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. Include some kind of results - 5% of ambiguous clusters 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

Surveys of microbial communities aim to quantify the abundance of constituent microbes through the measurement of amplified segments of genomic DNA. Genes that produce ribosomal RNA, specifically the 16S rRNA sub-sequence, are favored targets for amplification and measurement because their high degree of conservation makes these genes an attractive basis for phylogenetic analysis and taxonomic identification. 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 useful as a metric of pairwise similarity (Ghodsi, Liu, and Pop 2011). Clusters are defined as OTU, operational taxonomic units, and OTU counts are the number of sequences assigned to a cluster. Clusters are commonly termed as operational taxonomic units (OTUs); OTUs are considered to be proxies for organismal abundance (Wooley, Godzik, and Friedberg 2010). 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 population being sampled assigned to the cluster (Kembel et al. 2012). Databases have been developed to characterize the number of gene copies per prokaryotic genome have been developed (Stoddard et al. 2014,Perisin et al. (2015)), however these databases do not include the 16S rRNA gene copy sequences. A number of methods to normalize the OTU count values based on copy number estimates have been developed (Kembel et al. 2012,(**???**),Perisin et al. (2015)). All of these methods use a single copy number correction value for individual OTUs, therefore bacterial strains with different copy numbers assigned to a single OTU can bias the correction factor. The mutiple gene copy number and diversity of the gene copies within a genome either due to mutations or horizontal gene transfer (Pei et al. 2010,Koeppel and Wu (2013)) futher challenges copy number correction as assignment of gene copies to multiple clusters is not accounted for. Additionally, due to horizontal gene transfer a 16S rRNA gene copy is a genome can be more similar to a 16S rRNA gene from a unrealted taxa than to other gene copies within the same genome. No cluster taxonomic annotation methods take account for single genome multiple cluster assignment.

To address this issue we have developed a 16S rRNA gene copy sequence database. The extent to whic h the number of gene copies per genome assigned to a cluster and number of clusters individual genomes were

assigned was characterized using the gene copy sequence database. We then attempted to develop a new method for copy number correction that accounts for the different numbers of assigned gene copies per cluster and genomes with multiple cluster assignments.

Methods

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/cgi-bin/sw_request?rnammer)(Lagesen et al. 2007). The extracted 16S rRNA genes were clustered with dnaclust (Ghodsi, Liu, and Pop 2011). THe sequences were clustered for 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. Exploratory data analysis was performed in R.

See `multicopy_taxa_code.pdf` detailed methods including bash scripts and commands.

Results and Discussion

16S rRNA Multicopy Database

A database of 16S rRNA gene copy sequences was generated by extracting 16S rRNA gene sequences from all closed bacterial genomes in the NCBI RefSeq database. Number of 16S rRNA gene copies per genome ranged from 1 to 16 with a median of 3. The distribution of 16S rRNA gene copy numbers agrees with previous studies (**???**,(**???**)), exluding a genome with 16S rRNA gene copies. Only a single method was used to identify 16S rRNA gene copy sequences and therefore may contain sequences that were incorrectly identified as 16S rRNA genes. Therefore the results below are presented as a proof of concept. A more rigorus 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.

Cluster Copy Number

The number of 16S rRNA gene copies from a single genome within a cluster varies. The variabily in copy number within a cluster biases copy number correction when only a single value in copy number correction. Single cluster copy number value is currently the only method used in copy number correction (Kembel et al. 2012,Angly et al. (2014),(**???**),(**???**)). CopyRighter (Angly et al. 2014), attempts to address 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 confound 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. Lowering the clustering threshold has little effect on the number of clusters copies of the 16S rRNA gene is assigned to for a genome. 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

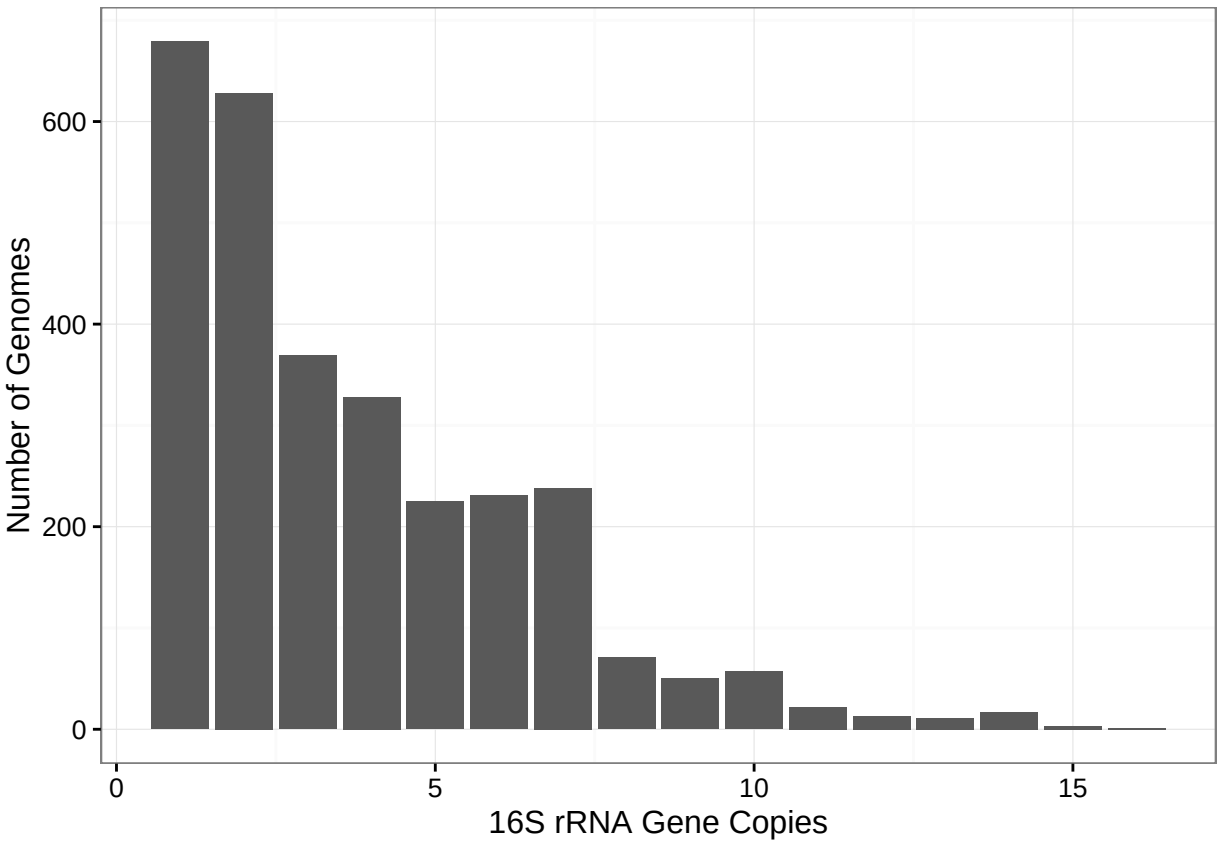

Figure 1: Distribution in the number of gene copies present in a bacterial genome.

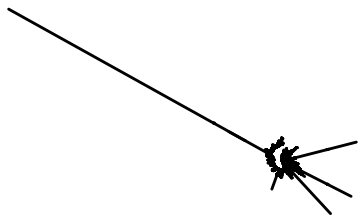

Figure 2: Phylogeny of 16S rRNA gene copy database used in this study.

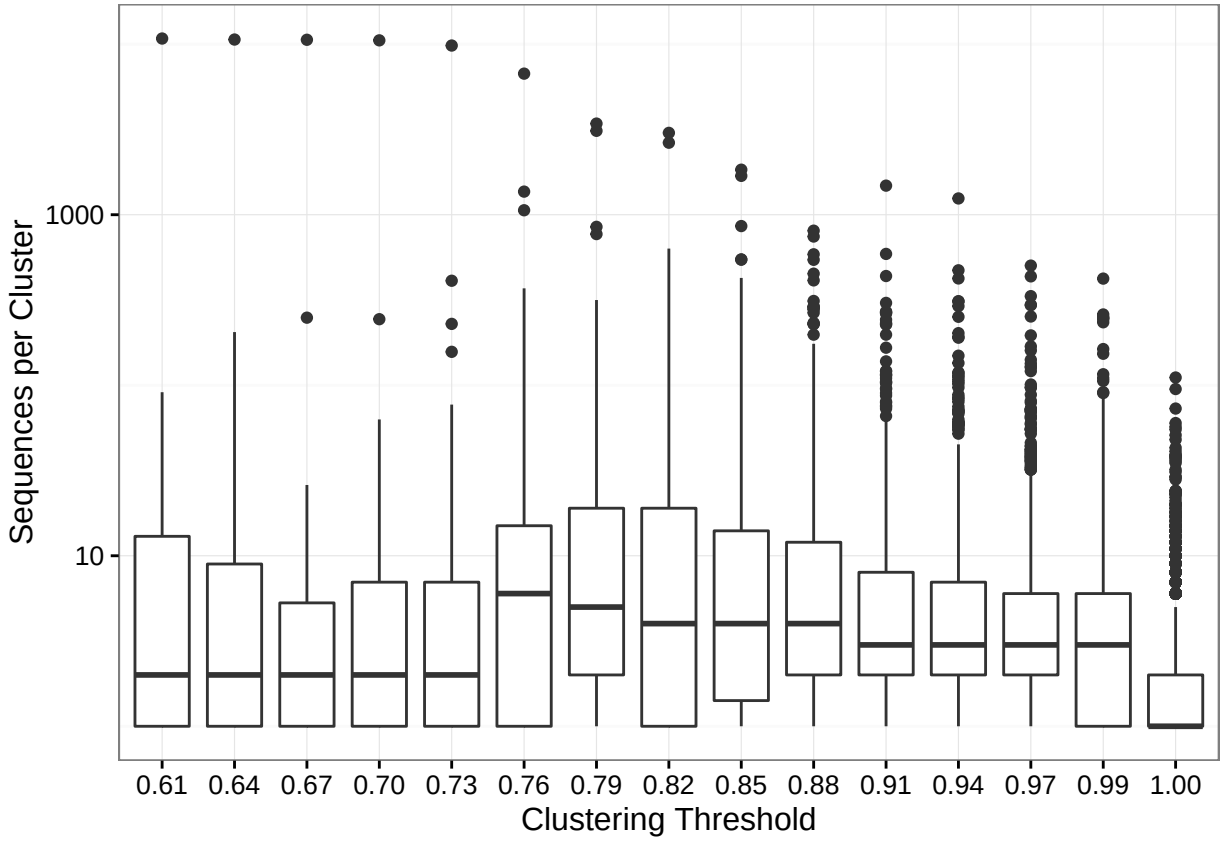

Figure 3: Distribution of cluster size, number of sequences assigned to a cluster, by clustering threshold.

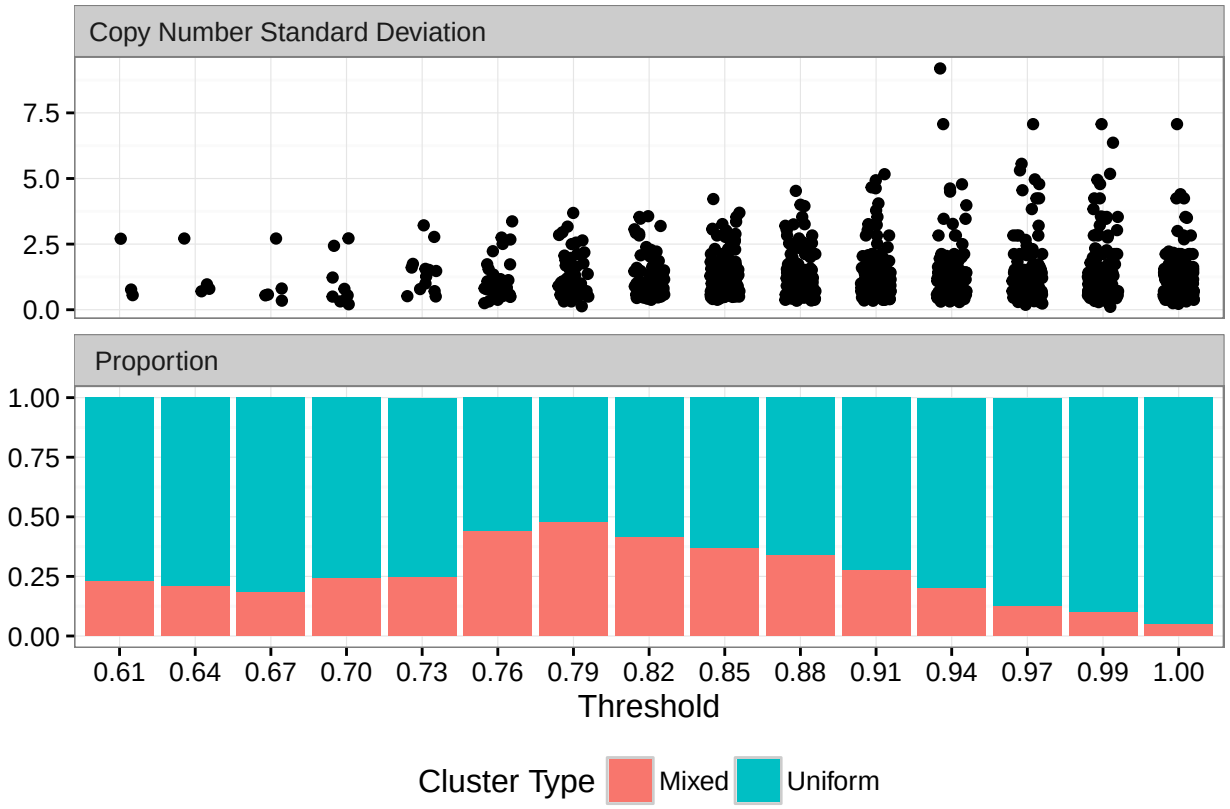

Figure 4: Variability number of 16S rRNA copies for sequences assigned to a cluster by genome. Top plot is the standard deviation of the copy numbers and bottom plot is the proportion of cluster with multiple copy numbers (Mixed), and single copy number values (Uniform)

included in the database which are likely responsible for a few of the multiple cluster genome sequence copy assignments, especially thoes with mutiple assigned clusters when using lower threshold values.

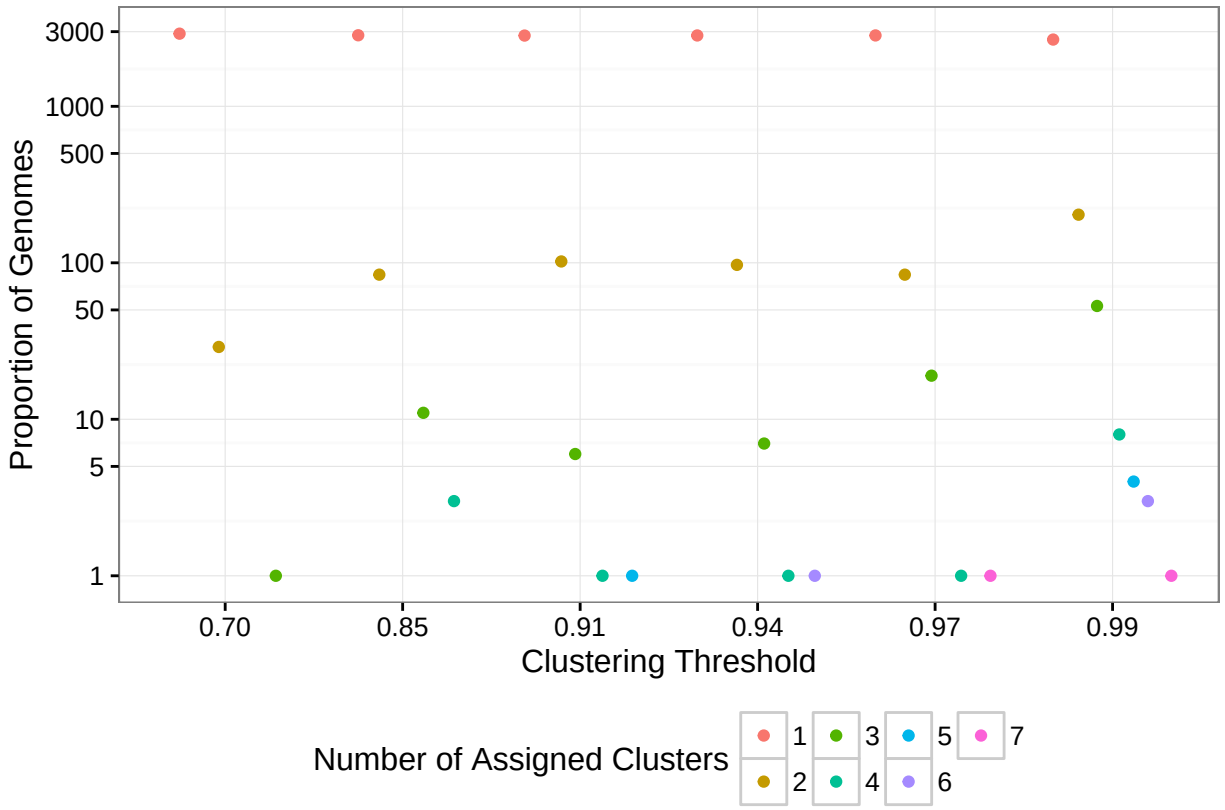

Figure 5: Proportion of genomes assigned to multiple clusters by clustering threshold.

Cluster Taxonomic Ambiguity

Taxonomic for 97% clustering threshold, >5% cluster classification error rate for genus and species level classifications. For clustering thresholds less than 0.97, the proportion of taxonomically ambiguous clusters is greater than 0.06. When analyzing 16S rRNA metagenomic datasets, the presence of taxonomically ambiguous clusters must be taken into consideration.

Phylogenetic Analysis

The phylogenetic distribution of ambiguous clusters. 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. The wide distribution of sequences assigned to ambiguous clusters indicates that ambiguity in cluster taxonomic annotation is challenges taxonomic classification issues for a majority of taxonomic lineages.

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

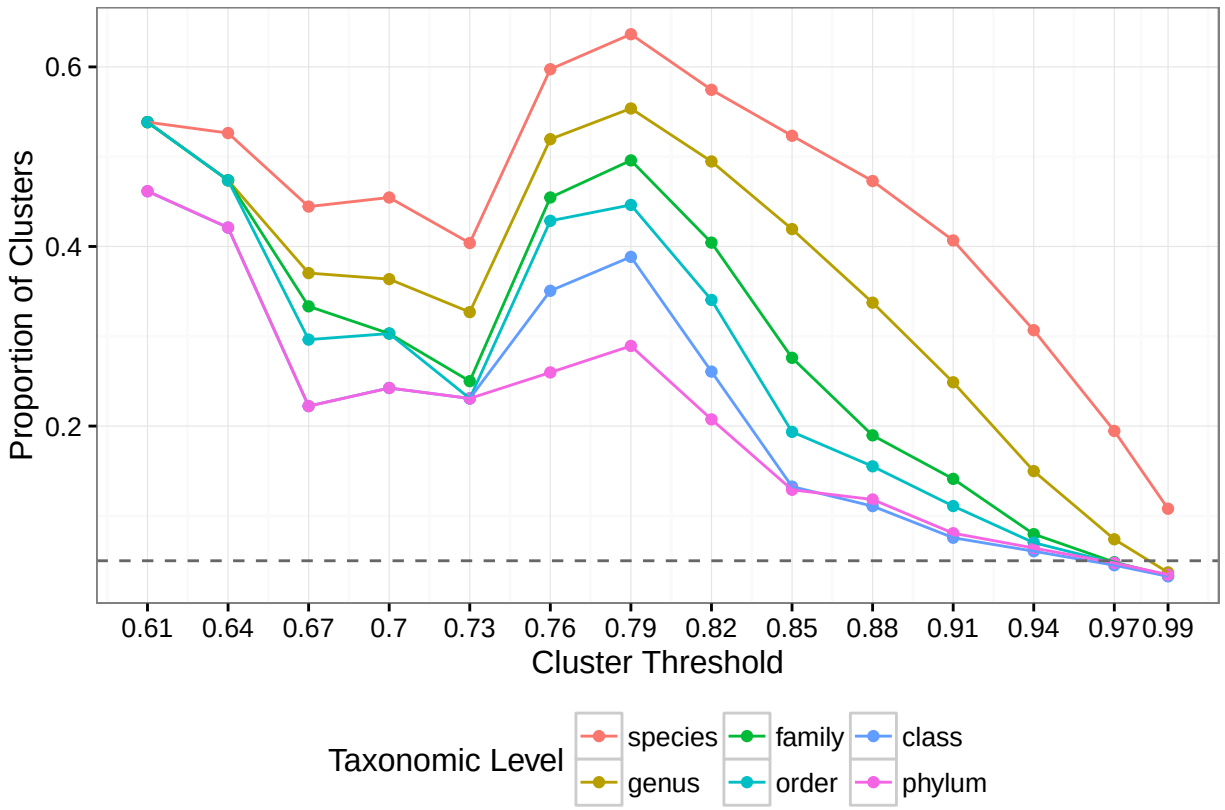

Figure 6: Proportion of cluster with more than one assigned taxa by clustering threshold and taxonomic level.

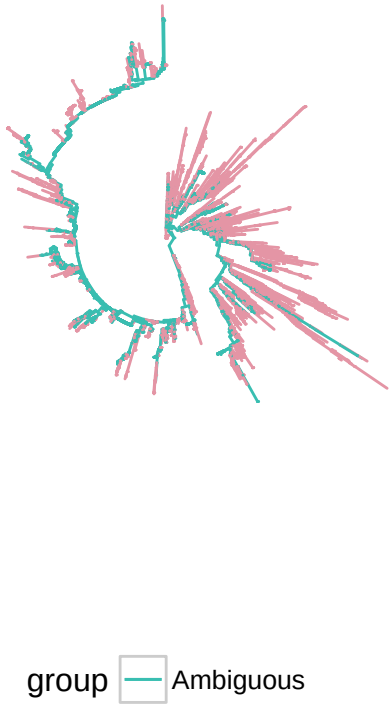

Figure 7: 16S rRNA gene phylogeny, teal braches indicate 16S rRNA gene sequences assigned to taxonomically ambiguous clusters, at the 0.97 clustering threshold and genus level taxonomic assignment. To provide better branch resolution in the phyogenetic tree, potential outlier 16S rRNA sequences were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

OTUs 1..n and reference based counts of species 1..m is: Equation (XXX): A simple linear relationship between 16S rRNA copy and species count underlies the physical system Organism identification could simply be treated as a linear optimization problem under these conditions if a suitable objective function c is defined.

However, several 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.

An alternative approach is to use

Equation (XXX): 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 PDij (the probability of detection in a transcript from species j), number of copies of the relevant sequence, and statistical noise in the observation.

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.

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.

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.

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.