Annals of Epidemiology 26 (2016) 360–365

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

The Microbiome and Epidemiology

Incorporating microbiota data into epidemiologic models: examples from vaginal microbiota research

Janneke H. van de Wijgert MPH, PhD [a,*], Vicky Jespers MD, MEpi, PhD [b]

[a] Department of Clinical Infection, Microbiology and Immunology, Institute of Infection and Global Health, University of Liverpool, Liverpool, UK
[b] HIV and Sexual Health Unit, Department of Public Health, Institute of Tropical Medicine, Antwerpen, Belgium

ARTICLE INFO

Article history:
Received 1 December 2015
Accepted 23 March 2016
Available online 1 April 2016

Keywords:
Microbiota
Microbiome
Vagina
Next generation sequencing
Quantitative PCR
Epidemiology
Biostatistics

ABSTRACT

Purpose: Next generation sequencing and quantitative polymerase chain reaction technologies are now widely available, and research incorporating these methods is growing exponentially. In the vaginal microbiota (VMB) field, most research to date has been descriptive. The purpose of this article is to provide an overview of different ways in which next generation sequencing and quantitative polymerase chain reaction data can be used to answer clinical epidemiologic research questions using examples from VMB research.
Methods: We reviewed relevant methodological literature and VMB articles (published between 2008 and 2015) that incorporated these methodologies.
Results: VMB data have been analyzed using ecologic methods, methods that compare the presence or relative abundance of individual taxa or community compositions between different groups of women or sampling time points, and methods that first reduce the complexity of the data into a few variables followed by the incorporation of these variables into traditional biostatistical models.
Conclusions: To make future VMB research more clinically relevant (such as studying associations between VMB compositions and clinical outcomes and the effects of interventions on the VMB), it is important that these methods are integrated with rigorous epidemiologic methods (such as appropriate study designs, sampling strategies, and adjustment for confounding).

Introduction

The improved availability and affordability of high-throughput molecular techniques is revolutionizing microbiota research [1], including vaginal microbiota (VMB) research [2]. VMB dysbiosis (also known by its clinical name bacterial vaginosis [BV]) has long been recognized as a common clinical condition with potentially devastating consequences (such as preterm birth), but its etiology and pathogenesis have never been fully understood. BV is treated empirically in most clinical settings, diagnosed by the Amsel criteria (clinical signs and microscopy) in some specialized clinics [3], and diagnosed by Gram stain Nugent scoring (microscopy) in research settings [4]. Microscopy and culture studies had already shown that the VMB of healthy asymptomatic women predominantly consists of lactobacilli, and that BV is associated with a reduction of lactobacilli and an overgrowth of other (facultative) anaerobic bacteria. However, high-throughput molecular techniques have characterized VMB compositions in much more detail, identified novel bacterial taxa in the vaginal niche, and allowed the field to get a better handle on determinants of VMB composition, VMB fluctuations over the menstrual cycle and over a lifetime, VMB associations with clinical outcomes, and the effects of interventions on the VMB [2].

While early studies between 2002 and 2013 used a variety of molecular techniques (DNA fingerprinting, DNA microarrays, quantitative polymerase chain reaction (qPCR), and sequencing of DNA isolated from culture colonies or directly from genital samples using many different sequencing platforms; [2]), studies in 2014 and 2015 almost exclusively used next generation sequencing (NGS) and/or (multiplex) qPCR of DNA extracted directly from genital samples. For that reason, we have focused this article on the latter two techniques. Furthermore, in the VMB field thus far, the vast majority of studies have targeted the 16S ribosomal DNA (rDNA) gene for bacterial identification. We therefore limited this review to NGS and qPCR of the 16S rDNA gene, but note that

* Corresponding author. Department of Clinical Infection, Microbiology and Immunology, Institute of Infection and Global Health, University of Liverpool, Ronald Ross Building, West Derby Street, Liverpool L69 7BE, UK. Tel.: +44-151-795-9613.
E-mail address: j.vandewijgert@liverpool.ac.uk (J.H. van de Wijgert).

http://dx.doi.org/10.1016/j.annepidem.2016.03.004
1047-2797/

shotgun sequencing is increasingly available and affordable and will likely increase in importance in future VMB research.

We wrote this article for epidemiologists who are interested in studying the effects of microbiota composition on clinical outcomes but are not experts in genomic laboratory methods or bioinformatics. Throughout the article, we used examples from VMB research. While the first 20 years of VMB genomics have been dominated by the development and initial applications of the technologies in relatively small, mostly descriptive studies, we believe that the time has now come for incorporation into clinical epidemiologic studies to answer biomedical research questions or test interventions on a much wider scale.

Basic NGS technical information of relevance to epidemiologists

This paragraph briefly summarizes the principles of 16S rDNA-based NGS, but more detailed explanations can be found in the Appendix. In microbiota studies, the conserved regions of the 16S rDNA gene are used for the initial amplification of 16S rDNA present in a sample, and portions of one or more variable regions are sequenced to allow for identification of bacterial species, genera, or higher order taxa (collectively referred to as taxa in this article; [5–7]). The ability to classify sequencing reads to species level depends on various factors including choice of NGS platform [8,9], variable region(s), and alignment databases (see in the following paragraphs). Most NGS platforms allow for multiplexing (the use of a unique barcode sequence to identify DNA originating from a specific sample), so that samples can be pooled during sequencing and subsequently sorted by barcode.

A multiplex 16S rDNA NGS run typically results in thousands of sequence reads per sample [8,9]. The sequence reads are first checked for quality and preprocessed, a process that is known to introduce biases (an observed microbiota composition, i.e., different from the actual microbiota composition; [10,11]). The processed reads are then used to identify bacterial taxa present in each sample by sequence alignment [12–14]. The reads are usually first assigned to operational taxonomic units (OTUs; based on a sequence similarity threshold—usually 97%—within the experimental data set), which are subsequently compared with known bacterial taxa sequences in publicly available databases [15–17]. These databases do not always allow for assignment of sequences at species level, and some laboratories have designed their own customized databases to fill the gaps (see e.g., [18]). Some researchers report phylotypes (based on sequence similarity with an external database) instead of OTUs. We will refer to OTUs in the remainder of the article, but all methods described also apply to phylotypes unless explicitly stated otherwise.

From the resulting sequence alignment, phylogenetic analysis can be conducted to assess the sequences' shared evolutionary origins. Methods for estimating phylogenies, each with their own strengths and weaknesses, include neighbor-joining, unweighted pair group method with arithmetic mean, maximum parsimony, maximum likelihood, and Bayesian inference of phylogeny (reviewed in [19]). Phylogenies are typically visualized using a dendrogram (Appendix: Fig. 1).

Rarefaction curves are used to determine whether most taxa present in a sample were in fact identified (Appendix: Fig. 2; [20]). Most articles only report on taxa that constitute at least 0.1% of the overall bacterial community; taxa constituting less than 0.1% are referred to as rare taxa. However, if a bacterial community has 10^8 bacteria per mL of biological sample, then rare taxa may represent up to 10^5 bacteria per mL. Such "rare" taxa could cause disease (e.g., if it produces toxins or has a high pathogenicity index for other reasons), play important roles in the bacterial community, or constitute a "seed bank" of taxa whose numbers increase under conditions that favor their growth [21].

The number of sequence reads per individual sample within one study can be vastly different for a number of reasons. This is usually dealt with by normalizing the data in the following ways: (1) base analyses on the relative abundance of each species; or (2) rarefy, which refers to the process of throwing away sequences from samples with high numbers of reads so that all samples have the same number of reads [22]. Although the former does not address heteroscedasticity (different species might have different variability), the latter omits potentially large amounts of available valid data. Some experts therefore object to both these options and argue in favor of a third option, which is to use negative binomial models to account for differences in read numbers between samples (for an in-depth discussion, see [22]).

Ecologic analyses

The field of microbial and/or environmental ecology existed long before human microbiota research soared, and ecologic terminology and methodology have been incorporated into human microbiota research. The term "richness" refers to the number of taxa present in an ecological community (not taking the abundance of each taxa into account), and "evenness" refers to how close in abundance these taxa are. Diversity takes both richness and evenness into account. The total diversity ("gamma diversity") consists of the diversity at one ecologic niche or in one (type of) sample ("alpha diversity") and the differentiation between ecologic niches or (types of) samples ("beta diversity"). Popular alpha diversity measures include the Shannon (also referred to as Shannon–Wiener) diversity index and (inverse) Simpson diversity index. Popular beta diversity measures include the Bray–Curtis dissimilarity (uses counts of shared and unshared OTUs between two samples), Jensen-Shannon divergence (measures the similarity between two probability distributions), and UniFrac measures (uses shared and unshared branches in a phylogenetic tree [23]). Diversity is often visualized by a heatmap showing each OTU (on the vertical axis) for each participant or sampling time point (on the horizontal axis) with the proportion of sequence reads assigned to each OTU (often referred to as the relative abundance) shown in a different color (Appendix: Fig. 1). Alternatively, an interpolated bar plot is shown, with the relative abundance on the vertical axis and the participant or time point on the horizontal axis and each OTU shown in a different color (Appendix: Fig. 3).

Using NGS data to answer biomedical research questions

After multiple samples have been sequenced and the sequencing data have been organized into OTUs for each sample, it is time to consider how these data can be used to answer biomedical research questions. We have divided this section into (1) methods that compare the presence or relative abundance of individual OTUs, or community compositions, between different groups of women or sampling time points; and (2) methods that first reduce the complexity of the data into a few variables followed by the incorporation of these variables into traditional biostatistical models. A third group of methods of potential interest, but not discussed further in this article, are bioinformatics methods such as sequence mining (which identifies statistically relevant patterns) and alignment-free sequence analysis (when alignment is not possible, e.g., because sequences are not closely related). The NGS data input for most methods in the second and third category is a distance matrix. A distance matrix in this context is a two-dimensional array containing the distances (the degree of similarity) of all pairwise sequences and/or OTUs in the data set.