HMP16SData: Efficient Access to the Human Microbiome Project through Bioconductor

Lucas Schiffer[1,2], Rimsha Azhar[1,2], Lori Shepherd[3], Marcel Ramos[1,2,3], Ludwig Geistlinger[1,2], Curtis Huttenhower[4,5], Jennifer B Dowd[1,6], Nicola Segata[7], Levi Waldron[1,2]

1. Graduate School of Public Health and Health Policy, City University of New York, New York, NY
2. Institute for Implementation Science in Population Health, City University of New York, New York, NY
3. Roswell Park Cancer Institute, University of Buffalo, Buffalo, NY
4. Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA
5. The Broad Institute of MIT and Harvard, Cambridge, MA
6. Department of Global Health and Social Medicine, King's College London, London, UK
7. Centre for Integrative Biology, University of Trento, Trento, Italy

Introduction

The NIH Human Microbiome Project (HMP) was one of the first large-scale population studies of microbiome variation outside of disease, including healthy American adults aged 18 to 40 and producing a comprehensive reference for the composition and variation of the “healthy” human microbiome.[1,2] Raw and processed 16S rRNA (16S) and metagenomic shotgun (MGX) sequencing data can be freely downloaded from the HMP data analysis and coordinating center (DACC) (https://www.hmpdacc.org/hmp/) or analyzed online using the HMP data portal (https://portal.hmpdacc.org/).

However, accessing and analyzing the data with statistical software still involves substantial bioinformatic and data management challenges. These include data import and merging of microbiome profiles with public and controlled-access participant data, integration with phylogenetic trees, potentially mapping microbial and participant identifiers for comparison between 16S and MGX data sets, and accessing controlled participant data. Specifically, access to most participant data is controlled, requiring authorization through the National Center for Biotechnology Information (NCBI) Database of Genotypes and Phenotypes (dbGaP), and requires the use of specialized software for download and decryption.

We thus developed the *HMP16SData* R/Bioconductor package to simplify access to and analysis of HMP 16S data. The design of the package follows our *curatedMetagenomicData* R/Bioconductor package, enabling comparative analysis with MGX samples from the HMP and dozens of other studies.[3] *HMP16SData* leverages Bioconductor’s *ExperimentHub* and the

SummarizedExperiment data class to distribute merged taxonomic and public-access participant data. It provides 16S gene sequencing data for variable regions 1–3 (V13) and 3–5 (V35), with merged participant data, and a function for automated merging of controlled-access data to researchers with a project approved by dbGaP. Methods for subsetting, coercion to the *phyloseq* class for ecological and differential abundance analysis, and comparative analyses and are also provided.[4] Finally, *HMP16SData* greatly simplifies access to and merging of restricted participant data. These simplifications enable epidemiologists with only basic R skills and limited knowledge of HMP DACC or dbGaP procedures to quickly make use of HMP data.

Methods

HMP16SData provides data from the HMP 16S compendium as processed through the HMP DACC QIIME pipeline (https://www.hmpdacc.org/HMQCP/).[5] All publicly-available participant data, as obtained from the HMP DACC, is also included and a function provides simplified access to and merging of controlled data from dbGaP for registered researchers. Use of *HMP16SData* begins with one of two functions: `V13` (to download data for 16S V13) or `V35` (to download data for 16S V35), each of which returns a *SummarizedExperiment* object. Each object contains (and can be accessed by): sequencing count data (`assay`) Selection of samples by body site, visit number, and taxonomic hierarchy is straightforward through standard *SummarizedExperiment* or *phyloseq* subsetting methods. Researchers with an approved dbGaP project can optionally use a second function, `attach_dbGaP`, to attach controlled participant data, prior to coercion to a *phyloseq* object for ecological analyses such as alpha and beta diversity. **Example 1** demonstrates the selection of only stool samples, attachment of controlled participant data, and coercion to a *phyloseq* object.

```
library(HMP16SData)

V35_stool <-
    V35() %>%
    subset(select = HMP_BODY_SUBSITE == "Stool")

V35_stool_protected <-
    attach_dbGaP(V35_stool, "~/prj_12146.ngc")

V35_stool_phyloseq <-
    as_phyloseq(V35_stool)
```

Example 1 *– Data access using the HMP16SData API. This example demonstrates subsetting by body subsite, attaching of controlled participant data from dbGaP, and coercion to a phyloseq class object.*

Controlled-Access Participant Data Analysis

Non-restricted participant data include only visit number, sex, run center, body site, and body subsite – an additional 248 participant data variables are available through dbGaP after project registration through dbGaP. After project approval, dbGaP provides researchers with a "repository key" that identifies and decrypts controlled-access participant data. The `attach_dbGaP` function takes the public *SummarizedExperiment* data set and the path to the dbGaP repository key as arguments; it performs download, decryption, and merging of controlled participant data, and returns another *SummarizedExperiment* with controlled-access participant data added to its *colData* slot. Internally, `attach_dbGaP` uses system calls to the NCBI SRA (Sequence Read Archive) Toolkit for download and decryption, and R functionality to load and merge the controlled data. A data dictionary describing the controlled-access participant data variables is incorporated into the package and is accessible by entering `data(dictionary)`.

Phyloseq Class Coercion

The *phyloseq* package is a commonly used tool for ecological analysis of microbiome data in R/Bioconductor. *HMP16SData* provides a function, `as_phyloseq`, to coerce its default *SummarizedExperiment* objects to *phyloseq* objects. The resulting objects contain taxonomic abundance count data, participant data, complete taxonomy, and phylogenetic trees, enabling computation of UniFrac and other ecological distances.[6]

Results

HMP16SData provides a total of 7,641 taxonomic profiles from 16S variable regions 1-3 and 3-5 for 239 participants in the HMP, for 18 body subsites and up to three visits. These profiles are provided as two Bioconductor *SummarizedExperiment*-class objects: `V13` and `V35` (**Table 1**), which integrate OTU count data, taxonomy, a phylogenetic tree, and public-use participant information. Each object includes both 16S and MGX sample identifiers, enabling mapping and comparison to MGX profiles distributed by our *curatedMetagenomicData* R/Bioconductor package.[3]. Such a comparison is illustrated in the phylum-level relative abundance plots of matched 16S and MGX sequencing samples in **Figure 1**. Code to reproduce Table 1 and Figure 1 are provided in the package "vignette" documentation.

	V13		V35	
	N	%	N	%
Sex				
Female	1,521	52.48	2,188	46.13
Male	1,377	47.52	2,555	53.87
HMP Body Subsite				
Tongue Dorsum	190	6.56	316	6.66
Supragingival Plaque	189	6.52	313	6.60
Right Retroauricular Crease	187	6.45	297	6.26
Stool	187	6.45	319	6.73
Left Retroauricular Crease	186	6.42	285	6.01
Palatine Tonsils	186	6.42	312	6.58
Buccal Mucosa	183	6.31	312	6.58
Subgingival Plaque	183	6.31	309	6.51
Attached Keratinized Gingiva	181	6.25	313	6.60
Hard Palate	178	6.14	302	6.37
Throat	170	5.87	307	6.47
Saliva	162	5.59	290	6.11
Anterior Nares	161	5.56	269	5.67
Right Antecubital Fossa	146	5.04	207	4.36
Left Antecubital Fossa	145	5.00	201	4.24
Mid Vagina	89	3.07	133	2.80
Posterior Fornix	88	3.04	133	2.80
Vaginal Introitus	87	3.00	125	2.64

Table 1 *– Select characteristics of 16S rRNA (16S) samples for variable regions 1–3 (V13) and 3–5 (V35) available through HMP16SData. All numbers represent samples rather than subjects, given that there are multiple samples per subject.*

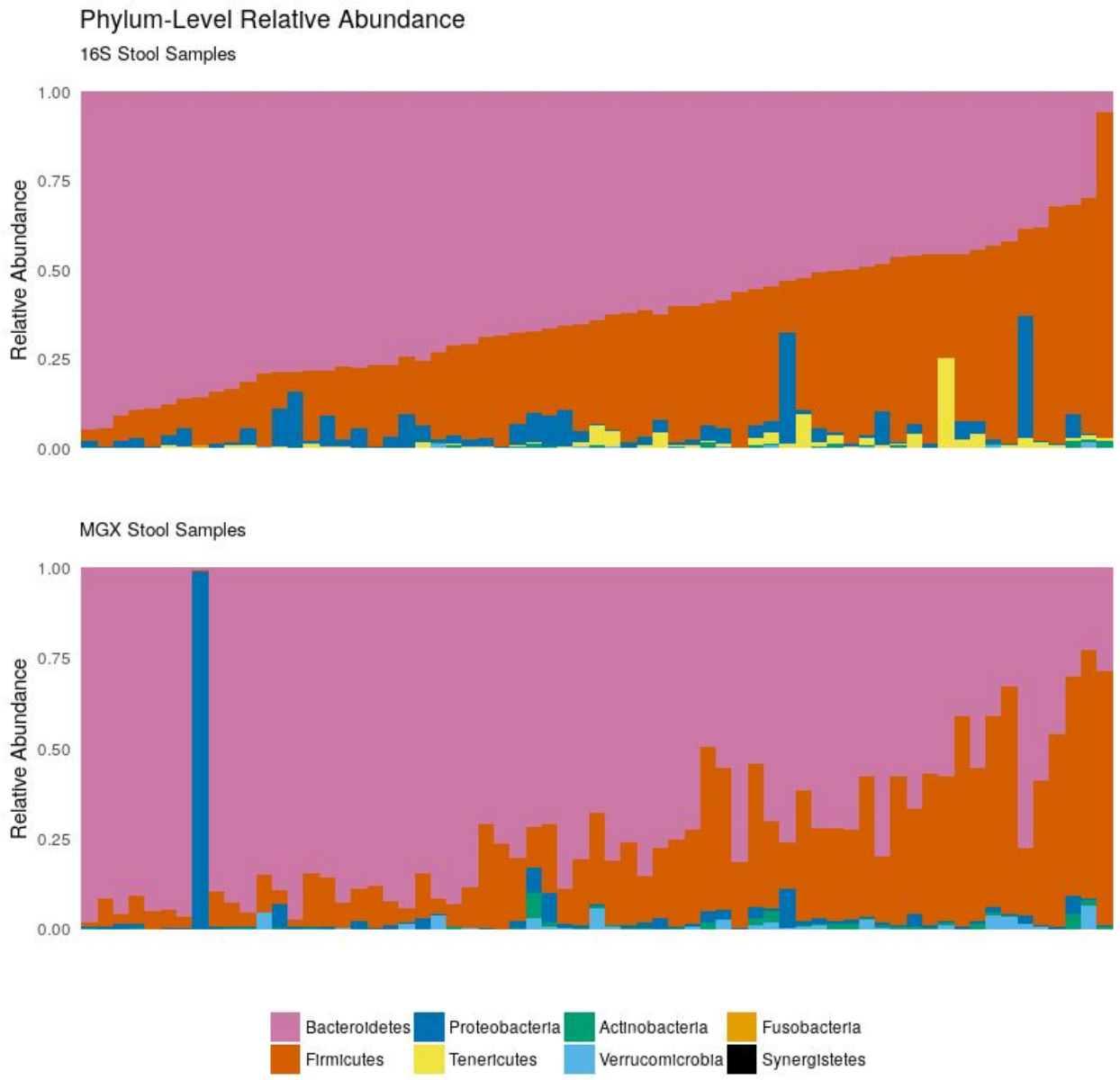

Figure 1 *– Phylum-level relative abundance of the eight most abundant phyla in matched 16S and MGX samples from the HMP. Samples are ordered in both plots by abundance of Bacteroidetes in 16S samples. Notably, the figure illustrates the Bacteroidetes/Firmicutes gradient with reasonable agreement between the 16S and MGX samples.*

Discussion

The HMP provides a comprehensive reference for the composition, diversity, and variation of the human microbiome in the absence of overt disease, making it a potential control or comparison cohort for many microbiome studies. The R/Bioconductor environment provides an extensive range of operations for data analysis, with documented workflows available for typical microbiome investigations.[7] The *HMP16SData* package thus integrates HMP 16S taxonomic abundance profiles, controlled and public participant data, and phylogenetic distances with R/Bioconductor. This greatly reduces the time and bioinformatics expertise required to analyze these data, particularly in the context of additional integrated microbiome population studies. Further, users of other analysis environments can export the resulting data and data products to other formats (SAS, SPSS, STATA, etc.) using the *haven* R package, or export to text files using built-in R/Bioconductor commands.[8] We hope this facilitates broader utilization of the data generated by the HMP among epidemiologists, statisticians, and computational biologists.

Some precautions should be noted when using *HMP16SData* in comparative metagenomic analyses. First, studies of the human microbiome are susceptible to batch effects which should be accounted for in making cross-study comparisons, along with other forms of technical variation.[9,10] Second, the V13 and V35 data sets are obtained from sequencing different variable regions of the 16S rRNA gene, and provide correlated but different estimates of taxonomic relative abundance.[11,12] In the case of the HMP, the samples sequenced in V13 are a subset of those that were sequenced in V35. With these precautions in mind, the HMP is a key reference data set for future microbiome studies. This work enables efficient access to and analysis of the HMP by greatly reducing previous hurdles of data access and management.

Acknowledgements

This research was supported by the National Institute of Allergy and Infectious Diseases (1R21AI121784-01 to J.B.D. and L.W.), the National Institute of Dental and Craniofacial Research (U54DE023798 to C.H.), the National Human Genome Research Institute (R01 HG005220 to Rafael Irizarry), the National Science Foundation (MCB-1453942 and DBI-1053486 to C.H.), and in part, under National Science Foundation Grants CNS-0958379, CNS-0855217, ACI-1126113 to the City University of New York High Performance Computing Center at the College of Staten Island.

The color pallet used in **Figure 1** is optimized for color-blind individuals as proposed by Wong.[13]

References

1. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*. 2012 Jun 13;**486**(7402):207–214.

2. Human Microbiome Project Consortium. A framework for human microbiome research. *Nature*. 2012 Jun 13;**486**(7402):215–221.

3. Pasolli E, Schiffer L, Manghi P, et al. Accessible, curated metagenomic data through ExperimentHub. *Nat Methods*. 2017 Oct 31;**14**(11):1023–1024.

4. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*. 2013 Apr 22;**8**(4):e61217.

5. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. nature.com; 2010 May;**7**(5):335–336.

6. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. *ISME J*. 2011 Feb;**5**(2):169–172.

7. Callahan BJ, Sankaran K, Fukuyama JA, McMurdie PJ, Holmes SP. Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses. *F1000Res*. 2016 Jun 24;**5**:1492.

8. Wickham H, Miller E. haven: Import and Export ‘SPSS’, ‘Stata’ and ‘SAS’ Files [Internet]. 2018. Available from: https://CRAN.R-project.org/package=haven

9. Huttenhower C, Knight R, Brown CT, et al. Advancing the microbiome research community. *Cell*. 2014 Oct 9;**159**(2):227–230.

10. Gibbons S, Duvallet C, Alm EJ. Correcting for batch effects in case-control microbiome studies [Internet]. *bioRxiv* 2018 [cited 2018 Mar 28]. p. 165910. Available from: https://www.biorxiv.org/content/early/2018/03/17/165910

11. Chakravorty S, Helb D, Burday M, Connell N, Alland D. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J Microbiol Methods*. 2007 May;**69**(2):330–339.

12. Yang B, Wang Y, Qian P-Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics*. 2016 Mar 22;**17**:135.

13. Wong B. Points of view: Color blindness. *Nat Methods*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2011 May 27;**8**:441.