Abstract

Alterations in the composition of the microbiota have been implicated in many diseases. The Human Microbiome Project (HMP) provides a comprehensive reference dataset of the "normal" human microbiome of 242 healthy adults at five major body sites. The HMP used both 16S ribosomal RNA gene sequencing and whole-genome metagenomic sequencing to profile the subjects' microbial communities. However, accessing and analyzing the HMP dataset still presents technical and bioinformatic challenges, as researchers must import the microbiome data, integrate phylogenetic trees, and access and merge public and restricted metadata. In this issue, the *HMP16SData* R/Bioconductor package developed by Schiffer and colleagues (Am J Epidemiol. XXX; XX (XX): XX–XXX) greatly simplifies access to the HMP data by combining 16S taxonomic abundance data, public patient metadata, and phylogenetic trees as a single data object. The authors also provide an interface for users with approved dbGaP projects to easily retrieve and merge the controlled-access HMP metadata. This package has a broad range of appeal to researchers across disciplines and with various levels of expertise in using R and/or other statistical tools. This will translate to improved data accessibility for public health research, with data from healthy individuals serving as a reference for disease-associated studies.

Keywords: Bioconductor, databases**,** Human Microbiome Project, microbiome, R

Large datasets such as the Human Microbiome Project are needed for robust epidemiology of microbiome-associated diseases

The human microbiota comprises the bacterial, fungal, archaeal and viral occupants of the human body. An increasing number of health conditions have been linked to the gastrointestinal (e.g. inflammatory bowel disease (1), colorectal cancer (2), obesity (3), type 2 diabetes (3), and rheumatoid arthritis (4)), oral (e.g. periodontitis (5), pancreatic cancer (6)), skin (e.g. dermatitis (7) and other cutaneous diseases (8)), and vaginal (e.g. bacterial vaginosis (9)) microbiota. In order to understand how the microbiome changes in disease, it is crucial to first understand the microbial composition and variance within healthy populations. MetaHIT (10) was the first large-scale survey of healthy adult stool microbiota, while the first phase of the Human Microbiome Project (HMP) (11) provided a comprehensive reference dataset of the "normal" human microbiome of healthy individuals at five major body sites (GI tract, nasal, skin, oral cavity, and vagina). Applying these data to create well-designed epidemiological studies and meta-analyses is an important step in determining how dysbiosis (deviation from the normal microbiome) in microbial composition and functional profiles may contribute to disease etiology. While the HMP data is freely available to researchers at the HMP Data Analysis and Coordination Center (HMPDACC) website (12), its format is not readily importable into widely-used statistical software such as R or Stata, presenting a significant technical challenge to researchers. Schiffer and colleagues have substantially reduced these technical challenges by creating the *HMP16SData* R/Bioconductor package (13) which greatly simplifies researcher access to HMP dataset. The *HMP16SData* removes technical hurdles to data processing and provides access to the HMP 16S rRNA data as an object that is ready to analyze. This will provide epidemiologists with a powerful tool to help to identify robust epidemiological links between the microbiome and disease.

The HMP data from HMPDACC is a valuable but technically challenging epidemiological resource

Between 2008 and 2010, the HMP phase I collected thousands of samples from 15 (male) or 18 (female) distinct body sites of 242 healthy adults between the ages of 18 and 40 years over multiple time points. To characterize the composition of samples' microbial communities, the 454 FLX Titanium platform was used to sequence the V35 hypervariable region of the 16S ribosomal RNA (rRNA) gene. The V13 hypervariable region of the 16S rRNA gene was also sequenced for a subset of the samples for a more complete picture of a community (14) and to allow for methods comparison. Finally, a subset of samples underwent whole-genome shotgun (WGS) or metagenomic sequencing of the whole community DNA using the Illumina GAIIx platform. Although 16S profiling is currently the most commonly used and most cost-effective tool for microbiome analysis, WGS (which uses sequencing with random primers to sequence overlapping regions of a genome) allows for more accurate definition of taxa at the species level, as well as allowing for identification and quantitation of functional genes.

A wealth of HMP data is available for download at the HMP's Data Analysis Coordination Center (HMPDACC) (12): 16S rRNA data for hypervariable regions V13 and V35 (e.g. OTU tables describing the abundance of taxa in samples, phylogenetic trees describing the relationships of OTUs, taxonomy data, WGS sequencing data, and some accompanying sample metadata such as sample type, patient sex, sequencing center, and visit number. Sensitive metadata such as subject age and medical history are kept confidential, but researchers who submit an application can be granted access through the National Center for Biotechnology Information (NCBI) database of Genotypes and Phenotypes (dbGAP). Despite its aim of having accessible data and providing user-friendly data retrieval, retrieval and downstream analysis of HMP data from the DACC in its current format still presents as a considerable

bioinformatic challenge, especially to researchers with limited knowledge of HMPDACC and dbGaP procedures. Researchers are required to import the microbiome taxonomy and abundance data, integrate phylogenetic trees, and correctly merge metadata from hundreds of patients with thousands of samples. Furthermore, using restricted metadata requires an application to dbGaP, after which the clinical metadata must be downloaded and merged. Substantial expertise is needed to successfully merge these large datasets into suitable formats for downstream analysis. In order to address this issue, the authors of the *phyloseq* R/Bioconductor package (15) have previously provided a guide on how to import the HMPv35 dataset from HMPDACC (16) and have made the 16S rRNA data from V35 hypervariable regions publicly available as a processed single R object (17). This is a very useful resource, but it does not include the V13 or WGS datasets, and it contains only the limited, publicly-available patient metadata. Here, Schiffer *et al* have improved on this by creating *HMP16SData*, an R dataset that integrates the HMPv13 and HMPv35 16S rRNA data and provides easy access of restricted patient data from dbGaP to researchers with approved dbGaP projects. Furthermore, *HMP16SData* is easily merged with the HMP WGS data, which is available via the *CuratedMetagenomeData* R/Bioconductor package(18).

Ready integration of metadata, taxonomy, abundance, and phylogenetic relationships facilitates quantitative analysis

In 16S rRNA analysis, raw sequencing data is processed by open source computational tools and pipelines such as mothur (19) and QIIME (20), which group sequences into operational taxonomic units (OTUs) based on sequence similarity (usually 97% identity, corresponding approximately but not exactly to species). The number of times each OTU sequence is detected in each sample is summarized as an OTU table. OTU counts can be summarized at various taxonomic levels (e.g., class, order, family, genus) for analysis.

The Bioconductor package *phyloseq (15)* in R (21) is the most popular tool for downstream analysis of microbial sequencing data. It facilitates statistical analysis and creation of publication-quality graphics, and enables reproducible research when used with documentation tools such as markdown (22). A distinctive feature of *phyloseq* is the integration of OTU-clustered data, taxonomic assignments, and associated sample data as a *phyloseq* object. This allows users to investigate the relative abundance and diversity of organisms at various taxonomic levels, which is especially useful in instances where analyses at taxonomic ranks higher than species provide more ecologically meaningful information. Another useful feature of *phyloseq* is the ability to easily agglomerate taxa by their taxonomic ranks. The data integration provided by *phyloseq* greatly simplifies analysis of microbial within the R environment.

A special feature of a microbiome data matrix is that the distribution of individual OTUs is highly skewed and often sparse (23). For example, the bacterial abundance in the human gut microbiome consists of a high proportion of zero counts at lower taxonomic levels (24). The presence of excess zeros presents a challenge when analyzing microbiome data, particularly when comparing between groups. Two common approaches to solve this problem include normalizing transformation (e.g., variance-stabilizing transformations and linear modelling) and log transformation using a generalized linear model (GLM). Several specialized R /Bioconductor packages greatly simplify this statistical modeling processes and are well-integrated with the *phyloseq* package; these include *MetagenomeSeq* (23), which was developed specifically for marker gene analysis, as well as RNA-Seq focused R packages such as *DESeq2 (25)* and *edgeR (26)*.

In this issue, Schiffer *et al*. (13) have provided an elegant way for researchers to quickly access and analyze the HMP phase 1 data. The authors developed the *HMP16SData* R/Bioconductor

package by combining HMP 16S taxonomic abundance data, public and (optionally) restricted patient metadata, and phylogenetic trees as a single data object, which can then be easily converted into a *phyloseq* object or alternatively exported in .csv, STATA, SAS, or SPSS formats for use with other statistical software. The *HMP16SData* package is easy to install via Bioconductor (27), and clear, helpful online documentation is also available in the package vignette (28) as well as the authors' github site (29). In addition, after researchers have requested and obtained access to dbGaP, the *HMP16SData* package includes the *attach_dbGaP* function, which allows decryption and attachment of restricted patient data from dbGaP into the other data. By removing barriers to data access and management, the authors have made microbiome data from the HMP much more accessible to the research community.

Limitations of the HMP16SData

Researchers who wish to compare their own datasets to the HMP datasets should be aware of inherent biases due to different sampling handling, DNA extraction protocols, 16S rRNA gene primer selection, sequencing platforms and bioinformatics processing pipelines (30). Sequencing of the 16S rRNA gene in the HMP samples relied on the now discontinued Roche 454 platform, whereas the majority of amplicon sequencing is now done on the Illumina platform due to its higher throughput and lower cost. The 454 platform yields longer but fewer sequences, while the Illumina provides shorter reads but at much greater sequencing depth. Furthermore, the platforms are prone to different types of sequencing error, which require different forms of error correction during bioinformatic processing of the sequencing data (31, 32). The difference in total number of sequences per sample must be considered when comparing alpha diversity (total number of taxa) in populations between Illumina and 454 data.

However, even if data is rarefied (subsampled to the same number of reads per sample), these platform differences will introduce some biases the researcher should be aware of.

In addition, the taxonomy reference databases and analysis pipelines which were used to process the HMP data have had many version updates (e.g., QIIME 1.3.0 vs. 2.2018.8 current, mothur 1.1.8 vs. 1.40 current); these software differences could introduce additional bias. Finally, it is becoming increasingly common to process Illumina amplicon sequencing data by inferring of amplicon sequence variants (e.g. exact sequence variants) rather than OTUs with 97% sequence similarity. Two examples of sequence variant-calling software include DADA2 (33) and Deblur (34); both of these are implemented in QIIME 2 (35). If the same sequencing data are processed by both QIIME 1.9 and QIIME 2.0, general patterns of taxonomic composition and beta diversity will be similar, but due to differences in correcting for sequencing error, data produced by QIIME 2.0 will give a more conservative and more accurate estimation of the alpha diversity of a sample (36).

Because there is still no batch normalization method for microbiome data, researchers carrying out meta-analyses using microbiome data such as the HMP data must be aware of these batch and data effects. For more comparable results across studies, certain protocol choices including using relative (not absolute) diversity measures, phylogenetic (not taxonomic) analyses, and quantitative (not presence or absence) measures may buffer the inter-protocol variations (30, 36). Schiffer *et al* recognize the biases that occur with cross-study comparisons, and acknowledge the need to reprocess raw HMP data with modern bioinformatics tools.

Conclusions

In summary, the *HMP16SData* R/Bioconductor package developed by Schiffer and colleagues (13) greatly simplifies access to the Human Microbiome Project data, a landmark data resource. In our opinion, this package has a broad range of appeal to researchers across

disciplines and with various levels of expertise in using R and/or other statistical tools. This will translate to improved data accessibility for public health research, with data from healthy individuals serving as a reference for disease-associated studies, or as a baseline for comparing the Western population with microbiomes from other geographic and ethnic cohorts. As additional large-scale datasets become available, *HMP16SData* will be an invaluable tool to the research community to easily access and analyze these resulting datasets and better understand host-microbe interactions.

References

1. Morgan XC, Tickle TL, Sokol H, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol* 2012;13(9):R79.
2. Louis P, Hold GL, Flint HJ. The gut microbiota, bacterial metabolites and colorectal cancer. *Nat Rev Microbiol* 2014;12(10):661-72.
3. Hartstra AV, Bouter KE, Backhed F, et al. Insights into the role of the microbiome in obesity and type 2 diabetes. *Diabetes Care* 2015;38(1):159-65.
4. Scher JU, Abramson SB. The microbiome and rheumatoid arthritis. *Nat Rev Rheumatol* 2011;7(10):569-78.
5. Jorth P, Turner KH, Gumus P, et al. Metatranscriptomics of the human oral microbiome during health and disease. *MBio* 2014;5(2):e01012-14.
6. Fan X, Alekseyenko AV, Wu J, et al. Human oral microbiome and prospective risk for pancreatic cancer: a population-based nested case-control study. *Gut* 2018;67(1):120-7.
7. Kong HH, Oh J, Deming C, et al. Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome Res* 2012;22(5):850-9.
8. Findley K, Grice EA. The skin microbiome: a focus on pathogens and their association with skin disease. *PLoS Pathog* 2014;10(10):e1004436.
9. Ma B, Forney LJ, Ravel J. Vaginal microbiome: rethinking health and disease. *Annu Rev Microbiol* 2012;66:371-89.
10. Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010;464(7285):59-65.
11. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* 2012;486(7402):207-14.
12. NIH Human Microbiome Project. The HMP Data Analysis and Coordination Center. ((https://hmpdacc.org/hmp/)). (Accessed 2 Nov 2018).
13. Schiffer L, Azhar R, Shepherd L, et al. HMP16SData: Efficient Access to the Human Microbiome Project through Bioconductor. *American Journal of Epidemiology* 2018.
14. Human Microbiome Project Jumpstart Consortium. Evaluation of 16S rDNA-based community profiling for human microbiome research. *PLoS ONE* 2012;7(6):e39315.
15. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 2013;8(4):e61217.
16. McMurdie PJ, Holmes S. Human Microbiome Project. (http://joey711.github.io/phyloseq-demo/HMP_import_example.html). (Accessed 2 Nov 2018).
17. McMurdie PJ, Holmes S. Already-imported HMPv35 dataset. (http://joey711.github.io/phyloseq-demo/HMPv35.RData). (Accessed 2 Nov 2018).
18. Pasolli E, Schiffer L, Manghi P, et al. Accessible, curated metagenomic data through ExperimentHub. *Nat Methods* 2017;14(11):1023-4.
19. Schloss PD, Westcott SL, Ryabin T, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology* 2009;75(23):7537-41.
20. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010;7(5):335-6.

21. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2018. (https://www.R-project.org/). (Accessed).
22. Allaire J.J., Horner J., Marti V., et al. markdown: 'Markdown' Rendering for R. 2017. (https://CRAN.R-project.org/package=markdown). (Accessed).
23. Paulson JN, Stine OC, Bravo HC, et al. Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* 2013.
24. Xia Y, Sun J. Hypothesis Testing and Statistical Analysis of Microbiome. *Genes Dis* 2017;4(3):138-48.
25. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;15(12):550.
26. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26(1):139-40.
27. Huber W, Carey VJ, Gentleman R, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* 2015;12(2):115-21.
28. Schiffer L, Azhar R, Shepherd L, et al. HMP16SData 2018. (https://bioconductor.org/packages/devel/data/experiment/vignettes/HMP16SData/inst/doc/HMP16SData.html). (Accessed 1 Nov 2018).
29. Schiffer L, Azhar R, Shepherd L, et al. 16S rRNA Sequencing Data from the Human Microbiome Project. ((https://github.com/waldronlab/HMP16SData)). (Accessed 2 Nov 2018).
30. Sinha R, Abu-Ali G, Vogtmann E, et al. Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. *Nat Biotechnol* 2017;35(11):1077-86.
31. Quince C, Lanzen A, Davenport RJ, et al. Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 2011;12:38.
32. Luo C, Tsementzi D, Kyrpides N, et al. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS One* 2012;7(2):e30087.
33. Callahan BJ, McMurdie PJ, Rosen MJ, et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016;13(7):581-3.
34. Amir A, McDonald D, Navas-Molina JA, et al. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems* 2017;2(2).
35. Bolyen E, Rideout JR, Dillon MR, et al. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. *PeerJ Preprints* 2018;6:e27295v1.
36. Nearing JT, Douglas GM, Comeau AM, et al. Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ* 2018;6:e5364.