AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

PERSPECTIVE
Clinical Science and Epidemiology

Data and Statistical Methods To Analyze the Human Microbiome

Levi Waldron[a,b]

[a]Graduate School of Public Health and Health Policy, City University of New York, New York, New York, USA
[b]Institute for Implementation Science in Population Health, City University of New York, New York, New York USA

ABSTRACT The Waldron lab for computational biostatistics bridges the areas of cancer genomics and microbiome studies for public health, developing methods to exploit publicly available data resources and to integrate -omics studies.

KEYWORDS machine learning, meta-analysis, metagenomics, statistical analysis

The rapidly developing field of human microbiome studies will benefit from adapting the statistical and computational methods of more mature areas of high-dimensional data analysis and from ongoing use of the growing catalog of publicly available microbiome data. This perspective discusses methods and resources for robust identification of differentially abundant microbes and predictive models of microbiome-linked health outcomes. I summarize lessons from high-dimensional data analysis for cancer genomics and efforts by my lab to leverage and adapt the Bioconductor project for analysis and comprehension of high-throughput genomic data (1) to bring value-added published data, meta-analysis, and methods for multiomic data analysis to the microbiome community.

COMPARATIVE ANALYSIS AND META-ANALYSIS FOR DIFFERENTIAL ABUNDANCE

Differential abundance analysis is probably the most common objective of microbiome profiling studies and genomics studies in general. The objective is to identify microbial taxa, anywhere on the tree of life, that are over- or underabundant in some condition relative to a reference condition. These conditions can be observed or experimentally determined. The most commonly used methods for differential abundance analysis are LEfSe (2) and a variety of tools based on log linear regression models with negative binomial (3) or zero-inflated Gaussian error models (4). Regression approaches involve a false-discovery rate estimation to correct for multiple-hypothesis testing. Log linear modeling approaches build on a large body of statistical and computational work and provide several practical advantages. First, regression approaches eliminate the need for rarefaction, a process that has been described as "inadmissable" for the identification of differentially abundant taxa (5) because it throws away potentially useful data, the extra reads from samples with greater sequencing depth. Second, they adapt empirical Bayesian methods developed to reduce false-positive results in microarray differential expression analysis by "borrowing" information across taxa on how taxa are distributed across samples. Finally, they accommodate multivariate models that can be used for causal inference, such as to control for confounding effects or to test hypotheses of the microbiome as a mediator between environmental exposure and health outcomes. Regression modeling, now the almost exclusive choice for differential expression analysis of RNA sequencing data, is also well suited to metatranscriptomic differential abundance analysis.

Received 18 November 2017 **Accepted** 7 December 2017 **Published** 13 March 2018

Citation Waldron L. 2018. Data and statistical methods to analyze the human microbiome. mSystems 3:e00194-17. https://doi.org/10.1128/mSystems.00194-17.

Address correspondence to levi.waldron@sph.cuny.edu.

Conflict of Interest Disclosures: L.W. has nothing to disclose.

mSystems® *vol. 3, no. 2, is a special issue sponsored by Janssen Human Microbiome Institute (JHMI).*

Public data and statistical methods to move human microbiome studies ahead

These efforts can be enhanced by the standardization and reuse of published data for meta-analysis, comparative analysis, and method development. Thus, my lab developed the curatedMetagenomicData database (6) in collaboration with the laboratories of Nicola Segata (MetaPhlAn2 [7] and other methods for metagenomics), Curtis Huttenhower (developers of the bioBakery [8] and many methods therein), and Martin Morgan (head of the Bioconductor project [1]). This database provides more than 6,000 human-associated shotgun metagenomic profiles, uniformly processed from raw sequencing data to provide taxonomic abundance (7) and metabolic functional potential (9). Samples are primarily from stool specimens but include the Human Microbiome Project and other data sets sampling from other human body sites. We developed a fully automated, cloud-based pipeline to facilitate ongoing addition and updating of the database as new metagenomes and reference genomes become available and to encourage community contributions and even creation of alternative and competing databases.

MULTIOMIC INVESTIGATION OF THE MICROBIOME

Metagenomic studies, as in other areas of genomics, increasingly incorporate multiple assays in an experiment. My lab recently published MultiAssayExperiment (10), software for the integration of multiomics experiments in Bioconductor. MultiAssayExperiment has enabled coordinated representation and manipulation of multiple -omics data types for 11,000 patients and 33 cancers studied as part of the Cancer Genome Atlas. A more complete picture of host-microbiome relationships may also be developed by collecting multiple -omics data types, and I have been involved in studies including metatranscriptomics (11) and host gene expression (12) in addition to taxonomic and functional microbiome abundance data. To overcome the complexity of reproducible data analysis and interpretation of such experiments, I am working with other Bioconductor microbiome package developers to create a common standard for representing microbiome data. This standard will provide compatibility with MultiAssayExperiment and with recent advances based on HDF5 and Google BigTable for on-disk data and remote representation of very large data. This will, for example, allow curatedMetagenomicData (6) to represent taxonomic, gene family, and metabolic functional profiles for more than 6,000 samples as a single Bioconductor object that users can interact with in almost the same way as they currently do with microbiome (4, 13) or gene expression data from a single study, even on a standard laptop.

PREDICTIVE MODELING/MACHINE LEARNING

Prediction of health outcomes is a complementary objective to differential abundance analysis. Although similar models are sometimes used for these different objectives, the objective of making accurate predictions motivates different methods for model development and assessment. A mainstream approach to prediction modeling in high-dimensional data is to apply multivariate penalized regression, or machine learning methods such as Support Vector Machine, in conjunction with cross-validation to assess prediction accuracy. These approaches have been quickly adopted for prediction of health status from microbiome data. Colleagues and I have previously shown in meta-analyses of cancer transcriptomes that such approaches are prone to overoptimistic estimation of prediction accuracy (14). There are numerous possible reasons for such overoptimism. The data used to develop prediction models are by necessity retrospective, meaning they are predicting the past and not the future. "Information leakage" in data set through incorrect cross-validation, "reverse causality" effects of treatment on the microbiome, batch effects introduced by knowledge of outcomes, for example by sequencing cases together and then sequencing controls in another batch. Most studies do not collect statistically random samples, and therefore, the samples are not representative of the population.

Even with these challenges, it is sometimes still possible to develop accurate models of disease state and outcome from high-dimensional data. Colleagues and I showed that systematic leave-one-data set-in cross-study validation (15) of independent pub-

licly available data sets provides a more realistic picture of generalizable prediction accuracy and that heterogeneous studies can be used to train robust prediction models through leave-one-data set-out cross-study validation (16). We have also shown the value of these approaches for metagenomic prediction problems (17). In systematic cross-study validation of gene expression-based models of cancer patient prognosis, we have shown even simple and suboptimal machine learning algorithms to be competitive with complex, theoretically optimal methods (18). Standardized databases like curatedMetagenomicData (6) and our in-development HMP16SData package (http://bioconductor.org/packages/HMP16SData/) will facilitate future work to find the limits of accuracy for disease prediction from all available microbiome profiles.

FUTURE OUTLOOK

Discoveries that are replicable across independent experiments are more likely to be valid and useful than those seen only in a single data set. My research aims to harness publicly available microbiome data through curation, integration and standardization, novel reanalysis, and methodological development. I aim to ensure that studies of the human microbiome benefit from concurrent methodological development in other areas of genomics and from the growing body of publicly available microbiome data. These benefits include more reliable identification of differentially abundant microbial species, strains, and community structure and the development of disease prediction models that hold up to independent validation across populations. I see the Bioconductor project as providing a unique opportunity for the microbiome community to leverage more than 15 years of development of statistical methods for -omics data and to integrate microbiome data with other types of high-throughput data. As such, I plan to continue developing the Bioconductor platform to the needs of the microbiome community, through the development of databases, promotion of standards for data representation, and development of needed methods for data manipulation and analysis.

ACKNOWLEDGMENTS

The work discussed in this perspective was funded by the National Cancer Institute (U24CA180996) and by the National Institute of Allergy and Infectious Diseases (1R21AI121784-01) of the National Institutes of Health.

REFERENCES

1. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MacDonald J, Obenchain V, Oleś AK, Pagès H, Reyes A, Shannon P, Smyth GK, Tenenbaum D, Waldron L, Morgan M. 2015. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods 12:115–121. https://doi.org/10.1038/nmeth.3252.
2. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. 2011. Metagenomic biomarker discovery and explanation. Genome Biol 12:R60. https://doi.org/10.1186/gb-2011-12-6-r60.
3. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550. https://doi.org/10.1186/s13059-014-0550-8.
4. Paulson JN, Stine OC, Bravo HC, Pop M. 2013. Differential abundance analysis for microbial marker-gene surveys. Nat Methods 10:1200–1202. https://doi.org/10.1038/nmeth.2658.
5. McMurdie PJ, Holmes S. 2014. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol 10:e1003531. https://doi.org/10.1371/journal.pcbi.1003531.
6. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, Beghini F, Malik F, Ramos M, Dowd JB, Huttenhower C, Morgan M, Segata N, Waldron L. 2017. Accessible, curated metagenomic data through ExperimentHub. Nat Methods 14:1023–1024. https://doi.org/10.1038/nmeth.4468.
7. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods 12:902–903. https://doi.org/10.1038/nmeth.3589.
8. McIver LJ, Abu-Ali G, Franzosa EA, Schwager R, Morgan XC, Waldron L, Segata N, Huttenhower C. 29 November 2017. bioBakery: a meta'omic analysis environment. Bioinformatics https://doi.org/10.1093/bioinformatics/btx754.
9. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, Rodriguez-Mueller B, Zucker J, Thiagarajan M, Henrissat B, White O, Kelley ST, Methé B, Schloss PD, Gevers D, Mitreva M, Huttenhower C. 2012. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol 8:e1002358. https://doi.org/10.1371/journal.pcbi.1002358.
10. Ramos M, Schiffer L, Re A, Azhar R, Basunia A, Rodriguez C, Chan T, Chapman P, Davis SR, Gomez-Cabrero D, Culhane AC, Haibe-Kains B, Hansen KD, Kodali H, Louis MS, Mer AS, Riester M, Morgan M, Carey V, Waldron L. 2017. Software for the integration of multiomics experiments in Bioconductor. Cancer Res 77:e39–e42. https://doi.org/10.1158/0008-5472.CAN-17-0344.
11. Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, Giannoukos G, Boylan MR, Ciulla D, Gevers D, Izard J, Garrett WS, Chan AT, Huttenhower C. 2014. Relating the metatranscriptome and metagenome of the human gut. Proc Natl Acad Sci U S A 111:E2329–E2338. https://doi.org/10.1073/pnas.1319284111.
12. Morgan XC, Kabakchiev B, Waldron L, Tyler AD, Tickle TL, Milgrom R, Stempak JM, Gevers D, Xavier RJ, Silverberg MS, Huttenhower C. 2015. Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. Genome Biol 16:67. https://doi.org/10.1186/s13059-015-0637-x.

13. McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8:e61217. https://doi.org/10.1371/journal.pone.0061217.
14. Waldron L, Haibe-Kains B, Culhane AC, Riester M, Ding J, Wang XV, Ahmadifar M, Tyekucheva S, Bernau C, Risch T, Ganzfried BF, Huttenhower C, Birrer M, Parmigiani G. 2014. Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. J Natl Cancer Inst 106:dju049. https://doi.org/10.1093/jnci/dju049.
15. Bernau C, Riester M, Boulesteix A-L, Parmigiani G, Huttenhower C, Waldron L, Trippa L. 2014. Cross-study validation for the assessment of prediction algorithms. Bioinformatics 30:i105–i112. https://doi.org/10.1093/bioinformatics/btu279.
16. Riester M, Wei W, Waldron L, Culhane AC, Trippa L, Oliva E, Kim S-H, Michor F, Huttenhower C, Parmigiani G, Birrer MJ. 2014. Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. J Natl Cancer Inst 106:dju048. https://doi.org/10.1093/jnci/dju048.
17. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. 2016. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Comput Biol 12:e1004977. https://doi.org/10.1371/journal.pcbi.1004977.
18. Zhao SD, Parmigiani G, Huttenhower C, Waldron L. 2014. Más-o-menos: a simple sign averaging method for discrimination in genomic data analysis. Bioinformatics 30:3062–3069. https://doi.org/10.1093/bioinformatics/btu488.