Supplemental Material - *metagenomeFeatures*: An R package for working with 16S rRNA reference databases and marker-gene survey feature data.

Nathan D. Olson and Nidhi Shah

2019-02-02

Contents

1 Background

16S rRNA amplicon sequencing is commonly used for microbial community characterization, including differential abundance and diversity analysis. A limitation to 16S rRNA amplicon sequencing is a lack of taxonomic resolution, where organisms are only identifiable to the genus or family level. We define taxonomic resolution as the ability to differentiate between groups within a taxonomic level, for example, differentiating between species within a genus. We are interested in determining whether the 16S rRNA region of interest contains sufficient information for species-level taxonomic assignment. Taxonomic resolution varies by clade and amplicon regions. However, the extent to which taxonomic resolution varies is not well characterized.

There are multiple 16S rRNA databases each using their own file format and whose maintainers use different curation approaches resulting in different taxonomic and sequence compositions. One can perform taxonomic or other 16S rRNA sequence analysis on an individual database but with `metagenomeFeatures` and `MgDb` annotation packages the same analysis can easily be performed on multiple databases, increasing the power of the analysis. Here we demonstrate how `metagenomeFeatures` and the `MgDb` annotation packages can be used to characterize taxonomic resolution for the *Paenibacillus* genus and V12 and V4 amplicon regions. Originally, *Paenibacillus* was classified under the *Bacillus* genus, a novel genus was formed based on the 16S rRNA gene similarity in the 1990s. *Paenibacillus* spp. are facultative anaerobic bacteria present in a variety of environments including the soil, water, and can act as opportunistic pathogens in humans (Ouyang et al. 2008). It has been shown that *Paenibacillus* spp. will play an important role in sustainable agricultural industries (Grady et al. 2016). Thus, an appropriate speciation of this genus is of an interest to the community.

We used V12 and V4 region as they represent two commonly used amplicons for 16S rRNA marker-gene surveys. We will use the Greengenes 13.5 database, accessed using the `greengenes13.5MgDb`, Silva 128 database, accessed using the `silva128.1MgDb`, and RDP 11.5 database, accessed using the `ribosomaldatabaseproject11.5MgDb`, as annotation packages for our analysis of the *Paenibacillus* genus. First, we show the distribution of sequences belonging to *Paenibacillus* in these three databases, and then a way to combine different subsets of database sequences to build a custom database object. This helps investigators to pool sequences from different databases, to create an accurate and most representative database for their sample. This is especially helpful when a certain clade of interest is not very well represented in the databases. We then evaluate 16S rRNA amplicon sequencing taxonomic resolution for *Paenibacillus* species for V12 and V4 regions. In many analyses, the Greengenes 13.5 database is used for demonstration purposes but the other `MgDb` annotation packages such as RDP 11.5, SILVA 128, can also be used.

While the databases provide a consistent interface to the different 16S rRNA databases, the databases use different approaches to formatting the taxonomic names. As a result, the majority of the code used to produce this document is tidying the data. This code is not shown in the document but is available in the source Rmarkdown file which is provided as supplemental material and available in the metagenomeFeatures GitHub repository.[1]

2 Required Packages

In addition to `metagenomeFeatures`, `greengenes13.5MgDb`, `ribosomaldatabaseproject11.5MgDb`, and `silva128.1MgDb`, the `DECIPHER`, `tidyverse`, and `ggpubr` packages are also used in the following analysis. Our analysis uses the `DECIPHER` package to extract the amplicon regions, perform multiple sequence alignment, and generate a pairwise sequence distance matrix (Wright 2015). The `tidyverse` and `ggpubr` packages will be used to reformat the taxonomic and distance matrix data and generate summary figures (Wickham 2017 @ggpubr). See Session Information Section for package version information.

```r
library(tidyverse)
library(ggpubr)
library(UpSetR)
library(DECIPHER)
library(metagenomeFeatures)
library(greengenes13.5MgDb)
library(silva128.1MgDb)
library(ribosomaldatabaseproject11.5MgDb)
```

3 16S rRNA Database Comparison

We developed 16S rRNA MgDb annotation packages for three 16S rRNA databases, Greengenes, SILVA, and RDP. RNAcentral, a meta-database for non-coding RNA sequences provides cross database unique identifiers for sequences in the databases. The number of sequences in the databases and fraction of sequences with RNAcentralIDs varies for the three databases (Table 1). Using the RNAcentral IDs we can evaluate the overlap between the three databases (Fig. 1).

```
## Warning: `data_frame()` is deprecated, use `tibble()`.
## This warning is displayed once per session.
```

[1] https://github.com/HCBravoLab/metagenomeFeatures/blob/master/inst/manuscript/manuscript_supplemental.Rmd

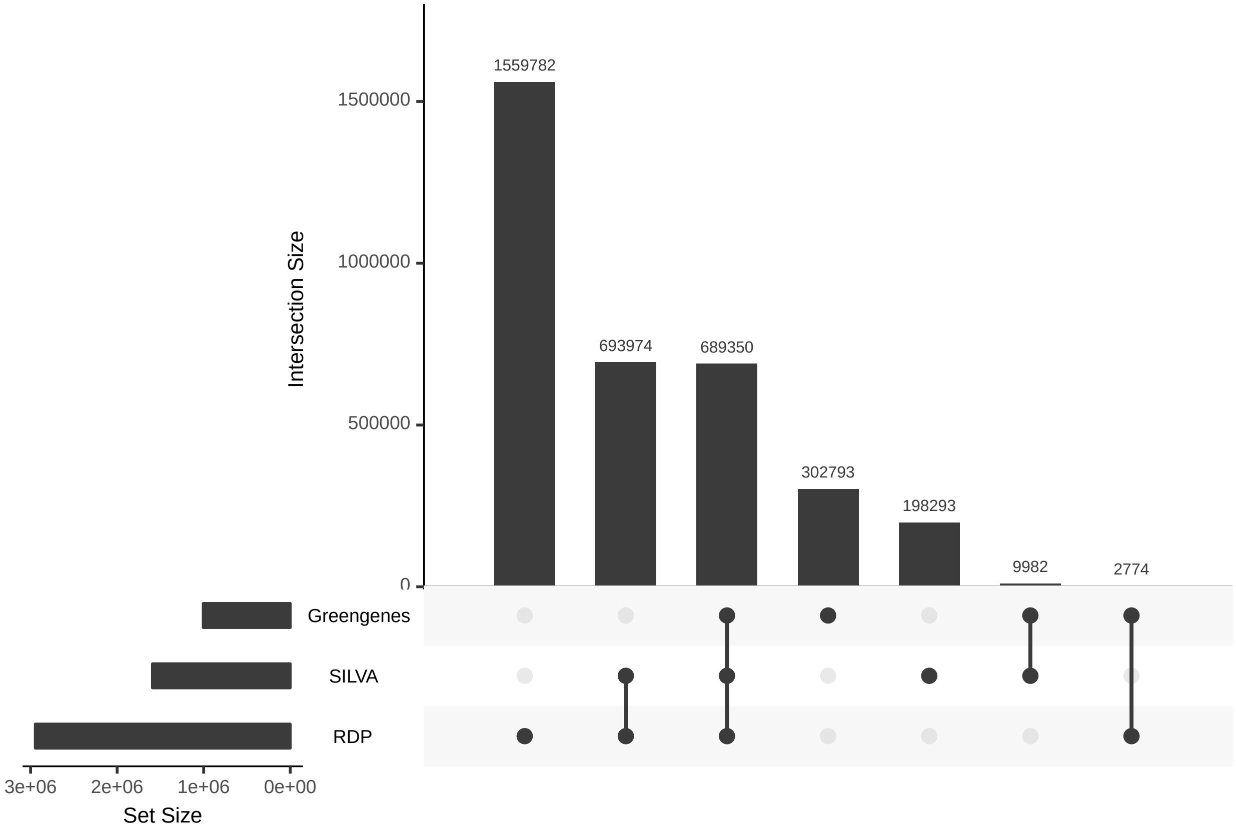

Figure 1: Overlap in sequence composition between the three 16S rRNA databases with MgDb annotation packages for sequences with RNAcentral IDs.

Table 1: Number of sequences (size) and fraction of sequences with RNAcentral ids (frac ids) for the three 16S rRNA databases with MgDb annotation packages.

Database	Size	Frac Ids
Greengenes	1,262,986	0.91
SILVA	1,922,213	0.98
RDP	3,356,808	0.99

4 *Paenibacillus* Taxonomic Resolution

For the taxonomic resolution analysis, we will: 1. characterize the taxonomic composition of the *Paenibacillus* genus in the three databases, 2. calculate pairwise distances between sequences for 30 species with the most sequences, 3. evaluate pairwise distances within and between species.

4.1 Taxonomic Characterization

First obtain the taxonomic information for *Paenibacillus* genus from the three databases using the `mgDb_select()` function.

```r
paeni_16S_greengenes <- metagenomeFeatures::mgDb_select(gg13.5MgDb,
                          type = c("taxa"),
                          keys = "Paenibacillus",
                          keytype = "Genus")
paeni_16S_silva <- metagenomeFeatures::mgDb_select(slv128.1MgDb,
                          type = c("taxa"),
                          keys = "Paenibacillus",
                          keytype = "Genus")
paeni_16S_rdp <- metagenomeFeatures::mgDb_select(rdp11.5MgDb,
                          type = c("taxa"),
                          keys = "Paenibacillus",
                          keytype = "Genus")
```

For *Paenibacillus* genus the databases vary in the number of sequences for each species with some databases not having any sequences for some species (Fig. 2). With a relatively small overlap between the databases, multiple database analysis provides additional power (Fig. 1). The three database maintainers differ in their curation approaches, taxonomic name formatting, and update frequency. There are benefits to using all three databases in a cross database analysis. The Greengenes database taxonomic name is more rigidly formatted, only including the major taxonomic levels, e.g. does not include intermediate level such as suborder, and only goes to the species level. As a result, species name parsing is unlikely to negatively impact the results. SILVA and RDP are consistently updated while Greengenes has not been updated since 2013, and therefore it does not contain sequences that were not available in 2013. As a result, the SILVA and RDP databases are significantly more comprehensive than Greengenes. Both, the SILVA and RDP taxonomic name formatting include intermediate taxonomic levels allowing for more fine grained taxonomic analysis. Additionally, the SILVA database is the only one of the three to include eukaryotic 16S sequences as well as prokaryotic sequences. For our analysis of the *Paenibacillus*, the and issue with all databases is that the sequences only classified to the genus level ("Unassigned" at species level) is the most abundant group, Greeengenes - 2308, SILVA - 3326, and RDP - 7025 (Fig. 2).

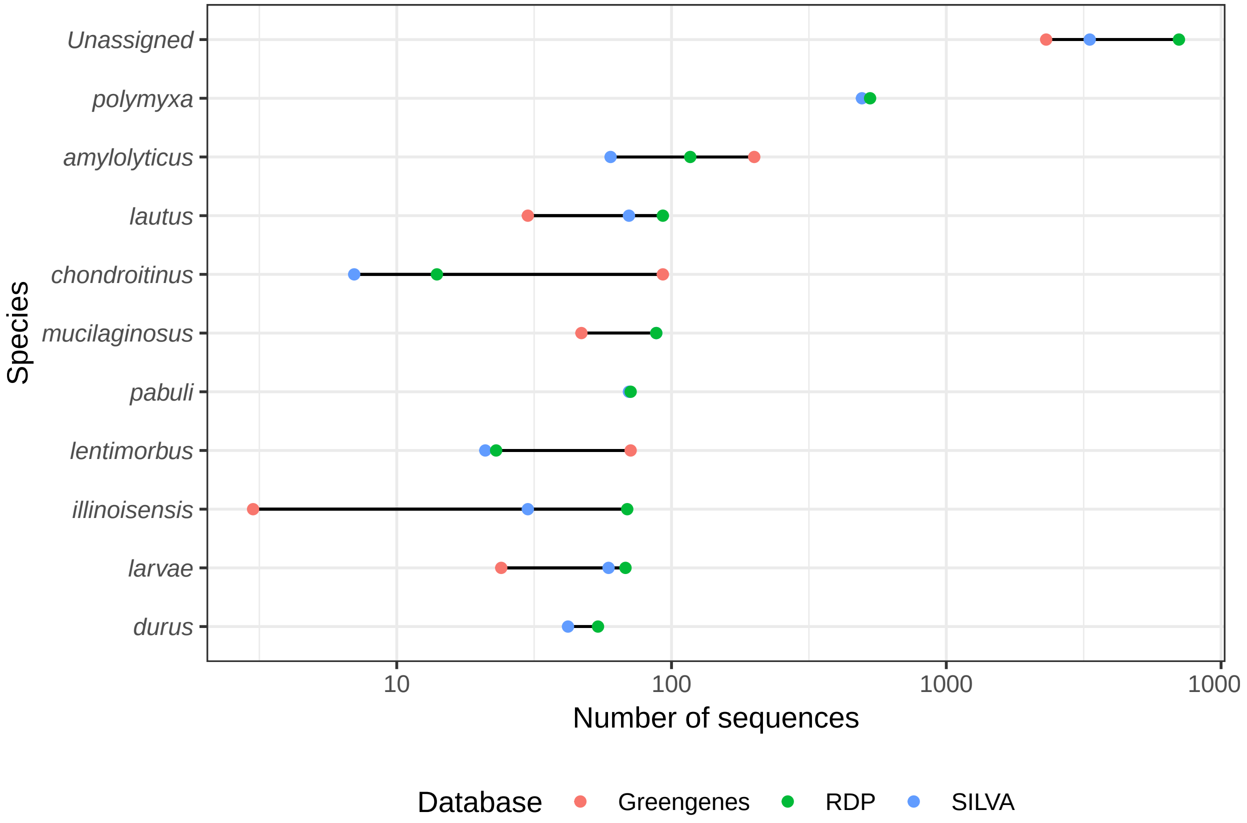

Figure 2: Number of sequences in the genus *Paenibacillus* assigned to the 10 most represented species as well as sequences not assigned at the species level for the Greengenes, SILVA, and RDP databases.

4.2 Taxonomic Resolution

Next, we evaluate the 16S rRNA amplicon sequencing taxonomic resolution for *Paenibacillus* Species by comparing within and between Species amplicon pairwise distance for the V12 and V4 regions. To differentiate between Species, the pairwise distances for within-Species amplicon sequences must be less than the between Species distances. Additionally, the difference in amplicon sequence pairwise distances between and within Species must be greater than the sequencing error rate to detect the difference.

4.2.1 Generating Distance Matrix for 16S Amplicons

Here, we will demonstrate the process for extracting the target amplicon region and calculating pairwise distances using the Greengenes 13.5 database for simplicity. See the source Rmarkdown file for code used to calculate pairwise distances for other databases and the V4 region as well as tidying the data for downstream analysis.

The V12 and V4 regions were extracted from the database sequences using pattern matching. Some sequences do not contain both forward and reverse primers as not all the 16S rRNA sequences are full length. Only sequences with both forward and reverse primers are included in the analysis.

The following PCR primers were used for our *in-silico* PCR:

Region	Direction	Primer
V12	Forward	27F - AGAGTTTGATCATGGCTCAG
	Reverse	336R - CACTGCTGCSYCCCGTAGGAGTCT
V4	Forward	515F - GTGCCAGCMGCCGCGGTAA
	Reverse	806R - GGACTACHVGGGTWTCTAAT

```r
forward_primer <- "AGAGTTTGATCATGGCTCAG"
## reverse complementing reverse primer
reverse_primer <- DNAString("CACTGCTGCSYCCCGTAGGAGTCT") %>%
    reverseComplement() %>%
    as.character()

## Selecting Paenibacillus taxa and seq
paeni_16S_greengenes <- metagenomeFeatures::mgDb_select(gg13.5MgDb,
                           type = c("taxa", "seq"),
                           keys = "Paenibacillus",
                           keytype = "Genus")

## Finding sequences with forward primer
forward_match <- Biostrings::vmatchPattern(forward_primer,
                                           subject = paeni_16S_greengenes$seq,
                                           max.mismatch = 2) %>%
    as.list() %>% map_dfr(as.data.frame,.id = "seq_id")

## Finding sequences with reverse primer
reverse_match <- Biostrings::vmatchPattern(reverse_primer,
                                           subject = paeni_16S_greengenes$seq,
                                           max.mismatch = 2,
                                           fixed = FALSE) %>%
    as.list() %>% map_dfr(as.data.frame,.id = "seq_id")

## sequences with both forward and reverse primers
```

```
seqs_to_use_ids <- intersect(forward_match$seq_id, reverse_match$seq_id)
seqs_to_use <- names(paeni_16S_greengenes$seq) %in% seqs_to_use_ids

## Trimming sequences with both primers
paeni_V12 <- TrimDNA(paeni_16S_greengenes$seq[seqs_to_use],
                     leftPatterns = forward_primer,
                     rightPatterns = reverse_primer,
                     type = "both")
```

```
## Finding left pattern: 100% internal, 0% flanking
##
## Finding right pattern: 100% internal, 0% flanking
##
## Time difference of 0.05 secs
```

```
## Excluding seqs with length 0
paeni_V12_seqs <- paeni_V12[[2]][width(paeni_V12[[2]]) != 0]
```

For more accurate distance estimates, a multiple sequence alignment is used to calculate pairwise distances. We will use the AlignSeqs function in the DECIPHER package to generate the multiple sequence alignment.

```
v12_align <- AlignSeqs(paeni_V12[[2]], verbose = FALSE)
```

The resulting alignment can be viewed using the BrowseSeqs function in the DECIPHER package.

```
BrowseSeqs(v12_align)
```

Next a pairwise distance matrix is generated using the DistanceMatrix function in the DECIPHER package for taxonomic resolution analysis and converting distance matrix to a data frame for analysis.

```
v12_dist <- DistanceMatrix(v12_align,
                           correction = "none",
                           verbose = FALSE,
                           includeTerminalGaps = FALSE)
```

4.2.2 Multiple Database Comparison

Now that we have our pairwise distance matrix, we can evaluate the taxonomic resolution of the genus *Paenibacillus* by comparing the within and between species distances. To reduce the complexity of our analysis, we are only going to look at the pairwise distances between sequences assigned to the 10 species with the most sequences (Fig. 2). With more sequences we will have better estimates for within species pairwise distances but is likely to result in the exclusion of closely related species. We first compare the distribution of within and between species pairwise distances for the genus. Then look at the pairwise distances within species *P. amylolyticus* compared distances between *P. amylolyticus* and the other nine most represented *Paenibacillus* species.

4.2.2.1 Genus Level Comparison

Pairwise distance is smaller within than between species, indicating that the V12 and V4 regions are potentially suitable for classifying members of the *Paenibacillus* genus to the Species level (Fig. 3). The median and interquartile pairwise distance range for within and between species comparisons in consistent across the three databases further supporting our observation. A large number of outliers, pairwise distances past the boxplot whiskers, were observed for all three databases. Outliers below the between species boxplots and above the within species are problematic for species level classification. The sequences and taxonomic assignment of

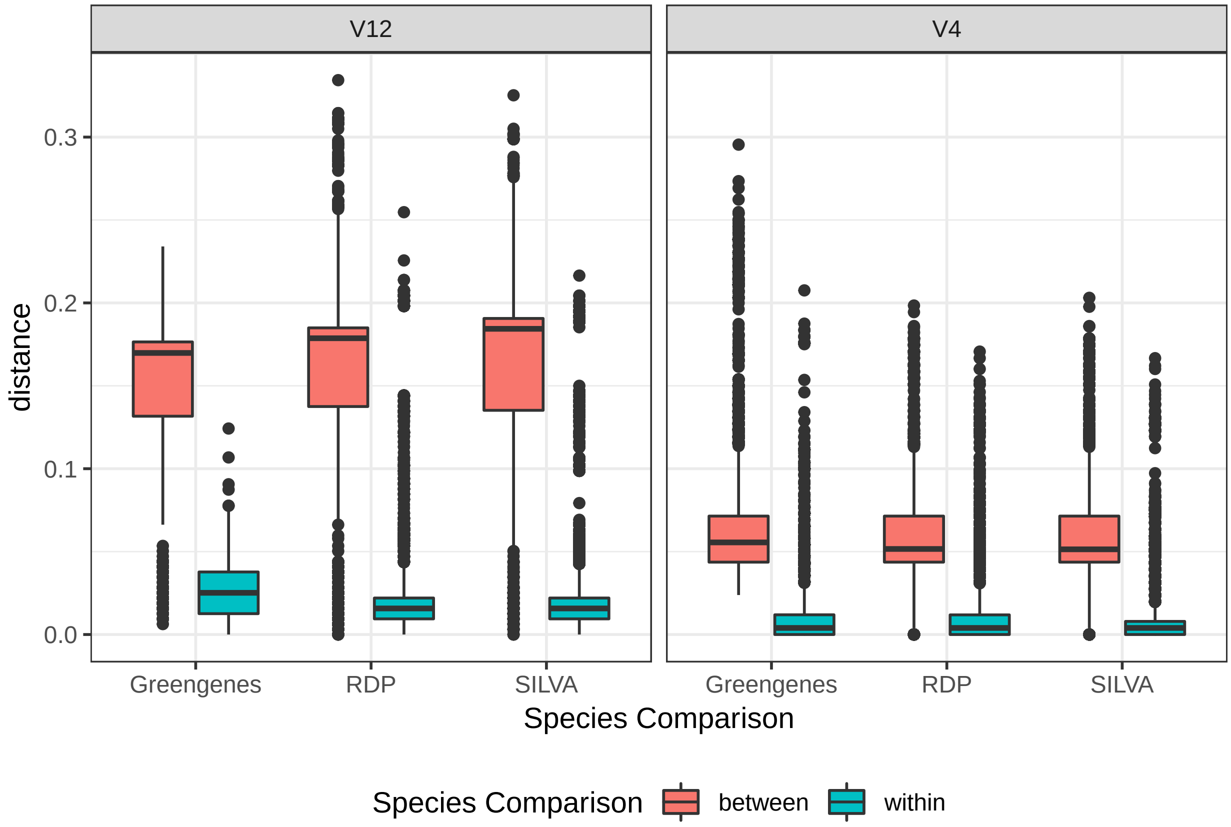

Figure 3: Distribution of within and between Species pairwise distances for the V12 and V4 16S rRNA region. Sequences not classified to the Species level were excluded from the analysis.

these outliers should be evaluated to ensure they are not errors in the database, a result of incorrect database parsing, or are examples where phylogeny and taxonomy are inconsistent.

Overall, the V12 region had greater pairwise distances than V4 for both within and between Species. It is important also to consider that we only included sequences with species assignments and the species with the most sequences in the database. Species-level information for the unclassified sequences and inclusion of less well-represented species might yield results that are inconsistent with our analysis. Additionally, our analysis does not identify the pairwise sequence distance required to classify a sequence as a novel *Paenibacillus* Species.

4.2.2.2 Species level comparison

While the overall pairwise distance is greater between than within species for the *Paenibacillus* genus, it is important to understand how within and between species pairwise distances compare for individual Species. The within species pairwise for *P. amylolyticus* distances are similar to between species distances for *P. pabuli* and *P. illnoisensis* (Fig. 4). When including outliers the within pairwise distances for *P. amylolyticus* is in the range of between pairwise distances for other *P. spp.*. Similar to the genus level comparison investigation of the outliers sequences and strains is required to evaluate the taxonomic resolution for *Paenibacillus*.

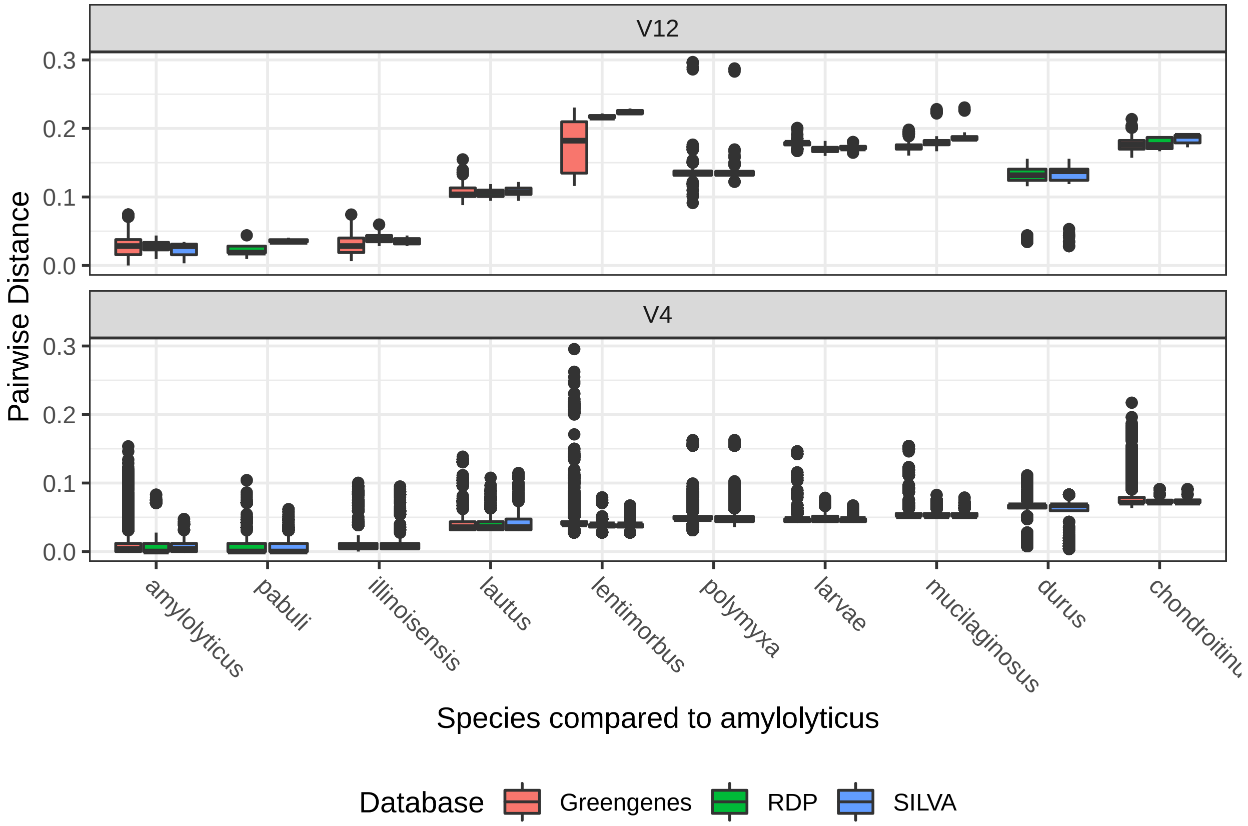

Figure 4: Pairwise distances within the *Paenibacillus amylolyticus* species and between *P. amylolyticus* and the 9 other most represented species.

5 Conclusion

The metagenomeFeatures package and associated 16S rRNA database packages greengenes13.5MgDb, silva128.1MgDb, ribosomaldatabaseproject11.5MgDb, provide a consistent interface for working with the different databases. Furthermore, through the use of the RNAcentralIDs we are able to evaluate database overlap and perform cross database analysis. We demonstrate how the *metagenomeFeatures* package in conjunction with one of the associated 16S rRNA database packages, and other R packages, can be used to evaluate whether species-level taxonomic classification is possible for a specific amplicon region. The approach used here can easily be extended to evaluate taxonomic groups (changing filtering parameters), or amplicon regions (changing primer sequences).

---

6 Session Information

6.1 System Information

```
sessioninfo::platform_info()
```

```
##  setting  value
##  version  R version 3.5.2 (2018-12-20)
##  os       macOS High Sierra 10.13.6
##  system   x86_64, darwin17.7.0
##  ui       unknown
##  language (EN)
##  collate  en_US.UTF-8
##  ctype    en_US.UTF-8
##  tz       America/New_York
##  date     2019-02-02
```

6.2 Package Versions

```
sessioninfo::package_info() %>%
  filter(attached == TRUE) %>%
  select(package, loadedversion, source) %>%
    knitr::kable(booktabs = TRUE)
```

package	loadedversion	source
bindrcpp	0.2.2	CRAN (R 3.5.0)
Biobase	2.40.0	Bioconductor
BiocGenerics	0.26.0	Bioconductor
Biostrings	2.48.0	Bioconductor
DECIPHER	2.8.1	Bioconductor
dplyr	0.7.8	CRAN (R 3.5.1)
forcats	0.3.0	CRAN (R 3.5.0)
ggplot2	3.1.0	CRAN (R 3.5.2)
ggpubr	0.2	CRAN (R 3.5.2)
greengenes13.5MgDb	2.0.0	Bioconductor
IRanges	2.14.12	Bioconductor
magrittr	1.5	CRAN (R 3.5.0)
metagenomeFeatures	2.3.3	Bioconductor
purrr	0.3.0	CRAN (R 3.5.2)
readr	1.3.1	CRAN (R 3.5.2)
ribosomaldatabaseproject11.5MgDb	1.00.1	Bioconductor
RSQLite	2.1.1	CRAN (R 3.5.0)
S4Vectors	0.18.3	Bioconductor
silva128.1MgDb	1.00.0	Bioconductor
stringr	1.3.1	CRAN (R 3.5.0)
tibble	2.0.1	CRAN (R 3.5.2)
tidyr	0.8.2	CRAN (R 3.5.1)
tidyverse	1.2.1	CRAN (R 3.5.0)
UpSetR	1.3.3	CRAN (R 3.5.1)
XVector	0.20.0	Bioconductor

7 References

Grady, Elliot Nicholas, Jacqueline MacDonald, Linda Liu, Alex Richman, and Ze-Chun Yuan. 2016. “Current Knowledge and Perspectives of Paenibacillus: A Review.” *Microbial Cell Factories* 15 (1): 203.

Kassambara, Alboukadel. 2017. *Ggpubr: ’Ggplot2’ Based Publication Ready Plots*. https://CRAN.R-project.org/package=ggpubr.

Ouyang, Jie, Zhiheng Pei, Larry Lutwick, Sharvari Dalal, Liying Yang, Nicholas Cassai, Kuldip Sandhu, et al. 2008. “Paenibacillus Thiaminolyticus: A New Cause of Human Infection, Inducing Bacteremia in a Patient on Hemodialysis.” *Annals of Clinical & Laboratory Science* 38 (4): 393–400.

Wickham, Hadley. 2017. *Tidyverse: Easily Install and Load the ’Tidyverse’*. https://CRAN.R-project.org/package=tidyverse.

Wright, Erik S. 2015. “DECIPHER: Harnessing Local Sequence Context to Improve Protein Multiple Sequence Alignment.” *BMC Bioinformatics* 16 (October): 322.