Metagenome-validated Parallel Amplicon Sequencing and Text Mining-based Annotations for Simultaneous Profiling of Bacteria and Fungi: Vaginal Microbiome and Mycobiota in Healthy Women

Seppo Virtanen (✉ seppo.virtanen@helsinki.fi)

University of Helsinki and Helsinki University Hospital <https://orcid.org/0000-0003-0261-8958>

Schahzad Saqib

University of Helsinki: Helsingin Yliopisto

Tinja Kanerva

University of Helsinki: Helsingin Yliopisto

Pekka Nieminen

University of Helsinki and Helsinki University Hospital

Ilkka Kalliala

University of Helsinki and Helsinki University Hospital

Anne Salonen

University of Helsinki: Helsingin Yliopisto <https://orcid.org/0000-0002-6960-7447>

Research Article

Keywords: Microbiome, mycobiome, parallel sequencing, 16S rRNA gene, ITS, text mining, metagenomics, vagina

Posted Date: March 18th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-321778/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: Amplicon sequencing of kingdom-specific tags such as 16S rRNA gene for bacteria and internal transcribed spacer (ITS) region for fungi are widely used for investigating microbial populations. So far most human studies have focused on bacteria while studies on host-associated fungi in health and disease have only recently started to accumulate. To enable cost-effective parallel analysis of bacterial and fungal communities in human and environmental samples, we developed a method where 16S rRNA gene and ITS-1 amplicons were pooled together for a single Illumina MiSeq or HiSeq run and analysed after primer-based segregation. Taxonomic assignments were performed with Blast in combination with an iterative text-extraction based filtration approach, which uses extensive literature records from public databases to select the most probable hits that were further validated by shotgun metagenomic sequencing.

Results: Using 50 vaginal samples, we show that the combined run provides comparable results on bacterial composition and diversity to conventional 16S rRNA gene amplicon sequencing. The text-extraction-based taxonomic assignment guided tool provided ecosystem specific annotations that were confirmed by Metagenomic Phylogenetic Analysis (MetaPhlAn). The metagenome analysis revealed distinct functional differences between the bacterial community types while fungi were undetected, despite being identified in all samples based on ITS amplicons. Co-abundance analysis of bacteria and fungi did not show strong between-kingdom correlations within the vaginal ecosystem of healthy women.

Conclusion: Combined amplicon sequencing for bacteria and fungi provides a simple and cost-effective method for simultaneous analysis of microbiota and mycobiota within the same samples. Text extraction-based annotation tool facilitates the characterization and interpretation of defined microbial communities from rapidly accumulating sequencing and metadata readily available through public databases.

Introduction

Profiling of bacterial communities using 16S rRNA gene amplicon sequencing is the standard tool for analysing the composition and diversity of microbiota in human and environmental samples. The fungal component of the microbiota, also called mycobiota, has a long history of research in mycorrhizal fungi and other environmental samples, but has only recently gained interest as part of human commensal microbiota. Based on recent molecular studies, there are an estimated $10^{12} - 10^{13}$ fungi compared to $10^{13} - 10^{14}$ bacteria in the human microbiota, across the gastrointestinal tract, oral cavity, vaginal mucosa, and skin [1]. The most abundant fungi colonizing humans belong to phyla Ascomycota and Basidiomycota. The common fungal genera found in the human mouth, gut, skin, lungs and vagina in healthy individuals as well as in selected diseases has been extensively reviewed elsewhere [1, 2]. Fungi can interact with the host as well as the rest of the microbiota in both mutualistic and antagonistic ways, highlighting their relevance for human health [3–6]. For the vaginal ecosystem, the bacterial composition is well described. For most healthy women, it predominantly consists of *Lactobacillus* species, namely *L.*

crispatus, *L. iners*, *L. gasseri*, or *L. jensenii*, while the presence of *Gardnerella vaginalis* and an assortment of other, typically anaerobic species is indicative of dysbiosis and prevail e.g. in bacterial vaginosis (BV) [7]. The fungal component of the vaginal ecosystem is poorly characterized, as after the pioneering research of Drell et al. in 2013, where they characterized the vaginal mycobiota and microbiota of 494 women, progress has been slow [8]. Only a handful of original papers have been published on the vaginal mycobiota, yet it has been associated with various diseases, most obviously yeast infections, but also with vulvodynia, bacterial dysbiosis, and cervical dysplasia [9, 10].

Despite the clear biological relevance of fungi and interest for cross-kingdom studies, there are limited tools available for such analyses. While metagenomics, i.e. whole genome shotgun sequencing (WGS) in principle allows sequencing of entire microbial communities, in practice the very low relative abundance of fungi (e.g. in the vagina in the range of $0.17 \pm 0.04\%$, if detected at all [11] and in the gut 0.01% [12]), which is a hindrance for a cost-effective analysis. There is a rapidly growing number of studies that have processed and sequenced both 16S rRNA gene and ITS amplicons separately for parallel characterization of fungal and bacterial communities in the same samples. In biomedical research, these include studies addressing the impact of sample storage and DNA extraction [13, 14], descriptive compositional analyses of human samples in cross-sectional [8] and longitudinal settings [9], and their correlation to (disease) phenotypes [15–18].

There is an increasing demand for affordable and robust methods to study dominant members of eukaryotic microbes that constitute the human microbiome i.e. fungi, alongside bacteria. To address this, we developed a parallel amplicon sequencing technique that combines bacterial 16S rRNA gene and the fungal ITS region in a single Illumina sequencing run. This study aims to determine the feasibility and performance of this method by applying it to vaginal samples from 50 healthy women and characterizing their bacterial and fungal profiles, with shotgun metagenomics as a reference. In addition, we also attempt to assign species level annotations from amplicon sequencing data using a custom text mining-based filtration approach.

Methods

Sample collection and processing

We sampled 50 non-pregnant Caucasian women aged 25-45 years attending population-based cervical cancer screening as previously described [19]. The exclusion criteria were vaginal intercourse 48 hours prior to sampling, pregnancy, previous hysterectomy, and the inability to tell the time of last menstrual period. Detailed subject characteristics have been previously reported [19]. All samples were collected with a sterile flocked swab (FLOQSwabs, Copan spa, Italy) from the right fornix of the vagina during a speculum exam. DNA extraction was done using bead beating as described [19]. The extracted DNA was stored at -20°C .

Sequencing of 16S rRNA gene and Internal transcribed spacer-1 (ITS-1) amplicons

Data generated from an earlier study [19] that targeted the V3-V4 region of the 16S rRNA gene was used as a control dataset (PRJEB25778) to determine any data loss or variation that might be brought on by parallel sequencing. This sequencing run was done using paired end mode (PE250 kit) and Illumina HiSeq as detailed previously [19]. For the parallel sequencing approach, the 16S rRNA gene and ITS amplicons were generated separately from the previously extracted frozen DNA and pooled for barcoding.

For bacteria, we used the same 16S rRNA gene primers for V3-V4 region as before, specifically 341F 5'-CCTACGGGNGGCWGCAG-3' and 785Rev 5'-GACTACHVGGGTATCTAATCC-3' [19,20]. These primers have been validated to provide high taxon coverage for vaginal bacteria [21]. The reaction comprised of 5 ng/µl template, 1X Phusion® Master Mix (ThermoFisher, catalog number: F-531L), 0.375 µM V3-V4 locus specific primers and DMSO. The PCR was run under the following settings: 98°C for 60 s, 40 cycles of 98°C for 10 s, 64°C for 40 s, 72°C for 40 s and finally 10 min at 72°C, whereafter the samples were stored at 4°C.

For fungi, a two-step PCR was applied due to the poor amplification when using amplicon-specific primers with overhang adapters. PCR-amplicons of the ITS-1 region were generated using ITS1F and ITS2 primers (5'-GGTCATTTAGAGGAAGTAA-3' and 5'-GCTGCGTTCTTCATCGATGC-3' [22,23]. The primers for the second PCR to add the adapters were adapted from [24]. Both ITS PCR reactions were carried out in a total reaction volume of 20 µl consisting of 10 µl of 1X Phusion® Master Mix, 4.4 µl of the template, 0.5 µl of each primer, 0.6 µl of DMSO, with water added to reach the final volume. Template concentration at the first ITS PCR was 22 ng of sample DNA and at the second the amplicon derived from the previous PCR step was used. The cycling parameters for both ITS PCRs were 98°C for 60 s, followed by 44 cycles of 98°C for 10 s, 58°C for 40 s, 72°C for 40 s and finally 10 min at 72°C, after which the samples were stored at 4°C.

Both 16S rRNA gene and ITS PCR products were visualized using agarose gel electrophoresis and then purified with AMPure XP beads (Beckman Coulter, Copenhagen, Denmark). After purification, both PCR products were quantified using Quant-it PicoGreen dsDNA assay kit, 2000 assays (Invitrogen, cat.nro P7589). The index PCR was performed to add the barcoded primers [24] for sequencing, which was done using the following settings: 98°C for 60 s, 30 cycles of 98°C for 10 s, 64°C for 30 s, 72°C for 30 s and finally 10 min at 72°C. The 16S rRNA gene amplicons were diluted to 40 nM and the two locus-specific PCR products were mixed 1:1 for volume to be used as a template. The PCR amplification was performed with 4 µl template, 10 µl 1X Phusion® Master Mix, 0.5 µM each barcoded primer, 0.6 µl DMSO, with water added to reach the final volume of 20 µl.

The PCR products were quantified using Quant-it PicoGreen dsDNA assay kit (Invitrogen, cat.nro P7589). A 50 nM pool was prepared, and its clean-up was performed with AMPure XP beads (Beckman Coulter, Copenhagen, Denmark). The pooled 16S rRNA gene V3-V4-ITS amplicon mixture was sequenced at the

Biomedicum Functional Genomics Unit (FuGU), Helsinki, Finland with an Illumina MiSeq instrument using paired end 2 × 300 bp reads and a MiSeq v3 reagent kit. The loading concentration was 7.5 pM with 15% PhiX spike-in.

Pre-processing and splitting of 16S rRNA gene and ITS reads

First, we removed reads with ambiguous (Ns) bases and used cutadapt version 2.6 to trim the primers from raw reads and to split them into separate 16S rRNA gene and ITS amplicon datasets (fastq-files) with “discard_untrimmed” option [25]. With this method the 16S rRNA gene or ITS library could be extracted from the mixed library and analysed independently. We used the same pre-processing steps for both sequencing runs including the library split. After the library split, removal of primers, and “Filter and Trim”, the two 16S rRNA gene libraries were pre-processed together with the DADA2 package version 1.12.1 in R version 3.6.1, following the DADA2 Pipeline Tutorial (1.12) [26–28]. The ITS library was pre-processed as in DADA2 ITS Pipeline Workflow (1.8) except the removal of primers and ambiguous bases was already done during the library split [29].

Taxonomic assignment

The vaginal microbiota in most women consist of one or more genomically and functionally distinct *Lactobacillus* species that drive the well-established categorical vaginal community state types and determine the various compositions of bacteria and fungi [30,31]. To study the composition of this ecosystem it was rendered necessary to be able to assign taxonomies at species level resolution. These assignments can be both difficult and unreliable when sequence homology is the only basis for distinction, as is the case with popular tools and databases such as SILVA [32], RDP [33], and GreenGenes [34]. This is primarily due to the limited sequence information contained within the individual hypervariable regions of the 16S rRNA gene/ITS region, as well as due to incomplete databases that also contain reported mis-annotations [35,36]. For accurate annotations, our requirements were 1) an extensive and up-to-date reference database that can capture the largest number of taxa, and 2) background information that can be utilized to determine the specific ecosystem that each taxon has been isolated from to support the choice of the most probable hit for species that are poorly or completely indistinguishable in terms of sequence homology alone. Aligning the sequences using NCBI BLAST against the nucleotide (nt) database fulfilled the first requirement. The tabular output obtained as a result (outfmt 6), allowed us to extract accession IDs for each hit, which serve as keys to NCBI's extensive network of databases that were accessed through the rentrez R package [37]. We extracted information such as isolation source, host, and publications (titles and abstracts) associated to each hit and concatenated this information into one large block of text. Next, we constructed word queries (collectively termed “word banks”, Figure 1), that consist of lists of words specifying the ecosystem of interest. These word banks were the basis for filtration, searching for key words within this block of text to

determine the most likely annotations. As an example, to obtain annotations for the “Human vagina”, a word bank would consist of the following words: “human”, “homo sapiens”, “woman”, “women”, “vagina”, “vaginal”, “cervicovaginal”, “female genital tract”, “vaginitis” etc. For complete contents of each word bank see Additional file 3.

This tool (“taxminer”) can be flexibly used to stack different sets of word banks in succession to gain the maximum number of annotations between interrelated ecosystems, resulting in varying degrees of relevance and confidence. This trickle-down approach prevents oversaturated entries from masking real results in a pool of multiple hits and apply a hierarchy between different ecosystems relevant for the study. In our study, we stacked the vaginal, gut + skin, and clinical word banks as shown in Figure 1. To ensure the veracity of the tool, the selected annotations were manually verified and confirmed. While we acknowledge that this still does not provide a complete solution to species-level taxonomic annotations, which is an inherent problem with 16S rRNA gene amplicon sequencing, this method allows for a customized search strategy to obtain the mostly likely knowledge-based results. By utilizing a rich resource of information, we can intuitively select the best hits and weed out mis-annotations, both of which would prove to be a significant challenge otherwise.

Statistical methods

The R package *vegan* (2.5-7) was used to calculate: 1) alpha diversity (Shannon), rarefaction, and 2) permutational multivariate analysis of variance (PERMANOVA) through the “adonis” function to quantify the variance explained by the sequencing run type and other technical as well as clinical and other host variables on the bacterial and fungal composition within the samples [30]. Principal coordinate analysis was done using R packages *phyloseq* (1.34.0) and *vegan*, and the figures were plotted with *ggplot2* (3.3.3) [38]. All beta-diversity measures were based on Bray-Curtis dissimilarity. To classify the samples into community state types (CSTs) based on the microbial compositions, we used a recently developed tool called VALENCIA (VAginal community state typE Nearest Centroid clAssifier) [31]. All bar plots are created using *ggplot2* and heatmaps are created with *pheatmap* [32,33]. Correlation analysis between bacteria and fungi was done using Spearman correlation using *rcorr*.

Shotgun metagenomic sequencing

To validate the performance of the combined 16S rRNA gene+ITS amplicon sequencing, shotgun metagenomic sequencing was carried out on a subset of 21 samples that had sufficiently DNA left (min 300 ng). The libraries were prepared using Illumina Nextera™ DNA Flex Library Prep Kit and sequenced on Illumina’s NovaSeq6000 S4 PE150 run. [39]. To evaluate the accuracy of the text mining-based annotation method on the 16S rRNA gene and ITS amplicon data, the metagenome data was analysed with MetaPhlAn 3.0 [40]. We also performed a kingdom level search with Kraken2 (version 2.0.8) and RefSeq bacterial, archaea, viral, fungi and protozoa database (krak_microb) [39]. Finally, the HUMAnN 3.0 pipeline [41] was used to obtain the functional profiles for each sample. The reads per kilobase (RPK)

values contained within the pathway abundance output file were converted into counts per million (CPM), using the renormalizing utility ("humann_renorm_table"). The gene families output files provided by the HUMAnN pipeline were regrouped into level 4 EC categories using the regrouping utility ("humann_regroup_table"), to visualize the enzyme level metabolic potential of the microbes. These were also renormalized and combined with the pathway abundance files using the unpacking utility ("humann_unpack_pathways"). For the heatmap(s) of functional profiles, complete linkage clustering was applied on both rows and columns and the CPM values mapped were log transformed. Further information associated to each EC number such as enzyme name, reaction, substrate etc was extracted using KEGGREST [42].

Results

High concordance for bacterial data between combined 16S rRNA gene+ITS versus conventional 16S rRNA gene amplicon sequencing

The main aim of this study was to determine the feasibility and performance of a combined method of sequencing for bacterial 16S rRNA gene and fungal ITS region amplicons (hereafter also referred as combined run). Since the 50 vaginal samples were already analysed in a previous publication for the bacterial compositions, the existing 16S rRNA gene amplicon data (also referred to as reference run) and the results obtained therein serve as a point of reference [19]. To make the previous and current runs comparable, the outputs were analysed together through the same pre-processing and annotations pipeline as described in Methods. The 16S rRNA gene+ITS run (hereafter referred also as combined run for short), produced a mean of 120,186 reads per sample (for 16S rRNA gene 38,093 and ITS 82,423 reads per sample) on high read quality (quality profiles Additional file 1: Supplement Fig. 1). Details of both sequencing runs can be found in the Table 1. Rarefaction curves of the combined sequencing show that the sequencing depth was more than sufficient to capture all the species diversity for both amplicon types. Maximum diversity for all samples was detected with ~38,000 and ~32,000 reads, while 75% of the samples reached plateau at ~17,000 and ~600 reads for bacteria and fungi respectively indicating that the samples had been sequenced to sufficient depth to capture their true diversity.

Table 1. Sequencing run information

Sequencing run	16S rRNA gene		16S rRNA gene+ITS	
General sequencing information				
Instrument	Illumina HiSeq 2500		Illumina MiSeq	
Chemistry	HiSeq Rapid SBS kit v2 PE250		MiSeq Reagent kit v3 PE300	
Multiplexing level	689 samples, 50% run capacity		177 samples	
PhiX	(other high diversity samples included)		15 %	
Sequencing information for the 50 samples used in this study				
Number of raw reads				
Total per 50 samples	2,872,346		6,009,343	
average (range)	57,447	(4,604 - 209,300)	120,187	(66,686 - 166,697)
Number of 16S rRNA gene reads*				
average (range)	55,909	(4,448 - 205,205)	27,777	(6,137 - 62,889)
% of reads	98 %		32 %	
After DADA2				
average (range)	48,505	(3,793 - 171,427)	16,084	(2,589 - 38,816)
Number of ITS reads*				
average (range)	827	(89 - 3,100)	60,378	(25,861 - 95,392)
% of reads	2 %**		68 %	
After DADA2				
average (range)	-		24,801	(1,363 - 54,865)

* The reads containing ambiguous bases (Ns) were removed before the library split.

** No PCR-amplification for ITS; likely reflecting cross-reactivity of 16S rRNA gene primers with 18S rRNA gene region upstream of ITS1.

In terms of bacterial composition, the profiles from the two runs (16S rRNA gene and 16S rRNA gene+ITS) were highly similar (Pearson's correlation 0.97), vast majority of the samples clustering tightly in Principal coordinate analysis (PCoA) plots (Figure 2) and negligible variance being explained by the run type in PERMANOVA ($R^2 = 0.004$, $p = 0.9$). As previously published [19], the most abundant bacteria within the sample were of the genus *Lactobacillus*, a majority of which were *L. iners* and *L. crispatus* (Figure 3; parallel stacked bar plots), both being comparable in terms of prevalence between the runs – *L. iners* (16S rRNA gene: 25/50, combined: 23/50), *L. crispatus* (16S rRNA gene: 21/50, combined 19/50). *Gardnerella vaginalis*, a Bifidobacterium-related species associated with bacterial vaginosis was observed in 17 samples in the 16S rRNA gene-only run and 16 samples in the combined run. *Fannyhessea vaginae* (formerly known as *Atopbium vaginae*) was seen to mostly co-occur with *G. vaginalis*. We also observed *Clostridiales genomosp.* (CP049781.1) in non-*Lactobacillus*-dominated samples that were seen to contain BV-associated bacteria upon visual inspection. This particular bacterium was previously known as Bacterial vaginosis associated bacteria 1 (BVAB1) and has recently been renamed as "*Candidatus Lachnocurva vaginae*" [36]. When the ASVs for this microbe were annotated against the silva database, they were assigned as *Shuttleworthia*, which is a known

mis-annotation for these dedicated databases [35,36]. For a full list of BLAST vs silva vs RDP annotations for the ASVs see Additional file 4. The bacterial profiles were also assigned to CSTs and subCSTs (Additional file 1: Supplement Fig. 2).

Vaginal mycobiota based on ITS amplicon sequencing

Based on the ITS-reads of the combined run, all 50 samples contained fungi (Fig. 4). A total of 146 fungal ASVs were successfully annotated and were further agglomerated by species and filtered based on prevalence ≥ 2 of the samples and abundance > 100 reads. This resulted in 23 unique fungal annotations (Figure 4), with the phylum level distribution between the samples as follows: Ascomycota (47.8%, 11/23), Basidiomycota (43.5%, 10/23), and unclassified (8.7%, 2/23). The most abundant phylum within the samples were the Ascomycota, accounting for 83.6% of the total reads, with the most abundant species being *Candida albicans* (49.6%), *Alternaria alternata* (13.2%), *Cladosporium species* (10%), *Epicoccum nigrum* (8.3%), and *Saccharomyces cerevisiae* (1.1%) (Figure 4, stacked barplots). In terms of prevalence, *Cladosporium sp.* were seen to be present in most of the samples [*C. anthropophilum*: 28/50 samples (56%), *C. allicinum*: 23/50 samples (48%)], *A. alternata* was observed in 14/50 samples (28%), *C. albicans* was seen in 9/50 samples (18%), while *S. cerevisiae* was observed in 2 samples (4%). Basidiomycota made up 9.5% of the total reads, with *Malassezia* species being the most abundant (4.3%), and most prevalent, with *M. restricta* being observed in 26/50 samples (52%), *M. globosa* in 14/50 samples (28%), *M. sympodialis* in 3/50 samples (6%), *M. dermatitis* in 2/50 samples (4%), and unclassified *Malassezia* in 2/50 samples (4%). Lastly, uncultured fungal species made up 9.5% of the total abundance and were present in 33/50 samples (66%).

In PERMANOVA analysis, among the 29 tested technical and host variables, the number of lifetime sex partners was most strongly associated to the fungal composition ($R^2 = 0.07$, $p = 0.0003$), while also history of BV ($R^2 = 0.04$, $p = 0.04$) and the day of menstrual cycle at sampling reached statistical significance ($R^2 = 0.03$, $p = 0.047$). To determine the co-occurrence patterns between vaginal bacteria and fungi, the abundance profiles of microbes that occurred in at least 5 samples were correlated and visualized (Figure. 5). We observed that there were no significant correlations between the dominant lactobacilli – *L. crispatus* and *L. iners* – and fungi. However, *L. acidophilus* and *L. jensenii* were positively correlated to *M. restricta* while *L. reuteri* was seen to be positively correlated to an uncultured *Peniophora*. Additionally, *M. restricta* was negatively correlated with *Prevotella spp.* and *Sneathia amnii*. Other correlations of note were the negative correlation between *G. vaginalis* and *C. allicinum* and the positive correlation between *C. albicans* and *Streptococcus anginosus*.

Metagenomics results in relation to the 16S rRNA gene and parallel 16S rRNA gene+ITS amplicon sequencing

To assess the performance and coverage of the parallel 16S rRNA gene+ITS sequencing, we compared the taxonomic outputs to metagenomic shotgun sequencing. We also needed to benchmark the text-mining based annotation pipeline against an independent method of species recognition. The 21 libraries consisted of total 557,102,162 read pairs 2x150 (per sample average 26,528,674; min 10,276,162; max 46,160,072). We found no falsely annotated species among abundant and well-known (cultured) species with our 16S rRNA gene annotation pipeline compared to clade-specific metagenomic species annotation (MetaPhlAn). Bacterial compositions were highly similar between the amplicon and metagenomic sequencing (Figure 6, Additional file 1: Supplement Fig. 5). There was some variation in *G. vaginalis* detection between amplicon sequencing and metagenomics, however: *G. vaginalis* was not detected at all with metagenome in two samples (1216 and 1261) and was slightly more abundant in the remaining samples when detected. Overall, there was high concordance for the bacterial compositions across all three sequencing runs (Pearson's correlation for combined vs WGS 0.93, PERMANOVA for all three runs $R^2 = 0.007$, $p = 0.88$).

Fungi could not be detected through metagenomics either with MetaPhlAn or Kraken2, which use clade-specific marker genes and k-mer compositions for taxonomic assignments, respectively. In kingdom level analysis the yield for fungal reads was only 0.08-0.17% and the fungal community compositions

could not be determined except *S. cerevisiae* was detected in a single sample (ID: 1192) through MetaPhlAn.

Functional attributes of the vaginal microbiota

After removal of the human reads, functional profiling with the HUMAnN pipeline identified 104 metabolic pathways and 381 enzymes, revealing the metabolic potential of the bacteria present within these samples (n=21). Hierarchical clustering of the functional profiles successfully rearranged and regrouped samples with similar bacterial compositions and resultant CSTs (Figure 7), suggesting that there are identifiable differences in the metabolic potential of the typical bacterial populations in the vagina. For a full list of individual pathways, their enzymes, and EC numbers see Additional file 5.

The first cluster contained pathways that are essential to life and were observed across all samples. This included coenzyme A biosynthesis (COA-PWY), tRNA charging (TRNA-CHARGING-PWY), which consists of a group of 19 ligases responsible for amino acid activation, and cell wall biosynthesis pathways, mainly UDP-N-acetyl-D-glucosamine (UDPNAGSYN-PWY) and peptidoglycan biosynthesis (PEPTIDOGLYCANSYN-PWY). Pyruvate fermentation to lactate and acetate (PWY-5100) was largely observed in samples containing lactobacilli, highlighting their capacity to produce lactic acid.

A second cluster, representing pathways observed in samples containing *L. crispatus* (or CST I), consisted of amino acid biosynthesis, flavin biosynthesis, folate transformations, amino sugar degradation, and acetylene degradation pathways. There were nine enzymes related to the superpathway of L-lysine, L-threonine and L-methionine biosynthesis (PWY-724), shared between *L. crispatus* and *Bifidobacterium* spp. but were absent in *L. iners*. A further 7 enzymes related to the L-lysine biosynthesis pathway (PWY-2941) were observed within this cluster, with pathway variant II, distinguished by the enzyme N-acetyl-diaminopimelate deacetylase (EC 3.5.1.47), exclusively seen in *L. crispatus*. Two other variants (III & VI) that share the remaining six enzymes were also observed in *Bifidobacterium* spp., while they were completely absent in *L. iners*. An amino sugar degradation pathway, superpathway of N-acetylglucosamine (GlcNAc), N-acetylmannosamine (ManNAc) and N-acetylneuraminate (NeuAc) degradation (GLCMANNANAUT-PWY), and acetylene degradation (P161-PWY) were also observed to be exclusive to *L. crispatus*.

Another group of sparsely distributed pathways was observed to be specific to samples containing *L. iners* (or CST III), which included CDP-diacylglycerol biosynthesis (PWY-5667), 6-hydroxymethyl-dihydropterin diphosphate biosynthesis (PWY-6147), and six enzymes from the mevalonate pathway (PWY-922). The last cluster of interest consisted of pathways observed only in samples containing *Bifidobacterium* spp. or *G. vaginalis* (CST IV C-3 and CST IV B). This included two pathways for the synthesis of inosine monophosphate (IMP); a super pathway of 5-aminoimidazole ribonucleotide biosynthesis (PWY-6277) and an inosine-5'-phosphate biosynthesis pathway (PWY-6123).

Discussion

Performance and output of the combined 16S rRNA gene and ITS amplicon sequencing

With the rapidly expanding list of microbes being characterized for their role in human health and disease, it is crucial to be able to identify and differentiate between i) commensal microbes that co-exist in a symbiotic state within various niches of interest within the human body and ii) pathogens that induce dysbiosis and drive several clinical conditions. Studies targeting colonization patterns of fungi are particularly scarce, especially for the vaginal ecosystem when compared to bacteria [5,6]. A simple text search performed on the NCBI nucleotide database shows that there are 209,784 entries for vaginal specific bacteria, while there are only 1,458 entries for fungi, highlighting the disparity. We show that a

combined sequencing method that targets the 16S rRNA gene for bacteria and the fungal ITS region is a convenient strategy to study two distinct and clinically relevant microbial kingdoms in parallel without loss of data or incurring notable extra costs.

To our knowledge, this work is the first to report an amplicon sequencing method to study bacteria and fungi simultaneously within human samples using Illumina sequencing technology. Previously, Kittleman et al. reported a pyrosequencing approach that combined analysis of phylogenetic marker genes for bacteria, archaea and eukaryotic micro-organisms in ruminal samples and Coller et al. combined 16S rRNA gene and fungal ITS1 amplicons to an Illumina MiSeq run to study soil samples [43,44]. In addition, many studies have generated 16S rRNA gene and fungal ITS amplicons in the same samples, but the analyses were done separately [8,9,13–18]. Also, bacterial ITS amplicons have been employed for improved taxonomic resolution in parallel to conventional 16S rRNA gene amplicon sequencing [45,46]. Kits for single library constructions for bacteria and fungi are also commercially available (at least from Swift, Qiagen and Perkin Elmer), but the costs per sample are substantially higher and hence our method provides a cost-effective option for laboratories that process high sample volumes and possess the basic expertise for the index-PCR. In earlier studies where a single Illumina run was used for sequencing of both bacteria and fungi, the index-PCR was performed separately for each type of amplicon and they were only combined for sequencing [43,44]. Hence, to our knowledge we are the first to report a protocol where the pooling of bacterial and fungal amplicons takes place already after the locus-specific PCRs, and hence saves time and costs as all downstream steps including index PCR for both amplicon types occur in a single well. Use of the same barcodes for bacterial and fungal amplicons further simplifies the design and improves the efficiency especially when there are limited number of barcodes available.

By combining 16S rRNA gene and ITS amplicon sequencing in a parallel setting we were able to achieve comprehensive community composition data for both bacterial and fungal communities with very little added cost compared to ordinary 16S rRNA gene sequencing. The 16S rRNA gene and ITS libraries can be separated easily with publicly available bioinformatic tools like cutadapt that was used here and was the only additional step in pre-processing. The rarefaction curves indicated that with the chosen input of 177 vaginal samples in a single MiSeq run, both bacterial and fungal species were sequenced to an extent sufficient to represent their true diversity (Additional file 1: Supplement Fig. 3). Comparisons to conventional 16S rRNA gene amplicon and metagenomic sequencing demonstrated that the newly developed method produced highly similar results for bacterial compositions and there was a high degree of reproducibility between the results despite some methodological differences (HiSeq vs MiSeq, PE250 vs. PE300 chemistry) as a possible source of variation.

The simultaneous analysis of bacterial and fungal communities from the same samples is not a trivial approach since efficient and unbiased cell disruption is required to retrieve representative nucleic acids for further analysis. We used bead beating method for DNA extraction that has been previously shown to perform well on disrupting the rigid cells walls of fungi [47]. The analysis of fungal components of the microbiota is much more challenging than for bacteria e.g. due to the natural length polymorphism of ITS1 length that may complicate both the actual sequencing and subsequent bioinformatic analyses.

Text-mining based annotation and prioritization of hits

Achieving species level taxonomic annotations is often desired and crucial when investigating homogenous ecological niches such as the female genital tract, which is predominantly populated by different species of lactobacilli, rendering it necessary to attempt for identification of individual species and their patterns of interaction with the host. Presently, sequence homology is the main criterium used for taxonomic classifications within amplicon sequencing workflows, which however cannot distinguish between highly similar sequences, resulting in only a fraction of the sequence queries being assigned to species-level. This is also challenging in BLAST-based annotations, which can potentially produce thousands of highly similar alignment results and have significant amount of noise that may mask real, more relevant annotations. Conventionally, selecting results from the BLAST output would either involve taking the “top-hit” from a sorted list of annotations or require manual scanning. We show that by utilizing readily available information from public databases the selection of the most probable hit can be facilitated by means of knowledge-based species level ecosystem specific taxonomic annotations. By using BLAST as the alignment tool of choice and the nucleotide database as the reference to maximize the number of species that are potentially detected, our approach automates this selection process. Taxonomic annotations are split into a two-tiered process: i) sequence homology is determined by BLAST algorithms, with a threshold set at 98% for both “percentage identity” and “query coverage” and ii) ambiguity between identical alignments is eliminated by introducing background information such as previously identified host and source of isolation as a selection criterion to identify the most relevant hit. Hypothetically, given a set of 10 indistinguishable species level annotations, this approach allows us to probabilistically select the 5th alignment indicated to have been isolated from human vaginal swabs, and disregarding the other 9 that came from diverse sources including chicken crop, beer malt, and giant pandas. The filtration can be strict enough to look for specific strains of a microbe investigated in an environmental niche, or vague enough to accept anything collected from human samples.

While we demonstrate this tool in combination with BLAST based annotations, it can essentially be used as an added layer to any number of existing tools that can provide a link to NCBI accession IDs or attach previously gathered information for each alignment. By allowing any combination of user defined text queries, it can be customized for each individual study and enhances existing tools by increasing confidence in the final taxonomic annotations. In addition, information such as publications, journals, authors, lineage, and strain level resolution are also made available in the process. Strain identification is of particular interest as it is needed to identify and study the established biological and functional disparity in microbial strains, as discussed later [11,48,49].

We acknowledge that this approach is limited by the inherent limitations of 16S rRNA gene amplicons on discriminating bacteria to species level so that in less characterized ecosystems even the best guess, based on homology and background information, can be of low confidence. Another limitation may come from data availability, especially in terms of identifying novel species for a particular environment. However, the latter is remedied by the trickle-down application of the selection process, where the most

important and most likely annotations are assigned first, followed by expanding the filtration criteria to include the remaining potentially rare/novel taxa with decreasing confidence.

Microbial profiles

Whole genome shotgun (WGS) and an individual 16S rRNA gene sequencing method validated the bacterial profiles obtained from our parallel sequencing approach. Since the samples were from healthy subjects, they were dominated by lactobacilli, while samples with mixed/BV-like microscopy results [19] were indeed populated with *G. vaginalis*, *F. vaginae*, *Veillonella spp.*, *uncultured bacteria*, and *C. genomosp. (BVAB1)*. Since we report relative abundances, the discrepancy observed in the proportion of *G. vaginalis* observed in some samples could be attributed to the absence of these uncultured bacteria from the metagenome results, which are otherwise make up to 10% of the total abundance of the samples in question. The inability of metagenome pipeline to detect these microbes is largely due to incomplete reference databases. While we were not specifically searching for individual strains, the final annotations obtained through the filtration approach contained 11 different strains of *G. vaginalis* (see Additional file 1: Supplement Fig. 5), a microbe associated to bacterial vaginosis (BV), which have previously been described to exhibit varying capacities for sialidase (an enzyme responsible for mucus degradation), and beta galactosidase production [50–53]. Different strains of *G. vaginalis* have been identified with varying degrees of virulence based on the presence/absence of the genetic machinery facilitating adherence, biofilm formation, cytotoxicity, glycan degradation, antibiotic resistance, and displacement of lactobacilli [52]. This is indicative of the tendency of some *G. vaginalis* strains to colonize a non-BV vaginal environment as a commensal microbe while other strains can exhibit pathogenicity, highlighting the importance of achieving strain level resolution to understand their biological characteristics and as a diagnostic tool. As our samples were from healthy subjects, this could represent an early or intermediate state of dybiosis.

While we obtained high concordance between the bacterial profiles obtained with our combined amplicon sequencing vs. WGS approach, fungi could only be detected with the former. This is due to the underrepresentation of fungal reads within the metagenomes as well as their coverage in the reference databases used for WGS approach. The yield for fungal reads within the metagenomes was 0.08 - 0.17% in our study and $0.17 \pm 0.04\%$ in a previous large scale vaginal microbiota study [11]. Eukaryotic genomes are also significantly larger and much more complex compared to prokaryotes, which further hinders metagenome analysis, especially with a restricted set of reference databases [54]. Apart from *S. cerevisiae* being detected in one sample (ID: 1192) with MetaPhlan, there were no positive hits in the WGS data while 23 unique species were detected across the 50 samples by ITS amplicon sequencing. This shows that even in this era of relatively low-cost shotgun metagenomic sequencing, amplicon sequencing still has the edge when studying complex low abundance communities. With the lack of studies focusing on fungi, deposited fungal sequences with valid human isolation sources are also scarce. In our samples only *C. albicans*, a well-characterized pathogenic yeast responsible for vulvovaginal candidiasis [5], was identified from vaginal samples. The remaining annotations were traced back to either the gut

(*Aspergillus spp.*), skin (*Malassezia spp.* and *Cladosporium spp.*), or other clinical isolates (*S. cerevisiae*, *Schizophyllum spp.*). Of these annotations, *Cladosporium spp.* were a possible wildcard because of the uncertainty surrounding their isolation source. On one hand, the metadata for *Cladosporium spp.* hits stated that they were isolated from clinical human samples and certain species have been associated to human and animal infections. On the other hand, they are also widely recognized as household fungi and are shown to be a consistent environmental contaminant in fungal studies [8,55,56].

Investigating the correlations between the micro- and mycobiota revealed some interesting trends, notably the positive correlation between *C. albicans* and *S. anginosus*. Both microbes are individually implicated in negative clinical outcomes and *Streptococci spp.* have been shown to co-aggregate with *C. albicans*, which allows it to form a thicker biofilm and obtain antibiotic resistance [57,58]. *Prevotella spp.* and *Malassezia spp.* have previously been shown to positively correlate on skin, which opposed our findings, indicating either a different colonization pattern in this ecological niche or presence of different strains [59]. Positive correlation between *L. reuteri* and an uncultured fungus of genus *Peniophora* could potentially be attributed to their shared ability to bind/sequester and degrade aflatoxin B₁, a mycotoxin produced by *Aspergillus spp.*, as a survival mechanism [60–62]. We speculate that this might be indicative of an evolutionary characteristic that these microbes possess to inhabit the same environment. Among the *Cladosporium* species, the only observed correlation was a negative one between *C. allicinum* and *G. vaginalis* [63,64]. Due to the scarcity of information on the vaginal fungal communities and their potential associations to clinical and other host factors, we carried out the same PERMANOVA analysis for fungi as previously done for the bacterial microbiota of the same samples [65]. We found number of sex partners (lifetime), period day, history of BV, and antibiotic and probiotic use to be associated with fungal abundance profiles while education was the strongest factor explaining bacterial variation [65]. This divergence suggests that the drivers of colonization patterns for vaginal bacteria and fungi may be independent of each other [65]. For the full PERMANOVA analysis results see Additional file 2.

Functional profiles

Even with a relatively small sample size, functional profiles obtained from the metagenome analysis revealed clear separation of bacterial profiles and their metabolic potential. The capacity for *L. crispatus* to synthesize lysine has been previously identified as well as its absence in *L. iners* [66]. An interesting observation was the co-occurrence of the almost complete mevalonate pathway (six enzymes), which produces cholesterol as an end-product, and inerolysin, a cholesterol dependent cytolysin (CDC) in samples containing *L. iners*. Presence of the mevalonate pathway has been indicated in urogenital bacterial populations and associated to the ability of certain lactic acid bacteria to cope with oxidative stress, which may be an additional adaptation that allows *L. iners* to colonize hostile environments as well as co-habit with other microbes, especially *G. vaginalis* [66–68]. CDCs such as inerolysin and vaginolysin (*G. vaginalis* specific) have been associated to membrane adhesion and subsequent pathogenicity. Pore formation in the host cell membranes, a characteristic of CDCs, has been speculated to allow the microbes to gain access to nutrients and potentially kill immune cells, allowing *L. iners* to

survive in a dysbiotic state [66,69,70]. A repertoire of enzymes involved in the biosynthesis and degradation of inosine monophosphate (IMP) were observed in samples containing *G. vaginalis*. Even though the enzymes associated with degradation could not be assigned to a single taxonomy, their exclusivity in these samples already shows an interesting trend. Other studies have discussed inosine in various contexts, including as a metabolite from related intestinal *Bifidobacterium pseudolongum* that modulates an enhanced immunotherapy response, while for some strains of *G. vaginalis* it has been indicated as a carbon source [71,72]. These findings warrant further investigations to understand the role of inosine metabolism among certain *G. vaginalis* strains within the vaginal ecosystem.

Conclusion

The main outcomes of this work are: 1) development and validation of a novel cost-effective method for simultaneous amplicon sequencing for bacteria and fungi in complex samples; 2) development of a novel annotation approach for microbial amplicon sequencing data where the conventional homology-based search is combined with text mining for knowledge-based selection of ecosystem-specific hits; 3) novel insights into the human vaginal ecosystem through description of the mycobiota as well as the functional attributes of the bacterial microbiome emerging from patterns of enzymatic and other metabolic capacities that are characteristic for the well-established vaginal community types (CSTs), each dominated by distinct prevalent bacteria.

Our sequencing method should also be directly applicable to environmental sciences, agriculture, biotechnology, food technology, and any other field of science where studying bacterial and fungal populations are of particular interest. Even with its limitations, the customizable nature the text mining-based filtration approach can be adapted to a multitude of ecosystems under investigation.

Declarations

Ethics approval and consent to participate

The study was approved by the ethical committee of The Hospital District of Helsinki and Uusimaa and Helsinki region hospital district (21/13/03/03/2014) and performed in accordance with the principles of the Helsinki Declaration. All participants signed an informed consent.

Consent for publication

Not applicable.

Competing Interests

The authors declare no competing interests.

Acknowledgements

We want to thank all women who volunteered to take part in our research and the nurses at HUSLAB who helped us to collect the samples. University of Helsinki and Helsinki University Hospital provided us financial support and facilities to complete this study. We thank Rebecka Ventin-Holmberg for literature search and discussions on selection of methodology for fungal analysis.

Funding

This study was supported by grants from the European Union's Horizon 2020 research and innovation programme (H2020 MSCA Sweet Crosstalk project under grant agreement No 814102) to AS, from Helsinki Uusimaa Hospital District, Academy of Finland, Finnish Medical Foundation, and Jalmari and Rauha Ahokas Foundation to IK, and from Helsinki Uusimaa Hospital District and Finnish Cancer Organisation to PN, and Finnish Medical Foundation and Doctoral School in Health Sciences, University of Helsinki and Helsinki Central Hospital to SV.

Availability of data and material

The sequences have been deposited at European Nucleotide Archive. ENA accession numbers: PRJEB25778 and PRJEB43447. The latest version of the R package "taxminer" is available on Github <https://github.com/SchahzadSaqib/taxminer>. The library split script has been added as a utility script to the same GitHub page.

Authors' contributions

PN, AS, IK and SV conceptualized the original project. SV recruited the study subjects and collected the samples. Development of parallel sequencing methodology: AS, SV, TK, SS. Development of annotation methodology: SS, AS, SV. Analysis and interpretation of data: SS, SV, AS. Writing, review, and revision of the manuscript: AS, SS, SV, IK, PN. All authors read and approved the final manuscript.

References

1. Rowan-Nash AD, Korry BJ, Mylonakis E, Belenky P. Cross-Domain and Viral Interactions in the Microbiome. *Microbiol Mol Biol Rev* [Internet]. American Society for Microbiology; 2019 [cited 2020 Jun 29];83. Available from: <https://mmbr.asm.org/content/83/1/e00044-18>.
2. Krüger W, Vielreicher S, Kapitan M, Jacobsen ID, Niemiec MJ. Fungal-bacterial interactions in health and disease [Internet]. *Pathogens*. MDPI AG; 2019 [cited 2020 Jun 29]. p. 70. Available from: <https://www.mdpi.com/2076-0817/8/2/70>.
3. Underhill DM, Iliev ID. The mycobiota: Interactions between commensal fungi and the host immune system [Internet]. *Nat. Rev. Immunol.* Nature Publishing Group; 2014 [cited 2020 Sep 29]. p. 405–16. Available from: [/pmc/articles/PMC4332855/?report = abstract](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4332855/?report=abstract).
4. Lai GC, Tan TG, Pavelka N. The mammalian mycobiome: A complex system in a dynamic relationship with the host. *Wiley Interdiscip Rev Syst Biol Med* [Internet]. Wiley-Blackwell; 2018 [cited 2020 Sep 29];11. Available from: [/pmc/articles/PMC6586165/?report = abstract](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6586165/?report=abstract).

5. Bradford LL, Ravel J. The vaginal mycobiome: A contemporary perspective on fungi in women's health and diseases [Internet]. Virulence. Taylor and Francis Inc.; 2017 [cited 2020 Sep 29]. p. 342–51. Available from: /pmc/articles/PMC5411243/?report = abstract.
6. Cui L, Morris A, Ghedin E. The human mycobiome in health and disease [Internet]. Genome Med. BioMed Central; 2013 [cited 2021 Feb 19]. p. 63. Available from: /pmc/articles/PMC3978422/.
7. Onderdonk AB, Delaney ML, Fichorova RN. The Human Microbiome during Bacterial Vaginosis. Clin Microbiol Rev American Society for Microbiology (ASM). 2016;29:223–38.
8. Drell T, Lillsaar T, Tummeleht L, Simm J, Aaspõllu A, Väin E, et al Characterization of the Vaginal Micro- and Mycobiome in Asymptomatic Reproductive-Age Estonian Women. Fredricks DN, editor. PLoS One [Internet]. Public Library of Science; 2013 [cited 2020 Mar 2];8:e54379. Available from: <http://dx.plos.org/10.1371/journal.pone.0054379>.
9. Ward TL, Dominguez-Bello MG, Heisel T, Al-Ghalith G, Knights D, Gale CA. Development of the Human Mycobiome over the First Month of Life and across Body Sites. mSystems [Internet]. American Society for Microbiology; 2018 [cited 2020 Jun 29];3. Available from: <https://msystems.asm.org/content/3/3/e00140-17>.
10. Godoy-Vitorino F, Romaguera J, Zhao C, Vargas-Robles D, Ortiz-Morales G, Vázquez-Sánchez F, et al. Cervicovaginal fungi and bacteria associated with cervical intraepithelial neoplasia and high-risk human papillomavirus infections in a hispanic population. Front Microbiol. 2018;9.
11. Ma B, France MT, Crabtree J, Holm JB, Humphrys MS, Brotman RM, et al. A comprehensive non-redundant gene catalog reveals extensive within-community intraspecies diversity in the human vagina. Nat Commun [Internet]. Nature Research; 2020 [cited 2020 Apr 13];11:1–13. Available from: <https://www.nature.com/articles/s41467-020-14677-3>.
12. Nash AK, Auchtung TA, Wong MC, Smith DP, Gesell JR, Ross MC, et al. The gut mycobiome of the Human Microbiome Project healthy cohort. Microbiome [Internet]. BioMed Central; 2017 [cited 2020 Sep 29];5:153. Available from: <https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-017-0373-4>.
13. Angebault C, Ghoulane A, Volant S, Botterel F, D'Enfert C, Bougnoux ME. Combined bacterial and fungal intestinal microbiota analyses: Impact of storage conditions and DNA extraction protocols. Coste AT, editor. PLoS One [Internet]. Public Library of Science; 2018 [cited 2020 Jun 29];13:e0201174. Available from: <https://dx.plos.org/10.1371/journal.pone.0201174>.
14. Angebault C, Payen M, Woerther PL, Rodriguez C, Botterel F. Combined bacterial and fungal targeted amplicon sequencing of respiratory samples: Does the DNA extraction method matter? Sturtevant J, editor. PLoS One [Internet]. Public Library of Science; 2020 [cited 2020 Jun 29];15:e0232215. Available from: <https://dx.plos.org/10.1371/journal.pone.0232215>.
15. Kaplan RC, Wang Z, Usyk M, Sotres-Alvarez D, Daviglius ML, Schneiderman N, et al. Gut microbiome composition in the Hispanic Community Health Study/Study of Latinos is shaped by geographic relocation, environmental factors, and obesity. Genome Biol [Internet]. BioMed Central Ltd.; 2019

- [cited 2020 Jun 29];20:219. Available from:
<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1831-z>.
16. Wu L, Zeng T, Deligios M, Milanesi L, Langille MGI, Zinellu A, et al. Age-Related Variation of Bacterial and Fungal Communities in Different Body Habitats across the Young, Elderly, and Centenarians in Sardinia. *mSphere* [Internet]. American Society for Microbiology; 2020 [cited 2020 Jun 29];5. Available from: <https://msphere.asm.org/content/5/1/e00558-19>.
 17. Usyk M, Zolnik CP, Castle PE, Porras C, Herrero R, Gradissimo A, et al Cervicovaginal microbiome and natural history of HPV in a longitudinal study. Silvestri G, editor. *PLoS Pathog* [Internet]. Public Library of Science; 2020 [cited 2020 Jun 29];16:e1008376. Available from: <https://dx.plos.org/10.1371/journal.ppat.1008376>.
 18. Kirjavainen PV, Karvonen AM, Adams RI, Täubel M, Roponen M, Tuoresmäki P, et al. Farm-like indoor microbiota in non-farm homes protects children from asthma development. *Nat Med* [Internet]. Nature Publishing Group; 2019 [cited 2020 Jun 29];25:1089–95. Available from: <https://www.nature.com/articles/s41591-019-0469-4>.
 19. Virtanen S, Rantsi T, Virtanen A, Kervinen K, Nieminen P, Kalliala I, et al. Vaginal Microbiota Composition Correlates Between Pap Smear Microscopy and Next Generation Sequencing and Associates to Socioeconomic Status. *Sci Rep* Nature Publishing Group. 2019;9:1–9.
 20. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* [Internet]. Oxford Academic; 2013 [cited 2020 Sep 29];41:e1–e1. Available from: .
 21. Hugerth LW, Pereira M, Zha Y, Seifert M, Kaldhusdal V, Boulund F, et al Assessment of In Vitro and In Silico Protocols for Sequence-Based Characterization of the Human Vaginal Microbiome. Zhu L, editor. *mSphere* [Internet]. American Society for Microbiology; 2020 [cited 2021 Feb 10];5. Available from: <https://msphere.asm.org/content/5/6/e00448-20>.
 22. Gardes M, Bruns TD. ITS primers with enhanced specificity for basidiomycetes - application to the identification of mycorrhizae and rusts. *Mol Ecol* [Internet]. John Wiley & Sons, Ltd; 1993 [cited 2020 Sep 29];2:113–8. Available from: <https://onlinelibrary.wiley.com/doi/full/10.1111/j.1365-294X.1993.tb00005.x>.
 23. White TJ, Bruns T, Lee S, Taylor J. AMPLIFICATION AND DIRECT SEQUENCING OF FUNGAL RIBOSOMAL RNA GENES. FOR PHYLOGENETICS. *PCR Protoc* [Internet]. Elsevier; 1990 [cited 2020 Nov 30]. p. 315–22. Available from: <https://www.scienceopen.com/document?vid=7dc39406-40bd-42de-b0d4-54ea74b3e063>.
 24. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. *Appl Environ Microbiol* [Internet]. American Society for Microbiology; 2013 [cited 2020 Sep 29];79:5112–20. Available from: .

25. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal [Internet]. 2011 [cited 2021 Feb 16];17:10. Available from: <http://www-huber.embl.de/users/an->.
26. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13:581–3.
27. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2019.
28. Callahan B. DADA2 Pipeline Tutorial (1.12). GitHub [Internet]. 2019;1. Available from: <https://benjjneb.github.io/dada2/tutorial.html>.
29. Callahan BJ. DADA2 ITS Pipeline Workflow (1.8) [Internet]. GitHub. 2019. Available from: https://benjjneb.github.io/dada2/ITS_workflow.html.
30. Mehta O, Ghosh TS, Kothidar A, Gowtham MR, Mitra R, Kshetrapal P, et al. Vaginal Microbiome of Pregnant Indian Women: Insights into the Genome of Dominant Lactobacillus Species. Microb Ecol [Internet]. Springer; 2020 [cited 2021 Feb 15];80:487–99. Available from: <https://doi.org/10.1007/s00248-020-01501-0>.
31. France M, Ma B, Gajer P, Brown S, Humphrys M, Holm J, et al. VALENCIA: A nearest centroid classification method for vaginal microbial communities based on composition. 2020 [cited 2020 Sep 22]; Available from: <https://doi.org/10.21203/rs.2.24139/v1>.
32. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Available from: <http://www.arb-silva.de>.
33. 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 [Internet]. 2014;42:D633–42. Available from: <http://rdp.cme.msu.edu/>.
34. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol [Internet]. 2006;72:5069–72. Available from: <http://greengenes.lbl.gov>.
35. Lennard K, Dabee S, Barnabas SL, Havyarimana E, Blakney A, Jaumdally SZ, et al. Microbial composition predicts genital tract inflammation and persistent bacterial vaginosis in South African adolescent females. Infect Immun. American Society for Microbiology; 2018;86.
36. Holm JB, France MT, Ma B, McComb E, Robinson CK, Mehta A, et al. Comparative Metagenome-Assembled Genome Analysis of “Candidatus Lachnocurva vaginae”, Formerly Known as Bacterial Vaginosis-Associated Bacterium – 1 (BVAB1). Front Cell Infect Microbiol [Internet]. Frontiers Media S.A.; 2020 [cited 2021 Feb 17];10:117. Available from: <https://www.frontiersin.org/article/10.3389/fcimb.2020.00117/full>.
37. Sayers E. E-utilities Quick Start. National Center for Biotechnology Information (US); 2018 [cited 2020 Jul 27]; Available from: <https://www.ncbi.nlm.nih.gov/books/NBK25500/>.
38. McMurdie PJ, Holmes S. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. Watson M, editor. PLoS One [Internet]. Public Library of Science; 2013

- [cited 2020 Mar 5];8:e61217. Available from: <https://dx.plos.org/10.1371/journal.pone.0061217>.
39. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol* [Internet]. BioMed Central Ltd.; 2019 [cited 2021 Feb 19];20:257. Available from: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1891-0>.
 40. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* [Internet]. Nature Publishing Group; 2012 [cited 2020 Jun 30];9:811–4. Available from: <https://www.nature.com/articles/nmeth.2066>.
 41. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods* [Internet]. Nature Publishing Group; 2018 [cited 2020 Jun 30];15:962–8. Available from: <https://www.nature.com/articles/s41592-018-0176-y>.
 42. Tenenbaum D, Maintainer B. KEGGREST. Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG) version 1.30.1 from Bioconductor [Internet]. 2020 [cited 2021 Mar 1]. Available from: <https://rdrr.io/bioc/KEGGREST/>.
 43. Kittelmann S, Seedorf H, Walters WA, Clemente JC, Knight R, Gordon JI, et al Simultaneous Amplicon Sequencing to Explore Co-Occurrence Patterns of Bacterial, Archaeal and Eukaryotic Microorganisms in Rumen Microbial Communities. White BA, editor. *PLoS One* [Internet]. Public Library of Science; 2013 [cited 2020 Sep 22];8:e47879. Available from: <https://dx.plos.org/10.1371/journal.pone.0047879>.
 44. Coller E, Cestaro A, Zanzotti R, Bertoldi D, Pindo M, Larger S, et al. Microbiome of vineyard soils is shaped by geography and management. *Microbiome* [Internet]. BioMed Central Ltd.; 2019 [cited 2021 Feb 22];7. Available from: <https://pmc/articles/PMC6839268/>.
 45. Milani C, Duranti S, Mangifesta M, Lugli GA, Turrone F, Mancabelli L, et al. Phylotype-level profiling of lactobacilli in highly complex environments by means of an internal transcribed spacer-based metagenomic approach. *Appl Environ Microbiol* [Internet]. American Society for Microbiology; 2018 [cited 2021 Mar 2];84:706–24. Available from: <http://aem.asm.org/>.
 46. Milani C, Lugli GA, Turrone F, Mancabelli L, Duranti S, Viappiani A, et al. Evaluation of bifidobacterial community composition in the human gut by means of a targeted amplicon sequencing (ITS) protocol. *FEMS Microbiol Ecol* [Internet]. Oxford University Press; 2014 [cited 2021 Mar 2];90:n/a-n/a. Available from: <https://academic.oup.com/femsec/article-lookup/doi/10.1111/1574-6941.12410>.
 47. Huseyin CE, Rubio RC, O'Sullivan O, Cotter PD, Scanlan PD. The fungal frontier: A comparative analysis of methods used in the study of the human gut mycobiome. *Front Microbiol* [Internet]. Frontiers Media S.A.; 2017 [cited 2021 Feb 22];8. Available from: <https://pmc/articles/PMC5534473/>.
 48. Greenblum S, Carr R, Borenstein E. Extensive strain-level copy-number variation across human gut microbiome species. *Cell* [Internet] Cell Press. 2015;160:583–94. [cited 2021 Feb 15 ;. Available from: <https://pmc/articles/PMC4507803/>.

49. Yan Y, Nguyen LH, Franzosa EA, Huttenhower C. Strain-level epidemiology of microbial communities and the human microbiome [Internet]. *Genome Med.* BioMed Central Ltd; 2020 [cited 2021 Feb 15]. p. 71. Available from: <https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-00765-y>.
50. Lewis WG, Robinson LS, Gilbert NM, Perry JC, Lewis AL. Degradation, foraging, and depletion of mucus sialoglycans by the vagina-adapted actinobacterium *Gardnerella vaginalis*. *J Biol Chem* [Internet]. American Society for Biochemistry and Molecular Biology; 2013 [cited 2021 Feb 9];288:12067–79. Available from: </pmc/articles/PMC3636892/>.
51. Cornejo OE, Hickey RJ, Suzuki H, Forney LJ. Focusing the diversity of *Gardnerella vaginalis* through the lens of ecotypes. *Evol Appl* [Internet]. Wiley-Blackwell; 2018 [cited 2021 Feb 9];11:312–24. Available from: <http://doi.wiley.com/10.1111/eva.12555>.
52. Castro J, Alves P, Sousa C, Cereija T, França Â, Jefferson KK, et al. Using an in-vitro biofilm model to assess the virulence potential of Bacterial Vaginosis or non-Bacterial Vaginosis *Gardnerella vaginalis* isolates. *Sci Rep* [Internet]. Nature Publishing Group; 2015 [cited 2021 Feb 9];5. Available from: </pmc/articles/PMC4481526/>.
53. Vaneechoutte M, Guschin A, Van Simaey L, Gansemans Y, Van Nieuwerburgh F, Cools P. Emended description of *Gardnerella vaginalis* and description of *gardnerella leopoldii* sp. Nov., *gardnerella piovii* sp. nov. and *Gardnerella swidsinskii* sp. nov., with delineation of 13 genomic species within the genus *Gardnerella*. *Int J Syst Evol Microbiol* [Internet]. Microbiology Society; 2019 [cited 2021 Feb 9];69:679–87. Available from: <https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijsem.0.003200>.
54. Gilbert JA, Dupont CL. Microbial metagenomics: Beyond the genome. *Ann Rev Mar Sci* [Internet]. Annual Reviews. 2011;3:347–71. [cited 2021 Feb 25];. Available from: www.annualreviews.org.
55. Aihara M, Tanaka T, Takatori K. *Cladosporium* as the main fungal contaminant of locations in dwelling environments. *Biocontrol Sci.* 2001;6:49–52.
56. Sandoval-Denis M, Gené J, Sutton DA, Wiederhold NP, Cano-Lira JF, Guarro J. New species of *Cladosporium* associated with human and animal infections. *Persoonia Mol Phylogeny Evol Fungi* [Internet]. Nationaal Herbarium Nederland; 2016 [cited 2020 Jul 31];36:281–98. Available from: </pmc/articles/PMC4988372/?report=abstract>.
57. Jenkinson HF, Lala HC, Shepherd' MG. Coaggregation of *Streptococcus sanguis* and Other *Streptococci* with *Candida albicans*. *Infect. Immun.* 1990.
58. Tavernier S, Sass A, De Bruyne M, Baeke F, De Rycke R, Crabbé A, et al. Decreased susceptibility of *Streptococcus anginosus* to vancomycin in a multispecies biofilm is due to increased thickness of the cell wall. *J Antimicrob Chemother* [Internet]. Oxford University Press; 2018 [cited 2021 Feb 19];73:2323–30. Available from: <https://academic.oup.com/jac/article/73/9/2323/5036452>.
59. Ward TL, Knights D, Gale CA. Infant fungal communities: Current knowledge and research opportunities [Internet]. *BMC Med.* BioMed Central Ltd.; 2017 [cited 2021 Feb 19]. Available from: </pmc/articles/PMC5304398/>.

60. Alberts JF, Gelderblom WCA, Botha A, van Zyl WH. Degradation of aflatoxin B1 by fungal laccase enzymes. *Int J Food Microbiol Elsevier*. 2009;135:47–52.
61. Hernandez-Mendoza A, González-Córdova AF, Vallejo-Cordoba B, Garcia HS. Effect of oral supplementation of *Lactobacillus reuteri* in reduction of intestinal absorption of aflatoxin B 1 in rats. *J Basic Microbiol*. 2011;51:263–8.
62. Kim S, Lee H, Lee S, Lee J, Ha J, Choi Y, et al. Invited review: Microbe-mediated aflatoxin decontamination of dairy products and feeds. *J Dairy Sci Elsevier Inc*. 2017;100:871–80.
63. Denning DW, O'Driscoll BR, Hogaboam CM, Bowyer P, Niven RM. The link between fungi and severe asthma: A summary of the evidence [Internet]. *Eur. Respir. J. European Respiratory Society*; 2006 [cited 2021 Feb 19]. p. 615–26. Available from: .
64. Hevia A, Milani C, López P, Donado CD, Cuervo A, González S, et al. Allergic Patients with Long-Term Asthma Display Low Levels of *Bifidobacterium adolescentis*. 2016 [cited 2021 Feb 19]; Available from: <http://www.idi.mineco.gob.es/>.
65. Virtanen S, Rantsi T, Virtanen A, Kervinen K, Nieminen P, Kalliala I, et al. Vaginal Microbiota Composition Correlates Between Pap Smear Microscopy and Next Generation Sequencing and Associates to Socioeconomic Status. *Sci Rep Nature Publishing Group*. 2019;9:7750.
66. France MT, Mendes-Soares H, Forney LJ. Genomic comparisons of *Lactobacillus crispatus* and *Lactobacillus iners* reveal potential ecological drivers of community composition in the vagina. *Appl Environ Microbiol* [Internet]. American Society for Microbiology; 2016 [cited 2021 Feb 8];82:7063–73. Available from: [/pmc/articles/PMC5118917/?report = abstract](https://pubmed.ncbi.nlm.nih.gov/27118917/).
67. Hagi T, Kobayashi M, Nomura M. Aerobic conditions increase isoprenoid biosynthesis pathway gene expression levels for carotenoid production in *Enterococcus gilvus*. *FEMS Microbiol Lett* [Internet]. Oxford University Press; 2015 [cited 2021 Feb 15];362:75. Available from: <https://academic.oup.com/femsle/article/doi/10.1093/femsle/fnv075/660701>.
68. Thomas-White K, Forster SC, Kumar N, Van Kuiken M, Putonti C, Stares MD, et al. Culturing of female bladder bacteria reveals an interconnected urogenital microbiota. *Nat Commun* [Internet] Nature Publishing Group. 2018;9:1–7. [cited 2021 Feb 15];. Available from: www.nature.com/naturecommunications.
69. Ragaliauskas T, Plečkaitytė M, Jankunec M, Labanauskas L, Baranauskiene L, Valincius G. Inerolysin and vaginolysin, the cytolytic proteins implicated in vaginal dysbiosis, differently impair molecular integrity of phospholipid membranes. *Sci Rep* [Internet] Nature Publishing Group. 2019;9:1–11. [cited 2021 Feb 15];. Available from: www.nature.com/scientificreports.
70. Rampersaud R, Planet PJ, Randis TM, Kulkarni R, Aguilar JL, Lehrer RI, et al. Inerolysin, a cholesterol-dependent cytolytic protein produced by *Lactobacillus iners* ∇. *J Bacteriol* [Internet]. American Society for Microbiology Journals; 2011 [cited 2021 Feb 15];193:1034–41. Available from: <http://jb.asm.org/>.
71. Mager LF, Burkhard R, Pett N, Cooke NCA, Brown K, Ramay H, et al. Microbiome-derived inosine modulates response to checkpoint inhibitor immunotherapy. *Science* (80-) [Internet]. American

Association for the Advancement of Science; 2020 [cited 2021 Feb 17];369:1481–9. Available from: <https://science.sciencemag.org/content/369/6510/1481>.

72. Khan S, Vancuren SJ, Hill JE. A Generalist Lifestyle Allows Rare *Gardnerella* spp. to Persist at Low Levels in the Vaginal Microbiome. *Microb Ecol* [Internet]. Springer; 2020 [cited 2021 Feb 17];1. Available from: [/pmc/articles/PMC7678777/](https://pubmed.ncbi.nlm.nih.gov/3287777/).

Figures

Figure 1

Workflow for the text-mining approach. A) An illustration of the annotation workflow implemented within this study. The alignments are performed with BLAST against the nucleotide data, which is restricted with an ID list for Bacteria/Fungi. The accession IDs from the alignment are keys to several NCBI databases used to extract isolation/ecosystem related information for each hit. Word banks that define the study specific host/ecological niche are then used as a filtration and selection criteria, resulting in a final list of

most relevant annotations. B) The accession ID based database restrictions are very flexible and can be stacked with different word banks. In our study we separately extract accession ID lists for the organism of interest (i.e. bacteria or fungi) + absence/presence of uncultured entries. The sequence alignment is performed against restricted database prior to stacking them with multiple word bank filtration in a trickle-down approach.

Figure 2

Principal coordinates analysis (PCoA) of the relative species abundances. All 50 + 50 samples are plotted, a line combining the same sample in the reference 16S rRNA gene data (HiSeq) and the combined 16S rRNA gene+ITS run (MiSeq). Dominant species of each sample are marked with colour, the sequencing method with shape and sample read count with background colour. The few samples with highest distance are labelled.

Figure 3

Bacterial relative abundances. Comparison of bacterial compositions derived from the combined 16S rRNA gene+ITS amplicon run (SEP-samples) versus a conventional 16S rRNA gene amplicon sequencing run (S-samples).

Figure 4

Fungal relative abundances. Fungal profiles from the combined 16S rRNA gene and ITS amplicon sequencing run.

Figure 5

Correlations between bacterial and fungi. Heatmap of the statistically significant ($p < 0.05$) spearman correlations between the bacterial and fungal profiles.

Figure 6

Comparison between bacterial profiles obtained from the three different approaches. Comparison of the bacterial compositions derived from the conventional 16S rRNA gene amplicon sequencing (HiSeq), combined 16S rRNA gene+ITS sequencing (MiSeq), and metagenomes (MTGN) for a subset of 21 samples.

Figure 7

Bacterial functional profiles. Heatmap illustrating the functional profiles obtained from the metagenomic analysis of the subset of 21/50 vaginal samples. The full list of genes and pathways obtained for each sample are listed in Additional file 4.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1.docx](#)
- [Additionalfile2.xlsx](#)
- [Additionalfile3.xlsx](#)
- [Additionalfile4.xlsx](#)
- [Additionalfile5.xlsx](#)