16S rRNA Multi-Copy Cluster and Analysis

Nathan Olson and Ethan Ertel

December 7, 2015

Abstract

Targeted genomic DNA amplification and sequencing of 16S rRNA is commonly used to characterize microbial communities in terms of identity and quantity. Current methods used to define community composition treat each sequence independently. However, microbial genomes may contain multiple copies of the 16S rRNA gene. Furthermore due to differential selective pressure, the gene sequence can vary between copies. The presence of both multiple different 16S sequences invalidates the assumption that clusters may be treated independently or used to uniquely identify species. Methods to correct for multiple copies only consider the gene copy number and not sequence diversity. We analyzed 16S rRNA genes sequences from 2943 GenBank reference genomes and show that 16S genes from one organism may be represented by many clusters, and that the number of gene copies in a cluster can vary by strain. As a initial attempt to develop a method that not only corrects for copy number but also taxonomic classification, we applied a simple set of linear constraints. The linear constriants were unable to model the copy number mixed taxonomy clusters, we conclude the report with a discussion of other potential models that can be applied to address this issue.

Introduction

Microbial community surveys aim to characterize the identity and abundance of prokaryotic organisms within a sample. One method for characterizing microbial communinties is by selectively sequencing the marker genes from prokaryotic genomes in the sample. Genes that produce ribosomal RNA, specifically the 16S rRNA gene, is most commonly marker gene because as it contain both conserved and variable regions which faciliate amplification and taxonomic identification respectively (Wooley, Godzik, and Friedberg 2010). Thus, these genes are generally favored over other possible sequences despite disadvantages including, multiple gene copies within a genome and taxonomic ambiguity due to horizontal gene transfer (Vos et al. 2012).

16S rRNA metagneomic datasets are commonly analyzed by clustering sequences based on a defined similarity. Edit distance, or the number of changes required to convert one sequence into another, is used for computing pairwise similarity (Ghodsi, Liu, and Pop 2011). Clusters are commonly termed as operational taxonomic units (OTUs); OTUs are considered to be proxies for organismal abundance (Wooley, Godzik, and Friedberg 2010). The number of sequences assigned to an OTU, OTU counts, are used to determine the relative abundance of OTUs within a community. Multiple 16S rRNA gene copies will bias the OTU counts as the number of sequences is not equivalent to the abundance of cells within the sampled population (Kembel et al. 2012). Databases characterizing the number of gene copies per prokaryotic genome exist (Stoddard et al. 2014; Perisin et al. 2015). However, these databases do not include the 16S rRNA gene copy sequences, only gene copy number. A number of methods to normalize OTU count values based on copy number estimates have been developed (Kembel et al. 2012; Angly et al. 2014; Perisin et al. 2015). All of these methods use a single copy number correction value for individual OTUs, therefore prokaryotic strains with different copy numbers assigned to a single OTU can bias the correction factor. Copy number correction is further challenged as gene copies within a genome, either due to mutations or horizontal gene transfer, may be assigned to multiple OTUs (Pei et al. 2010; Koeppel and Wu 2013). Additionally, due to horizontal gene transfer, a 16S rRNA gene copy in a genome can be more similar to a 16S rRNA gene from unrelated taxa than to other gene copies within the same genome. No cluster taxonomic annotation methods take account single genome multi-cluster assignments.

To address this issue, we developed a 16S rRNA gene copy sequence database. The database was then used to characterize the extent to which the number of gene copies per genome assigned to a cluster and number of clusters individual genomes were assigned. We then attempted to develop a new method for copy number

correction accounting for the differences in the number of assigned gene copies per cluster and genomes with multiple cluster assignments.

Methods

To generate the database we downloaded 2943 microbial genomes and their taxonomic information from NCBI RefSeq (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria) (Pruitt et al. 2012). 16S rRNA gene sequences were extracted from the genomes using rnammer (http://www.cbs.dtu.dk)(Lagesen et al. 2007). The extracted 16S rRNA genes were then clustered with dnaclust (Ghodsi, Liu, and Pop 2011). The sequences were clustered at a range of threshold values, where similarity scores for all pairwise comparisons of sequences within a cluster must be above the threshold value. Similarity score equals 1 - (edit distance)/length(shortest string), and ranges from 0 to 1. Typical threshold values used in 16S rRNA metagenomic studies range from 0.95 to 0.99; however, we investigate thresholds as low as 0.61 and as high as 1.00. The 16S rRNA gene sequences were annotated using the NCBI taxonomy database. An SQLite database was first generated using the taxit python package (http://fhcrc.github.io/taxtastic). Characterization of the number of gene copies per genome assigned to a cluster and number of assigned clusters gene copies for a genome was done using the R statistical programming language (R Core Team 2015). A phylogenetic tree was used to further characterize the phylogenetic distribution to genomes with 16S rRNA gene copies assigned to multiple clusters. Multiple sequence alignment and phylogeny reconstruction methods used were based on thoes use in (Kembel et al. 2012). Multiple sequence alignment performed using infernal (Nawrocki and Eddy 2013) using the reference alignment for bacterial small subunit rRNA (http://rfam.xfam.org).

See `multicopy_taxa_code.pdf` for methods used to generate the database including bash scripts and commands. Source code for this report and scripts used in the study are available at https://github.com/nate-d-olson/multicopy_cluster_analysis.

Results and Discussion

16S rRNA Multicopy Database

A database of 16S rRNA gene copy sequences was generated by extracting 16S rRNA sequences from all closed bacterial genomes in the NCBI RefSeq database. The number of 16S rRNA gene copies per genome ranged from 1 to 16 with a median of 3 (Figure 1). The distribution of 16S rRNA gene copy numbers agrees with previous studies (Angly et al. 2014; Větrovsky and Baldrian 2013), excluding a genome with 16 gene copies. Only a single method was used to identify 16S rRNA gene copy sequences, and therefore, may contain incorrectly classified sequences, as indicated by outliers in the database phylogenetic tree (Figure 2). Therefore, the results below are presented as a proof of concept. A more rigorous approach to generating the sequence copy database similar to that used in Angly et al. (2014) would help validate the database.

Clustering

The extracted 16S rRNA genes were clustered with dnaclust using a range of threshold values from 1.00 (identical gene sequences) to 0.61 (sequences with 61% similarity). For lower clustering (< 0.73) thresholds the majority of the sequences were assigned to a single cluster (Figure 3).

Cluster Copy Number

The number of 16S rRNA gene copies from a single genome within a cluster varies. The variability in copy number within a cluster biases copy number correction when only a single value is used (Figure 4). Single cluster copy number is currently the only method used in copy number correction (Kembel et al. 2012; Angly et al. 2014; Stoddard et al. 2014; Perisin et al. 2015). CopyRighter (Angly et al. 2014), attempts to address

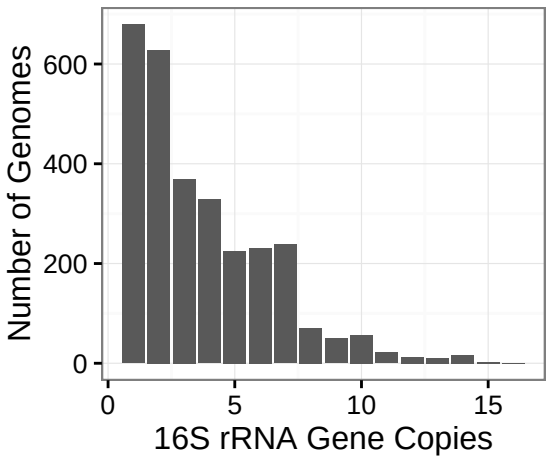

Figure 1: Distribution in the number of gene copies present in a bacterial genome.

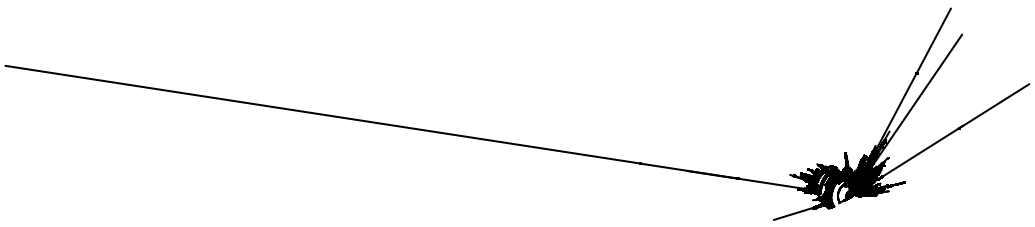

Figure 2: Phylogeny of 16S rRNA gene copy database used in this study. Long branch length indicate potential sequences in the database that were incorrectly identified as 16S rRNA genes by rnammer.

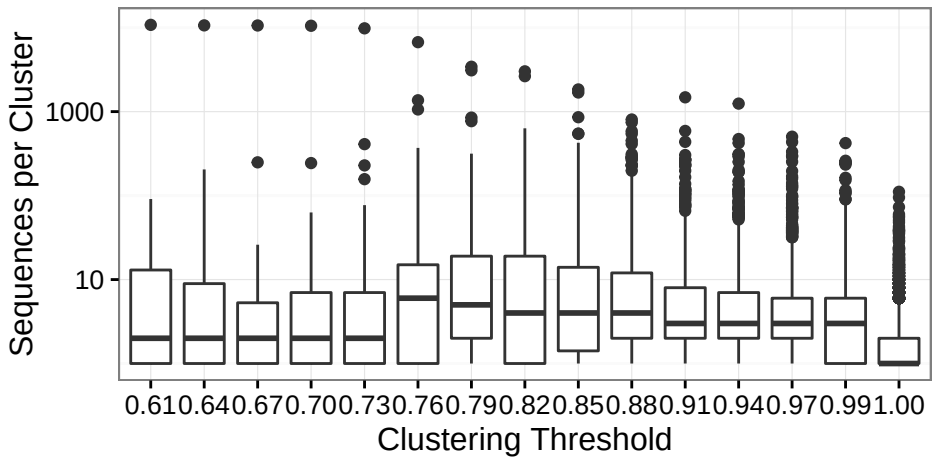

Figure 3: Distribution of cluster size, number of sequences assigned to a cluster, by clustering threshold.

this issue by defining the cluster copy number correction value proportionally based on the copy number of the species assigned to a cluster.

Assignment of 16S rRNA gene sequences from a single genome to multiple clusters further confounds the use of single copy nuber correction values. While 16S rRNA gene copies from a single genome are assigned to only one cluster, a number of genomes have 16S rRNA gene sequences assigned to multiple clusters (Figure 5). Lowering the clustering threshold has little effect on the number of clusters 16S rRNA gene copies within a a genome are assigned. The database used in this analysis was generated using only a single method for 16S rRNA gene extraction. Visualization of the database phylogeny indicates that a number of outlier sequences are included in the database and is likely responsible for a few of the multiple cluster genome assignments, especially those with multiple assigned clusters when using lower threshold values.

Cluster Taxonomic Ambiguity

Greater than 5% cluster classification error rate was observed for genus and species level classifications when using for 97% clustering threshold, (Figure 6). For clustering thresholds less than 0.97, the proportion of taxonomically ambiguous clusters is greater than 0.06. This taxonomic ambiguity can lead to biases in cluster taxonomic classifications.

Phylogenetic Analysis

When clustering at the 0.97 threshold and using genus level taxonomic annotation, sequences belonging to clusters with multiple assigned taxa are distributed throughout the phylogenetic tree (Figure 7). The wide distribution of sequences assigned to ambiguous clusters indicates that ambiguity in cluster taxonomic annotation challenges taxonomic classification issues for a majority of the taxonomic lineages.

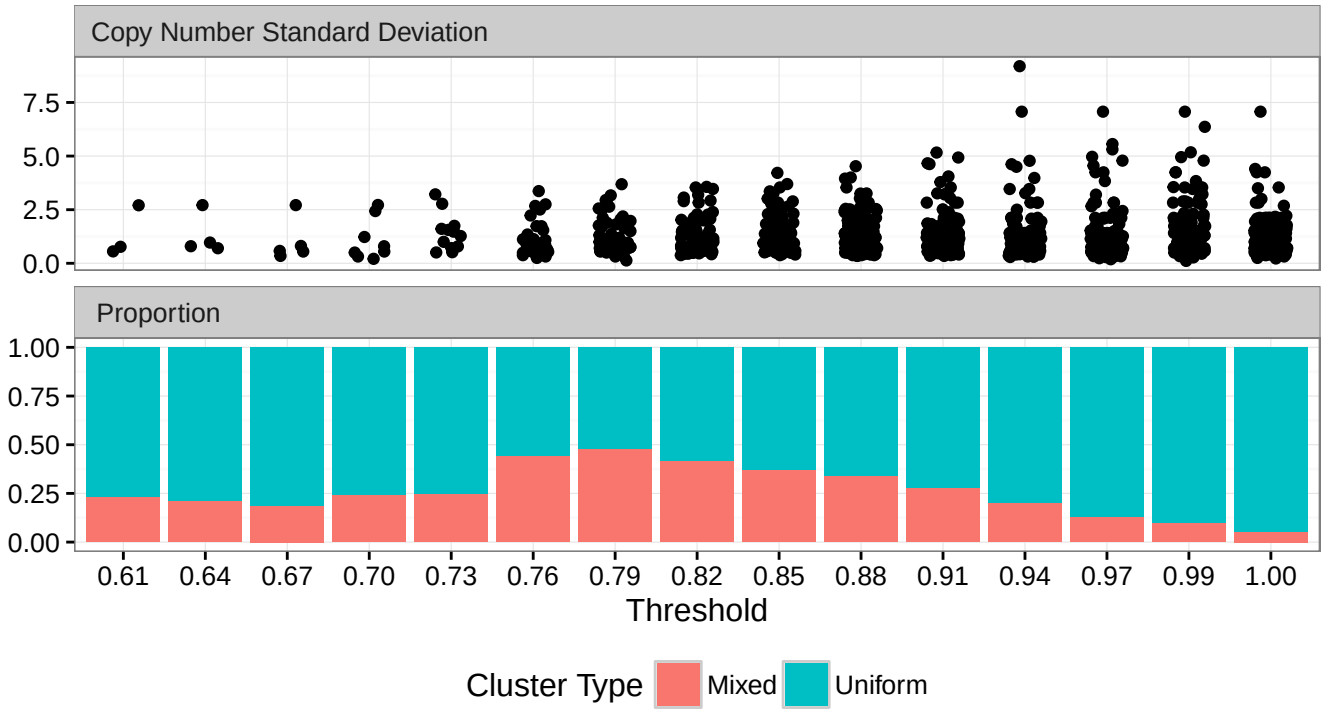

Figure 4: Variability number of 16S rRNA copies for sequences assigned to a cluster by genome. Top plot is the standard deviation of the copy numbers and bottom plot is the proportion of cluster with multiple copy numbers (Mixed), and single copy number values (Uniform)

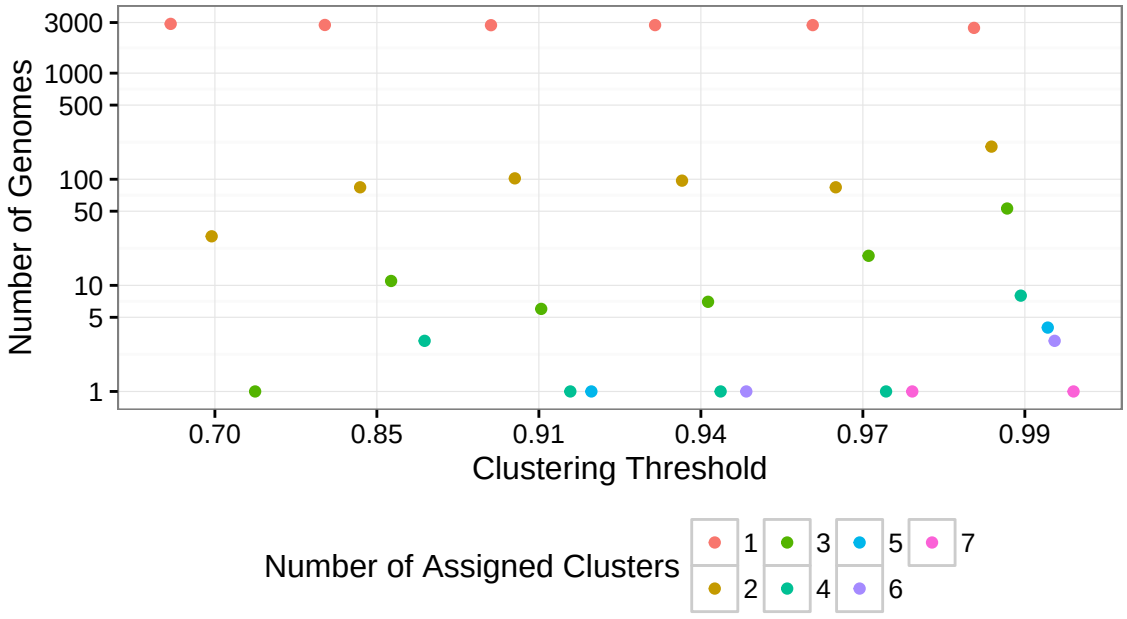

Figure 5: Number of genomes assigned to multiple clusters by clustering threshold.

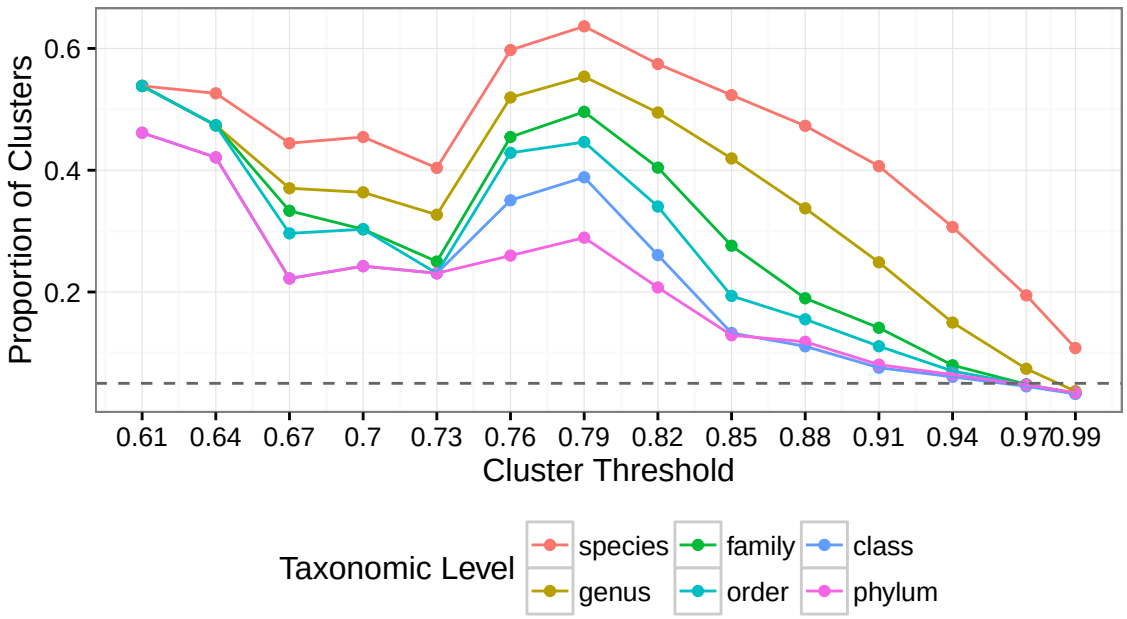

Figure 6: Proportion of clusters with more than one assigned taxa by clustering threshold and taxonomic level.

Resolving Copy Number and Taxonomic Cluster Ambiguity

A number of approaches may use reference-based 16S multicopy information to guide taxonomic binning. If one assumes consistent and unbiased detection of 16S PCR products, the relationship between clustered OTUs 1..n and reference based counts of species 1..m follows the structure presented in Equation 1.

$$\begin{bmatrix} copy_{OTU\#1,Species\#1} & \dots & copy_{OTU\#1,Species\#m} \\ \vdots & \ddots & \vdots \\ copy_{OTU\#n,Species\#1} & \dots & copy_{OTU\#n,Species\#m} \end{bmatrix} \begin{bmatrix} count_{Species\#1} \\ \vdots \\ count_{Species\#m} \end{bmatrix} = \begin{bmatrix} count_{OTU\#1} \\ \vdots \\ count_{OTU\#n} \end{bmatrix}$$

Equation 1 A simple linear relationship between 16S rRNA copy and species count underlies the physical system

In ideal conditions, organism identification and cluster disambiguation could be treated as a linear optimization problem under these conditions by defining a suitable objective function (c). Even in this case, the n by m matrix on the left hand side of this equation represents a very high dimensional and sparse space that will in a vast majority of situations have multiple optimal solutions. So, the linear optimization approach to multicopy guided taxon quantification may have limitations even under ideal measurement conditions. Indeed, our implementation of this approach (shown in classify_by_constraint.R) yielded incorrect classifications.

Several other major shortcomings hinder the success of this naive approach. PCR amplification is not uniform or linear, and so certain sequences will be overrepresented in such a way that the linear relationships no longer hold. In addition, the number of different OTUs detected is known to be a function of sequencing depth (Paulson et al. 2013). Finally, the prediction of species counts is biased by the numbers and types of species present in the GenBank reference database.

One possible structure of the relationship between measured counts and taxa present is given in Equation 2. Using a relationship of this type requires a substantial amount of careful work and a great deal of additional

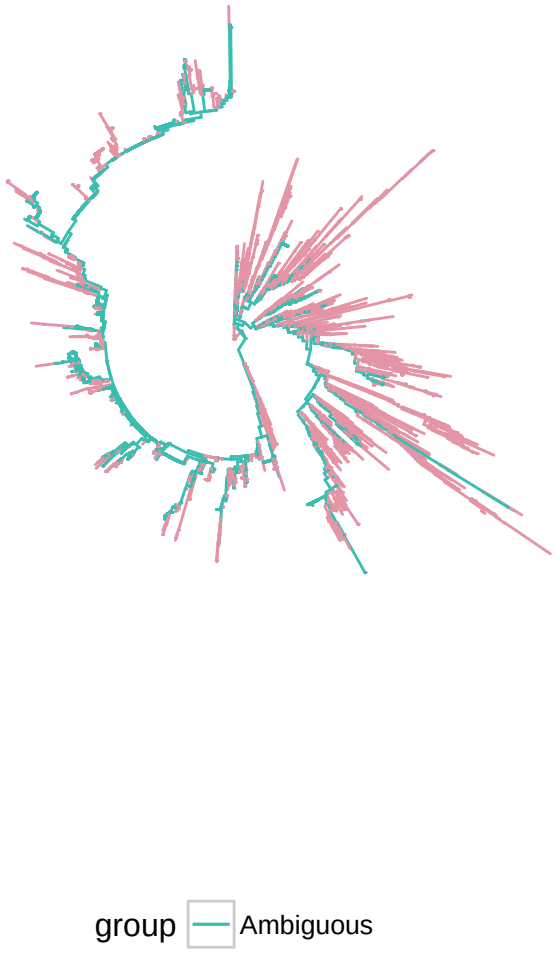

Figure 7: 16S rRNA gene phylogeny, teal braches indicate 16S rRNA gene sequences assigned to taxonomically ambiguous clusters, at the 0.97 clustering threshold and genus level taxonomic assignment. To provide better branch resolution in the phyogenetic tree, potential outlier 16S rRNA sequences were excluded from the phylogenetic tree below

data for the purpose of estimating parameters with each taxon/OTU pairing. Additional mathematical work is required to determine effective estimators for each parameter and to formulate the problem more precisely. The information gained by this model would support the use of expectation maximization to optimize taxonomic binning.

$$E\left[count_{OTU\#i}\right] = E\left[\sum_{j=1}^{m} PD_{ij}(depth) \cdot (c_{ij} + \epsilon) \cdot x_j\right]$$

Equation 2 The expected number of sequences in OTU i equals the sum of the weighted expectation for each component mapping to OTU i. These expectations depend on PD_{ij} (the probability of detection in a transcript from species j), number of copies of the relevant sequence, and statistical noise in the observation.

Simpler models may be effective in directing quantification and binning. An approach used for RNA-Seq quantification(Patro, Mount, and Kingsford 2014), uses relatively simple expectations to guide quantification of sequences based on even distribution of k-mers on those reads. In our case, OTUs may be collapsed into OTU equivalence classes, consisting of taxa mapping to an OTU sequence which have similar copy number of 16S sequences clustered to that OTU. The approach seeks to attain nearly even distribution of OTU counts over a single taxon; thus, assignment of equivalence OTUs to any taxon supports the assignment of other equivalence OTUs to that taxon.

Conclusion

Based on our initial analysis of 16S rRNA multi-copy sequences, clusters can be contain sequences from genomes with different copy numbers, sequences from a single genome maybe assigned to different clusters, and sequences from multiple taxa may be assigned to the same cluster. These three issues indicate that single-value cluster copy number correction methods may produce biased results, though the extent to this bias was not assessed. Although our attempt to develop methods to correct for this bias were unsuccessful, a copy number correction method that accounts for copy number and taxonomic cluster ambiguity could improve the accuracy of 16S rRNA microbial community analysis. The 16S rRNA gene sequences in our database were only identified using a single method and likely includes non-16S rRNA sequences. Finally, 16S rRNA sequencing methods only target part of the gene. Further analysis of taxonomic and copy number ambiguity should include a charaterization of the 16S rRNA gene regions most commonly used in 16S rRNA metagenomic studies.

References

Angly, Florent E, Paul G Dennis, Adam Skarshewski, Inka Vanwonterghem, Philip Hugenholtz, and Gene W Tyson. 2014. "CopyRighter: A Rapid Tool for Improving the Accuracy of Microbial Community Profiles Through Lineage-Specific Gene Copy Number Correction." *Microbiome* 2 (1). Springer: 1–13.

Ghodsi, Mohammadreza, Bo Liu, and Mihai Pop. 2011. "DNACLUST: Accurate and Efficient Clustering of Phylogenetic Marker Genes." *BMC Bioinformatics* 12 (1). BioMed Central Ltd: 271.

Kembel, Steven W, Martin Wu, Jonathan A Eisen, and Jessica L Green. 2012. "Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance." *PLoS Computational Biology* 8 (10). Public Library of Science: e1002743. doi:10.1371/journal.pcbi.1002743.

Koeppel, Alexander F, and Martin Wu. 2013. "Surprisingly Extensive Mixed Phylogenetic and Ecological Signals Among Bacterial Operational Taxonomic Units." *Nucleic Acids Research*. Oxford Univ Press, gkt241.

Lagesen, Karin, Peter Hallin, Einar Andreas Rødland, Hans-Henrik Staerfeldt, Torbjørn Rognes, and David W Ussery. 2007. "RNAmmer: consistent and rapid annotation of ribosomal RNA genes." *Nucleic Acids*

Research 35 (9): 3100–3108. doi:10.1093/nar/gkm160.

Nawrocki, Eric P, and Sean R Eddy. 2013. "Infernal 1.1: 100-fold faster RNA homology searches." *Bioinformatics (Oxford, England)* 29 (22): 2933–5. doi:10.1093/bioinformatics/btt509.

Patro, Rob, Stephen M Mount, and Carl Kingsford. 2014. "Sailfish Enables Alignment-Free Isoform Quantification from RNA-Seq Reads Using Lightweight Algorithms." *Nature Biotechnology* 32 (5). Nature Publishing Group: 462–64.

Paulson, Joseph N, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. 2013. "Differential Abundance Analysis for Microbial Marker-Gene Surveys." *Nature Methods* 10 (12). Nature Publishing Group: 1200–1202.

Pei, Anna Y, William E Oberdorf, Carlos W Nossa, Ankush Agarwal, Pooja Chokshi, Erika A Gerz, Zhida Jin, et al. 2010. "Diversity of 16S RRNA Genes Within Individual Prokaryotic Genomes." *Applied and Environmental Microbiology* 76 (12). Am Soc Microbiol: 3886–97.

Perisin, Matthew, Madlen Vetter, Jack A Gilbert, and Joy Bergelson. 2015. "16Stimator: Statistical Estimation of Ribosomal Gene Copy Numbers from Draft Genome Assemblies." *The ISME Journal*. Nature Publishing Group.

Pruitt, Kim D, Tatiana Tatusova, Garth R Brown, and Donna R Maglott. 2012. "NCBI Reference Sequences (RefSeq): Current Status, New Features and Genome Annotation Policy." *Nucleic Acids Research* 40 (D1). Oxford Univ Press: D130–D135.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Stoddard, Steven F, Byron J Smith, Robert Hein, Benjamin RK Roller, and Thomas M Schmidt. 2014. "RrnDB: Improved Tools for Interpreting RRNA Gene Abundance in Bacteria and Archaea and a New Foundation for Future Development." *Nucleic Acids Research*. Oxford Univ Press, gku1201.

Větrovsky, Tomáš, and Petr Baldrian. 2013. "The Variability of the 16S RRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses." *PLoS One* 8 (2): e57923.

Vos, Michiel, Christopher Quince, Agata S Pijl, Mattias de Hollander, and George A Kowalchuk. 2012. "A Comparison of RpoB and 16S RRNA as Markers in Pyrosequencing Studies of Bacterial Diversity." *PLoS One* 7 (2). Public Library of Science: e30600.

Wooley, John C, Adam Godzik, and Iddo Friedberg. 2010. "A Primer on Metagenomics." *PLoS Comput Biol* 6 (2): e1000667.