**Appendix belonging to: ‘Incorporating microbiota data into epidemiological models: Examples from vaginal microbiota research’**

Janneke van de Wijgert, Vicky Jespers

**Basic information about NGS data of relevance to epidemiologists**

16S ribosomal RNA (rRNA) is a component of the 30S small subunit of prokaryotic ribosomes; the prokaryotic gene encoding for it is referred to as 16S rDNA. Some regions of this gene are highly conserved among all prokaryotes (bacteria and archaea) while other regions are variable. In microbiota studies, the conserved regions 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) [1-3]. The ability to classify reads to species level depends on various factors including choice of NGS platform [4, 5], 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.

***Preprocessing of sequence reads:*** A multiplex 16S rDNA NGS run typically results in thousands of sequence reads per biological sample [4, 5]. As explained above, these sequence reads represent short pieces of the targeted variable region(s) of the bacterial 16S rDNA gene. The sequence reads are checked for quality and preprocessed. The exact preprocessing procedures depend on the sequencing method used but will likely include quality control procedures (to filter out low quality and short reads, reads without barcode or primer sequences, chimeric sequences, etc.), piecing together of smaller reads into larger reads, sorting sequences by barcode, and removing barcode and primer sequences. The final reads after completion of these preprocessing procedures are used for identification of the bacterial taxa present in each sample. Unfortunately, all laboratory and reads preprocessing procedures are known to introduce biases, resulting in an observed microbiota composition that is different from the actual microbiota composition [6, 7]. Studies using mock samples (with known quantities of known bacterial taxa) have shown that biases introduced by the initial DNA extraction and PCR amplification methods are larger than those due to the subsequent sequencing and classification methods [10]. Some laboratories estimate their error rates in sequencing runs by sequencing such mock samples alongside the experimental samples.

***Sequence alignment and assignment of reads to bacterial taxa:*** In bioinformatics, a sequence alignment is a way of arranging sequences of DNA (or RNA or protein) to identify regions of similarity. Alignment can be done pairwise or for multiple sequences (multiple sequence alignment or MSA). Popular alignment tools are BLAST [8], ClustalX [9], and the UCLUST tool in QIIME [10]. In microbiota research, alignment is mostly used to assign sequence reads to bacterial taxa. The reads are usually first assigned to operational taxonomic units (OTUs), which are subsequently compared to known bacterial taxa sequences in publicly available databases. These databases include quality-controlled databases that only collect 16S rDNA sequences (such as the ribosomal database project [11] and Greengenes [12]), or databases containing any publicly available nucleotide sequences and their protein translations (such as the National Center for Biotechnology Information (NCBI)’s Genbank [13]). It is not always possible to assign sequences at species level using these publicly available databases, and some laboratories specialized in the VMB have designed their own customized databases to fill the gaps (see for example [14]). Some researchers report on phylotypes instead of OTUs. OTUs use a percent sequence similarity threshold (usually 97% for species level) within the experimental dataset, and phylotypes compared to an external database, for classification. While OTUs are sensitive to sequencing errors, phylotypes are dependent on the accuracy and completeness of external databases; the ability to classify reads to species level does not depend on whether OTUs or phylotypes are used.

***Phylogenetic analyses:*** From the resulting sequence alignment, sequence homology can be inferred and 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 [15]). Phylogenies are typically visualized using a dendrogram (Figure 1).

***Rarefaction curves and rarefying:*** The total number of sequence reads, and the number of detected OTUs, increases with the number of samples sequenced. Rarefaction curves plot the number of taxa as a function of the number of sequences per sample (Figure 2) [16]. A steep slope indicates that a large fraction of the taxa diversity remains to be discovered but if the curve becomes flatter to the right, deeper sequencing is unlikely to yield a significant number of additional taxa.

***Visualising diversity:*** 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 (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 (Figure 3).

**References**

1. 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.
2. 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.
3. Weinstock GM. Genomic approached to studying the human microbiota. Nature 2012;489(7415):250-6. doi: 10.1038/nature11553.
4. Ronaghi M, Uhlén M, Nyrén P. A sequencing method based on real-time pyrophosphate. Science 1998;281(5375):363.
5. 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.
6. 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.
7. 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.
8. Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. J Mol Biology 1990;215(3):403-10.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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.
15. Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nat Rev Genet 2012;13(5):303-14. doi: 10.1038/nrg3186.
16. Gotelli NJ, Colwell RK. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters 2001;4(4):379-91.
17. 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.
18. 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.
19. 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.

**Figure 1: Heatmap with dendrogram and Shannon diversity index (NGS data)**



Example of a heatmap (courtesy: Jacques Ravel), using data published in [17]. The heatmap shows each phylotype on the vertical axis for each participant on the horizontal axis with the relative abundance of each phylotype shown in different colors.**Figure 2: Rarefaction curves (NGS data)**

Rarefaction curves of the vaginal microbiota of 8 women participating in the HELIUS study in Amsterdam, the Netherlands (unpublished data). Samples were sequenced using the MiSeq platform. Samples 1 and 2 have high richness, samples 3-5 intermediate richness and 6-8 low richness. The curves of samples 2, 5, and 8 have (almost) reached a plateau and deeper sequencing will therefore unlikely reveal more OTUs in these samples. However, the curves of samples 1, 3, 4, 6, and 7 have not reached a plateau and likely contain OTUs that have not been sequenced.

**Figure 3: Interpolated bar plot over time (NGS data)**



Example of an interpolated bar plot (courtesy: Jacques Ravel), using data published in [18]. The bar plot shows the relative abundance of phylotypes on the vertical axis (with each phylotype shown in a different color) and time point on the horizontal axis for one woman.**Figure 4: PCA plot (qPCR data)**



The figure displays a scatter plot of the second versus first PC for 426 samples of South African, Kenyan and Rwandan women [19]. The colors of the individual dots show the BV status based on the Nugent score for each sample. The arrows show the factor loadings of the microbiota species included in the two PCs. The first PC represents a gradient from a *Lactobacillus*-dominated VMB to a *G. vaginalis* and *A. vaginae*-dominated VMB. The second PC represents the concentration of *L. crispatus* versus *L. iners.*