scientific reports

Check for updates

OPEN

Evaluation of microbiome enrichment and host DNA depletion in human vaginal samples using Oxford Nanopore's adaptive sequencing

Mike Marquet[1]✉, Janine Zöllkau[6,7], Jana Pastuschek[6,7], Adrian Viehweger[2], Ekkehard Schleußner[6,7], Oliwia Makarewicz[1,4], Mathias W. Pletz[1,4], Ralf Ehricht[3,4,5] & Christian Brandt[1,4]

Metagenomic sequencing is promising for clinical applications to study microbial composition concerning disease or patient outcomes. Alterations of the vaginal microbiome are associated with adverse pregnancy outcomes, like preterm premature rupture of membranes and preterm birth. Methodologically these samples often have to deal with low relative amounts of prokaryotic DNA and high amounts of host DNA (> 90%), decreasing the overall microbial resolution. Nanopore's adaptive sampling method offers selective DNA depletion or target enrichment to directly reject or accept DNA molecules during sequencing without specialized sample preparation. Here, we demonstrate how selective 'human host depletion' resulted in a 1.70 fold (± 0.27 fold) increase in total sequencing depth, providing higher taxonomic profiling sensitivity. At the same time, the microbial composition remains consistent with the control experiments. The complete removal of all human host sequences is not yet possible and should be considered as an ethical approval statement might still be necessary. Adaptive sampling increased microbial sequencing yield in all 15 sequenced clinical routine vaginal samples, making it a valuable tool for clinical surveillance and medical-based research, which can be used in addition to other host depletion methods before sequencing.

Long read sequencing technologies, such as nanopore sequencing, allows for fast, real-time, and culture-free metagenomic sequencing[1]. However, in samples containing relatively high proportions of host DNA (> 90%) like saliva, throat, buccal mucosa, and vaginal samples, the detection of low abundant species is expected to be impaired[2]. Host DNA depletion prior to sequencing by selective lysis of host and microbial cells or selective removal of CpG-methylated host DNA are complex wet-lab procedures[3]. For instance, host DNA depletion via, e.g., saponin or the "MolYsis Complete5" kit improves the sensitivity of pathogen detection after sequencing[4–6].

Oxford Nanopore Technologies (ONT) recently (November 2020) introduced target enrichment or depletion of unwanted DNA molecules directly during sequencing by simply providing any target DNA fasta sequence. While a DNA molecule is sequenced in the nanopore, the data is already compared live with references to decide whether the DNA molecule should be sequenced further (accepted or no decision yet) or removed directly from the pore (rejected). Each pore is individually addressable and can reverse the voltage on its pore to reject DNA molecules and sequence another one instead, increasing the sequencing capacity for molecules of interest[7]. The main advantage is that this depletion or enrichment method can be combined in addition to wet-lab depletion or enrichment methods and does not prolong the overall sequencing run time.

[1]Institute for Infectious Diseases and Infection Control, Jena University Hospital, Jena, Germany. [2]Department of Medical Microbiology and Virology, University Hospital Leipzig, Leipzig, Germany. [3]Leibniz Institute of Photonic Technology (IPHT), Jena, Germany. [4]InfectoGnostics Research Campus, Jena, Germany. [5]Institute of Physical Chemistry, Friedrich-Schiller-University Jena, Jena, Germany. [6]Department of Obstetrics, Jena University Hospital, Jena, Germany. [7]Center for Sepsis Control and Care (CSCC), Jena University Hospital, Jena, Germany. ✉email: mike.marquet@med.uni-jena.de

The host depletion by adaptive sequencing and thus enriching microbial organisms is likely to be significant for clinical samples, particularly if high levels of human DNA is expected (e.g., vaginal microbiome samples). The vaginal microbiome is characterized by low-alpha diversity (number of taxonomic groups) with a high relative abundance of *Lactobacillus* species. *Lactobacilli* promote vaginal and reproductive health producing specific metabolites (e.g., lactic acid, hydrogen peroxide, or bacteriocins) that inhibit colonization of the vaginal microenvironment by harmful microbiota[8]. Disruption or imbalance of the composition of vaginal microbiome during pregnancy (e.g., by antibiotic treatment) can result in complications such as preterm premature rupture of membranes (PPROM), cervical insufficiency, pregnancy loss, or preterm birth[9]. The latter is associated with early-onset neonatal sepsis (EONS) and risk of neonatal morbidity, mortality and may lead to long-term complications and deficits for the newborn[10–12]. Increased organism diversity and relative abundance of organisms like Group B streptococci, *Escherichia coli*, *Pseudomonas aeruginosa*, or *Ureaplasma parvum* and the depletion of certain *Lactobacilli* are indicators for bacterial vaginosis[8,13]. However, their identification via conventional culture-based, microbiological diagnostic techniques suffers from long reporting delays and low sensitivity and specificity[14,15]. On the other hand, pathogen identification based on 16S ribosomal RNA gene sequencing gives insights into the metagenomic composition with a lower resolution than metagenomics and suffers from various inherent biases to interpret abundance data properly: e.g., primer mismatches, different gene copy numbers, recombinations, sequence- and primer-dependent polymerase efficiency, or choice of hypervariable regions[16–19]. Nanopore sequencing has become a widely used method for metagenomic sequencing with similar taxonomic classification performance as short read sequencer (e.g., Illumina)[20–22]. The lower raw read accuracy of 97.8% is compensated via the higher information richness due to the longer reads. In the present study, we evaluated the overall performance of adaptive sampling to enrich bacterial reads and deplete human DNA in 15 human vaginal samples. The aim was to evaluate the strength and limitations of this technology and its potential metagenomics-based clinical diagnostic routines in culture-free metagenomic samples.

Results and discussion

Vaginal samples from swabs mostly yield small amounts of DNA (< 40 ng/µl) for library preparation and subsequent sequencing, making PCR amplification often mandatory, which can introduce amplification bias and alter the microbial composition.

Influence of nonspecific amplification-based library preparation for the determination of microbial communities. The nanopore amplification-based library preparation kit (RPB004) uses transposase-mediated cleaving of DNA molecules to attach the primer binding sites for PCR amplification, which should reduce PCR amplification bias.

We initially assessed this bias by determining the microbial composition of the ZymoBIOMICS Microbial Community Standard[23] (control) by sequencing using the RPB004 PCR-based library preparation kit and compared the abundance of the different species to the native PCR-free library preparation kit (LSK109). DNA of the mock community cells was isolated simultaneously in three replicates to address experimental variations. Each replicate was sequenced with the LSK109 and the RPB004 kit. Accordingly, all samples have the same "lysis and DNA isolation" bias; therefore, the library preparation kits are the only parameter that differentiates the sequenced samples.

The reads were mapped against the microbial genomes via minimap2 v.2.19[24], counted via samtools depth v1.11[25] (bases sequenced per organism), and summarized via ggplot2 (Fig. 1). All reads could be mapped to the reference genomes of the mock community.

Both sequencing kits detected all ten organisms of the ZymoBIOMICS Microbial Community Standard and the results of the sample's replicates exhibited only negligible deviation.

Compared to the expected abundance in the control, Gram-positive bacteria and yeast were underrepresented in the sequencing data obtained by both library preparation methods (amplification free: average 0.60 fold, min 0.34 fold, max 0.79 fold; amplification-based: average 0.70 fold, min 0.35 fold, max 0.97 fold), while Gram-negative bacteria were overrepresented (amplification free: average 1.84 fold, min 1.80 fold, max 1.88 fold; amplification-based: average 1.64 fold, min: 1.01 fold, max: 1.99 fold). The amplification-based library preparation approach shows a considerable difference to the PCR-free library preparation method for *Pseudomonas aeruginosa* (0.55 fold of PCR-based), *Lactobacillus fermentum* (0.47 fold of PCR-based), *Staphylococcus aureus* (2.21 fold of PCR-based), and *Cryptococcus neoformans* (1.57 fold of PCR-based). Six organisms show minor differences to the PCR-free library preparation *(Bacillus subtilis*, *Enterococcus faecalis*, *Escherichia coli*, *Listeria monocytogenes*, *Saccharomyces cerevisiae*, *Salmonella enterica*).

We expected the PCR-free library preparation approach to represent the microbial community standard more accurately since it was previously validated by other groups[26]. However, it showed clear variation in abundances compared to the control, which might be attributed to the different cell disruption device used in this work. The description of the ZymoBIOMICS Microbial Community Standard states that it mimics a mixed microbial community of well-defined composition, containing three easy-to-lyse Gram-negative bacteria, five tough-to-lyse Gram-positive bacteria, and two tough-to-lyse yeasts. Thus, Gram-positive bacteria and yeast were underrepresented due to differential lysis rather than differences in the library preparation protocols. We did not observe a significant advantage or disadvantage in choosing the amplification-based library preparation method over the PCR-free library preparation method to assess the microbial composition as both similarly overrepresent Gram-negative bacteria (Fig. 1), but interestingly the PCR amplification-based library preparation seemed to represent the control slightly better than the PCR-free library preparation.

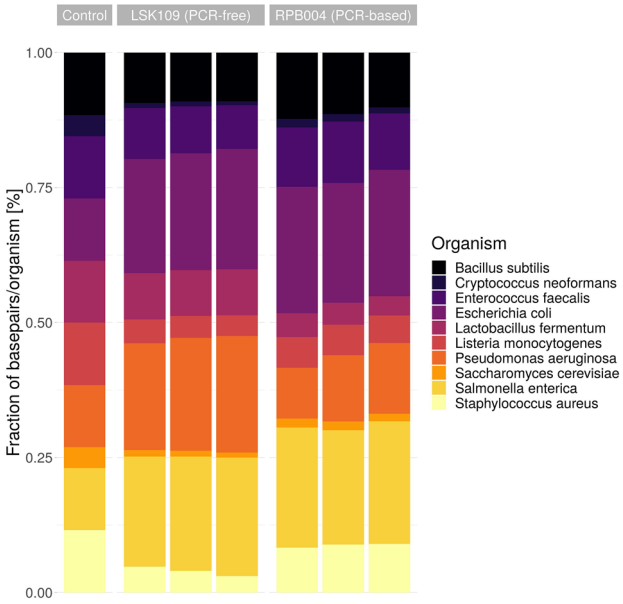

Figure 1. Abundance of the ten sequenced organisms of the ZymoBIOMICS Microbial Community Standard for the native PCR-free library preparation (LSK109) and the nanopore amplification-based library preparation (RPB004). The expected fraction for the microbial standard is shown on the left (control). The fraction of sequenced bases was determined by mapping the sequenced reads against the ten organisms via minimap2.

Adaptive sampling for metagenomes: enrichment or depletion? ONT's adaptive sampling method enables it to either enrich or deplete DNA. To evaluate if microbial target enrichment or host depletion is more suitable for human vaginal metagenome sequencing (providing more microbial sequencing data), we compared both methods against a control experiment without adaptive sampling.

We sequenced a human vaginal metagenome (87.93% human host contamination) from a pregnant woman to derive species information for the enrichment process first. In a second step, we performed a depletion experiment using a human genome as reference (GCF_000001405.39). Finally, we performed an enrichment experiment using nine bacterial genomes downloaded from NCBI as reference based on the most abundant identified species from the first control sequencing experiment (see "Methods" section: "Nanopore sequencing").

Each read passing the nanopore during adaptive sampling was mapped against a single or multiple reference genome(s) (e.g., human reference genome or multiple bacterial genomes) while sequencing. The mapping occurred in intervals of several bases, and three types of decisions were made: (1) 'no_decision'—the read has been continued and mapped against the reference(s) after several bases again ('no decision'), (2) 'stop_receiving'—the read was accepted and fully sequenced ('accepted'), (3) 'unblock'—the sequencing was immediately stopped and the read was rejected by reversing of the voltage ('rejected'). The base pairs required until a decision has been made were summarised in Fig. 2 B for all reads. For both methods, read rejections occurred within approx. 400–800 bp. Accepting reads started at approx. 400 bp or 4000 bp for the enrichment or depletion protocols, respectively. More generally, read lengths of at least 400 bp were required for both adaptive sampling methods to start the individual reads' decision-making process.

The enrichment experiment yielded a higher total reads' number (5.67 million, 1.50 fold more than depletion, of which 5.44 million were rejected reads), followed by the depletion experiment (3.79 million reads, 1.39 fold more than the control, of which 3.07 million reads were rejected). The control yielded 2.73 million reads. One should note that experimental variations affect the total sequencing performance, but the yield increase via depletion of human sequences was further validated (see "Performance of human host depletion via adaptive sampling in human vaginal metagenomic samples").

Due to short read lengths, which result from the high rejection rate and the fast decision process (Fig. 2B.2), the enrichment experiment yielded the least amount of total bases and microbial bases while 'human depletion' yields the most microbial bases (Table 1). Without adaptive sampling, the proportion of sequenced human reads was unsurprisingly highest (87.93%) but could be strongly reduced by the depletion approach to 34.73% and by the bacterial enrichment down to 8.29% (Fig. 2A). The 'human depletion' method rejected almost 81.01% of all reads, which was lower than the total abundance of human DNA in the control experiment, suggesting that the chosen human genome might be insufficient for a complete depletion of all human reads or the adaptive sampling process itself is prone to error. The bacterial enrichment method rejected 95.93% of all reads, which indicates that some bacterial reads were also rejected. We identified 5.48% of *Gardnerella* reads, 2.41% of *Lactobacillus* reads, and 2.20% of other microbial reads in the 'rejected' fraction of the bacterial enrichment experiment. Simultaneously, the proportions of essential vaginal microorganisms, like *Lactobacillus* and *Gardnerella,* could be increased by both methods but higher by the enrichment protocol.

We compared the proportions of bacterial genera of the experiments identified from the reads of the 'accepted' and 'no decision' category to validate whether the overall microbial composition was retained despite adaptive

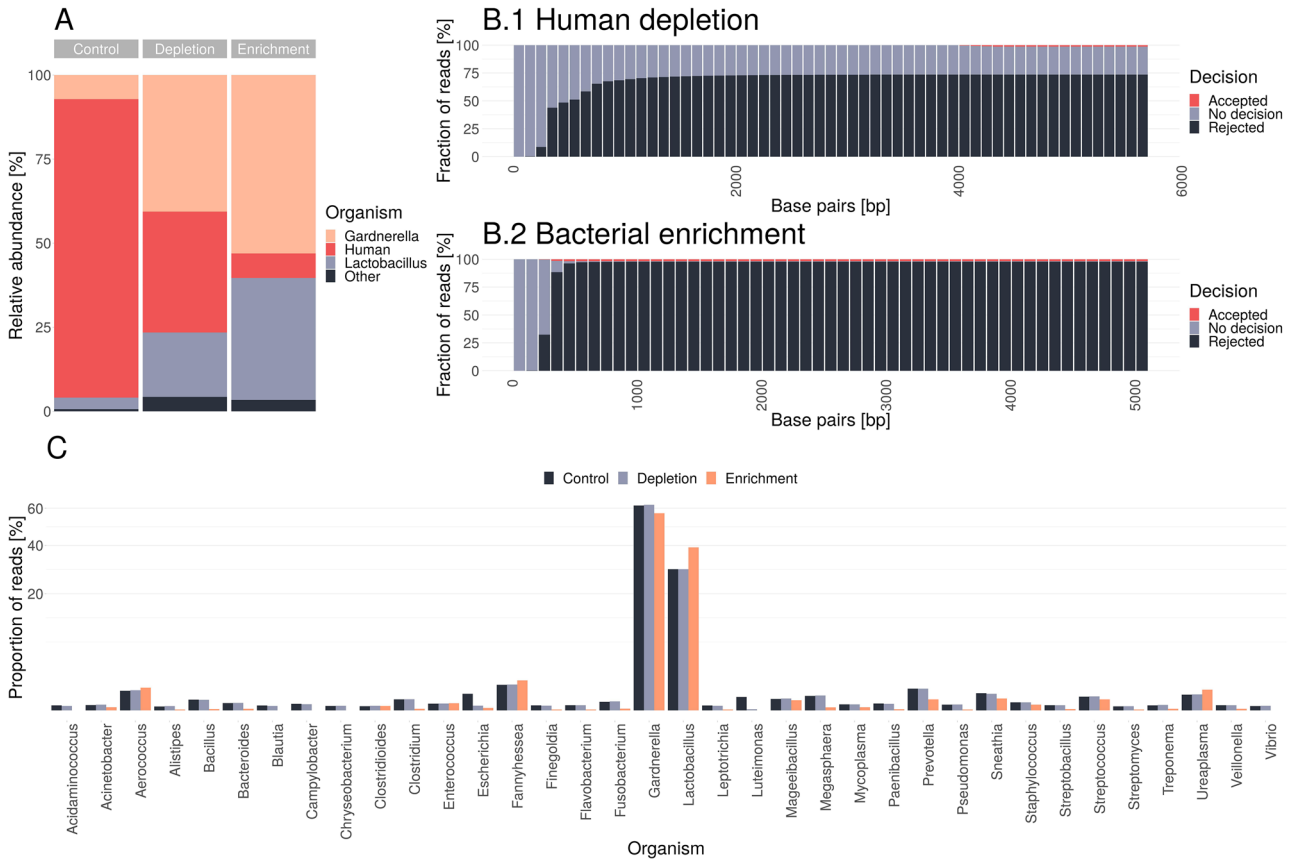

Figure 2. (**A**) Proportions of sequenced human, Gardnerella and Lactobacillus reads for the control, depletion and enrichment sequencing experiments using the 'accepted' and 'no decision' fractions for the adaptive sampling experiments only. Reads were taxonomically classified with centrifuge v1.0.4. (**B**) Base pairs required until a decision has been made in the depletion (**B.1**) and enrichment (**B.2**) experiment for all sequenced reads. (**C**) Proportion of sequenced genera for each of the experiments.

Experiment	Total bases (Gigabases)	Total reads (million)	Rejected reads (million)	Rejected reads (%)	microbial bases (Gigabases)
Control	9.91	2.73	None	0	1.06
Depletion	5.2	3.79	3.07	81.01	1.43
Enrichment	4.26	5.67	5.44	95.93	0.77

Table 1. Summary of sequencing yield generated for the control, depletion and enrichment experiment using the PCR-based (RPB004) library preparation kit with a median read length of 2500 bp due to the amplification step. The calculated microbial bases are based on the 'accepted' & 'no_decision' fractions.

sampling (Fig. 2C). The human depletion method clearly showed very similar proportions to the control for 34 of 36 microorganisms except for *Escherichia* and *Luteimonas*. The difference in the proportions of these two organisms could be attributed to experimental variations, especially since their frequency in the control experiment was only 0.05% (*Escherichia*) and 0.03% (*Luteimonas*). Conversely, the enrichment method shows significant differences in most genera, including important vaginal microorganisms like *Gardnerella*, *Lactobacillus*, and *Ureaplasma*.

Therefore, we assume that an enrichment approach might be unsuitable for investigating the microbial composition between metagenomic samples if not all species can be reliably provided as target sequences during the enrichment. On the other hand, the 'human depletion' experiments maintained a comparable microorganism composition as the control experiments and considerably (53.20%) reduced the number of human reads, making it a robust choice for clinical metagenomic samples with high amounts of human host DNA.

Performance of human host depletion via adaptive sampling in human vaginal metagenomic samples. We collected 15 vaginal samples of pregnant women (see "Methods" section: "Sample selection") with high proportions of host DNA (>90%).

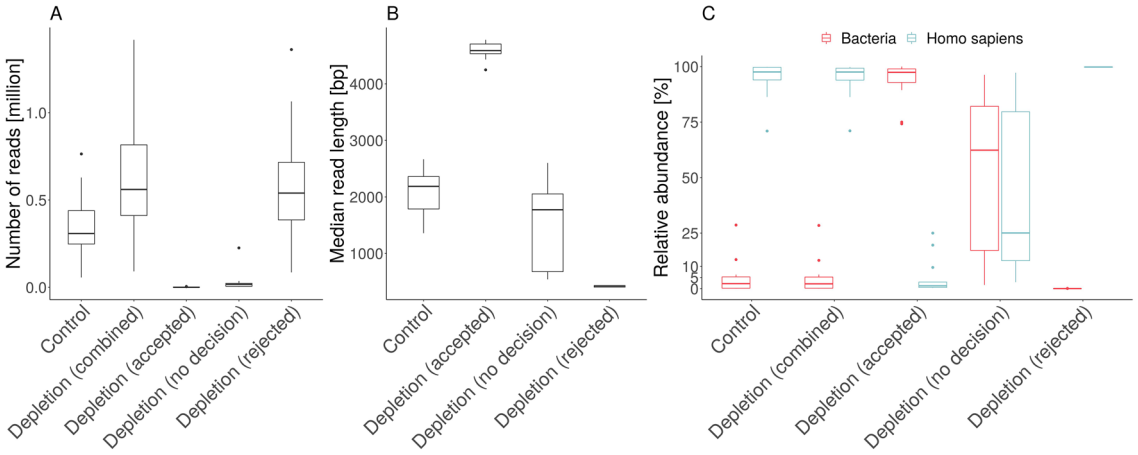

Figure 3. Overall performance of human depletion experiments compared to control experiments for 15 vaginal metagenomic samples. (**A**) Total number of sequenced reads for depletion and control experiments. Depletion experiments were additionally split into the three decision categories: 'rejected', 'no decision', and 'accepted'. (**B**) Median read length distribution of the three depletion decision categories compared to the control experiments. (**C**) Human (blue) and bacterial (red) proportions for each sample of the control and depletion experiments. Depletion experiments were additionally split into the three decision categories: 'rejected', 'no decision', and 'accepted'.

First, each of the 15 samples were sequenced without adaptive sampling serving as a control experiment and ground truth of their metagenomic composition to track possible changes introduced via adaptive sampling. We then sequenced the same isolated DNA from the control experiments while using adaptive sampling (human DNA depletion) and compared the overall sequencing performance to the previously sequenced controls (Fig. 3). Additionally, a negative control of a swab without patient material following the same sample gathering and sequencing approach yielded 126 reads (ranging from 10 to 4000 bp) but none of the reads were classifiable and might be attributed to some PCR-primer and sequencing adapter constructs.

All reads were taxonomically classified via centrifuge v1.0.4[27] to investigate their taxonomic composition (centrifuge database: Human-Virus-Bacteria-Archaea 01.2021)[28]. 99.43–99.85% of the reads generated in the control experiments could be taxonomically classified, while 97.44–99.83% of the reads in the depletion experiments could be taxonomically assigned.

On average, depletion experiments yielded 1.71 fold (± 0.27 fold) more reads (including rejected reads) than the corresponding control experiments (Fig. 3A). This corresponds to a yield of ~ 1.7 flow cells from a standard Nanopore sequencing experiment, with the only difference being that unwanted DNA molecules (human) were only partially sequenced. Reads of the categories 'no decision' (average: 5.38%, ± 7.24%) and 'accepted' (average: 0.23%, ± 0.35%) contributed a small overall proportion of all sequenced reads due to the high amount of human DNA. On average, adaptive sampling categorized 92.05% (± 7.42%) of reads as 'rejected'. 'Accepted' reads were comparably long (Fig. 3C) with a median read length of ~ 4000 bp, which is in line with previous results (Fig. 2B).

Samples of the control experiments contained 97.59% (± 7.63%) human reads, while bacteria reads made up to 2.22% (± 7.49%) (Fig. 3C). Without splitting into the three read categories, the depletion runs showed similar proportions of species like the control. The 'rejected' fraction of the depletion experiments contained 99.80% (± 0.09%) human reads and 0.02% bacterial reads (± 0.02%), indicating a very selective depletion process. Reads of the category 'accepted' contained low amounts of human reads 1.23% (± 8.13%) and 97.42% (± 8.38%) bacterial reads. Three samples contained 4, 20, and 92 reads only within the 'accepted'-fractions and were excluded in the previous calculation. The 'no decision'-fractions contained 62.37% bacterial reads (± 34.76%), but also 25.06% human reads (± 36.19%).

In summary, most of the not 'rejected' reads were placed into the 'no decision' category as the 'accepted' decision was rarely made. The 'no decision' category also included many human reads and most of the bacterial reads. Combining the 'accepted' and 'no decision' fractions of the bacterial reads generally yielded more microbial reads compared to the control experiment underlining the capabilities of adaptive sampling to increase the sequencing depth (Fig. 3C). Furthermore, the decision made for the 'rejected' category was remarkably accurate as it contained almost exclusively human reads. However, those human reads were still identified in the 'accepted' and 'no decision' fractions as observed in previous experiments. Thus, to reliably remove all human reads during sequencing seems rather elusive.

Depletion does not alter the species distribution in samples. Adaptive sampling selectively depletes human reads while simultaneously enriching microbial reads due to the increased sequencing depth but may alter species representation and metagenomic composition. We, therefore, compared the taxonomically classified reads of the 15 vaginal metagenomic samples in detail, to compare the individual abundance between control and the host depletion experiments (Fig. 4, Supplementary Figures S1, S2).

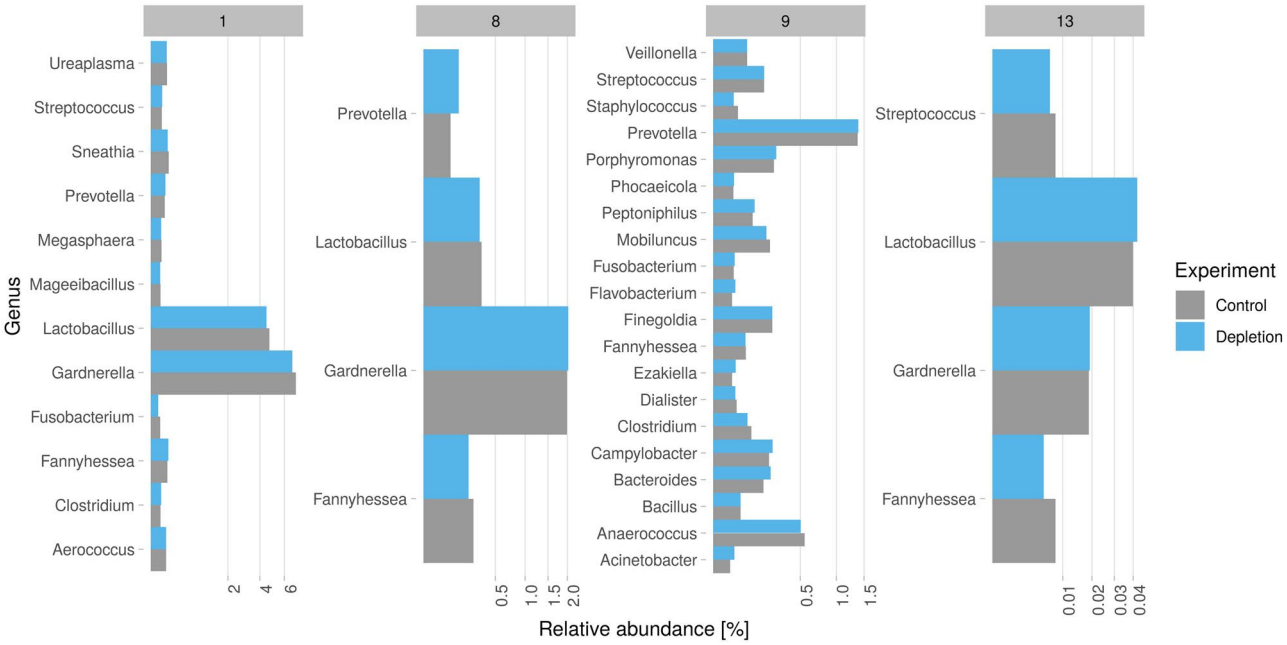

Figure 4. Comparison of the relative abundance of bacterial genera for four of the 15 vaginal samples (control in grey, depletion in blue). The full overview is shown in Supplementary Figure S1 for genus and Supplementary Figure S2 for species classifications.

The proportion of each genus was calculated in relation to the total amount of reads generated for each sample, including 'rejected' reads for the depletion experiment. We only included genera with at least 30 reads to avoid over-interpreting uncertain taxonomic classifications. This resulted on average in five bacterial genera (min 1, max 20) and in six bacterial species (min 1, max 26; Supplementary Figure S2), which correspond well with the expected vaginal microbiome[29]. Across all 15 samples, the bacterial proportion varied from the corresponding control experiments on average by 1.03 fold, indicating a similar representation of genus abundance levels by the depletion experiment (Fig. 4). Low abundance genera with reads counts between > 30 and < 100 showed higher discrepancies (± 0.23 fold). Higher read counts showed higher reliability (e.g. > 500 < 1000 reads (± 0.03 fold) and > 1000 reads (± 0.04 fold). This higher discrepancy in genera with a small number of reads is expected, as the experimental variability's influence is more prominent. We did not detect any organisms that were found solely by only one method using the applied cutoff. Overall, both experiment groups performed highly similarly in relation to the detected genus abundance levels between the control and depletion experiments.

In addition to the abundance comparison, we assessed the Bray–Curtis-Dissimilarity and the Spearman correlation for a more robust statistical analysis. Low Bray–Curtis dissimilarity values for all samples (0.005–0.077) indicate high similarity between each pair of metagenomes (control and depletion), underlining previous findings using only the abundance comparison (Table 2). The Spearman correlation test resulted in statistical significance (2.02E−05 and 1.92E−06) and positive coefficients (rho: 0.98 and 0.93) for samples 1 and 9 (Table 2). In other words: if, e.g. Lactobacillus is highly abundant in the control of sample 1, it is also highly abundant in the corresponding depletion experiment. Samples 3, 10 and 15 were not included in the Spearman correlation calculation since only one pairwise case (organism) occurs in these samples. The P value indicates no statistical significance for the remaining samples, probably due to few pairwise cases (< 10 pairwise cases).

Conclusion

The enrichment of targets in human cells[30–34] or species in mock communities[31,35] or fecal samples of lions[36] via adaptive sampling was previously demonstrated. In the present study, we used clinical metagenomic samples obtained from vaginal swabs of pregnant women to evaluate the performance to deplete the high content of human DNA that heavily impairs downstream microbiome analyses. Our results demonstrated that ONT's unique adaptive sequencing feature has reliably increased the overall sequencing depth of bacterial sequences in clinical metagenomic samples without changing the microbial composition when providing a human reference genome to deplete the human DNA during sequencing. However, the enrichment experiment showed significantly higher 'human depletion', but changed the overall identified bacterial composition by also depleting other microbial sequences, illustrating that the enrichment method may be poorly suited for some microbiome studies.

Currently, to increase the sequencing depth for metagenomic samples with low microbial material, several sequencing runs per sample or sequencers with higher throughput (e.g., PromethION in case of ONT) are necessary, besides wet laboratory methods to deplete the host DNA. ONT's adaptive sampling method demonstrated in our work a 1.7 fold increase in sequencing depth for samples with high human DNA contamination, which increases sensitivity of taxonomic profiling by providing more sequencing data[2]. Moreover, adaptive sampling can be used in addition to wet lab procedures to increase sensitivity further[37]. This makes molecular monitoring

Sample	Spearman correlation		Bray–Curtis dissimilarity
	p value	rho	
1	2.02E−05	0.98	0.02452134
2	1	1	0.06627899
3	–	–	0.05145795
4	0.53	0.87	0.02238958
5	0.27	0.96	0.005473575
6	1	1	0.02300497
7	0.33	1	0.01389923
8	0.83	1	0.02963051
9	1.92E−06	0.93	0.03501389
10	–	–	0.02023301
11	0.83	1	0.04614055
12	0.45	0.5	0.02863529
13	0.513	0.94	0.04716042
14	0.83	1	0.03748569
15	–	–	0.07734091

Table 2. Calculated Bray–Curtis dissimilarity and Spearman correlation for each pair of metagenomes (control and depletion).

of human reservoirs with low microbial concentrations and high host DNA loads (e.g., nasal swabs, sputum, or skin swabs) more feasible. In this manner, the adaptive sampling improved the detected organisms beneficial to the vaginal microbiota, such as *Lactobacillus,* and pathogenic microbiota, such as *Streptococcus, Gardnerella*, or *Ureaplasma*, potentially harmful in pregnancy. Therapeutic measures can be derived based on the presence or ratio of certain species improving vaginal microecological diagnostics in the future, which enables clinically relevant insights into eubiosis or dysbiosis during pregnancy. On a side note, due to the increased sequencing depth via adaptive sampling, more samples can be barcoded and sequenced simultaneously, reducing the total cost per sample in a diagnostic laboratory.

However, human DNA could not be completely removed, so that the raw data always contained human sequences, which poses an ethical problem. Although the ONT adaptive sequencing could still be used for diagnostic purposes, patient consent is required for scientific purposes or data upload to public repositories like the National Center for Biotechnology Innovation (NCBI) or the European Nucleotide Archive (ENA). In turn, removing human sequences *by* different bioinformatics approaches poses hurdles for institutions and hospitals without adequate bioinformatics support. However, the research field of nanopore sequencing is evolving rapidly and dynamically, and ONT may address current limitations in the foreseeable future. In addition, continuous improvements in raw data accuracy and new chemistries mean that less sequencing depth is needed for reliable results[38].

The present work can help to decide which adaptive sampling approach is best suited for analyzing specific clinical samples and questions. For example, when sequencing metagenomes with unknown microbial composition and host contamination, the depletion method is a better choice. On the other hand, the enrichment method might be helpful for metagenomics if only certain bacterial species are of interest as the higher rejection rates increase the sequencing depth further. Combining the real-time sequencing data stream of ONT with automated analysis pipelines, the turnaround time from sample collection to analysis and an appropriate treatment strategy can be reduced.

A few limitations must be noted. Our results were based on the library preparation with PCR amplification, for which we did not observe a noticeable bias when sequencing the microbial community standard. Still, these results might be different in other specimens. Due to the DNA isolation via bead beating and the PCR-based library preparation, we sequenced short DNA fragments of approximately 2500 bp only. Longer DNA fragments might improve the adaptive sampling's decision-making, further increasing overall sequencing depth. Furthermore, the user should carefully select the provided reference genome(s) during adaptive sampling as, e.g., another human reference might slightly improve or worsen the overall depletion performance. Finally, raw read accuracy is currently at around 97.5% and might impact the read to reference mapping during sequencing and thus the adaptive sampling accuracy.

We strongly believe that adaptive sampling will prove exceptionally useful within clinical research and the individual microbiological and microbiological diagnostic approach in routine diagnostics. The increased information depth for compartment-specific human microbiomes in a physiologic and pathophysiologic context may change paradigms of antiinfective therapies in a personalized risk stratifying manner.

Methods

All methods were performed in accordance with the relevant guidelines and regulations.

Sample selectionn. Vaginal swabs of adult pregnant women were admitted to the hospital with PPROM (hospitalization between 22 + 0 and 34 + 0 weeks) and collected with sterile FLOQSwabs (Copan, Italy) within the PEONS trial (ClinicalTrials.gov NCT03819192) between September 2019 and March 2020. The institutional review board approved the study, and all participants signed a written informed consent. Swabs were immediately frozen at − 80 °C until analysis was performed.

DNA extraction. DNA was extracted from 75 µl ZymoBIOMICS Microbial Community Standard (Zymo Research Corporation, Irvine, CA, USA. Product D6300, Lot ZRC190633) using the ZymoBIOMICS DNA Miniprep extraction kit according to the manufacturer's instructions. Similarly, the DNA from vaginal swabs was extracted using the ZymoBIOMICS DNA Miniprep extraction kit. The cell disruption was conducted for five minutes with the Speedmill Plus (Analytik Jena, Germany).

Library preparation. DNA quantification steps were performed using the dsDNA HS assay for Qubit (Invitrogen, US). DNA of the microbial community standard was size-selected by cleaning up with 0.45 × volume of Ampure XP buffer (Beckman Coulter, Brea, CA, USA) and eluted in 50 µl EB buffer (Qiagen, Hilden, Germany). The library was prepared from 1 µg input DNA using the SQK-LSK109 kit (Oxford Nanopore Technologies, Oxford, UK) and 5 ng using the SQK-RPB004 kit (Oxford Nanopore Technologies, Oxford, UK), according to the manufacturer's protocol.

The sequencing library of clinical vaginal samples was prepared from 5 ng input DNA using the SQK-RPB004 kit (Oxford Nanopore Technologies, Oxford, UK), according to the manufacturer's protocol.

Nanopore sequencing. The microbial community standard was sequenced on the GridION using FLO-MIN106D Flow cells and the minknow-core-gridiron:4.1.2 software (all Oxford Nanopore Technologies). The Standard 48-h script with active channel selection was applied.

Vaginal samples were sequenced by a standard 72-h run script with and without adaptive sampling (depletion/enrichment). For 'human depletion experiments', the human reference genome GCA_000001405.28_GRCh38.p13 was used. A fasta reference file containing eight different bacteria species (*Aerococcus christensenii* (NZ_CP014159.1), *Fannyhessea vaginae* (NZ_UFSV01000001.1), *Gardnerella vaginalis* (NZ_PKJK01000001.1), *Lactobacillus iners* (NZ_AEKI01000028.1), *Mageeibacillus indolicus* (NC_013895.2), *Prevotella intermedia* (NZ_CP024727.1), *Prevotella jejuni* (NZ_CP023863.1), and *Ureaplasma parvum* (NC_010503.1)) was used for enrichment experiments. We performed a 'Flow cell refuel' step after approx. 18–20 h of runtime by adding 70 µl of a 1:1 water-SQB buffer (Oxford Nanopore Technologies) to the flow cell SpotON port.

Nanopore basecalling. Reads of the microbial community standard were basecalled using Guppy v4.2.2 GPU basecaller (Oxford Nanopore Technologies) during sequencing using the high accuracy basecalling model.

Bioinformatics analysis. Sequencing data of the microbial community standard were mapped using minimap2 v.2.19 against the reference organisms' sequences provided by Zymobiomics[23]. Subsequently, the base pairs sequenced per organism and per sequencing method were compared to each other.

The basecalled reads generated during adaptive sampling were extracted from the fastq files via samtools v1.19 and split based on the read_until.csv-file into three categories: 'accepted', 'rejected' and 'no decision'. The reads were taxonomically classified via centrifuge v1.0.4 (updated centrifuge database h.v.b.a 01.2021[28]).

Ethical approval and patient consent. This study was approved by ethical committees at University Hospital Jena (No. 2018-1183), University Hospital Halle/Saale (No. 2019-012), and University Hospital Rostock (No. A 2019-0055). Eligible women, who participated in the study, were informed about the study, applied procedures, and any risks due to sampling by a physician and gave their written consent. Participants were also informed that they could withdraw from the study at any time.

Data availability

The centrifuge database available at: https://osf.io/5zv8t/. The read data without human reads is available at the following Bioproject accession number: PRJNA799199 (SRA Accession numbers: SRR17688764-SRR17688740) .

Received: 14 October 2021; Accepted: 1 March 2022
Published online: 07 March 2022

References

1. Brandt, C., Bongcam-Rudloff, E. & Müller, B. Abundance tracking by long-read nanopore sequencing of complex microbial communities in samples from 20 different biogas/wastewater plants. *Appl. Sci.* **10**, 7518. https://doi.org/10.3390/app10217518 (2020).
2. Pereira-Marques, J. *et al.* Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. *Front. Microbiol.* **10**, 1277. https://doi.org/10.3389/fmicb.2019.01277 (2019).
3. Host DNA depletion efficiency of microbiome DNA enrichment methods in infected tissue samples—ScienceDirect. [cited 13 Jul 2021]. https://www.sciencedirect.com/science/article/pii/S0167701219310693?via%3Dihub.
4. Charalampous, T. *et al.* Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nat. Biotechnol.* **37**, 783–792. https://doi.org/10.1038/s41587-019-0156-5 (2019).
5. Hasan, M. R. *et al.* Depletion of human DNA in spiked clinical specimens for improvement of sensitivity of pathogen detection by next-generation sequencing. *J. Clin. Microbiol.* **54**, 919–927. https://doi.org/10.1128/JCM.03050-15 (2016).
6. Yap, M. *et al.* Evaluation of methods for the reduction of contaminating host reads when performing shotgun metagenomic sequencing of the milk microbiome. *Sci. Rep.* **10**, 21665. https://doi.org/10.1038/s41598-020-78773-6 (2020).

7. Loose, M., Malla, S. & Stout, M. Real-time selective sequencing using nanopore technology. *Nat. Methods* **13**, 751–754. https://doi.org/10.1038/nmeth.3930 (2016).
8. Chee, W. J. Y., Chew, S. Y. & Than, L. T. L. Vaginal microbiota and the potential of Lactobacillus derivatives in maintaining vaginal health. *Microb. Cell Fact.* **19**, 203. https://doi.org/10.1186/s12934-020-01464-4 (2020).
9. Amir, M. *et al.* Maternal microbiome and infections in pregnancy. *Microorganisms* **8**, 1996. https://doi.org/10.3390/microorganisms8121996 (2020).
10. Anderson, P. J. *et al.* Attention problems in a representative sample of extremely preterm/extremely low birth weight children. *Dev. Neuropsychol.* **36**, 57–73. https://doi.org/10.1080/87565641.2011.540538 (2011).
11. Singh, G. K., Kenney, M. K., Ghandour, R. M., Kogan, M. D. & Lu, M. C. Mental health outcomes in US children and adolescents born prematurely or with low birthweight. *Depress. Res. Treat.* **2013**, e570743. https://doi.org/10.1155/2013/570743 (2013).
12. McCormick, M. C., Litt, J. S., Smith, V. C. & Zupancic, J. A. F. Prematurity: An overview and public health implications. *Annu. Rev. Public Health* **32**, 367–379. https://doi.org/10.1146/annurev-publhealth-090810-182459 (2011).
13. Simonsen, K. A., Anderson-Berry, A. L., Delair, S. F. & Davies, H. D. Early-onset neonatal sepsis. *Clin. Microbiol. Rev.* **27**, 21–47. https://doi.org/10.1128/CMR.00031-13 (2014).
14. Ng, P. C. & Lam, H. S. Diagnostic markers for neonatal sepsis. *Curr. Opin. Pediatr.* **18**, 125–131. https://doi.org/10.1097/01.mop.0000193293.87022.4c (2006).
15. Russell, A. R. B. & Kumar, R. Early onset neonatal sepsis: Diagnostic dilemmas and practical management. *Arch. Dis. Child. Fetal. Neonatal. Ed.* **100**, F350–F354. https://doi.org/10.1136/archdischild-2014-306193 (2015).
16. Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41**, e1–e1. https://doi.org/10.1093/nar/gks808 (2013).
17. Suzuki, M. T. & Giovannoni, S. J. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol.* **62**, 625–630. https://doi.org/10.1128/aem.62.2.625-630.1996 (1996).
18. Acinas, S. G., Sarma-Rupavtarm, R., Klepac-Ceraj, V. & Polz, M. F. PCR-induced sequence artifacts and bias: Insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl. Environ. Microbiol.* **71**, 8966–8969. https://doi.org/10.1128/AEM.71.12.8966-8969.2005 (2005).
19. Teng, F. *et al.* Impact of DNA extraction method and targeted 16S-rRNA hypervariable region on oral microbiota profiling. *Sci. Rep.* **8**, 16321. https://doi.org/10.1038/s41598-018-34294-x (2018).
20. Pearman, W. S., Freed, N. E., & Silander, O. K. The advantages and disadvantages of short- and long-read metagenomics to infer bacterial and eukaryotic community composition. 2019. p. 650788. https://doi.org/10.1101/650788.
21. Foox, J. *et al.* Performance assessment of DNA sequencing platforms in the ABRF next-generation sequencing study. *Nat. Biotechnol.* **39**, 1129–1140. https://doi.org/10.1038/s41587-021-01049-5 (2021).
22. Somerville, V. *et al.* Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. *BMC Microbiol.* **19**, 143. https://doi.org/10.1186/s12866-019-1500-0 (2019).
23. ZymoBIOMICS Microbial Community Standard. In: Zymo Research Europe [Internet]. [cited 13 Jul 2021]. https://www.zymoresearch.de/products/zymobiomics-microbial-community-standard.
24. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191 (2018).
25. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* https://doi.org/10.1093/gigascience/giab008 (2021).
26. Nicholls, S. M., Quick, J. C., Tang, S. & Loman, N. J. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *GigaScience* https://doi.org/10.1093/gigascience/giz043 (2019).
27. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Res.* https://doi.org/10.1101/gr.210641.116 (2016).
28. Brandt, C. Centrifuge database: Human-virus-bacteria-archaea. 2021. https://osf.io/5zv8t/.
29. Ma, B., Forney, L. J. & Ravel, J. The vaginal microbiome: Rethinking health and diseases. *Annu. Rev. Microbiol.* **66**, 371–389. https://doi.org/10.1146/annurev-micro-092611-150157 (2012).
30. Payne, A. *et al.* Nanopore adaptive sequencing for mixed samples, whole exome capture and targeted panels. *bioRxiv* https://doi.org/10.1101/2020.02.03.926956 (2020).
31. Payne, A. *et al.* Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat. Biotechnol.* **39**, 442–450. https://doi.org/10.1038/s41587-020-00746-x (2021).
32. Giannuzzi, G. *et al.* Alpha satellite insertions and the evolutionary landscape of centromeres. *bioRxiv* https://doi.org/10.1101/2021.03.10.434819 (2021).
33. Miller, D. E. *et al.* Targeted long-read sequencing identifies missing disease-causing variation. *Am. J. Hum. Genet.* https://doi.org/10.1016/j.ajhg.2021.06.006 (2021).
34. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED|Nature Biotechnology. 2021. https://www.nature.com/articles/s41587-020-0731-9.
35. Martin, S. *et al.* Nanopore adaptive sampling: A tool for enrichment of low abundance species in metagenomic samples. *bioRxiv* https://doi.org/10.1101/2021.05.07.443191 (2021).
36. Wanner, N., Larsen, P. A., McLain, A. & Faulk, C. The Mitochondrial genome and epigenome of the golden lion tamarin from fecal DNA using nanopore adaptive sequencing. *bioRxiv* https://doi.org/10.1101/2021.05.27.446055 (2021).
37. Gan, M. *et al.* Combined nanopore adaptive sequencing and enzyme-based host depletion efficiently enriched microbial sequences and identified missing respiratory pathogens. *BMC Genom.* **22**, 732. https://doi.org/10.1186/s12864-021-08023-0 (2021).
38. At NCM, announcements include single-read accuracy of 99.1% on new chemistry and sequencing a record 10 Tb in a single PromethION run. In: Oxford Nanopore Technologies [Internet]. 2021. http://nanoporetech.com/about-us/news/ncm-announcements-include-single-read-accuracy-991-new-chemistry-and-sequencing.

Acknowledgements

We sincerely thank the participating pregnant women for their support of this study. We also thank Yvonne Heimann and Dr. Kristin Dawcynszki from the PEONS study team for sampling and providing the swabs.

Author contributions

Conceptualization, M.M. and C.B. Experiment conduction, M.M. Figures created by M.M. and C.B. J.Z., J.P., A.V., E.S., O.M., M.W.P., R.E. All authors actively participated in the writing and editing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. The PEONS project was funded by the Federal Ministry of Education and Research (BMBF), Germany, FKZ 01EO1502. Funding is provided by the Interdisciplinary Center of Clinical Research of the Medical Faculty Jena.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-08003-8.

Correspondence and requests for materials should be addressed to M.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.