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 16S rRNA genes which have diverged in evolution and sequence, which 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 analyze 16S sequences from 2943 GenBank reference genomes and show that one organism may be represented by many clusters, and that one cluster may also contain multiple species. [[XXX summary statistics]]. A simple set of linear constraints [[XXX may be total rubbish / might save the day and hope for all of humanity]].

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 and attractive basis for phylogenetic analysis and taxonomic identification. Thus, these genes are generally favored over other possible sequences despite disadvantages including (Vos et al. 2012). In practice, measured sequences are clustered by similarity and these clusters are given a taxonomic identification. 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 commonly termed as operational taxonomic units (OTUs); OTUs are considered to be proxies for organismal abundance (Wooley, Godzik, and Friedberg 2010).

However, 16S sequences are known to be present in multiple copies (Pei et al. 2010) and the presence of paralogous sequences presents the possibility that distantly related sequences may be identified as highly similar (Koeppel and Wu 2013). A number of issues arise from this fact. Firstly, a single organism may be counted multiple times in many unique clusters. Secondly, paralogous (XXX) genes present in multiple species may cluster together, confounding identification; genes from species A may be labeled as species B if both map to a cluster labeled with the taxonomic information of species B.

- Other studies that have characterized multi-copies
- Available methods correcting for gene copy number
- Sequence matching
    - rnammer - how it works and what it does
- Multiple sequence alignment
    - infernal - how it works and what it does
- Phylogenetic Tree construction
    - Maximum likelihood tree
    - Evolutionary model

- Sequence clustering
  - dnaClust - how it works and what is does

Need to read http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2432038/pdf/pone.0002566.pdf http://bergelson.uchicago.edu/?p=886

Methods

We downloaded 2943 microbial genomes from GenBank (Benson et al. 2000). 16S sequences were extracted using rnammer (http://www.cbs.dtu.dk/cgi-bin/sw_request?rnammer)(Lagesen et al. 2007). Duplicate genomes were removed using script XXX and sequences were indexed according to the GenInfo Identifier (gi) each read's respective source genome. These sequences were then XXX.

Extracted sequences were then clustered using DNAClust (Ghodsi, Liu, and Pop 2011) to determine OTU mapping of each sequence. Clustering was evaluated at various threshold values, where similarity score of each sequence must be above the threshold value to be admitted to a cluster. Similarity score equals 1 - (edit distance)/length[shortest string], and ranges from 0 to 1. Typical threshold values for 16S clusters range from 0.95 to 0.99; however, we investigate thresholds as low as 0.61 and as high as 1.00.

XXX Nate, do you know what you want to say about diversity analysis?

Methods

Generating 16S Copy Database

- Downloaded reference genomes from NCBI RefSeq (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria)
- Extracted 16S rRNA gene sequences from whole genomes using rnammer (http://www.cbs.dtu.dk/cgi-bin/sw_request?rnammer)(Lagesen et al. 2007).

```
extract_16S(){
    prefix=$(echo $fa | sed 's/.fna//')
    if [ ! -e "$prefix.fasta" ]
    then
        echo $prefix
        rnammer -S bacterial -m ssu -xml $prefix-16S.xml -gff $prefix-16S.gff -f $prefix-16S.fasta < $fa
    fi
}

N=8
(
for fa in */*fna; do
   ((i=i%N)); ((i++==0)) && wait
   extract_16S "$fa" &
done
)
```

- Annotated Sequences with NCBI taxonomy database
  - Generate taxonomic database using the taxit python package (http://fhcrc.github.io/taxtastic/index.html), database created on 11/30/2015.
  - `taxit new_database -d ncbi_taxa_db.sqlite`

    `16S-multicopy-cluster-annotation.Rmd`

Clustering 16S sequences

- Sequences clustered using dnaclust (http://dnaclust.sourceforge.net/)(Ghodsi, Liu, and Pop 2011)
  - Sequences clustered at multiple thresholds, to relate to different taxonomic levels

    ```
    for i in 1.00 0.99 0.97 0.94 0.91 0.88 0.85 0.82 0.79 0.76 0.73 0.70 0.67 0.64 0.61;
    do
    ../dnaclust_repo_release3/dnaclust -d -l -t 8 -e 999 -i all_refseq_16S.fasta -s $i > all_refsec
    done
    ```

Multiple Sequence alignment and phylogenetic analysis

- Multiple sequence alignment performed using infernal (Nawrocki and Eddy 2013).
  - Reference alignment for bacterial small subunit rRNA http://rfam.xfam.org.

```
cmalign --verbose --ifile info.txt --elfile seq_el.txt --sfile score.txt --tfile parsetrees.txt -o ../in
```

Phylogenetic tree construction

- Generation of phlyogenetic tree using RAxML (http://sco.h-its.org/exelixis/web/software/raxml/index.html)(Stamatakis 2014), based on methods used by (Kembel et al. 2012).

```
esl-reformat --rename ID phylip infernal_filtered/cmd_line_infernal_1.1_pfiltered.fasta > infernal_filte
```

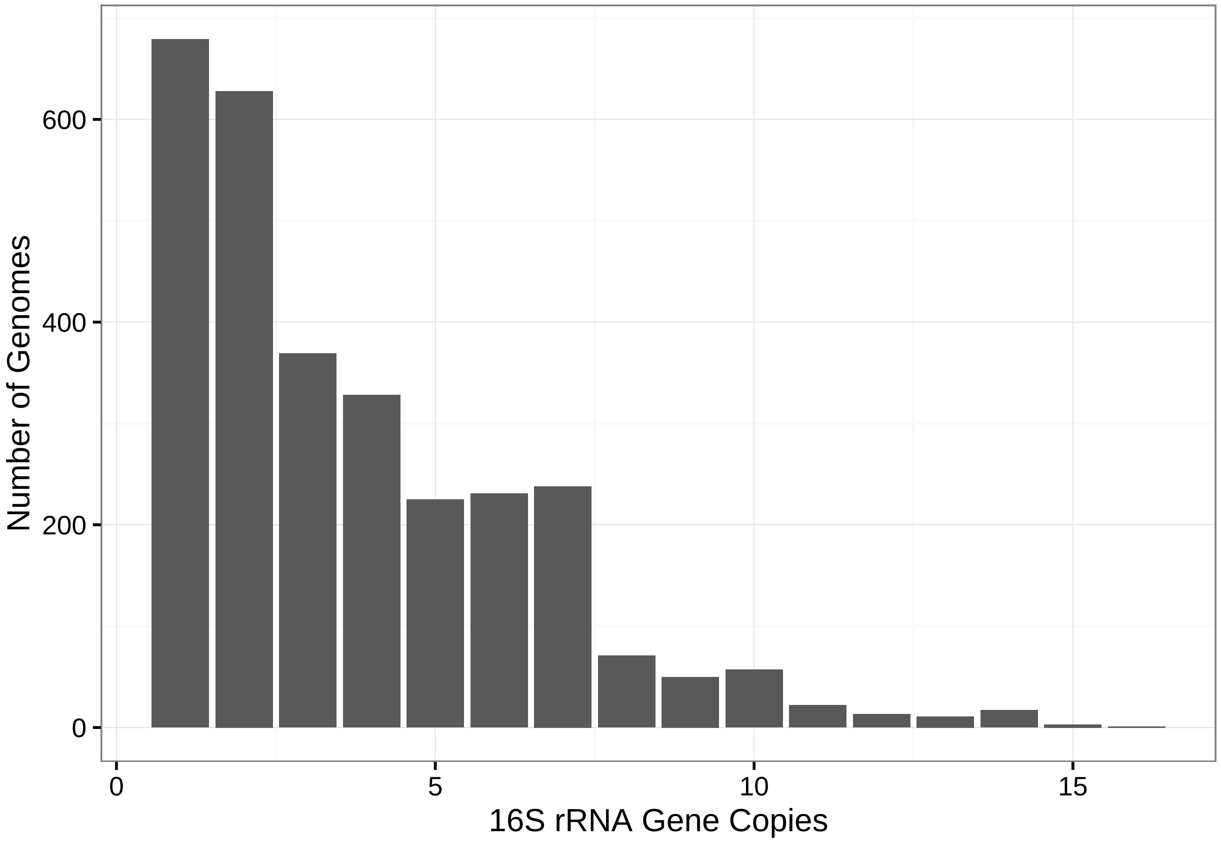
```
raxmlHPC-PTHREADS-SSE3 -m GTRGAMMA -T 7 -p 16 -n infernal_16S_tree -s infernal_filtered/cmd_line_inferna
```

Results

Summary of Dataset

Number of 16S rRNA gene copies per genome

```
ggplot(copy_count) + geom_bar(aes(x = copies)) +
    theme_bw() + labs(y = "Number of Genomes", x = "16S rRNA Gene Copies")
```



- 16S rRNA gene sequence diversity
  - scatter/ boxplot
    - y-axis = within genome diversity /similarity
    - x-axis = taxa levels
    - use boxplots to show level diversity summary

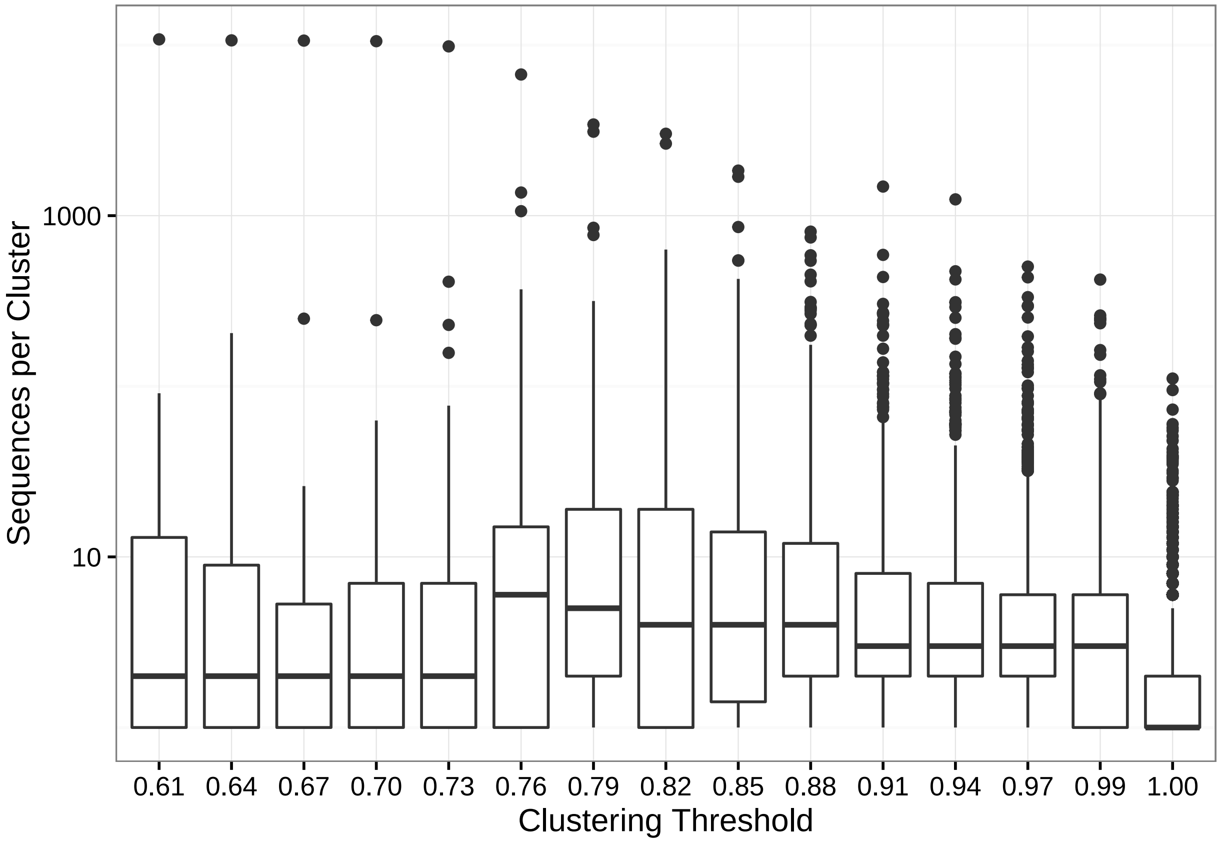
Clustering

```
* breakdown of number of clusters
* sequences assigned per cluster
```

Number of sequences in each cluster

```r
cluster_df_size <- cluster_df %>% group_by(threshold, id) %>% summarise(size = n())
```

```r
ggplot(cluster_df_size) + geom_boxplot(aes(x = threshold, y = size)) +
    scale_y_log10() + theme_bw() +
    labs(x = "Clustering Threshold", y = "Sequences per Cluster")
```



Number of gene copies per genome by cluster

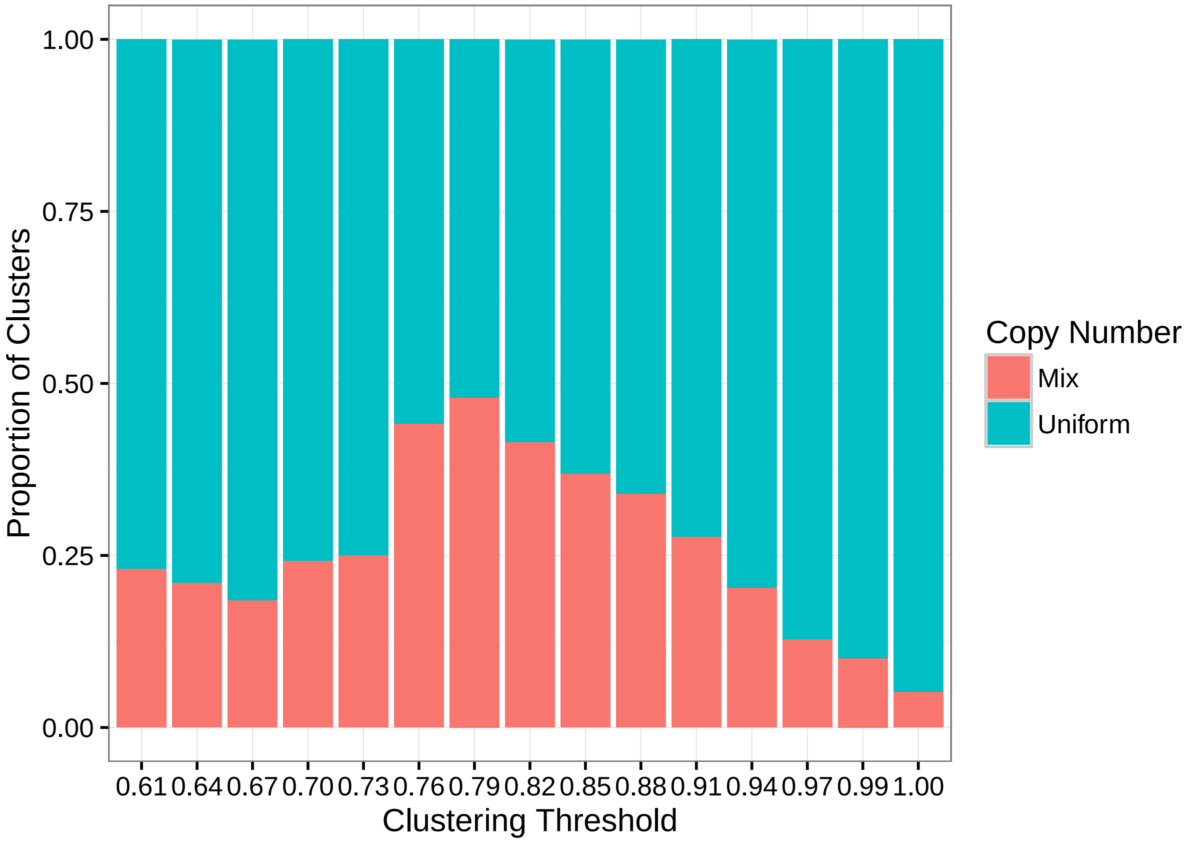
```
cluster_copy_number <- cluster_df %>% group_by(threshold, id, gi, count) %>% summarise(size = n())
    # check size should equal count
with(cluster_copy_number, sum(!(count == size)))
```

```
## [1] 0
```

```
## checks out
cluster_copy_number %<>% select(-size)
```

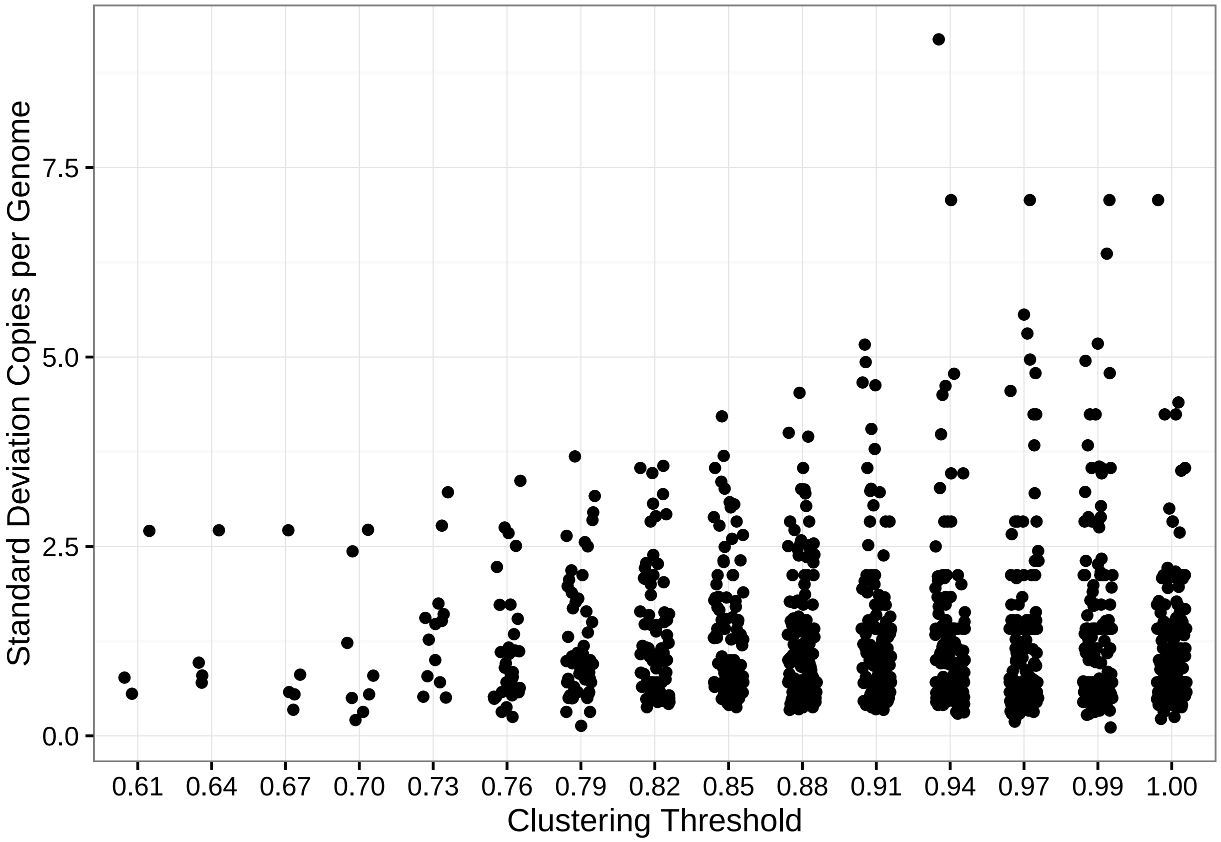
Variability in the number of 16S copies in a genome per cluster. This variability leads to biases when only a single value is used to correct abundance values for a cluster.
Proportion of clusters with multiple copy number values.

```
cluster_copy_number %>%
    group_by(threshold, id) %>%
    summarize(number_sd =sd(count)) %>%
    mutate(mixed_count = ifelse(number_sd == 0 | is.na(number_sd), "Uniform", "Mix")) %>%
ggplot() + geom_bar(aes(x = threshold, fill = mixed_count), position ="fill") + theme_bw() + labs(fill
```



Variability in copy number values

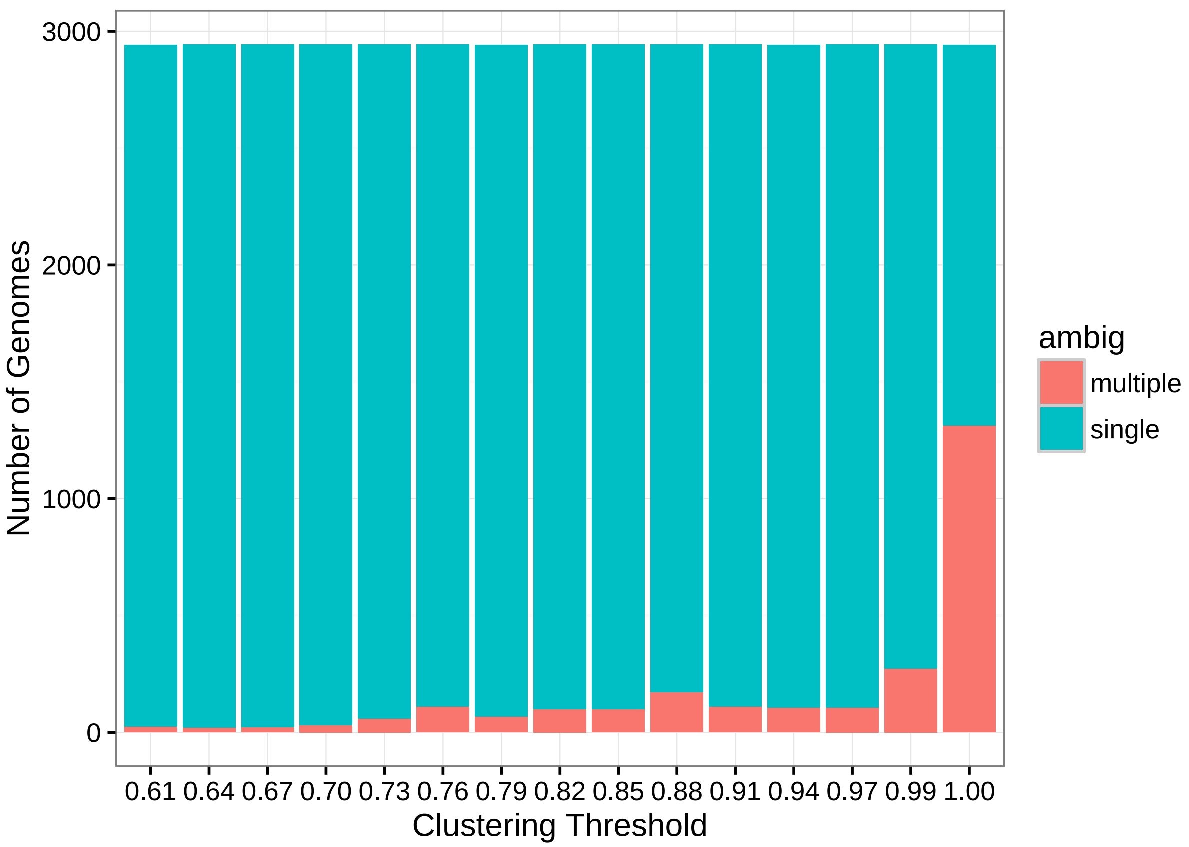
```
cluster_copy_number %>%
    group_by(threshold, id) %>%
    summarize(number_sd =sd(count)) %>%
    filter(number_sd > 0, !is.na(number_sd)) %>%
ggplot() + geom_jitter(aes(x = threshold,y = number_sd), position = position_jitter(width = 0.5)) +
    theme_bw() + labs(x = "Clustering Threshold", y = "Standard Deviation Copies per Genome")
```



Ambiguous clusters, clusters with more than one genome assigned

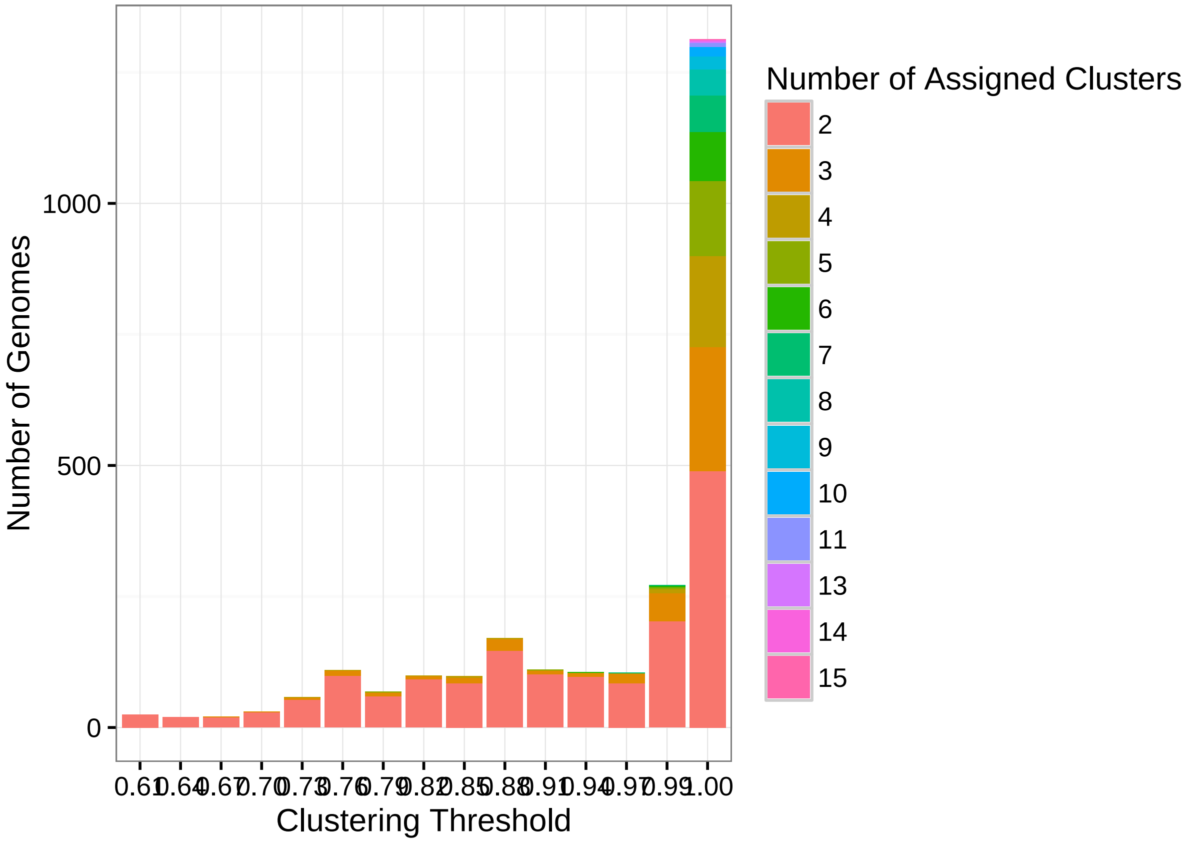
Proportion of genomes with 16S rRNA gene copies assigned to more than one cluster.

```r
cluster_df_ambig %>% ggplot() +
    geom_bar(aes(x = threshold, fill = ambig)) + theme_bw() +
    labs(x = "Clustering Threshold", y = "Number of Genomes")
```



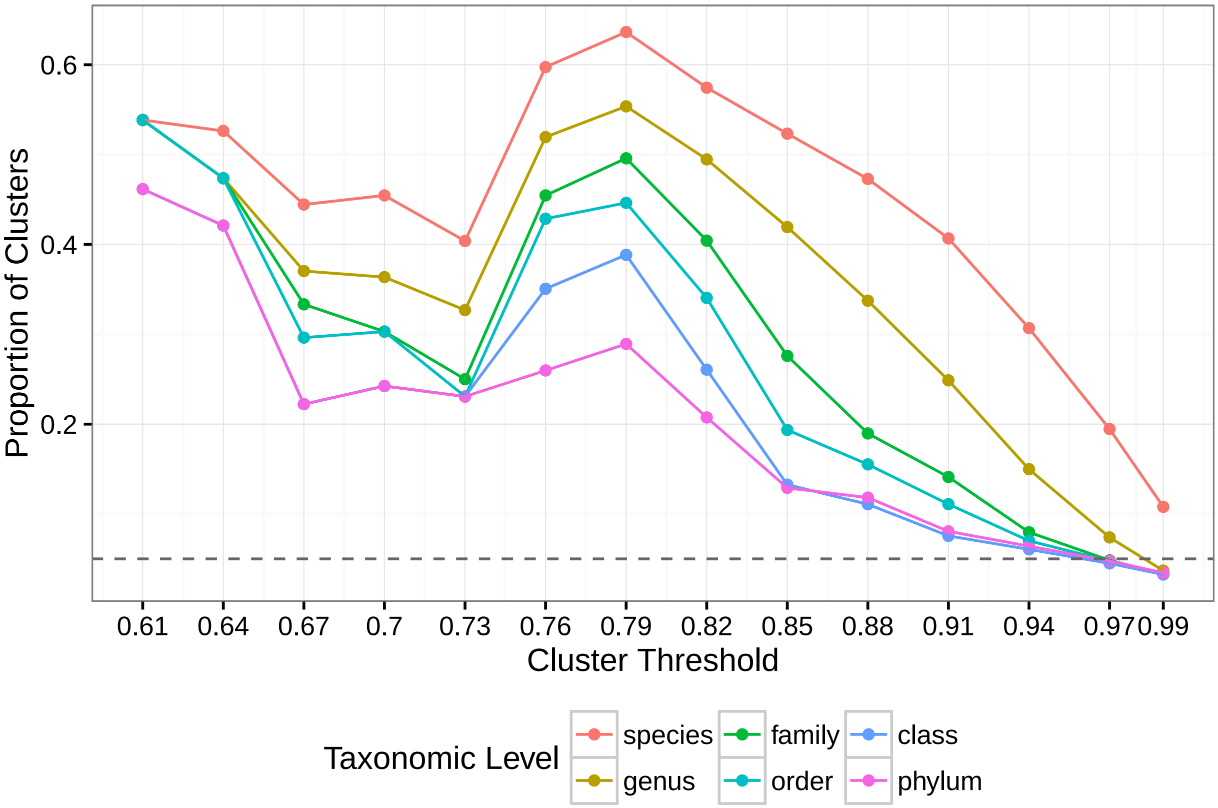
Number of clusters assigned, for genomes with more than one assigned gene copy.

```
cluster_df_ambig %>% filter(n_assigned > 1) %>%
    ggplot() + geom_bar(aes(x = threshold, fill = factor(n_assigned))) + theme_bw() + labs(x = "Clusteri
```



Taxonomic breakdown of clusters

Taxonomic for 97% clustering threshold, >5% cluster classification error rate for genus and species level classifi-



cations.

Phylogenetic tree - showing relationship between clusters and taxa assignments

Discussion

- Brief summary of findings
- Comparison of copy number results to published results
- When copy number correction can be applied ### When copy number correction does not work - ambiguous clusters
- Primer regions
  - This work focuses on the whole 16S gene
  - What are the expected results for single gene copies????

Recommendations

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 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

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.

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.

Stamatakis, Alexandros. 2014. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." *Bioinformatics (Oxford, England)* 30 (9): 1312–3. doi:10.1093/bioinformatics/btu033.

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.