DATA NOTE

16S rRNA sequencing of samples from universal stool bank donors

Marina Santiago and Scott W. Olesen[*]

Abstract

Objectives: Universal stool banks provide stool to physicians for use in treating recurrent *Clostridioides difficile* infection via fecal microbiota transplantation. Stool donors providing the material are rigorously screened for diseases and disorders with a potential microbiome etiology, and they are likely healthier than the controls in most microbiome datasets. 16S rRNA sequencing was performed on samples from a selection of stool donors at a large stool bank, OpenBiome, to characterize their gut microbial community and to compare samples across different timepoints and sequencing runs.

Data description: 16S rRNA sequencing was performed on 200 samples derived from 170 unique stool donations from 86 unique donors. Samples were sequenced on 11 different sequencing runs. We are making this data available because rigorously screened, likely very healthy stool donors may be useful for characterizing and understanding microbial community differences across different populations and will help shed light into the how the microbiome community promotes health and disease.

Keywords: Human microbiome, 16S rRNA sequencing, Stool donor, Fecal microbiota transplant, Donors, Stool sample, Feces

Objective

Universal stool banks provide rigorously-screened stool to physicians treating patients with recurrent *Clostridioides difficile* infection using fecal microbiota transplantation under US Food and Drug Administration enforcement discretion [1, 2] as well as for research purposes. These stool banks provide centralized donor screening and material preparation, which increases the quality and accessibility of fecal microbiota transplantation as a therapy. Rigorous screening of donors is required to prevent transmission of pathogens or other microbiome-mediated diseases from the donor to the recipient.

The dataset described below is sourced from stool donors from a large, non-profit stool bank (OpenBiome, Cambridge, MA). The bank uses a rigorous screening process [1] that includes (i) an online pre-screen survey where candidates are excluded based on common criteria including body mass index, logistic constraints, and recent antimicrobial use; (ii) an in-person clinical assessment and interview where candidates are excluded for reasons like medication use, infectious disease risk factors, and potentially microbiome-mediated indications such as psychiatric illness; and (iii) a battery of laboratory tests to confirm health status. This results in an average of 3% of candidates accepted as donors [3].

This dataset will complement and extend previously-published sequencing from a subset of the bank's donors [4, 5]. This dataset will be important for understanding how microbiome communities vary across different populations and contribute to health and disease. We are making it available for use by the scientific community for use on its own or as a healthy control comparison population in studies of disease.

*Correspondence: solesen@openbiome.org
OpenBiome, 2067 Massachusetts Ave, Cambridge, MA 02140, USA

Table 1 Overview of data files/datasets

Label	Name of data file/dataset	File types (file extension)	Data repository and identifier (DOI or accession number)
Data set 1	Raw sequencing files	.fastq	*European Nucleotide Archive*; https://identifiers.org/ena.embl:PRJEB41316
Data file 1	exclusion_criteria_comparison	.xlsx	*Zenodo*; https://doi.org/10.5281/zenodo.4282615
Data file 2	donor_health_data	.xlsx	*Zenodo*; https://doi.org/10.5281/zenodo.4282615
Data file 3	otu_table	.tsv	*Zenodo*; https://doi.org/10.5281/zenodo.4282615
Data file 4	metadata	.csv	*Zenodo*; https://doi.org/10.5281/zenodo.4282615
Data file 5	jsd	.pdf	*Zenodo*; https://doi.org/10.5281/zenodo.4282615
Data file 6	pcoa	.pdf	*Zenodo*; https://doi.org/10.5281/zenodo.4282615

Data description

As a result of the extensive screening, this population is healthier compared to other sequenced healthy populations like the Human Microbiome Project or the American Gut Project [6, 7]. The criteria used by these large projects describe different portions of the healthy population but do not screen out as many participants as universal stool banks. A full comparison of these criteria is included in Data File 1. The 86 stool donors that have provided these samples are 71% male and 29% female. Their average age is 27.7, and their average body mass index is 23.1. A full table of available donor health data is in Data File 2.

This dataset consists of 200 samples that have been characterized using 16S rRNA sequencing. These samples come from 170 unique donations from 86 individual donors and were sequenced on 11 sequencing runs. Donations from 48 donors were sequenced more than once. Some of these samples have been included as replicates on the same or on different sequencing runs. 11 donations from 9 donors were sequenced more than once on the same run, and 15 donations from 10 donors were sequenced more than once on different runs.

The samples were sequenced by the University of Michigan DNA Sequencing Core on an Illumina MiSeq. The resulting fastq files (Data set 1) were processed using Qiime 2 (version 2020.8) [8] to create an OTU (operational taxonomic unit) table (Data File 3). Briefly, forward and reverse reads were demultiplexed, joined (using *vsearch join-pairs* with default settings), quality filtered (using *quality-filter q-score* with default parameters), and denoised using Deblur (using *deblur denoise-16S* with a trim length of 253 bp and minimum requirement of 1 read per sequence) [9]. Taxonomies were assigned to unique sequences using a naïve Bayesian classifier [10] trained on the 99% OTUs in the Greengenes database (version 13_8, using *feature-classifier classify-sklearn*) [11–13]. Beta diversity was computed using the Jensen-Shannon divergence (using *diversity beta* with 1 pseudocount). Data File 4 is a metadata file describing these samples. 3 samples did not have any denoised reads and were discarded from downstream analysis.

To confirm that the community composition of each donor remains consistent between sequencing runs, we examined the beta diversity between samples from the same donor but different runs, from the same run but different donors, and from the same donor and run. Samples from the same donor but different runs were more similar to one another relative to samples from different donors sequenced on the same run (medians of 0.608 vs. 0.612, $p = 0.03$, Mann–Whitney U test; Data file 5). Furthermore, donors explained more of the observed beta diversity than sequencing runs (R^2 0.72 vs. 0.02, PERMANOVA by marginal effects; Data file 6), confirming that donor microbiota composition remains stable over time and the biological and technical replicates in this dataset.

Limitations

Although this is a unique and high-quality dataset, no comparator samples from other populations were sequenced along with these samples, so we cannot compare the bank's stool donor population with the healthy community in general or with any specific disease state. Furthermore, only a subset of samples was sequenced multiple times; a more robust dataset would have additional biological and technical replicates.

Authors' contributions
MS and SWO conceived of the manuscript. SWO processed the data and created the initial OTU table. MS further processed the OTU table. MS drafted the manuscript. MS and SWO edited the manuscript. Both authors read and approved the final manuscript.

Availability of data and materials
The data described in this Data Note can be freely and openly accessed on the *European Nucleotide Archive* under accession https://identifiers.org/ena.embl:PRJEB41316 [14] and on the *Zenodo* repository under https://doi.org/10.5281/zenodo.4282615 [15]. See Table 1 for details and links to the data.

Declarations

Ethics approval and consent to participate
Samples were collected from donors enrolled in the OpenBiome donor program. The donor program operates under the New England IRB (reference number 120160907). Written informed consent was obtained from participants. The study was submitted to and approved by OpenBiome's Research Review Panel.

Consent for publication
Written informed consent was obtained from participants.

Competing interests
MS and SWO are employed as consultants by OpenBiome. MS has shares in Finch Therapeutics, Inc.

References
1. Chen J, Zaman A, Ramakrishna B, Olesen SW. Stool banking for fecal microbiota transplantation: methods and operations at a large stool bank. medRxiv. 2020. https://doi.org/10.1101/2020.09.03.20187583.
2. Quality & Safety. OpenBiome. https://www.openbiome.org/safety. Accessed 15 Sep 2020.
3. Kassam Z, Dubois N, Ramakrishna B, Ling K, Qazi T, Smith M, et al. Donor screening for fecal microbiota transplantation. N Engl J Med. 2019;381:2070–2.
4. Poyet M, Groussin M, Gibbons SM, Avila-Pacheco J, Jiang X, Kearney SM, et al. A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. Nat Med. 2019;25:1442–52.
5. Santiago M, Eysenbach L, Allegretti J, Aroniadis O, Brandt LJ, Fischer M, et al. Microbiome predictors of dysbiosis and VRE decolonization in patients with recurrent C. difficile infections in a multi-center retrospective study. AIMS Microbiol. 2019;5:1–18.
6. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS, Giglio MG. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.
7. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, et al. American gut: an open platform for citizen science microbiome research. mSystems. 2018. https://doi.org/10.1128/msystems.00031-18.
8. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37:852–7.
9. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Xu ZZ, et al. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems. 2017. https://doi.org/10.1128/msystems.00191-16.
10. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73:5261–7.
11. 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. 2006;72:5069–72.
12. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome. 2018;6:90.
13. Nicholas Bokulich, Mike Robeson, Ben Kaehler, Matthew Dillon. bokulich-lab/RESCRIPt: 2020.6.1. Zenodo; 2020. https://doi.org/10.5281/zenodo.3945228.
14. Olesen S. openbiome/donors-16 s v1.0. Zenodo. 2020. https://doi.org/10.5281/zenodo.4282615.
15. European Nucleotide Archive. 2020. https://identifiers.org/ena.embl:PRJEB41316.

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