Data in Brief 55 (2024) 110681

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Intestinal microbiota dataset revealed by high-throughput sequencing of 16S rRNA in children with anemia in southern Peru

Jani Pacheco-Aranibar[a,b], Karla Diaz-Rodriguez[a], Rosemary Zapana-Begazo[a], Steven Criollo-Arteaga[a], Jose A. Villanueva-Salas[a], Julio C. Bernabe-Ortiz[a,*]

[a] Universidad Católica de Santa María, Urb. San José s/n, Umacollo, Arequipa 04013, Peru
[b] Instituto de Biotecnología del ADN Uchumayo, Arequipa, Perú

ARTICLE INFO

Article history:
Received 11 October 2023
Revised 19 June 2024
Accepted 20 June 2024
Available online 2 July 2024

Dataset link: Data identification number: PRJNA935779 (Original data) (National Center for Biotechnology Information (NCBI))

Keywords:
16S rRNA
Intestinal microbiota
Children with anemia
Metagenomics
Next-generation sequencing

ABSTRACT

Anemia is the most common hematological disorder affecting humans. In Peru, anemia is a pressing issue that present the most significant concern due to its adverse effects, such as delayed growth and psychomotor development, in addition to a deficiency in cognitive development. Anemia is a significant public health issue in Peru, which has one of the highest prevalence rates in infants in the Latin American and Caribbean (LAC) region, affecting approximately 43.6 % of children under three years nationally as of 2017, with rural areas experiencing a higher prevalence of approximately 53.3 %. In 2019, the prevalence was highest in the Sierra (48.8 %) and Selva regions (44.6 %), whereas the coast had a lower rate of 33.9 % in children under 36 months.

Although the composition of the gut microbiota is relatively well described in children, there is little information on the identification of the microbiota in iron-deficiency anemia. There is evidence that diseases or health conditions can change the microbiota, or vice versa. This study aimed to identify the microbiota in children with anemia who did not recover after iron treatment. In a previous study, we found that the phylum Actinobacteria was predominant in the microbiota of children with anemia. These data will be useful

* Corresponding author.
E-mail address: jbernabe@ucsm.edu.pe (J.C. Bernabe-Ortiz).

https://doi.org/10.1016/j.dib.2024.110681

for understanding the functionality of the most important bacteria found in each group at the genus or species level, especially the metabolic pathways in which they participate and their links with iron metabolism.
Microbial composition data were obtained through next-generation 16S rRNA sequencing (NGS) of stool samples from children with anemia in southern Peru. Numerous studies have underscored the importance of early symbiotic development in infant health and its long-term impact on health. From infancy, modulation of the gut microbiota can promote long-term health. According to the National Institute of Health (NIH), iron-deficiency anemia may cause serious complications, such as fatigue, headaches, restless legs syndrome, heart problems, pregnancy complications, and developmental delays in children. The development of the gut microbiota is regulated by a complex interplay between host and environmental factors. The bidirectional link between the gut microbiota and anemia plays an important role in tracking the gut microbiota and will be useful in understanding the composition of the intestinal microbiota and its implications in anemia, which has now become a public health problem. Our previous study investigated the microbial composition in children with iron-deficiency anemia and revealed the presence of several bacterial groups, including *Proteobacteria, Actinobacteria, Firmicutes*, and *Chloroflexi*.
In addition, these data may be useful for investigating the association between the intestinal microbiota of children with persistent anemia and those who have recovered.

Specifications Table

Subject	***Microbiology: Microbiome***
Specific subject area	Metagenotyping, Molecular biology, Bioinformatics
Type of data	Fastq
How the data were acquired	Illumina MiSeq platform was used to conduct the 16S rRNA gene sequencing
Data format	Fastq Raw sequence reads
Description of data collection	Eighteen fecal samples (approximately 1 g each) were collected from children with anemia in Arequipa Region, southern Peru. Bacterial metagenomic DNA from the representative sample was extracted using the DNeasy PowerSoil kit (Qiagen, Germany), and V3-V4, regions of the 16S rRNA genes were amplified. Libraries of 16S rRNA gene, amplicons were prepared using the Swift amplicon 16S plus ITS panel kit (Swift Biosciences, USA). Finally, Illumina MiSeq platform (250 bp paired ends) was used to sequence the prepared library [4].
Data source location	• Institution: Universidad Catolica Santa Maria • District/Province/Region: Arequipa, Arequipa, Arequipa • Country: Peru
Data accessibility	Repository name: The National Center for Biotechnology Information (NCBI) Data identification number: PRJNA935779 Direct URL to data: http://www.ncbi.nlm.nih.gov/bioproject/935779

(*continued on next page*)

Related research article | Díaz-Rodríguez K, Pacheco-Aranibar J, Manrique-Sam C, Ita-Balta Y, Carpio-Toia AMD, López-Casaperalta P, Chocano-Rosas T, Fernandez-F F, Villanueva-Salas J, Bernabe-Ortiz JC. Intestinal Microbiota in Children with Anemia in Southern Peru through Next-Generation Sequencing Technology. Children (Basel). 2022 Oct 25;9(11):1615.
doi: 10.3390/children9111615. PMID: 36360343; PMCID: PMC9688611.

1. Value of the Data

- The data provided information on the intestinal microbial community of samples collected from children with anemia in southern Peru, its taxonomy, and a comparison between children who recovered from anemia and clinically healthy children [1].
- The data could be used to compare the intestinal microbiome profiles obtained from children with anemia who did not respond to treatment with those obtained from others who recovered from anemia after treatment and clinically healthy children.
- Our previous study showed the microbial composition of children with iron deficiency anemia, revealing the presence of several bacterial groups including *Proteobacteria, Actinobacteria, Firmicutes, and Chloroflexi,* the data could be used for further studies on bacteria involved in treatment resistance and, on the other hand, to identify biomarkers for anemia.

2. Objective

- This study aimed to characterize the microbiota of children with anemia that persisted after treatment with iron. In our previous study, we found that Actinobacteria was predominant in the microbiota of children with anemia. The data shown here will be useful for understanding the functionality of the most important bacteria found in each group at the genus or species level, especially the metabolic pathways in which they participate and their links with iron metabolism [1,2,3,5].

3. Data Description

3.1. OTU identification and taxonomic annotation

To obtain more accurate and reliable results, raw data were merged and filtered to obtain clean data (Phred score $>$ 20). To study the composition of the microbial communities, clustering into Operational Taxonomic Units (OTU) with 97 % sequence identity was performed through clustering. In the OTU construction process, the summary information of the samples was collected, including the total reads in each sample, reads classified as specific taxa, and unclassified reads. A summary is shown in Fig. 1.

3.2. Relative abundance of genus in the group

Samples from the three groups (AA, RA, and C) were analyzed to determine the distribution and abundance of the intestinal microbiota. Fig. 2 shows the 10 main taxa of each group in the lowest taxonomic rank achieved: genus, distributed in a histogram of relative abundance, in which the taxa with the highest relative abundance in each group can be visualized.

The results show a greater relative abundance in the three groups of the genera: *Faecalibacterium, Bacteroides, Prevotella* and *CAG-352*, and other less abundant genera that are grouped under the label “Others;” however, for group C, the *Asteroleplasma* genus that was not observed in the rest of the groups.

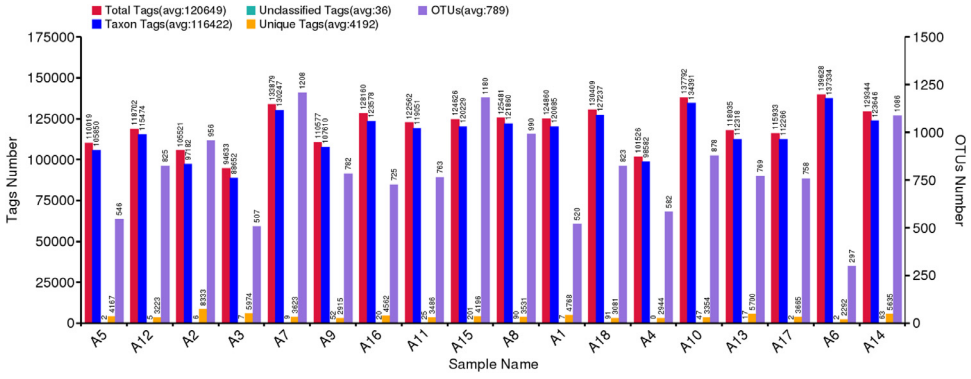

Fig. 1. Summary of tags and OTUs numbers of each sample. The Y axis titled "Tags Number" represents the total number of reads; "Total tags" (red bars) represent the total reads in each sample; "Taxon Tags" (blue bars) indicate the reads classified as specific taxa; "Unclassified Tags" (green bars) indicates the number of unclassified reads; "Unique Tags" (orange bars) indicates the number of unique reads present in each sample. The Y2 axis titled "OTUs Numbers" shows the number of OTUs (Operational Taxonomic Units) shown as "OTU" (purple bars) in the figure to identify the numbers of OTUs in different samples.(For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

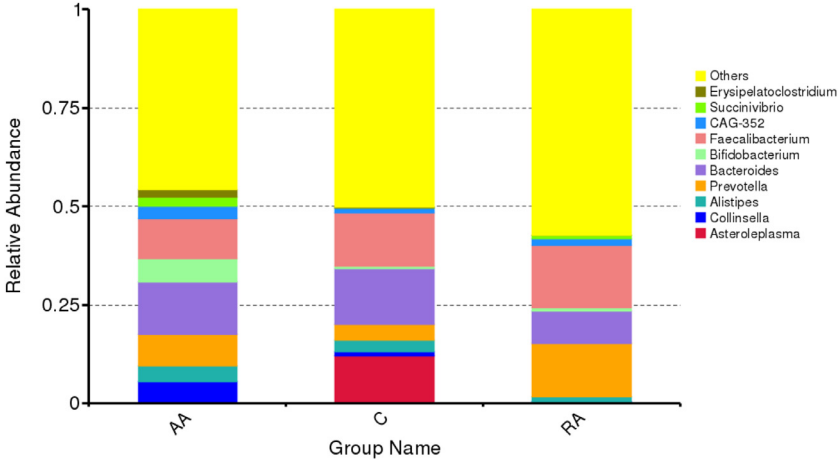

Fig. 2. Distribution histogram of relative abundance of genus. The relative abundance of the top 10 most representative genera in each study group is presented. (AA) anemia, (C) control group, (RA) recovered from anemia.

3.3. Between-group variation analysis of species: Análysis LEfSe

Linear discriminant analysis (LDA) and Effect Size (LEfSe) analyses identified taxa that were significantly more abundant in a group and were used to analyze and detect biomarkers. In the Cladogram illustrated in Fig. 3, taxa (biomarkers) with significant differences are colored according to the corresponding group color. Red nodes, represented by the Rikenellaceae family, contributed significantly to the group anemia (AA); green nodes, represented by the genera *Agathobacter* and species *Eubacterium_rectale*, contributed significantly to the control group (C); blue nodes, represented by the genera *Alloprevotella* and species *Prevotella_melaninogenica* contributed significantly to the group recovered from anemia (RA).

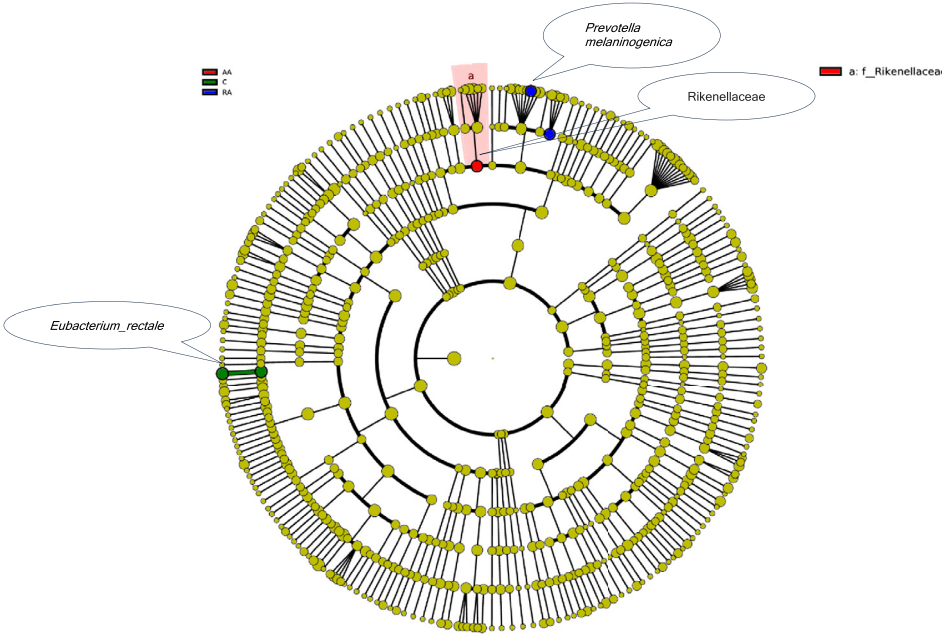

Fig. 3. LEfSe analysis. The cladogram represents biomarkers whose abundance shows significant differences among groups. The selection criteria is that LDA scores are larger than the set threshold (4 set by default). The length of each bin, namely, the LDA score, represents the effect size (the extent to which a biomarker can explain the differentiating phenotypes among groups). Circles radiating from inner side to outer side represents taxonomic rank from phylum to genus. Each circle represents a distinct taxon at a corresponding taxonomic rank. The diameter of each circle represents proportionally the relative abundance of each taxon. Yellow indicates taxa with no significant differences.(For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Methods

4.1. Experimental design

A cross-sectional study was performed by recruiting 18 children aged $<$ 10 years from Southern Peru. The participants were categorized into three distinct groups based on their hemoglobin levels. The first group, designated as the “AA” group, consisted of 10 children diagnosed with anemia, characterized by hemoglobin levels falling below 11 g/dL. The second group, termed the “RA” group, included four children who had successfully recovered from anemia, displaying hemoglobin values surpassing 11 g/dL. The control group, denoted as “C,” comprised four children who had never presented with anemia. Before initiating the study, informed consent was obtained from the parents of all the participating children, and the study was approved by the research ethics committee of the university. Sample collection procedures involved the meticulous collection of fecal samples, which were subsequently stored in hermetically sealed containers (mwe fecal transwab from Medical Wire & Equipment Co.) and kept cold for conservation.

4.2. Isolation of microbial DNA

A preliminary step was performed in a 15 mL Falcon tube, wherein 1 g of fecal sample was resuspended in 3 mL of 0.9 % NaCl and vigorously agitated by vortexing (Eurolab, Madrid, Spain)

for 30 s. Subsequently, the mixture was centrifuged at 3000 rpm for 2 min, and the resulting supernatant was carefully transferred to a microcentrifuge tube and centrifugated at 14,000 rpm for 3 min to obtain the bacterial pellet. Subsequently, a second wash was conducted using 1 mL phosphate-buffered saline (PBS) and centrifuged at 3000 rpm for 2 min. The supernatant obtained in this step was centrifuged again at 14,000 rpm for 1 min, and the resulting pellet was retained in a microcentrifuge tube. Subsequently, 100 µg of glass beads were added to the tube containing the pellet for vortex agitation for 5 min. Next, 500 µL of phenol:chloroform:isoamyl alcohol; 25:24:1 (> pH 7.0) was added for the extraction of bacterial DNA. The aqueous fraction containing nucleic acids was then precipitated with 750 µL of isopropyl alcohol (35 %). The concentration and purity of the DNA were assessed using a NanoDrop spectrophotometer (NanoDrop One, Thermo Scientific, Walthman, MA USA), and its integrity was evaluated using 1.5 % agarose gel electrophoresis.

4.3. Illumina sequencing

To sequence the V3-V4 variable region of the 16S rRNA gene in the bacteria present in the fecal sample, sequencing libraries were generated using the NEBNext Ultra Library Prep Kit for Illumina following the manufacturer's protocol and including index codes. Subsequently, for library evaluation, the Agilent Bioanalyzer 2100 system, and the Qubit 2.0 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) were utilized. Finally, sequencing was performed on an Illumina platform, yielding 250 bp paired-end sequences.

4.4. Bioinformatics

The FLASH program v. 1.2.7 was employed to merge the paired-end sequences. Next, the quality of the raw tags was assessed by filtering them under specific conditions following the QIIME quality control process, resulting in high-quality clean tags. Subsequently, the chimeric sequences were detected and removed using the UCHIME algorithm. Further sequence analysis was conducted using Uparse version 11.0.667. Phylogenetic analysis was performed between each representative sequence and different operational taxonomic units (OTUs) using the GreenGene database. GreenGene was used to study the differences between the dominant species in different samples. The sequences were aligned using the MUSCLE program. Finally, alpha and beta diversity were calculated in weighted and unweighted UniFrac using QIIME software (QIIME 2 q2studio-2022.8.0).

4.5. Statistical analysis and data handling

Principal component analysis (PCA) and hierarchical clustering were conducted using the FactoMineR package (Vienna, Austria) in R software version 2.15.3, in combination with ggplot2. Principal coordinate analyses (PCoA) using weighted and unweighted UniFrac distances were performed using the integrated WGCNA, stat, and ggplot2 packages in R. Effect size assessment in linear discriminant analysis was performed using LEfSe software, which was found as a module within MicrobiomeAnalyst. The *p*-values were calculated using the permutation test method, whereas q-values were obtained using the Benjamini–Hochberg false discovery rate method. Anosim, MRPP, and Adonis analyses were conducted using the vegan package in R, and molecular variance analysis (AMOVA) was performed using the open-source Mothur software v. 1.48.0, using the AMOVA function. Additionally, t-tests were performed using R.

Ethics Statement

Informed consent was obtained from the parents of all the participating children, and the study was approved by the Research Ethics Committee of the Universidad Católica de Santa

María (DICTAMEN-196-2020, December 21, 2020) in Arequipa, Peru. The study also complies with the ethical principles of the Declaration of Helsinki.

Data availability

Intestinal microbiome dataset from children with anemia from south Perú. https://www.ncbi.nlm.nih.gov/bioproject/935779.

CRediT Author Statement

Jani Pacheco-Aranibar: Conceptualization, Methodology, Resources, Investigation, Formal analysis, Software, Data curation, Validation, Visualization, Writing – original draft, Writing – review & editing; **Karla Diaz-Rodriguez:** Conceptualization, Methodology, Resources, Investigation, Formal analysis, Software, Data curation, Validation, Visualization, Writing – original draft, Writing – review & editing; **Rosemary Zapana-Begazo:** Investigation, Formal analysis; **Steven Criollo-Arteaga:** Investigation, Formal analysis, Writing – review & editing, Resources; **Jose A. Villanueva-Salas:** Investigation, Formal analysis, Writing – review & editing, Resources; **Julio C. Bernabe-Ortiz:** Conceptualization, Methodology, Resources, Investigation, Formal analysis, Software, Data curation, Validation, Visualization, Writing – original draft, Writing – review & editing.

Acknowledgments

The authors thank the Universidad Católica de Santa Maria for their support.

The study was conducted at the Laboratory of Research the Postgraduate school and financed by vice-rector for research of Universidad Católica de Santa Maria, Arequipa, Perú.

Funding

This study was funded by internal funds for research at the Universidad Católica Santa María, RESOLUTION 26617-R2019.

Declaration of Competing Interest

The authors declare that they have no competing financial interests or personal relationships that may have influenced the work reported in this study.

References

[1] K. Díaz-Rodríguez, J. Pacheco-Aranibar, C. Manrique-Sam, Y. Ita-Balta, C.-T. Agueda, P. López-Casaperalta, T. Chocano-Rosas, F. Fernandez-F, J. Villanueva- Salas, J.C. Bernabe-Ortiz, Intestinal microbiota in children with anemia in southern Peru through next-generation sequencing technology, Children 9 (11) (2022) 1615, doi:10.3390/children9111615.

[2] INEI;, & Ministerio de Economía y Finanzas. (2020). Perú: indicadores de resultados de los programas presupuestales 2014–2019. https://doi.org/10.1017/CBO9781107415324.004

[3] F. Asnicar, G. Weingart, T.L. Tickle, C. Huttenhower, N. Segata, Compact graphical representation of phylogenetic data and metadata with GraPhlAn, PeerJ 3 (2015) e1029.

[4] E. Avershina, T. Frisli, K. Rudi, De novo semi-alignment of 16S rRNA gene sequences for deep phylogenetic characterization of next generation sequencing data, Microb. Environ. 28 (2) (2013) 211–216.

[5] S. Huey, L. Jiang, M. Fedarko, D. McDonald, et al., Nutrition and the gut microbiota in 10- to 18-month-old children living in urban slums of Mumbai, India, MSphere 5 (5) (2020) 1–18.