UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Posgrados

Unraveling Strain-Level Variation in the Gut Microbiome using Metagenome-Assembled Genomes

Tesis en torno a una hipótesis o problema de investigación y su contrastación

Galo David Flores Cuadrado

Paúl Cárdenas M.D., Ph.D.
Director de Trabajo de Titulación

Trabajo de titulación de posgrado presentado como requisito
para la obtención del título de Magister en Microbiología

Quito, septiembre del 2023

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

COLEGIO DE POSGRADOS

HOJA DE APROBACIÓN DE TRABAJO DE TITULACIÓN

Unraveling Strain-Level Variation in the Gut Microbiome using Metagenome-Assembled Genomes

Galo David Flores Cuadrado

Nombre del Director del Programa:	Patricio Rojas
Título académico:	M.D., Ph.D.
Director del programa de:	Maestría de Microbiología

Nombre del Decano del colegio Académico:	Carlos Valle
Título académico:	Ph.D.
Decano del Colegio:	COCIBA

Nombre del Decano del Colegio de Posgrados:	Hugo Burgos
Título académico:	Ph.D.

Quito, septiembre 2023

© DERECHOS DE AUTOR

Nombre del estudiante:	Galo David Flores Cuadrado
Código de estudiante:	00326722
C.I.:	1722764832
Lugar y fecha:	Quito, 4 de septiembre de 2023.

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following graduation project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

DEDICATORIA

A mi familia que me ha dado las herramientas para llegar hasta aquí y seguir mejorando.

AGRADECIMIENTOS

A mis tutores, al centro de bioinformática, al Instituto de Microbiología y a la USFQ.

RESUMEN

El análisis metagenómico proporciona información valiosa sobre la composición y la dinámica de las comunidades microbianas. En este estudio, utilizamos un enfoque de genomas ensamblados a partir de metagenomas (MAGs por sus siglas en inglés) para investigar la variación en las cepas presentes en el microbioma intestinal infantil en diferentes momentos. Recopilamos 122 muestras de heces de 39 niños de 1 a 7 años cada 5 a 12 meses. Los MAGs se obtuvieron mediante el ensamblaje *de novo* utilizando MEGAHIT y el binning con Metawrap. Se identificaron taxonómicamente un total de 1126 MAGs, que representaban diversos filos, clases, órdenes, familias y géneros. El análisis a nivel de cepa se centró en *Prevotella copri* y *Bacteroides fragilis*, revelando variaciones en SNPs e inserciones/deleciones (*indels*) entre muestras del mismo individuo en diferentes momentos. Nuestro análisis muestra que cada individuo tiene por lo menos 1 a 3 cepas de *P. copri* y 1 a 4 cepas de *B. fragilis* en cualquier momento, de igual manera encontramos evidencia de intercambio de cepas en las dos especies. Estos hallazgos confirman un alto dinamismo en poblaciones bacterianas de dos de los taxones más grandes en el microbioma humano.

Palabras clave: MAGs, microbioma, intestino, metagenomas, Prevotella, Bacteroides.

ABSTRACT

Metagenomic analysis provides valuable insights into the composition and dynamics of microbial communities. In this study, we employed a metagenome-assembled genomes (MAGs) approach to investigate the variation in strains present in the infant gut microbiome over different time points. We collected 122 faecal samples from 39 children aged 1 to 7 every 5 to 12 months. MAGs were obtained through de novo assembly using MEGAHIT and binning with Metawrap. A total of 1126 MAGs were taxonomically identified, representing various phyla, classes, orders, families, and genera. We carried out a strain-level analysis focusing on *Prevotella copri* and *Bacteroides fragilis*. Our results revealed SNPs and insertions/deletions (indels) within the sample from the same individual and among samples collected at different time points. Our analysis showed that each individual had at least 1-3 strains of *P. copri* and 1-4 strains of *B. fragilis* at any given time. We also found evidence of strain turnover in both species. We found evidence that confirms highly dynamic bacterial populations in the two of the major taxa in the human microbiome.

Key words: MAGs, microbiome, gut, metagenome, Prevotella, Bacteroides.

Tabla de contenido

GENERAL INTRODUCTION

The microbiome refers to the repertoire of microorganisms, known as microbiota, and their genomes, such as bacteria, fungi, viruses, and other microbes, that inhabit an organism or an environment [1]. It is crucial in maintaining various ecosystems, including the human microbiome. The human microbiome is a complex and highly diverse community of microorganisms that reside in human mucosal and skin surfaces like the gut, skin, mouth, etc. [2]. These microorganisms have a symbiotic relationship with their host and may impact various aspects of human physiology, immunity, metabolism, and even mental health [3]. Current knowledge of gut microbiome composition has grown significantly in recent years. Research has revealed that the gut microbiome is a compound community consisting of a vast array of microorganisms, including primarily bacteria, that can make up to 90% of the gut microbiome, viruses, fungi, archaea, and protists [1]. The gut microbiome's composition can vary a lot from individual to individual in response to changes in diet, lifestyle, environment, and genetics [4]. The relationship between core and accessory genomes and how they affect the microbiome has been studied [5]. A bacterial species' core is made up of genes shared by all the members of a given species; it is vertically inherited, contains the housekeeping genes, and, bioinformatically, is used to determine the microbial species [5]. On the other hand, the accessory genome consists of genes present in some strains but not all, providing adaptive potential to the bacterial cell [6]. These accessory genes can confer antibiotic resistance, stress tolerance, novel or different metabolic capabilities, and niche-specific adaptations, affecting the functional diversity of the microbiome and its capacity to adapt to environmental changes [5]. Most of the genes in the accessory genome are thought to be acquired by horizontal gene transfers (HGT) [7].

The HGT is a process by which genetic material can be transferred between genetically different organisms, enabling the acquisition of new traits. This phenomenon plays a significant role in

shaping bacterial evolution and can contribute to developing pathogenic strains, providing virulence factors and adaptive features modifying its behavior and capabilities [7]. This is different from genetic recombination, a process where two DNA molecules or segments exchange genetic material resulting in the rearrangement of the segment, since this process usually takes place with DNA from the same individual or members of the same species, either by homologous regions or by specific site recognition [8].

Microbiome composition

Studies have identified several predominant bacterial phyla in the gut microbiome, including Firmicutes with a presence from 50% up to 80%, Bacteroidetes from 20% to 50%, Actinobacteria, and Proteobacteria ranging from 1 to 10% each [9]. Within these phyla, there is substantial diversity at the genus and species levels. For example, *Bacteroides* spp, *Faecalibacterium* spp, and *Ruminococcus* spp are common bacterial genera in the human gut [10]. These bacteria may be crucial for human metabolism because they synthesize vitamins, metabolize fiber, and produce short-chain fatty acids that can be used by the animal host [11]. The composition of the gut microbiome at different taxa levels is dynamic and can alter over time or in response to age [12], geography [12][13], diet [13], and other unidentified causes [12]. In diseases including irritable bowel syndrome (IBS), obesity, and inflammatory bowel disease (IBD), for example, changes in the gut microbiome have been noted [14]. These diseases have been linked to imbalances in the relative abundance of some bacterial groups. For instance, it has been observed that people with IBD had lower amounts of helpful bacteria like *Bifidobacterium* and *Lactobacillus* and higher levels of potentially hazardous bacteria like *Escherichia coli* pathotypes [15]. We must remember that the correlation between microbiome change and disease doesn't imply causation. Additionally, the assignment of functions to

bacterial species ignores that the accessory genome can cause severe changes in the bacterium-animal host interaction; for instance, a commensal bacterium can become pathogenic [16].

Studies have shown that despite day-to-day fluctuations, and while having a degree of individuality, the overall composition and diversity of the gut microbiome tend to remain relatively stable, at the species level, within an individual over months or even years [17][18]. The stability of the gut microbiome referring to species and strains is a subject of ongoing research. Phyla, such as Firmicutes and Bacteroidetes, tend to exhibit relatively high stability in the gut microbiome, with certain core members consistently present [9]. Factors like diet, lifestyle, and environmental influences can impact species composition within an individual's gut microbiome [9][14]. Strains may undergo clonal expansion or decline based on selective pressures or competitive interactions, while certain strains with specific traits can exhibit higher stability, the overall strain-level composition tends to be more dynamic than phyla and species [9].

Some researchers have also indicated that certain core microbial species (not to be confused with core genome) and strains persistently colonize the gut over extended periods. These core microbiome members are thought to play essential roles in homeostasis and contribute to the stability of the gut microbiome [19]. However, it's worth noting that the stability of specific species or strains within the gut microbiome can be influenced by various factors, such as dietary changes, medication use, and host genetics [20].

Evidence also points to the possibility that specific circumstances or treatments may cause instability in the species present in the gut microbiome. For instance, research has demonstrated

that antibiotics can significantly modify the gut microbiome, resulting in decreased species diversity and changes to the microbial makeup [21]. Similar variables can affect the stability of the gut microbiome, perhaps causing dysbiosis and related health effects. These factors include food, infections, and disease states [22].

Dietary elements also influence how the gut microbiota functions. For example, short-chain fatty acids (SCFAs) are produced by particular bacteria in the gut when dietary fiber is used as a fuel source. SCFAs aid in regulating metabolism, lower inflammation, and support the soundness of the intestinal barrier [23]. On the other hand, meals high in fat and sugar have been linked to changes in the gut microbial ecology and a reduction in SCFAs synthesis that bacteria can produce [24].

Moreover, diet changes can rapidly affect the gut microbiome composition regarding the species present. Some studies have shown that shifting from a vegan diet to a Western-style diet can lead to noticeable changes in microbial composition within days [20][25]. These changes highlight how dynamic the gut microbiome can be. It was discovered that various diets significantly affect the gut microbiome's structure and operation. According to those studies, food habits can affect the variety and number of microbial species in the gut. An increase in genera like *Alistipes* spp. and *Bacteroides* spp. have been linked to a Western diet, for instance, which is known for its high intake of processed foods, sugar, and saturated fats [20][13]. On the other hand, a more varied gut microbiome has been associated with plant-based diets high in fiber, fruits, and vegetables [22]. All these factors can influence the gut microbiome at the species or strain level changing the behavior of the bacteria from commensal to pathogenic or vice versa.

One significant challenge in understanding the microbiome is its immense diversity and complexity. New insights in DNA sequencing technologies allow researchers to explore better and characterize the microbiome. Classical microbiology, which involves isolating and culturing individual microbial species in the laboratory [26], differs from metagenomics, which focuses on studying the genetic material directly extracted from environmental samples [27]. Classical microbiology often targets specific microorganisms of interest or known pathogens, studying their phenotypic characteristics and specific traits [26]. In contrast, metagenomics takes a more comprehensive approach by analyzing the entire microbial community present in a sample, providing a better understanding of the functional potential of microbial communities through the analysis of collective genetic content [26][27]. Metagenomics overcomes the bias of classical microbiology toward culturable microorganisms by directly examining genetic material from environmental samples [28][29]. It enables high-throughput analysis of large datasets to study complex microbial communities [30].

The composition of the gut microbiome has been thoroughly investigated thanks to developments in DNA sequencing methods, notably metagenomic sequencing. Researchers can pinpoint individual bacterial species or strains that are present in the gut by examining the genetic makeup of the microbial community [19]. Furthermore, functional profiles of the gut microbiome have been uncovered by metagenomic research, offering insight into the genes and metabolic pathways involved in a range of microbial activities and interactions with the host [10].

REFERENCES

1. Sender R, Fuchs S, Milo R. Revised estimates for the number of human and bacteria cells in the body. PLoS biology. 2016 Aug 19;14(8):e1002533.
2. Marchesi JR, Adams DH, Fava F, Hermes GD, Hirschfield GM, Hold G, Quraishi MN, Kinross J, Smidt H, Tuohy KM, Thomas LV. The gut microbiota and host health: a new clinical frontier. Gut. 2016 Feb 1;65(2):330-9.
3. Kamada N, Chen GY, Inohara N, Núñez G. Control of pathogens and pathobionts by the gut microbiota. Nature immunology. 2013 Jul;14(7):685-90.
4. Rajilić-Stojanović M, De Vos WM. The first 1000 cultured species of the human gastrointestinal microbiota. FEMS microbiology reviews. 2014 Sep 1;38(5):996-1047.
5. Konstantinidis KT, Ramette A, Tiedje JM. The bacterial species definition in the genomic era. Philosophical Transactions of the Royal Society B: Biological Sciences. 2006 Nov 29;361(1475):1929-40.
6. Whelan FJ, Hall RJ, McInerney JO. Evidence for Selection in the Abundant Accessory Gene Content of a Prokaryote Pangenome. Mol Biol Evol. 2021 Aug 23;38(9):3697-3708. doi: 10.1093/molbev/msab139.
7. Juhas M, Van Der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW. Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS microbiology reviews. 2009 Mar 1;33(2):376-93.
8. Krogh BO, Symington LS. Recombination proteins in yeast. Annu. Rev. Genet.. 2004 Dec 15;38:233-71.
9. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. Nature. 2012 Sep 13;489(7415):220-30.
10. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012 Jun 13;486(7402):207-214.
11. Flint HJ, Scott KP, Duncan SH, Louis P, Forano E. Microbial degradation of complex carbohydrates in the gut. Gut microbes. 2012 Jul 14;3(4):289-306.
12. Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. Genome medicine. 2016 Dec;8(1):1-1.
13. De Filippis F, Pellegrini N, Vannini L, Jeffery IB, La Storia A, Laghi L, Serrazanetti DI, Di Cagno R, Ferrocino I, Lazzi C, Turroni S. High-level adherence to a Mediterranean diet beneficially impacts the gut microbiota and associated metabolome. Gut. 2016 Nov 1;65(11):1812-21.
14. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR. A human gut microbial gene catalogue established by metagenomic sequencing. nature. 2010 Mar 4;464(7285):59-65.
15. Frank DN, St. Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proceedings of the national academy of sciences. 2007 Aug 21;104(34):13780-5.
16. Clermont O, Christenson JK, Denamur E, Gordon DM. The C lermont E scherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. Environmental microbiology reports. 2013 Feb;5(1):58-65.
17. Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, Clemente JC, Knight R, Heath AC, Leibel RL, Rosenbaum M. The long-term stability of the human gut microbiota. Science. 2013 Jul 5;341(6141):1237439.
18. Claesson MJ, Cusack S, O'Sullivan O, Greene-Diniz R, de Weerd H, Flannery E, Marchesi JR, Falush D, Dinan T, Fitzgerald G, Stanton C. Composition, variability, and temporal stability of the intestinal microbiota of the elderly. Proceedings of the National Academy of Sciences. 2011 Mar 15;108(supplement_1):4586-91.

19. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, Peng Y. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012 Oct 4;490(7418):55-60.
20. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 2014 Jan 23;505(7484):559-63.
21. Dethlefsen L, Huse S, Sogin ML, Relman DA. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. PLoS biology. 2008 Nov;6(11):e280.
22. Sonnenburg JL, Bäckhed F. Diet–microbiota interactions as moderators of human metabolism. Nature. 2016 Jul 7;535(7610):56-64.
23. Den Besten G, Van Eunen K, Groen AK, Venema K, Reijngoud DJ, Bakker BM. The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. Journal of lipid research. 2013 Sep 1;54(9):2325-40.
24. Murphy EF, Cotter PD, Healy S, Marques TM, O'sullivan O, Fouhy F, Clarke SF, O'toole PW, Quigley EM, Stanton C, Ross PR. Composition and energy harvesting capacity of the gut microbiota: relationship to diet, obesity and time in mouse models. Gut. 2010 Dec 1;59(12):1635-42.
25. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, Sinha R. Linking long-term dietary patterns with gut microbial enterotypes. Science. 2011 Oct 7;334(6052):105-8.
26. Madigan MT, Martinko JM, Bender KS, Buckley DH, Stahl DA. Brock Biology of Microorganisms, 14th Edn London.
27. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. Microbiology and molecular biology reviews. 2004 Dec;68(4):669-85.
28. Harwani D. The great plate count anomaly and the unculturable bacteria. Microbiology. 2013 Sep;2(9):350-1.
29. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & biology. 1998 Oct 1;5(10):R245-9.
30. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 2004 Mar 4;428(6978):37-43.

ARTICLE

Unraveling Strain-Level Variation in the Gut Microbiome using Metagenome-Assembled Genomes

Galo Flores C.[1], Sara G. Cifuentes[1] , Gabriel Trueba[1] and Paúl A. Cárdenas[1]

[1] Instituto de Microbiología, Universidad San Francisco de Quito USFQ, Quito, Ecuador.

Keywords: MAGs, microbiome, gut, metagenome, Prevotella, Bacteroides.

Abstract's word count: 173 words.

Text's word count: 2283.

ABSTRACT

Metagenomic analysis provides valuable insights into the composition and dynamics of microbial communities. In this study, we employed a metagenome-assembled genomes (MAGs) approach to investigate the variation in strains present in the gut microbiome over different time points. We collected 125 faecal samples from 39 children aged 1 to 7 every 5 to 12 months. MAGs were obtained through de novo assembly using MEGAHIT and binning with Metawrap. A total of 1126 MAGs were taxonomically identified, representing various phyla, classes, orders, families, and genera. We carried out a strain-level analysis focusing on *Prevotella copri* and *Bacteroides fragilis*. Our results revealed SNPs and insertions/deletions (indels) within the sample from the same individual and among samples collected at different time points. Our analysis showed that each individual had at least 1-3 strains of *P. copri* and 1-4 strains of *B. fragilis* at any given time. We also found evidence of strain turnover in both species. We found evidence that confirms highly dynamic bacterial populations in the two of the major taxa in the human microbiome.

INTRODUCTION

Our understanding of bacterial communities and how they respond to environmental changes has been completely transformed by culture-independent genetic and genomic data, including high-throughput sequencing and metagenomics [1]. In oceanic microbial communities, metagenomic analysis has revealed distinct changes in microbial taxa and functional genes in response to temperature gradients [2], and the identification of specific genes linked to pollutant degradation in soil microbial communities has shed light on their responses to environmental contaminants [3]. The intestinal microbiome's complexity, diversity, and dynamics are known to respond to intestinal perturbations [2][3]. Even though many studies describe the diversity

and stability of bacterial phyla, genera, and species, there is very little information about the stability of the strain diversity.

The study of strain-level dynamics in a microbiome is crucial to have a better understanding of complex microbial communities [4]. For instance, some horizontally transferred genes (accessory genome) can change the phenotype of any strain: a commensal strain into a pathogenic one [5]. The accessory genome can contain genes involved in metabolic adaptation, virulence factors, and antibiotic resistance [6][7].

On the other hand, it has been suggested that microbiome influences various physiological processes (such as glycemic response) and different strains could have different impact for the host physiology [8]. Strain-level analyses can give insights on these interactions whereas core genome (16S RNA gene metagenomics) can only give taxonomic information from which we can infer a limited number of metabolic functions. The core genome codes for housekeeping processes, DNA replication, translation, central metabolism [9], cell division, cell wall synthesis, and responses to stress [10]. In this study, we use the MAGs approach to see if there is a variation in the strains present in the gut microbiome at different time points.

MATERIALS AND METHODS

The faecal samples used in this study were collected from August 2018 to September 2021 according to the methodology discussed by Cifuentes *et al*. [11]. A total of 39 children from ages 1 to 7 participated in the study, providing 1 to 5 samples corresponding to at least one of the 5 sampling cycles performed in 5 months to a year interval. The samples were sequenced using the Illumina Nova-Seq platform. For quality control and removing the host DNA from the raw reads the tools FastQC [12], Trimmomatic [13], BWAtools [14] and Samtools [15] were used. To obtain the MAGs from the samples an adaptation of the MAG Snakemake

pipeline was implemented following the protocol used by Saheb, Almeida, Segre & Finn [16]. The reads were assembled with MEGAHIT [17], and the binning of the resulting assemblies was conducted using the corresponding module of Metawrap [18]. We also used the bin refinement module of the software. The steps discussed above are the basis for obtaining MAGs. The steps to assess and guarantee the quality of the genomes were performed as indicated in the MAG Snakemake pipeline with 20% of the total samples analyzed, dereplication of MAGs and bottlenecks evaluation were also carried out with 20%. The Genome Taxonomy Database and its toolkit were used to classify MAGs according to bacterial and archaeal taxonomy.

Samples corresponding to the same individual in different collection time points were grouped. We focused on individuals presenting *Prevotella copri* and *Bacteroides fragilis*, the two most common species of their respective genera, resulting in 20 and 12 respectively. To obtain the consensus sequences of the bins corresponding to *Prevotella copri,* we used minimap2 [19], Samtools [15] and Ivar [20], using as a reference of *Prevotella copri* (NCBI reference NZ_GG703857.1) and *Prevotella stercorea* (RefSeq GCF_003473415.1). The same steps were followed for member of *Bacteroides fragilis* using the reference NZ_CP069563.1 from RefSeq. Mafft software [21] was used to align the sequences from both groups. Insertions-deletions, SNPs and overall variants were quantified among each individual using Snippy [22]. To identify different strains among samples, PanPhlan3 [23] was used.

RESULTS

Metagenome Assembled Genomes

From 122 metagenomic samples we obtained 1126 MAGs identified to species level, or an individual OTU code, and 69 genomes not identified with the database used. Of the ones that

were identified the three most representative at phylum level were Firmicutes A with 44.3% representation, Bacteroidota with 26.7% and Actinobacteria with 10%, at Class level the biggest group were Clostridia with 44.4%, Bacteroidia with 29.7% and Actinomycetia with 7.4%. In the Order level, Bacteroidales had 29.7%, Oscillospirales 24.4% and Lachnospirales with 18.8%; Bacteroidaceae with 19.7%, Lachnospiraceae with 18.8% and Ruminococcaceae at 12.2% were the biggest groups at Family level. *Prevotella*, *Faecalibacterium* and *Bifidobacaterium* were the most predominant groups at genus level with 9.5%, 7.9% and 7.4% respectively. At every taxa level there were groups with little representation individually, those groups are joined under the label others in Figure 1.

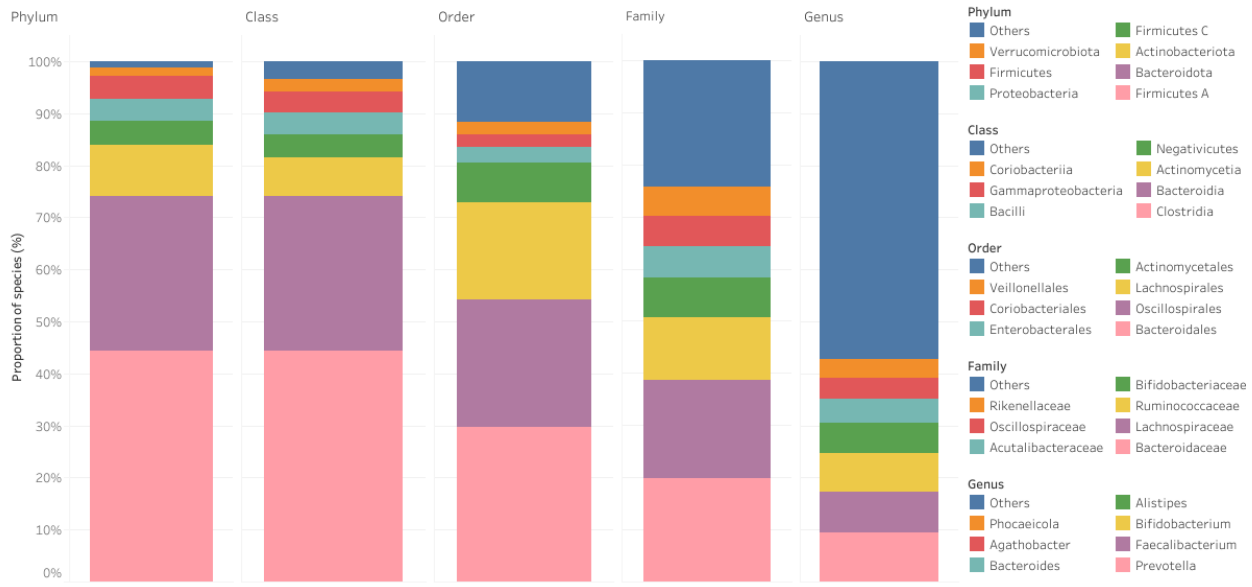

Figure 1 Proportion of species at Phylum, Class, Order, Family and Genus levels of the 1126 MAGs obtained, taxonomically identified using GTDB-tk.

Strain-level analysis of *Prevotella* and *Bacteroides*

From the genomes resulting from the pipeline used, we focused on the strains belonging to the species *Prevotella copri* and *Bacteroides fragilis*, from the same person. To assess the number of variations between time points, we compared the genome variants (putative strains) from

samples collected at different time points. We obtained values for SNPs and indels between the strains samples as well as the total number of variations using Snippy. When the software found no indels there are blank spaces on the tables (Table 1 and Table 2). Strains differ from point mutations by the total number of variations present between genomes.

We also wanted to see the composition of these groups at a strain level comparing all the samples, from the 20 individuals with the species *Prevotella copri* and the 12 with *Bacteroides fragilis* from the previous step, only 16 and 10 respectively passed the coverage threshold to be analyzed in Panphlan. In addition to comparing all the genomes obtained that corresponded to one of the groups we ordered the samples by the individuals they belong to. The presence/absence matrix obtained was visualized as a heatmap where the rows are the genes annotated by PanPhlan3. Both in *Prevotella copri* and *Bacteroides fragilis*, there is variation among all the genomes recovered as well as among each individual.

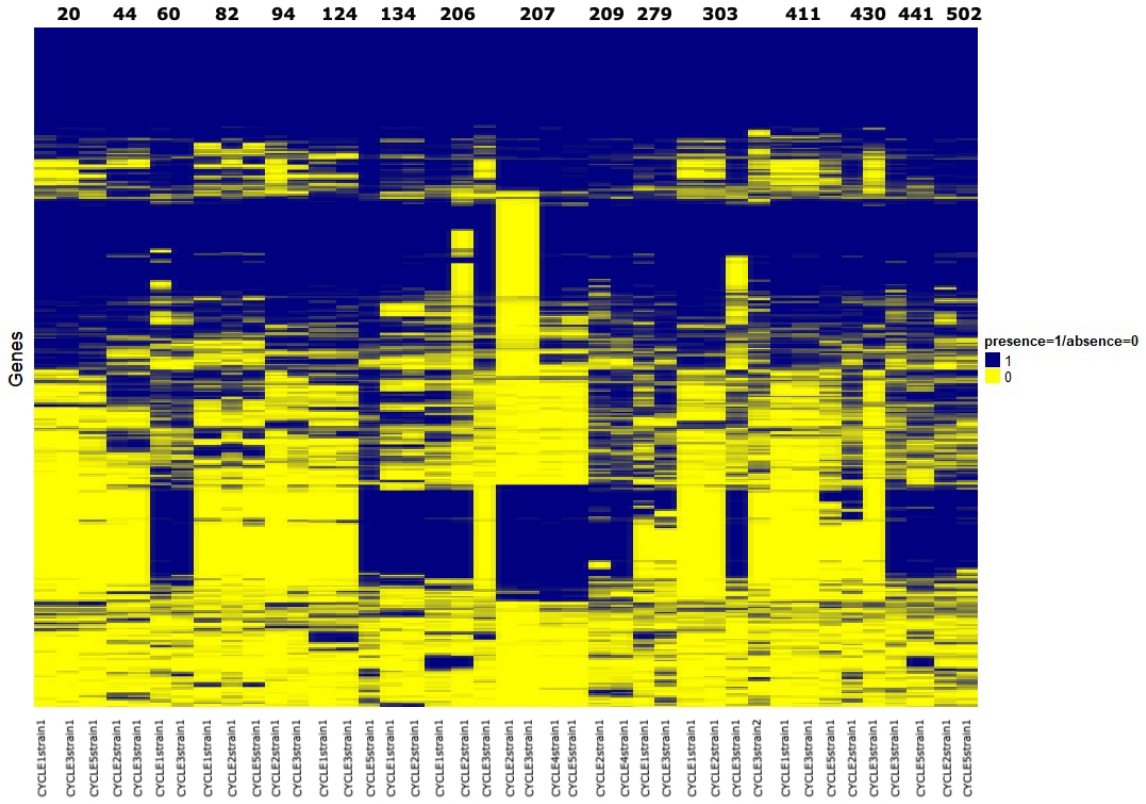

Figure 2 Heatmap of putative strains present of Prevotella copri ordered by individuals. Columns representing the strains and rows the genes of the pangenome used, numbers at the top correspond to the individuals.

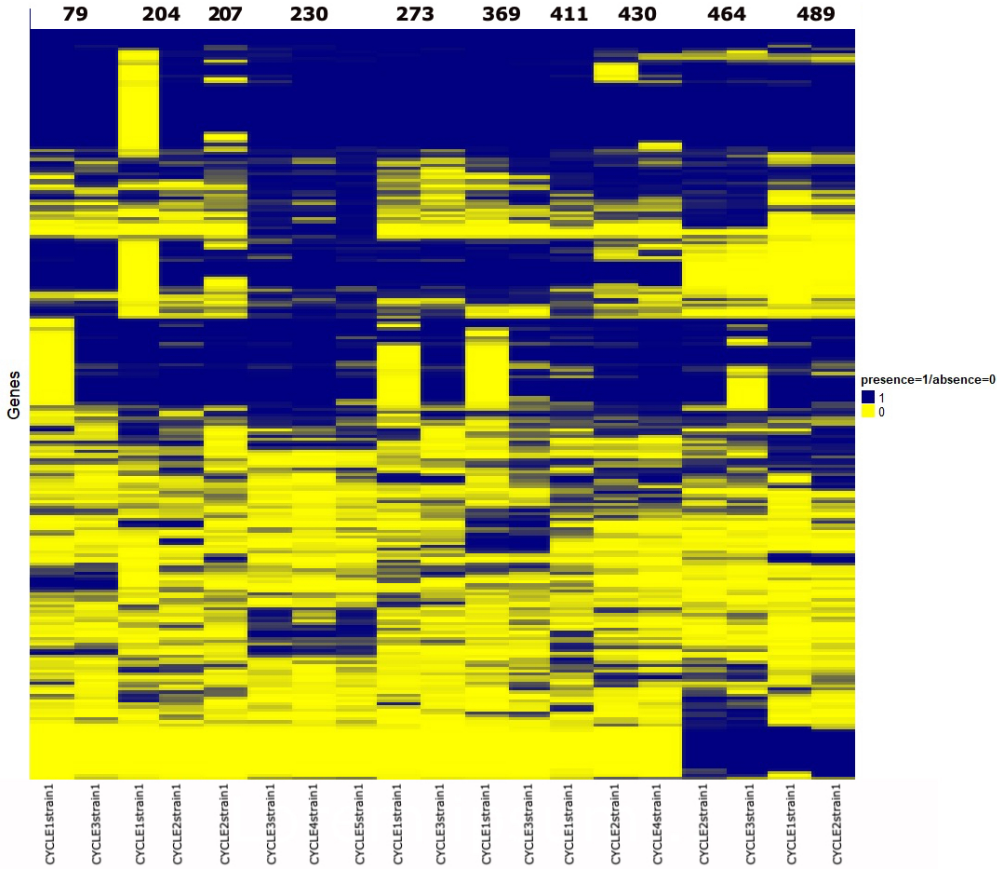

Figure 3 Heatmap of putative strains present of Bacteroides fragilis ordered by individuals. Columns representing the strains and rows the genes of the pangenome used, numbers at the top correspond to the individuals.

Table 1 Prevotella copri strain variation in the same individual at different sampling periods. In each individual, at the top there is the strain with which we compared the rest of the strains.

	TOTAL	SNPs	IN/DELS
INDIVIDUAL207			
CYCLE2strain1	-	-	-
CYCLE3strain1	334	229	16
CYCLE4strain2	20492	14021	54
CYCLE4strain3	512	344	11
CYCLE5strain1	513	326	11
CYCLE5strain2	29294	20458	111
INDIVIDUAL369			
CYCLE1strain1	-	-	-
CYCLE1strain2	187	122	-
CYCLE1strain3	256	147	1
CYCLE1strain4	6794	4440	11
CYCLE2strain1	29	20	-
CYCLE3strain1	152	102	1
CYCLE3strain2	197	116	-
INDIVIDUAL411			
CYCLE1strain1	-	-	-
CYCLE3strain1	12999	10383	197
CYCLE5strain1	9715	7738	123
INDIVIDUAL430			
CYCLE2strain1	-	-	-
CYCLE3strain1	123	75	-
CYCLE3strain2	143	98	-
CYCLE5strain1	29632	21076	159
INDIVIDUAL20			
CYCLE1strain1	-	-	-
CYCLE3strain1	5303	4566	130
CYCLE5strain1	10955	7446	32
CYCLE5strain2	4584	3901	124
INDIVIDUAL33			
CYCLE2strain1	-	-	-
CYCLE3strain1	181	118	-
INDIVIDUAL44			
CYCLE2strain1	-	-	-
CYCLE2strain2	12756	8911	34
CYCLE3strain1	11839	8165	41
INDIVIDUAL60			
CYCLE1strain1	-	-	-
CYCLE1strain2	408	274	-
CYCLE1strain3	40	27	-
CYCLE3strain1	521	335	1
CYCLE3strain2	22755	18086	271
INDIVIDUAL82			
CYCLE1strain1	-	-	-
CYCLE2strain1	10098	6834	27
CYCLE2strain2	200	125	-
CYCLE5strain1	11120	7603	30
CYCLE5strain2	869	572	-
CYCLE5strain3	63	42	-

	INDIVIDUAL84		
CYCLE3strain1	-	-	-
CYCLE5strain1	1075	757	10
CYCLE5strain2	1214	803	3
	INDIVIDUAL94		
CYCLE2strain1	-	-	-
CYCLE3strain1	11247	7664	32
CYCLE3strain2	50	31	1
CYCLE3strain3	365	286	19
	INDIVIDUAL124		
CYCLE1strain1	-	-	-
CYCLE1strain2	154	105	-
CYCLE1strain3	169	104	4
CYCLE3strain1	442	286	4
CYCLE3strain2	33	29	2
CYCLE5strain1	287	183	3
CYCLE5strain2	241	150	1
	INDIVIDUAL134		
CYCLE1strain1	-	-	-
CYCLE2strain1	7531	5051	11
	INDIVIDUAL206		
CYCLE1strain1	-	-	-
CYCLE2strain1	463	301	-
CYCLE2strain2	1487	1072	44
CYCLE3strain1	191	129	-
CYCLE5strain1	27931	19783	115
CYCLE5strain2	453	303	-
	INDIVIDUAL209		
CYCLE2strain1	-	-	-
CYCLE3strain1	140	94	-
CYCLE4strain1	5234	3540	14
CYCLE5strain1	191	128	-
	INDIVIDUAL279		
CYCLE1strain1	-	-	-
CYCLE1strain2	558	375	-
CYCLE3strain1	1848	1222	6
CYCLE3strain2	5335	4626	28
	INDIVIDUAL303		
CYCLE1strain1	-	-	-
CYCLE2strain1	3945	3386	79
CYCLE3strain1	104	66	1
CYCLE3strain2	26807	18908	107
	INDIVIDUAL441		
CYCLE3strain1	-	-	-
CYCLE5strain1	432	287	-
CYCLE5strain2	3410	2293	9
	INDIVIDUAL464		
CYCLE3strain1	-	-	-
CYCLE5strain1	131	91	1

	INDIVIDUAL502		
CYCLE2strain1	-	-	-
CYCLE5strain1	181	110	-

Table 2 Bacteroides fragilis strain variation in the same individual at different sampling periods. In each individual, at the top there is the strain with which we compared the rest of the strains.

	TOTAL	SNPs	IN/DELS
	INDIVIDUAL207		
CYCLE2strain1	-	-	-
CYCLE3strain1	500	398	11
	INDIVIDUAL369		
CYCLE1strain1	-	-	-
CYCLE3strain1	923	588	14
	INDIVIDUAL411		
CYCLE1strain1	-	-	-
CYCLE3strain1	785	500	3
	INDIVIDUAL430		
CYCLE2strain1	-	-	-
CYCLE4strain1	269	166	5
CYCLE4strain2	1321	850	19
	INDIVIDUAL79		
CYCLE1strain1	-	-	-
CYCLE1strain2	1174	746	12
CYCLE3strain1	1114	723	10
	INDIVIDUAL204		
CYCLE1strain1	-	-	-
CYCLE2strain1	362	254	13
CYCLE3strain1	271	142	-
CYCLE5strain1	7878	6587	68
	INDIVIDUAL230		
CYCLE3strain1	-	-	-
CYCLE4strain1	372	248	7
CYCLE5strain1	247	159	-
CYCLE5strain2	326	217	7
	INDIVIDUAL232		
CYCLE2strain1	-	-	-
CYCLE4strain1	1418	891	18
CYCLE5strain1	1495	988	18
	INDIVIDUAL273		
CYCLE1strain1	-	-	-
CYCLE3strain1	883	589	9
CYCLE3strain2	17624	16197	271
CYCLE3strain3	2279	1476	26
	INDIVIDUAL464		
CYCLE2strain1	-	-	-
CYCLE3strain1	350	254	8
	INDIVIDUAL489		
CYCLE1strain1	-	-	-

CYCLE1strain2	752	443	2
CYCLE2strain1	490	282	2
CYCLE2strain2	571	367	1
	INDIVIDUAL534		
CYCLE2strain1	-	-	-
CYCLE3strain1	35	16	4

DISCUSSION

Our MAG analysis suggested that there are 1-4 strains of *Prevotella copri* and 1-3 strains of *Bacteroides fragilis* in the intestine of infants at any given time. These numbers may represent only the numerically dominant strains and may not be the total number of strains. A more exhaustive study is required to obtain the total diversity [24]. Our results contradict previous reports indicating that individuals carry 1 strain of *Bacteroides fragilis* [25]. We also found evidence of strain turnover in *Prevotella copri* and *Bacteroides fragilis* over 5-12 months. These results are in contrast with studies showing species and strain stability in the gut microbiome over time and where an individual maintains particular strains for extended periods [25][26][27]. When focusing on *P. copri* strains (Fig. 2), we show that on each of the 16 individuals, different strains are found at different time points, considering the number of variations between strains (Table 1). We also found evidence that there is a constant change in the strain present in an individual, and it doesn't seem to return to a strain from a previous time point, the same can be said about the 10 individuals corresponding to *B. fragilis* (Fig. 3, Table 2). Other authors have found evidence of different strains of *P. copri* and *B. fragilis* present in an individual simultaneously [28]. These results could resemble recent findings in *E. coli,* with many numerically dominant and satellite strains changing over time in the human microbiome [24].

Accessory genomes could be as large as the core genomes, and the accessory genome is probably the main source of genetic innovation in strains displaying different phenotypes [29]. A large proportion of the accessory genome is formed by horizontally transferred genes [29]. Given that different strains could have a high diversity of accessory genes differentiating one from the other, strain turnover could have a relevant impact on several physiological processes linked to the microbiome. We show the presence of different strains present in an individual at

different time points. Still, more studies at the strain level of the microbiome composition are needed to assess the impact of its variations, the relevance of the changes in different conditions or diseases, and the potential therapeutic benefit that a controlled modification could have on an individual. Based on the strains' variability, the microbiome stability, and the microbiome dynamics found in this study, it is essential to consider a whole genome approach. The 16S rRNA gene sequencing only should be used to reveal the composition down to the species level of the microbiome and to show the overall ecological composition of bacterial communities. Focusing only down to the species level could ignore relevant interactions and turnovers that could significantly affect the microbiome as a whole and the role it plays within the environment.

We highlight the potential of MAGs in understanding microbial communities without the need for individual isolation and culturing, revealing the presence of diverse microorganisms at different taxonomic levels, including common groups found in the gut microbiome. This genomic information contributes to a better understanding of the human microbiota. However, there are limitations in the assembly and binning processes, affecting the accuracy of the obtained metagenome-assembled genomes (MAGs) [30][31].

Microbiomes are complex/dynamic microbial communities that at strain level could significantly impact the microbiome's functionality, resulting in different outcomes for the animal host. There is a need for more studies at the strain level to improve our understanding of the gut microbiome, as the tools needed for such studies also continue to improve.

REFERENCES

1. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. Microbiology and molecular biology reviews. 2004 Dec;68(4):669-85.
2. Fuhrman JA, Steele JA, Hewson I, Schwalbach MS, Brown MV, Green JL, Brown JH. A latitudinal diversity gradient in planktonic marine bacteria. Proceedings of the National Academy of Sciences. 2008 Jun 3;105(22):7774-8.
3. Riesenfeld CS, Goodman RM, Handelsman J. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. Environmental microbiology. 2004 Sep;6(9):981-9.
4. Louca S, Doebeli M, Parfrey LW. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. Microbiome. 2018 Dec;6:1-2.
5. Denamur E, Clermont O, Bonacorsi S, Gordon D. The population genetics of pathogenic Escherichia coli. Nat Rev Microbiol. 2021 Jan;19(1):37-54.
6. He Y, Wang H, Chen L. Comparative secretomics reveals novel virulence-associated factors of Vibrio parahaemolyticus. Frontiers in Microbiology. 2015 Jul 17;6:707.
7. Touchon M, Perrin A, De Sousa JA, Vangchhia B, Burn S, O'Brien CL, Denamur E, Gordon D, Rocha EP. Phylogenetic background and habitat drive the genetic diversification of Escherichia coli. PLoS genetics. 2020 Jun 12;16(6):e1008866.
8. Martiny JB, Jones SE, Lennon JT, Martiny AC. Microbiomes in light of traits: a phylogenetic perspective. Science. 2015 Nov 6;350(6261):aac9323.
9. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS genetics. 2009 Jan 23;5(1):e1000344.
10. Zwick ME, Joseph SJ, Didelot X, Chen PE, Bishop-Lilly KA, Stewart AC, Willner K, Nolan N, Lentz S, Thomason MK, Sozhamannan S. Genomic characterization of the Bacillus cereus sensu lato species: backdrop to the evolution of Bacillus anthracis. Genome research. 2012 Aug 1;22(8):1512-24.
11. Cifuentes SG, Graham J, Loayza F, Saraiva C, Salinas L, Trueba G, Cárdenas PA. Evaluation of changes in the faecal resistome associated with children's exposure to domestic animals and food animal production. Journal of global antimicrobial resistance. 2022 Dec 1;31:212-5.

12. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. https://github.com/s-andrews/FastQC
13. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114-20.
14. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997. 2013 Mar 16.
15. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. Gigascience. 2021 Feb;10(2):giab008.
16. Saheb Kashaf S, Almeida A, Segre JA, Finn RD. Recovering prokaryotic genomes from host-associated, short-read shotgun metagenomic sequencing data. Nature Protocols. 2021 May;16(5):2520-41.
17. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015 May 15;31(10):1674-6.
18. Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome. 2018 Dec;6(1):1-3.
19. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep 15;34(18):3094-100.
20. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, Tan AL, Paul LM, Brackney DE, Grewal S, Gurfield N. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome biology. 2019 Dec;20(1):1-9.
21. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution. 2013 Jan 16;30(4):772-80.
22. Seemann T. Snippy: rapid haploid variant calling and core genome alignment. GitHub https://github. com/tseemann/snippy. Accessed. 2020 Feb;10.
23. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, Mailyan A, Manghi P, Scholz M, Thomas AM, Valles-Colomer M. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. elife. 2021 May 4;10:e65088.

24. Hu D, Fuller NR, Caterson ID, Holmes AJ, Reeves PR. Single-gene long-read sequencing illuminates Escherichia coli strain dynamics in the human intestinal microbiome. Cell Rep. 2022 Jan 11;38(2):110239.
25. Zhao S, Lieberman TD, Poyet M, Kauffman KM, Gibbons SM, Groussin M, Xavier RJ, Alm EJ. Adaptive Evolution within Gut Microbiomes of Healthy People. Cell Host Microbe. 2019 May 8;25(5):656-667.e8.
26. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, Brady A, Creasy HH, McCracken C, Giglio MG, McDonald D. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature. 2017 Oct 5;550(7674):61-6.
27. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, Waller A, Mende DR, Kultima JR, Martin J, Kota K. Genomic variation landscape of the human gut microbiome. Nature. 2013 Jan 3;493(7430):45-50.
28. Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. ConStrains identifies microbial strains in metagenomic datasets. Nature biotechnology. 2015 Oct;33(10):1045-52.
29. Zhu A, Sunagawa S, Mende DR, Bork P. Inter-individual differences in the gene content of human gut bacterial species. Genome Biol. 2015 Apr 21;16(1):82. doi: 10.1186/s13059-015-0646-9. PMID: 25896518; PMCID: PMC4428241.
30. Yang C, Chowdhury D, Zhang Z, Cheung WK, Lu A, Bian Z, Zhang L. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. Computational and Structural Biotechnology Journal. 2021 Jan 1;19:6301-14.
31. Chen LX, Anantharaman K, Shaiber A, Eren AM, Banfield JF. Accurate and complete genomes from metagenomes. Genome Research. 2020 Mar 1;30(3):315-33.