BMC Microbiology

RESEARCH

Open Access

Full-length 16S rRNA Sequencing Reveals Gut Microbiome Signatures Predictive of MASLD in children with obesity

Yu-Cheng Lin[1,2,3*], Chi-Chien Wu[3], Yun-Er Li[1], Chun-Liang Chen[1], Chia-Ray Lin[4] and Yen-Hsuan Ni[4]

Abstract

Background The gut microbiota plays a crucial role in metabolic dysfunction-associated steatotic liver disease (MASLD). Next-generation sequencing technologies are essential for exploring the gut microbiome. While recent advancements in full-length 16S (FL16S) rRNA sequencing offer better taxonomic resolution, whether they establish stronger associations with the risk of MASLD remains to be determined.

Method This study utilized long-read FL16S and short-read V3-V4 16S rRNA sequencing to profile gut microbiome compositions in age-, sex-, and BMI-matched case–control pairs of obese children with and without MASLD. A random forest predictive model was employed, using gut-microbiota features selected based on the top 35 most abundant taxa or a linear discriminant analysis score greater than 3. The model's performance was evaluated by comparing the area under the receiver operating characteristic curve (AUC) through a tenfold cross-validation method.

Results Subjects with MASLD exhibited significantly elevated serum alanine aminotransferase, triglycerides, and homeostasis model assessment of insulin resistance levels compared to controls. At the genus level, the gut microbiome compositions detected by both FL16S and V3-V4 sequencing were similar, predominantly comprising *Phocaeicola* and *Bacteroides*, followed by *Prevotella*, *Bifidobacterium*, *Parabacteroides*, and *Blautia*. The AUC for the model based on FL16S sequencing data (86.98%) was significantly higher than that based on V3-V4 sequencing data (70.27%), as determined by DeLong's test ($p = 0.008$).

Conclusion FL16S rRNA sequencing data demonstrates stronger associations with the risk of MASLD in obese children, highlighting its potential for real-world clinical applications.

Keywords MASLD, Gut microbiota, 16S rRNA, Sequencing, Random forest

Background

Metabolic dysfunction-associated steatotic liver disease (MASLD) represents a significant health challenge worldwide [1]. Identifying MASLD early is challenging due to its often-asymptomatic nature. Commonly, MASLD comes to light during incidental findings from ultrasound exams or liver function tests, underscoring the necessity for early and proactive screening among those at risk to ensure prompt diagnosis and intervention [2].

Advancements in gut microbiota research have surged with the introduction of culture-independent next-generation sequencing technologies. Amplicon sequencing

*Correspondence:
Yu-Cheng Lin
yclin116@vghtpe.gov.tw
[1] Department of Pediatrics, Taipei Veterans General Hospital, Taipei City, Taiwan
[2] Department of Healthcare Administration, Asia Eastern University of Science and Technology, New Taipei City, Taiwan
[3] Department of Pediatrics, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[4] Departments of Pediatrics, National Taiwan University Hospital, Taipei, Taiwan

of the 1500 bp 16S rRNA gene, encompassing 9 variable regions (V1-V9), has become a powerful approach for bacterial identification [3]. Changes in gut microbiota have been linked to various metabolic disorders, such as obesity and MASLD [4–6]. A notable study from our team showed that an imbalance in gut *Desulfovibrio* can worsen MASLD by making the intestinal barrier more permeable and increasing liver CD36 expression [7]. This highlights the potential of gut microbiome profiles to serve as biomarkers for MASLD identification [8, 9].

Traditionally, short-read sequencing, such as the V3-V4 (V3V4) approach, classify sequences with more than 97% similarity into operational taxonomic units (OTUs). These OTUs are then matched against reference databases to determine their taxonomic classification. However, this approach can lead to taxonomic ambiguities [10]. For instance, V3V4 sequencing may fail to differentiate between closely related species like *Escherichia coli* and *Shigella* serogroups due to their high sequence identity exceeding 99% [11]. Recent advances allow high-throughput gene sequencing data to be analyzed using amplicon sequence variants (ASVs), which provide exact resolution down to single-nucleotide differences. ASV-based approach offers greater precision and taxonomic resolution than OTU-based approach and has thus gained broader adoption in microbial community sequencing [12, 13].

Another recent advancement is that long-read full-length 16S (FL16S) rRNA sequencing provides high taxonomic resolution for profiling gut microbiota [14]. Compared to traditional short-read V3V4 sequencing, long-read sequencing enables complete 16S rRNA gene analysis, allowing for a more accurate ASV inference and species annotations [15]. The enhanced taxonomic resolution is particularly advantageous for analyzing samples with complex microbial communities [16], such as human fecal samples. Currently, FL16S rRNA sequencing offers comparable costs and significantly better taxonomic resolution, making it a preferable choice for clinical use. However, due to a lack of direct comparison studies, its superiority over short-read sequencing in clinical gut microbiome analysis remains uncertain.

In this study, we aimed to directly compare the effectiveness of these two 16S rRNA sequencing methods in clinical settings, using MASLD as a representative clinical disease for analysis. We assessed the utility of FL16S by comparing the performance of random forest (RF) models that used microbial signatures from either FL16S or V3V4 sequencing data to distinguish between obese children with and without MASLD in an age-, sex-, and BMI-matched case–control cohort.

Methods

Patient enrollment and data collection

This study was based on our previous cohort of obese children aged 6–18 years who underwent screening for hepatic steatosis using ultrasonography [17]. Data collected included age, gender, and body mass index (BMI). Fasting venous blood samples were analyzed for alanine aminotransferase (ALT), aspartate aminotransferase (AST), triglycerides (TG), total cholesterol (TCHO), gamm-glutamyltransferase (GGT), and high-density lipoprotein (HDL) cholesterol. Insulin resistance was quantified using the homeostatic model assessment for insulin resistance (HOMA-IR) index. All subjects met the criteria for obesity, defined as BMI values > 95th percentile for age and sex, according to standards from the Ministry of Health and Welfare, Taiwan [18]. MASLD was defined according to consensus criteria endorsed by multiple pediatric societies, requiring the presence of steatosis along with cardiometabolic risk factors, including overweight or obesity, with abnormalities in glucose levels, blood pressure, triglycerides, and HDL cholesterol [19].

To develop a predictive model for MASLD, we designed a case–control study by selecting 26 children with MASLD and 26 matched controls without MASLD from this cohort. The pairs were matched by age, sex, and BMI, as these factors are known to influence gut microbiome composition [20–22]. This matched-pair design establishes a robust framework for investigating the independent role of the gut microbiome in MASLD. The study was approved by the ethics committee of Far Eastern Memorial Hospital, and informed consent was obtained from all parents.

Fecal DNA extraction and sample preparation

Total fecal DNA was extracted using a QIAamp PowerFecal Pro DNA Kit (Cat. No.: 51804, Qiagen, Germany). The DNA's quantity and quality were assessed using an NanoPhotometer™ (Implen, Germany) and stored at -80°C until further use. The 16S genes were amplified from the extracted fecal DNA using specific primers (Table S1).

For short read sequencing, the V3V4 region was amplified using the primer set 341F (5'-CCTACGGGNGGC WGCAG-3') and 806R (5'-GACTACHVGGGTAT CTAATCC-3') following Illumina's 16S Metagenomic Sequencing Library Preparation protocol. PCR was performed with 12.5 ng of gDNA and KAPA HiFi HotStart ReadyMix (Roche, REF: 07958935001) under the following conditions: 95°C for 3 min; 25 cycles of 95°C for 30 s, 55°C for 30 s, and 72°C for 30 s; followed by a final extension at 72°C for 5 min and a hold at 4°C. The PCR products were analyzed on a 1.5% agarose gel to verify the

presence of a distinct band approximately 500 bp in size. Subsequently, the products were purified using AMPure XP beads (Beckman, PN: A63882).

For full-length 16S rRNA gene amplification (V1-V9 regions), barcoded primers containing a 5' buffer sequence (GCATC), a 16-base barcode, and degenerate 16S-specific sequences (Forward: 5'Phos/GCATC-16-base barcode-AGRGTTYGATYMTGGCTCAG-3'; Reverse: 5'Phos/GCATC-16-base barcode-RGYTAC CTTGTTACGACTT-3', where R = A/G, Y = C/T, M = A/C) were used. PCR was carried out with 2 ng of gDNA and KAPA HiFi HotStart ReadyMix (Roche, REF: 07958935001) under the following conditions: 95°C for 3 min; 20–27 cycles (depending on the sample) of 95°C for 30 s, 57°C for 30 s, and 72°C for 60 s; followed by a final extension at 72°C for 5 min and a hold at 4°C. The PCR products were analyzed on a 1% agarose gel to verify the presence of a distinct band approximately 500 bp in size. Subsequently, the products were purified using AMPure PB beads (PacBio, PN: 100–265-900).

For the V3V4 sequencing, a secondary PCR was performed by using the 16S rRNA V3V4 region PCR amplicon and Nextera XT Index Kit with dual indices and Illumina sequencing adapters. The indexed PCR product quality was verified using Qubit 4.0 Fluorometer (Thermo Scientific) and Qsep100TM system. Library was sequenced on an Illumina MiSeq platform and paired 300-bp reads were generated.

For FL16S sequencing, the SMRTbell library was incubated with sequencing primer v4 and sequel II Binding Kit (2.1) for the primer annealing and polymerase binding. Sequencing was performed in the circular consensus sequence (CCS) mode on a PacBio Sequel IIe instrument to generate the HiFi reads with Predicted Accuracy (Phred Scale) of 30. The ZymoBIOMICS Microbial Community DNA standard (D6306, Zymo Reaearch) was added as a sample in the library and used as positive control to evaluate the sequencing quality.

Sequencing data analysis

For FL16S sequencing, CCS reads were processed using the SMRT Link software with a minimum predicted accuracy threshold of 0.9. For V3V4 sequencing, 300 bp paired-end raw reads were generated for amplicon sequencing, with each sample demultiplexed based on its unique barcode. The resulting sequences from both FL16S and V3V4 sequencing were trimmed using the QIIME2 Cutadapt plugin (v2021.4; https://qiime2.org/) [23]. Next, both platforms were analyzed using the DADA2 plugin in a workflow that included quality filtering, dereplication, error model learning, ASV inference, and chimera removal [24, 25].

Taxonomy classification was primarily conducted using the NCBI reference database (version 2020/7), with the vsearch feature classifier plug in of QIIME2 [26, 27]. In our study, because the V3-4 sequence reads were relatively short, they could not be accurately annotated for *Faecalibacterium* using the NCBI database alone. To address this, we incorporated the SILVA database (version 132) specifically for genus-level identification of *Faecalibacterium*.

Validation of taxonomic accuracy

To ensure the accurate quantification and validation of taxonomic assignments derived from sequencing data, PCR-based validation was performed for seven key taxa: *Bacteroides ovatus*, *Prevotella copri*, *Faecalibacterium prausnitzii*, *Bacteroides stercoris*, *Phascolarctobacterium*, *Bifidobacterium*, and *Roseburia*. These taxa were selected for validation based on their high relative abundance observed in the FL16S 16S rRNA sequencing results. Specific primers targeting each taxon were designed (Table S2), and PCR amplification followed by electrophoresis was performed according to standard protocols. The "relative PCR quantity (%)" for each taxon was calculated as a normalized signal intensity, using the formula: relative PCR quantity = [(signal intensity of sample—signal intensity of background) ÷ (signal intensity of the highest value in each taxon—signal intensity of background)] × 100.

In addition to PCR validation, the ZymoBIOMICS Microbial Community DNA standard (D6306, Zymo Research) was used to compare the accuracy of V3V4 amplicon sequencing and FL16S rRNA sequencing. The sequencing results from both methods were evaluated against the theoretical composition of the DNA standard. This comparative analysis further validated the robustness of the metagenomic data.

Statistical analysis

Potential microbial biomarkers were identified using Linear Discriminant Analysis Effect Size (LEfSe) analysis [28], with linear discriminant analysis (LDA) employed to assess the effect size of differentially abundant taxa. Taxa with an LDA score (log10) > 3 were considered significant. The R package pROC (version 1.18.0) was employed to generate the receiver operating curves (ROCs) and assessed the AUCs' 95% confidence intervals. Delong's test was conducted to compare different ROCs' AUCs [29].

Results

Baseline characteristics of subjects

The baseline characteristics of the 26 matched case–control pairs are shown in Table 1. Obese children with

Table 1 Clinical characteristics of age-, sex-, and BMI-matched obese children with and without MASLD

Variable	Non-MASLD	MASLD	P-value
Age (year)	12.7 ± 2.3	12.5 ± 2.1	0.723
Male [n (%)]	80.8%	76.9%	0.740
BMI (kg/m^2)	29.0 ± 2.9	29.8 ± 3.4	0.360
AST (U/L)	18.7 ± 4.3	32.5 ± 14.5	< 0.001
ALT (U/L)	16.5 ± 6.1	53.7 ± 33.2	< 0.001
GGT (U/L)	15.9 ± 6.1	31.7 ± 17.8	< 0.001
T-CHO (mg/dl)	158.3 ± 22.5	174.1 ± 34.5	0.057
TG (mg/dl)	87.4 ± 33.2	122.3 ± 60.7	0.013
HDL (mg/dl)	47.0 ± 10.3	44.9 ± 10.6	0.485
HOMA-IR	4.2 ± 2.3	6.6 ± 3.5	0.006

Non-MASLD subjects were matched with MASLD subjects based on age, sex, and BMI. All values are presented as mean ± SEM

Abbreviations: ALT alanine aminotransferase, *AST* aspartate aminotransferase, *BMI* body mass index, *HDL* high-density lipoprotein cholesterol, *HOMA-IR* homeostatic model assessment for insulin resistance index, *GGT* gamma-glutamyltransferase, *T-CHO* total cholesterol, *TG* triglyceride

MASLD exhibited significantly higher levels of ALT, AST, GGT, TG, and HOMA-IR index compared to non-MASLD controls. By matching participants based on age, sex, and BMI [28]—factors known to influence gut microbiome composition—this study design effectively minimizes potential confounding, allowing for a more accurate evaluation of the relationship between the gut microbiome and MASLD.

Gut-microbiota profile comparison across different sequencing platforms

This study utilized two sequencing methodologies, FL16S and V3V4, to decode the gut-microbiota profile from identical fecal DNA specimens. The generated sequences were aligned against the NCBI database for taxonomic classification. We selected the NCBI database as the reference for this study to ensure the inclusion of newly identified species, a critical consideration for our analysis. In addition, after quality filtering at Q30, the average sampling depth for FL16S rRNA sequencing exceeded 10,000 high-quality reads per sample, while V3V4 sequencing achieved over 30,000 reads per sample. Rarefaction curve analysis confirmed that a depth of 10,000 reads per sample was sufficient to capture the microbial diversity.

A Venn diagram depicted the gut microbiota's quantity, elucidating unique and shared taxa across both sequencing methods. The analysis revealed 48 families common to both platforms (Fig. 1A), with 23 families uniquely detected by FL16S and 25 families by V3V4. At the genus level, FL16S sequencing identified a broader spectrum

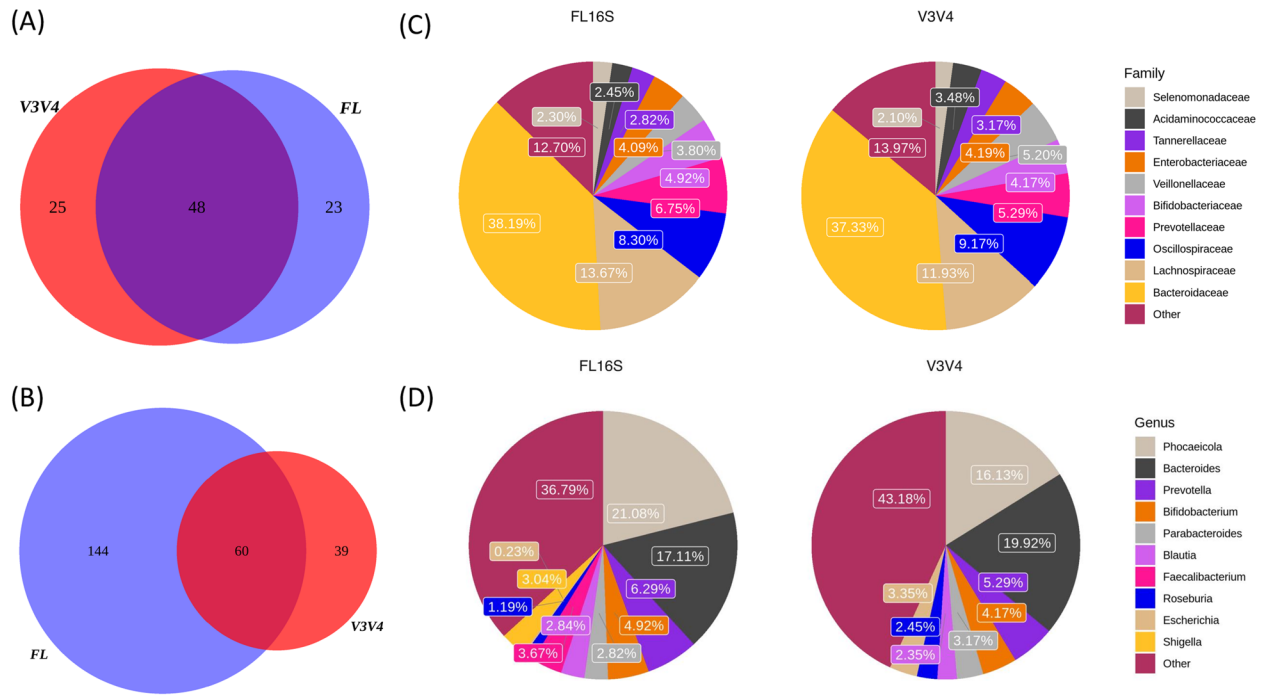

Fig. 1 Bacterial taxonomy comparison between FL16S and V3V4 sequencing methods. This figure presents a Venn diagram illustrating the overlap and unique bacterial taxa identified by FL16S and V3V4 sequencing at the (**A**) family and (**B**) genus levels. Accompanying pie charts detail the taxonomic composition at the (**C**) family and (**D**) genus levels, showcasing the relative abundance (%) of each taxon. FL16S, full-length 16S rRNA sequencing; V3V4, V3-4 16S rRNA sequencing

of gut microbiota, detecting 144 genera compared to 39 genera identified by V3V4 sequencing (Fig. 1B).

At the family level, the bacterial compositions showed comparable similarities between FL16S and V3V4 sequencing (Fig. 1C). At the genus level, *Phocaeicola* (FL16S: 21.08% vs. V3V4: 16.13%) and *Bacteroides* (FL16S: 17.11% vs. V3V4: 19.92%) were predominant in the cohort. Additionally, genera such as *Prevotella*, *Bifidobacterium*, *Parabacteroides*, and *Blautia* were also abundant but with relatively lower proportions compared to the top genera (Fig. 1D).

Next, Subjects were grouped based on MASLD status and sequencing platforms ($n=26$ per group) as follows: (1) MASLD + FL16S (FL16S/F), (2) MASLD + V3V4 (V3V4/F), (3) Non-MASLD + FL16S (FL16S/N), and (4) Non-MASLD + V3V4 (V3V4/N). The phylum and genus-level taxonomic compositions of the top 10 classified taxa are detailed in Fig. 2. At the phylum level, both sequencing methods predominantly identified *Bacteroidetes* and *Firmicutes*. At the genus level, both FL16S and V3V4 sequencing methods showed broadly similar trends in the relative abundances of key taxa, including *Prevotella*, *Bacteroides*, and *Faecalibacterium*. Boxplots comparing the relative abundances of these genera across groups demonstrated concordance between the two platforms (Fig. 2C), with notable enrichment of *Prevotella* in the MASLD groups and significant differences in *Bacteroides* between FL16S/N and FL16S/F groups. These results indicate that, despite differences in sequencing methodologies, both platforms provide comparable insights into the genus-level composition of gut microbiota associated with MASLD.

Confirmation of taxonomic annotations

PCR-based validation confirmed the presence of seven key taxa, which were selected based on their high

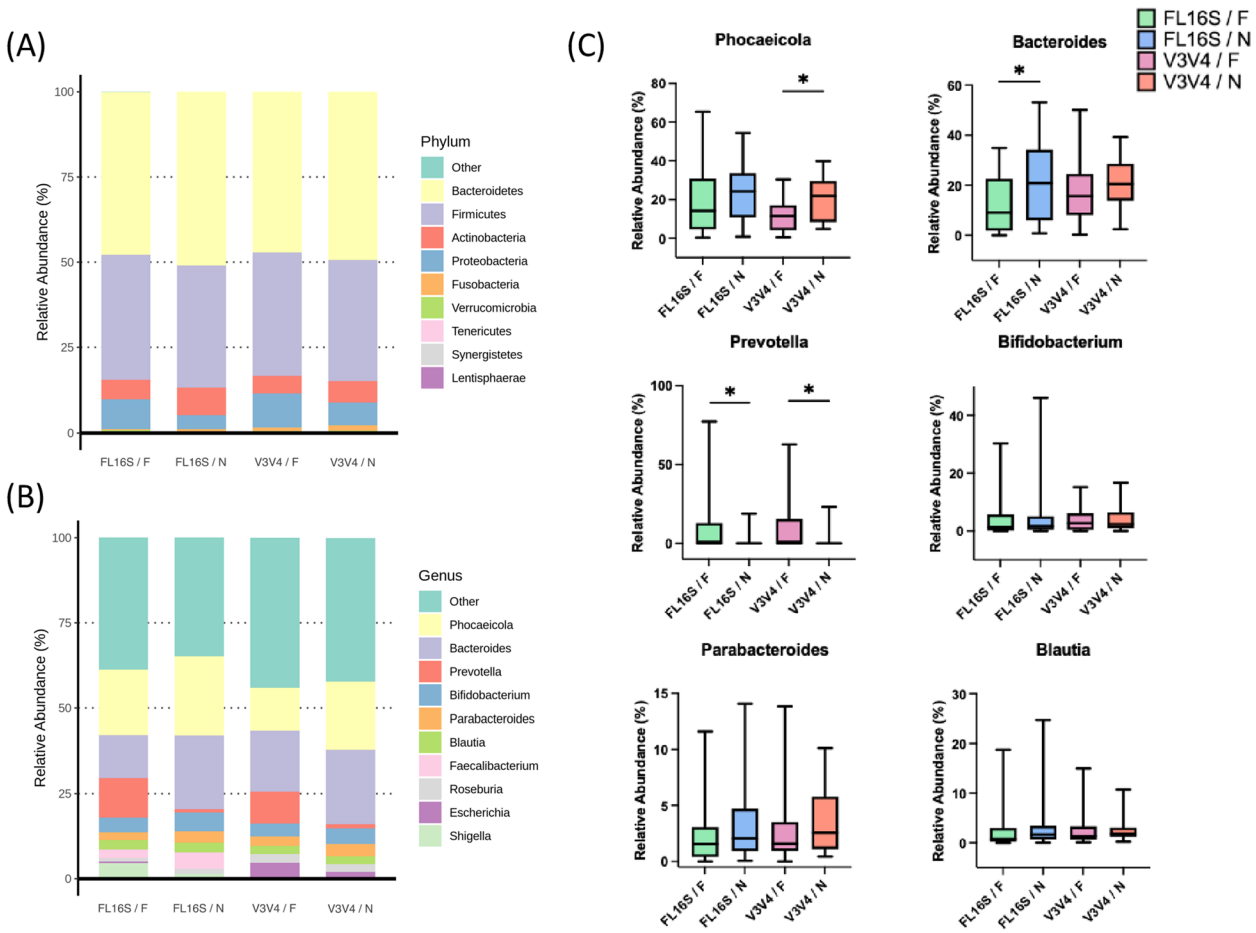

Fig. 2 Distinct composition of gut microbiota in obese children revealed by FL6S and V3V4 sequencing data. Stacked bar charts illustrate the compositions in the gut microbiome at the (**A**) phylum and (**B**) genus levels, as determined by the FL16S and V3V4 sequencing methods; (**C**) Box plot comparing the relative abundance of key bacterial genera across the two sequencing platforms. FL16S, full-length 16S rRNA sequencing; V3V4, V3-4 16S rRNA sequencing; F, Fatty liver group; N, Non-fatty liver group. Statistical significance: $*p<0.05$

relative abundance in 16S rRNA sequencing. As shown in Fig. 3, the relative PCR quantity correlated well with the relative abundances obtained from FL16S sequencing, categorized into three groups: low (< 0.1%), medium (0.1–5%), and high (> 5%). This alignment between FL16S sequencing and PCR results demonstrates the accuracy of taxonomic annotations and reinforces the reliability for profiling gut microbiota.

The ZymoBIOMICS Microbial Community DNA standard (D6306, Zymo Research) was used to evaluate and compare the accuracy of V3V4 amplicon sequencing and FL16S rRNA sequencing. As shown in Fig. S1, the microbial compositions obtained from both methods correlated well with the theoretical composition of the DNA standard, which includes DNA from eight bacterial genera: *Pseudomonas* (4.2%), *Escherichia-Shigella* (10.1%), *Salmonella* (10.4%), *Lactobacillus* (18.4%), *Enterococcus* (9.9%), *Staphylococcus* (15.5%), *Listeria* (14.1%), and *Bacillus* (17.4%). These values were then compared with the results obtained from platforms using full-length and short-read sequencing technologies. While minor variability was observed, both methods are sufficiently accurate for general microbial profiling. The comparison of gut microbiota profiles between the two approaches in this study is unlikely to introduce significant bias, supporting the robustness of the analyses.

Random forest models for MASLD detection using FL16S or V3V4 sequencing data

We developed predictive models employing gut-microbiota features to discern MASLD in obese children. Figure 4 outlines the bioinformatics process utilized for MASLD identification. To create a RF model with broad applicability, we initially selected gut microbiota features that ranked in the top 35 for relative abundance (Table S3) and those with an LDA score greater than 3 (Table S4) at both the genus and species levels. These selected features were then utilized to train the RF classification model. The model's development involved a training set enhanced through tenfold cross-validation, which was employed to maximize data utilization, reduce the impact of random bias, and ensure the model's stability and generalizability, particularly given the limited sample size in our study.

Further refinement of the model was informed by the mean decrease accuracy (MDA) values and the relative abundance of the taxa, aiding in the construction of more precise predictive models. This process led to a focus on the most influential 20% of gut-microbiota features, determined by their MDA ranking [30].

Consequently, the number of features considered in the model was reduced from 121 to 24 for the FL16S sequencing, and from 49 to 10 for the V3V4 sequencing (as Fig. 5A, B), enhancing the model's efficiency and accuracy.

The FL16S-derived RF model highlighted several gut microbiotas with less than 1% abundance indicating their

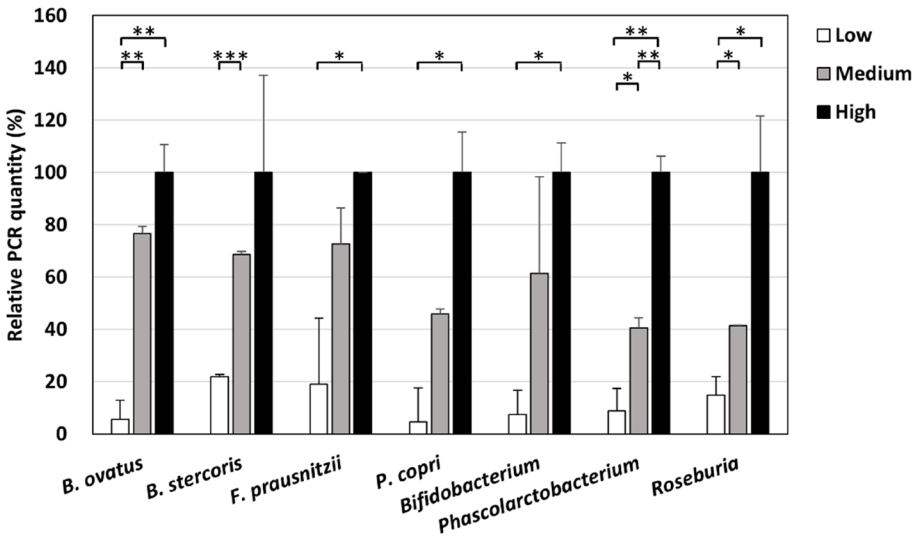

Fig. 3 Validation of FL16S sequencing data by PCR. The bacteria were grouped based on their relative abundance observed in the FL16S rRNA sequencing results: Low (< 0.1%), Medium (0.1%–5%), and High (> 5%). The "relative PCR quantity (%)" for each taxon was calculated as a normalized signal intensity, using the formula: relative PCR quantity = [(signal intensity of sample—signal intensity of background) ÷ (signal intensity of the highest value in each taxon—signal intensity of background)] × 100. Statistical significance: *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$

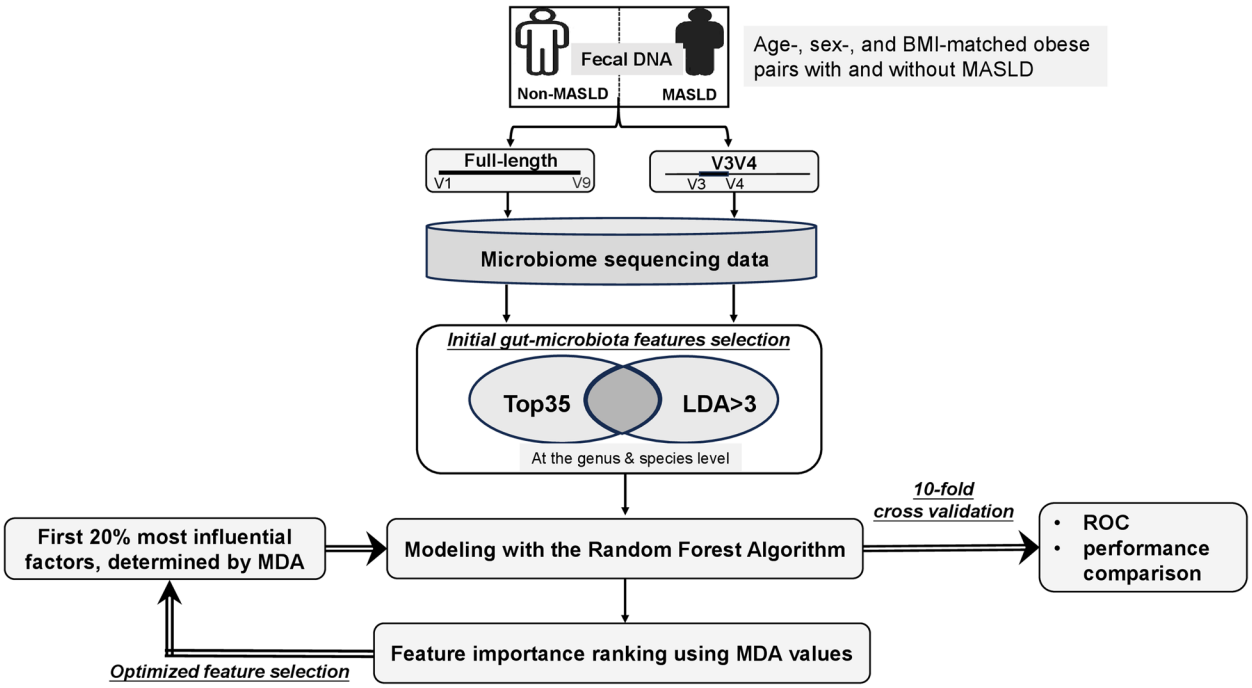

Fig. 4 An overview of the workflow for MASLD prediction in obese children involves selecting gut microbiota features for a two-step random forest modeling process. Initially, we identified key features based on their top 35 relative abundances and LDA scores greater than 3, across both genus and species levels. The model was then optimized by focusing on features with the highest MDA values, specifically targeting the most influential top 20% of features to boost the model's predictive accuracy. LDA, linear discriminant analysis; MDA, mean decrease accuracy; ROC, receiver-operating characteristic

significance despite low abundance. This list includes *Oscillibacter* (0.70%), *Veillonella* (0.59%), *Blautia wexlerae* (0.79%), *Phascolarctobacterium faecium* (0.68%), *Phocaeicola coprophilus* (0.59%), *Bacteroides thetaiotaomicron* (0.47%), and *Bacteroides caccae* (0.39%). Notably, all of these microbiotas, except *Phocaeicola coprophilus*, were also present in the dataset defined by an LDA score greater than 3, further supporting their potential importance.

The RF model demonstrated significantly superior performance with the FL16S sequencing data, achieving an AUC of 86.98%, compared to 70.27% for the V3V4 sequencing data (DeLong's test, $p=0.0079$) (Fig. 5C). Notably, V3V4 sequencing failed to detect *Faecalibacterium* when the sequencing data were aligned with the NCBI database. However, upon alignment with the SILVA database, ASVs were identified as *Faecalibacterium* at the genus level, with a relative abundance of 5.01% in the Non-MASLD group and 2.84% in the MASLD group (Fig. S2).

At the family level, these ASVs were classified as belonging to *Oscillospiraceae*. To address the potential bias caused by using different databases for alignment, the relative abundance of *Oscillospiraceae* identified by the NCBI database was deliberately included in the V3V3 data-based RF model, as shown in Fig. S3A-B. Despite this calibration, there was no significant difference in the performance of the V3V4 data-based RF model, as demonstrated by DeLong's test ($p=0.8157$), illustrated in Fig. S3C.

Discussion

Our study aimed to evaluate the effectiveness of different 16S rRNA sequencing methods for detecting MASLD. Consistent with prior research [31], our findings confirm the role of gut microbiota dysregulation in MASLD development, supporting the potential of gut microbiome profiles as clinically relevant biomarkers for MASLD detection. In addition, using random forest models with FL16S and V3V4 sequencing data, we found that FL16S sequencing demonstrated superior ability to differentiate MASLD in an age-, sex-, and BMI-matched cohort of obese children. FL16S sequencing offers greater resolution at both the genus and species levels, making it a more robust tool for clinical applications.

Recent advances in microbiome research have shifted the focus from phylum-level analyses to deeper taxonomic resolutions, enabling the identification of specific microbial species associated with disease mechanisms [32]. Early studies predominantly focused on broad

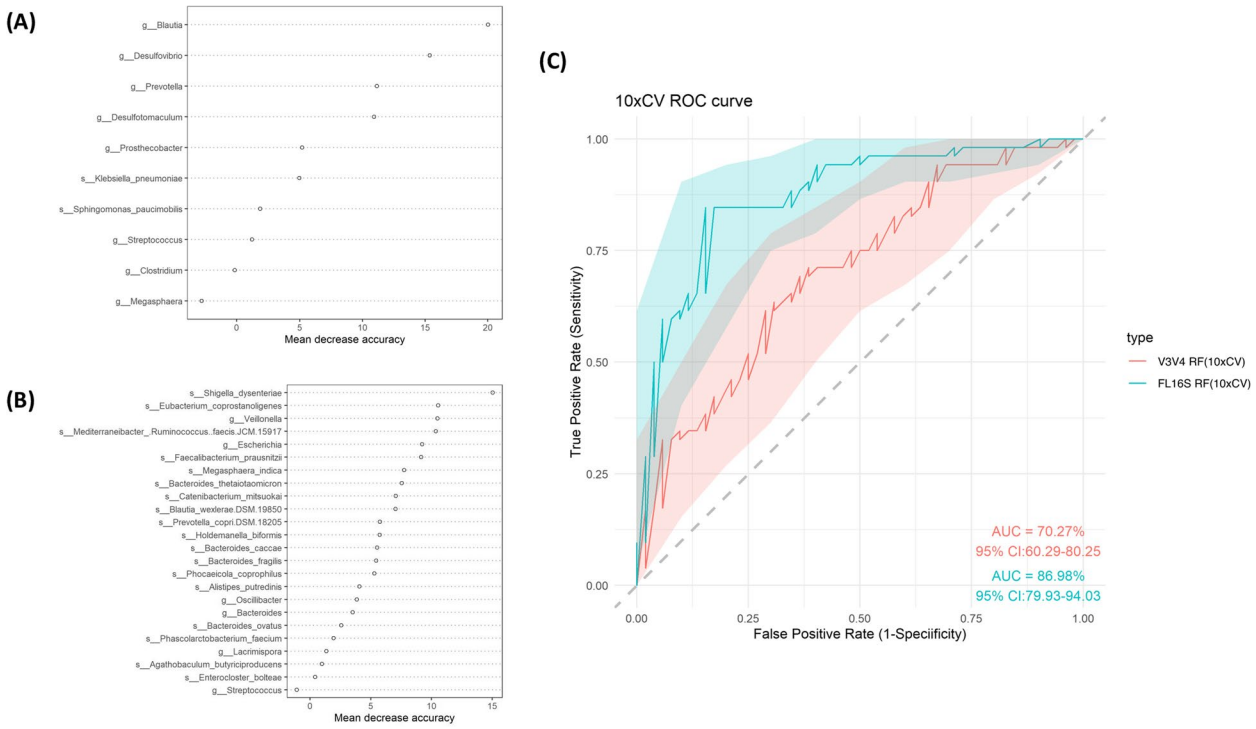

Fig. 5 Gut microbiota features via random forest modeling in MASLD and Non-MASLD Groups. Displayed are the importance scores based on MDA for both (**A**) V3V4 and (**B**) FL16S rRNA sequencing methods. ROC curves (**C**-**D**) were employed to evaluate the prediction accuracy for MASLD, with the calculation of AUCs providing a measure of performance. MDA, mean decrease accuracy; ROC, receiver-operating characteristic; AUC; the area under the ROC curve

taxonomic shifts at the phylum level, reporting increases in *Firmicutes* and *Proteobacteria* and decreases in *Bacteroidetes* among patients with MASLD [33]. However, phylum-level analyses offer limited insights into the specific microbial players driving disease processes. Improved sequencing technologies have facilitated genus-level analyses, identifying microbial groups like *Prevotella* and *Escherichia* that better correlate with disease severity in MASLD [34]. More recently, species-level resolution achieved through advanced methods such as FL16S sequencing may enable more accurate associations with disease stages and patient outcomes, advancing precision diagnostics for complex diseases like MASLD [35].

Our findings highlight the association between decreased *Faecalibacterium* and MASLD, consistent with previous studies [36]. In the FL16S data-based RF model, *Faecalibacterium* demonstrated a MDA value of 8.74, underscoring its importance in detecting MASLD. In contrast, V3V4 sequencing data required alignment with the SILVA database to improve annotation accuracy for *Faecalibacterium*, as the shorter read lengths were insufficient for precise annotation using the NCBI database. This limitation highlights the differences in taxonomic accuracy when using different reference databases [37], particularly with short-read sequencing methods. It underscores the advantage of FL16S sequencing, which offers much longer read lengths, enabling more accurate alignment and annotation.

A limitation of our study lies in the potential uncertainty in taxonomic accuracy. To address this, we used the DNA standard during FL16S sequencing and performed PCR-based validation for key bacterial taxa to ensure reliable 16S rRNA data. While shotgun metagenomic sequencing offers greater precision, its clinical use is constrained by high costs, intensive data processing, and interpretation challenges. FL16S sequencing provides a practical alternative, offering sufficient resolution for clinical applications. Another limitation is the feature selection process in our random forest model, which used the entire dataset. Ideally, separate training and test sets would reduce bias, but given our small sample size, we applied tenfold cross-validation to maximize data utilization and model stability. This process reduces random bias and supports generalizability. In addition, our model construction involved a two-step feature selection process: selecting taxa ranked in the top 35 by relative abundance or with an LDA score > 3, and refining these features using the MDA metric to retain the top 20% most impactful taxa. This streamlined approach minimized feature complexity, reduced overfitting, and enhanced

model performance and generalizability. Future studies with larger datasets may explore incorporating independent test sets to complement our approach.

In conclusion, our study demonstrates the improved ability of FL16S sequencing over V3V4 sequencing to identify associations between gut microbiota and MASLD among obese children. These findings underscore the potential of FL16S sequencing for advancing clinical research and applications.

Abbreviations

ALT	Alanine aminotransferase
AST	Aspartate aminotransferase
ASV	Amplicon sequence variant
AUC	Area under the receiver operating characteristic curve
FL16S	Full-length 16S
GGT	Gamma-glutamyltransferase
HOMA-IR	Homeostasis model assessment of insulin resistance
LEFSe	Linear discriminant analysis effect size
LDA	Linear discriminant analysis
MASLD	Metabolic dysfunction-associated steatotic liver disease
MDA	Mean decrease accuracy
OTU	Operational taxonomic unit; ROC, receiver-operating characteristic curve
TG	Triglyceride
V3V4	V3-V4

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12866-025-03849-0.

Supplementary Material 1.

Acknowledgements
Special thanks to the Core Laboratory of Far Eastern Memorial Hospital, Taiwan, for providing support and equipment.

Authors' contributions
The authors' responsibilities were as follows — Y-CL and Y-HN designed research; Y-CL, C-CW, Y-EL, C-LC, and C-RL conducted research; Y-CL, C-CW, Y-EL, C-LC, and C-RL analyzed data; Y-CL and C-CW wrote the paper; Y-CL had primary responsibility for the final content of the manuscript. All authors read and approved the final manuscript.

Funding
This research was funded by the Council of National Science and Technology, Executive Yuan, Taiwan (MOST 107–2314-B-418–012-MY2, MOST 111–2314-B-418–010, NSTC 113–2314-B-075–081, and NSTC 112–2314-B-075–083-MY3).

Data availability
Sequence data that support the findings of this study have been deposited in SRA repository with the accession code PRJNA1126338.

Declarations

Ethics approval and consent to participate
Our study adhered to the Declaration of Helsinki. The study was approved by the ethics committee of Far Eastern Memorial Hospital, and all parents provided their informed consent.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 23 May 2024 Accepted: 26 February 2025
Published online: 17 March 2025

References

1. Lazarus JV, Mark HE, Anstee QM, Arab JP, Batterham RL, Castera L, et al. Advancing the global public health agenda for NAFLD: a consensus statement. Nat Rev Gastroenterol Hepatol. 2022;19(1):60–78.
2. Vos MB, Abrams SH, Barlow SE, Caprio S, Daniels SR, Kohli R, et al. NASPGHAN Clinical Practice Guideline for the Diagnosis and Treatment of Nonalcoholic Fatty Liver Disease in Children: Recommendations from the Expert Committee on NAFLD (ECON) and the North American Society of Pediatric Gastroenterology, Hepatology and Nutrition (NASPGHAN). J Pediatr Gastroenterol Nutr. 2017;64(2):319–34.
3. Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol. 2013;31(9):814–21.
4. Leung C, Rivera L, Furness JB, Angus PW. The role of the gut microbiota in NAFLD. Nat Rev Gastroenterol Hepatol. 2016;13(7):412–25.
5. Saltzman ET, Palacios T, Thomsen M, Vitetta L. Intestinal Microbiome Shifts, Dysbiosis, Inflammation, and Non-alcoholic Fatty Liver Disease. Front Microbiol. 2018;9:61.
6. Jayakumar S, Loomba R. Review article: emerging role of the gut microbiome in the progression of nonalcoholic fatty liver disease and potential therapeutic implications. Aliment Pharmacol Ther. 2019;50(2):144–58.
7. Lin YC, Lin HF, Wu CC, Chen CL, Ni YH. Pathogenic effects of Desulfovibrio in the gut on fatty liver in diet-induced obese mice and children with obesity. J Gastroenterol. 2022;57(11):913–25.
8. Aron-Wisnewsky J, Vigliotti C, Witjes J, Le P, Holleboom AG, Verheij J, et al. Gut microbiota and human NAFLD: disentangling microbial signatures from metabolic disorders. Nat Rev Gastroenterol Hepatol. 2020;17(5):279–97.
9. Leung H, Long X, Ni Y, Qian L, Nychas E, Siliceo SL, et al. Risk assessment with gut microbiome and metabolite markers in NAFLD development. Sci Transl Med. 2022;14(648):eabk0855.
10. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, et al. Experimental and analytical tools for studying the human microbiome. Nat Rev Genet. 2011;13(1):47–58.
11. Devanga Ragupathi NK, Muthuirulandi Sethuvel DP, Inbanathan FY, Veeraraghavan B. Accurate differentiation of Escherichia coli and Shigella serogroups: challenges and strategies. New Microbes New Infect. 2018;21:58–62.
12. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017;11(12):2639–43.
13. Fasolo A, Deb S, Stevanato P, Concheri G, Squartini A. ASV vs OTUs clustering: Effects on alpha, beta, and gamma diversities in microbiome metabarcoding studies. PLoS ONE. 2024;19(10): e0309065.
14. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 2019;10(1):5029.
15. Matsuo Y, Komiya S, Yasumizu Y, Yasuoka Y, Mizushima K, Takagi T, et al. Full-length 16S rRNA gene amplicon analysis of human gut microbiota using MinION nanopore sequencing confers species-level resolution. BMC Microbiol. 2021;21(1):35.
16. Singer E, Bushnell B, Coleman-Derr D, Bowman B, Bowers RM, Levy A, et al. High-resolution phylogenetic microbial community profiling. ISME J. 2016;10(8):2020–32.
17. Lin YC, Chang PF, Lin HF, Liu K, Chang MH, Ni YH. Variants in the autophagy-related gene IRGM confer susceptibility to non-alcoholic fatty liver disease by modulating lipophagy. J Hepatol. 2016;65(6):1209–16.
18. Chen W, Chang MH. New growth charts for Taiwanese children and adolescents based on World Health Organization standards and health-related physical fitness. Pediatr Neonatol. 2010;51(2):69–79.
19. European Society for Pediatric Gastroenterology H, Nutrition, Vajro P, European Association for the Study of the L, North American Society for Pediatric Gastroenterology H, Nutrition, et al. Paediatric steatotic liver disease has unique characteristics: A multisociety statement endorsing the new nomenclature. J Pediatr Gastroenterol Nutr. 2024;78(5):1190–6.
20. Odamaki T, Kato K, Sugahara H, Hashikura N, Takahashi S, Xiao JZ, et al. Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. BMC Microbiol. 2016;16:90.

21. Valeri F, Endres K. How biological sex of the host shapes its gut microbiota. Front Neuroendocrinol. 2021;61: 100912.
22. Xu Z, Jiang W, Huang W, Lin Y, Chan FKL, Ng SC. Gut microbiota in patients with obesity and metabolic disorders - a systematic review. Genes Nutr. 2022;17(1):2.
23. Martin M. Cutadapt removes adapter sequences from high throughput sequencing reads. EMBnet J. 2011;17(1):3.
24. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13(7):581–3.
25. Quin C, Estaki M, Vollman DM, Barnett JA, Gill SK, Gibson DL. Probiotic supplementation and associated infant gut microbiome and health: a cautionary retrospective clinical comparison. Sci Rep. 2018;8(1):8283.
26. 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(1):90.
27. 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(8):852–7.
28. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biol. 2011;12(6):R60.
29 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–45.
30. Han. H, Guo. X, Yu. H. Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest. 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, 2016: 219–224.
31. Sharpton SR, Schnabl B, Knight R, Loomba R. Current Concepts, Opportunities, and Challenges of Gut Microbiome-Based Personalized Medicine in Nonalcoholic Fatty Liver Disease. Cell Metab. 2021;33(1):21–32.
32. Buetas E, Jordan-Lopez M, Lopez-Roldan A, D'Auria G, Martinez-Priego L, De Marco G, et al. Full-length 16S rRNA gene sequencing by PacBio improves taxonomic resolution in human microbiome samples. BMC Genomics. 2024;25(1):310.
33. De Minicis S, Rychlicki C, Agostinelli L, Saccomanno S, Candelaresi C, Trozzi L, et al. Dysbiosis contributes to fibrogenesis in the course of chronic liver injury in mice. Hepatology. 2014;59(5):1738–49.
34. Boursier J, Mueller O, Barret M, Machado M, Fizanne L, Araujo-Perez F, et al. The severity of nonalcoholic fatty liver disease is associated with gut dysbiosis and shift in the metabolic function of the gut microbiota. Hepatology. 2016;63(3):764–75.
35. Kim C, Pongpanich M, Porntaveetus T. Unraveling metagenomics through long-read sequencing: a comprehensive review. J Transl Med. 2024;22(1):111.
36. Iino C, Endo T, Mikami K, Hasegawa T, Kimura M, Sawada N, et al. Significant decrease in Faecalibacterium among gut microbiota in nonalcoholic fatty liver disease: a large BMI- and sex-matched population study. Hepatol Int. 2019;13(6):748–56.
37. Camilla. C, Marco. S. A comparison between Greengenes, SILVA, RDP, and NCBI reference databases in four published microbiota datasets. bioRxiv. 2023. https://doi.org/10.1101/2023.04.12.535864.

Publisher's Note