The metagenome of the female upper reproductive tract

Li, Fei; Chen, Chen; Wei, Weixia; Wang, Zirong; Dai, Juanjuan; Hao, Lilan; Song, Liju; Zhang, Xiaowei; Zeng, Liping; Du, Hui; Tang, Huiru; Liu, Na; Yang, Huanming; Wang, Jian; Madsen, Lise; Brix, Susanne; Kristiansen, Karsten; Xu, Xun; Li, Junhua; Wu, Ruifang; Jia, Huijue

Published in:
GigaScience

DOI:
10.1093/gigascience/giy107

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
RESEARCH

The metagenome of the female upper reproductive tract

Fei Li¹,²,³,†, Chen Chen²,³,†, Weixia Wei⁴,⁵,†, Zirong Wang², Juanjuan Dai⁴,⁵, Lilan Hao²,³, Liju Song²,³, Xiaowei Zhang²,³, Liping Zeng⁴,⁵, Hui Du⁴,⁵, Huiru Tang⁴,⁵, Na Liu⁶, Huanming Yang²,⁹, Jian Wang²,⁹, Lise Madsen²,⁷,¹¹, Susanne Brix¹², Karsten Kristiansen²,⁷, Xun Xu²,³, Junhua Li²,³,⁸,¹³, Ruifang Wu⁴,⁵,* and Huijue Jia²,³,⁸,¹⁰,*

¹BGI Education Center, University of Chinese Academy of Sciences, Shenzhen 518083, China, ²BGI-Shenzhen, Shenzhen 518083, China, ³China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China, ⁴Peking University Shenzhen Hospital, Shenzhen 518036, China, ⁵Shenzhen Key Laboratory on Technology for Early Diagnosis of Major Gynecological diseases, Shenzhen, PR China, ⁶BGI genomics, BGI-Shenzhen, Shenzhen 518083, China, ⁷Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, Universitetsparken 13, 2100 Copenhagen, Denmark, ⁸Shenzhen Key Laboratory of Human Commensal Microorganisms and Health Research, BGI-Shenzhen, Shenzhen 518083, China, ⁹James D. Watson Institute of Genome Sciences, Hangzhou310000, China, ¹⁰Macau University of Science and Technology, Taipa, Macau 999078, China, ¹¹Institute of Marine Research (IMR), Postboks 1870, Nordnes, N-5817, Bergen, Norway, ¹²Department of Biotechnology and Biomedicine, Technical University of Denmark, Soltofts Plads, 2800 Kongens. Lyngby, Denmark and ¹³School of Bioscience and Biotechnology, South China University of Technology, Guangzhou 510006, China

∗Correspondence address. Huijue Jia. BGI-Shenzhen, Shenzhen 518083, China. E-mail: jiahuijue@genomics.cn or http://orcid.org/0000-0002-3592-126X or Ruifang Wu. Peking University Shenzhen Hospital, Shenzhen 518036, China. E-mail: wurf100@126.com

†These authors contributed equally to this work.

Abstract

Background: The human uterus is traditionally believed to be sterile, while the vaginal microbiota plays an important role in fending off pathogens. Emerging evidence demonstrates the presence of bacteria beyond the vagina. However, a microbiome-wide metagenomic analysis characterizing the diverse microbial communities has been lacking. Results: We performed shotgun-sequencing of 52 samples from the cervical canal and the peritoneal fluid of Chinese women of reproductive age using the Illumina platform. Direct annotation of sequencing reads identified the taxonomy of bacteria, archaea, fungi and viruses, confirming and extending the results from our previous study. We replicated our previous findings in another 24 samples from the vagina, the cervical canal, the uterus and the peritoneal fluid using the BGISEQ-500 platform revealing that microorganisms in the samples from the same individuals were largely shared in the entire reproductive tract. Human sequences made up more than 99% of the 20GB raw data. After filtering, vaginal microorganisms...
were well covered in the generated reproductive tract gene catalogue, while the more diverse upper reproductive tract microbiota would require greater depth of sequencing and more samples to meet the full coverage scale. **Conclusions:** We provide novel detailed data on the microbial composition of a largely unchartered body site, the female reproductive tract. Our results indicated the presence of an intra-individual continuum of microorganisms that gradually changed from the vagina to the peritoneal fluid. This study also provides a framework for understanding the implications of the composition and functional potential of the distinct microbial ecosystems of the female reproductive tract in relation to health and disease.

**Keywords:** metagenomics; microbiota; female upper reproductive tract

## Background

Evolution of the female reproductive tract has resulted in complex and unique structures such as the uterus, cervix and the vagina. The human vagina hosts trillions of bacteria that can significantly impact the health of women and their neonates. The cervix has traditionally been regarded to function as a perfect barrier between the vagina and uterus leading to the assumption that the upper reproductive tract constitutes a sterile environment. However, judging from evidence in insects and other animals, humans are probably no exception with regard to possible vertical transmission of the mothers’ microbiota before birth [1]. Thus, in humans, bacterial DNA has been detected in the placenta [2, 3]. Based on our recent analyses using 16S rRNA gene amplicon sequencing, the upper reproductive tract, including cervix, uterus, fallopian tubes, and peritoneal fluid, harbors diverse communities of bacteria, though at low abundance [4]. Recent studies of female reproductive tract microbiota have mainly focused on the vagina using 16S rRNA gene amplicon sequencing [5–7]. Studies using 16S rRNA gene amplicon sequencing have limitations in relation to lower taxonomic resolution and the lack of ability to perform species-specific functional inference. Metagenomic shotgun sequencing can address these limitations, but only a few studies have applied metagenomic shotgun sequencing to the vaginal microbiota [8], and no studies have characterized the compositional range of the upper reproductive tract microbiome using metagenomic analysis. The present study is the first to provide metagenomic data from the female upper reproductive tract.

## Data Description

Samples of six locations (CL, lower third of vagina; CU, posterior fornix; CV, cervical mucus drawn from the cervical canal; ET, endometrium; FLL and FRL, left and right fallopian tubes; and PF, peritoneal fluid from the pouch of Douglas) throughout the female reproductive tract from 137 Chinese women of reproductive age undergoing surgery for conditions not known to involve the male reproductive tract from 137 Chinese women of reproductive age undergoing surgery for conditions not known to involve the female upper reproductive tract. The sample selection was based on the data from CV and PF samples [4], which we identified as robust representations of the overall samples. Since higher amounts of DNA is required for shotgun-sequencing compared to 16S rRNA gene amplicon sequencing, more stringent criteria for selection of samples were applied including individual sub-clusters representation and sufficient DNA amounts (see details in the Methods section). According to the cluster criterion, clustering results based on the relative abundance of operational taxonomic units (OTUs) in the PF and CV samples showed that the samples marked with red (all containing DNA > 1 μg) were well distributed among all collected samples (Supplementary Fig. S2), and therefore selected for shotgun sequencing in this study. Hence, 25 PF and 25 CV samples were selected for sequencing on the BGISEQ-500 platform. After quality control, high-quality reads were aligned to hg 19 using SOAP and GRCh38 using DeconSeq to remove human reads (see details in the Methods section). We therefore abandoned the strategy of host DNA removal prior to shotgun metagenomics sequencing.

The sample selection was based on the data from CV and PF samples [4], which we identified as robust representations of the overall samples. Since higher amounts of DNA is required for shotgun-sequencing compared to 16S rRNA gene amplicon sequencing, more stringent criteria for selection of samples were applied including individual sub-clusters representation and sufficient DNA amounts (see details in the Methods section). According to the cluster criterion, clustering results based on the relative abundance of operational taxonomic units (OTUs) in the PF and CV samples showed that the samples marked with red (all containing DNA > 1 μg) were well distributed among all collected samples (Supplementary Fig. S2), and therefore selected for shotgun sequencing in this study. Hence, 25 PF and 25 CV samples were selected for sequencing on the Illumina HiSeq 4000 platform. After quality control, high-quality reads were aligned to hg 19 using SOAP and GRCh38 using DeconSeq to remove host reads (see details in the Methods section). We observed an average host contamination of 99.72% for CV and 79% for PF samples (Supplementary Fig. S1a). Furthermore, the bacterial composition was altered following kit treatment when compared with the control group (Supplementary Fig. S1b). We therefore abandoned the strategy of host DNA removal prior to shotgun metagenomics sequencing.

## Analyses and Discussion

### Metagenomic sequencing

According to shotgun sequencing of vaginal samples in the Human Microbiome Project and of placental samples by Aagaard et al., more than 90% of the sequences were derived from human host DNA [2, 9]. To overcome this problem, we first tested a commercial kit that removes human DNA by binding and precipitating CpG-methylated DNA. Unfortunately, after the kit treatment, a considerable amount (99.9% for CV sample and 79% for PF sample) of host DNA still remained (Supplementary Fig. S1a). Furthermore, the bacterial composition was altered following kit treatment when compared with the control group (Supplementary Fig. S1b). We therefore abandoned the strategy of host DNA removal prior to shotgun metagenomics sequencing.

### Metagenomic sequencing

Analyses and Discussion

We provide novel detailed data on the microbial composition of a largely unchartered body site, the female reproductive tract. Our results indicated the presence of an intra-individual continuum of microorganisms that gradually changed from the vagina to the peritoneal fluid. This study also provides a framework for understanding the implications of the composition and functional potential of the distinct microbial ecosystems of the female reproductive tract in relation to health and disease.
The overall microbiome composition of the cervical canal and the peritoneal fluid of reproductive-age women. Cumulative bar charts of the main taxa at domain (A) and family (B) levels in CV and FF samples. (C) Compositional overlap at the family level of CV and PF samples from the same individuals. Relative number of reads was calculated as \( N_p \times \frac{\phi_p}{m} \), where \( \phi_p \) is the number of reads within \( p \) taxa in a sample, \( \phi \) is the total number of reads within a sample, and \( m \) is the median number of reads within all 50 samples. When \( p \) taxa is shared by CV and FF samples from the same individuals and at the same time, both \( N_p \) values are higher than 0.1% \( \times m \), the \( p \) taxa is included in the cumulative bar charts. Taxa names (B, C) in black, purple, and blue denote bacteria, eukaryotes, and viruses, respectively.
A diverse microbiome in the cervical canal and the peritoneal fluid of reproductive age women

To obtain an overview of the overall composition of the vagino-uterine microbiome, we used Kraken to directly assign sequencing reads to all types of microbial taxa [10]. The dominant Lactobacillus spp. in CV and Pseudomonas spp. in PF were detected in the present study in accordance with those found in the previous study [4]. In addition, methane-producing archaea, yeasts, herpesviruses, papillomaviruses, and bacteriophages were also identified (Fig. 1A, 1B).

The abundance of these taxonomic units varied among samples, and those constituting more than 0.1% of the total reads number and identified in the CV and PF samples from the same individual are shown in Fig. 1C.
We annotated the genes in the gene catalog according to the Kyoto Encyclopedia of Genes and Genomes (KEGG) [14]. The matched genes in the PF samples (15,316 genes) were all covered within the CV samples (39,087 genes). Comparing the CV and the PF samples in the distribution of KEGG pathways, the PF samples showed a greater proportion of genes involved in carbohydrate metabolism, replication and repair, membrane transport, and drug resistance, whereas genes involved in translation, energy metabolism, and metabolism of cofactors and vitamins were enriched in the CV samples (Fig. 4). In relation to KEGG orthology (KO) modules, CV samples showed enrichment of transport systems for thiamine, cystine, taurine, and putative ABC transport systems compared to the PF samples. Regulatory systems of aerobic and anaerobic respiration, osmotic stress response, and multicellular behavior control were also enriched in the CV samples (Supplementary Table S3).

Methods

Sample description
A total of 137 Chinese women of reproductive age undergoing surgery for conditions not known to involve infection (hysteromyoma, adenomyosis, endometriosis, and salpingectomy) were enrolled in this study (Supplementary Table S1). Samples were taken from the CL, CU, and CV on the day of the clinical visit without any prior disturbance. Depending on the clinical conditions, laparoscopy or laparotomy was performed, and samples from the ET, FLL, FRL, and PF were taken during surgery (Supplementary Table S1). The study was approved by the institutional review boards at Peking University Shenzhen Hospital and BGI-Shenzhen, and all women provided written informed consent. The subject exclusion criteria, sampling, and DNA extraction methods can be found in [4].

To test the effect of experimental removal of human DNA, one CU sample and one CV sample were used for shotgun sequencing on the Illumina HiSeq2000 platform with or without prior removal of human DNA. The NEBNext Microbiome DNA Enrichment Kit was used according to the manufacturer’s instructions, with 10 μg input DNA per sample.

We performed a prior selection of samples to undergo shotgun-sequencing. The selection was based on the data from CV and PF samples [4] using the following two criteria: samples should represent individual subclusters when subjected to hierarchical (centroid-linkage) clustering based on relative abundances of OTUs from 16S rRNA gene amplicon sequencing and the amount of DNA should be greater than 1 μg. The samples with good scattering in different clusters based on the relative abundance of OTUs in the PF and CV samples were selected for shotgun sequencing on the Illumina HiSeq4000 platform.

We replicated the findings in 24 additional samples on the BGISEQ-500 platform, where additional sites (CL, CU, CV, ET, and PF) of six women were analyzed. To meet the need of library construction, the amount of DNA in all 24 samples was greater than 1 μg. Three qualified samples from each woman were set as a threshold.

Metagenomic shotgun sequencing
Library construction and shotgun sequencing using the Illumina HiSeq2000/4000 platforms (insert size 350 bp; 100 bp of PE reads; two replicate libraries were constructed for each lane) and BGISEQ-500 (100 bp of SE reads; one library was constructed for each lane) were performed as previously described [15] (and...
Figure 4: KEGG pathway classification of the vagino-uterine microbiome. Comparison of CV (red) and PF (blue) data based on KEGG annotation, which emphasizes functional similarity of the CV and PF microbiota.

see protocol in protocols.io [16]). The quality control of sequencing data from the HiSeq and BGISEQ platforms was also performed as previously described [15]. Human sequences were eliminated by alignment to the hg19 reference genome using SOAP2.22 (SOAPaligner/soap2, RRID:SCR_005503). As the resulting data still contained human sequences, a more stringent procedure using DeconSeq by aligning data to the GRCh38 reference genome was applied [17].

**Taxonomic assignment of sequencing reads**

High-quality, non-human sequences were tentatively assigned to microbial taxa using Kraken using default parameters (Kraken, RRID:SCR_005484) [10]. For pair-end reads Kraken automatically concatenated the pairs together with a single N between the sequences using default parameters, and according to the manual this software raises the sensitivity by about 3 percentage points over classifying the sequences as single-end reads.

**Construction of a gene catalog**

The high-quality, non-human sequencing reads of the 52 samples sequenced using the Illumina HiSeq platform were de novo assembled into contigs using IDBA-UD (IDBA-UD (RRID:SCR_011912)) [18]. We used the same strategy as described in previous studies [12, 13], where genes were predicted from the contigs by MetaGeneMark [19], and highly similar genes (95% identity, 90% overlap) were removed as redundant using CD-HIT (CD-HIT, RRID:SCR_007105) [20]. Functional annotations were made by BLASTP (v2.2.24) based on the KEGG (v76) databases (KEGG, RRID:SCR_012773)[14].
Availability of supporting data

The sequencing data after filtering out low-quality and host reads are available via the EBI database using the accession number PRJEB24147. Additional supporting data are available via the GigaScience GigaDB database [21].

Additional files

Supplementary Figure S1: Evaluation of the NEBNext Microbiome DNA Enrichment Kit by two comparative strategies. Sample names suffixed by “-HR” represent DNA samples that were subjected to host DNA removal before shotgun sequencing, while sample names suffixed with A represent DNA samples that were subjected to shotgun sequencing directly (a). The table data shows the obtained read number, and remaining reads after removal of host DNA reads in the two samples. b) Influence of host DNA presence on bacterial DNA identification during shotgun sequencing. The plots display the compositional difference amongst major bacteria genera in samples with and without (-HR) host DNA presence. Data were analyzed by mapping reads to the ICG bacterial reference gene catalog [12].

Supplementary Figure S2: Samples selected for metagenomic sequencing. Hierarchical clustering of CV (a) and PF (b) samples based on the relative abundances of OTUs. Samples which represent individual sub-clusters and hold DNA amounts above 1 μg were selected for shotgun-sequencing (red).

Supplementary Table S1: Phenotypic information for the 137 subjects.

Supplementary Table S2: Statistics for each shotgun-sequenced sample.

Supplementary Table S3: The distribution of the Modules in the female reproductive tract.

Abbreviations

CL: lower third of vagina; CU: posterior fornix; CV: cervical mucous drawn from the cervical canal; ET: endometrium; FLL: left fallopian tube; FRL: right fallopian tube; KEGG: Kyoto Encyclopedia of Genes and Genomes; OTU: operational taxonomic units; PE: paired-end; PF: peritoneal fluid from the pouch of Douglas; SE: single-end.

Competing interests

The authors declare that they have no competing interests.

Funding

The study was supported by the Shenzhen Municipal Government of China (JCYJ20160229172757249, JCYJ2015061090833370) and a grant from the Macau Technology Development Fund (102/2016/A3).

Author contributions

H.J. and R.W. conceived and directed the project. W.W., J.D., L.Z., H.D., H.T., and R.W. performed the clinical diagnosis and sample collection. F.L., C.C., Z.W., and L.H. performed the bioinformatic analyses and prepared display items. C.C., F.L., Z.W., X.Z., J.L., and H.J. wrote the first version of the manuscript. L.M., S.B., and K.K. revised the manuscript. All authors contributed to the final revision of the manuscript.

Acknowledgements

We gratefully acknowledge colleagues at BGI-Shenzhen for DNA quality control, library construction, sequencing, and helpful discussions.

References

20. Li W, Godzik A. Cd-hit: a fast program for clustering and