Published in final edited form as:
Cell. 2014 October 9; 159(2): 227–230. doi:10.1016/j.cell.2014.09.022.

Advancing the Microbiome Research Community

Curtis Huttenhower[1,2,*,†], **Rob Knight**[3,4,5,*], **C. Titus Brown**[6], **J. Gregory Caporaso**[7,8], **Jose C. Clemente**[9,10], **Dirk Gevers**[4], **Eric A. Franzosa**[1], **Scott T. Kelley**[11], **Dan Knights**[12,13], **Ruth E. Ley**[14], **Anup Mahurkar**[15], **Jacques Ravel**[15,16], **The Scientists for Advancement of Microbiome Research**[17], and **Owen White**[15,18,*]

[1]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

[2]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

[3]Department of Chemistry & Biochemistry, University of Colorado, Boulder, CO 80309, USA

[4]BioFrontiers Institute, University of Colorado, Boulder, CO 80309, USA

[5]HHMI, University of Colorado, Boulder, CO 80309, USA

[6]Department of Microbiology and Molecular Genetics, and Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA

[7]Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ 86001, USA

[8]Institute for Genomics and Systems Biology, Argonne National Laboratory, Argonne, IL 60439 USA

[9]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

[10]Immunology Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

[11]Department of Biology, San Diego State University, San Diego, CA 92182, USA

[12]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455, USA

[13]Biotechnology Institute, University of Minnesota, Saint Paul, MN 55108, USA

[14]Department of Microbiology, Cornell University, Ithaca, NY 14853, USA

[15]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA

[16]Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, MD 21201, USA

[18]Center for Health-Related Informatics and Bioimaging, University of Maryland School of Medicine, Baltimore, MD 21201, USA

[*]Correspondence to: owhite@som.umaryland.edu.
[17]See http://microbiomescientists.org/.
[†]These authors contributed equally to the manuscript.

Abstract

The human microbiome has become a recognized factor in promoting and maintaining health. We outline opportunities in interdisciplinary research, analytical rigor, standardization, and policy development for this relatively new and rapidly developing field. Advances in these aspects of the research community may in turn advance our understanding of human microbiome biology.

It is now widely recognized that disturbances in our normal microbial populations may be linked to acute infections such as *Clostridium difficile* and to chronic diseases such as heart disease, cancer, obesity, and autoimmune disorders (Clemente et al., 2012). This has prompted substantial interest in the microbiome from both basic and clinical perspectives. Although our genome is relatively static throughout life, each of our microbial communities changes profoundly from infancy through adulthood, continuing to adapt through ongoing exposures to diet, drugs and environment. Understanding the microbiome and its dynamic nature may be critical for diagnostics and, eventually, interventions based on the microbiome itself. However, several important challenges limit the ability of researchers to enter the microbiome field and/or conduct research most effectively.

FUNDAMENTAL CHALLENGES

Many microbiome studies to date have focused on finding patterns, and moving towards mechanism remains a major challenge. Once the “natural history” is better characterized (research to date has focused on a few locations in the Western world leaving much to described), the next step is to test for causality: when cases and controls differ, does the microbiome cause the phenotypic change, does the phenotype drive a change in the microbiome, or are there feedback loops between the two? Determining which factors in a complex ecosystem are most associated with important differences is necessary for the development of diagnostic or therapeutic strategies. For example, is the species membership, gene functional profile, transcript or protein expression, metabolite profile, or a combination thereof indicated in a particular condition? In this context, study designs that allow causal inference, such as prospective longitudinal studies and randomized, controlled experimental designs are crucial.

Current microbiome studies tend to take either top--down or bottom--up perspectives. The former constitutes ecological or systems--level investigations of entire microbial communities, while the latter focus on mechanistic examinations of the effects of individual microbes, genes, or metabolites. For example, observations of whole--community changes associated with obesity are now quite robust (Ley et al., 2005). The latter focuses on a more detailed level, where several representative studies have been very successful in identifying microbial effects in drug responses, such as the role of specific strains of the gut Actinobacterium *Eggerthella lenta* in inactivating the cardiac glycoside drug digoxin (Haiser et al., 2013) and of p-cresol production by certain gut bacteria interfering with host detoxification of acetaminophen (Clayton et al., 2009). The dynamic nature of the microbiome thus requires scientific approaches that incorporate aspects both of genetics and of functional molecular studies into the experimental design. For example, integration of ecology with molecular mechanism has identified gut microbial metabolism as a potential

impediment in the use of therapeutic food for treatment of severe malnutrition in Malawi (Smith et al., 2013), for example. Connecting top--down and bottom--up strategies to determine specific mechanism as well as patterns of association is thus a key goal for the field moving forward.

ASSAYING AND UNDERSTANDING THE MICROBIOME

Studies of the microbiome share, and in some cases magnify, hurdles common to many current ‘omics fields. The cost of sequencing is dropping much faster than the cost of analysis, creating a bottleneck in computation. Improved algorithms, increased personnel trained in analysis of microbiome data, and access to free or inexpensive computing power such as cloud-based resources would all help. Other technical challenges are unique to the study of microbial communities. For example, because of the remarkable variation in the microbiome between body sites, ages, locations, lifestyles, diets, and host genetics, our definitions of “baseline” must continue to be expanded to survey the worldwide microbiome in health and its perturbations in disease. This is true for all microbial components: viruses, phage, eukaryotes, and archaea, as well as bacteria.

Neither the data generation platforms nor the analysis methods used with the microbiome have yet reached the level of refinement necessary for translational applications and systematic meta--analyses, as has been achieved in other ‘omics areas such as gene expression or genetics over many years of study. Unfortunately, there are not as yet uniform standards for how data are deposited and how experiments are described. Data centralization efforts such as the NCBI Short Read Archive (SRA), database of Genotypes and Phenotypes (dbGaP), and BioSample must balance extremely broad accessibility - being all things to all people - with the practical concerns necessary to easily deposit and retrieve individual studies’ files. The diversity and lack of standards in human microbiome research has resulted in little consistency in how data are deposited in these repositories, and many incompatible file formats and conventions are currently in use. Consequently, it is very difficult to reconcile data from different studies, even when the same phenotypes are available. At the level of sequences deposited within these resources, field--specific considerations such as barcoding and primers are not a part of the overall repository design and may not be described well in metadata, leading to considerable challenges in interpretation. The dataset resulting from the Human Microbiome Project (HMP) is one of the largest such examples to date, where the automated deposition pipelines of multiple sequencing centers resulted in a variety of files, some from re-sequencing of the same sample and some containing as few as a single read after human read filtering (Human Microbiome Project Consortium, 2012). For scientists who want to use such data products for downstream research, even large datasets from individual projects thus pose a major data integration challenge.

The microbiome of each subject exists in a demographic, environmental and clinical context; the more precise the definition of clinical phenotypes and natural history, the greater the analytic potential, particular in comparisons between studies. Comparison of data and metadata between human microbiome studies is also susceptible to batch effects (where samples processed at the same time appear to be different due to technical variation) and

other technical challenges. Precise descriptions of phenotype, reproducible study designs, and standardizing sampling techniques are thus important for assessing variability due to technical effects, sampling bias, and other factors.

Standardization of phenotype and sample processing is of critical importance, but the development of controlled vocabularies (and tools to applying controlled vocabularies) is not complete. One ontology used with the microbiome, EnvO, was originally developed for environmental microbial communities and only partially resolves these problems (Hirschman et al., 2008); synonyms or near- synonyms are common, such as “stool”, “feces”, “faeces”, “gut”, etc. Likewise the gut has been annotated as a “Moist Tropical Environment” in some datasets, but this is likely not the intended biome description. Documentation supporting the use of the MIxS standard for the human microbiome community and improved user interfaces for tools that allow annotation and deposition of standards--compliant data could resolve a major bottleneck in current studies (Yilmaz et al., 2011). Similarly, the PhenX project, which identifies a common set of phenotype variables that are useful across many studies (Pan et al., 2012), provides a model for how microbiome metadata could be annotated. Adherence to the PhenX standard and to obtaining BioSample identifiers that are stable across multiple analyses of the same specimen will be especially useful for complex multi’omic studies (Barrett et al., 2012), as well as for systematic meta--analysis of datasets where statistical power is limited due to small population sizes in individual studies. This is especially important if microbiome data are to become more rapidly applicable in clinical settings and in large--scale epidemiological studies.

HUMAN STUDIES ISSUES AND POTENTIAL SOLUTIONS

Particularly in the United States, many opportunities exist to streamline microbiome research efforts among institutions, at the national level, and for international collaborations. For example, there are significant duplications of effort and inconsistencies resulting as individual microbiome researchers consult with their local IRBs (Institutional Review Boards) or other ethics committees, in part because microbiome studies are so new and do not exactly fit the model of either human genetics or microbiological research. This is particularly true for fecal microbiota transplantation, which has been increasingly implemented into clinical practice with neither clear regulatory guidelines nor a transparent facilitation of the associated research opportunities for making causal connections between the microbiota and host physiology. Efforts initiated by the NIH’s Clinical and Translational Science Award program, such as IRBShare, may be particularly applicable to the microbiome to increase communication and sharing of best practices between IRBs in multi-site studies. Registries designed to simplify recruiting clinical research volunteers are now common and provide the added benefit of linking diverse projects across a national research network (Richesson and Vehik, 2010). As a research community, we should consider systems such as these to streamline subject recruitment, because they have been shown to increase study enrollment and lower costs. Methods used in combination with automated eligibility screening to identify clinical participants could also be employed to simplify recruitment (Beauharnais et al., 2012; Pressler et al., 2012).

Privacy concerns unique to human-associated microbial communities introduce another challenge for microbiome research. For example, in the HMP, the identifier of the sequencing machine generating each dataset was access--restricted, because this seemingly--innocuous information could be associated with a sequencing center and thus the location of the donor individual (although *de minimis* risk guidelines have since been developed (Rhodes et al., 2011)). Although consistency and streamlining of IRBs is an ongoing effort in many fields, there is as yet little understanding of the subject protection ramifications of releasing individual sets of host--associated microbial sequences. For many subjects enrolled under earlier protocols, it is often possible to release only aggregate data, not detailed clinical information that could theoretically be combined with ‘omics to allow the identification of individual subjects. dbGAP, the protected--access database for sensitive biomedical information (Mailman et al., 2007), plays an important role but can be cumbersome to work with due to regulatory compliance and implementation complexity. The generation of large, free, and open IRB--approved high--dimensional datasets will lead to substantial advances across the board, both in the human microbiome and other areas of modern genomic medicine.

Human microbiome studies are not unique to any one NIH institute or center (IC), and they are currently supported by over a dozen ICs. This diversity in research initiatives is exciting, but cultural differences between ICs with respect to data sharing, accessibility and patient confidentiality are a concern. A recently formed Trans--NIH Microbiome Working Group is expected to be especially valuable in harmonized policy development between ICs, as well as identifying opportunities to that broaden access to data and increase reusability of results. Additional instruction to federal grant review committees on the interpretation and benefit of ‘omics approaches that complement traditional genetic approaches would also help advance microbiome research, as would dedicated study sections with members that span the broad range of expertise required to adequately assess such studies. Negotiating interoperability within and across the NIH and other federal agencies will have a disproportionately large and positive effect on microbiome research because it will eliminate the need for a large number of pairwise negotiations on a case--by--case basis.

PROSPECTS FOR THE FUTURE

Despite all the challenges, there is immense potential for microbiome research. Significant gains will be achieved with modest investments in training, improved submission tools, increased metadata utilization, and resources such as standardized reagents, protocol registries or reference datasets. Online tutorials with example data, webinars, virtual machines and packaged software encapsulating data and methods for reproducibility, and public computing environments such as the DIAG [http://diagcomputing.org] - which is specifically designed for data-rich tasks such as those encountered in metagenomics - will all play important roles. Experimental design guidelines, adequate power calculations, and basic improvements to data submission tools are critical - yet very difficult to achieve in the current funding climate - we must facilitate communication within the human microbiome research community to overcome this. When we do, we will make it much easier for investigators at all levels to enter the field, and to propagate standards and best practices within the field.

Thus while diverse microbial communities inhabit many locations of our bodies - and appear to be associated with a spectrum of diseases - it will be the organization of our communities of researchers and funding agency program managers that will ultimately improve human health. Practically speaking, standardization at every level will enhance the application of both top--down and bottom--up microbiome research. We believe that if the recommendations we propose are implemented, the field will simultaneously be in a position to make efficient use of existing resources, to consistently design, execute, and share new study results, and to realize the full potential of improved outcomes for a broad range of human diseases.

Acknowledgments

We thank all members of the Human Microbiome Consortium, International Human Microbiome Consortium, International Nucleotide Sequence Databases, the Genome Standards Consortium, and the NIH Office of the Director Roadmap Initiative. This activity was supported in part by National Institutes of Health grants U01HG004866 (OW), R01HG005969 and U54DE023798 (CH), U54HG003067 (DG), and U19AI08404, UH2AI083264, RO1AI089878, RO1GM103604 and RO1NR014826 (JR), and the Howard Hughes Medical Institute (RK), and by Army Research Office grant W911NF-11-1-0473 (CH).

References

Human Microbiome Project Consortium. A framework for human microbiome research. Nature. 2012; 486:215–221. [PubMed: 22699610]

Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, Kimelman M, Pruitt KD, Resenchuk S, Tatusova T, et al. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic Acids Res. 2012; 40:D57–63. [PubMed: 22139929]

Beauharnais CC, Larkin ME, Zai AH, Boykin EC, Luttrell J, Wexler DJ. Efficacy and cost-effectiveness of an automated screening algorithm in an inpatient clinical trial. Clinical trials. 2012; 9:198–203. [PubMed: 22308560]

Clayton TA, Baker D, Lindon JC, Everett JR, Nicholson JK. Pharmacometabonomic identification of a significant host-microbiome metabolic interaction affecting human drug metabolism. Proc Natl Acad Sci U S A. 2009; 106:14728–14733. [PubMed: 19667173]

Clemente JC, Ursell LK, Parfrey LW, Knight R. The impact of the gut microbiota on human health: an integrative view. Cell. 2012; 148:1258–1270. [PubMed: 22424233]

Haiser HJ, Gootenberg DB, Chatman K, Sirasani G, Balskus EP, Turnbaugh PJ. Predicting and manipulating cardiac drug inactivation by the human gut bacterium Eggerthella lenta. Science. 2013; 341:295–298. [PubMed: 23869020]

Hirschman L, Clark C, Cohen KB, Mardis S, Luciano J, Kottmann R, Cole J, Markowitz V, Kyrpides N, Morrison N, et al. Habitat-Lite: a GSC case study based on free text terms for environmental metadata. OMICS. 2008; 12:129–136. [PubMed: 18416669]

Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A. 2005; 102:11070–11075. [PubMed: 16033867]

Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet. 2007; 39:1181–1186. [PubMed: 17898773]

Pan H, Tryka KA, Vreeman DJ, Huggins W, Phillips MJ, Mehta JP, Phillips JH, McDonald CJ, Junkins HA, Ramos EM, et al. Using PhenX measures to identify opportunities for cross-study analysis. Hum Mutat. 2012; 33:849–857. [PubMed: 22415805]

Pressler TR, Yen PY, Ding J, Liu J, Embi PJ, Payne PR. Computational challenges and human factors influencing the design and use of clinical research participant eligibility pre-screening tools. BMC medical informatics and decision making. 2012; 12:47. [PubMed: 22646313]

Rhodes R, Azzouni J, Baumrin SB, Benkov K, Blaser MJ, Brenner B, Dauben JW, Earle WJ, Frank L, Gligorov N, et al. De minimis risk: a proposal for a new category of research risk. Am J Bioeth. 2011; 11:1–7. [PubMed: 22047112]

Richesson R, Vehik K. Patient registries: utility, validity and inference. Adv Exp Med Biol. 2010; 686:87–104. [PubMed: 20824441]

Smith MI, Yatsunenko T, Manary MJ, Trehan I, Mkakosya R, Cheng J, Kau AL, Rich SS, Concannon P, Mychaleckyj JC, et al. Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. Science. 2013; 339:548–554. [PubMed: 23363771]

Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol. 2011; 29:415–420. [PubMed: 21552244]