

Bacterial identification and analytic challenges in clinical microbiome studies

J. Kirk Harris, PhD,^a and Brandie D. Wagner, PhD^b Aurora, Colo

Key words: Microbiome, statistics, feature selection

The importance of the human microbiome is rapidly emerging.¹ Our bodies host complex communities of microbes with much greater genetic diversity than our own genome.² This complexity is variable across subjects,³ and there is still much to learn about the processes that drive the establishment and maintenance of these communities.⁴ More importantly, these communities fundamentally affect the normal function of our bodies and thus are critical to understand as modifiers of human processes, including disease.^{5,6}

The gut is an important interface with the environment and a major source of antigen exposure throughout life.⁷ In this issue of the *Journal*, Abrahamsson et al⁸ report on the relationship between bacterial diversity in the gut and the development of IgE-positive atopic eczema in the article entitled “Low diversity of the gut microbiota in infants with atopic eczema.” Three stool samples were available for study from the first year of life from 20 infants with eczema and 20 control subjects with no allergic manifestation before 2 years of age. Infants with eczema had lower stool diversity than control subjects at 1 month. Furthermore, particular subsets of bacteria differed between the groups, suggesting a specific relationship between these organisms and disease. These results are supported by several studies showing relationships between eczema and specific constituents of the infant gut microbiota.^{9–11} However, larger prospective trials have found that no particular bacterial groups were associated with the development of eczema.^{12,13} This might reflect geographic differences that confound larger studies but do not affect more localized recruitment. Resolving the differences between different studies requires consideration of how microbiome studies are analyzed, particularly with respect to feature selection, statistical approaches to higher-dimensional datasets, and taxonomic identification.

MICROBIOME DATA AND FEATURE SELECTION

Microbiome datasets are complex high-dimensional data structures with several key features not present in other

high-dimensional data. Sequence counts, or relative abundance, are used to represent the community present in each sample. These counts are nonnegative and often skewed. In addition, given the high variability between subjects, many taxa are seen in only a subset of subjects and thus result in a larger than expected number of zero values. These characteristics distinguish microbiome data from other types of high-dimensional data and are not adequately addressed by using common statistical approaches.

Studies of the microbiome relative to disease are important and require appropriately tailored analytic approaches to identify taxa that distinguish patients from control subjects. These approaches are critical to select the data features that differentiate the groups. Abrahamsson et al⁸ circumvented the data issues described (apparent in Table III; eg, *Erysipelotrichi*) by analysis of Shannon diversity between patients and control subjects. This approach is interesting, but like all community metrics, it requires conversion of the primary data into a single metric, and thus the detail is lost. The analysis at various taxonomic levels to identify lineages differing between groups is an interesting approach to attempt to uncover the contribution of related organisms to the difference seen in diversity. We note a few other feature-selection approaches applicable to sequence count data.^{14–16} More general analytic approaches based on statistical modeling are needed to allow incorporation of clinical information but require development to accommodate microbiome data features.

TAXON IDENTIFICATION

Names are important. The primary means for linking microbiome data to the rich culture-based clinical literature is through identification of the organism that contributed the sequence. However, sequence-based approaches provide information that allows variable levels of identification. For a sequence that is identical to reference, we have high confidence of the source organism. As the sequence diverges, the confidence one can place on the name shifts away from the specific to more general terms at the genus or sometimes phylum level: thus the use of taxon to identify sequence groups instead of the term organism. This also highlights the strong reliance on approaches developed around acute infectious disease, in which treatment of single causal organisms results in disease resolution. This approach will likely prove fruitful in microbiome research but might not accommodate all microbiome-related problems in human disease.

Most existing systems do not attempt species-level identification based on sequence data at this time. One reason for this is that the short sequence reads produced by high-throughput sequencing platforms provide less information and are sometimes inadequate for species-level identification. Abrahamsson et al⁸ used the Basic Local Alignment Search Tool (BLAST; National Center for Biotechnology Information, Bethesda, Md) for sequence identification. This approach is capable of very specific identification of sequences but lacks guidelines on how similar a sequence must

From ^athe Department of Pediatrics, University of Colorado Anschutz Medical Campus, and ^bthe Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado.

Disclosure of potential conflict of interest: The authors declare that they have no relevant conflicts of interest.

Received for publication December 1, 2011; accepted for publication December 15, 2011.

Corresponding author: J. Kirk Harris, PhD, Department of Pediatrics, University of Colorado Anschutz Medical Campus, 13123 E 16th Ave, Box B395, Aurora, CO 80045. E-mail: Jonathan.Harris@ucdenver.edu.

J Allergy Clin Immunol 2012;129:441–2.

0091-6749/\$36.00

© 2012 American Academy of Allergy, Asthma & Immunology

doi:10.1016/j.jaci.2011.12.969

be for taxonomic assignment. In this study the taxonomic rank of the query sequence was assigned if the BLAST result was 95% or greater identical over more than 180 nucleotides (approximately 91% coverage), which is fairly broad but likely appropriate for the genus-level or higher assignments used. Other systems (eg, Ribosomal Database Project classifier¹⁷) do have guidelines for assignment of taxonomic rank and use bootstrapping to establish specific CIs for the assignments but do not attempt to classify beyond genus. The limitation is that no existing system has all the capability required to attempt reliable assignment of species-level names.

Systems are needed to identify potential targets for modulation within communities. The approach for identification of organisms might be based on taxonomic classification or phylogenetic analysis but likely will require the ability to communicate effectively without reliance on traditional organism names. We continue to learn more about the genomic potential of the vast human microbiome and the corresponding metabolic pathways; however, much work remains to uncover the contribution of shifts in communities or function to human health. Ultimately, control of communities through combinations of antibiotics, prebiotics, and probiotics might provide the best approach for managing or preventing disease. This would represent an ecologic approach that currently is not possible. Further investigation that identifies divergence from healthy community structure is required. Once we understand what leads to health, modulation of communities toward the healthy state can be studied.

REFERENCES

1. Relman DA. Microbial genomics and infectious diseases. *N Engl J Med* 2011;365:347-57.
2. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, et al. Metagenomic analysis of the human distal gut microbiome. *Science* 2006;312:1355-9.
3. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut microbiome in lean and obese twins. *Nature* 2009;457:480-5.
4. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, et al. Succession of microbial consortia in the developing infant gut microbiome. *Proc Natl Acad Sci U S A* 2011;108(suppl 1):4578-85.
5. Fujimura KE, Slusher NA, Cabana MD, Lynch SV. Role of the gut microbiota in defining human health. *Expert Rev Anti Infect Ther* 2010;8:435-54.
6. Frank DN, Zhu W, Sartor RB, Li E. Investigating the biological and clinical significance of human dysbioses. *Trends Microbiol* 2011;19:427-34.
7. McLoughlin RM, Mills KHG. Influence of gastrointestinal commensal bacteria on the immune responses that mediate allergy and asthma. *J Allergy Clin Immunol* 2011;127:1097-107.
8. Abrahamsson TR, Jakobsson HE, Andersson AF, Björkstén B, Engstrand L, Jenmalm MC. Low diversity of the gut microbiota in infants with atopic eczema. *J Allergy Clin Immunol* 2012;129:434-40.
9. Björkstén B, Naaber P, Sepp E, Mikelsaar M. The intestinal microflora in allergic Estonian and Swedish 2-year-old children. *Clin Exp Allergy* 1999;29:1-5.
10. Watanabe S, Narisawa Y, Arase S, Okamatsu H, Ikenaga T, Tajiri Y, et al. Differences in fecal microflora between patients with atopic dermatitis and healthy control subjects. *J Allergy Clin Immunol* 2003;111:587-9.
11. Kalliomäki M, Kirjavainen PV, Eerola E, Kero P, Salminen SJ, Isolauri E. Distinct patterns of neonatal gut microflora in infants in whom atopy was and was not developing. *J Allergy Clin Immunol* 2001;107:129-34.
12. Adlerberth I, Strachan DP, Matricardi PM, Ahrné S, Orfei L, Åberg N, et al. Gut microbiota and development of atopic eczema in 3 European birth cohorts. *J Allergy Clin Immunol* 2007;120:343-50.
13. Penders J, Stobberingh EE, Thijs C, Adams H, Vink C, van Ree R, et al. Molecular fingerprinting of the intestinal microbiota of infants in whom atopic eczema was or was not developing. *Clin Exp Allergy* 2006;36:1602-8.
14. Wagner BD, Robertson CE, Harris JK. Application of two-part statistics for comparison of sequence variant counts. *PLoS ONE* 2011;6:e20296.
15. Zhou Y, Xia K, Wright FA. A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics* 2011;27:2672-8.
16. White JR, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol* 2009;5:e1000352.
17. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 2007;73:5261-7.