Vaginal microbial profiling in a preterm birth high-risk cohort using shallow shotgun metagenomics

Alishum Ali\textsuperscript{A,B,C}, Claus T Christophersen\textsuperscript{A,E} and Jeffrey A Keelan\textsuperscript{B,D,F}

\textsuperscript{A}Western Australia Human Microbiome Collaboration Centre (WAHMCC), Trace and Environmental DNA (TrEnD) Laboratory, School of Molecular Life Sciences, Curtin University of Technology, Perth, WA, Australia
\textsuperscript{B}School of Biomedical Sciences and Division of Obstetrics and Gynaecology, The University of Western Australia, Perth, WA, Australia
\textsuperscript{C}Department of Neonatology, Fiona Stanley Hospital, Perth, WA, Australia
\textsuperscript{D}King Edward Memorial Hospital, Perth, WA, Australia
\textsuperscript{E}School of Medical and Health Sciences, Edith Cowan University, Perth, WA, Australia
\textsuperscript{F}Email: jeff.keelan@uwa.edu.au

Abstract. Preterm birth (PTB) is a significant health problem globally, with an estimate of 15 million cases annually. Approximately 10% of neonates born early will die prematurely, while a subset will develop severe life-long morbidities. Unfortunately, preterm birth’s syndromic nature has evaded prevention strategies, and it continues to impose a high burden on healthcare systems and families. The role of vaginal bacteria in triggering biomolecular causes of PTB has been recognised for years. However, translating this knowledge to practical diagnostic and therapeutic strategies has remained elusive. New techniques in high-throughput sequencing have improved our understanding of the nature and role of the vaginal microbiome during pregnancy. Several multi-ethnic and multi-geographical studies into the vaginal microbiome have identified five distinct bacterial profiles termed community state types (CSTs), one of which is positively associated with dysbiosis and increased risk of PTB. In a small pilot study of first-trimester vaginal microbial DNA obtained from pregnant women at high-risk of PTB, we compared the CST profiles generated using standard 16S amplicon sequencing with shallow shotgun metagenomics (SSM). Both methods identified the presence of the five CSTs as has been reported previously, although the metagenomic data showed greater taxonomic resolution and more accurate CST assignation. These findings suggest that SSM is a cost-effective and potentially superior alternative to 16S sequencing for vaginal microbiome analysis.

Received 24 April 2021, accepted 20 May 2021, published online 15 June 2021

Introduction

Preterm birth (PTB), defined by the World Health Organization as all deliveries between 20–37 weeks of completed gestation, is a complex syndrome. The condition is divided into four groups based on gestational age (GA) at of birth: extreme PTB (<28 GA), very PTB (28-32 GA), moderate PTB (32-34 GA) and late PTB (34-37 GA)\textsuperscript{1}. PTB impacts the lives of 15 million families annually, with an approximate 10% mortality rate in the first month after delivery\textsuperscript{2}. Despite advances in neonatal care and improved survival and reduced morbidity, preterm infants are at high risk of developing metabolic disorders and debilitating neurological conditions, such as blindness, deafness, neurodevelopmental delays, and behavioural issues well into adulthood\textsuperscript{3}. A recent meta-analysis of PTB hospitalisation costs in the US, Canada, and The Netherlands reported that the individual healthcare costs for extreme PTB were between $111 152–$576 972 per delivery\textsuperscript{4}.

PTB is a syndrome that is both difficult to predict and to prevent\textsuperscript{2}. Multiple methods and approaches for PTB prediction have been developed and evaluated, with varying success\textsuperscript{6–9}. Similarly, preventative treatments are limited and lack the required efficacy, applicability and precision. Women identified as at high risk of PTB (typically due to either a previous PTB and/or a short cervical length defined as <25 mm) typically receive one of two clinically-recommended preventive interventions at the discretion of the treating obstetrician, namely exogenous progesterone (vaginally, orally or intramuscularly) or cervical cerclage surgery\textsuperscript{10}.

A meta-analysis from 2018 with large high-risk pregnancies cohorts report that vaginal progesterone (VP) use resulted in a pooled relative risk ratio (RR) of 0.29–0.68, while cervical cerclage had a RR of 0.64–0.70\textsuperscript{11}. The effectiveness of VP appears to be particularly robust in high-risk women with short cervical length (<25 mm), as has been recently demonstrated in the EPPPIC meta-analysis\textsuperscript{12}.

PTB has long been known to be associated with ascending intrauterine infections originating from a dysbiotic (sub-optimal) lower vaginal tract microbiome\textsuperscript{5,13,14}. In 2–27% of pregnant women, the microbiome composition shifts to an increase in species...
diversity, leading to a dysbiotic vaginal microbiome associated with a disease state. Several studies have now show that an increase in bacterial diversity is linked to reproductive tract inflammation and increased risk of PTB\textsuperscript{15–19}. Despite numerous studies investigating the predictive usefulness of vaginal microbiome analysis, the diagnostic utility of this approach remains elusive. In a recent large study of low-risk Australian women, a high-risk microbial profile in the 2nd trimester was identified based on the presence/absence and combinations of known bacterial species\textsuperscript{18}. Notwithstanding this study’s clinical relevance to PTB management, it is important to point out that this particular work is based on analysis of a selected number of risk-associated bacteria, not the entire microbiome per se.

In 2011, Ravel and colleagues classified the vaginal microbiome of healthy reproductive-age women into five distinct community state types (CST), conditional on the dominance of one of four \textit{Lactobacillus} spp. or lack thereof. CST-I is dominated by \textit{Lactobacillus crispatus}, CST-II by \textit{Lactobacillus gasseri}, CST-III by \textit{Lactobacillus iners}, CST-IV by diverse anaerobic bacteria resembling the clinically diagnosed condition of bacterial vaginosis (BV), and CST-V by \textit{Lactobacillus jensenii}\textsuperscript{20}. The robustness of CST classifications has since been confirmed in many human microbiome studies, regardless of ethnicity, geographical location or sequencing methodology\textsuperscript{21}.

More recently, the Ravel laboratory developed the tool ‘VALENCIA’ (VAGinaL community state type Nearest Centroid clAssifier), which unbiasedly affirmed the presence of the original broad five CST profiles while defining an additional set of 13 subCST groups\textsuperscript{22}. Importantly, VALENCIA links CST profiles with clinical descriptors across multiple ethnicities, plus provides researchers with the ability to accurately differentiate between known subtypes of CST-IV. The new CST-IV classification now takes into account the presence and abundance of \textit{Lactobacillus} spp. and the following clinically essential bacteria: \textit{Gardnerella vaginalis}, \textit{Bifidobacterium} spp., \textit{Atopobium vaginae}, \textit{Prevotella} spp., \textit{Enterococcus} spp., \textit{Streptococcus} spp., and \textit{Staphylococcus} spp.\textsuperscript{22}.

Notably, in the context of the vaginal microbiome, studies have shown that CST-IV and CST-III dominance early in pregnancy increase the risk of PTB\textsuperscript{18,23}, and it is now believed that microbial diversity assessment and CST profiling may help screen women for PTB risk\textsuperscript{17}. Despite solid evidence associating PTB with CST suboptimal profiles, it is worth mentioning that most of the evidence was from studies with a predominantly Caucasian cohort. In African American cohorts, the associated significance was weak or disappeared altogether\textsuperscript{24}.

With the acceptance of ethnicity as a significant confounding factor, we know that vaginal microbiome dysbiosis is a substantial risk factor for uterine ascending infection, which has been causally linked to up to 40% of all preterm births\textsuperscript{8}. However, the techniques used to generate microbiome data are often constrained by limited financial and bioinformatic resources, limiting their clinical and/or diagnostic utility. Therefore, employing methods that increase taxonomic resolution at a reasonable cost have the potential to enable CST profiling to be conducted for PTB risk prediction and treatment in high-risk pregnancies, as well as increasing the accuracy and resolution of the data.

Presently, the vaginal microbiome is typically studied via two DNA-based approaches and one RNA-based strategy: metabar-coding (DNA), metagenomics (DNA) and, to a lesser extent, metatranscriptomics (RNA, not discussed further)\textsuperscript{25}. Metabar-coding (also known as metataxonomics or amplicon sequencing) is the most commonly used technique for microbiome analysis, partly due to its simplicity, but primarily because of the low cost (typically <$100 per sample) and well established analysis pipelines (e.g. USEARCH/DADA2). Amplicon sequencing involves the PCR amplification of a small hypervariable region or regions (250–500 bp) of the taxonomically informative 16S rRNA gene expressed in all bacterial species. Typically, microbiome specialists would design primers that can amplify a set of variable regions that allow the taxonomic discrimination and identification of bacterial genera – in some cases to the species level; this is necessary for CST profiling, although bias can be introduced through primer design, the selected 16S rRNA gene region and its coverage\textsuperscript{25}. To eliminate obvious bias, primers may need to be redesigned to increase the species detection within the same taxonomic kingdom, or if separate domains are to be targeted, such as when characterising the prokaryotes, fungi and microeukaryotes communities present in the human vaginal tracts\textsuperscript{25,26}.

In contrast, metagenomics or shotgun sequencing has significant advantages over amplicon sequencing. It can remove detection bias by sequencing all DNA present in a sample, providing taxonomy to strain-level accuracy. Furthermore, it provides the researcher with the ability to assess metabolic functional potential of the genomes by conducting pathway analysis based on the sequenced genes. Although standard shotgun sequencing has advantages over amplicon sequencing, it carries some critical disadvantages: (1) the amount of DNA required is at least 1 μg; (2) analysis is expensive ($500–$1000 each); and (3) there is a requirement to have access to specialist bioinformatics resources and high-performance computing\textsuperscript{37}.

We have recently completed a pilot study assessing the taxonomic resolution resulting from a recent methodological advance in metagenomic analysis called shallow shotgun metagenomics (SSM)\textsuperscript{28,29}. In SSM a sample is typically sequenced to a depth <1 million reads, which is an order of magnitude or more lower than the depth expected in a standard metagenomics study (depth between 10 million to 2.5 billion reads)\textsuperscript{29}. The reduction in sequencing depth reduces the cost of SSM to those similar to amplicon sequencing, while retaining broad taxonomic coverage at higher taxonomic resolution with functional genetic information. Hillman and colleagues recently showed that a sequencing depth as low as 100 000 reads can mirror >90% of the alpha diversity and gene functional capacity relative to that mapped by ultra-deep metagenomics\textsuperscript{28}. 
These attributes make SSM ideal for the study of microbiomes in large, longitudinal cohorts.

SSM also has two important practical limitations. Firstly, if a sample type contains a very high host:microbe DNA ratio, such as found in blood or tissue biopsies, then SSM may not be the method of choice, because the dominance of host DNA will swamp the reads assigned to microbes and low abundance species may be missed. Secondly, there are bioinformatic constraints, as most available tools are not designed to meet the particular requirements of SSM-generated data; this can result in the generation of false positives and negatives. Additionally, the entire metagenomics field is limited by the availability of high-quality genome databases due to the infancy of this field. Thus, rare or non-sequenced organisms are reported as negative or unassigned, potentially losing important taxonomic information and compromising interpretation. Although these points are all important limitations to consider in study design, in some microbiomes such as the skin or the vaginal microbiome (our research area) that contain a higher host DNA but low-to-medium biomass, SSM may still offer significant advantages due to the medical importance of identifying bacteria, fungi, viruses and micro-eukaryotes to species or strain resolution, which is not provided by amplicon sequencing.

In this study, we compared the bacterial taxonomic profile of SSM to standard 16S amplicon sequence in the context of the vaginal microbiota. The comparison was made using two sample sets: (1) a mock vaginal community consisting of six vaginal bacterial species with an even abundance of 16.7% to validate the robustness of the pipeline; and (2) DNA from 22 high vaginal swabs collected from women at high risk of PTB during their first trimester in Perth, Western Australia; the swabs were obtained from the Western Australian Pregnancy Biobank, with informed consent and institutional ethical approval. Our swabs yielded DNA concentrations between 1–40 ng/μL; two samples and the negative controls did not have enough DNA for sequencing, and thus were eliminated from analysis. The host DNA in the remaining 20 samples acquired on average 89% of the MiSeq Illumina sequenced reads, leaving only 2.1 million reads for the analysis of 20 samples (plus a mock community control).

Mock community analysis

First, we gauged the performance of our methods using the American Type Culture Collection (ATCC) standardized even abundance vaginal bacteria mock community (ATCC MSA-1007 medically relevant species). The sequencing comparison yielded a highly correlative bacterial composition (Figure 1). The data generated by amplicon sequencing (using primers targeting the v4 16S rRNA gene region) vs. SSM showed excellent taxonomic agreement, although there were some minor differences in relative abundance. However, it is worth mentioning that the (515f/806r) v4 primers used here were designed to enable detection of all six species and thus would be expected to amplify them preferentially. *Mycoplasma hominis* was markedly underrepresented in the *Met* (metagenomics) group where it represented only 1% of total species, while in the *Amp* (amplicon) group it was detected at 19% – very similar to the expected 16.7% in the mock community. We attribute this discrepancy in the *Met* group to the unavoidable stochasticity/compositionality introduced during sequencing, where the abundance of a species can be heavily skewed at random. Additionally, in this study we applied a completely PCR-free library preparation method to avoid amplification bias; however, this approach required a considerable amount (>100–1000 ng total) of starting genomic DNA, more than that provided with the ATCC

![Figure 1](image-url) Mock community bacterial species relative abundance differences between amplicon (Amp) and SSM (Met) sequencing methods. The table on the right corresponds to the relative abundance on a scale from 0–1 (rounded to 2 decimal places).
product (4 ng/µL). This does not offer a full explanation as to why the rest of the species in the mock community were not also detected at lower proportions than expected. We believe the difference is most likely driven by the fundamentally different genome-based reference database and tool used to assign taxonomy in SSM compared to the widely used extensive options available for bacterial amplicon sequencing. Interestingly, other studies comparing the outcome of mock communities and metagenomics also showed that amplicon sequencing seem to provide closer compositional agreement\(^{20}\). Importantly, this artificial situation would be unlikely to be replicated in a real-life analysis of complex, natural bacterial communities.

**Vaginal swab analysis**

Although the amplicon method showed considerable agreement on the taxonomic assignments of mock species, the SSM approach when applied to vaginal samples provided a species or strain level taxonomic assignment with high confidence as required for accurate vaginal CST determination. Figure 2 shows the relative abundance of the top 30 species in the 20 vaginal samples according to the two methods. While there was general agreement in the relative abundance of the most common species, several less abundant species were absent in the Amp group (e.g. *Neisseria gonorrhoeae*). In addition, amplicon sequencing could not resolve the genus *Bifidobacterium* to species level, while SSM identified the species as *B. longum*. We also found that *L. iners* abundance was overrepresented in amplicon sequencing profiles. In contrast, SSM was able to resolve the same samples to either *L. jensenii* or *L. ultunensis* dominance. Enrichment of *L. iners* detection in the Amp group can be explained by preferential primer amplification.

As shown by amplicon sequencing, taxonomic uncertainty can be problematic to vaginal microbiome profiling, because it can distort the accurate picture of community composition and structure. In our analysis of CST profiles, we identified that these inaccuracies can result in CST-V or CST-IV being wrongly labelled as CST-III. This was evident in the sample from one patient (M65), whose profile was dominated by *L. ultunensis* as detected by SSM, but designated CST-III by amplicon sequencing (refer to Figure 3).

Although the detection of atypical CST types such as those dominated by species *L. ultunensis/amylovorus* posed a challenge during the allocation of CSTs, the fact that *Gardnerella vaginalis* seems to co-exist in these atypical communities prompted us to
allocate them to CST-IV (mixture of facultative anaerobes with a moderate abundance of G. vaginalis). We took this approach to help in differentiating the atypical group dominant samples from other Lactobacillus dominated CST types commonly associated with vaginal microbial health. Although amplicon sequencing generates considerably lower taxonomic resolution than SSM, we believe it remains helpful as a tool for vaginal microbiome characterisation because it can broadly differentiate between CST types on Lactobacillus spp. dominance. Nonetheless, this comparison highlights the limitations of using amplicon sequencing in accurately distinguishing between closely related CST profiles such as those dominated by the Lactobacillus genus.

In conclusion, our pilot study suggests that shallow shotgun metagenomics is a superior method compared to amplicon sequencing in the context of species-level vaginal microbiome characterisation related to health and disease. Importantly, while standard (deep) metagenomics is cost-prohibitive for large studies, in this pilot study we show that the benefits associated with sequencing all DNA in a sample can be achieved at costs similar to amplicon sequencing. Our study also suggests that the vaginal microbiome data and CST demographics generated by high-resolution shotgun metagenomics may need to be re-examined in the context of microbial health and disease risk. Our follow-up work intends to improve our microbiome data accuracy and confidence by complementing shallow metagenomics laboratory workflow with a site-specific, multi-kingdom reference database combined with alternative bioinformatics algorithms.

**Conflicts of interest**

The authors declare no conflicts of interest.

**Declaration of funding**

This research was funded by grants from Channel 7 Telethon Trust (WA Pregnancy Biobank), Telethon-Perth Children’s Hospital Research Fund and Western Australia Department of Health (WA Human Microbiome Collaborative Centre).

**References**


**In Focus**

![PCoA ordination using Bray–Curtis distance](image)


Berman, H.L. et al. (2020) Understanding and interpreting community sequencing measurements of the vaginal microbiome. BJOG 127, 139–146. doi:10.1111/1471-0528.15978


Biographies

Aliashum Ali is a Research Officer at Fiona Stanley Hospital Department of Neonatology and PhD student at Curtin University, with over 10 years’ experience in molecular diagnostics. His PhD project aims to shed light on the links between extreme prematurity and vaginal microbiome and inflammation through the application of precision medicine principles and multi-omic analytical strategies. Ultimately, his aim is to identify predictive biomarkers to help improve pregnancy outcomes. Ali’s motivation stems from his personal experience following the premature birth of his son.

Dr Claus T Christophersen, PhD, MSc, is a molecular microbiologist specialising in the role and impact of the microbiome in health. He participates in multidisciplinary research to understand how the microbiome interacts with the host and how to manipulate it to can improve health or prevent diseases. He has an MSc from the University of Copenhagen and a PhD from the University of Western Australia. He then undertook a post-doctoral appointment and later became a research scientist and team leader in CSIRO Food and Nutrition before returning to Perth and taking up positions at ECU and Curtin University. He leads the WA Human Microbiome Collaboration Centre (WAHMCC) at Curtin University.

Jeffrey Keelan is Head of the School of Biomedical Sciences and Professor in the Division of Obstetrics and Gynaecology, University of Western Australia. He has 38 years’ experience in biomedical research and medical science and has worked and published in the areas of obstetrics, reproductive biology, endocrinology, pharmacology, toxicology, immunology, microbiology and nanomedicine. His current research is focused on the pharmacological treatment of intraamniotic infection/inflammation for the prevention of preterm birth; placental health and dysfunction; nanoparticle-based drug delivery in pregnancy; the intrauterine microbial and endocrine environment; and the role of the microbiome in pregnancy, parturition and early life.