Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Metagenome datasets from women with polycystic ovary syndrome from Irkutsk, Eastern Siberia, Russia

Natalia Belkova*, Elizaveta Klimenko, Naliia Vilson, Alexsandra Sambialova, Daria Markova, Ilia Igumnov, Larisa Suturina

Scientific Centre for Family Health and Human Reproduction Problems, 16, Timiryazeva Street, Irkutsk 664003, Russia

ARTICLE INFO

Article history:
Received 18 May 2020
Revised 29 July 2020
Accepted 31 July 2020
Available online 6 August 2020

Keywords:
Metagenomics
Human microbiome
Gut
Feces
Saliva
Serum
Women
Polycystic ovary syndrome
PCOS

ABSTRACT

For the metagenomic characterization of potential taxonomic and functional diversity of microorganisms associated with polycystic ovary syndrome (PCOS) in women, we surveyed five women with PCOS and collected samples of feces, saliva, and serum. After quality processing, we have obtained from 915,594 to 3,880,379 reads; these 16,693 sequences had ribosomal RNA genes, 2,091,990 sequences contained predicted proteins with known functions, and 3,750,261 sequences had predicted proteins with unknown functions. Host DNA accounted for ca. 0.03% and less in datasets of fecal samples, from 1.41 to 24.94% in saliva samples; the remaining sequences were attributed to archaeal, bacterial, or viral DNA. In serum, from 38.18 to 75.77% were characterized as fragments of the human genome, but the remaining sequences were unidentified. Among microbes, a total of one archaeal and eight bacterial phyla were revealed. Viral DNA was detected in several fecal and one saliva sample and was classified as C2likevirus, Flavivirus, and *Streptococcus* bacteriophage. The metagenome sequence data were deposited at NCBI SRA as BioProject No. PRJNA625611.

* Corresponding author.
E-mail address: nlbelkova@gmail.com (N. Belkova).

https://doi.org/10.1016/j.dib.2020.106137

Specifications Table

Subject	Microbiology
Specific subject area	Metagenomics
Type of data	Table Figure Metagenome sequences
How data were acquired	Shotgun DNA sequencing using Illumina NextSeq 550 platform.
Data format	Raw data Analyzed
Parameters for data collection	Feces, saliva, and serum sampled from five untreated women with PCOS, identified at the outpatient department of the Scientific Center for Family Health and Human Reproduction Problems (SCFHHRP) between September and November 2019. PCOS was verified by the presence of at least two criteria from hyperandrogenism, oligo-anovulation, and polycystic ovarian morphology. The data collection was approved by the Ethics Committee of the SCFHHRP.
Description of data collection	Total DNA from fecal and saliva samples was extracted using HostZERO Microbial DNA Kit (Zymo Research, USA). Additionally, fecal samples were treated with the Quick-DNA Fecal/Soil Microbe Miniprep Kit (Zymo Research, USA). Total DNA from serum was isolated using AmpliSens® DNA-sorb-B (InterLabService, Russia). Paired-end libraries were prepared using a Nextera XT DNA Library Preparation Kit, and shotgun sequencing was done using NextSeq 500/550 High Output Kit (300 Cycles) with the Illumina NextSeq 550 platform.
Data source location	Institution: Scientific Center for Family Health and Human Reproduction Problems City: Irkutsk Country: Russia
Data accessibility	Raw data were deposited to NCBI Repository name: SRA Data identification number: BioProject PRJNA625611, BioSamples from SAMN14603886 to SAMN14603890 Direct URL to data: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA625611

Value of the Data

- These are the first shotgun metagenome datasets on feces, saliva, and serum sampled from women with polycystic ovary syndrome (PCOS) from the Eastern Siberia of Russia.
- The data provides valuable information about the diversity and functional potential of archaeal, bacterial, and dsDNA viral communities in the gut and saliva microbiome of women with PCOS.
- These data are useful for the definition of human microbiome peculiarities associated with PCOS.
- Raw sequence data can be used for various additional bioinformatics processing.

1. Data Description

Polycystic ovary syndrome (PCOS) is the most frequent endocrine and metabolic disorder in premenopausal women [1]. Direct interactions of the gut microbiome (GM) and PCOS have been shown in animal model studies [2]. In contrast, the peculiarities of GM associated with PCOS have been studied for female adolescents and women of reproductive years [2,3]. In recent years, pronounced dysbiosis of gut microbial communities has been shown in metabolic diseases (obesity and insulin resistance) associated with PCOS [4,5]. Although the search for correlations be-

Table 1.
Annotation of datasets from the feces, saliva, and serum from women with PCOS represented by Metaphlan2.

Sample ID	Human sequences (%)	Sequences annotated to microorganisms (%)		
		Achaea	Eubacteria	Virus
Feces				
48-f1*	0.003	0	99.997	0
49-f1	0.01	0.19	99.79	0
50-f1	0.01	0	99.92	0.06
51-f1	0.03	0.23	99.66	0.07
48-f2**	0.01	0	99.99	0
49-f2	0.01	0.04	99.94	0
50-f2	0.01	0	99.68	0.30
52-f2	0	0.27	99.72	0
Saliva				
48-sl	3.21	0	96.79	0
49-sl	4.46	0	95.54	0
50-sl	1.41	0	98.59	0
51-sl	24.94	0	75.06	0
52-sl	1.95	0	94.95	3.09
Serum				
48-sr	51.91	ND	ND	ND
49-sr	75.77	ND	ND	ND
50-sr	67.48	ND	ND	ND
51-sr	38.18	ND	ND	ND

Comments: * fecal samples with a marker –f1 were isolated with HostZERO Microbial DNA Kit (Zymo Research, USA), ** fecal samples with a marker –f2 were isolated with Quick-DNA Fecal/Soil Microbe Miniprep Kit (Zymo Research, USA); ND, not detected.

tween the GM and metabolic diseases has become a hot point for recent research, the geography of the studies on PCOS is very narrow and included mainly Chinese studies. A few studies have been carried out in the USA, Austria, and Spain [2–8].

The dataset contains raw sequencing data obtained by the shotgun sequencing of feces, saliva, and serum from five women from Irkutsk, Eastern Siberia, Russia. The data files (reads in FASTQ format) were deposited to NCBI SRA as BioProject No. PRJNA625611. Raw data contained 38,261,075 reads totaling 577,737,325 bp with an average length of 151 bps, 127,895 reads failed to pass the QC pipeline. While of the sequences that passed QC, 16,693 sequences had ribosomal RNA genes, 2,091,990 sequences contained predicted proteins with known functions, and 3,750,261 sequences had predicted proteins with unknown functions.

Fecal samples contained from 1,275,709 to 3,044,387 sequences, saliva, from 1,701,648 to 3,880,379 sequences, whereas serum contained from 915,594 to 2,765,266 sequences. Host DNA accounted for ca. 0.03% and less in datasets of fecal samples, from 1.41 to 24.94% in saliva samples. The remaining sequences were annotated to archaeal, bacterial, or viral DNA (Table 1). In serum from 38.18 to 75.77% of sequences were characterized as fragments of the human genome, but the remaining sequences were unidentified.

Metaphlan2 allowed the revealing of a total of one archaeal and eight bacterial phyla in the gut and saliva of women with PCOS. Information about the structure of microbial communities in gut and saliva metagenome is presented in Fig. 1. Abundant genera in the gut metagenome were *Faecalibacterium, Bacteroides, Dialister, Prevotella, Alistipes*, and *Subdoligranulum*, representing from 8 to 20% of total sequences. At the same time, *Prevotella, Veillonella, Neisseria, Haemophilus*, and *Porphyromonas* were dominant in the saliva metagenome and varied from 5 to 31%. Viral DNA was detected in several fecal and one saliva sample and was classified as C2likevirus, Flavivirus, and *Streptococcus* bacteriophage.

These are the first datasets on the diversity of archaeal, bacterial, and dsDNA viral communities in the gut and saliva metagenome of women with PCOS from Eastern Siberia of Russia, based on Illumina sequencing technology. Datasets were deposited to the SRA NCBI database as BioProject No. PRJNA625611.

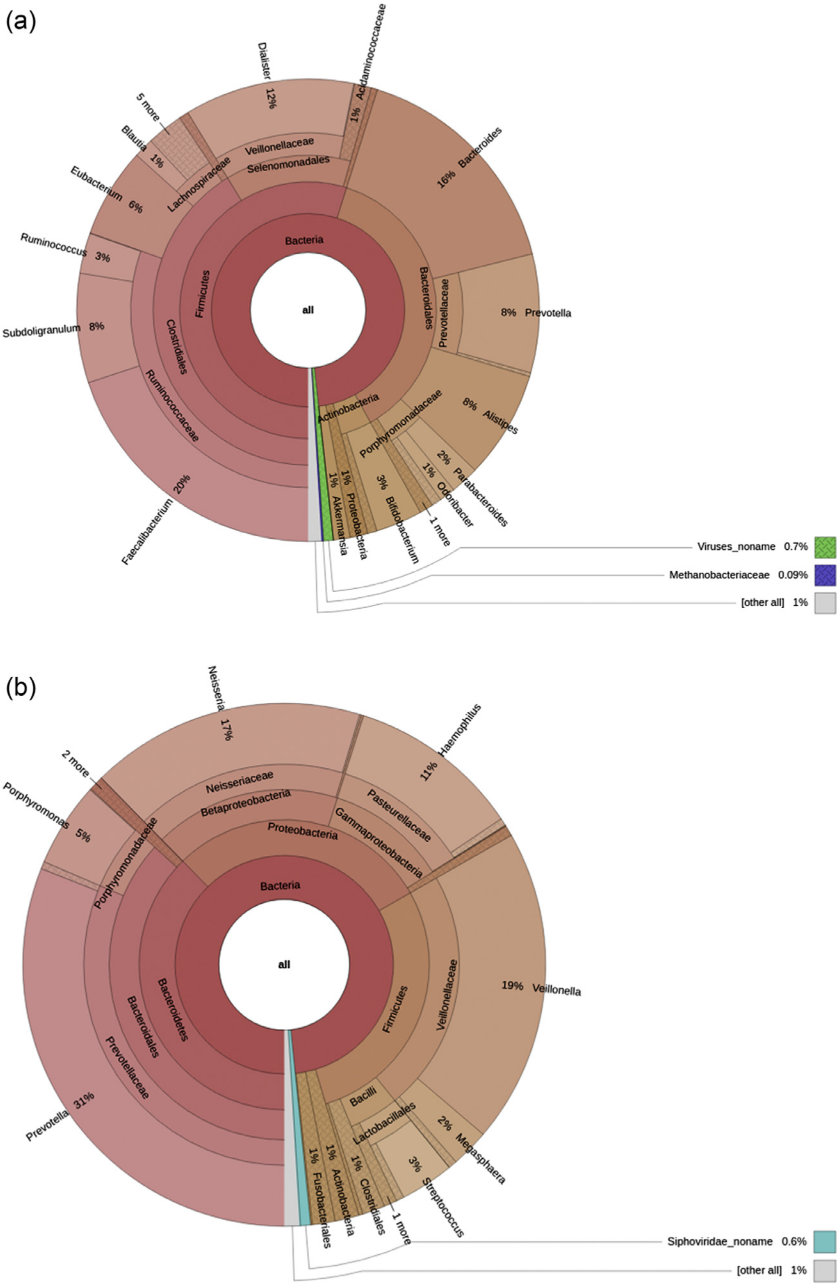

Fig. 1. The structure of microbial communities in gut (**A**) and saliva (**B**) metagenome.

Table 2
The clinical characteristics of studied participants.

Clinic parameter	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5
Age, years	29	24	23	24	20
BMI, kg/m^2	19.13	22.8	20.0	21.8	21.97
SBP/DBP, mm Hg	120/80	128/79	114/70	127/79	127/72
Oligo-anovulation (+/–)	+	+	+	+	+
Hyperandrogenemia (+/–)	+ (TT ↑; FAI ↑; DHEAS ↑)	+ (FT ↑)	+ (TT ↑; DHEAS ↑)	+ (TT ↑)	+ (TT ↑; FT ↑; DHEAS ↑)
Hirsutism (+/–)	+ (FG score = 5)	– (FG score = 2)	+ (FG score = 6)	+ (FG score = 7)	+ (FG score = 5)
PCOM (+/–)	+ (AFC > 25; OV = 21; 19.7 mm^3)	+ (AFC = 15; OV=8.7; 10.2 mm^3)	+ (AFC > 20; OV=24; 22 mm^3)	+ (AFC=28; 32 OV=19.7; 22 mm^3)	– (AFC = 10; 10 OV=8.58; 9.94 mm^3)

Comments: BMI, body mass index; PCOM, polycystic ovarian morphology; TT, total testosterone; FT, free testosterone; FAI, free androgen index; DHEAS, dehydroepiandrosterone sulfate; FG, Ferriman–Gallwey score; AFC, antral follicle count; OV, ovarian volume.

2. Experimental design, materials, and methods

2.1. Study cohort

Feces, saliva, and serum were sampled from women referred to the outpatient department ("Center of Innovative Medicine") of the Scientific Center for Family Health and Human Reproduction Problems (Irkutsk). The datasets comprised participants who met the inclusion criteria and did not meet the exclusion criteria. Inclusion criteria were as follows: premenopausal women with a voluntary signed informed consent and the presence of any two or three PCOS criteria (hyperandrogenism, oligo-anovulation, and polycystic ovarian morphology by pelvic ultrasound) [9]. The exclusion criteria included: hyperprolactinemia, hypothyroidism, none-classic adrenal hyperplasia (NCAH), tumors, obesity, diabetes, glucose intolerance, intake of oral contraceptives, and other hormonal medicines, as well as antimicrobial preparations and insulin sensitizers. General characteristics of the patients included in the datasets are shown in Table 2. The data collection was approved by the Ethics Committee of SCFHHRP (protocol N 6 from 06.09.2019).

2.2. Sample collection and treatment

Fecal sampling and total DNA isolation were conducted according to the requirements of International Human Microbiome Standards, IHMS. Patients were instructed in the proper feces and saliva sample collection. Patients were provided with refrigerants for cold transportation of biomaterial to the laboratory within 2 h after collection. For fecal samples, the instruction was done to prevent contamination with urine and toilet water. Saliva samples were collected in the morning after an overnight fast. Patients were instructed not to brush their teeth and to drink only water before saliva sampling. Saliva was collected in the mouth for several minutes and then voided into sterile plastic tubes. This process was repeated until the desired volume of 1–2 ml was reached. Saliva samples were immediately cooled on ice and transported to the laboratory. In the laboratory, all samples were aliquoted, one of which was used directly for DNA extraction; the others were frozen and stored at −80 °C.

Venous blood was collected in the morning after an overnight fast from the ulnar vein into a tube with SiO_2. The sample was evaluated for hemolysis. A sample of the patients without hemolysis was defecated at room temperature for 30 min until a clot formed completely. Then,

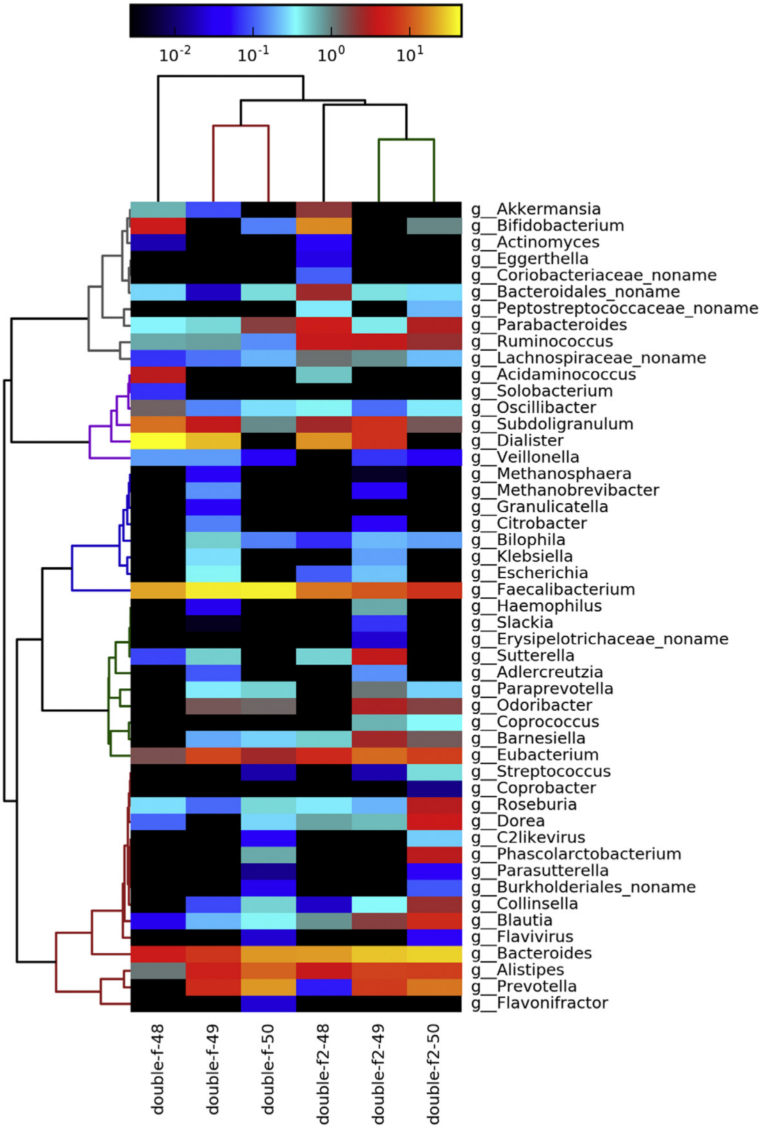

Fig. 2. Heat map of bacterial communities structure of gut samples. DNA was isolated with two commercial kits.

centrifugation was performed at 3000 rpm for 15 min. The resulting serum was poured into a sterile tube and stored in a refrigerator; DNA was isolated within 2 h after sampling.

Total DNA from fecal and saliva samples was extracted using HostZERO Microbial DNA Kit (Zymo Research, USA), which supplies a step to allow the elimination of host DNA. Additionally, fecal samples were treated with the Quick-DNA Fecal/Soil Microbe Miniprep Kit (Zymo Research, USA). The differences in the host and bacterial DNA content between the fecal samples treated with two commercial kits presented in Table 1 and Fig. 2. Total DNA from serum was isolated using AmpliSens® DNA-sorb-B (InterLabService, Russia). DNA quality was tested by 0.8% agarose gel electrophoresis, while the concentration was determined using a Qubit 4 Fluorometer (Thermo Scientific, USA).

2.3. Library preparation and sequencing

Paired-end libraries were prepared using a Nextera XT DNA Library Preparation Kit, and shotgun sequencing was done using a NextSeq 500/550 High Output Kit (300 Cycles) with the Illumina NextSeq 550 platform according to the manufacturer instruction.

2.4. Analysis of metagenome datasets

Following quality control using FastQC v0.11.8 [10], sequences were trimmed (Trimmomatic [11]) and merged. The host sequence was detected and filtered using BWA [12] and SAMtools [13]. Analysis and taxonomy annotation of output data was performed through Metaphlan2 [14] (Table 1) and Krona [15] (Fig. 1).

Declaration of Competing Interest

The authors declare that they have no known competing for financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Acknowledgments

This work was done according to the program of the Ministry of Science and Higher Education of the Russian Federation [program number 0416–2020–008].

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2020.106137.

References

[1] D. Lizneva, L. Suturina, W. Walker, S. Brakta, L. Gavrilova-Jordan, R. Azziz, Criteria, prevalence, and phenotypes of polycystic ovary syndrome, Fertil. Steril. 106 (2016) 6–15, doi:10.1016/j.fertnstert.2016.05.003.

[2] X. Qi, C. Yun, L. Sun, J. Xia, Q. Wu, Y. Wang, L. Wang, Y. Zhang, X. Liang, L. Wang, F.J. Gonzalez, A.D. Patterson, H. Liu, L. Mu, Z. Zhou, Y. Zhao, R. Li, P. Liu, C. Zhong, Y. Pang, C. Jiang, J. Qiao, Gut microbiota-bile acid-interleukin-22 axis orchestrates polycystic ovary syndrome, Nat. Med. 25 (2019) 1225–1233, doi:10.1038/s41591-019-0509-0.

[3] B. Jobira, D.N. Frank, L. Pyle, L.J. Silveira, M.M. Kelsey, Y. Garcia-Reyes, C.E. Robertson, D. Ir, K.J. Nadeau, M. Cree-Green, Obese adolescents with PCOS have altered biodiversity and relative abundance in gastrointestinal microbiota, J. Clin. Endocrinol. Metab. 105 (2020) dgz263, doi:10.1210/clinem/dgz263.

[4] L. Zhou, Z. Ni, W. Cheng, J. Yu, S. Sun, D. Zhai, C. Yu, Z. Cai, Characteristic gut microbiota and predicted metabolic functions in women with PCOS, Endocr. Connect. 9 (2020) 63–73, doi:10.1530/EC-19-0522.

[5] B. Zeng, Z. Lai, L. Sun, Z. Zhang, J. Yang, Z. Li, J. Lin, Z. Zhang, Structural and functional profiles of the gut microbial community in polycystic ovary syndrome with insulin resistance (IR-PCOS): a pilot study, Res. Microbiol. 170 (2019) 43–52, doi:10.1016/j.resmic.2018.09.002.

[6] L. Lindheim, M. Bashir, J. Münzker, C. Trummer, V. Zachhuber, B. Leber, A. Horvath, T.R. Pieber, G. Gorkiewicz, V. Stadlbauer, B. Obermayer-Pietsch, Alterations in gut microbiome composition and barrier function are associated with reproductive and metabolic defects in women with polycystic ovary syndrome (PCOS): a pilot study, PLoS ONE 12 (2017) e0168390, doi:10.1371/journal.pone.0168390.

[7] M. Insenser, M. Murri, R. Del Campo, M.Á. Martínez-García, E. Fernández-Durán, H.F. Escobar-Morreale, Gut microbiota and the polycystic ovary syndrome: influence of sex, sex hormones, and obesity, J. Clin. Endocrinol. Metab. 103 (2018) 2552–2562, doi:10.1210/jc.2017-02799.

[8] P.J. Torres, M. Siakowska, B. Banaszewska, L. Pawelczyk, A.J. Duleba, S.T. Kelley, V.G. Thackray, Gut microbial diversity in women with polycystic ovary syndrome correlates with hyperandrogenism, J. Clin. Endocrinol. Metab. 103 (2018) 1502–1511, doi:10.1210/jc.2017-02153.

[9] H.J. Teede, M.L. Misso, M.F. Costello, A. Dokras, J. Laven, L. Moran, T. Piltonen, R.J. Norman, International PCOS Network, Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome, Fertil. Steril. 110 (2018) 364–379, doi:10.1016/j.fertnstert.2018.05.004.

[10] S. Andrews, FastQC: A Quality Control Tool For High Throughput Sequence Data, (2010) Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ Accessed 6 October 2011.

[11] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014) 2114–2120, doi:10.1093/bioinformatics/btu170.

[12] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (2009) 1754–1760, doi:10.1093/bioinformatics/btp324.

[13] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 genome project data processing subgroup, the sequence alignment/map (SAM) format and SAMtools, Bioinformatics 25 (2009) 2078–2079, doi:10.1093/bioinformatics/btp352.

[14] D.T. Truong, E.A. Franzosa, T.L. Tickle, M. Scholz, G. Weingart, E. Pasolli, A. Tett, C. Huttenhower, N. Segata, MetaPhlAn2 for enhanced metagenomic taxonomic profiling, Nat. Methods 12 (2015) 902–903, doi:10.1038/nmeth.3589.

[15] B.D. Ondov, N.H. Bergman, A.M. Phillippy, Interactive metagenomic visualization in a Web browser, BMC Bioinform. 12 (2011) 385, doi:10.1186/1471-2105-12-385.