Annals of Epidemiology 26 (2016) 311–321

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

The Microbiome and Epidemiology

Intricacies of assessing the human microbiome in epidemiologic studies

Courtney K. Robinson MS [a], Rebecca M. Brotman PhD, MPH [a,b,*], Jacques Ravel PhD [a,c,**]

[a] Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore
[b] Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore
[c] Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore

ARTICLE INFO

Article history:
Received 4 December 2015
Accepted 5 April 2016
Available online 12 April 2016

Keywords:
16S ribosomal RNA gene
Metagenomics
Microbiota

ABSTRACT

Purpose: In the past decade, remarkable relationships have been documented between dysbiosis of the human microbiota and adverse health outcomes. This review seeks to highlight some of the challenges and pitfalls that may be encountered during all stages of microbiota research, from study design and sample collection, to nucleic acid extraction and sequencing, and bioinformatic and statistical analysis.
Methods: Literature focused on human microbiota research was reviewed and summarized.
Results: Although most studies have focused on surveying the composition of the microbiota, fewer have explored the causal roles of these bacteria, archaea, viruses, and fungi in affecting disease states. Microbiome research is in its relatively early years and many aspects remain challenging, including the complexity and personalized aspects of microbial communities, the influence of exogenous and often confounding factors, the need to apply fundamental principles of ecology and epidemiology, the necessity for new software tools, and the rapidly evolving genomic, technological, and analytical landscapes.
Conclusions: Incorporating human microbiome research in large epidemiologic studies will soon help us unravel the intricate relationships that we have with our microbial partners and provide interventional opportunities to improve human health.

Introduction

It is believed that in the human body, microorganisms are more numerous than human somatic and germ cells [1]. Together, the genomes of these microbial mutualists (collectively defined as the metagenome) provide traits and services to humans, and in some cases, are associated with disease pathogenesis [2]. Over the past 10 years, with the advent of high-throughput sequencing technologies, there has been an exponential increase in molecular studies of the human microbiome. Rather than relying solely on bacterial cultivation for identification, partial sequencing of the bacterial 16S rRNA gene has become the standard in cataloging organisms in biological samples.

If humans are thought of as a composite of microbial and human cells, and the human genetic landscape as an aggregate of the human genome, the microbiota (bacteria, archaea, and lower eukaryotes), and the virome (the collective set of bacteriophages and viruses), then the picture that emerges is one of a human "supra-organism" [3]. It therefore becomes necessary to consider human health and disease outcomes in the context of our microbial partners. Microbiology is now entering a new era where the focus moves from the properties of single organisms in isolation to the operations of whole communities. The new field of metagenomics involves the genomic characterization of the entire microbial communities and not just cultivation of single organisms.

Molecular methods for interrogating microbial communities have led to a better understanding of the organisms present at specific sites on the human body and their potential roles in human health. The respective microbiota in each body niche can influence a wide variety of health outcomes including obesity [4], brain chemistry [5], ulcerative colitis [6], gynecologic and obstetric health [7], and periodontal disease [8]. Efforts to describe a "core" human microbiome, in the hopes of providing a baseline for comparisons [9], have proven to be challenging because bacterial communities

* Corresponding author. Institute for Genome Sciences, University of Maryland School of Medicine, 801 West Baltimore Street, 6th Floor, Baltimore, MD 21201. Tel.: +1-(410)-706-6767; fax: +1-(410)-706-1482.
** Corresponding author. Institute for Genome Sciences, University of Maryland School of Medicine, 801 West Baltimore Street, 6th Floor, Baltimore, MD 21201. Tel.: +1-(410)-706-5674; fax: +1-(410)-706-1482.
E-mail addresses: rbrotman@som.umaryland.edu (R.M. Brotman), jravel@som.umaryland.edu (J. Ravel).

http://dx.doi.org/10.1016/j.annepidem.2016.04.005
1047-2797/

show high intersubject variability in species composition [10], whereas functional gene expression is more conserved [9].

With large data sets capturing many dimensions of the microbiota, including diversity, relative abundance and absolute abundance of bacterial taxa, as well as functional measurements of the microenvironment, there are tremendous opportunities for epidemiologic studies to describe the microbiota's role in transitions between healthy and disease states. To date, most studies have focused on quantifying the statistical associations between the compositions of the human microbiota with health outcomes; however, fewer have been able to document how microbial changes are part of the causal chain leading to disease [8,11].

Early studies on microbes were constrained to culture-based methods that were limited by the large numbers of species that resisted cultivation. Although cultivation of microbes has improved and the proportion of organisms not yet cultivated is rapidly decreasing, the development of molecular methods for characterizing the microbiota, including marker gene amplicon, metagenomic, and metatranscriptomic sequencing, brought about rapid access to the identification and genomic information of previously uncultivated organisms. Marker gene amplicon (mainly 16S rRNA gene) sequencing involves interrogating a single gene to identify which species are present. Combined with broad and species-specific quantitative polymerase chain reaction (PCR), this approach affords cataloging species and their abundance in biological samples. For an overview of the human microbiome and 16S rRNA gene-based analyses for characterizations of the human microbiota, as well as terminology in this rapidly evolving field, we refer the reader to an excellent review by Tyler et al. [12] as well as an editorial by Marchesi and Ravel [13].

To gain insight into the functional make up of microbial communities, metagenomic sequencing is applied by sequencing all the DNA recovered from a sample. Analyzing these reads can identify what organisms are present and the community's genomic content and functional potential. Metatranscriptomic sequencing, which surveys expressed genes in a sample, defines the function of the community at the time of sampling. These approaches could be further expanded by looking at the metaproteome [14–16] or the community metabolic outcomes, the metabolome [17–19]. Recent technological advances in high-throughput sequencing have enabled the parallel processing of large number of samples at affordable costs. As a consequence, these methodologies can now be integral to large-scale epidemiologic studies.

In this review, we seek to detail what is involved with analyses of the human microbiota from an epidemiologic perspective, with specific attention to the associated difficulties in designing, executing, and interpreting studies of the human microbiome. Figure 1 presents a sample workflow for conducting a 16S rRNA sequencing study, and while the details would differ when conducting a metagenomic or metatranscriptomic study, this flow chart highlights the issues to consider at each step of the process.

Sample collection and storage conditions

One of the first issues that arises when planning epidemiologic studies of the human microbiome is determining collection methods for the samples. Collection should recover samples that are representative of the true microbiota present at the site while limiting sampling biases and contamination. Less invasive sampling methods encourage recruitment and retention of study participants, and a pilot study can help inform and validate sampling methods. For example, recent studies on the methods for sampling the sinonasal microbiota [20] and intestinal mucosa [21] found the less invasive methods provided samples that had consistent microbiota profiles with samples obtained using classical sampling methods. In contrast, fecal transport swabs recovered less DNA and showed altered microbiota profiles compared to that of fecal material samples [22], stressing the importance of validating collection methods.

An important aspect of sampling strategy also includes sampling frequencies, which if performed in a clinical setting is often limited by the willingness of participants to return to the study site frequently as well as staffing requirements. However, participants are capable and willing to perform self-sampling at home and with high compliance rates [7,23–29], thus enabling large field-based longitudinal epidemiologic studies. Numerous groups have validated the use of self-collected samples compared to clinician-collected samples for microbiome studies and pathogen detection, as well as confirmed uniformity from repeated sampling at the same sitting [30–33]. The number of samples to be collected at each time point should also be considered. Excessive sampling can be difficult from a human subject perspective, and may in itself disturb the microenvironment thus introducing compounding biases over time, making it potentially difficult to interpret longitudinal patterns of change.

After sample collection, it is then important to take into consideration methods for sample transport and both short-term and long-term storage. Delays often occur between sampling and final storage because of logistical issues, and it is not always possible to process samples immediately after collection. Numerous studies have evaluated the effect of temperature and duration of storage on fecal samples and have found conflicting results in terms of the effect on microbiota composition based on 16S rRNA gene profiling, with some samples showing little change [22,34–37] and others showing significant differences [38,39]. Amies transport media has been a successful choice for preserving fecal [40,41], vaginal [7,31,42], and nasal [43] samples for DNA extraction and sequencing. Samples taken for transcriptomic analysis need to be stored appropriately to minimize RNA degradation, so preservation with guanidine thiocyanate is usually used to prevent nucleases from degrading RNA molecules [44]. RNA later has been used successfully for recovery of DNA and RNA from fecal samples [38,44,45] and saliva [46].

DNA/RNA extraction, 16S rRNA gene amplification, and library preparation

A critical step to microbiome analyses is DNA extraction, as in principle this is where most biases could be introduced, mostly from uneven cell lysis across the microbial community. Cell lysis, typically achieved through enzymatic and/or mechanical manipulations, would ideally work on all cell types equally, resulting in DNA being representative of the composition of the starting material. However, cells can vary in their susceptibility to lysing methods, with some lysing under fairly gentle conditions, and others, particularly gram-positive organisms or spores, needing much harsher conditions that may result in shearing of DNA from easily lysed organisms. Several studies have shown the use of mechanical lysis gives the highest bacterial diversity in 16S rRNA gene surveys [47,48], and performs particularly well in the recovery of gram-positive organisms in fecal communities [49]. Oral samples extracted using either mechanical or enzymatic lysis steps have shown overall similar microbiota profiles based on 16S rRNA gene amplicon sequencing but with higher recovery of certain taxa with either method [50]. It is therefore important to consider what types of organisms are expected in a specific sample when choosing an extraction method, and noting that no methods are inherently free of biases [48]. Similar considerations apply to RNA extraction methods.