AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

PERSPECTIVE
Host-Microbe Biology

Exploring Linkages between Taxonomic and Functional Profiles of the Human Microbiome

Morgan G. I. Langille[a]

[a]Department of Pharmacology, Dalhousie University, Halifax, Nova Scotia, Canada

ABSTRACT Microbiome studies typically focus on characterizing the taxonomic and functional profiles of the microbes within a community. Functional profiling is generally thought to be superior to taxonomic profiling for investigating human-microbe interactions, but there are several limitations and challenges to existing approaches. This Perspective discusses the current sequencing and bioinformatic methods for producing taxonomic and functional profiles, recent studies utilizing and comparing these technologies, and the existing challenges and limitations of these data. In addition, functional versus taxonomic conservation across the population is questioned, while future research that focuses on investigating the taxonomic diversity of microbial functions is proposed.

KEYWORDS 16S, function, human microbiome, metagenomics, taxonomy

The characterization of microbiomes can be undertaken using various technologies. The most prevalent are 16S rRNA gene surveys (16S) and metagenome studies. Although the 16S approach is less expensive, it is routinely seen as inferior to metagenomics since the former is limited to identifying only the taxa that can be amplified by the chosen set of "universal" primers, thus biasing particular clades of bacteria and archaea and often missing all of microbial eukaryotes and viruses. Further, metagenomics provides insights into the functional capabilities of the microbiome by profiling relative abundances of genes within the microbial community, while 16S is primarily limited to describing taxonomic changes. However, for many host-associated microbiome environments, 16S is the only feasible profiling technique due to host contamination swamping the majority of microbial signal from metagenomic techniques. Therefore, 16S is likely to be a commonly used method, even as sequencing costs continue to decrease.

Predictive methods that leverage large reference genome databases and ancestral state reconstruction methods such as PICRUSt (1) provide insight into the functional repertoire based on 16S profiles. These predictive methods provide functional hypotheses that, like many other technologies, can be validated with specific sequencing primers or through metabolomic analysis. PICRUSt 2.0 (https://github.com/picrust/picrust2) is being actively developed, with a final release being planned in 2018, and provides several new features in comparison to its predecessor. These include predictions based on any 16S sequence originating from an operational taxonomic unit (OTU) clustering approach or from an amplicon sequence variant approach (2), which provides resolution to the level of single nucleotide differences in the 16S rRNA gene instead of the previous 97% sequence identity cutoff commonly used in OTUs. In addition, PICRUSt 2.0 will be based on over 39,000 genomes (a nearly 10-fold increase in reference genomes) and will provide predictions that integrate with the MetaCyc functional framework (3). The accuracy of predicting the eukaryotic functional proportion of the microbiome based on 18S rRNA gene profiles is also being tested and validated. A major limitation of both metagenomics and PICRUSt inferences is that they

Received 30 October 2017 **Accepted** 16 January 2018 **Published** 27 March 2018

Citation Langille MGI. 2018. Exploring linkages between taxonomic and functional profiles of the human microbiome. mSystems 3:e00163-17. https://doi.org/10.1128/mSystems.00163-17.

Address correspondence to morgan.g.i.langille@dal.ca.

Conflict of Interest Disclosures: M.G.I.L. has nothing to disclose.

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

depend on accurate gene annotations. Previous research has shown that microbial gene annotations are notoriously inaccurate, making biological interpretations of microbiome community function uncertain (4). In addition, these genes may not be transcribed or translated, limiting the impact of their annotated function. Conclusions about microbiome function derived either from metagenomics or from PICRUSt should be treated as hypotheses that require further in-depth validation through functional assays.

Overall, it is intuitive and generally thought that function is much more informative than taxonomic information since it is what the organisms do that we care about and not who they are (5). Indeed, it has been noted by several groups that function seems more highly conserved across samples than across taxa, suggesting that function is more resilient across communities than the individual strains that come and go (6). However, comparisons of taxa and function conservation that were more technical and philosophical in nature have suggested that these comparisons are not meaningful due to their being based on completely different scales (7). For example, are metabolic pathways equivalent to taxonomic phyla, genera, or species? The problem is that, although taxa and function are linked, it is impossible to access them on similar scales. As expected, when using comprehensive gene families instead of broadly conserved functional pathways, functional conservation disappears (7). Therefore, describing functions as being more conserved than taxa is an artifact of the methods and databases used in the comparison rather than an actual biological statement. Nonetheless, function provides information on possible mechanisms present between microbes and in microbe-host interactions. These interactions are essential for understanding and modeling microbial communities, especially with respect to the various microbiome-related diseases.

Considering that the majority of disease-microbiome relationships are not defined by a single species but rather by a complex community of microbes, machine learning methods are an obvious approach for understanding these complex data sets. Machine learning methods take as the input sets of features such as abundances of taxa, genes, transcripts, etc., and a training data set for classifications such as those resulting from comparisons of disease patients to healthy controls, responses to treatment, etc., and output a classifier that can be used on a novel data set. Machine learning methods based on metagenomic taxonomic profiles have shown promise for the classification of various diseases such as colorectal cancer, obesity, diabetes, and Crohn's disease (8). Taxonomic profiles based on 16S data have also been successful in the identification of subtle differences in samples from the gut microbiome of moderately exercised mice (9). It would seem intuitive that inclusion of relative abundances of genes in these models would help improve classification accuracy. Indeed, there are several examples where changes in gene abundances have been more informative than taxonomic differences in examples such as predicting the gut colonization of a strain of *Bifidobacterium longum* (10). However, a previous observation indicated that machine learning models built with 16S taxa were just as accurate as gene abundance profiles predicted using PICRUSt (11). Further, we recently showed that the levels of accuracy of data from metagenomic-based taxonomic profiles and gene abundance profiles were very similar in looking at predicting disease and treatment outcome in pediatric Crohn's disease (12). These results suggest that there may be major limitations in how we are currently defining and using metagenomic gene abundance profiles. One limitation is that these analyses are based on data in the KEGG database, which is well annotated but not very comprehensive. The relevant functional differences could easily be hidden within those genes that are currently "unknown." Although genes of unknown function are initially limiting in their biological interpretation, they could provide completely novel insights if they are among the major features used in classifying a microbiome-related disease. Again, inaccuracies in gene annotation, along with not measuring levels of RNA, protein, or metabolites, could also be further hampering the accuracy of classification.

One side benefit of many machine learning methods is that they often provide

insight into the most important features for classification. These features provide identification of possible species or functions to be further investigated and are not limited to only those previously characterized. Further, these different types of microbiome features can be combined into a single model to test improved accuracy or to determine the relative levels of importance of different feature types. For example, we recently investigated whether host genetics, metagenomics-based taxon profiles that included viruses and microbial eukaryotes, metagenomic-based gene abundances, 16S taxon profiles, or simply alpha diversity were the most useful for predicting disease and treatment outcome from gut biopsy samples in a pediatric Crohn's disease cohort (12). We found that, while host genetics and alpha diversity profiles were statistically significant in detecting Crohn's disease, they had much lower accuracy than 16S genera profiles and was not predictive at all in determining treatment outcome. Further, we found that combining features from different technologies produced a more accurate classifier. Choosing the correct features for machine learning use is an active area of research. Some diseases can be optimally predicted using simple 16S taxonomic profiles, while others require gene abundance information. Other diseases may require going beyond genomics to measuring levels of RNA transcripts, proteins, and metabolites.

Another obvious factor in disease classification from microbiome data is that we need not use taxon information and functional information independently as is currently done in most bioinformatic analyses. Methods that provide linkages between taxa and their respective functions provide much richer and biological relevant information. PICRUSt has the ability to output the taxonomic contributions of its predictions, while Humann2 (http://huttenhower.sph.harvard.edu/humann2) provides links between taxa and functional annotations for metagenomic data. In addition, methods like FishTaco provide methods to link taxonomy to function using statistical and modeling frameworks (13). These methods provide novel information not simply on how the relative abundance of a particular gene changed but on what organism contributed to those changes. For example, the loss of a particular function could be more biologically interesting if that function had been contributed by 10 different species than by a single species. Functions within the human microbiome were recently characterized as core pathways (one body site), multicore pathways (several body sites), and supercore pathways (all body sites), and the taxonomic contributions of each of these were characterized as being different depending on the body site (14). This type of research forms the foundation for further characterizing taxonomic contributions to various functions in the healthy human microbiome. Further research is needed to determine if these functional contributions are consistent across various diseases or whether there exist unique signatures representing correspondences between certain functions and taxa in particular diseases.

A major hurdle that remains for developing microbiome signatures for precision medicine is the limited robustness of classifiers across data sets. Due to the heterogeneity of DNA extraction, library preparation, sequencing, and bioinformatic techniques, meta-analyses are often difficult to conduct. However, despite technical differences, several studies have demonstrated that biological signal does often remain in 16S (15, 16) and metagenomic (8, 17) data sets. These types of studies depend on the public release of microbiome data, and it is in the best interest of all researchers to uphold open data standards (18). In addition, well-established and open standard operating procedures (19) and the development of standards for sample collection and sequencing (20) and of statistical methods for correcting for batch effects (21) will likely make meta-analysis more powerful in the future.

Microbiome studies will continue to rely on existing sequencing technologies. However, the issue of how we leverage this information to move beyond simply listing and cataloging individual microbial taxa and gene abundances is at the forefront of understanding and modeling microbial communities and harnessing the full potential of the human microbiome.

REFERENCES

1. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower C. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 31:814–821. https://doi.org/10.1038/nbt.2676.
2. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods 13:581–583. https://doi.org/10.1038/nmeth.3869.
3. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Subhraveti P, Weaver DS, Karp PD. 2016. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 44:D471–D480. https://doi.org/10.1093/nar/gkv1164.
4. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, Pandey G, Yunes JM, Talwalkar AS, Repo S, Souza ML, Piovesan D, Casadio R, Wang Z, Cheng J, Fang H, Gough J, Koskinen P, Törönen P, Nokso-Koivisto J, Holm L, Cozzetto D, Buchan DWA, Bryson K, Jones DT, Limaye B, Inamdar H, Datta A, Manjari SK, Joshi R, Chitale M, Kihara D, Lisewski AM, Erdin S, Venner E, Lichtarge O, Rentzsch R, Yang H, Romero AE, Bhat P, Paccanaro A, Hamp T, Kaßner R, Seemayer S, Vicedo E, Schaefer C, et al. 2013. A large-scale evaluation of computational protein function prediction. Nat Methods 10:221–227. https://doi.org/10.1038/nmeth.2340.
5. Doolittle WF, Booth A. 2017. It's the song, not the singer: an exploration of holobiosis and evolutionary theory. Biol Philos 32:5–24. https://doi.org/10.1007/s10539-016-9542-2.
6. Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207–214. https://doi.org/10.1038/nature11234.
7. Inkpen SA, Douglas GM, Brunet TDP, Leuschen K, Doolittle WF, Langille MGI. 2017. The coupling of taxonomy and function in microbiomes. Biol Philos. https://doi.org/10.1007/s10539-017-9602-2.
8. 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.
9. Lamoureux EV, Grandy SA, Langille MGI. 2017. Moderate exercise has limited but distinguishable effects on the mouse microbiome. MSystems 2:e00006-17. https://doi.org/10.1128/mSystems.00006-17.
10. Maldonado-Gómez MX, Martínez I, Bottacini F, O'Callaghan A, Ventura M, van Sinderen D, Hillmann B, Vangay P, Knights D, Hutkins RW, Walter J. 2016. Stable engraftment of Bifidobacterium longum AH1206 in the human gut depends on individualized features of the resident microbiome. Cell Host Microbe 20:515–526. https://doi.org/10.1016/j.chom.2016.09.001.
11. Xu Z, Malmer D, Langille MGI, Way SF, Knight R. 2014. Which is more important for classifying microbial communities: who's there or what they can do? ISME J 8:2357–2359. https://doi.org/10.1038/ismej.2014.157.
12. Douglas GM, Hansen R, Jones CMA, Dunn KA, Comeau AM, Bielawski JP, Tayler R, El-Omar EM, Russell RK, Hold GL, Langille MGI, Van Limbergen J. 2018. Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. Microbiome 6:13. https://doi.org/10.1186/s40168-018-0398-3.
13. Manor O, Borenstein E. 2017. Systematic characterization and analysis of the taxonomic drivers of functional shifts in the human microbiome. Cell Host Microbe 21:254–267. https://doi.org/10.1016/j.chom.2016.12.014.
14. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, Brady A, Creasy HH, McCracken C, Giglio MG, McDonald D, Franzosa EA, Knight R, White O, Huttenhower C. 2017. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 550:61–66. https://doi.org/10.1038/nature23889.
15. Lozupone CA, Stombaugh J, Gonzalez A, Ackermann G, Wendel D, Vázquez-Baeza Y, Jansson JK, Gordon JI, Knight R. 2013. Meta-analyses of studies of the human microbiota. Genome Res 23:1704–1714. https://doi.org/10.1101/gr.151803.112.
16. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. 2017. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun 8:1784. https://doi.org/10.1038/s41467-017-01973-8.
17. Sze MA, Schloss PD. 2016. Looking for a signal in the noise: revisiting obesity and the microbiome. MBio 7:e01018-16. https://doi.org/10.1128/mBio.01018-16.
18. Langille MGI, Ravel J, Fricke WF. 2018. "Available upon request": not good enough for microbiome data! Microbiome 6:8. https://doi.org/10.1186/s40168-017-0394-z.
19. Comeau AM, Douglas GM, Langille MGI. 2017. Microbiome helper: a custom and streamlined workflow for microbiome research. mSystems 2:e00127-16. https://doi.org/10.1128/mSystems.00127-16.
20. Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, Tramontano M, Driessen M, Hercog R, Jung FE, Kultima JR, Hayward MR, Coelho LP, Allen-Vercoe E, Bertrand L, Blaut M, Brown JRM, Carton T, Cools-Portier S, Daigneault M, Derrien M, Druesne A, de Vos WM, Finlay BB, Flint HJ, Guarner F, Hattori M, Heilig H, Luna RA, van Hylckama Vlieg J, Junick J, Klymiuk I, Langella P, Le Chatelier E, Mai V, Manichanh C, Martin JC, Mery C, Morita H, O'Toole PW, Orvain C, Patil KR, Penders J, Persson S, Pons N, Popova M, Salonen A, Saulnier D, et al. 2017. Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol 35:1069–1076. https://doi.org/10.1038/nbt.3960.
21. Gibbons SM, Duvallet C, Alm EJ. 2017. Correcting for batch effects in case-control microbiome studies. bioRxiv https://www.biorxiv.org/content/early/2017/07/24/165910.