Response to Commentary: Improving accessibility of the Human Microbiome Project data through integration with R/Bioconductor

Levi Waldron, Lucas Schiffer, Rimsha Azhar, Marcel Ramos, Ludwig Geistlinger, and Nicola Segata

Correspondence to Dr. Levi Waldron, Graduate School of Public Health and Health Policy, City University of New York, 55 West 125th Street, New York, NY 10027 (e-mail: levi.waldron@sph.cuny.edu)

This research was supported by the National Institute of Allergy and Infectious Diseases (1R21AI121784-01 to J.B.D. and L.W.), the National Cancer Institute (5U24CA180996 to Martin Morgan), 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.

Conflicts of interest: none declared.

Running head: HMP16SData

Word count: 600 / 600

We thank Griffith and Morgan (1) for their excellent summary of the Human Microbiome Project phase 1 dataset and our efforts to remove technical hurdles to its use by epidemiologists. Their commentary provides a clear overview of our *HMP16SData* (2) Bioconductor (3) package and the necessary precautions for users of these data. In this reply, we expand more generally on the need to lower barriers to reuse of public-access genomic datasets.

The importance of public availability of published data is already broadly accepted across disciplines from perspectives of reproducibility, transparency, and further scientific discovery. Open resistance to data sharing and reuse policies (e.g. to "research parasites" (4)) has been overwhelmed, and the prevalence of data sharing has expanded due to journal policies (such as this journal, which adopts recommendations of the International Committee for Medical Journal Editors (5)), funding policies (such as the NIH genomic data sharing policy (6) and the European Commission Open Research Data Pilot (7)), and recognition of its importance by authors and peer reviewers. The benefit of data sharing, however, comes "not from providing access to data or depositing them somewhere, but from making it possible for others to find and reanalyze the data in a meaningful way." (8) Towards this objective, however, there is less consensus about how to move forward.

Our manuscript and the commentary by Griffith and Morgan highlight technical barriers to utilizing the HMP 16S rRNA gene sequencing dataset, but such barriers are by no means limited to this dataset. De-centralized researcher-driven studies provide a majority of publicly available genomic data, and present additional challenges of standardization and completeness. For example, the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database enforces provision only of a minimal set of mandatory metadata that are relevant across all areas of genomic investigation (such as library and instrument information, and species) (9), whereas the participant metadata of critical interest

in epidemiology are provided with no requirements for inclusion or vocabulary. Key attributes such as age, sex, and disease status may be missing, and when present they must be cleaned and standardized. Our related *curatedMetagenomicData* project (10) developed a system for manual standardization and automatic syntax-checking of participant metadata when made possible by the voluntary provision of key metadata by the researchers who upload data. The adoption of more specific standards for how metadata from health studies are shared would make such manual standardization unnecessary, but significant practical work and consensus-building remains. Groups like the Society for Epidemiological Research may be able to play a leadership role in establishing such community standards.

The growth of multi'omic datasets, where multiple types of molecular data are collected on the same specimens, raises additional bioinformatic hurdles to reanalysis. Such datasets may require multiple data processing pipelines and complex data linkage. The "Integrative Human Microbiome Project" (iHMP) (11) is providing longitudinal measurements of metagenomics, metatranscriptomics, metabolomics, metaproteomics, and other data, presenting an even greater data integration challenge than the current project. Such complex datasets can leave error-prone and non-generalizable sets of tasks to perform for every analysis, exposing limitations in traditional approaches to data management. We and others are working to use recent software for multi'omic data integration in Bioconductor (12) to provide a similar level of usability for the iHMP data.

In summary, the sharing of research data is key to allowing reproducibility of existing studies and to maximizing research investments in public health. However, the details of that sharing and ongoing community efforts towards standardization will determine the extent to which hard-earned and expensive research data are used to their full potential for public good.

References

1. Jocelyn C Griffith and Xochitl C Morgan. Invited Commentary: Improving accessibility of the Human Microbiome Project data through integration with R/Bioconductor. *Am. J. Epidemiol.* XXX(XX):XX–XXX.

2. Schiffer L, Azhar R, Shepherd L, et al. HMP16SData: Efficient Access to the Human Microbiome Project through Bioconductor. *Am. J. Epidemiol.* XXX(XX):XX–XXX.

3. Huber W, Carey VJ, Gentleman R, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*. 2015;12(2):115–121.

4. Longo DL, Drazen JM. Data Sharing. *N. Engl. J. Med.* 2016;374(3):276–277.

5. Taichman DB, Sahni P, Pinborg A, et al. Data Sharing Statements for Clinical Trials - A Requirement of the International Committee of Medical Journal Editors. *N. Engl. J. Med.* 2017;376(23):2277–2279.

6. National Institutes of Health. NIH genomic data sharing policy. 2014;(Notice number NOT-OD-14-124)

7. Guedj D, Ramjoué C. European Commission Policy on Open-Access to Scientific Publications and Research Data in Horizon 2020. *Biomed Data J*. 2015;01(1):11–14.

8. Haug CJ. From Patient to Patient--Sharing the Data from Clinical Trials. *N. Engl. J. Med.* 2016;374(25):2409–2411.

9. Leinonen R, Sugawara H, Shumway M, et al. The sequence read archive. *Nucleic Acids Res.* 2011;39(Database issue):D19–21.

10. Pasolli E, Schiffer L, Manghi P, et al. Accessible, curated metagenomic data through ExperimentHub. *Nat.*

Methods. 2017;14(11):1023–1024.

11. Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe*. 2014;16(3):276–289.

12. Ramos M, Schiffer L, Re A, et al. Software for the Integration of Multiomics Experiments in Bioconductor. *Cancer Res.* 2017;77(21):e39–e42.