Annals of Epidemiology 26 (2016) 301–305

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

The Microbiome and Epidemiology

The importance of the microbiome in epidemiologic research

Blake M. Hanson MS, PhD *, George M. Weinstock PhD

The Jackson Laboratory for Genomic Medicine, Farmington, CT

ARTICLE INFO

Article history:
Received 11 January 2016
Accepted 23 March 2016
Available online 7 April 2016

Keywords:
Microbiota
Microbial Consortia
Epidemiology

ABSTRACT

Purpose: The human microbiome is the community of microorganisms that live on and in the body. Currently, most applications of microbiome analysis derive from the perspective of discovery and characterization. The completion of the NIH Human Microbiome and the European MetaHIT projects will change the focus to studying the role of the microbiome on human health and disease.
Methods: Recent developments in technology and bioinformatics have afforded an opportunity to explore more fully the importance of community structure, detection of pathogens, and community interactions. The current state of microbiome research in terms of effect size, power calculations, how stratification on community classes can increase this power, and the importance of study design and power in reproducibility is reviewed.
Results: Work is needed to characterize microbiome development, ecological stability, and variation. Development and implementation of variance stabilization techniques should replace rarefaction of data, which reduces study power, in future research.
Conclusions: Epidemiologists have most of the tools necessary to explore the relationship between the microbiome and human health. Further development of tools for large-scale multivariate data sets will be helpful. Applying the methods of epidemiology will be critical in translating research results to preventive interventions and population health.

The term “microbiome” was originally coined by Joshua Lederberg and represents “the ecological community of commensal, symbiotic and pathogenic microorganisms that literally share our body space” [1,2]. This microbial community is composed of roughly 100 trillion microbial cells [3], which is between three and 10 times the number of human cells [4,5]. Lederberg put a strong emphasis on viewing the microbiome not as a distinct element, dividing the microbial cells and *Homo sapiens* cells, but as a superorganism composed of all cells present, both microbial and *H. sapiens* [1]. Currently, there are an estimated 19,797 protein-encoding genes in the human genome according to GENCODE release 23 [6,7], and there are 536,112 unique microbial genes within an average individual’s gut [8] resulting in roughly 27 times as many microbial genes present as human genes in this tissue alone. Of course, microbes are not limited to the gut but are present on every body surface that comes into contact with the environment [9–13] and fulfill a number of metabolic roles unfulfilled by human cells [14]. This superorganism can be investigated using a modification of the conceptual epidemiologic framework of the epidemiologic triangle where the microbiome is added in the form of a fourth node [15]. Traditionally, the epidemiologic triangle has been used to organize the relationship between three potential causal determinants: the host, the agent, and the environment. The microbiome is impacted by and impacts all three of these determinants and must be considered not as a component of one these three nodes, but as its own distinct node (Fig. 1).

This microbiome has historically been outside the reach of in-depth scientific inquiry because culture-based methods are not sufficient for sampling communities of hundreds of different taxa present at a range of abundances. With the advent of sampling by Sanger DNA sequencing of the ensemble community (“culture-independent” sampling), the microbiome began to come into focus. However, recent technological developments (“next-generation” sequencing) provided much deeper and cost-effective sampling. A typical microbiome data point from a human sample (stool, saliva, vaginal swab, skin scrape, nasal lavage, etc) is a set of from thousands to hundreds of millions of sequences, assigned to a list of hundreds of taxa or thousands of genes to provide abundances for each. Such data points are typically produced from tens of samples, with larger studies up to hundreds of samples. This with the accompaniment of appropriate bioinformatics has opened up new research avenues for analysis of the microbiome, such as details of

* Corresponding author. The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT 06030. Tel.: +1-860-837-2416; fax: +1-860-837-2398.
E-mail address: blake.hanson@jax.org (B.M. Hanson).

http://dx.doi.org/10.1016/j.annepidem.2016.03.008
1047-2797/

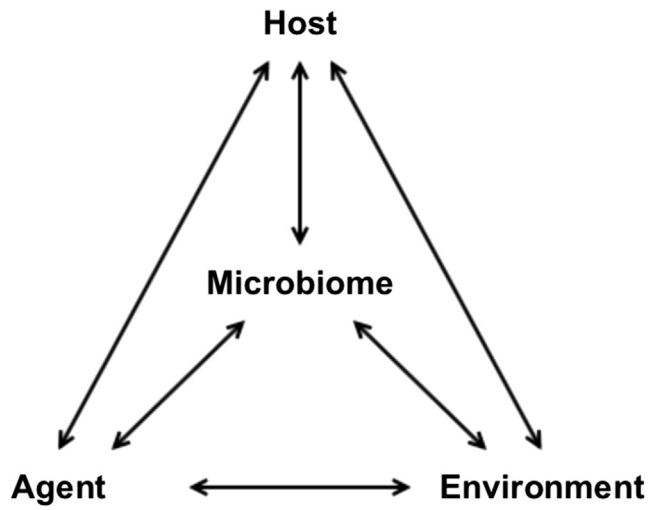

Fig. 1. An epidemiologic triangle incorporating the microbiome as a fourth important and distinct node. Adapted from Foxman and Rosenthal, 2013 [15].

community structure, detection of pathogens existing within the microbiome and their associated virulence mechanisms, as well as community interactions, such as commensalism, mutualism, and amensalism.

With the completion of the first NIH Human Microbiome Project (HMP; [13]) and the European MetaHIT project [8], the focus has turned from characterization and cataloging of the microbiome to investigating the interaction of the microbiome and human health. Human microbiome biomarkers have potential clinical utility for predicting disease risk, identifying disease onset, predicting treatment response, determining treatment success, guiding preventative measures, and developing therapeutics based on microbiome manipulation. However, to date, most applications of microbiome analysis to clinical situations have had low diagnostic accuracy, so the results are primarily useful from a discovery perspective but not for general clinical use. This is in part due to a need for statistical tools and experimental designs to move the microbiome from the laboratory to the clinic. As concluded in a recent Annual Review of Statistics and Its Application article: "unique characteristics of the (microbiome) data produced by the new technologies, as well as the sheer magnitude of these data, make drawing valid biological inferences from microbiome studies difficult [16]. Analysis of these big data poses great statistical and computational challenges." The microbiome data and analysis are all part of the general trend in genomic big data. As summarized by a National Biomarker Development Alliance report: "new 'omics' technologies are generating data with rapidly escalating volume, velocity, and variety, and data analysis usually requires complex deconvolution of low signal-to-noise signatures [17]. In addition, the large-scale molecular data sets that these new technologies generate differ fundamentally from traditional biological and clinical data sets. Traditional clinical and epidemiological data sets comprise a small number of variables tracked across a proportionately larger number of samples, while today's 'omics' technologies are measuring variables per sample whose numbers far exceed the typical number of samples. Together, these factors create a need for new data analytics and infrastructure that are only now being developed …"

Epidemiologists have the tools to investigate and elucidate the relationship between the microbiome and human health, but it will require consideration of the microbiome not just as an external entity but also as a component of our self. Epidemiologists often analyze large, integrated data sets that contain different types of data and are therefore well equipped to work with this type of data. As microbiome development, ecological stability, and variation become better understood, epidemiologists are well poised to further characterize the microbiome and aid in the translation of research results to interventions and population health.

We will present a few key topics in the current state of microbiome research, focusing on the assessments of power calculation, how stratification can increase this power, and the importance of study design selection.

Effect size

To develop power calculation tools, one considers what an effect size is in microbiome studies. This can be addressed in two ways: how much of a difference between the microbiome does one need to differentiate two groups, and how much of an effect size is biologically relevant. In these analyses, lists of taxa can be compared at different phylogenetic levels (phylum to species) where the number of categories increases the lower taxonomic level, or list of genes can be compared as single genes or grouped into pathways also reducing categories. In addition, we can consider more aggregate measures such as beta-diversity [18]—an assessment of the similarity between two populations or samples—or measures with finer resolution such as changes in specific organisms or operational taxonomic units. When considering effect size, identifying the number of individuals for each group and the number of sequences that need to be generated for each sample are essential [19,20].

Power calculation

In characterizing the microbiome and generating new hypotheses, many studies compare the microbial communities of groups with different exposures or interventions. When designing these studies, special attention must be paid to statistical power—the ability to detect an expected effect and reject the null hypothesis—and recruiting enough participants to achieve this power.

One way to represent the microbial community structure and compare them between individuals and groups is through the use of distance matrices. One such method uses UniFrac [21,22] or Jaccard [23] pairwise distance matrices for permutational multivariate analysis of variance (PERMANOVA) using distance matrices [24]. Statistical power of PERMANOVA relies on the number of exposure or intervention groups, the number of subjects in each group, the size of the effect, and the distances between subjects within each group [25]. In addition, because PERMANOVA uses a pseudo-*F* ratio, the power estimation techniques for parametric ANOVA are not applicable to PERMANOVA [24]. Kelly et al. recently published a novel power and sample size calculation tool, implemented in the R programming language [26] as the *micropower*package that allows researchers to simulate and model different effect sizes within microbiome composition [25] (Table 1).

Another representation of a microbial community is to consider it as a whole assemblage of distinct microbes that can be assessed using multivariate tools. In addition to assessments of microbial diversity, multivariate analysis tools allow researchers to investigate the effects of multiple exposures or interventions and the association with individual members of the microbial community while still considering the other members of the microbial community [27]. Many nonparametric multivariate tests have the limitation, however, that either the size of the effect cannot be quantified, or that the dispersion of specific taxa within a set of samples is consistent. The Dirichlet-multinomial distribution model provides adjustments over a general multinomial distribution that limit type I error (rejecting the null hypothesis incorrectly) due to overdispersion in specific taxa frequencies [27]. La Rosa et al. have published a power, sample size, and parameter estimation