Recovery of Human Gut Microbiota Genomes Substantially With Third-generation Sequencing

Yanfei Li

Shanghai Pudong New District Zhoupu Hospital

Yueling Jin

Shanghai Pudong New District Zhoupu Hospital

Haoying Pan

Shanghai Pudong New District Zhoupu Hospital

Jianming Zhang

Shanghai OE biotech company

Lan Wu

Shanghai University of Medicine and Health Sciences

Dingsheng Liu

Shanghai Pudong New District Zhoupu Hospital

Jinlong Liu

Shanghai University of Medicine and Health Sciences

Junwei Shen (✉ xingkezh5@163.com)

Tongji University Affiliated East Hospital: Shanghai East Hospital <https://orcid.org/0000-0002-2542-3291>

Jing Hu

Tongji University

Methodology

Keywords: human gut microbiota, genome, third-generation sequencing, bacteria

Posted Date: October 9th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-87441/v1>

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

[Read Full License](#)

Abstract

Background

Human gut microbiota modulates normal physiological functions, such as the maintenance of barrier homeostasis and the modulation of metabolism, and various chronic diseases including type 2 diabetes and gastrointestinal cancer. Despite decades of researches, the composition of the gut microbiota remains unexplored and unidentified.

Results

Here we established an effective extraction method to obtain high-quality gut microbiota genomic DNA and detected the samples with third-generation sequencing technology. We acquired a quite big data from each sample and assembled many reliable contigs. Not only enormous unknown genes, but also several new bacteria subspecies or species were identified.

Conclusions

This work provides a novel and reliable framework to recover gut microbiota genomes substantially, facilitating the understanding of the roles of the microbiome that underlie in human health and disease.

Background

There are growing evidences that gut microbiota-the human commensals-influences normal physiological functions from the maintenance of barrier homeostasis to the modulation of metabolism, inflammation, immunity, and development [1-3]. Further, its disorder is a potential risk factor of various chronic diseases such as type 2 diabetes, gastrointestinal cancer, and brain disorders [4-6]. Furthermore, the gut microbiota is one of the crucial factors which influence drug effectiveness owing to its effects on drug metabolism [7, 8]. Despite decades of researches, the composition of the gut microbiota remains unidentified. For example, through analyzing more than 150,000 human microbial genomes, Edoardo et al recapitulated nearly 5000 species-level genome bins, in which 77% without genomes in public repositories [9]. One of the main difficulties is that many microbes are uncultivable and hard to be isolated [9]. In addition, culture-dependent genomic researches are rare to large cohorts [10]. High-through sequencing that do not need to cultivate microbes, overcomes these shortage and becomes a powerful method for bacterial metagenome study.

There are two main methods for bacteria metagenome sequencing: the 16s rRNA-based sequencing and whole-metagenome shotgun sequencing (WGS). 16S rRNA-based sequencing is widely used in assessing microbial communities due to its low cost, time efficiency, and the ability to provide a full overview of the community [11, 12]. However, as only a single region of 16s rRNA in the genome is detected, the information it provides about the microbial community is very limited [13, 14]. Furthermore, it underestimates the diversity of the microbes. Compared with 16s rRNA sequencing, WGS technology

examines the whole bacterial genome, and provides a more accurate detection both at the species and diversity levels[15]. Through WGS technology, it is feasible to assemble the entire bacteria metagenome. However, there are several drawbacks in WGS technology owing to small sequencing reads (mostly less than 200 bp). For example, the sequence information is too little to assemble the de novo genome efficiently [16, 17]. As a result, many microbiota genomes are fragmented into lots of contigs [17]. Moreover, great phenotypic differences exist even between highly related strains of the same species [18, 19], and the differences between these strains are difficult to be distinguished by WGS. Therefore, there is a dire need to detect the bacteria metagenome through a more powerful sequencing method.

The third-generation sequencing (TGS) technology, also known as long-read sequencing, detects the isolated genome DNA without amplification, and produces surprising long reads (average 10–20 kb)[20]. Compared with the second-generation sequencing technology of 16s rRNA and WGS, it can detect much longer fragments, so TGS produces genome assemblies of unprecedented quality [20]. For example, Hui et al de novo assembled a chromosome-level reference genome of red-spotted grouper with TGS [21]. Besides, TGS has been used in bacterial genome detection yet. For instance, Johanna et al assemble a certain specific microbial species living in the vaginal with an abundance of more than 75% [22]. Thidathip et al compared the taxonomic abundance of gut microbiota of head and neck cancer patients; while the mean length is about 1 kb, which is much smaller than that of the buffalo genome (11.5 kb) [23, 24]. However, one main difficult for gut microbiota genomes detection is that it is tough to obtain high-quality genome DNA since the complex fecal environment. Therefore, there are few studies on gut microbiota with TGS yet. In this study, we have tried multiple methods and successfully extracted high-quality gut microbial genome DNA. Though detecting the gut microbiota metagenome via Single Molecule Real-Time (SMRT) Sequencing of Pacibio, we assembled the bacteria genomes efficiently and discovered many new contigs without genomes in public repositories and new bacteria species.

Results

Gut microbiota genome isolation and the whole characteristics of the TGS results

We analyzed the data of prokaryotes genomes in the National Center for Biotechnology Information (NCBI) and found that complete genomes account for only a small portion (<https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/>) (Fig. 1a, Supplementary Fig. 1a), which was consistent with the previous research [9]. In order to assemble de novo complete genomes of gut microbes, we tried to detect the gut microbiota metagenome through the TGS technology. To obtain high-quality gut microbiota genome samples for TGS, we tried our best to improve the integrity and the quantity of the genome DNA. In the beginning, the isolated genome DNA was fragmented (Supplementary Fig. 1b). After trying many different kits, we acquired a complete genome DNA with the kit from the MPbio company. Then we tried to improve the yield of the genome DNA. We isolated the top band and purified the gut microbiota genome DNA with a DNA Purification Kit. Finally, we acquired two high-quality samples: a 34-year old male and a 10-month old baby.

We then detected the whole metagenome of the two samples through TGS technology. We acquired 53G data from the baby sample, and 47G data from the adult sample. Through analyzing the genome database in NCBI, we found that most of the bacterial genomes are large than 0.5 Mb and the smallest genome is about 0.1 Mb (Supplementary Fig. 1b). Therefore, we chose 0.1 Mb and 0.5 Mb as cutoff values to analyze the assembly contigs (Fig. 1b). In the baby sample, there were 43 contigs larger than 0.5 Mb, and 248 contigs larger than 0.1 Mb. Similarly, in the adult sample, there are 33 contigs larger than 0.5 Mb, and 237 contigs larger than 0.1 Mb (Supplementary Tab. 1, Fig. 1b). To assess the reliability of our methods, we analyzed the sequences of 16S rRNA from the longest contig (Contig_511). All the six 16S rRNA sequences of contig 511 are almost the same, with only two different bases in the over 1.5 kb sequence (Supplementary Fig. 1c). Interestingly, the 16S sequence in contig 511 has fewer differences than that in the *E. coli* (U00096.3) (Supplementary Fig. 1c). These results suggest that the methods we used are available and reliable (Supplementary Fig. 1c). Moreover, the contig length was significantly related to the number of the coding regions (CDSs) (Fig. 1c). Through analyzing the function potential of the CDSs, we found that many CDSs were correlated with the metabolism of amino acids and carbohydrates. Interestingly, the function potential of a large number many genes were unknown, indicating that there are still numerous valuable genes worth investigating in gut microbes (Fig. 1d).

The aspects of the contigs in the two samples

Since most of bacteria genomes are over 0.5 Mb, we further analyzed the contigs over 0.5 Mb in the two samples. In these contigs, ten of them are circular ((Supplementary Tab. 2, Fig. 2a); and twelve contigs are more than 99% complete (Fig. 2b). Nine contigs are both circular and complete (>99%) (Fig. 2c). Therefore, we focus on the nine contigs and analyzed their features. The full-length 16S rRNA blast analyses showed that less than 50% (four of nine contigs) of these contigs could be completely matched with previous sequences in the NCBI databases (Fig. 2d and 2e) (indicated by the red arrow). Five of them are not completely matched, indicating that they might be new species or new subspecies (Fig. 2e). To investigate the taxonomic location of the nine contigs, we analyzed them with a phylogenetic analysis using the full-length 16S rRNA. Comparing with some common bacteria, including *Bacterioides fragilis* and *Escherichia coli*, we found that their taxonomic location in the phylogenetic tree successfully (Fig. 2e).

The genome analysis of the five non-match genomes.

Then we further analyzed the five genomes that cannot be completely matched. Their genome length were from 1.5 Mb to 3.5 Mb ((Supplementary Tab. 3, Fig. 3a). Compared with the huge difference in genome length, the average length of each gene was very similar (Fig. 3a). On the contrary, the difference of GC contents between these genomes were quite large (Fig. 3b). And afterwards, we analyze some special structures of the genomes, including non-coding RNA (ncRNA) and repeat sequences. Interestingly, the number of rRNA and tRNA in each genome was correlated with each other (Fig. 3c). Each

genome had all types of repeat sequences, and simple repeats were the major repeat type (Fig. 3d). Intriguingly, we found that all the five bacteria carry streptomycin resistance, which might result from long-term antibiotic use (Fig. 3e), suggesting that the extensively use of antibiotics might have irreversibly transformed the human gut microbial. To determine the classification they belong to, we blasted the full-length 16S rRNAs in NCBI databases (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome). Four of them (contig_82, contig_242, contig_318 and contig_511) had very high identities (>99%). On the contrary, the identity of contig_638 is less than 97% (Fig. 3f).

Phylogeny of four high identity non-matched genomes.

Then we analyzed the bacteria with identities more than 99%, and detected their accurate taxonomic location by phylogenetic tree analysis. We blasted the full-length 16s rRNAs of these genomes with the NCBI databases, and then selected the most related species (rank by identity) for further analysis. Contig_82 belongs to *Collinsella erofaciens* and might be a new subspecies (Fig. 4a). Similarly, contig_242 is a subspecies of *Megasphaera micronuciformis*; contig_318 is a subspecies of *Lactobacillus dextrinicus*; and contig_511 is a subspecies of *Lactobacillus smilis* (Fig. 4b-4d). The analysis of non-redundant database annotations confirmed these results (Supplementary Fig. 3a-3d). Then we compared the full-length 16s rRNAs of these genomes with that of the related bacterial strains (indicated by the red arrow) and found that the differences between them were very small (Fig. 4e). The most considerable difference in contig_318 was only five bases among more than 1500 bases (Fig. 4e). Interestingly, the *Megasphaera micronuciformis* (contig_242) is a conditional pathogen; therefore we sought to evaluate whether healthy people also carry other conditional pathogens or pathogens. We selected 27 common pathogens and conditional pathogens, including *Escherichia coli*, *Staphylococcus aureus*, *Shigella flexneri*, and *Helicobacter pylori*. Only *Shigella flexneri* was found in the gut of adult and baby, while other pathogenic bacteria were not found (Fig. 4f).

Genomice aspects of the new species of *Candidatus Enterococcus shanghaius*.

Then we analyze the contig_638 with a poor identity. Through the analysis of non-redundant database annotations, we found that it is difficult to find a bacterium matched contig_638, since the most similar *Enterococcus faecalis* had only 8% similarity (Fig. 5a). Then we sought to identify its classification through NCBI blast and the 16s rRNA phylogenetic tree analysis. Excitingly, we discovered that this bacterium belongs to the *Enterococcus* (Fig. 5b). Since it was found in Shanghai, we named it as *Candidatus Enterococcus shanghaius*. Comparing with the most similar bacteria, *Enterococcus saccharolyticus* and *Enterococcus faecalis*, we found that there were considerable differences in the 16s rRNA sequence among them (Fig. 5c). Then we analyzed the function potential of the genome and found that many genes might be correlated to carbohydrate metabolism and lipid metabolism (Fig. 5d). Interestingly, the function potential of many genes were not clear (Fig. 5d). Further analysis of sugar metabolism shows that the most genes were associated with hydrolysis (Fig. 5e). Finally, we analyzed the whole genome of the contig_638: it contains a total of 2.1M bases, and the GC content is relatively

low, about 34.49%, which is also consistent with other species of *Enterococcus* (Fig. 5f). It has 2111 genes, 12 rRNA and 63 tRNA (Fig. 5f).

Discussion

Extensive studies on the human gut microbiota have been reported in the past decades. However, large numbers of species of in the gut microbiota remain unknown. Through improving the extraction method, we successfully acquired the high-quality genomic DNA of the gut microbiota, and obtained approximately 50 Gb TGS data each sample, which is much bigger than previous studies [23, 25, 26]. We have successfully assembled nine bacteria genomes and large numbers of contigs through these methods. Interestingly, more than 50% of the genomes (5 of 9 bacteria genomes) might be new human species or subspecies. Via large-scale metagenomic assembling, Edoardo et al and Alexandre et al uncovered thousands of new bacteria species [9, 27]. These results, together with our work, indicate that it is still of great valuable to investigate the gut microbiota in the future. Furthermore, the methods used in this study help to assemble the genome of unknown bacteria, in consequence, might facilitate to the Human Microbiome Project and the Genomic Encyclopedia of Bacteria and Archaea[28, 29]. Furthermore, bacteria are widely distributed in various environments, such as natural lakes, oceans, soils and some polluted environments. Therefore, it is of great interest to investigate bacteria living in different environments with our methods.

Compared with metagenome, the genes in the microbial genome has traditionally been underestimated. For example, Hila et al found that there are a fairly large number of unknown small proteins in the human microbiome [30]. Intriguingly, our work had discovered more than 10,000 unknown genes with no known domain (Fig. 1d). These findings illuminate that a vast number of unknown genes are need to explore in the human microbiome. Importantly, some special genes of bacteria are very important for human health. For example, an anti-inflammatory protein from *Faecalibacterium prausnitzii* could inhibit the NF- κ B pathway in intestinal epithelial cells [31]. Analyzing with the KEGG, we also found that the genes of the two samples were closely associated with our basic physiological activities (Supplementary Fig. 4a and 4b). It is reported that several proteins from bacteria are potential diagnostic disease indicators since they can pass through intestinal epithelial cell and enter into the plasma [32]. Besides, the exploring of the drug-metabolizing enzymes from gut microbiota might be very useful in drug development and personalized medicine [7, 33]. Our research helps to discover more bacteria proteins linking to human health and diseases.

In the nine assemble genomes, we luckily discovered a new bacterium-Candidatus *Enterococcus shanghaius*. We discovered that it belongs to the *Enterococcus*, a ubiquitous Gram-positive genus with low-GC genomes. It is reported that many members of this genus are pathogenic bacteria or conditional pathogens for their role as primary causative agents of health care-associated infections [34, 35]. Therefore, this bacterium might be a conditional pathogen. Besides this type of bacteria, we have also found other two conditional pathogens: *Megasphaera micronuciformis* and *Shigella flexneri*. There is an

important question: what matters human health: quality, quantity of microbes [36]. The results in these work suggest that quality and quantity of microbes are both important for our health.

Despite the promise that this study holds for gut microbiota, it is important to note its limitations. First, the lengths of bacteria genome we acquired were between 1.5 Mb and 3.5 Mb, which is much smaller than that of *Escherichia coli* str. K-12 (4.64 Mb). We found out that the contig lengths were closely correlated with the coverage (Supplementary Fig. 5). Therefore it is very important to improve the data quantity of each sample to acquire longer contigs. With the continuous development of TGS technology, it is probable to acquire more sequencing data in a single sample. Second, although we have established an effective method, it is still not easy to acquire sufficient high-quality genome DNA, and we have harvested two samples. It is of urgent need to improve these methods and to detect more human feces samples.

Conclusions

In summary, we established an effective extraction method to obtain high-quality gut microbiota genomic DNA and detected the samples with TGS technology. Not only a large number of unknown genes, but also several new subspecies and species were identified with our methods. This work provides a novel and reliable framework for exploring gut microbiota genomes, facilitating the understanding of the mechanisms that underlie the role of the microbiome in health and disease.

Methods

Sample collection and bacteria genome DNA extraction

All methods in this study were approved by the Research Medical Ethics Committee of Shanghai University of Medicine & Health Sciences affiliated Zhoupu Hospital. The feces samples were collected from a 34-year old male and a 10-month old baby and then stored in the Ultra-Low Temperature Freezer (Haier, Qingdao, China). The genome DNA extraction were performed with FastDNA Spin Kit for Feces (MPbio, California, USA). To acquire sufficient high-quality gut microbiota genome DNA, we improved the experimental procedure, as follows: we added 500 mg feces in a 2 ml Lysing Matrix E tube, then mixed the feces with 825 μ l Sodium Phosphate Buffer and 275 μ l PLS solution, then shook the mix and vibrated for 15 seconds. Afterwards we centrifuged the samples at 14,000 g for 5 minutes at room temperature and decanted supernatant. Subsequently, we added 978 μ l Sodium Phosphate Buffer and shook the mix and vibrated the mixture for 15 seconds, then added 122 μ l MT Buffer and shook up and down gently for 5 minutes. Then we placed the samples in the shaker at 4 centigrade for 30 minutes; centrifuged samples at 14,000 g for 5 minutes and then transferred the supernatant to a clean EP tube; added 250 μ l of PPS solution, shook vigorously to mix, and incubated at 4°C for 10 minutes and centrifuged samples at 14,000 g for 2 minutes; transferred supernatant to the Binding Matrix Solution in a 15 ml conical tube and shook gently for 5 minutes. Then we centrifuged samples at 14,000 g for 2 minutes and decanted the supernatant. Afterwards, we washed the binding mixture pellet with 1 ml Wash Buffer #1 and transferred

the binding mixture to a SPIN Filter tube and centrifuge at 14,000g for 1 minute. We emptied the catch tube and added 500 µl of prepared Wash Buffer #2 to the SPIN Filter tube and gently resuspended the pellet. Afterwards, we centrifuged the samples at 14,000 g for two times to extract residual ethanol. Finally, we transfer the SPIN Filter bucket to a clean 1.9 ml Catch Tube and add 100 µl TES to resuspend the genome DNA. The DNA were detected with agarose gel electrophoresis, and the top bands were isolated and purified with a DNA Purification Kit (Finegene, Shanghai, China). The DNA concentration and integrity were assessed by a NanoDrop2000 spectrophotometer (Thermo Fisher Scientific, Waltham, USA).

library construction

For Pacific Biosciences sequencing library preparation and SMRT sequencing, DNA was fragmented by a Covaris g-TUBE device (10 kb) and was concentrate DNA with AMPure PB beads following the manufacturer' protocol (Beckman Coulter Co., USA). The DNA damage and ends were repaired in a LoBind microcentrifuge tube. Blunt ligation reaction was performed by adding 1 µL of blunt adaptor (20 µM) and 1 µL of ligase to the 30 µL of DNA and then incubation was performed at room temperature for 15 min. SMRTbell™ templates were purified with AMPure PB beads and then the concentration was measured by Qubit. Sequencing was performed on a PacBio Sequel instrument by OE Biotech Co., Ltd (Shanghai, China).

Bioinformatic analysis

Metagenome assembly was performed using flye software after getting valid reads. ORF prediction of assembled scaffolds using prodigal was performed and translated into amino acid sequences. The non-redundant gene sets were built for all predicted genes using CD-HIT. The clustering parameters were 95% identity and 90% coverage. The longest gene was selected as representative sequence of each gene set. The gene set representative sequence (amino acid sequence) was annotated with NR, KEGG, COG, SWISSPROT and GO database with an e-value of 1e-5. The taxonomy of the species was obtained as a result of the corresponding taxonomy database of the NR Library.

NCBI prokaryotes genome databases

The prokaryotes genome data were acquired from the databases in the NCBI (<https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/>). There are 266319 prokaryote genomes up to now (Chromosome (3,186), Complete (19,702), Contig(141,127), Scaffold(102,304))(Supplementary Fig. 1). There are 108,506 prokaryote genomes that are associated with human (Chromosome(1,539), Complete(8,179), Contig(57,034), Scaffold(41,754)) (Fig. 1a).

Blast in the NCBI

We acquired the full-length 16s rRNAs from the assembled contigs. Then the 16s rRNAs were blasted in the database of the NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome), then we screened similar

sequences in the order of identity, and downloaded the relative sequences. Then we identified their bacterial species of the assembled contigs and analyzed their evolutionary relationship.

ClustalW and phylogenetic tree analysis

To examine the differences between the full-length 16s rRNAs, we compared the 16s rRNAs and visualized the differences using bioedit software (Borland Software Corporation, Scotts Valley, USA). For phylogenetic tree analysis, we blasted the full-length 16s rRNAs in the NCBI database (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome), and selected the most related full-length 16s rRNAs by identity. With these 16s rRNAs, we analyzed the phylogenetic tree with mega7[37].

Single bacteria analysis

The single bacteria analysis included gene prediction, ncRNA prediction, repeat sequence prediction, non-redundant analysis and the common function potential analyses. We performed the gene prediction with Prokaryotic Dynamic Programming Genefinding Algorithm (prodigal(v2.6.3))[38]; The results included gene number, average gene length(bp) and GC% (gene region). ncRNA predictions were harnessed with three softwares (tRNA(tRNAscan-SE(v1.3.1))[39], rRNA(RNAmmer (v1.2))[40], sRNA(Rfam(v10.0)))[41]. Repeat sequence prediction was analyzed with RepeatMasker(v4.0.7)[42]. The common function potential analyses were included Non-redundant (<https://www.ncbi.nlm.nih.gov>), Swissprot (<http://www.uniprot.org>), KEGG (<http://www.genome.jp/kegg/pathway.html>), Cluster of Orthologous Groups of proteins (<https://www.ncbi.nlm.nih.gov/COG/>), comprehensive antibiotic resistance database (CARD) (<https://card.mcmaster.ca>)[43] and carbohydrate-Active enzymes database (<http://www.cazy.org>) [44].

Statistical analysis

R programming language v. 3.4.3 was used for statistical analysis. Statistical significance between two groups was determined using an unpaired two-tailed Student's t test. Data are presented as mean $\pm$ SD (standard deviation) or mean $\pm$ SEM (standard error of the mean) as indicated in the figure legends. P values were considered statistically significant at $P < 0.05$.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Datasets from metagenomic sequencing are archived at the Sequence Read Archives (SRA) at NCBI.

Competing interests

The authors declare no competing financial interests.

Funding

This work was supported by Special clinical research project of Shanghai Municipal Health Commission (20194Y0458); Pandeng Plan of Shanghai University of Medicine and Health Sciences (B3-0200-20-311007-6); China Postdoctoral Science Foundation funded project (Grant No. 2018M632169).

Author contributions

J.H. and J.S. conceived and designed the research. Y.L. performed sampling. Y.L., J.H. extracted the bacteria genome DNA. J.Z. sequenced DNA and assembled the contigs. J.Z. and J.H., analyzed and interpreted the data. J.H., Y.J., H.P., L.W., D.L., and J.L. performed microbiome analysis. Y.L, J.H., J.S prepared the first manuscript, which was revised by all of the authors.

Acknowledgements

The authors thank the contributors to the databases in the NCBI for their publicly available data. This work could not have been possible without their open sharing of data.

Additional information

Extended data is available for this paper.

Supplementary information is available for this paper.

Correspondence and requests for materials should be addressed to J.S. or J.H.

References

1. Dominguez-Bello MG, Godoy-Vitorino F, Knight R, Blaser MJ: **Role of the microbiome in human development.** *Gut* 2019, **68**(6):1108-1114.
2. Visconti A, Le Roy CI, Rosa F, Rossi N, Martin TC, Mohnhey RP, Li WZ, de Rinaldis E, Bell JT, Venter JC *et al.*: **Interplay between the human gut microbiome and host metabolism.** *Nat Commun* 2019, **10**.
3. D'Amelio P, Sassi F: **Gut Microbiota, Immune System, and Bone.** *Calcified Tissue Int* 2018, **102**(4):415-425.
4. Gopalakrishnan V, Helmink BA, Spencer CN, Reuben A, Wargo JA: **The Influence of the Gut Microbiome on Cancer, Immunity, and Cancer Immunotherapy.** *Cancer Cell* 2018, **33**(4):570-580.

5. Ma C, Han MJ, Heinrich B, Fu Q, Zhang QF, Sandhu M, Agdashian D, Terabe M, Berzofsky JA, Fako V *et al*: **Gut microbiome-mediated bile acid metabolism regulates liver cancer via NKT cells.** *Science* 2018, **360**(6391).
6. Zhu SB, Jiang YF, Xu KL, Cui M, Ye WM, Zhao GM, Jin L, Chen XD: **The progress of gut microbiome research related to brain disorders.** *J Neuroinflamm* 2020, **17**(1).
7. Zimmermann M, Zimmermann-Kogadeeva M, Wegmann R, Goodman AL: **Mapping human microbiome drug metabolism by gut bacteria and their genes.** *Nature* 2019, **570**(7762):462-+.
8. Javdan B, Lopez JG, Chankhamjon P, Lee YCJ, Hull R, Wu QH, Wang XJ, Chatterjee S, Donia MS: **Personalized Mapping of Drug Metabolism by the Human Gut Microbiome.** *Cell* 2020, **181**(7):1661-+.
9. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P *et al*: **Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle.** *Cell* 2019, **176**(3):649-+.
10. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N: **Shotgun metagenomics, from sampling to analysis (vol 35, pg 833, 2017).** *Nat Biotechnol* 2017, **35**(12):1211-1211.
11. Diwan V, Albrechtsen HJ, Smets BF, Dechesne A: **Does universal 16S rRNA gene amplicon sequencing of environmental communities provide an accurate description of nitrifying guilds?** *J Microbiol Meth* 2018, **151**:28-34.
12. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, Leopold SR, Hanson BM, Agresta HO, Gerstein M *et al*: **Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis.** *Nat Commun* 2019, **10**.
13. Poretsky R, Rodriguez-R LM, Luo CW, Tsementzi D, Konstantinidis KT: **Strengths and Limitations of 16S rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics.** *Plos One* 2014, **9**(4).
14. Laudadio I, Fulci V, Palone F, Stronati L, Cucchiara S, Carissimi C: **Quantitative Assessment of Shotgun Metagenomics and 16S rDNA Amplicon Sequencing in the Study of Human Gut Microbiome.** *OMICS* 2018, **22**(4):248-254.
15. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL: **Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing.** *Biochem Biophys Res Commun* 2016, **469**(4):967-977.
16. Somerville V, Lutz S, Schmid M, Frei D, Moser A, Irmeler S, Frey JE, Ahrens CH: **Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system.** *Bmc Microbiol* 2019, **19**.
17. Stewart RD, Auffret MD, Warr A, Wiser AH, Press MO, Langford KW, Liachko I, Snelling TJ, Dewhurst RJ, Walker AW *et al*: **Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen.** *Nat Commun* 2018, **9**.
18. Abdullah WZW, Mackey BM, Karatzas KAG: **High Phenotypic Variability among Representative Strains of Common Salmonella enterica Serovars with Possible Implications for Food Safety.** *J Food Protect* 2018, **81**(1):93-104.

19. Knopp M, Andersson DI: **Predictable Phenotypes of Antibiotic Resistance Mutations.** *Mbio* 2018, **9**(3).
20. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C: **The Third Revolution in Sequencing Technology.** *Trends Genet* 2018, **34**(9):666-681.
21. Ge H, Lin KB, Shen M, Wu SQ, Wang YL, Zhang ZP, Wang ZY, Zhang Y, Huang Z, Zhou C *et al*: **De novo assembly of a chromosome-level reference genome of red-spotted grouper (*Epinephelus akaara*) using nanopore sequencing and Hi-C.** *Mol Ecol Resour* 2019, **19**(6):1461-1469.
22. Holm JB, France MT, Ma B, McComb E, Robinson CK, Mehta A, Tallon LJ, Brotman RM, Ravel J: **Comparative Metagenome-Assembled Genome Analysis of "Candidatus Lachnocurva vaginae", Formerly Known as Bacterial Vaginosis-Associated Bacterium-1 (BVAB1).** *Front Cell Infect Mi* 2020, **10**.
23. Wongsurawat T, Nakagawa M, Atiq O, Coleman HN, Jenjaroenpun P, Allred JI, Trammel A, Puengrang P, Ussery DW, Nookaew I: **An assessment of Oxford Nanopore sequencing for human gut metagenome profiling: A pilot study of head and neck cancer patients.** *J Microbiol Meth* 2019, **166**.
24. Low WY, Tearle R, Bickhart DM, Rosen BD, Kingan SB, Swale T, Thibaud-Nissen F, Murphy TD, Young R, Lefevre L *et al*: **Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity.** *Nat Commun* 2019, **10**.
25. Suzuki Y, Nishijima S, Furuta Y, Yoshimura J, Suda W, Oshima K, Hattori M, Morishita S: **Long-read metagenomic exploration of extrachromosomal mobile genetic elements in the human gut.** *Microbiome* 2019, **7**(1):119.
26. Song WZ, Thomas T, Edwards RJ: **Complete genome sequences of pooled genomic DNA from 10 marine bacteria using PacBio long-read sequencing.** *Mar Genom* 2019, **48**:35-43.
27. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, Lawley TD, Finn RD: **A new genomic blueprint of the human gut microbiota.** *Nature* 2019, **568**(7753):499-504.
28. Mukherjee S, Seshadri R, Varghese NJ, Eloie-Fadrosh EA, Meier-Kolthoff JP, Goker M, Coates RC, Hadjithomas M, Pavlopoulos GA, Paez-Espino D *et al*: **1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life (vol 35, pg 676, 2017).** *Nat Biotechnol* 2018, **36**(4):368-368.
29. Proctor LM, Creasy HH, Fettweis JM, Lloyd-Price J, Mahurkar A, Zhou WY, Buck GA, Snyder MP, Strauss JF, Weinstock GM *et al*: **The Integrative Human Microbiome Project.** *Nature* 2019, **569**(7758):641-648.
30. Sberro H, Fremin BJ, Zlitni S, Edfors F, Greenfield N, Snyder MP, Pavlopoulos GA, Kyrpides NC, Bhatt AS: **Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes.** *Cell* 2019, **178**(5):1245-+.
31. Quevrain E, Maubert MA, Michon C, Chain F, Marquant R, Tailhades J, Miquel S, Carlier L, Bermudez-Humaran LG, Pigneur B *et al*: **Identification of an anti-inflammatory protein from *Faecalibacterium prausnitzii*, a commensal bacterium deficient in Crohn's disease.** *Gut* 2016, **65**(3):415-425.
32. Abasiyanik MF, Wolfe K, Phan HV, Lin J, Laxman B, White SR, Verhoef PA, Mutlu GM, Patel B, Tay S: **Ultrasensitive digital quantification of cytokines and bacteria predicts septic shock outcomes.** *Nat*

Commun 2020, **11**(1).

33. Zimmermann M, Zimmermann-Kogadeeva M, Wegmann R, Goodman AL: **Separating host and microbiome contributions to drug pharmacokinetics and toxicity.** *Science* 2019, **363**(6427):600-+.
34. Ben Braiek O, Smaoui S: **Enterococci: Between Emerging Pathogens and Potential Probiotics.** *Biomed Res Int* 2019, **2019**.
35. Garcia-Solache M, Rice LB: **The Enterococcus: a Model of Adaptability to Its Environment.** *Clin Microbiol Rev* 2019, **32**(2).
36. Cani PD: **Human gut microbiome: hopes, threats and promises.** *Gut* 2018, **67**(9):1716-1725.
37. Kumar S, Stecher G, Tamura K: **MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets.** *Mol Biol Evol* 2016, **33**(7):1870-1874.
38. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: prokaryotic gene recognition and translation initiation site identification.** *Bmc Bioinformatics* 2010, **11**:119.
39. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 1997, **25**(5):955-964.
40. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW: **RNAmmer: consistent and rapid annotation of ribosomal RNA genes.** *Nucleic Acids Res* 2007, **35**(9):3100-3108.
41. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR: **Rfam: an RNA family database.** *Nucleic Acids Res* 2003, **31**(1):439-441.
42. Tarailo-Graovac M, Chen N: **Using RepeatMasker to identify repetitive elements in genomic sequences.** *Curr Protoc Bioinformatics* 2009, **Chapter 4**:Unit 4 10.
43. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, Lago BA, Dave BM, Pereira S, Sharma AN *et al*: **CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database.** *Nucleic Acids Res* 2017, **45**(D1):D566-D573.
44. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B: **The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics.** *Nucleic Acids Research* 2009, **37**:D233-D238.

Figures

Figure 1

The whole characteristics of the two samples. a, Pie chart comprising 4 classes of human bacteria sequence data in the NCBI databases: Chromosome, Complete, Scaffold and Contig. b, Numbers of contigs with different lengths acquired from the two samples, relative to the adult sample (up) and the baby sample (down). c, The whole characteristics of the two samples. The corresponding contig length and CDS number were depicted in the first outer layer and the second outer layer, respectively; the GC

content was depicted in the third outer layer; the contig number and the coverage were depicted in the first inner layer and the second inner layer, respectively. d, Function analyses of all the CDSs of the two samples.

Figure 2

The different aspects of the contigs (0.5 Mb) in the two samples. a, The length and coverage of the contigs in the two samples (0.5 Mb). b, Complete analysis of the contigs in the two samples. c, Wenn

diagram shows the number of contigs (Complete $\geq 99\%$ or circle). d, Pie diagram shows the number of the nine genomes whether they match public genomes. e, phylogenetic tree of the nine assembled genomes.

Figure 3

The genome analysis of the five non-match genomes. a, The genome length and gene number of the five non-match genomes. b, The GC content(total and gene) of the contigs in the two samples of the five

genomes. c, The ncRNA analysis (rRNA, tRNA and sRNA) of the five genomes. d, The repeat sequences of the five genomes. e, The antibiotic potential of the five genomes. f, The blasted result of the full-length 16S rRNAs of the five genomes with the NCBI database.

Figure 4

Phylogeny of four high identity non-match genomes. a, phylogenetic tree of the high identity non-match genomes through full-length 16s rRNAs, showing their accurate taxonomic location. b, ClustalW multiple

alignment of the full-length 16s rRNAs of these genomes with that of the related bacterial strains. c, Conditional pathogens *Shigella flexneri* lived in the gut of adult and baby.

Figure 5

Genome aspects of the new species of *Candidatus Enterococcus shanghaius*. a, Non-redundant database annotations of contig_638. b, Phylogenetic tree of the high contig_638 through full-length 16s rRNAs. c, ClustalW multiple alignment of the full-length 16s rRNAs of *Candidatus Enterococcus shanghaius* with

that of the related bacterial strains. d, Function analyses of the CDSs of the *Candidatus Enterococcus shanghaius*. e, Carbohydrate-Active enzymes analysis of the *Candidatus Enterococcus shanghaius*. f, Circos diagram of the whole genome of the *Candidatus Enterococcus shanghaius* (From the inner to the outer are GC skew, GC content, non-coding RNA (rRNA red, tRNA blue, sRNA green), lagging strand COG annotation, leading strand COG annotation).

Supplementary Files

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

- [SupplementaryTable3.xlsx](#)
- [SupplementaryTable2.xls](#)
- [SupplementaryTable1.xls](#)
- [FigureS4.tif](#)
- [FigureS3.tif](#)
- [FigureS2.tif](#)
- [FigureS1.tif](#)