding analyses with PhyloChip, DGGE, or clone libraries). The search was completed between 23 and 28 March 2018.

**Results**

All of the 164 sequences identified as Pseudoalteromonadaceae in the Greengenes database were incorrectly assigned to the Vibrionales order, rather than the Alteromonadales order (the accepted next level of taxonomic rank for this family (Ivanova et al., 2014)). Of these 164 representative sequences in Greengenes, 115, 118, and 117 were confirmed by SILVA, RDP, and NCBI, respectively, as Pseudoalteromonadaceae sequences and were assigned (correctly) to the order Alteromonadales. Furthermore, RDP and NCBI identified 45, and SILVA 43, of the 164 sequences as *Vibrio* spp. that were incorrectly assigned to the Pseudoalteromonadaceae family in Greengenes; therefore, because Greengenes assigned all
members of the Pseudoalteromonadaceae family to the Vibrionales order, these 43-46 sequences were only incorrect at the family level (Figure 1). These sequences mismatched by family represented 23 *Vibrio* spp. and included common species such as *V. alginolyticus*, *V. cholerae*, and *V. harveyi* (Supplementary Data File 1). This means that while these sequences did belong in the order Vibrionales, if the family level was chosen for downstream analysis, they would have been analyzed incorrectly as Pseudoalteromonadaceae. There were several sequences for which we were unable to confirm taxonomic identity using NCBI, SILVA, or RDP or were attributed to the bacterial families Shewanellaceae, Colwelliaceae, Oceanospirillaceae within the order Alteromonadales (Figure 1).

We found 85 published papers that met our search criteria in journals such as *PLoS One*, *Frontiers in Microbiology*, *ISME J*, among others that used either 13_5 or 13_8 versions of Greengenes to assign their taxonomy (Table 1). Of these articles, 20 reported results clearly stating the incorrect taxonomic assignments or listing the mismatched taxonomy in a table (in text or in supplemental material). An additional 35 papers were presumed to have misclassification errors based on the database version reported. Only 17 of the evaluated articles reported either the correct taxonomic assignment directly or were presumed to have the correct assignment based on the use of earlier Greengenes versions (or the explicitly stated use of other databases for assignment), which did not have the mismatch error. The remaining 13 papers did not report sufficient detail on the database or bioinformatics pipeline used to determine if the mismatch error was likely. In all, 55 of the 85 papers (~64.7%) since 2013 likely reported relative abundance levels of Vibrionales or Alteromonadales that were incorrect due to improper taxonomic assignments.
Discussion and Conclusions

Results presented here demonstrate the extent of taxonomic annotation errors associated with the family Pseudoalteromonadaceae present in the Greengenes database versions 13_8 and 13_5 for the order Vibrionales. Although the Gammaproteobacteria are known to include a number of polyphyletic groups, there is a clear distinction between the Vibrionales and Alteromonadales (Williams et al., 2010). Given that this assignment of Pseudoalteromonadaceae to Vibrionales only occurred in the Greengenes databases (13_5 and 13_8), this suggests an issue of curation rather than an underlying change in real taxonomic assignment. The Vibrionales and Alteromonadales are often associated with similar niches in marine systems; however, correct taxonomic classification is imperative, especially when attribution is important, such as the identification of potential pathogens.

Issues with taxonomic assignments among databases have been the subject of previous and on-going research (e.g., Kovlov et al. 2016, Edgar 2018b), given its importance in studies where identification is needed. Results with erroneous attribution can alter our understanding or interpretation of the ecology of those microbial assemblages. For example, given that 27% of Vibrio spp. sequences in Greengenes were incorrectly attributed to Pseudoalteromonadaceae instead of Vibrionaceae, we could be making inferences on artificially low relative abundances at the family level. Likewise, at the order level Pseudoalteromonadaceae sequences could artificially inflate the relative abundance of Vibrionales because of these assignment errors. Although it is not clear how this incorrect assignment error in arose between Greengenes version 12_10 and 13_5, it does not seem to be due to a real or inferred dispute in taxonomic classification. We acknowledge that potential taxonomic errors are problematic for other databases, as well (Edgar et al., 2018a; Edgar et al., 2018b). Therefore, support for curation of
taxonomic databases that are foundational to microbial ecology research using next-generation sequencing approaches is critical. This is an opportunity for field-wide collaboration to ensure that databases reviewed and curated.

Acknowledgments

We thank M. Ghazaleh Bucher, J. Westrich, and E. Ottesen for comments and advice on this manuscript.

References

Austin B and Zhang XH. 2006. Vibrio harveyi: a significant pathogen of marine vertebrates and invertebrates. Letters in applied microbiology, 43(2), pp.119-124.

Balvočiūtė M, Huson DH. 2017. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? BMC Genomics, 18(Suppl 2), 114. doi:10.1186/s12864-017-3501-4

Beiko RG. 2016. Microbial malaise: How can we classify the microbiome? Trends Microbiol. 23(11):671–9. doi: 10.1016/j.tim.2015.08.009

Beurmann S, Ushijima B, Videau P, Svoboda CM, Smith AM, Rivers OS, Aeby GS and Callahan SM, 2017. Pseudoalteromonas piratica strain OCN003 is a coral pathogen that causes a switch from chronic to acute Montipora white syndrome in Montipora capitata. PloS one, 12(11), p.e0188319.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7(5): 335-336.
Choudhury JD, Pramanik A, Webster NS, Llewellyn LE, Gachhui R and Mukherjee J. 2015. The pathogen of the Great Barrier Reef sponge Rhopaloeides odorabile is a new strain of Pseudoalteromonas agarivorans containing abundant and diverse virulence-related genes. Marine biotechnology, 17(4), pp.463-478.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al., 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis Nucl Acids Res 42(Database issue): D633-D642.

Davis AR, Targett NM, McConnel OJ, and CM Young. 1989. Epibiosis of marine algae and benthic invertebrates: natural products chemistry and other mechanisms inhibiting settlement and overgrowth. Bioorg. Mar. Chem. 3:85–114

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol 72: 5069-5072.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792-1797.

Edgar RC. 2018a. Taxonomy annotation errors in 16S rRNA and fungal ITS sequence databases. bioRxiv Accessed 27 March 2018; doi: 10.1101/288654

Edgar, R.C., 2018b. Taxonomy annotation and guide tree errors in 16S rRNA databases. bioRxiv, p.288654. Accessed 25 May 2018: doi: 10.1101/288654

Federhen S. 2012. The NCBI taxonomy database. Nucleic Acids Res. 40(D1):136–43. doi: 10.1093/nar/gkr1178
Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Systematic Biology 59:307-321.

Holmström C and Kjelleberg S, 1999. Marine Pseudoalteromonas species are associated with higher organisms and produce biologically active extracellular agents. FEMS microbiology ecology, 30(4), pp.285-293.

Ivanova EP, Ng HJ, Webb HK. 2014. The Family Pseudoalteromonadaceae. In: Rosenberg E., DeLong E.F., Lory S., Stackebrandt E., Thompson F. (eds) The Prokaryotes. Springer, Berlin, Heidelberg)

Kozlov AM, Zhang J, Yilmaz P, Glöckner FO, Stamatakis A. 2016. Phylogeny-aware identification and correction of taxonomically mislabeled sequences, Nucleic Acids Research, 44:5022–5033.

Letunic I and Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23:127-128.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, 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.

Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA. 2008. Database indexing for production MegaBLAST searches. Bioinformatics 15:1757-1764.

Morya VK, Choi W, Kim EK. 2014. Isolation and characterization of Pseudoalteromonas sp. from fermented Korean food, as an antagonist to Vibrio harveyi. Applied microbiology and biotechnology. 98(3):1389-95.
Newton A, Kendall M, Vugia DJ, Henao OL, Mahon BE. 2012. Increasing rates of vibriosis in the United States, 1996–2010: Review of surveillance data from 2 systems. Clin Infect Dis 54:S391–S395. doi:10.1093/cid/cis243

Okonechnikov K, Golosova O, Fursov M, team U. 2012. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics 28:1166-1167.

Pruesse E, Peplies J, and Glöckner FO. 2012. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics 28:1823-1829.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO. 2007. SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35: 7188–7196.

Richards GP, Watson MA, Needleman DS, Uknalis J, Boyd EF, Fay JP. 2017. Mechanisms for Pseudoalteromonas piscicida-induced killing of vibrios and other bacterial pathogens. Applied and environmental microbiology. 2017 Jun 1;83(11):e00175-17.

Takemura AF, Chien DM, Polz MF. 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 JR and Polz MF. 2006. Dynamics of Vibrio populations and their role in environmental nutrient cycling. In The biology of vibrios (pp. 190-203). American Society of Microbiology.

U.S. Center for Disease Control and Prevention (CDC). 2016 National Surveillance of Bacterial Foodborne Illnesses (Enteric Diseases): National Cholera and Vibriosis Surveillance. 2016. Retrieved from the Center for Disease Control website: https://www.cdc.gov/vibrio/surveillance.html
Vezzulli L, Grande C, Reid PC, Hélaouët P, Edwards M, et al. 2016. Climate influence on Vibrio and associated human diseases during the past half-century in the coastal North Atlantic. P Natl Acad Sci USA 113:E5062–71. doi:10.1073/pnas.1609157113

Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, Capoaso JG, et al. 2012. Impact of training sets on classification of high-throughput bacterial 16S rRNA gene surveys. ISME J 6: 94-103.

Williams, K. P., Gillespie, J. J., Sobral, B. W., Nordberg, E. K., Snyder, E. E., Shallom, J. M., & Dickerman, A. W. (2010). Phylogeny of gammaproteobacteria. Journal of bacteriology, 192(9), 2305-2314.
Figure 1 (on next page)

Phylogenetic Tree of Pseudoalteromonadaceae representative sequences in the Greengenes Database (n=164).

Each branch is labeled by the assigned ID in the representative sequences within Greengenes. Color strips indicated the assigned taxonomy (Order and Family) from different curated databases: Greengenes (GG), SILVA, RDP, and NCBI. The accepted Order for the family Pseudoalteromonadaceae is Alteromonadales (Ivanova et al., 2003) not Vibrionales, as featured in all Greengenes taxonomic assignments.
Incorrect taxonomic annotation of Pseudoalteromonadaceae in Vibrionales as published in peer reviewed literature.

Search was conducted between 23 and 28 March 2018 using Google Scholar. Search results included in the analysis were peer-reviewed papers published between 2013 and 2018 that were accessible in English and used a 16S rRNA gene next-generation sequencing approach.
| Search Term Queries                                      | Greengenes+ Vibrionales+ Pseudoalteromonadaceae | Greengenes+ Vibrionales \(^a\) | Total |
|---------------------------------------------------------|--------------------------------------------------|---------------------------------|-------|
| All papers fitting search criteria                      | 22                                               | 63                              | 85    |
| Papers confirmed using Greengenes versions 13_5 or 13_8 with known taxonomy errors | 14                                               | 41                              | 55    |
| Papers explicitly stating taxonomic mismatch (in text or supplemental material) \(^b\) | 10                                               | 10                              | 20    |
| Papers stating correct assignment or using earlier version of Greengenes | 7                                                | 10                              | 17    |
| Papers using Greengenes but with no information on database version. | 1                                                | 12                              | 13    |

\(^a\) Counts do not include papers appearing in the “Greengenes + Vibrionales + Pseudoalteromonadaceae” search

\(^b\) These papers are also included in the tally for papers using Greengenes 13_5 or 13_8