Databases and ontologies

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

Nathan D. Olson[1,2,3*], Nidhi Shah[2,3,4], Jayaram Kancherla[2,3], Justin Wagner[2,3,4], Joseph N. Paulson[5], and Hector Corrada-Bravo[2,3,4]

[1]Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA, [2]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA, [3]University of Maryland Institute for Advanced Computer Studies, College Park, MD 20742, USA, [4]Department of Computer Science, University of Maryland, College Park, MD 20742, USA, [5] Department of Biostatistics, Product Development, Genentech Inc., South San Francisco, CA 94080, USA

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

We developed the *metagenomeFeatures* R Bioconductor package along with annotation packages for the three primary 16S rRNA databases (Greengenes, RDP, and SILVA) to facilitate working with 16S rRNA sequence databases and marker-gene survey feature data. The *metagenomeFeatures* package defines two classes, `MgDb` for working with 16S rRNA sequence databases, and `mgFeatures` for working with marker-gene survey feature data. The associated annotation packages provide a consistent interface to the different 16S rRNA databases facilitating database comparison and exploration. The `mgFeatures` represents a crucial step in the development of a common data structure for working with 16S marker-gene survey data in R.

Availability: https://bioconductor.org/packages/release/bioc/html/metagenomeFeatures.html

Contact: nolson@nist.gov

1 Introduction

16S rRNA marker-gene surveys have significantly advanced our understanding of the diversity and structure of prokaryotic communities present in ecosystems including the human gut, open ocean, and even the international space station (Lang *et al*., 2017; Thompson *et al*., 2017; Human Microbiome Project Consortium 2012). For a 16S rRNA marker-gene survey, the 16S rRNA gene is sequenced using a targeted assay. The raw sequence data is processed using a bioinformatic pipeline where the sequences are grouped into features, e.g., operational taxonomic units (OTUs) or sequence variants (SVs), yielding a set of representative sequences (Callahan, McMurdie, and Holmes 2017; Beiko 2015).

A critical step in 16S rRNA marker-gene surveys is comparing representative sequences to a reference database for taxonomic classification or phylogenetic placement (Nguyen *et al*., 2016). There are numerous 16S rRNA reference databases of which Greengenes, RDP, and SILVA are arguably the most commonly used (DeSantis et al. 2006; Cole et al. 2014; Quast et al. 2012; McDonald et al. 2012). Additionally, there are smaller system-specific databases such as HOMD for the human oral microbiome (Chen et al. 2010)(http://www.homd.org/) and soil reference database (Choi et al. 2017). System-specific databases can improve taxonomic assignments for microbial communities not well represented in the major databases (Rohwer et al. 2017).

16S rRNA databases differ in the number and diversity of sequences, the taxonomic classification system, and the inclusion of intermediate ranks (Balvočiūtė and Huson 2017, Table 1). Databases format their data differently and use sequence identification systems unique to their database, challenging membership and composition comparisons. For example, Yang et al. (2016) used the SILVA database to evaluate how different 16S rRNA variable regions impact phylogenetic analysis. Similarly, Martinez-Porchas et al. (2017) also used the SILVA database to evaluate sequence similarity between 16S rRNA gene conserved regions. Differences in database formatting present a significant barrier to performing the same analysis using multiple databases. Additionally, taxonomic assignments can be database-dependent, providing further justification for database comparisons (Pettengill and Rand 2017). To facilitate database comparisons RNACentral (http://rnacentral.org/) a resource combining non-coding

RNA databases, provides unique identifiers for the sequences (The RNAcentral Consortium 2017).

Table 1. 16S rRNA gene sequence databases with Bioconductor annotation packages we developed.

Database	Version	Sequences	Taxonomic System[1]
Greengenes	13.5[2]	1,262,986	NCBI
RDP	11.5	3,356,809	Bergey's
SILVA	128.1	1,922,213	Bergey's

[1] Primary taxonomic system see (see Balvočiūtė and Huson 2017) and references therein for additional information on databases.
[2]Greengenes 13.8 85% OTUs is included in the *metagenomeFeatures* package as an example `MgDb` formatted database.

The statistical programming language, R provides a rich environment and software for data analysis (R Core Team, n.d.). Additionally, Bioconductor, the R bioinformatic software resource (Huber et al. 2015) includes a number of packages for working with DNA sequence data and 16S rRNA marker-gene survey data such as *phyloseq* (McMurdie and Holmes 2013) and *metagenomeSeq* (Paulson et al. 2013). While a number of software tools are available for working with 16S rRNA marker-gene survey feature data including Mothur (Schloss et al. 2009) as well as QIIME, specifically the q2-feature-classifier plugin (Bokulich et al. 2018). There are no existing tools for working with multiple 16S rRNA databases. Furthermore, tools for working with 16S rRNA marker-gene survey feature data in R all use different data structures. Therefore, an R package defining consistent data structures for working with 16S rRNA database and marker-gene survey feature data is needed.

To address this need we developed the R package *metagenomeFeatures* for working with both 16S rRNA gene database and marker-gene survey feature data. *metagenomeFeatures* provides a common data structure for working with the 16S rRNA databases and marker-gene survey feature data. Additionally, this package is the first step towards the development of a common data structure for use in analyzing metagenomic and marker-gene survey data using R packages such as *phyloseq* (McMurdie and Holmes 2013) and *metagenomeSeq* (Paulson et al. 2013).

2 *metagenomeFeatures* Package

The *metagenomeFeatures* package defines two data structures, `MgDb` for working with 16S rRNA databases, and `mgFeatures` for working with marker-gene survey feature data. There are three types of relevant information for both `MgDb` and `mgFeatures` class objects, (1) the sequences themselves, (2) sequence taxonomic lineage, and (3) a phylogenetic tree representing the evolutionary relationship between features. `MgDb` and `mgFeatures` data structures are both S4 object-oriented classes with slots for taxonomy, sequences, phylogenetic tree, and metadata.

The `MgDb-class` provides a consistent data structure for working with different 16S rRNA databases. As shown in Table 1 ,16S rRNA databases contain hundreds of thousands to millions of sequences, therefore an SQLite database is used to store the taxonomic and sequence data. Using an SQLite database prevents the user from loading the full database into memory. The database connection is managed using the *RSQLite* R package (Müller et al. 2017), and the taxonomic data are accessed using the *dplyr* and *dbplyr* packages (Wickham 2017; Wickham et al. 2017). The *DECIPHER* package is used to format the sequence data as an SQLite database (Wright 2016) and provides functions for working directly with the sequence data in the SQLite database. The `phylo` class, from the *APE* R package, defines the tree slot (Paradis, Claude, and Strimmer 2004). We developed Bioconductor annotation packages for commonly used databases, Greengenes, RDP, and SILVA (Table 1, Cole et al. 2014; Quast et al. 2012; DeSantis et al. 2006). Along with database specific sequence identifiers, RNAcentral indentifers are included in the SQLite table for inter-database comparisons.

`mgFeatures-class` is used for storing and working with marker-gene survey feature data. Similar to the `MgDb-class`, the `mgFeatures-class` has four slots, for taxonomy, sequences, phylogenetic tree, and metadata. As the number of features in a marker-gene survey dataset is significantly smaller than the number of sequences in a reference database, `mgFeatures` uses common Bioconductor data structures, `DataFrames` and `DNAStringSets` to define the taxonomic and sequence slots (Pagès et al. 2008; Pagès, Lawrence, and Aboyoun 2017). Similar to `MgDb-class`, a `phylo` class object is used to define the tree slot. For both `MgDb` and `mgFeatures` classes the tree slot is optional and the metadata are stored as a list.

The *metagenomeFeatures* package includes vignettes as example use cases for the *metagenomeFeatures* package and associated reference database annotation packages.

- Retrieving sequence and phylogenetic data for OTUs from closed-reference clustering.
- Exploring diversity for a taxonomic group of interest.

The R command `browseVignettes("metagenomeFeatures")` provides a list of vignettes associated with the package and `vignette("x")` is used to view specific vignettes, where "x" is the vignette name.

To further demonstrate the utility of the package, the manuscript supplemental information demonstrates using *metagenomeFeatures*, *greengenes13.5MgDb* annotation package, and *DECIPHER* to evaluate the potential for species-level taxonomic classification using 16S rRNA V12 and V4 sequence data.

3 Conclusions

The *metagenomeFeatures* package provides data structures and functions for working with 16S rRNA gene sequence reference databases and marker-gene survey feature data. The data structure provided by the `MgDb-class` in conjunction with the shared sequence identifier system developed by RNACentral facilitates comparisons between 16S rRNA databases. The `mgFeatures-class` provides the groundwork for the development of a common data structure for working with metagenomic and marker-gene sequence data in R which will increase interoperability between R packages developed for working with metagenomic sequence data. Additionally, while the data structures were developed for 16S rRNA gene sequence data they can be used for any marker-gene sequence data without modification and can be extended to work with shotgun metagenomic sequence data and databases.

4 Acknowledgements

The authors would like to thank Drs. Mihai Pop, Marc Salit, Samuel Forry, and Arlin Stolzfus for feedback on the manuscript. The Bioconductor core team provided valuable feedback during the package submission and update process. Opinions expressed in this paper are the authors' and do not necessarily reflect the policies and views of NIST or affiliated venues. Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendations or endorsement by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose. Official contribution of NIST; not subject to copyrights in USA.

5 Funding

This work was partially supported by National Institutes of Health (NIH) [NIH RO1GM114267 to J.W., J.K., H.C.B. and NIH R01HG005220 to H.C.B.]

Conflict of Interest: none declared.

6 References

Balvočiūtė, Monika, and Daniel H. Huson. 2017. "SILVA, RDP, Greengenes, NCBI and OTT - How Do These Taxonomies Compare?" *BMC Genomics* 18 (Suppl 2): 114.

Beiko, Robert G. 2015. "Microbial Malaise: How Can We Classify the Microbiome?" *Trends in Microbiology* 23 (11): 671–79.

Bokulich, Nicholas A., Benjamin D. Kaehler, Jai Ram Rideout, Matthew Dillon, Evan Bolyen, Rob Knight, Gavin A. Huttley, and J. Gregory Caporaso. 2018. "Optimizing Taxonomic Classification of Marker Gene Amplicon Sequences." e3208v2. PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.3208v2.

Callahan, Benjamin J., Paul J. McMurdie, and Susan P. Holmes. 2017. "Exact Sequence Variants Should Replace Operational Taxonomic Units in Marker-Gene Data Analysis." *The ISME Journal* 11 (12): 2639–43.

Chen, Tsute, Wen-Han Yu, Jacques Izard, Oxana V. Baranova, Abirami Lakshmanan, and Floyd E. Dewhirst. 2010. "The Human Oral Microbiome Database: A Web Accessible Resource for Investigating Oral Microbe Taxonomic and Genomic Information." *Database: The Journal of Biological Databases and Curation* 2010 (July): baq013.

Choi, Jinlyung, Fan Yang, Ramunas Stepanauskas, Erick Cardenas, Aaron Garoutte, Ryan Williams, Jared Flater, et al. 2017. "Strategies to Improve Reference Databases for Soil Microbiomes." *The ISME Journal* 11 (4): 829–34.

Cole, James R., Qiong Wang, Jordan A. Fish, Benli Chai, Donna M. McGarrell, Yanni Sun, C. Titus Brown, Andrea Porras-Alfaro, Cheryl R. Kuske, and James M. Tiedje. 2014. "Ribosomal Database Project: Data and Tools for High Throughput rRNA Analysis." *Nucleic Acids Research* 42 (Database issue): D633–42.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. 2006. "Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB." *Applied and Environmental Microbiology* 72 (7): 5069–72.

Huber, Wolfgang, Vincent J. Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S. Carvalho, Hector Corrada Bravo, et al. 2015. "Orchestrating High-Throughput Genomic Analysis with Bioconductor." *Nature Methods* 12 (2). Nature Research: 115–21.

Human Microbiome Project Consortium. 2012. "Structure, Function and Diversity of the Healthy Human Microbiome." *Nature* 486 (7402): 207–14.

Lang, Jenna M., David A. Coil, Russell Y. Neches, Wendy E. Brown, Darlene Cavalier, Mark Severance, Jarrad T. Hampton-Marcell, Jack A. Gilbert, and Jonathan A. Eisen. 2017. "A Microbial Survey of the International Space Station (ISS)." *PeerJ* 5 (December). PeerJ Inc.: e4029.

Martinez-Porchas, Marcel, Enrique Villalpando-Canchola, Luis Enrique Ortiz Suarez, and Francisco Vargas-Albores. 2017. "How Conserved Are the Conserved 16S-rRNA Regions?" *PeerJ* 5 (February): e3036.

McDonald, Daniel, Morgan N. Price, Julia Goodrich, Eric P. Nawrocki, Todd Z. DeSantis, Alexander Probst, Gary L. Andersen, Rob Knight, and Philip Hugenholtz. 2012. "An Improved Greengenes Taxonomy with Explicit Ranks for Ecological and Evolutionary Analyses of Bacteria and Archaea." *The ISME Journal* 6 (3): 610–18.

McMurdie, Paul J., and Susan Holmes. 2013. "Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data." *PloS One* 8 (4): e61217.

Müller, Kirill, Hadley Wickham, David A. James, and Seth Falcon. 2017. "'SQLite' Interface for R [R Package RSQLite Version 2.0]." Comprehensive R Archive Network (CRAN). https://cran.rstudio.com/web/packages/RSQLite/index.html.

Nguyen, Nam-Phuong, Tandy Warnow, Mihai Pop, and Bryan White. 2016. "A Perspective on 16S rRNA Operational Taxonomic Unit Clustering Using Sequence Similarity." *Npj Biofilms and Microbiomes* 2 (April). Nature Publishing Group: 16004.

Pagès, H., R. Gentleman, P. Aboyoun, and S. DebRoy. 2008. "Biostrings: String Objects Representing Biological Sequences, and Matching Algorithms." *R Package Version* 2 (0): 160.

Pagès, H., M. Lawrence, and P. Aboyoun. 2017. "S4Vectors: S4 Implementation of Vectors and Lists." *R Package Version 0. 13* 15.

Paradis, Emmanuel, Julien Claude, and Korbinian Strimmer. 2004. "APE: Analyses of Phylogenetics and Evolution in R Language." *Bioinformatics* 20 (2): 289–90.

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): 1200–1202.

Pettengill, James B., and Hugh Rand. 2017. "Segal's Law, 16S rRNA Gene Sequencing, and the Perils of Foodborne Pathogen Detection within the American Gut Project." *PeerJ* 5 (June): e3480.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2012. "The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools." *Nucleic Acids Research*, November. https://doi.org/10.1093/nar/gks1219.

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

Rohwer, Robin Rebecca, Joshua J. Hamilton, Ryan J. Newton, and Katherine D. McMahon. 2017. "TaxAss: Leveraging Custom Databases Achieves Fine-Scale Taxonomic Resolution." *bioRxiv*, January. http://biorxiv.org/content/early/2017/11/05/214288.abstract.

Schloss, Patrick D., Sarah L. Westcott, Thomas Ryabin, Justine R. Hall, Martin Hartmann, Emily B. Hollister, Ryan A. Lesniewski, et al. 2009. "Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities." *Applied and Environmental Microbiology* 75 (23): 7537–41.

The RNAcentral Consortium. 2017. "RNAcentral: A Comprehensive Database of Non-Coding RNA Sequences." *Nucleic Acids Research* 45 (D1): D128–34.

Thompson, Luke R., Jon G. Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J. Locey, Robert J. Prill, et al. 2017. "A Communal Catalogue Reveals Earth's Multiscale Microbial Diversity." *Nature* 551 (7681): 457–63.

Wickham, Hadley. 2017. "A 'Dplyr' Back End for Databases [R Package Dbplyr Version 1.1.0]." Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=dbplyr.

Wickham, Hadley, Romain Francois, Lionel Henry, and Kirill Müller. 2017. "A Grammar of Data Manipulation [R Package Dplyr Version 0.7.4]." Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/dplyr/index.html.

Wright, Erik S. 2016. "Using DECIPHER v2. 0 to Analyze Big Biological Sequence Data in R." *The R Journal* 8 (1). https://journal.r-project.org/archive/2016-1/wright.pdf.

Yang, Bo, Yong Wang, and Pei-Yuan Qian. 2016. "Sensitivity and Correlation of Hypervariable Regions in 16S rRNA Genes in Phylogenetic Analysis." *BMC Bioinformatics* 17 (March): 135.