**Incorporating microbiota data into epidemiological models: Examples from vaginal microbiota research**

Janneke van de Wijgert1\*, Vicky Jespers2

1. University of Liverpool, Institute of Infection and Global Health

Ronald Ross Building, West Derby Street

Liverpool L69 7BE, United Kingdom

Email: j.vandewijgert@liverpool.ac.uk

1. Institute of Tropical Medicine, HIV and Sexual Health Unit,

Nationalestraat 155

2000 Antwerpen, Belgium

Email: vjespers@itg.be

\* Corresponding author

Word counts:

Abstract: 199 words (max 200 words)

Main text: 4,217 words (max 5,000 words)

Number of appendices: 1

Number of tables and figures (in Appendix): 4

**Abstract**

**Purpose:** Next generation sequencing (NGS) and quantitative polymerase chain reaction (qPCR) 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 paper is to provide an overview of different ways in which NGS and qPCR data can be used to answer clinical epidemiological research questions using examples from VMB research.

**Methods:**  We reviewed relevant methodological literature as well as VMB papers (published between 2008 and 2015) that incorporated these methodologies.

**Results:** VMB data have been analysed 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:** In order 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 epidemiological methods (such as appropriate study designs, sampling strategies and adjustment for confounding).

**Keywords**

Microbiota, microbiome, vagina, next generation sequencing, quantitative PCR, epidemiology, biostatistics

**List of abbreviations**

ALDEx ANOVA-like differential expression

ANOVA analysis of variance

BV bacterial vaginosis

CCA canonical correlation analysis

GEE generalized estimating equations

Geq gene equivalents

LDA linear discriminant analysis

LEfSe linear discriminant analysis effect size method

MSA multiple sequence alignment

NCBI National Center for Biotechnology Information

NGS next generation sequencing

NMDS non-metric multidimensional scaling

OTUs operational taxonomic units

PC principal component

PCA principal components analysis

PCoA principal coordinates analysis

QIIME quantitative insights into microbial ecology

qPCR quantitative polymerase chain reaction

rRNA ribosomal RNA

STI sexually transmitted infection

UPGMA unweighted pair group method with arithmetic mean

VMB vaginal microbiota

**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 pre-term 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 employed 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 employed 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 shotgun sequencing is increasingly available and affordable and will likely increase in importance in future VMB research.

We wrote this paper 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 paper, 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 epidemiological 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 manuscript) [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 below). 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 that is 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 dataset), which are subsequently compared to 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 for example [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 manuscript 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 (UPGMA), maximum parsimony, maximum likelihood, and Bayesian inference of phylogeny (reviewed in [19]). Phylogenies are typically visualized using a dendrogram (Appendix: Figure 1).

Rarefaction curves are used to determine whether most taxa present in a sample were in fact identified (Appendix: Figure 2) [20]. Most papers 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 108 bacteria per mL of biological sample, then rare taxa may represent up to 105 bacteria per mL. Such ‘rare’ taxa could cause disease (for example, 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]. While 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 of 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/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: Figure 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: Figure 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 paper, are bioinformatics methods such as sequence mining (which identifies statistically relevant patterns) and alignment-free sequence analysis (when alignment is not possible, for example 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/OTUs in the dataset.

***Comparing the relative abundances of individual OTUs:*** The first step is usually to determine the relative abundances of OTUs of interest. In the VMB field, the focus is often on the relative abundance of the *Lactobacillus* genus, specific *Lactobacillus* species, and/or the most common BV-associated bacteria (*Gardnerella vaginalis*, *Atopobium vaginae*, among others) because of their known association with vaginal health. It is important to note that the relative abundances derived from NGS data are semi-quantitative and can be biased by laboratory and sequence processing methods used as discussed earlier. If answering the research question requires quantitative data, and the bacterial taxa of interest are known a priori, qPCR assays of those predefined taxa are likely to be more appropriate (see below). The mean or median relative abundance of an OTU of interest for multiple samples from a group of study participants (for example, HIV-positive versus HIV-negative women), and/or samples collected at a specific time point (for example, pre- and post-treatment), can also be calculated. Traditional biostatistical methods can then be used to compare these mean/median relative abundances of individual taxa between groups or sampling time points, or determine correlations or associations with other variables of interest (such as Gram stain Nugent scores, the presence of clinical signs and symptoms or behaviors; see for example [24-26]). This approach works well when focusing the analysis on just a few taxa but does not make optimal use of all the available data related to entire bacterial communities.

***Comparing community composition using diversity measures:*** A relatively easy way to compare community compositions is to compare alpha or beta diversity measures (see for example [27]). One should keep in mind, however, that diversity measures only provide information about richness and evenness of multiple taxa and not about which taxa are present. Whether this is clinically meaningful depends on the ecological niche and the research question. After a decade of molecular VMB studies, we now know that a healthy VMB is dominated by lactobacilli and therefore has low diversity, whereas BV is always associated with high diversity [2]; diversity measures can therefore be clinically meaningful in this context. In fact, some studies have shown ‘dose response’ relationships between VMB diversity and clinically relevant outcomes such as prevalence of sexually transmitted infections, shedding of HIV in the female genital tract, and vaginal mucosal degradation and inflammation [28, 29]. However, caution is needed because some organisms that can cause symptomatic vaginitis or adverse outcomes can be present in a low diverse *Lactobacillus*-dominated VMB (such as *Candida* yeasts, which are not measured by 16S techniques, and *Streptococcus agalactiae*, which is often present in low relative abundance).

The alpha diversity Shannon and reverse Simpson indices are continuous variables with a lower bound of 0 and no upper bound (with a higher value indicating increased diversity). A mean or median diversity for a group can therefore be determined, and groups can be compared using traditional biostatistical methods such as analysis of variance (ANOVA) or correlation methods. Beta diversity dissimilarity or distance measures (referred to as distance measures in the remainder of the manuscript) have a value between 0 and 1, with a score of 0 indicating that the two sequence profiles are exactly the same in bacterial presence and abundance, whereas a score of 1 represents profiles with no overlapping sequences. The distance measures themselves are an indication of the degree of similarity between two samples, but many research questions involve the comparison of groups containing multiple samples. This requires the compilation of a distance matrix including the distance measures of all possible sample pairs, which can subsequently be used for clustering or other dimensionality reduction methods, as described below.

***Identifying taxa (or other biomarkers) responsible for community composition differences:***

Indicator value analysis is an ecologic analysis method with an emphasis on identifying taxa responsible for community composition differences [30]. Indicator value analysis quantifies the fidelity (the proportion of samples in a group that contain the taxa) and specificity (relative abundance) of taxa in each group of samples in a user-specified classification of these groups, and tests for the statistical significance of the associations by permutation. It was used by Hickey and colleagues to determine if individual taxa were more strongly associated with the vaginal or the vulval environment [31].

The main aim of most metagenomic studies is to identify organisms, genes, or pathways that consistently explain the differences between two or more microbial communities. All of the methods described above, derived from either biostatistics or microbial ecology, have important limitations in that regard: they do not make use of all of the available data, and do not adequately address the multidimensionality of the data, the considerable experimental variance embedded in NGS procedures, the considerable inter-subject variability, and adjustment for multiple comparisons. Therefore, in recent years, many new analysis tools combining statistics and microbial ecology have been developed. Explaining each of these in detail is beyond the scope of this paper, but we will mention a few that have been used in VMB research in recent years. These include the Integral-LIBSHUFF tool [32], the linear discriminant analysis effect size (LEfSe) method [33], and the ANOVA-like Differential Expression (ALDEx) tool [34]. The Integral-LIBSHUFF tool uses the exact integral form of the Cramer-von Mises statistic, which tests the quality of a curve fit, and Monte Carlo sampling to calculate the probability that the observed differences between two sequence profiles are due to chance. It was used by Kim and colleagues to determine whether VMB composition was statistically significantly different between women, and the microbiota between body sites within women [35]. LEfSe determines the taxa (or genes or functions) most likely to explain differences between groups by first detecting taxa with significant differential abundance between groups using the non-parametric factorial Kruskal-Wallis sum-rank test, then investigating association consistency using (unpaired) Wilcoxon rank-sum tests, and finally estimating the effect size of each differentially abundant taxa by linear discriminant analysis (LDA). It was used to determine which VMB taxa were associated with race [36], with high risk HPV infection [37], and with the vaginal versus seminal microbiome within sexual relationships [27]. The ALDEx tool decomposes sample-to-sample variation into four parts (within-condition variation, between-condition variation, sampling variation, and general unexplained error) using Monte-Carlo sampling from a Dirichlet distribution. The outcome is a log2-transformed abundance value per taxa, which represents its abundance relative to the mean abundance of all taxa in the sample. It was used by Macklaim and colleagues, in combination with Welch’s t-tests and Benjamini Hochberg false discovery rate correction, to determine differentially abundant VMB taxa pre- and post-treatment of BV with oral tinidazole and probiotics [38].

***Dimensionality reduction - clustering:***

While the methods described in the previous paragraph make optimal use of NGS data to analyze differences in bacterial communities, they can only incorporate a limited number of demographic, behavioral or clinical covariates. They are therefore not ideal for addressing certain clinical epidemiological research questions, for example, those that require adjustment for multiple confounders. Dimensionality reduction techniques use all of the available NGS data (usually in the form of a distance matrix) but reduce the data to a small number of random variables that can subsequently be incorporated into bivariable or multivariable biostatistical models, including models with repeated measures over time.

A commonly used dimensionality reduction method in microbiome research is clustering. Many clustering methods and software tools are available, each with their own advantages and disadvantages [39]. VMB studies to date have identified 3-9 clusters per study (the number depending on the sample size and study population), usually including multiple clusters dominated by one *Lactobacillus* species (*L. crispatus*, *L. iners*, *L. gasseri*, *L. jensenii*, and/or *L. vaginalis*), at least one cluster with intermediate diversity (typically dominated by lactobacilli and *G. vaginalis*), and at least one cluster containing mixtures of anaerobes with or without *L. iners* [2]. After clustering, each sample (each woman at each sampling time point) can be assigned to a microbiota cluster, and biostatistical methods can be used to determine correlations or associations between variables of interest and membership of specific clusters [21, 28, 31, 40-43]. Cluster membership can be incorporated into multivariable statistical models as a categorical or ordinal variable (if the clusters are ranked, for example based on diversity) [44], but can also be analyzed as indicator variables with the cluster of interest coded as 1 and all other clusters as 0 [37, 45, 46]. Mehta and colleagues used multinomial mixed effects modeling with the VMB clusters ranked by relative *Lactobacillus* abundance as multinomial outcome, a subject-specific random slope and intercept, and various independent variables to determine cluster membership trends over time by HIV status and other factors [44]. Romero and colleagues used generalized estimating equations (GEE) models and different types of linear mixed effects models to determine associations between VMB cluster membership (with indicator variables for each cluster of interest) and pregnancy status [45].

Microbiome cluster membership may change over time. Gajer and colleagues introduced the concept of ‘community classes’, which they defined as the dominant VMB cluster over 16 weeks of observation [47]. This new variable was assessed as a predictor of HPV status by Fisher’s exact test [42]. They also calculated the frequency of transitions between each pair of clusters, as well as the rate of change of the log of Jensen-Shannon distances between consecutive clusters [47]. The latter was used as the independent variable in a linear mixed-effects model, which included ‘subject’ as the random effect (to account for correlations between repeated measurements in the same subject) and various dependent variables as fixed effects to determine their associations with cluster transition.

***Other dimension reduction methods:***

Other dimension reduction methods include those that are used when the number of dimensions in the data is not known (such as principle components analysis (PCA), principal coordinates analysis (PCoA), non-metric multidimensional scaling (NMDS), and factor analysis) or is known (such as LDA and canonical correlation analysis (CCA)). PCA and PCoA are by far the most used in VMB research [18, 21, 27, 31, 37, 43, 46-49] but NMDS has also been used [46, 50]. PCA and PCoA use an eigenvector-based approach to represent multidimensional data in as few dimensions (the principal components (PCs)) as possible. They aim for the first PC to account for as much of the variability in the data as possible with each succeeding component in turn accounting for the highest variance possible under the constraint that it is orthogonal to the preceding components. The PCs are often plotted in two or three-dimensional plots to visualize groupings (Appendix: Figure 4) and the first PC is often used as a dependent variable in statistical models to identify factors associated with it. The difference between PCA and PCoA is in the types of matrix input that can be used. PCA requires the input of a covariance or correlation matrix whereas PCoA can use matrices based on any distance metric.

**Using qPCR data to answer biomedical research questions**

Whereas NGS characterizes ‘all’ 16S rDNA present in a sample, qPCR quantifies one specific taxon (or, in the case of a multiplex qPCR assay, a few specific taxa). qPCR assays use primers directed to the taxon of interest to amplify this taxon (usually resulting in higher sensitivity and specificity for this taxon than NGS), and the fluorescent DNA-binding dye SYBRG-green for determination of its concentration by comparing the fluorescence level of the sample to a standard curve. The bacterial load of specific taxa is sometimes a better indicator of health or disease, or the effect of an intervention, than the mere presence or absence of the taxa. When this is the case, and when the taxa of interest are known a priori, it is likely better to use (multiplex) qPCR than NGS. qPCR assays have been developed for several VMB taxa such as the *Lactobacillus* genus, several *Lactobacillus* species, and several BV-associated taxa (*G. vaginalis*, *Atopobium vaginae*, *Leptotrichia* genus, *Sneathia* genus, *Prevotella* genus, among others) [51-53]. Similar to NGS data, qPCR data can be used to answer biomedical questions that: 1) compare the presence or concentrations of taxa (also referred to as bacterial loads) between groups or sampling time points; and 2) combine the prevalence and/or concentrations of a number of taxa into one or more variables.

***Comparing the presence or concentrations of taxa:*** The outcome of a qPCR assay is the concentration of the targeted bacterial taxa (number of 16S rDNA gene equivalents (geq) or copies/ml). This information can be reduced to a binary variable (indicating presence or absence of the taxa in the sample) or categorical variable (for example: not quantifiable, <106 geq/ml, ≥106 geq/ml [53]), but is usually used as a continuous variable, often log10-transformed. These binary, categorical or continuous variables can subsequently be compared between groups using standard biostatistical tests or used in bivariable or multivariable biostatistical models. In VMB research, such qPCR-based studies compared the presence and/or concentrations of taxa of interest between women with or without BV by Amsel criteria or Nugent scoring (for example, [53-58], women with or without vaginal symptoms [59, 60], pre-and post-menopausal women [61], in young women pre- and post-sexual debut [62], in women with pelvic inflammatory disease [63] or HIV [64, 65], and longitudinally over menstrual cycles [55, 66] and in pregnancy [67]. Current research is beginning to zoom in on pathogenesis of dysbiosis, such as associations between the presence and/or concentrations of taxa of interest and vaginal mucosal inflammatory markers [68, 69] or bacterial metabolites [70], the potential use of multiplex qPCR as a new diagnostic tool for BV [71, 72], and the inclusion of qPCR data in multivariable statistical models (typically mixed linear effects models). When including qPCR data in multivariable models, a value of half the lower limit of detection is sometimes assigned to women who do not have detectable concentrations of a particular taxon in their VMB to avoid losing a large proportion of the data [73]. The use of qPCR is particularly useful when studying interventions, such as antibiotic treatment [74, 75]. For example, unchanged concentrations of taxa that should be sensitive to the antibiotic used might suggest antibiotic resistance whereas declining but still detectable concentrations might lead to recurrence [74].

***Combining the prevalence and/or concentrations of a number of taxa into one or more variables:***

The dimensionality-reduction techniques described above for NGS data can also be applied to qPCR data, reducing multiple taxa concentrations to a few variables. For example, Jespers and colleagues conducted PCA analysis on log10-transformed qPCR data of *L. crispatus*, *L. iners*, *L. jensenii*, *L. gasseri*, *L. vaginalis*, *A. vaginae* and *G. vaginalis* (Figure 4), and found that the first and second PC explained 58% of the variance, with the first PC describing a gradient from a *Lactobacillus*-dominated VMB to a *G. vaginalis* and *A. vaginae*-dominated VMB and the second PC the concentration of *L. crispatus* versus *L. iners* [72]. This study also tested various combinations of individual taxa concentrations in their ability to diagnose BV (using receiver operating characteristics analyses) and concluded that a combination of *Lactobacillus* genus, *G. vaginalis*, and *A. vaginae* qPCR performed best with a sensitivity of 93.4% and a specificity of 83.6% compared to a Nugent score of 7-10.

***qPCR as an addition to NGS or other -omics methods:*** In recent years, NGS and qPCR are increasingly being used in combination. These studies typically employ NGS first to identify the taxa that are most important in the context of the research question, followed by qPCR assays of those taxa of interest to quantify their relative importance or evaluate the quantitative effects of interventions (see for example [73] for use of this strategy to study cervicitis). qPCR concentrations can also be correlated with results from –omics methods, such as proteomics and metabolomics. For example, Srinivasan and colleagues recently correlated the concentrations of key vaginal bacteria with the 30% most variable metabolites in their vaginal samples using Pearson correlation coefficients [70].

**Conclusions**

NGS and (multiplex) qPCR technologies are now widely available and affordable and research incorporating these methods is growing exponentially. However, in the VMB field, most research to date is still laboratory and bioinformatics-focused, and does not include rigorous epidemiological methods such as adequate statistical power to answer clinically relevant research questions, appropriate sampling methods, study designs that minimize biases, and statistical models that address confounding. We believe that the time has now come for incorporation of microbiome data into properly designed and powered clinical epidemiological studies to study the associations between microbiota and clinical outcomes properly, and to test interventions on a much wider scale.

**Acknowledgments**

We thank Hanneke Borgdorff for critically reviewing the manuscript and providing figures, and Jacques Ravel for providing figures.

**References**

1. Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. Nat Rev Microbiol 2015;13(12):787-94. doi: 10.1038/nrmicro3565.
2. van de Wijgert JHHM, Borgdorff H, Verhelst R, Crucitti T, Francis S, Verstraelen H, et al. The vaginal microbiota: what have we learned after a decade of molecular characterization? PLoS One 2014;9(8):e105998. doi: 10.1371/journal.pone.0105998.
3. Amsel R, Totten PA, Spiegel CA, Chen KCS, Eschenbach D, Holmes KK. Nonspecific vaginitis: diagnostic criteria and microbial and epidemiologic associations. Am J Med 1983;74(1):14-22.
4. Nugent RP, Krohn MA, Hillier SL. Reliability of diagnosing bacterial vaginosis is improved by a standardized method of Gram stain interpretation. J Clin Microbiol 1991;29(2):297-301.
5. Woese CR, Kandler O, Wheelis M. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 1990;87(12):4576-9.
6. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D et al. Experimental and analytical tools for studying the human microbiome. Nat Rev Genet 2011;13(1):47-58. doi: 10.1038/nrg3129.
7. Weinstock GM. Genomic approached to studying the human microbiota. Nature 2012;489(7415):250-6. doi: 10.1038/nature11553.
8. Ronaghi M, Uhlén M, Nyrén P. A sequencing method based on real-time pyrophosphate. Science 1998;281(5375):363.
9. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 2012;13(1):341. doi: 10.1186/1471-2164-13-341.
10. Brooks JP, Edwards DJ, Harwich Jr MD, Rivera MC, Fettweis JM, Serrano MG, et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. BMC Microbiol 2015;15:66. doi: 10.1186/s12866-015-0351-6.
11. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biology 2014;12:87. doi: 10.1186/s12915-014-0087-z.
12. Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. J Mol Biology 1990;215(3):403-10.
13. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL\_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 1997;25(24):4876-82.
14. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nature Methods 2010; doi:10.1038/nmeth.f.303.
15. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 2014; 42(Database issue):D633-42. doi: 10.1093/nar/gkt1244.
16. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 2012;6(3):610–8. doi: 10.1038/ismej.2011.
17. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. Genbank. Nucleic Acids Res 2013;41(Database issue):D36-42. doi: 10.1093/nar/gks1195.
18. Srinivasan S, Hoffman NG, Morgan MT, Matsen FA, Fiedler TL, Hall RW, et al. Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. PLoS One 2012;7:e37818. doi: 10.1371/journal.pone.0037818.
19. Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nat Rev Genet 2012;13(5):303-14. doi: 10.1038/nrg3186.
20. Gotelli NJ, Colwell RK. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters 2001;4(4):379-91.
21. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL, et al. Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci USA 2011;108(Suppl 1):4680-7. doi: 10.1073/pnas.1002611107.
22. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol 2014;10(4):e1003531. doi: 10.1371/journal.pcbi.1003531.
23. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 2005;71(12):8228-35.
24. Hummelen R, Fernandes AD, Macklaim JM, Dickson RJ, Changalucha J, Gloor GB, et al. Deep sequencing of the vaginal microbiota of women with HIV. PLoS One 2010;5:e12078. doi: 10.1371/journal.pone.0012078.
25. Hummelen R, Macklaim JM, Bisanz JE, Hammond JA, McMillan A, Vongsa R, et al. Vaginal microbiome and epithelial gene array in post-menopausal women with moderate to severe dryness. PLoS One 2011;6:e26602. doi: 10.1371/journal.pone.0026602.
26. Muzny CA, Sunesara IR, Kumar R, Mena LA, Griswold ME, Martin DH, et al. Characterization of the vaginal microbiota among sexual risk behavior groups of women with bacterial vaginosis. PLoS One 2013;8(11):e80254. doi: 10.1371/journal.pone.0080254.
27. Mändar R, Punab M, Borovkova N, Lapp E, Kiiker R, Korrovits P, et al. Complementary seminovaginal microbiome in couples. Res Microbiol 2015;166(5):440e447. doi: 10.1016/j.resmic.2015.03.009.
28. Borgdorff H, Tsivtsivadze E, Verhelst R, Marzorati M, Jurriaans S, et al. Lactobacillus-dominated cervicovaginal microbiota associated with reduced HIV/STI prevalence and genital HIV viral load in African women. ISME J 2014;8(9):1781-93. doi: 10.1038/ismej.2014.26.
29. Borgdorff H, Gautam R, Armstrong SD, Xia D, Ndayisaba GF, van Teijlingen NH, et al. Cervicovaginal dysbiosis is associated with proteome changes related to immune activation and cell death. Mucosal Immunol 2015;doi: 10.1038/mi.2015.86.
30. Dufrene M, Legendre P. Species assemblages and indicator species: The need for a flexible asymmetrical approach. Ecological Monographs 1997;67:345-66.
31. Hickey RJ, Zhou X, Settles ML, Erb J, Malone K, Hansmann MA, et al. Vaginal microbiota of adolescent girls prior to the onset of menarche resemble those of reproductive-age women. mBio 2015;6(2):e00097-15. doi: 10.1128/mBio.00097-15.
32. Schloss PD, Larget BR, Handelsman J. Integration of microbial ecology and statistics: a test to compare gene libraries. Appl Environ Microbiol 2004;70(9):5485-92.
33. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. Metagenomic biomarker discovery and explanation. Genome Biology 2011;12(6):R60. doi: 10.1186/gb-2011-12-6-r60.
34. Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. PLoS One 2013;8(7):e67019. doi: 10.1371/journal.pone.0067019.
35. Kim TK, Thomas SM, Ho M, Sharma S, Reich CI, Frank JA, et al. Heterogeneity of vaginal microbial communities within individuals. J Clin Microbiol 2009;47(4):1181-9. doi: 10.1128/JCM.00854-08.
36. Fettweis JM, Brooks P, Serrano MG, Sheth NU, Girerd PH, Edwards DJ, et al. Differences in vaginal microbiome in African American women versus women of European ancestry. Microbiology 2014;160(Pt 10):2272-82. doi: 10.1099/mic.0.081034-0.
37. Dareng EO, Ma B, Famooto AO, Akarolo A, Offiong RA, Olayniyan O, et al. Prevalent high-risk HPV infection and vaginal microbiota in Nigerian women. Epidemiol Infect 2015: doi:10.1017/S0950268815000965.
38. Macklaim JM, Jose C. Clemente JC, Knight R, Gloor GB, Reid G. Changes in vaginal microbiota following antimicrobial and probiotic therapy. Microb Ecol Health Dis 2015;26:27799. doi: 10.3402/mehd.v26.27799.
39. Wiwie C, Baumbach J, Röttger R. Comparing the performance of biomedical clustering methods. Nat Methods 2015;12(11):1033-8.
40. Brotman RM, Bradford LL, Conrad M, Gajer P, Ault K, Peralta L, et al. Association between Trichomonas vaginalis and vaginal bacterial community composition among reproductive-age women. Sex Transm Dis 2012;39(10):807-12.
41. Brotman RM, He X, Gajer P, Fadrosh D, Sharma E, Emmanuel F, et al. Association between cigarette smoking and the vaginal microbiota: a pilot study. BMC Infect Dis 2014;14:471. doi: 10.1186/1471-2334-14-471.
42. Brotman RM, Shardell MD, Gajer P, Tracy JK, Zenilman JM, Ravel J, et al. Interplay between the temporal dynamics of the vaginal microbiota and human papillomavirus detection. J Infect Dis 2014;210(11):1723-1733. doi: 10.1093/infdis/jiu330.
43. Albert AYK, Chaban B, Wagner EC, Schellenberg JJ, Links MG, van Schalkwyk J, et al. A study of the vaginal microbiome in healthy Canadian women utilizing cpn60-based molecular profiling reveals distinct Gardnerella subgroup community state types. PLoS One 2015;10(8):e0135620. doi: 10.1371/journal.pone.0135620.
44. Mehta SD, Donovan B, Weber KM, Cohen M, Ravel J, Gajer P, et al. The vaginal microbiota over an 8- to 10-year period in a cohort of HIV-infected and HIV-uninfected women. PLoS One 2015;10(2):e0116894. doi: 10.1371/journal.pone.0116894.
45. Romero R, Hassan SS, Gajer P, Tarca AL, Fadrosh DW, Nikita L, et al. The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. Microbiome 2014;2(1):4. doi: 10.1186/2049-2618-2-4.
46. MacIntyre DA, Chandiramani M, Lee YS, Kindinger L, Smith A, Angelopoulos N, et al. The vaginal microbiome during pregnancy and the postpartum period in a European population. Sci Rep 2015;5:8988. doi: 10.1038/srep08988.
47. Gajer P, Brotman RM, Gyoyun B, Sakamoto J, Schütte UM, Zhong X, et al. Temporal dynamics of the human vaginal microbiota. Sci Transl Med 2012;4(132):132ra52. doi: 10.1126/scitranslmed.3003605.
48. Walther-Antonio MR, Jeraldo P, Berg Miller ME, Yeoman CJ, Nelson KE, Wilson BA, et al. Pregnancy’s Stronghold on the vaginal microbiome. PLoS One 2014;9(6):e98514. doi: 10.1371/journal.pone.0098514.
49. Ursell LK, Gunawardana M, Chang S, Mullen M, Moss JA, Herold BC, et al. Comparison of the vaginal microbial communities in women with recurrent genital HSV receiving acyclovir intravaginal rings. Antiviral Res 2014;102:87-94. doi: 10.1016/j.antiviral.2013.12.004.
50. Hickey RJ, Abdo Z, Zhou X, Nemeth K, Hansmann M, Osborn TW3rd, et al. Effects of tampons and menses on the composition and diversity of vaginal microbial communities over time. BJOG 2013;120(6):695-704. doi: 10.1111/1471-0528.12151.
51. Ferris MJ, Masztal A, Martin DH. Use of species-directed 16S rRNA gene PCR primers for detection of Atopobium vaginae in patients with bacterial vaginosis. J Clin Microbiol 2004;42(12):5892-4.
52. Fredricks DN, Fiedler TL, Thomas KK, Oakley BB, Marrazzo JM. Targeted PCR for detection of vaginal bacteria associated with bacterial vaginosis. J Clin Microbiol 2007;45(10):3270-6.
53. Jespers V, van de Wijgert J, Cools P, Verhelst R, Verstraelen H, Delany-Moretlwe S, et al. The significance of Lactobacillus crispatus and L. vaginalis for vaginal health and the negative effect of recent sex: a cross-sectional descriptive study across groups of African women. BMC Infect Dis 2015;15:115. doi: 10.1186/s12879-015-0825-z.
54. Zozaya-Hinchliffe M, Lillis R, Martin DH, Ferris MJ. Quantitative PCR assessments of bacterial species in women with and without bacterial vaginosis. J Clin Microbiol 2010;48(5):1812-9. doi: 10.1128/JCM.00851-09.
55. Srinivasan S, Liu C, Mitchell CM, Fiedler TL, Thomas KK, Agnew KJ, et al. Temporal variability of human vaginal bacteria and relationship with bacterial vaginosis. PLoS One 2010;5(4):e10197. doi: 10.1371/journal.pone.0010197.
56. Shipitsyna E, Roos A, Datcu R, Hallen A, Fredlund H, Jensen JS, et al.: Composition of the vaginal microbiota in women of reproductive age - sensitive and specific molecular diagnosis of bacterial vaginosis is possible? PLoS One 2013;8(4):e60670. doi: 10.1371/journal.pone.0060670.
57. Datcu R, Gesink D, Mulvad G, Montgomery-Andersen R, Rink E, Koch A, et al. Vaginal microbiome in women from Greenland assessed by microscopy and quantitative PCR. BMC Infect Dis 2013;13:480. doi: 10.1186/1471-2334-13-480.
58. Ling Z, Liu X, Luo Y, Wu X, Yuan L, Tong X, et al. Associations between vaginal pathogenic community and bacterial vaginosis in Chinese reproductive-age women. PLoS One 2013;8(10):e76589. doi: 10.1371/journal.pone.0076589.
59. Biagi E, Vitali B, Pugliese C, Candela M, Donders GG, Brigidi P. Quantitative variations in the vaginal bacterial population associated with asymptomatic infections: a real-time polymerase chain reaction study. Eur J Clin Microbiol Infect Dis 2009;28(3):281-5. doi: 10.1007/s10096-008-0617-0.
60. Jespers V, Menten J, Smet H, Poradosu S, Abdellati S, Verhelst R, et al. Quantification of bacterial species of the vaginal microbiome in different groups of women using nucleic acid amplification tests. BMC Microbiol 2012;12:83. doi: 10.1186/1471-2180-12-83.
61. Zhang R, Daroczy K, Xiao B, Yu L, Chen R, Liao Q. Qualitative and semiquantitative analysis of Lactobacillus species in the vaginas of healthy fertile and postmenopausal Chinese women. J Med Microbiol 2012;61(Pt 5):729-39. doi: 10.1099/jmm.0.038687-0.
62. Mitchell CM, Fredricks DN, Winer RL, Koutsky L. Effect of sexual debut on vaginal microbiota in a cohort of young women. Obstet Gynecol 2012;120(6):1306-13. doi: http://10.1097/AOG.0b013e31827075ac.
63. Haggerty CL, Totten PA, Ferris M, Martin DH, Hoferka S, Astete SG, et al. Clinical characteristics of bacterial vaginosis among women testing positive for fastidious bacteria. Sex Transm Infect 2009;85(4):242-8. doi: 10.1136/sti.2008.032821.
64. Pepin J, Deslandes S, Giroux G, Sobela F, Khonde N, Diakite S, et al. The complex vaginal flora of West African women with bacterial vaginosis. PLoS One 2011;6(9):e25082. doi: 10.1371/journal.pone.0025082.
65. Mitchell C, Balkus JE, Fredricks D, Liu C, McKernan-Mullin J, Frenkel LM, et al. Interaction between lactobacilli, bacterial vaginosis-associated bacteria, and HIV Type 1 RNA and DNA genital shedding in U.S. and Kenyan women. AIDS Res Hum Retroviruses 2013;29(1):13-9. doi: 10.1089/AID.2012.0187.
66. Santiago GL, Cools P, Verstraelen H, Trog M, Missine G, El Aila N, et al. Longitudinal study of the dynamics of vaginal microflora during two consecutive menstrual cycles. PLoS One 2011;6(11):e28180. doi: 10.1371/journal.pone.0028180.
67. Vitali B, Cruciani F, Picone G, Parolin C, Donders G, Laghi L. Vaginal microbiome and metabolome highlight specific signatures of bacterial vaginosis. Eur J Clin Microbiol Infect Dis 2015:doi 10.1007/s10096-015-2490-y.
68. Kyongo JK, Jespers V, Goovaerts O, Michiels J, Menten J, Fichorova RN, et al. Searching for lower female genital tract soluble and cellular biomarkers: defining levels and predictors in a cohort of healthy Caucasian women. PLoS One 2012;7(8):e43951. doi: 10.1371/journal.pone.0043951.
69. Kyongo JK, Crucitti T, Menten J, Hardy L, Cools P, Michiels J, et al. A cross-sectional analysis of selected genital tract immunological markers and molecular vaginal microbiota in Sub-Saharan African women with relevance to HIV risk and prevention. Clin Vaccine Immunol 2015;22(5):526-38. doi: 10.1128/CVI.00762-14.
70. Srinivasan S, Morgan MT, Fiedler TL, Djukovic D, Hoffman NG, Raftery D, et al. Metabolic signatures of bacterial vaginosis. MBio 2015;6(2)pii:e00204-15. doi: 10.1128/mBio.00204-15.
71. Menard JP, Fenollar F, Henry M, Bretelle F, Raoult D. Molecular quantification of Gardnerella vaginalis and Atopobium vaginae loads to predict bacterial vaginosis. Clin Infect Dis 2008;47(1):33-43. doi: 10.1086/588661.
72. Jespers V, Crucitti T, van de Wijgert J, Vaneechoutte M , Delany-Moretlwee S, Mwaura M, et al. A DNA tool for early detection of vaginal dysbiosis in African women. Res Microbiol 2015;pii:S0923-2508(15)00180-1. doi: 10.1016/j.resmic.2015.10.006.
73. Gorgos LM, Sycuro LK, Srinivasan S, Fiedler TL, Morgan MT, Balkus JE, et al. Relationship of specific bacteria in the cervical and vaginal microbiotas with cervicitis. Sex Transm Dis 2015;42(9):475-81. doi: 10.1097/OLQ.0000000000000318.
74. Fredricks DN, Fiedler TL, Thomas KK, Mitchell CM, Marrazzo JM. Changes in vaginal bacterial concentrations with intravaginal metronidazole therapy for bacterial vaginosis as assessed by quantitative PCR. J Clin Microbiol 2009;47(3):721-6. doi: 10.1128/JCM.01384-08.
75. Mayer BT, Srinivasan S, Fiedler TL, Marrazzo JM, Fredricks DN, et al. Rapid and profound shifts in the vaginal microbiota following antibiotic treatment for bacterial vaginosis. J Infect Dis 2015;212(5):793-802. doi: 10.1093/infdis/jiv079.