

<https://doi.org/10.1038/s41522-025-00686-x>

Refining microbiome diversity analysis by concatenating and integrating dual 16S rRNA amplicon reads

Check for updates

Kyoung Su Kim¹, Jihye Noh², Bong-Soo Kim³, Hong Koh²✉ & Dong-Woo Lee¹✉

Understanding the role of human gut microbiota in health and disease requires insights into its taxonomic composition and functional capabilities. This study evaluates whether concatenating paired-end reads enhances data output for gut microbiome analysis compared to the merging approach across various regions of the 16S rRNA gene. We assessed this approach in both mock communities and Korean cohorts with or without ulcerative colitis. Our results indicate that using the direct joining method for the V1-V3 or V6-V8 regions improves taxonomic resolution compared to merging paired-end reads (ME) in post-sequencing data. While predicting microbial function based on 16S rRNA sequencing has inherent limitations, integrating sequencing reads from both the V1-V3 and V6-V8 regions enhanced functional predictions. This was confirmed by whole metagenome sequencing (WMS) of Korean cohorts, where our approach improved taxa detection that was lost using the ME method. Thus, we propose that the integrated dual 16S rRNA sequencing technique serves as a valuable tool for microbiome research by bridging the gap between amplicon sequencing and WMS.

Recent explorations into the human gut microbiome have captured widespread interest due to its complex composition, functional capabilities, and significant influence on human health and disease states^{1,2}. The surge in research activity is largely attributed to advancements in next-generation sequencing (NGS) technologies, which have transformed our ability to discern gut microbiota variances associated with a broad range of diseases such as cancer³, obesity⁴, diabetes⁵, inflammatory bowel diseases (IBD)^{6,7}, neurological disorders⁸, and antibiotics resistance^{9,10}. These technological advances have enabled large-scale population studies, providing deeper insights into the epidemiology of infectious diseases¹¹ and facilitating the analysis of extensive microbiome datasets^{12,13}.

Predominantly, 16S rRNA amplicon sequencing and whole metagenome sequencing (WMS) are pivotal in unraveling gut microorganism diversity and exploring the epidemiological factors that influence microbiome configurations^{14,15}. These methods have greatly advanced our understanding of the dynamics that shape the human gut microbiome, encompassing microbial taxa, epidemiological impacts, evolutionary patterns, and demographic variables such as ethnicity, environmental conditions, dietary habits, and age¹⁶⁻¹⁸. However, gut microbiome studies often face challenges due to inherent experimental biases. Such biases in taxonomic identification may stem from the choice of taxonomic marker genes

(e.g., 16S rRNA for bacteria, 18S rRNA for eukaryotes, and ITS regions for fungi) and their target regions^{19,20}, diversity in sequencing platforms^{21,22}, inconsistencies in data quality²³, and variations in reference databases²⁴. For example, the selection of the 16S rRNA regions critically affects the resolution and the precision in bacterial detection and classification²⁵, leading to discrepancies in estimating the presence of certain bacterial groups^{26,27}. Notably, V4-V5 region should be avoided in the infant feces²⁸, whereas the V1-V3 region is recommended for soil and saliva samples²⁹. Utilizing the full read length (V1-V9 region) is also recommended to reduce sequencing error rates³⁰.

Both 16S rRNA sequencing and WMS have their unique benefits and face distinct challenges. WMS provides in-depth insights into microbial communities and functional data but requires substantial computational resources and ongoing reference database updates³¹⁻³³. It also deals with challenges, such as host DNA depletion and variability in 16S rRNA primer coverage³⁴⁻³⁶. In contrast, 16S rRNA sequencing is a cost-effective and efficient alternative for specific applications, particularly when using methodologies that minimize inherent biases³⁷. Our study compares analytical methodologies within 16S rRNA sequencing, focusing on merging paired-end reads (ME) and direct joining (DJ). These methods aim to broaden the range of captured microbial data and reduce biases associated

¹Department of Biotechnology, Yonsei University, Seoul, South Korea. ²Department of Pediatrics, Yonsei University College of Medicine, Severance Fecal Microbiota Transplantation Center, Severance Hospital, Seoul, South Korea. ³Department of Nutritional Science and Food Management, Ewha Womans University, Seoul, South Korea. ✉e-mail: khong@yuhs.ac; leehicam@yonsei.ac.kr

with merging methods. ME merges reads based on overlapping sequences, potentially losing valuable genetic information when overlaps are minimal. DJ, however, concatenate forward and reverse reads directly, retaining all genetic information and enhancing the dataset completeness—essential for accurately depicting microbial communities^{38,39}.

We compare the quality of sequencing data between concatenated and merged reads, focusing on sequencing errors and the impact of different 16S rRNA regions on identifying rare microbial taxa in diverse cohorts, including healthy individuals and patients with ulcerative colitis (UC). Using correction formulas derived from mock community datasets²⁴, we have refined taxonomic classifications precision, aiding in the identification of unique metabolic pathways associated with health and UC. Through comparative functional profiling with multiple analytical pipelines based on 16S rRNA sequencing and WMS, we seek potential diagnostic markers and therapeutic targets. This comprehensive approach elucidates the role of the

gut microbiome in health and disease, utilizing dual 16S rRNA amplicon sequencing to improve clarity and specificity, advancing our understanding of microbial ecosystems and promoting targeted interventions that could profoundly affect patient care and therapeutic outcomes.

Results

Comparative validation of concatenation and merging methods for gut microbiome analysis

Our research assessed the effectiveness of concatenating versus merging pair-end reads across various 16S rRNA regions, using the ZIEL-II mock community datasets (SRP291583), which includes 19 bacteria across 18 genera. We applied ME and DJ alongside inside-out (IO) concatenation techniques (Fig. 1). Observations revealed a decline in sequence quality towards the 3'-end across all regions (Supplementary Fig. 1), with concatenation generally achieving better alignment of non-chimeric reads with

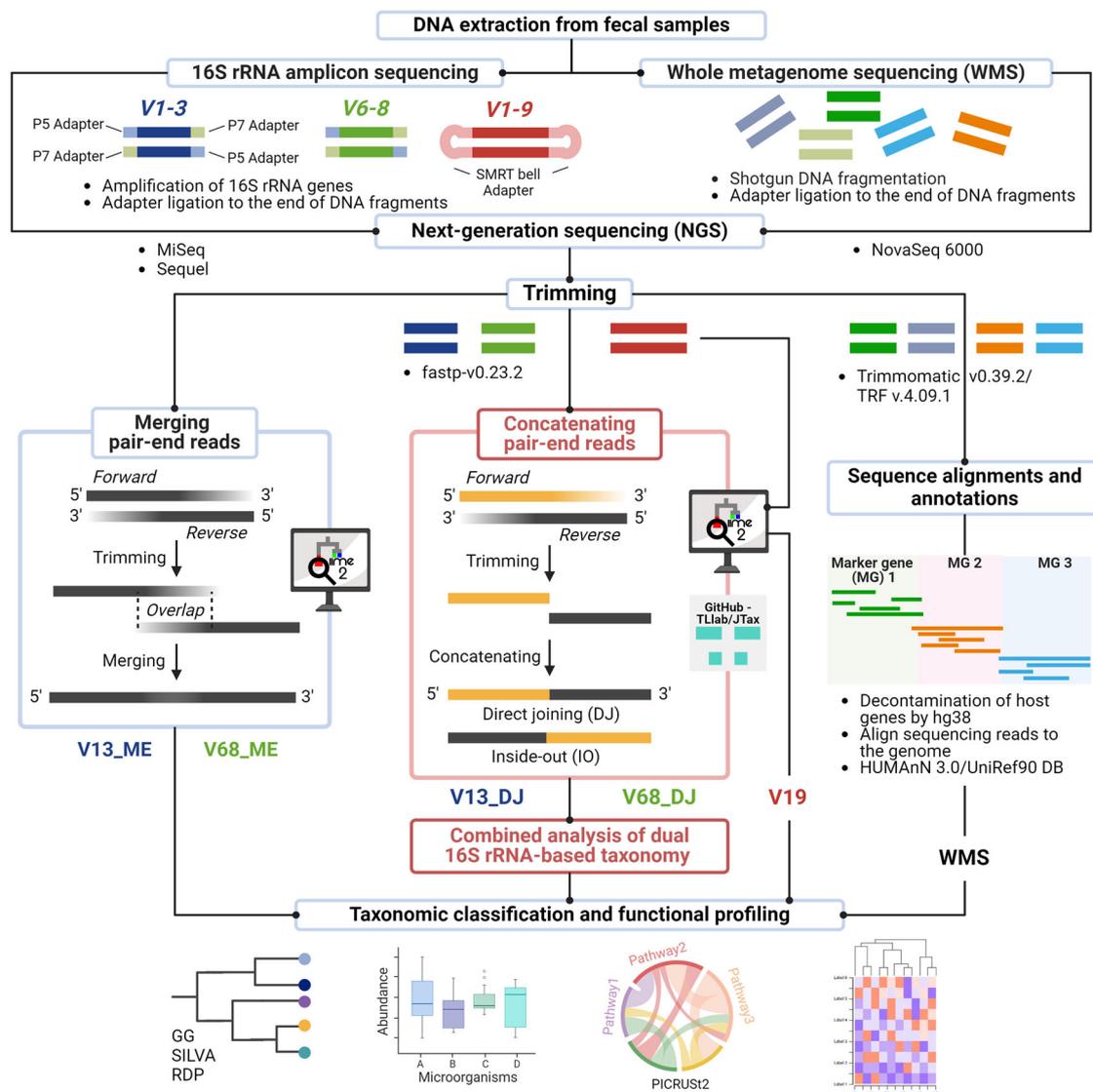


Fig. 1 | Analytical strategy for region-specific 16S rRNA amplicon sequencing and whole metagenome sequencing (WMS). This workflow outlines the analysis of amplified 16S rRNA sequences (V1-V3, V6-V8, and V1-V9). The steps include: 1) Adapter trimming with fastp; 2) Merging paired-end sequences for regions excluding V1-V9 using DADA2 in QIIME2, while V1-V9 single-end sequences undergo separate analysis; 3) Concatenating paired-end sequences through JTax using both direct joining (DJ) and inside-out (IO) techniques, with lengths trimmed based on median quality score 20 (Detailed trim positions in Supplementary Fig. 1);

4) Subjecting all region-derived sequences—merged, concatenated, or intact (V1-V9)—to quality filtering, denoising, and chimera removal via DADA2; 5) Classifying amplicon sequence variants (ASVs) produced across three analytical pipelines against 16S rRNA DBs (GG2, SILVA, and RDP); 6) Conduct functional profiling via PICRUSt2. For WMS, data processing ranged from shallow (4 GB) to deep (36 GB) sequencing reads, using Trimmomatic and TRF, aligned against hg38 with default settings, and analyzed for taxonomic and functional profiling with HUMAnN 3.0. This figure was created using Biorender.com.

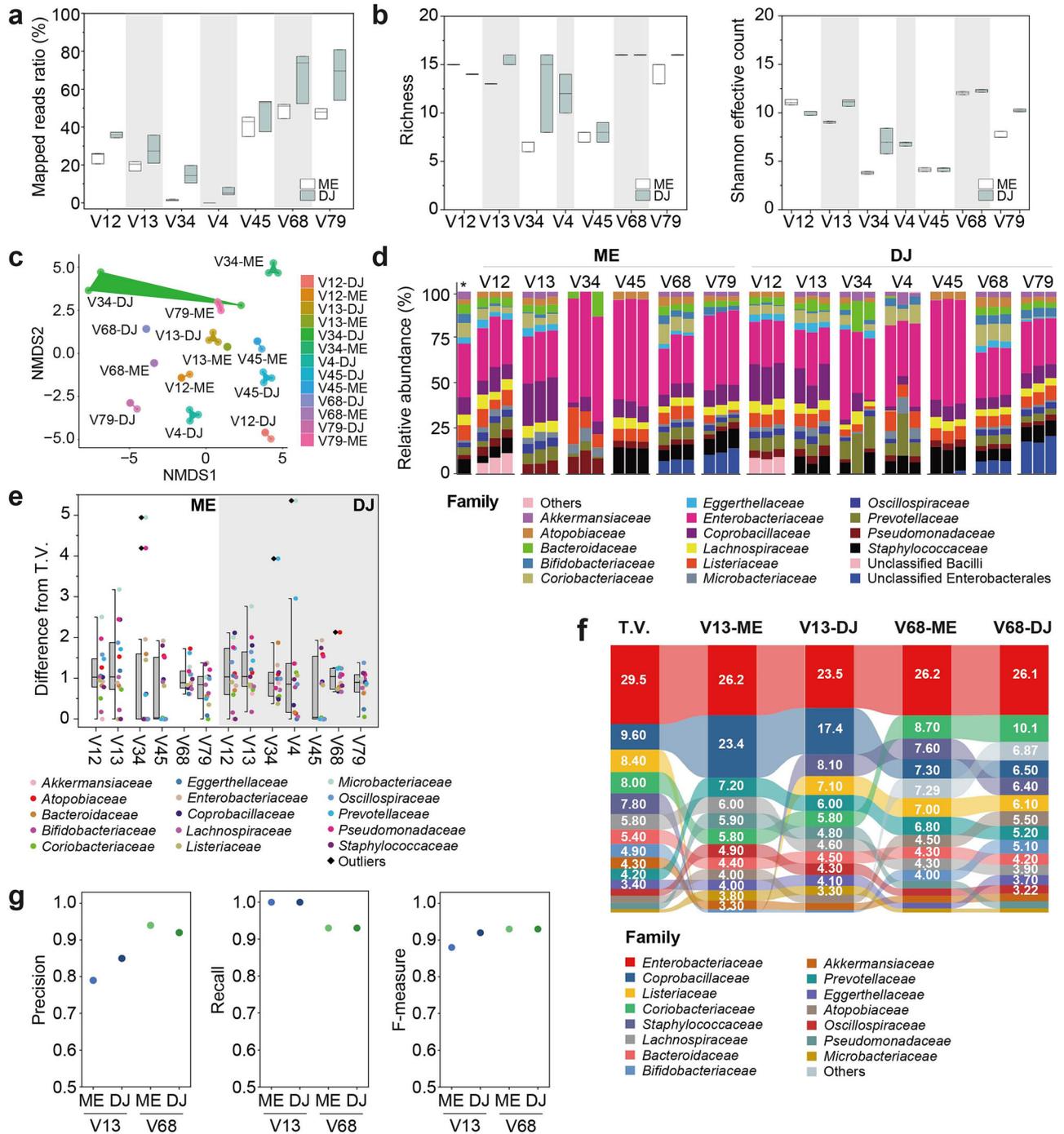


Fig. 2 | Validation of taxonomic resolution and richness of 16S rRNA sequencing reads using mock community datasets. **a** The comparison of the mapped reads ratio from the SRP291583 dataset. ME: the method merging raw paired-end reads, DJ: concatenating raw paired-end reads. **b** Alpha diversity metrics across methods, including observed species and Shannon index. **c** NMDS plot based on Jaccard distance across the analytical methods. **d** Family-level relative abundance, with “Others” indicating taxa

below 1% relative abundance. The relative abundance for the theoretical composition of ZIEL-II mock community²⁴ is indicated by an asterisk (*). **e** The distribution of difference values between theoretical and actual values for each family. **f** Comparison of relative abundance between theoretical values (T.V.) and specific 16S rRNA regions from ZIEL-2 mock datasets. **g** Evaluation of precision, recall, and F-measure for the ME and DJ pipelines using the V1-V3 and V6-V8 regions, based on the SILVA DB.

the SILVA database across all tested 16S rRNA regions (Fig. 2a and Supplementary Data 1).

Concatenation using the DJ method notably enhanced microbial diversity and evenness, evidenced by higher Richness and Shannon effective numbers compared to the ME method, particularly in the V1-V3, V3-V4 and V7-V9 regions (Fig. 2b). Non-metric multidimensional scaling (NMDS) suggested that adjacent regions within the 16S rRNA gene exhibit similar microbial communities, with significant differences in the V34 and

V79 regions (Fig. 2c), highlighting each region’s distinct response to the concatenation and merging techniques. The ME method particularly overestimated *Enterobacteriaceae* abundance in the V3-V4 (1.95-fold) and V4-V5 (1.92-fold) regions—discrepancies largely corrected by the DJ method, though not entirely in the V4-V5 region (Fig. 2d).

To substantiate the performance benefits of concatenation over the merging approach, we conducted a detailed correlation analysis by comparing the theoretical and actual measured relative abundances across

different region-specific methods, excluding the less reliable V1-V2 and V4-V5 regions (Supplementary Fig. 2). The DJ method improved the detection accuracy of several microbial families not identified by the ME method in the V3-V4 region, though it continued to face challenges with overestimations in the V4-V5 region (Fig. 2e). Detailed comparisons to theoretical values (TVs) revealed that specific DJ methods, especially V13-DJ (1.021) and V68-DJ (1.023), achieved median values closest to the ideal of 1.0 (Fig. 2e). However, there were notable discrepancies in other regions, indicating some inherent limitations in family-specific detection accuracy. For instance, V13-DJ significantly underdetected *Bifidobacteriaceae* (0.17), and V68-DJ overestimated *Atopobiaceae* (2.13) (Fig. 2e). The results showed that both V68-DJ and V13-DJ methods provided a more accurate and consistent representation of microbial abundances, enhancing the quality of taxonomic and functional insights derived from gut microbiome analyses (Supplementary Fig. 2).

Additional issues were observed in the V6-V8 and V7-V9 regions, where unclassified Enterobacterales were detected. Moreover, the V34-ME analysis demonstrated poor performance with significant outliers such as *Microbacteriaceae* and *Pseudomonadaceae*. V34-DJ (0.88) improved but still presented outliers, such as *Prevotellaceae* (Fig. 2e). In addition, V4-DJ still presented one outlier, such as *Microbacteriaceae* and V45-DJ still did not detect several families. Given these findings, we excluded V1-V2, V4, V4-V5, and V7-V9 from further analysis due to their lower correlation values (<0.66) and the presence of outliers or undetected families, which could skew the gut microbiome analysis.

In conclusion, our results indicate the importance of selecting suitable 16S rRNA regions for analysis, advocating for the use of V1-V3 and V6-V8 regions when employing concatenating methods with the SILVA database (DB) to increase accuracy and reduce biases in analyzing the gut microbial community (Fig. 2f, g). This approach highlighted that the V1-V3 region consistently achieved higher recall values than the V6-V8 region. The ME method exhibited the lowest F-measure values, with significant discrepancies observed in the detection of families like *Listeriaceae*, *Bifidobacteriaceae*, and *Eggerthellaceae* (Fig. 2g). Remarkably, *Coprobacillaceae* detection was excessively high in the V13-ME method (23.4%), compared to the ideal (9.60%) and V13-DJ (17.4%) approaches (Fig. 2d). The V13-DJ method notably increased precision by 8% and the F-measure value by 5% relative to the V13-ME method. Despite challenges in estimating relative abundance, the V6-V8 region demonstrated superior precision in amplifying gut microbial 16S rRNA genes, underscoring the crucial role of method selection in microbiome analysis (Fig. 2g).

Optimizing accuracy in gut microbiome analysis: the role of concatenated method and selection of 16S rRNA databases

The accuracy of estimating microbial relative abundance critically depends on the choice of 16S rRNA gene regions and read processing methodologies. We conducted an in-depth analysis using mock community data, focusing on the V1-V3 and V6-V8 regions, and calibrated coefficient values for specific family groups, selecting the most appropriate 16S rRNA DBs: Greengenes2 (GG2), SILVA, and the Ribosomal Database Project (RDP). Sequences were trimmed and processed for database matching using either the merging or concatenating method using V1-V3 and V6-V8 regions (Supplementary Fig. 3).

Our findings showed that the V13-ME method consistently overestimates relative abundance, particularly inflating families like *Enterobacteriaceae_A* and *Pseudomonadaceae* up to 93% relative to their expected values (24.7%) in the Zymo mock dataset. In contrast, concatenating methods—DJ and IO—yielded more accurate estimations at 22.0% and 24.1%, respectively (Supplementary Fig. 4a). Further assessments across different databases and primer sets revealed that the ME method consistently displayed the lowest correlation coefficients (R-values), particularly in the ZIEL-I mock dataset with the lowest R-values linked to the GG2 database (Supplementary Figs. 5–7). The ME method exhibited biases in the V1-V3 16S rRNA region, notably underrepresenting families such as *Bacillaceae* (1.1%), *Enterococcaceae* (0.3%), *Lachnospiraceae* (0.0%), and

Staphylococcaceae (0.8%) (Supplementary Figs. 4a and 6a). Conversely, concatenation approaches, particularly using SILVA and RDP databases, markedly improved accuracy over the ME method. While the ME method achieved the highest R-value with the V6-V8 region and SILVA database in the ZIEL-II dataset, it faced challenges, particularly with the GG2 database (Supplementary Fig. 7). Notably, updates to family names like *Eggerthellaceae*, *Erysipelatoclostridiaceae*, and *Verrucomicrobiaceae* were observed, and the ME method's failure to detect *Listeria welshimeri* within the V1-V3 region contrasted with the successful identification of *Listeriaceae* by the concatenating methods (Supplementary Fig. 4).

Comparative efficacy of concatenation versus merging in gut microbiome analysis

In our analysis, we evaluated the effectiveness of ME and concatenating methods using two significant datasets: SRP131748, which includes 60 samples from fecal and oronasal secretion, and SRP115494, containing 69 rectal samples, serving as supplementary data (Supplementary Data 2)^{7,40}. We investigated the performance of concatenating methods (DJ and IO) over ME in the primary dataset (SRP131748), targeting the V1-V3 region. These methods achieved significantly higher non-chimeric read alignment with 16S rRNA databases, marking 58.64% for DJ and 59.12% for IO, compared to only 47.42% for ME, demonstrating their enhanced efficiency in SRP131748 ($P < 0.05$) (Supplementary Data 3). In addition, in the SRP115494 dataset targeting the V4 region, the concatenation methods marked 66.05% for DJ and 68.17% for IO, compared to only 42.12% for ME, underlining their improved detection capabilities ($P < 0.05$).

Further investigations within SRP131748 demonstrated that DJ and IO methods not only provided higher alpha diversity but also depicted more distinct microbial profiles than ME (Supplementary Figs. 8a, 9a–d, and Supplementary Data 3). A notable finding was the detection of an *Enterococcus* strain in the oronasal secretion-prediabetes group (disease) by the concatenating methods, which was absent in the control group (healthy) analyzed by ME (Supplementary Fig. 9e, f). Additionally, *Phascolarctobacterium*, previously undetectable by ME, was significantly identified with the DJ and IO methods, highlighting their increased detection capabilities (Supplementary Fig. 10).

The supplementary dataset, SRP115494, targeting the V4 region, also indicated a decline in sequence quality towards the 3'-end (Supplementary Fig. 8b). Like findings in SRP131748, we reaffirmed the consistency of microbiome diversity between concatenating methods (DJ and IO) and ME across different sample types and conditions (Supplementary Fig. 11a and Supplementary Data 3). Microbial alpha diversity analysis revealed distinct differences across patient groups. IBD patients, including those with Crohn's disease (CD) and UC, showed lower microbial diversity than non-IBD controls, with ME showing the least diversity (Supplementary Figs. 11b). At the family level, the DJ method enriched taxa such as *Veillonellaceae*, *Erysipelotrichales*, *Pseudomonadales*, and *Staphylococcaceae*, contrasting with ME which predominantly identified *Lachnospiraceae_NK4A136_g* (Supplementary Figs. 11c–e). Moreover, taxa like *Oscillospirales*, *[Eubacterium] eligens_g*, and *Rombutsia* were detected only by the concatenating methods (Supplementary Fig. 12a, b).

In conclusion, the concatenation approach has demonstrated its potential to enhance microbial community diversity analysis, enabling a more comprehensive identification of specific microbial families and genera across both IBD and non-IBD cohorts (Supplementary Fig. 10). This method effectively might improve the resolution and accuracy of gut microbiota analysis, bridging the gap between traditional sequencing approaches and the nuanced demands of modern microbiome research.

Impact of 16S rRNA region selection on taxonomic assignment in gut microbiota

To evaluate the effect of different 16S rRNA regions on taxonomic assignments, we analyzed fecal samples from a Korean cohort, comprising healthy individuals ($n = 8$) and patients with UC ($n = 8$) (Supplementary Data 4).

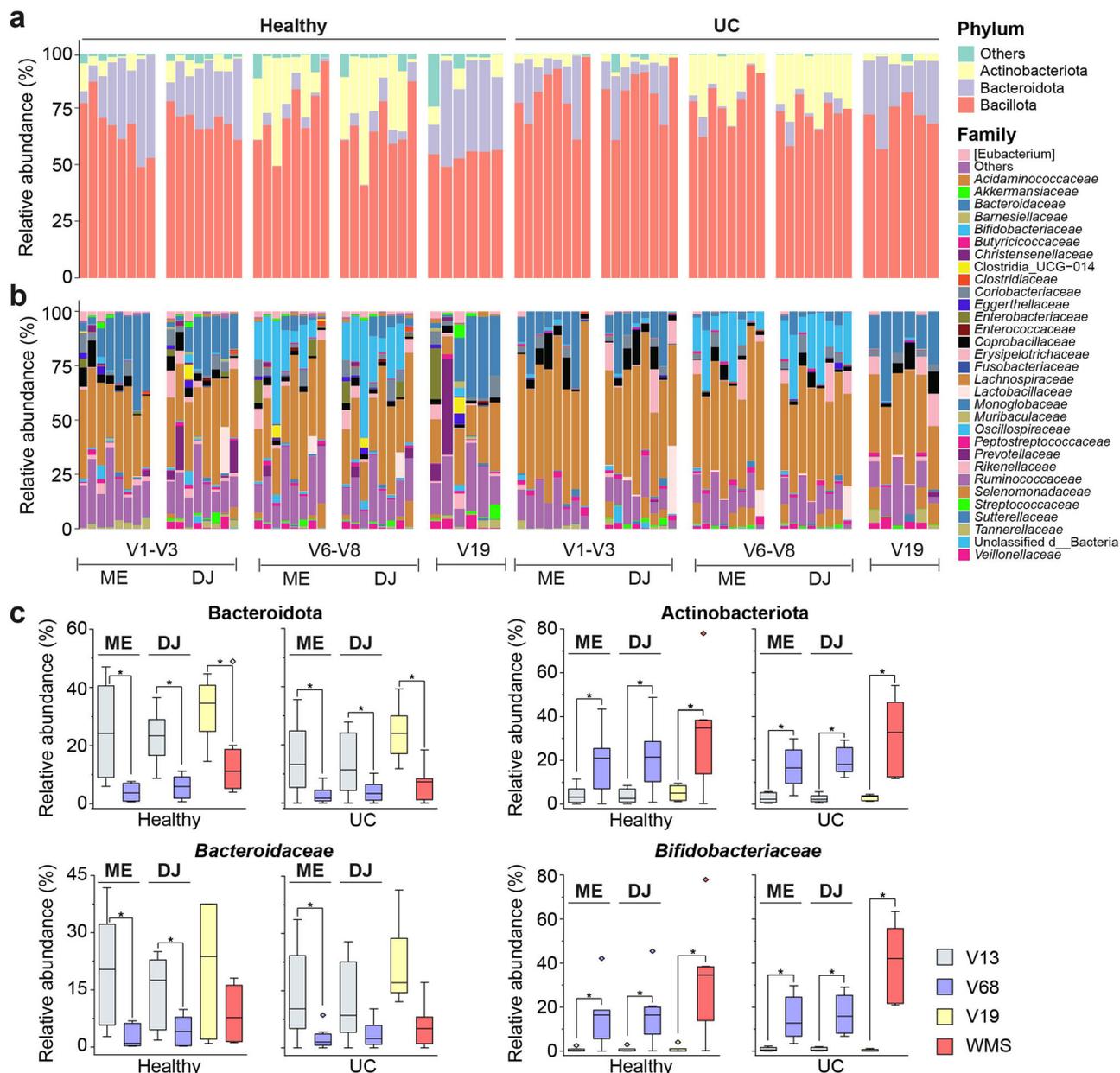


Fig. 3 | Comparison of richness and taxa classification between analytical methods. a Phylum-level relative abundance comparison. **b** Family-level relative abundance, with “Others” denoting taxa below 1% relative abundance.

c Comparative relative abundances of Bacteroidota, *Bacteroidaceae*, Actinobacteriota, and *Bifidobacteriaceae* across ME and DJ methods. Asterisk (*) means a *P* value < 0.05.

We employed primer pairs targeting V1-V3, V6-V8, and V1-V9 regions, analyzing sequence quality and read integrity across these regions (Supplementary Fig. 13a). Quality analysis indicated sequence deterioration towards the 3'-end in both forward and reverse reads, which could introduce biases, particularly in the reverse (Supplementary Fig. 13b-d).

The concatenation methods (DJ and IO) showed increased non-chimeric reads compared to the ME method, particularly in the V1-V3 and V6-V8 regions, demonstrating their effectiveness in managing sequence quality impacts (Supplementary Fig. 14a, b). Differences in non-chimeric reads between DJ and IO were not significant, nor were differences in richness and taxonomic resolution from previous datasets analyses (Supplementary Fig. 9). Alpha diversity analyses further supported the increased sensitivity of the DJ method in the V1-V3 region over ME (Supplementary Fig. 14c). Significant variations were observed in the detection of Bacteroidota and Actinobacteriota between in healthy and UC samples across the studied regions. The V1-V3 region showed less variability in detecting

Actinobacteriota, whereas the V6-V8 region was more consistent for Bacteroidota (Fig. 3a-c). At the family level, the V1-V3 and V1-V9 regions more consistently identified *Bacteroidaceae*, whereas the V6-V8 region was more sensitive to *Bifidobacteriaceae*, irrespective of the method used (Fig. 3b, c).

Genus-level heat tree visualizations between healthy and UC groups highlighted the influence of 16S rRNA region selection on taxonomic assignments, showing the DJ method's ability to reduce bias for families like *Akkermansiaceae* and *Lactobacillaceae* in the V1-V3 region (Fig. 4a, b). However, biases persisted for families such as *Clostridiaceae* and *Bifidobacteriaceae* based on the V6-V8 region. The V1-V3 region analysis also enriched *Enterobacteriaceae* and Bacteroidales, including families like *Bacteroidaceae*, *Rikenellaceae*, and *Marinifilaceae*, showing a clear contrast to the V6-V8 region.

Comparative taxonomic resolution with the WMS method revealed significant discrepancies in the relative abundance of Bacteroidota,

Actinobacteriota, and *Bifidobacteriaceae* between V1-V3, V1-V9, and WMS methods for both healthy individuals and UC patients (Fig. 3c). Eleven families detected by WMS were missed by the V13-ME method in the healthy group, whereas V13-DJ identified nine families (e.g., *Barnesiellaceae*, *Lactobacillaceae*, *Streptococcaceae*, and *Sutterellaceae*) also seen in WMS (Fig. 4c). However, *Barnesiellaceae*, *Lactobacillaceae* and *Sutterellaceae*, detected in WMS, were absent in both V68-ME and V68-DJ analyses. In the UC group, families like *Eggerthellaceae*, *Lactobacillaceae*, *Streptococcaceae*, *Oscillospiraceae*, and *Selenomonadaceae*, missed by V13-ME, were identified by V13-DJ. Notably, V68-DJ uniquely detected *Oscillospiraceae*, absent in WMS. Analysis using V19 indicated overestimated abundances of *Erysipelotrichaceae*, *Veillonellaceae*, and *Selenomonadaceae* compared to other methods (Fig. 4d). While concatenation-based methods demonstrated superior microbial detection relative to merging methods when validated with WMS, they still exhibited biased taxonomic resolution.

In conclusion, our findings underscore the critical importance of selecting appropriate 16S rRNA regions and analytical methods to better represent gut microbial diversity and minimize taxonomic biases. Despite the advantages of the concatenating method, relying solely on one 16S rRNA region may still result in biased outcomes, emphasizing the need for comprehensive methodological approaches in microbiome research.

Advancing gut microbiota profiling accuracy with correction coefficient-based adjustments for dual 16S rRNA reads

To improve the precision of gut microbiota analysis, we developed a methodology based on correction coefficients that integrates reads from both the V1-V3 and V6-V8 16S rRNA regions. We introduced the adjusted 16S rRNA (Adj-16S) sequencing method, applying correction coefficients based on analyses of dual 16S rRNA regions to more accurately adjust relative abundance values (Fig. 5a). For instance, the correction coefficient for *Enterobacteriaceae* using V13-DJ ($\omega_i^{13.En}$) was calculated as 1.26 by comparing the theoretical value (29.5) with the observed value (23.5), and similar for V68-DJ ($\omega_i^{68.En}$) (Fig. 5b). These adjustments provided a more accurate representation of taxonomic profiles. Utilizing mock datasets (Zymo, ZIEL-I, ZIEL-II) and the SILVA DB, we calculated correction coefficients for each 16S rRNA region, using weighted averages⁴¹ across 22 families. Weighted coefficient values for the V1-V3 ($\omega_i^{13.f}$) and V6-V8 ($\omega_i^{68.f}$) regions were determined by dividing the relative abundance from the respective regions, V1-V3 ($x_i^{13.f}$) and V6-V8 ($x_i^{68.f}$), by the total abundance from both regions. The means of these weighted coefficient values ($\bar{\omega}_i^{13.f}$ and $\bar{\omega}_i^{68.f}$) were computed using data from eight independent datasets across three mock datasets. The adjusted relative abundances for both the V1-V3 ($x_i^{13.f}$) and V6-V8 ($x_i^{68.f}$) regions were then calculated, leading to a formula representing the total adjusted relative abundance $\approx \sum_{i=1}^n (x_i^{13.f} + x_i^{68.f})$. These adjustments showed that the V6-V8 region more accurately reflected ideal community compositions, particularly for families like *Actinomycetaceae*, *Bifidobacteriaceae*, and *Tannerellaceae*. Conversely, the V1-V3 region was more responsive to *Coprobacillaceae*, *Microbacteriaceae*, and *Pseudomonadaceae* (Fig. 5c).

Based on these findings, we recommend adopting the Adj-16S for more precise profiling of 16 Korean gut microbial communities (Fig. 5d, e and Supplementary Data 5). This approach showed that *Bacteroidaceae* was more prevalent in the V1-V3 region (16.44%) compared to the V6-V8 region (3.50%). In the UC cohort, *Bacteroidaceae* levels were 8.91% in V1-V3 and only 2.27% in V6-V8. However, the Adj-16S revealed relative abundances of *Bacteroidaceae* at 9.06% for healthy individuals and 5.22% for UC patients, values that closely align with those obtained from WMS, which were 7.74% for healthy individuals and 5.08% for UC patients. Similarly, *Bifidobacteriaceae* was more abundant in the V6-V8 region (20.97%) than in the V1-V3 region (0.29%) among healthy individuals. Applying the calculated coefficients to balance discrepancies between the two regions resulted in a uniform representation of abundances, with improved concordance evidenced by correlation metrics at the family and genus levels compared to WMS data (Supplementary Fig. 15). Furthermore, the Adj-16S method detected 21 families, including *Butyrivibrionaceae*,

Clostridia_UCG-014, *Coprobacillaceae*, and *Monoglobaceae*, which were not identified in WMS analyses (Supplementary Data 5).

The Adj-16S method significantly delineated the microbial differences between healthy individuals and UC patients, revealing disparities in the detection of families like *Odoribacteraceae* and *Bacillota_unclassified* that were pronounced in WMS but not in 16S rRNA analyses (Fig. 5f). Families such as [Eubacterium] and *Rikenellaceae* distinctly categorized healthy from UC groups. *Oscillospiraceae* and *Akkermansiaceae*, predominantly found in the healthy cohort, illustrate the nuanced capability of concatenated 16S rRNA methods alongside WMS in detecting critical microbial differences. *Mariniflavaceae*, *Ruminococcaceae*, and *Anaerovoracaceae* were only significantly detected in the merging methods. These findings underscore the necessity of methodological precision in 16S rRNA-based profiling, affirming the concatenated approach for its improved accuracy and consistency in representing microbial abundances, closely aligned with the comprehensive insights provided by WMS analyses.

Comparative functional profiling in gut microbiota: insights from Adj-16S and WMS analysis

To further explore the functional capabilities of the gut microbiota, we conducted a comparative analysis using the Adj-16S method alongside traditional 16S rRNA amplicon sequencing techniques and WMS (Fig. 1). For the 16S rRNA data, predictive functional profiling was performed using PICRUSt2, while the WMS data were analyzed using HUMAnN 3.0. Our study identified several key functional pathways that were significantly different (P value < 0.05) between healthy individuals and UC patients (Fig. 6a). The V13-ME method identified 28 pathways in healthy subjects and 35 in UC patients, numbers which were at least twice those identified by other analytical methods. Conversely, the Adj-16S method pinpointed fewer pathways—14 in healthy subjects and 11 in UC patients. The WMS approach uniquely detected 12 pathways not identified by any 16S-based methods, with some pathways found to be common across 16S rRNA methods and WMS (Supplementary Fig. 16a, b). A direct comparison revealed that five pathways were common between Adj-16S and WMS analyses, indicating 25 unique pathways in WMS and 20 unique to Adj-16S across both subject groups.

Further validation through quantitative real-time PCR (qRT-PCR) analysis of 50 genes representing these pathways confirmed 18 significantly divergent pathways between the groups, with 12 additional pathways differing in the reverse comparison (Fig. 6b and Supplementary Data 7). This validation underlined the predictive accuracy of our methods, with the concatenation-based approach using V13-DJ and V68-DJ demonstrating relatively higher accuracy and F1 scores compared to merged methods (Fig. 6c). The false positive rate (FPR) for Adj-16S was 0.36, showcasing its precision relative to 0.53 for V13-DJ (FRR: 0.53) and V68-DJ (FPR: 0.38). However, the V19 method, despite having the lowest FPR, demonstrated limitations in detecting a higher number of true positives (TPs) or true negatives (TNs) (Fig. 6c, d).

A Venn diagram analysis emphasized the efficacy of the Adj-16S method in detecting the highest number (12) of TPs and TNs pathways compared to other techniques (Fig. 6e). The Adj-16S method detected all pathways except for SULFATE-CYS-PWY, uniquely identified by the V13-DJ method. While the V13-DJ method missed three pathways (PYR-IDOXYN-PWY, P23-PWY, and PWY-3781), the V68-DJ method failed to detect five (PWY-5855, PWY-7456, PWY-6467, PWY-6590, and SULFATE-CYS-PWY). Notably, the Adj-16S method closely aligns with other methods in capturing nearly all TN pathways for UC.

Collectively, our findings emphasize the importance of selecting appropriate 16S rRNA regions and employing concatenating methods to enhance accuracy and reduce biases in the functional profiling of gut microbiota. This study not only clarifies the differences between functional profiles derived from Adj-16S compared to the ME and DJ methods but also highlights the potential for discovering unique biomarkers or therapeutic targets within these methodologies (Fig. 6e).

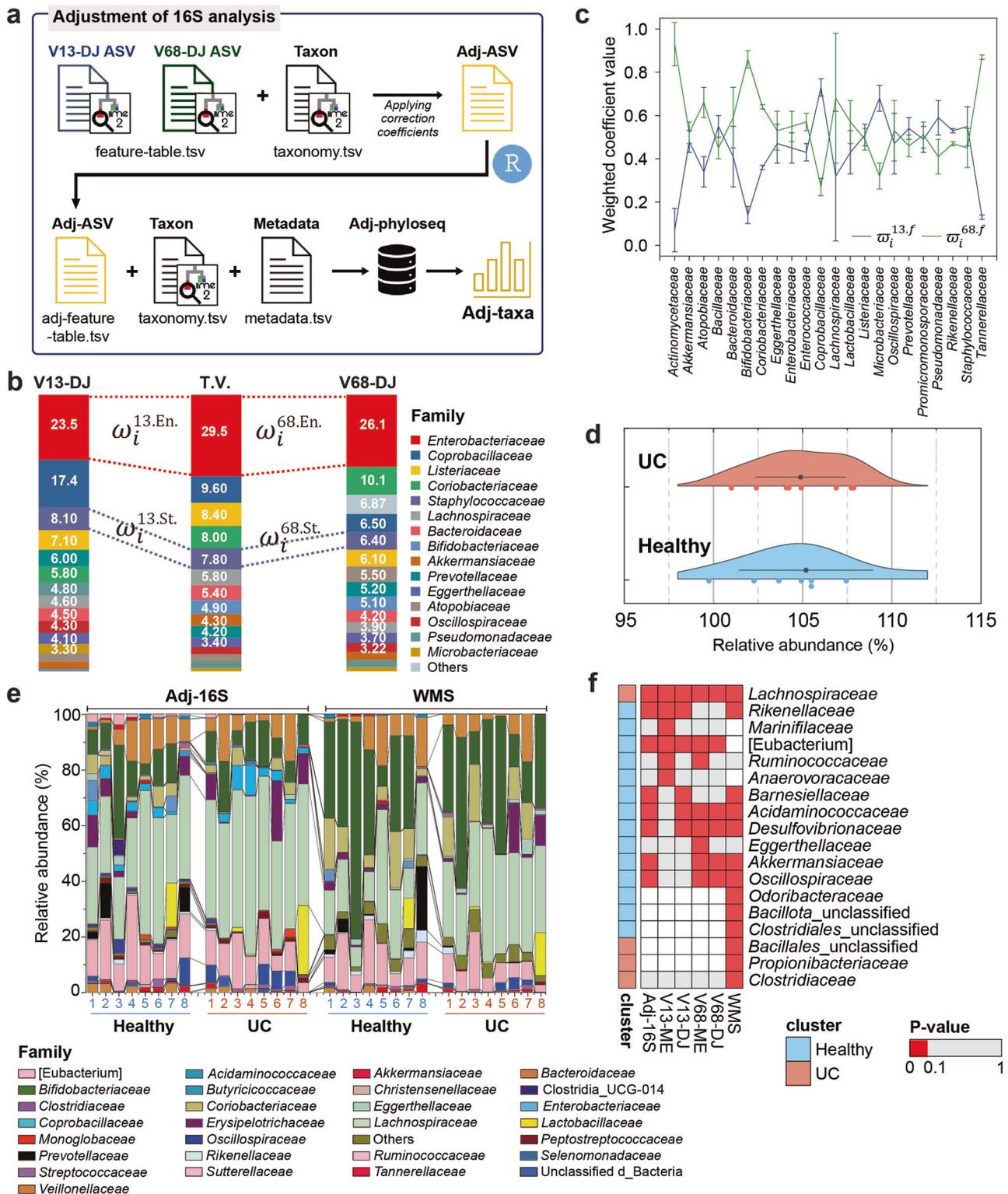


Fig. 5 | Adjusted relative abundance in Korean gut microbial communities.

a Schematic diagram illustrates a systematic approach to adjusting microbiome composition by applying correction coefficients, leading to a desired microbial balance. **b** Bar graphs representing detailed taxonomic resolution derived from V1-V3 and V6-V8 regions compared to theoretical values using mock community dataset. **c** Weighted coefficient values derived from the V1-V3 and V6-V8 regions using eight independent datasets for 22 families. **d** A raincloud plot showing the adjusted total relative

abundance by applying weighted coefficient values for the Korean cohort. **e** Comparison of adjusted 16S rRNA (Adj-16S) and WMS profiles for healthy and UC cohorts. [Eubacterium]: [Eubacterium]_coprostanoligenes_group. **f** Heatmap of family-level gut microbiota differences between healthy and UC groups by analytical method. The color scale is gray ($0.1 \leq P$ value), red (P value < 0.1), and white (Not detected). Statistical analysis was performed using Welch's t test.

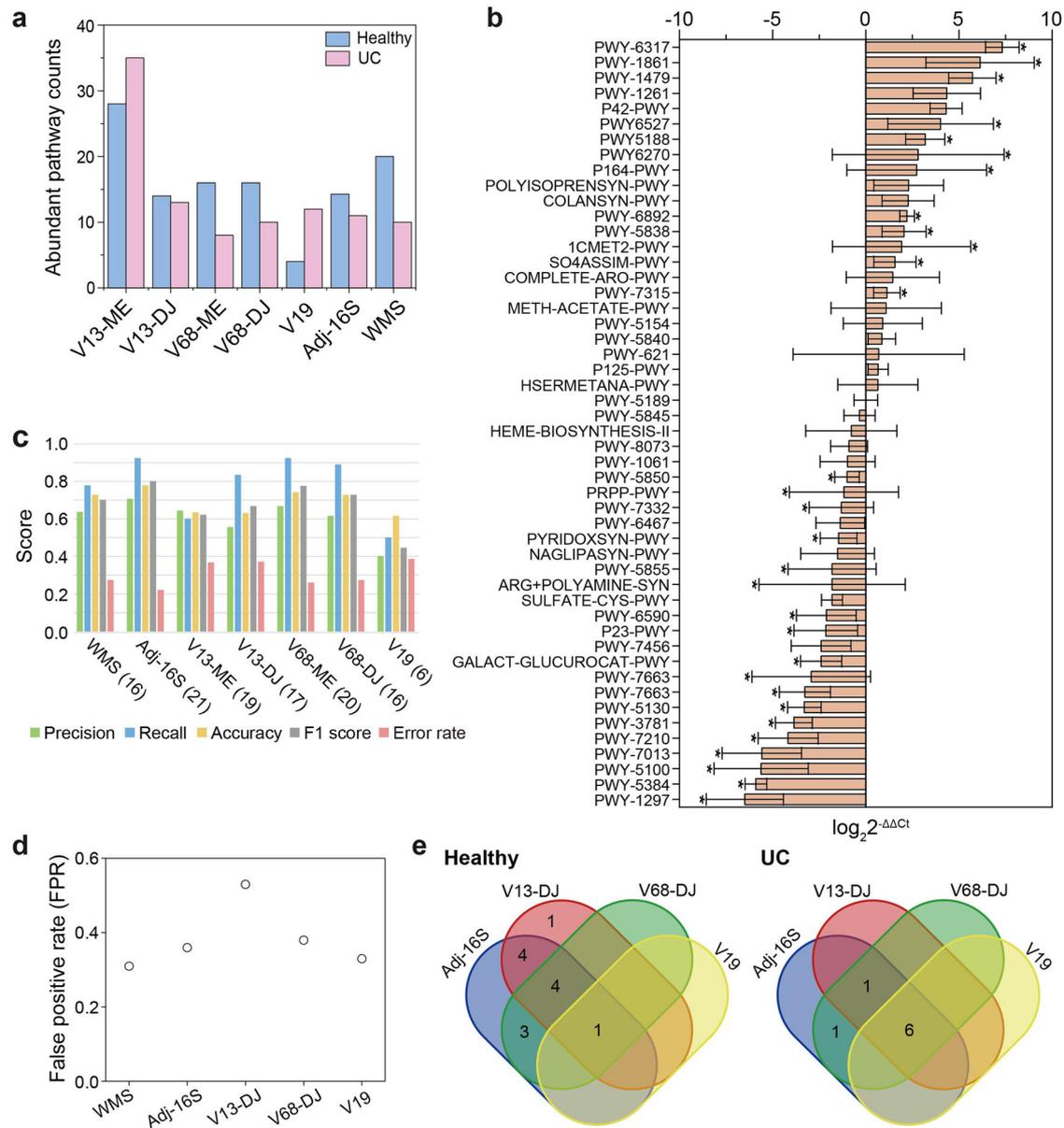


Fig. 6 | Functional pathway analysis in healthy and UC groups by WMS and 16S rRNA-derived PICRUSt2. **a** Number of functional pathways associated with each group. **b** Comparison of 50 genes between healthy ($n = 8$) and UC ($n = 8$) groups using qRT-PCR. The analysis aimed to calculate $\log_2^{-\Delta\Delta C_t}$ values (UC vs. Healthy) for candidate genes relative to 16S Ct values. The significance of differences between healthy and UC samples was assessed using a t test (*, P value < 0.1). **c** Histogram of precision, recall, accuracy, misclassification rate, and

F1 score for the datasets. Values in parentheses indicate the sum of true positive (TP) plus true negative (TN) results. For score calculation, metabolic pathways with a P value < 0.1 were used. **d** The false positive (FP) rate was determined by dividing FP by the sum of FP and TN. **e** Venn diagrams illustrating the distribution of TPs (left/healthy) and TNs (right/UC) identified by various analytical methods.

Discussion

In this study, we enhanced gut microbiota profiling through a concatenation approach using pivotal 16S rRNA gene regions, V1-V3 and V6-V8. This method diverges from conventional practices that rely primarily on single-region amplicons and merged reads. Our goal was to refine taxonomic assignment and deepen functional characterization of the gut microbiome, crucial for deciphering its role in health and disease. Previous research has often failed to provide robust experimental validation linking specific microbes to health outcomes or distinguishing functional differences between diseased and healthy states^{6,42,43}.

Our Adj-16S method aimed to minimize biases inherent in using single 16S rRNA regions. This approach significantly increased mapped read ratios and microbial identification resolution, thereby improving our

understanding of taxonomic structures within the gut microbiota (Supplementary Figs. 4 and 9e–g). For instance, it clarified the presence of specific taxa such as Oscillospirales and *Romboutsia* in non-IBD individuals and taxa such as the Family_XII_AD3011_group in CD, with greater clarity compared to conventional methods. Notably, *Romboutsia*, known for its beneficial acetate and propionate production⁴⁴, is diminished in CD compared to healthy individuals^{45,46}. Similarly, the *Oscillospiraceae* family, associated with anti-inflammatory valeric acid, shows higher abundance in healthy individuals than in those with CD⁴⁷. Notably, beneficial taxa associated with anti-inflammatory and metabolic benefits showed differential abundance in healthy versus CD individuals, highlighting potential therapeutic targets. Additionally, our analysis has refined the taxonomic resolution for taxa such as *Roseburia* and *Phascolarctobacterium* in the V1-V3

region (Supplementary Figs. 9 and 10). Furthermore, the concatenating method improved differentiation between gut microbiota of healthy individuals and those with UC, emphasizing the value of multiple regions for a comprehensive analysis (Fig. 3). By applying correction coefficients derived from mock community datasets, we aligned our relative abundance profiles more closely with WMS data, enhancing the accuracy of our analyses (Fig. 5 and Supplementary Fig. 15).

Our predictive functional profiling further delineated significant metabolic pathways associated with both healthy individuals and UC patients. Notably, pathways such as HEME-BIOSYTHESIS-II, PWY-5845 (menaquinol-9 biosynthesis), and PWY-5850 (menaquinol-6 biosynthesis) were prevalent in healthy individuals, which were not detected by other 16S rRNA-based methods and WMS (Supplementary Fig. 16b). These pathways are implicated in vitamin K deficiencies commonly observed in IBD patients⁴⁸ and may serve as vital biomarkers for IBD diagnosis^{48,49}. Interestingly, these pathways did not appear significant in WMS findings, underscoring the unique strengths of the Adj-16S method. Conversely, pathways such as NAGLIPASYN-PWY were identified as more prevalent in healthy individuals than in UC, which contradicts some reports associated with CD⁵⁰. For UC, consistent detection of six metabolic pathways such as PWY-5189 (tetrapyrrole biosynthesis II), PWY-621 (sucrose degradation III), METH-ACETATE-PWY, and PWY-7315 (dTDP-N-acetylthomosamine biosynthesis) across all analytical methods, including Adj-16S and various concatenated 16S rRNA sequencing methods, corresponded with literature suggesting altered metabolic states in UC patients (Fig. 6e). These findings align with observations that healthy individuals have higher levels of tetrapyrrole and its derivatives compared to the UC group⁵¹, suggesting a compensatory biosynthesis in UC might instigate heightened biosynthesis (PWY-5189). The β -fructofuranosidase gene linked to *Eubacterium rectale* was found in healthy samples and genomes related to *Lachnospiraceae bacterium* isolate MGYG-HGUT-02492, which is more abundant in UC. Furthermore, we observed the activation of starch degradation pathways in individuals with inflammatory bowel syndrome with diarrhea⁵², possibly indicating a connection to gut dysbiosis in UC. Pathways involved in the biosynthesis of compounds linked to inflammation, such as kynurenic acid (METH-ACETATE-PWY) and lipopolysaccharide biosynthesis (PWY-7315) were found to be elevated in UC^{53,54}, suggesting potential involvement in inflammatory processes. Interestingly, these pathways did not manifest as significant in WMS data, except for NAGLIPASYN-PWY and METH-ACETATE-PWY, indicating nuanced differences in the detection capabilities of various sequencing methodologies. Additionally, ARG + POLYAMINE-SYN, GALACT-GLUCURO-CAT-PWY, PWY-5130, PWY-7663, and PWY-8073, all associated with a healthy status, and 1CMET2-PWY, HSERMETANA-PWY, P125-PWY, PWY-1861, PWY-6270, PWY-6527 associated with UC, were corroborated by the WMS method.

The metabolic pathways uncovered using the Adj-16S and WMS methodologies offer promising avenues for a deeper understanding of UC, providing potential pathways for diagnostics and therapeutic development. The advanced 16S rRNA-based analytical method and WMS could illuminate our understanding of gut microbiota structure^{55,56}. Despite exploring deep shotgun sequencing analysis, our findings indicate that this technique did not notably improve our discrimination between the functional pathways of healthy individuals and those with UC (Supplementary Fig. 17). Even when varying sequencing depths—4 GB, 18 GB, and 36 GB—were employed, the ability to distinguish between these states did not significantly change⁵⁷. However, WMS is recognized for its capability to capture both taxonomic and functional features of bacteria and fungi, which may remain elusive with 16S rRNA sequencing⁵⁵. Recent advancements in metagenome-assembled genomes (MAGs) present a promising opportunity for exploring the ‘dark matter’ of the human gut microbiota⁵⁸. Yet, the WMS approach in this study, based on reference-based methods like MetaPhlan3, faces limitations in detecting only cataloged species, thus missing a vast array of uncultivated microbes. To overcome these challenges, newer methods such as MetaPhlan 4 integrate both reference genomes and MAGs

to expand species-level genome bins, enabling more comprehensive taxonomic profiling⁵⁹. Continuous updates to databases (e.g., GTDB⁶⁰ and UHGG⁶¹) and hybrid approaches that blend reference-based and assembly-based strategies, are crucial for refining metagenomic analysis. Although functional analysis was partially validated through qRT-PCR in this study, the predictive results from 16S rRNA-based functional analysis using PICRUSt2 rely on a limited set of reference data. Therefore, to enhance the efficiency of the Adj-16S method, updating transitional ecological classifications to align with continuously updated databases would be beneficial. Additionally, assessing these computational methods with integrated multiomics data is critical for advancing our understanding of microbial functions and interactions in the gut microbiota.

Our study acknowledges the limitations inherent in the scale and diversity of our mock community dataset. To refine the accuracy of our equations, expanding this dataset with a more extensive and diverse range of mock communities is essential. Such expansion would strengthen our foundation for using concatenated V1-V3 and V6-V8 16S rRNA regions to achieve a thorough gut microbiome analysis. While this method is optimized for analyzing the adult gut microbiome, its applicability to other environments (e.g., soil and marine environments) or human body sites (e.g., skin, saliva, and urinary tract) might require tailored analytical approaches. Additionally, the acquisition of robust results regarding differences in bacterial compositions between healthy individuals and UC patients requires the application of multiple differential abundance methods (e.g., ALDEx2 and ANCOM-II)⁶² rather than just LefSe method used in this study. However, a small sample size (fewer than 10 per group) can lead to a higher false discovery rate with bias correction algorithms such as ANCOM-II compared to Wilcoxon⁶³. In addition, the method’s performance can be optimized by implementing pipelines (e.g., TIC pipeline) that enhance the clustering of unclassified taxa⁶⁴.

In summary, concatenating unmergeable reads has fine-tuned the resolution of our gut microbiome profiling, allowing us a more detailed representation of the gut ecosystem. We have identified distinct metabolic pathways that differentiate healthy individuals from those with UC. Our approach offers an efficient, cost-effective, and labor-intensive approach for unraveling the complex interactions between hosts and microbes in the gut. This advancement enhances our ability to accurately map and comprehend these interactions is poised to make substantial impacts on developing targeted interventions, potentially revolutionizing patient care and therapeutic approaches.

Methods

Microbial community datasets

We utilized the SRP115494 (Longitudinal Multiomics of the Human Microbiome in IBD)⁷ and SRP131748 (Human Metagenome on pre-diabetic humans)⁴⁰ datasets from NCBI for our microbiome analysis strategy. Additionally, the SRP291583 dataset²⁴, comprising mock community datasets, was employed to validate gut microbiome analysis and develop correction coefficient formulas.

Preparation of human gut stools

Stool samples were collected from healthy Korean individuals ($n = 8$) and UC patients ($n = 8$), aged 19 to 45, randomly recruited following approval by the institutional review board (IRB) of Severance Hospital (IRB No. 4-2020-1487) (details in Supplementary Data 4). Written consent was obtained, and participants underwent a survey capturing basic information, including demographics, medical history, current medications, and gastrointestinal symptoms potentially affecting gut microbiota composition. Exclusion criteria included diagnosis with the disease during the study period, underlying diseases (e.g., malignancy, multiorgan failure, or peptic ulcer), current medications affecting the gastrointestinal tract that could not be discontinued seven days prior (e.g., proton pump inhibitors, antacids, and antibiotics), pregnancy, or failure to pass blood and stool screening tests. Stool samples were collected in conical tubes and transported to Yonsei

Severance FMT center, Seoul, Korea, where they were stored at -80 °C until DNA extraction.

DNA extraction and 16S rRNA amplicon sequencing

Fecal DNA was extracted using the QIAamp PowerFecal Pro DNA Kits (QIAGEN, Germany) following the manufacturer's instructions. Amplicon libraries were prepared according to Illumina's 16S metagenomic sequencing library preparation protocol using 12.5 ng of DNA from each sample⁶⁵. Various 16S rRNA partial regions were amplified with specific primers using MiSeq platform: V1-V3 with 27 F/534 R (27 F: 5'-AGAGTTT-GATCCTGGCTCAG-3', 534 R: 5'-ATTACCGCGGCTGCTGG-3'), V6-V8 with 968 F/1378 R (968 F: 5'-AACGCGAAGAACCTTAC-3', 1378 R: 5'-CGGTGTGTACAAGGCCCGGAAC-G-3'). The 16S rRNA V1-V9 region was sequenced using the Pacbio Sequel platform, excluding four samples (N022, N031, UC007, UC009). The 16S rRNA full length was amplified with V1-V9 region using 27 F/1492 R (27 F: 5'-AGAGTTT-GATCMTGGCTCAG-3', 1492 R: 5'-TACGGYTACCTTGTTAYGACTT-3'). For long-read sequencing, DNA libraries were prepared using the Procedure & Checklist—Amplification of Full-Length 16S Gene with Bar-coded Primers for Multiplexed SMRTbell® Library Preparation and Sequencing⁶⁶. The median reads per sample were 116,528 for the V1-V3 region, 115,738 for V6-V8, and 19,426 for V1-V9.

Analytical pipelines

We employed five analytical pipelines for 16S rRNA and WMS raw data. After trimming adapter sequences using fastp v0.23.2, 16S rRNA reads were merged with DADA2 plugged in QIIME2⁶⁷. The sequence length was trimmed based on QIIME2 Phred quality score (Q score) plots (Supplementary Figs. 1, 3, 8, and 13). The sequence trim was determined as the position where a median Q score lower than Q20 was first found. The reads were merged with at least a 12 bp overlap as a default value in QIIME2⁶⁷ and a quality score of Q20. In the V1-V9 region, we used all single-end sequences for which Q scores were above 20. Denoising and removal of chimera sequences were performed using DADA2.

In the concatenating method (DJ and IO), sequences were trimmed with fastp v0.23.2⁶⁸, and the processed forward and reverse reads were then subjected to concatenation via JTax⁶⁹. DJ method: 5'-forward reads-3'-NNNNNNNN-3'-reverse complement of reverse reads-5', IO method: 3'-reverse complement of reverse reads-5'-NNNNNNNN-5'-forward reads-3'. The concatenated position is connected to NNNNNNNN. During the concatenation of forward and reverse sequences, an overlapping portion of the sequences was generated, prioritizing the use of the forward sequence whenever possible. Reverse sequences were employed to fill gaps in the concatenation where the ideal PCR fragment length was not achieved. For example, in previous datasets^{7,40}, where the trim position of forward sequences was 249 and that of reverse sequences was 43 on the V4 target region, concatenated sequences of 292 bp were generated (Supplementary Fig. 8). Similarly, for datasets^{7,40}, where the trim position of forward sequences was 301 and that of reverse sequences was 207 on the V1-V3 target region, concatenated sequences of 508 bp were produced. Subsequently, the same trimming approach was applied to other 16S rRNA regions (Supplementary Figs. 3 and 13). Single-end sequences obtained from the concatenating method were then used for further analysis. Amplicon sequence variants (ASVs) were generated by DADA2 for taxa classification and functional profiling.

Shotgun sequencing

We performed shallow to deep WMS at 4 GB, 18 GB, and 36 GB levels to compare gut microbiota structures with 16S rRNA-based metagenome sequencing. WMS libraries, prepared with at least 100 ng of total DNA using the Illumina TruSeq Nano DNA protocol for 350 bp libraries (Illumina, San Diego, CA), were sequenced on the Illumina Novaseq 6000 platform. This generated 2 × 150 bp paired-end reads with a minimum of 27.7 million reads per sample. Sequencing depths achieved a median of 29.9 million reads for 4 Gbp samples, 139.6 million for 18 Gbp, and 255.9 million for

36 Gbp. KneadData v0.12.0, incorporating Trimmomatic v0.39.2⁷⁰ and TRF v.4.09.1⁷¹, was used to filter low-quality and adapter-laden WMS reads. Human-origin reads were removed using the human reference genome GRCh38⁷².

Taxonomic classification

Non-chimeric reads obtained from each method (ME, DJ, and IO) were aligned to databases: Greengenes2 (v2022.10), SILVA (v138.1), and RDP (v11). Taxa classification of ASVs was performed using *feature-classifier classify-sklearn* plugged in QIIME2. Relative abundance and richness metrics, such as Richness, Shannon effective numbers, and Simpson effective numbers, were visualized using phyloseq⁷³, vegan⁷⁴, and ggplot2 R packages. Before calculating Hill numbers⁷⁵ and relative abundance, we implemented a 0.25% cutoff retaining only ASVs observed at a relative abundance >0.25% in at least one sample⁷⁶. We used the SILVA database for taxonomic assignments, except in Supplementary Figs. 5–7, which illustrate differences in relative abundance depending on the 16S rRNA database. Additionally, we updated the taxonomic assignments from an older version by using <https://lpsn.dsmz.de>. Differential heat trees of taxonomic compositions, provided by feature-table with QIIME2 and phyloseq in each sample, were visualized using Metacoder⁷⁷. Linear discriminant analysis effect size (LEfSe)⁷⁸ was used to evaluate differential ASV abundance by analytical methods and healthy/diseased states.

Coefficient-based adjustments for dual 16S rRNA reads

Utilizing mock datasets (Zymo, ZIEL-I, ZIEL-II) and the SILVA DB, we calculated coefficient values for each 16S rRNA region using weighted averages⁴¹, covering 22 families. Weighted coefficient values for the V1-V3 ($\omega_i^{13,f}$) and V6-V8 ($\omega_i^{68,f}$) regions were determined by dividing the relative abundance from the V1-V3 ($x_i^{13,f}$) and V6-V8 ($x_i^{68,f}$) regions by the total abundance [the sum of the relative abundances from the V1-V3 region ($x_i^{13,f}$) and the V6-V8 region ($x_i^{68,f}$)], respectively. The means of these weighted coefficient values ($\bar{\omega}_i^{13,f}$ and $\bar{\omega}_i^{68,f}$) were computed using data from eight independent datasets across three mock datasets. The equations applied are as follows:

- For the V1-V3 region:

$$\omega_i^{13,f} = \frac{x_i^{13,f}}{x_i^{13,f} + x_i^{68,f}} \quad (1)$$

and the mean

$$\bar{\omega}_i^{13,f} = \frac{1}{n} \sum_{i=1}^n \omega_i^{13,f} \quad (2)$$

Similarly, we calculated weighted coefficient values for the V6-V8 region ($\omega_i^{68,f}$) as the relative abundance from the V6-V8 region ($x_i^{68,f}$) divided by the total abundance ($x_i^{13,f} + x_i^{68,f}$), with the mean of these values from the V6-V8 region ($\bar{\omega}_i^{68,f}$) presented as follows:

- For the V6-V8 region:

$$\omega_i^{68,f} = \frac{x_i^{68,f}}{x_i^{13,f} + x_i^{68,f}} \quad (3)$$

and the mean

$$\bar{\omega}_i^{68,f} = \frac{1}{n} \sum_{i=1}^n \omega_i^{68,f} \quad (4)$$

The adjusted relative abundances for both the V1-V3 ($x_i^{13,f'}$) and V6-V8 ($x_i^{68,f'}$) regions were then calculated as follows:

$$x_i^{13,f'} = \bar{\omega}_i^{13,f} \cdot x_i^{13,f} \quad (5)$$

and

$$x_i^{68.f'} = \hat{\omega}_i^{68.f'} \cdot x_i^{68.f} \quad (6)$$

These adjusted relative abundances lead to the formula representing the total adjusted relative abundance:

$$\text{Total relative abundance} \simeq \sum_{i=1}^n (x_i^{13.f'} + x_i^{68.f'}) \quad (7)$$

Functional profiling

For function profiling prediction, processed 16S rRNA sequencing ASVs were analyzed based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) database⁷⁹ using Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt2)¹⁴, with limitations noted for the IO method. WMS reads were analyzed by HUMAnN 3.0, including MetaPhlan3⁸⁰, mapping to the UniRef90 DB.

Pipeline performance measures with mock community

Performance was evaluated at the family level using TP, FP, false negative (FN), precision, recall, and F-measure³⁸. TP was calculated when relative abundance values from each pipeline matched ideal values from the mock community dataset. FP was calculated when values from each pipeline were overestimated and misclassified compared to the mock community. FN was calculated when values were not detected. The calculations for performance metrics are defined as follows: Precision = TP/(TP + FP), Recall = TP/(TP + FN), F-measure = 2 × precision × recall/(precision + recall).

Quantitative PCR validation

We performed quantitative PCR analysis on samples from eight healthy individuals and 8 patients with UC to quantify the abundance of genes from selected pathways. Primers were designed using the Primer-BLAST tool⁸¹. Primers listed in Supplementary Data 7 targeted representative genes from each pathway, with standard primers (515 F: 5'-GTGCCAGCMGCCGCGGTAA-3'/806 R: 5'-GGAC-TACHVGGGTWTCTAAT-3') for 16S rRNA⁸². The qRT-PCR reaction was conducted with a final primer concentration diluted to 0.5 μM, including 5 ng of genomic DNA in a 10 μl final reaction volume, using the iQ SYBR Green Supermix (BIO-RAD). The quantitative PCR conditions were as follows: pre-denaturation at 95 °C for 3 min; denaturation at 95 °C for 10 s for 40 cycles; and annealing at 55 °C for 30 s, followed by melt curve analysis. The qRT-PCR analysis aimed to calculate 2^{-ΔΔCt} values between candidate genes and 16S Ct values. The statistical significance of the differences between healthy individuals and UC samples was assessed using a *t* test (*P* < 0.1).

Statistical analysis and visualization

We used the *t* test, ANOVA of Origin software, and two-sided Welch's *t* test of STAMP software⁸³ to assess the significance of differences in the abundant microbiome and functional profiles between the healthy and UC patient groups.

Availability of supporting source code and requirements

Programming languages: Python 3.9.7, R 4.2.2.

Home page: <https://github.com/TLLab/JTax>.

Other requirements: BioPython module, R packages (ggplot2, phyloseq, vegan, metacoder, ggpubr, and microbiomeMarker⁸⁴), STAMP, and Origin software.

Data availability

The SRP131748 and SRP115494 datasets are available on the NCBI or HMP portal (<https://portal.hmpdacc.org/>). The SRP291583 dataset is also available on the NCBI. Our 16S rRNA sequencing and WMS raw data

supporting the results of this article are available from the Sequence Read Archive, with project ID: PRJNA1088906, PRJNA1088910 in the NCBI.

Code availability

We have uploaded the codes necessary to perform the Adj-16S method (<https://github.com/kyoung-su/Adj-16S>).

Received: 18 April 2024; Accepted: 25 March 2025;

Published online: 12 April 2025

References

- Zheng, D., Liwinski, T. & Elinav, E. Interaction between microbiota and immunity in health and disease. *Cell Res.* **30**, 492–506 (2020).
- Almeida, A. et al. A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).
- Helmink, B. A., Khan, M. A. W., Hermann, A., Gopalakrishnan, V. & Wargo, J. A. The microbiome, cancer, and cancer therapy. *Nat. Med.* **25**, 377–388 (2019).
- Maruvada, P., Leone, V., Kaplan, L. M. & Chang, E. B. The human microbiome and obesity: moving beyond associations. *Cell Host Microbe* **22**, 589–599 (2017).
- Crudele, L., Gadaleta, R. M., Cariello, M. & Moschetta, A. Gut microbiota in the pathogenesis and therapeutic approaches of diabetes. *EBioMedicine* **97**, 104821 (2023).
- Clooney, A. G. et al. Ranking microbiome variance in inflammatory bowel disease: a large longitudinal intercontinental study. *Gut* **70**, 499–510 (2021).
- Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
- Cryan, J. F., O'Riordan, K. J., Sandhu, K., Peterson, V. & Dinan, T. G. The gut microbiome in neurological disorders. *Lancet Neurol.* **19**, 179–194 (2020).
- Diebold, P. J. et al. Clinically relevant antibiotic resistance genes are linked to a limited set of taxa within gut microbiome worldwide. *Nat. Commun.* **14**, 7366 (2023).
- Worby, C. J. et al. Gut microbiome perturbation, antibiotic resistance, and *Escherichia coli* strain dynamics associated with international travel: a metagenomic analysis. *Lancet Microbe* **4**, e790–e799 (2023).
- Chiu, C. Y. & Miller, S. A. Clinical metagenomics. *Nat. Rev. Genet.* **20**, 341–355 (2019).
- Dai, D. et al. GMrepo v2: a curated human gut microbiome database with special focus on disease markers and cross-dataset comparison. *Nucleic Acids Res.* **50**, D777–D784 (2022).
- Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
- Douglas, G. M. et al. PICRUSt2 for prediction of metagenome functions. *Nat. Biotechnol.* **38**, 685–688 (2020).
- Truong, D. T. et al. MetaPhlan2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
- Abdill, R. J., Adamowicz, E. M. & Blekhan, R. Public human microbiome data are dominated by highly developed countries. *PLoS Biol.* **20**, e3001536 (2022).
- Pareek, S. et al. Comparison of Japanese and Indian intestinal microbiota shows diet-dependent interaction between bacteria and fungi. *NPJ Biofilms Microbiomes* **5**, 37 (2019).
- Lee, J. E., Kim, K. S., Koh, H., Lee, D. W. & Kang, N. J. Diet-induced host-microbe interactions: personalized diet strategies for improving inflammatory bowel disease. *Curr. Dev. Nutr.* **6**, nzac110 (2022).
- Walker, A. W. et al. 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. *Microbiome* **3**, 26 (2015).
- Johnson, J. S. et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* **10**, 5029 (2019).

21. Kwon, S., Lee, B. & Yoon, S. CASPER: context-aware scheme for paired-end reads from high-throughput amplicon sequencing. *BMC Bioinformatics* **15**, S10 (2014).
22. Schloss, P. D., Gevers, D. & Westcott, S. L. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* **6**, e27310 (2011).
23. Bokulich, N. A. et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* **10**, 57–59 (2013).
24. Abellan-Schneyder, I. et al. Primer, pipelines, parameters: issues in 16S rRNA gene sequencing. *mSphere* **6**, e01202–e01220 (2021).
25. McLaren, M. R., Willis, A. D. & Callahan, B. J. Consistent and correctable bias in metagenomic sequencing experiments. *Elife* **8**, e46923 (2019).
26. Chen, Z. et al. Impact of preservation method and 16S rRNA hypervariable region on gut microbiota profiling. *mSystems* **4**, e00271–18 (2019).
27. Eren, A. M., Borisy, G. G., Huse, S. M. & Mark Welch, J. L. Oligotyping analysis of the human oral microbiome. *Proc. Natl Acad. Sci. USA* **111**, E2875–E2884 (2014).
28. Alcon-Giner, C. et al. Optimisation of 16S rRNA gut microbiota profiling of extremely low birth weight infants. *BMC Genomics* **18**, 841 (2017).
29. Soriano-Lerma, A. et al. Influence of 16S rRNA target region on the outcome of microbiome studies in soil and saliva samples. *Sci. Rep.* **10**, 13637 (2020).
30. Matsuo, Y. et al. Full-length 16S rRNA gene amplicon analysis of human gut microbiota using MinION nanopore sequencing confers species-level resolution. *BMC Microbiol.* **21**, 35 (2021).
31. Wemheuer, F. et al. Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. *Environ. Microbiome* **15**, 11 (2020).
32. Comeau, A. M., Douglas, G. M. & Langille, M. G. Microbiome helper: a custom and streamlined workflow for microbiome research. *mSystems* **2**, e00127–16 (2017).
33. Kim, N. et al. Genome-resolved metagenomics: a game changer for microbiome medicine. *Exp. Mol. Med.* **56**, 1501–1512 (2024).
34. McIntyre, A. B. R. et al. Correction to: comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol.* **20**, 72 (2019).
35. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
36. Peterson, D. et al. Comparative analysis of 16S rRNA gene and metagenome sequencing in pediatric gut microbiomes. *Front. Microbiol.* **12**, 670336 (2021).
37. Callahan, B. J. et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
38. Dacey, D. P. & Chain, F. J. J. Concatenation of paired-end reads improves taxonomic classification of amplicons for profiling microbial communities. *BMC Bioinformatics* **22**, 493 (2021).
39. Ramakodi, M. P. Merging and concatenation of sequencing reads: a bioinformatics workflow for the comprehensive profiling of microbiome from amplicon data. *FEMS Microbiol. Lett.* **371**, fnae009 (2024).
40. Zhou, W. et al. Longitudinal multi-omics of host-microbe dynamics in prediabetes. *Nature* **569**, 663–671 (2019).
41. Clark-Carter D. *Measures of central tendency*. Elsevier (2010).
42. Pisani, A. et al. Dysbiosis in the gut microbiota in patients with inflammatory bowel disease during remission. *Microbiol. Spectr.* **10**, e0061622 (2022).
43. Sankarasubramanian, J., Ahmad, R., Avuthu, N., Singh, A. B. & Guda, C. Gut microbiota and metabolic specificity in ulcerative colitis and crohn's disease. *Front. Med.* **7**, 606298 (2020).
44. Gerritsen, J. et al. Genomic and functional analysis of *Romboutsia ilealis* CRIB(T) reveals adaptation to the small intestine. *PeerJ* **5**, e3698 (2017).
45. Qiu, X. et al. Characterization of fungal and bacterial dysbiosis in young adult Chinese patients with Crohn's disease. *Ther. Adv. Gastroenterol.* **13**, 1756284820971202 (2020).
46. Hu, J. et al. Correlation between altered gut microbiota and elevated inflammation markers in patients with Crohn's disease. *Front. Immunol.* **13**, 947313 (2022).
47. Chen, X. et al. Polysaccharides from *Sargassum fusiforme* after UV/H₂O₂ degradation effectively ameliorate dextran sulfate sodium-induced colitis. *Food Funct.* **12**, 11747–11759 (2021).
48. Weisshof, R. & Chermesh, I. Micronutrient deficiencies in inflammatory bowel disease. *Curr. Opin. Clin. Nutr. Metab. Care* **18**, 576–581 (2015).
49. Gisbert, J. P. & Gomollon, F. Common misconceptions in the diagnosis and management of anemia in inflammatory bowel disease. *Am. J. Gastroenterol.* **103**, 1299–1307 (2008).
50. Ananthkrishnan, A. N. et al. Gut microbiome function predicts response to anti-integrin biologic therapy in inflammatory bowel diseases. *Cell Host Microbe* **21**, 603–610 e603 (2017).
51. Franzosa, E. A. et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* **4**, 293–305 (2019).
52. Wang, Y. et al. Diet and gut microbial associations in irritable bowel syndrome according to disease subtype. *Gut Microbes* **15**, 2262130 (2023).
53. Palmieri, O. et al. Microbiome analysis of mucosal ileoanal pouch in ulcerative colitis patients revealed impairment of the pouches immunometabolites. *Cells* **10**, 3243 (2021).
54. Wang, D. et al. GPR35-mediated kynurenic acid sensing contributes to maintenance of gut microbiota homeostasis in ulcerative colitis. *FEBS Open Bio* **13**, 1415–1433 (2023).
55. Usyk, M. et al. Comprehensive evaluation of shotgun metagenomics, amplicon sequencing, and harmonization of these platforms for epidemiological studies. *Cell Rep. Methods* **3**, 100391 (2023).
56. Tessler, M. et al. Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Sci. Rep.* **7**, 6589 (2017).
57. Mas-Lloret, J. et al. Gut microbiome diversity detected by high-coverage 16S and shotgun sequencing of paired stool and colon sample. *Sci. Data* **7**, 92 (2020).
58. Yang, C. et al. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Comput Struct. Biotechnol. J.* **19**, 6301–6314 (2021).
59. Blanco-Miguez, A. et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlan 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).
60. Parks, D. H. et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).
61. Almeida, A. et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
62. Nearing, J. T. et al. Microbiome differential abundance methods produce different results across 38 datasets. *Nat. Commun.* **13**, 342 (2022).
63. Lin, H. & Peddada, S. D. Analysis of compositions of microbiomes with bias correction. *Nat. Commun.* **11**, 3514 (2020).
64. Kioukis, A., Pourjam, M., Neuhaus, K. & Lagkouvardos, I. Taxonomy informed clustering, an optimized method for purer and more informative clusters in diversity analysis and microbiome profiling. *Front. Bioinform.* **2**, 864597 (2022).
65. Illumina 16S metagenomic sequencing library preparation (Illumina Technical Note 15044223). *Illumina*, <https://sapac.support.illumina>.

- [com/downloads/16s_metagenomic_sequencing_library_preparation.html](#) (2013).
66. PACBIO. Procedure & Checklist - Preparing SMRTbell® Libraries using PacBio® Barcoded Universal Primers for Multiplexing Amplicons <https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Preparing-SMRTbell-Libraries-using-PacBio-Barcoded-Universal-Primers-for-Multiplexing-Amplicons.pdf>.
 67. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
 68. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
 69. Liu, T. et al. Joining Illumina paired-end reads for classifying phylogenetic marker sequences. *BMC Bioinformatics* **21**, 105 (2020).
 70. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
 71. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
 72. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
 73. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**, e61217 (2013).
 74. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).
 75. Jost, L. Entropy and Diversity. *Oikos* **113**, 363–375 (2006).
 76. Reitmeier, S. et al. Handling of spurious sequences affects the outcome of high-throughput 16S rRNA gene amplicon profiling. *ISME Commun.* **1**, 31 (2021).
 77. Foster, Z. S., Sharpton, T. J. & Grunwald, N. J. Metacoder: an R package for visualization and manipulation of community taxonomic diversity data. *PLoS Comput. Biol.* **13**, e1005404 (2017).
 78. Segata, N. et al. Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).
 79. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
 80. Beghini, F. et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* **10**, e65088 (2021).
 81. Ye, J. et al. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**, 134 (2012).
 82. Wu, Y. et al. Identification of microbial markers across populations in early detection of colorectal cancer. *Nat. Commun.* **12**, 3063 (2021).
 83. Parks, D. H., Tyson, G. W., Hugenholtz, P. & Beiko, R. G. STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics* **30**, 3123–3124 (2014).
 84. Cao, Y. et al. microbiomeMarker: an R/Bioconductor package for microbiome marker identification and visualization. *Bioinformatics* **38**, 4027–4029 (2022).

Acknowledgements

We express our gratitude to the Human Microbiome Project (HMP) and the laboratories of George M. Weinstock, Michael Snyder, Andrea

Mariani, and Nicholas Chia for providing the raw 16S rRNA data essential for this study. We also extend our thanks to Klaus Neuhaus's laboratory for supplying the mock community datasets. This work was partly supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) of Korea grant (RS-2021-NR056579 to D.W.L.), funded by the Ministry of Science and ICT (MSIT), Republic of Korea. This work is also partly supported by the Bioindustrial Technology Development Program of Korea grant (20018770), funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea). This research is also partly supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2023-KH141436).

Author contributions

K.S.K., H.K., and D.W.L. formulated the research plan. K.S.K. and J.H.N. performed the experiments. K.S.K., B.S.K., and D.W.L. analyzed the data. K.S.K. and D.W.L. wrote the manuscript. H.K. and D.W.L. conceived, planned, supervised, and managed the study.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41522-025-00686-x>.

Correspondence and requests for materials should be addressed to Hong Koh or Dong-Woo Lee.

Reprints and permissions information is available at <http://www.nature.com/reprints>

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025