

Characterization of the Gut Virome in Patients with Inflammatory Bowel Disease and Non-Alcoholic Fatty Liver Disease

Shuangshuang Lu^{1,*}, Yang Xia^{1,*}, Qiuyue Sun², Yuhui Sun³, Ruiyan Chen⁴, Hanshu Jin³, Jingxian Zhang³, Wenjia Liu^{1,3}, Jin Huang^{1,3}

¹Department of Gastroenterology, The Second People's Hospital of Changzhou, The Third Affiliated Hospital of Nanjing Medical University, Changzhou, Jiangsu, 213161, People's Republic of China; ²Nanjing Central Hospital, Nanjing, Jiangsu, 210018, People's Republic of China; ³Department of Gastroenterology, Graduate School of Nanjing Medical University, Nanjing, Jiangsu, 211166, People's Republic of China; ⁴Suzhou Medical College of Soochow University, Soochow University, Suzhou, Jiangsu, 215123, People's Republic of China

*These authors contributed equally to this work

Correspondence: Jin Huang; Wenjia Liu, Email hj042153@hotmail.com; 13813016290@126.com

Objective: The dysbiosis of the gut microbiota is a well-known correlate in the pathogenesis of inflammatory bowel disease (IBD). However, the microbiome characteristics of patients with IBD who also have non-alcoholic fatty liver disease (NAFLD) are understudied, particularly the potential pathogenic mechanisms of the gut virome.

Materials and Methods: In this study, we conducted a comprehensive gut virome correlation study, along with serum metabolomics analysis, by performing virus-like particle (VLP) and metagenomic sequencing on fecal samples from patients with inflammatory bowel disease and non-alcoholic fatty liver disease (IBD-NAFLD) and NAFLD (MASLD) controls without gastrointestinal diseases.

Results: The results showed that changes in the fecal virome were associated with IBD-NAFLD (MASLD), particularly with an increase in the abundance of Caudovirales in IBD-NAFLD (MASLD) patients. Subsequent analysis of the gut virome identified Bacteroides as the top predicted host for the viruses. Additionally, we identified the pathways involved in all differential metabolites through KEGG annotation analysis, with the highest correlation being the galactose metabolism pathway.

Conclusion: In conclusion, by using a customized integrated gut virome catalog tailored for IBD, we revealed the fundamental changes in the gut virome of IBD-NAFLD (MASLD) patients. This study is the first to uncover the specificity of the gut virome in IBD-NAFLD (MASLD) patients and predict Bacteroides as a potential host, suggesting a microbial signature primarily influenced by intestinal inflammation.

Keywords: gut virome, inflammatory bowel disease, nonalcoholic fatty liver disease, viral-like particle virome, IBD-NAFLD (MASLD) characterized viruses

Introduction

Inflammatory bowel disease (IBD)—an idiopathic, chronic, relapsing disorder—encompasses two major phenotypes, ulcerative colitis and Crohn's disease.¹ Once confined to the industrialised West, IBD is now marching across Asia alongside urbanisation and rising living standards. China illustrates the trend: incidence has climbed steeply over the past two decades and modelling predicts 1.5 million Chinese residents will be living with IBD by 2050.² The inexorable increase in patient numbers translates into prolonged medical care, recurrent hospitalisation, and mounting direct and indirect costs, shifting the global burden eastward. Although the trigger remains elusive, converging evidence implicates a quartet of interlocking factors—host genotype, environmental exposures, mucosal immunity, and the enteric microbiome.³ Among these, dysbiosis appears to act as a central hub, shaping the onset, trajectory, and behaviour of IBD in genetically susceptible individuals.

The gut microbiota is not a static collection of organisms but a fluid, self-regulating ecosystem that harbours an estimated 10^{14} microbial cells and their milieu—epithelial surfaces, mucus layer, and associated lymphoid tissue.^{3,4}

Beyond bacteria, this community encompasses fungi, archaea, and a vast array of viruses whose collective genomes outnumber our own. Differences in viral assemblages have been noted in individuals with IBD, yet when and why these signatures emerge remains unresolved.⁵ The human gut virome-dominated by bacteriophages but also encompassing eukaryotic and archaeal viruses-fluctuates in response to host genotype, diet, antibiotic exposure, and environmental triggers.⁶ These viral shifts are no longer considered mere bystanders; they are implicated in the onset or progression of colorectal cancer, chronic inflammatory disorders, and even cardiometabolic disease.³ By interacting with bacterial communities, sculpting early immune development, and modulating persistent inflammatory circuits, the virome is thought to fine-tune disease severity in concert with host genetics and immune status.⁷

Fatty liver disease is now the commonest chronic hepatic disorder worldwide, fueled either by alcohol or by metabolic risk. Its non-alcoholic subset (NAFLD), recently renamed metabolic dysfunction-associated steatotic liver disease (MASLD) according to international consensus, is diagnosed when $\geq 5\%$ of hepatocytes contain lipid droplets in the absence of excessive alcohol use.⁸ The condition forms a histological continuum: bland steatosis at one end and non-alcoholic steatohepatitis (NASH) at the other, where ballooned hepatocytes, lobular inflammation, and progressive fibrosis presage possible evolution to cirrhosis, hepatic failure, or hepatocellular carcinoma.⁸ NAFLD (MASLD) arises from a tangled interplay of inherited traits, nutrient excess, sedentary lifestyle, and insulin resistance, yet the gut–liver axis has emerged as a critical accelerator. Recent metagenomic work reveals that both alcoholic liver disease and NAFLD (MASLD) are accompanied by marked perturbations of the intestinal virome.⁹ Epidemiological mirrors of this biology show NAFLD (MASLD) prevalence climbing from 25.5% before 2005 to 37.8% in surveys conducted after 2016.¹⁰

Recent population-level surveys report that NAFLD (MASLD) occurs in roughly one-quarter of adults with IBD (median prevalence 26%), a figure that keeps rising even after stratification for obesity, diabetes, or other classical metabolic traits.^{11–13} This epidemiological divergence has refocused attention on gut-derived cues that might simultaneously fuel colonic inflammation and hepatic fat accumulation. In IBD, persistent dysbiosis, epithelial leakiness, and low-grade portal bacteraemia create a “second-hit” milieu that can ignite steatosis thousands of hepatocytes away from the inflamed mucosa.¹⁴ Yet microbial disruption extends beyond bacteria: Glassner et al noted that IBD-NAFLD (MASLD) subjects lacked the usual cardiometabolic stigmata seen in NAFLD-only controls, implying that non-traditional drivers-viral, genetic, or epigenetic-tip the liver toward lipid overload.¹⁵ Aggarwal’s biopsy-based cohort further revealed that Crohn’s disease, not ulcerative colitis, carries the heavier hepatic toll, with higher rates of ballooning and stage ≥ 2 fibrosis.¹⁶ Adding a virological layer, Lang et al recently demonstrated a stepwise erosion of phage diversity as NAFLD (MASLD) advanced from simple steatosis to NASH; conversely, the eukaryotic viral fraction remained stable, underscoring a selective collapse of bacteriophage populations that may uncouple microbial homeostasis and exacerbate both intestinal and hepatic injury.¹⁰

Although epidemiological bridges between NAFLD (MASLD) and IBD are now well documented, the molecular handshake that links a fatty liver to an inflamed colon remains fragmentary. Viral constituents of the gut are gaining scrutiny, yet how they orchestrate-or merely witness-this comorbidity is unresolved. Two main technical lenses are currently used: bulk metagenomic sequencing of total nucleic acids and virus-like particle (VLP) enrichment followed by deep sequencing.^{17,18} Both have catalogued phage expansion, contraction, or “blooms” in IBD cohorts, but these surveys stop at descriptive taxonomy and rarely ask whether virome deviation predicts, or even parallels, hepatic involvement. Consequently, most reports remain cross-sectional inventories that hunt for universal IBD-associated viral signatures,¹⁹ leaving the IBD-NAFLD (MASLD) intersection largely unexplored. Moreover, the exercise is technically handicapped: reference databases harbour $< 10\%$ of estimated global viral diversity, so the typical stool virome yields a torrent of dark matter—reads that cannot be assigned even at the family level.²⁰ Until expanded, curated phage and eukaryotic-virus databases are married to long-read and strain-resolved algorithms, the true extent to which viral predators or commensals drive the IBD-NAFLD (MASLD) axis will stay obscured.

Here we deployed a dual-layered strategy. We used VLP-enriched viromics and untargeted serum metabolomics to interrogate the converging gut–liver axis in individuals bearing both IBD and NAFLD (MASLD). By merging these data streams we generated a bespoke, disease-tuned virome catalogue that revealed numerous viral signatures absent from generic reference sets, and leveraged in-silico host-prediction tools to dissect the virus–bacteria–host circuitry underlying this comorbidity.

Methods

Human Subjects

All subjects involved in this study were rigorously selected from patients visiting the Department of Gastroenterology at the Third Affiliated Hospital of Nanjing Medical University from January 1, 2024, to June 30, 2024. Inclusion and screening were conducted by two gastroenterologists with over 10 years of clinical experience. A total of 32 subjects (18 with IBD-NAFLD (MASLD) as the disease group; 14 with only-NAFLD (MASLD) matched as the control group) consented to participate in the study; ultimately, 15 subjects (5 with CD-NAFLD (MASLD), 5 with UC-NAFLD (MASLD); 5 with only-NAFLD (MASLD), controls) provided fecal and blood samples.

Diagnostic criteria: IBD diagnosis: Strictly in accordance with the ECCO-ESGAR inflammatory bowel disease diagnostic assessment guidelines (2019),²¹ including clinical symptoms, endoscopic findings, histopathological examination and imaging results. NAFLD (MASLD) diagnosis: In accordance with the 2020 Asia-Pacific Association for the Study of the Liver (APASL) guidelines for the diagnosis and management of metabolic dysfunction-associated fatty liver disease (MAFLD),²² with imaging evidence of hepatic steatosis and exclusion of other causes of fatty liver (eg., excessive alcohol intake, viral hepatitis, drug-induced liver injury).

All included disease group subjects were experiencing their first episode with no prior history and were over 18 years old. For all subjects, IBD was diagnosed prior to NAFLD (MASLD) in the comorbid group, with a diagnosis interval of 3–12 months. Concurrently, all participants had not taken antibiotics/probiotics/prebiotics, or antiviral medications within three months prior to sample collection, had not undergone intestinal surgery, infectious diseases, or neoplastic diseases, and also excluded siblings with a history of IBD and NAFLD (MASLD). Control group subjects were matched with the disease group in terms of age, gender, and BMI to ensure consistency of results. All control group subjects underwent endoscopy, and the intestinal mucosa was confirmed to be free of inflammation/abnormalities. Individuals with heart, kidney, metabolic diseases (such as diabetes, moderate to severe hypertension), and cancer were excluded from this study. Disease severity: All IBD patients were in the mild-to-moderate stage (Harvey-Bradshaw Index for CD ≤ 7 ; Mayo Score for UC ≤ 6), and all NAFLD (MASLD) patients were in the simple steatosis stage (no evidence of NASH or liver fibrosis via imaging). Detailed clinical data of the study individuals, including disease scores, disease extent, medication use, complications, and serum indicators are provided in [Supplementary Table 1](#).

Ethical Standards

The study protocol complied with the ethical standards of the 1975 Declaration of Helsinki and was approved by the Ethics Committee of the Third Affiliated Hospital of Nanjing Medical University (The Second People's Hospital of Changzhou) with the ethical approval number [2024]KY204-01, and written informed consent was obtained from all participants. The study was designed and reported in accordance with the STROBE guidelines for observational studies, and the completed STROBE checklist is attached as [Supplementary Materials 1](#) (each checklist item is cross-referenced to the corresponding section of the manuscript).

Sample Collection

At least 10g of fresh fecal samples were collected from each subject in sterile fecal containers. The collected fecal samples were transported to the laboratory within 2 hours. All samples were stored at -80°C . Concurrently, all subjects provided at least 6mL of fasting peripheral blood, which was transported to the laboratory within 2 hours for centrifugation and serum retention, and then immediately stored in a -80°C low-temperature refrigerator. All samples used for virome and metabolome analysis were frozen for less than 1 month before processing.

Please refer to the supporting information for detailed information on serum metabolite detection, analysis data, virus-like particle enrichment, nucleic acid extraction and quality check, library preparation and sequencing, data assembly, identification and quantification of viral fragments, gene prediction, annotation and quantification, diversity analysis, metagenomic sequencing, and metagenomic data analysis, as well as statistical analysis. The overall technical route of this study is shown in [Figure 1](#).

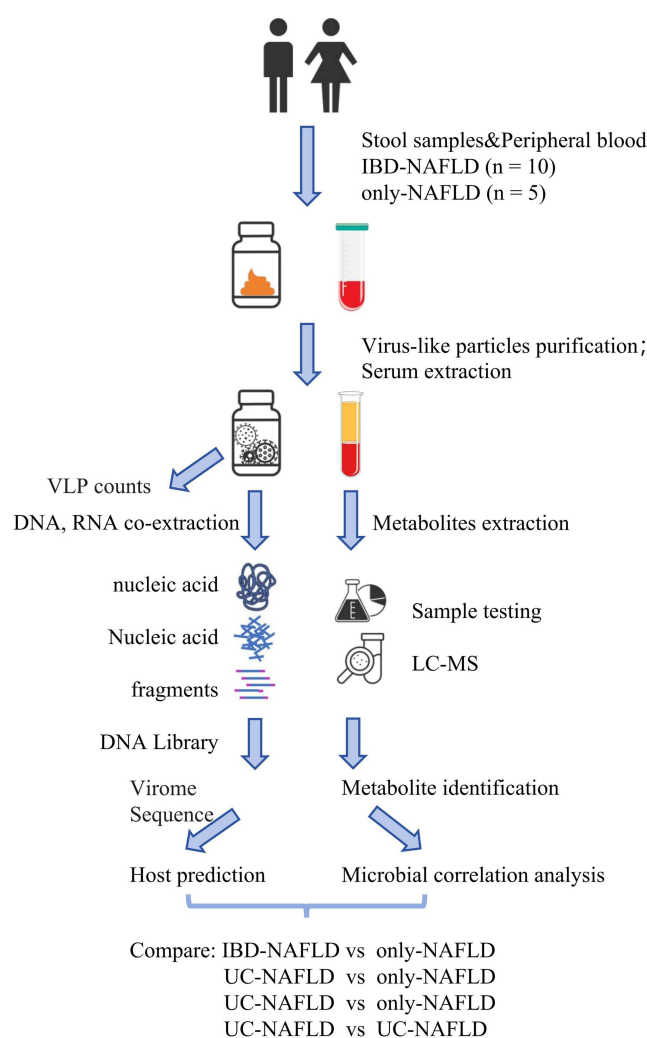


Figure 1 Overview of cohort characteristics and data analysis.

Serum Metabolite Extraction and Detection

This experiment was completed by Guangdong Mega Biotechnology Co., Ltd. ① Transfer 100μL of the sample into an EP tube, add 400μL of extraction solution (methanol: acetonitrile = 1:1 (V/V), containing a mixture of isotope-labeled internal standards), vortex mix for 30 seconds; ultrasonicate for 10 minutes (in an ice bath); stand at -40°C for 1 hour; centrifuge the sample at 4°C, 12000 rpm (centrifugal force 13800($\times g$), radius 8.6cm) for 15 minutes; ② Take the supernatant into the injection bottle for machine detection; all samples take an equal amount of supernatant mixed into a QC sample for machine detection. ③ Use an ultra-high-performance liquid chromatography system for chromatographic separation of the target. Use a ThermoFisher Q Exactive HF-X mass spectrometer for mass spectrometry data collection. ④ Use the ProteinLynx Global software to convert raw data and annotate with a mass spectrometry database after processing. ⑤ Differential metabolites are visualized in the form of volcano plots and radar plots. ⑥ By analyzing the enrichment pathways and topological structure of differential metabolites, the pathways most related to metabolite differences are determined and displayed in a rectangular dendrogram.

Virus Particle (VLPs) Enrichment and DNA Extraction

We profiled the enteric virome of 15 volunteers by coupling VLP enrichment with a dual-nucleic-acid workflow. After DNase/RNase treatment to minimise free nucleic acids, viral particles were lysed and DNA (dsDNA + ssDNA) and RNA

(ssRNA + dsRNA) were co-extracted.²³ Genome-type-specific libraries were prepared: RNA was reverse-transcribed with random hexamers, then both DNA and cDNA were sonicated to 300–500 bp and processed through the TruSeq Nano DNA protocol. Paired-end 150-bp sequencing was performed on an Illumina NovaSeq; reads were deposited as FASTQ files. Raw data were trimmed with Trimmomatic (SLIDINGWINDOW:4:20 MINLEN:50),²³ and host reads were subtracted by aligning to hg38 with BWA-MEM (default) plus SOAPaligner (-v 5 -r 2).^{24,25} Host-depleted, high-quality reads were assembled de novo for each sample with IDBA-UD (k 40–120),²⁶ metaSPAdes (—meta),²⁷ MEGAHIT (—k-list 21–141)²⁸ and Trinity (—seqType fq —SS_lib_type F).²⁹ Contigs ≥ 1 kb were retained; N50, L50 and cumulative length were calculated with QUAST, and read-recruitment rates were estimated by re-mapping cleaned reads to contigs using BWA-MEM. The assembly pipeline that maximised N50 while maintaining >80% read recovery was chosen for downstream analyses (details in Results).

Virus Sequence Identification and Classification Annotation

First, the CheckV³⁰ software was used to predict potential viral sequence sets in the assembled sequences, at the same time, the Virsorter2³¹ software was used to re-identify viruses in the original assembled sequences, identifying viral sequences from gene content and genomic structure characteristics, further confirming the identification results of CheckV, and screening high-confidence sequences as a supplement to the results of CheckV. After obtaining the above potential viral sequence sets, the PhaGCN2³² software was used for viral taxonomy annotation of the sequences. By combining the annotation results of PhaGCN2 and the annotation information of the target alignment sequences of CheckV in the database, the species information of the viral sequences was jointly confirmed. Finally, based on the virus identification method, confidence, and completeness information, the viral sequences were classified. For specific selection rules, please refer to the description in the results section.

Virus Abundance Statistics and Gene Prediction Analysis

Reads were aligned to the identified viral contigs to calculate the RPKM of each contig, Reads Per Kilobase of exon model per Million mapped reads (reads per thousand bases of the transcript per million mapped reads), for comparative analysis between samples. Prokka³³ was used to predict gene sequences on the viral contigs, evaluating the number and length of the predicted genes. The predicted protein sequences were compared with the viral sequences in the UniProtKB/Swiss-Prot/KEGG database (ViralZone,³⁴ reviewed proteins, <https://viralzone.expasy.org/>) to obtain functional annotation information.

Diversity Analysis and Statistical Analysis

Alpha diversity analysis of individual samples is used to reflect the abundance and diversity of viruses. Beta diversity analysis is employed to compare the differences in size among different samples across various groups, obtaining sample clustering heatmaps and sample PCA, PCoA plots based on distance matrices. Depending on the number of groups and repetitions, different statistical testing methods are selected (for details, refer to the results section), to analyze the differences in each species at various taxonomic levels between groups and identify biomarkers with statistical differences.

Virus Host Prediction

The identified viral sequences are analyzed using the CHERRY³⁵ software and PHP³⁶ software to predict the hosts and to statistically summarize the host species information. The best prediction results obtained from the CHERRY software are filtered by a score threshold to obtain species-level virus host prediction results. The PHP software is a prokaryotic virus host prediction tool based on a Gaussian Model, which calculates the host probability of 60,105 prokaryotic genomes to obtain the prokaryotic genome with the best matching score as the predicted host, and filters the results by setting a threshold to obtain virus host prediction results with higher credibility (up to the genus level).

Result

Patients' Characteristics

Thirty-two subjects (18 with IBD-NAFLD (MASLD) as the disease group; 14 with only-NAFLD (MASLD) matched as the control group) consented to participate in the study; ultimately, 15 subjects (5 with CD-NAFLD (MASLD), 5 with UC-NAFLD (MASLD); 5 with only-NAFLD (MASLD), controls) provided fecal and blood samples. Compared to the healthy group, there were no significant differences in the demographic or laboratory parameters of the disease group, including age, gender, weight, BMI, education level, aspartate aminotransferase (AST), alanine aminotransferase (ALT), alkaline phosphatase (AP), gamma-glutamyl transferase (GGT), bilirubin, and albumin. The results are shown in [Table 1](#).

3.2 Comprehensive analysis of the microbiome and microbiota-derived metabolites in IBD-NAFLD (MASLD) and only-NAFLD (MASLD) patients.

Screening of Differential Metabolites

The cutoff value used in this project was a P-value from the Student's *t*-test of less than or equal to 0.05, along with a Variable Importance in the Projection (VIP) score of greater than or equal to 1 from the OPLS-DA model's first principal component, and the absolute value of the logarithm (base 2) of the fold change was greater than or equal to 0. The results of the differential metabolite screening are described in [Supplementary Table 1](#). We visualized the results of the differential metabolite screening in the form of a volcano plot and bar chart of differential metabolite fold differences, as shown in [Figure 2](#).

Correlation Analysis of Differential Metabolites

To investigate the correlation between specific metabolites and the relative abundance of differential microbial species, we calculated the Pearson/Spearman correlation coefficients and P-values between metabolites and microbial communities to identify the relationships between particular metabolites and microbial populations. There were differential metabolites between IBD-NAFLD (MASLD) (including UC and CD) and only-NAFLD (MASLD), with significant relative abundance; additionally, we conducted pairwise comparisons between groups, and the results showed the presence of differential metabolites with significant relative abundance. As shown in [Figure 3A](#).

Table 1 Anthropometric and Clinical Characteristics of the Study Groups

Variables	UC-NAFLD (n=5)	CD-NAFLD (n=5)	Only-NAFLD (n=5)	P
Sex (M/F)	5(2/3)	5(3/2)	5(3/2)	—
Age (years, $\bar{x} \pm s$)	49.00 $\pm$ 13.62	47.80 $\pm$ 15.51	52.80 $\pm$ 5.71	0.839
Education (years, $\bar{x} \pm s$)	12.40 $\pm$ 3.14	13.00 $\pm$ 1.55	12.20 $\pm$ 2.23	0.887
Weight (kg, $\bar{x} \pm s$)	65.40 $\pm$ 8.40	64.00 $\pm$ 8.29	69.80 $\pm$ 9.62	0.634
BMI ($\bar{x} \pm s$)	22.68 $\pm$ 2.21	22.08 $\pm$ 1.39	24.92 $\pm$ 2.16	0.138
Albumin (g/L, $\bar{x} \pm s$)	39.76 $\pm$ 1.59	41.34 $\pm$ 7.74	42.30 $\pm$ 3.99	0.781
Tot. Bilirubine (mg/dL, $\bar{x} \pm s$)	9.28 $\pm$ 3.39	9.52 $\pm$ 4.19	9.98 $\pm$ 6.62	0.979
Dir. Bilirubine (mg/dL, $\bar{x} \pm s$)	4.4 $\pm$ 1.34	3.84 $\pm$ 1.17	3.90 $\pm$ 2.35	0.879
ALT (U/L, $\bar{x} \pm s$)	11.00 $\pm$ 7.07	19.52 $\pm$ 6.77	14.18 $\pm$ 7.99	0.286
AST (U/L, $\bar{x} \pm s$)	16.20 $\pm$ 4.83	18.34 $\pm$ 2.50	16.38 $\pm$ 3.94	0.696
GGT (U/L, $\bar{x} \pm s$)	15.60 $\pm$ 5.24	22.00 $\pm$ 16.80	18.20 $\pm$ 13.88	0.785
AP (U/L, $\bar{x} \pm s$)	77.80 $\pm$ 16.79	79.80 $\pm$ 46.92	73.20 $\pm$ 27.31	0.958
Tot. Cholesterol (mg/dL, $\bar{x} \pm s$)	3.48 $\pm$ 0.17	4.04 $\pm$ 0.64	4.56 $\pm$ 0.80	0.075

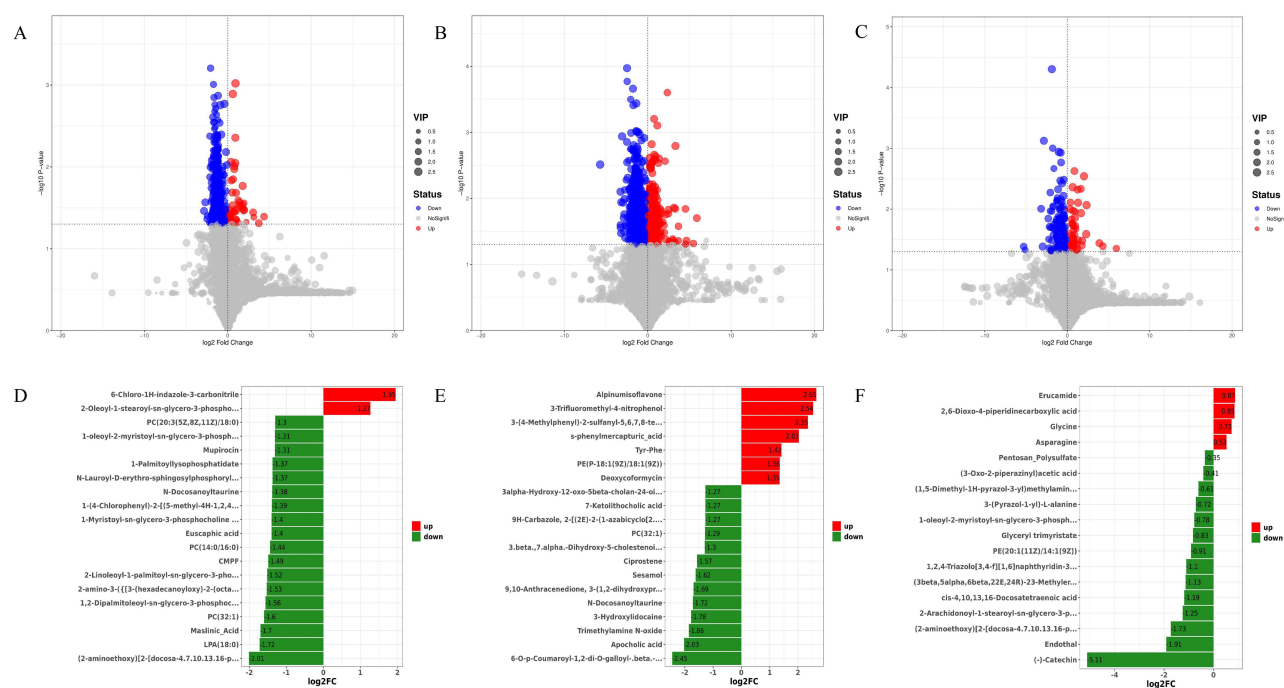


Figure 2 Volcano plot and bar chart of differential metabolite fold differences for the screening of differential metabolites among the three groups. Each point in the volcano plot represents a metabolite, the horizontal axis represents the fold change of each substance in the comparison between groups, and the vertical axis indicates the P-value of the Student's t-test. The size of the dots represents the VIP (Variable Importance in the Projection) value of the OPLS-DA model; the larger the dot, the higher the VIP value. The color of the dots represents the final selection results, with significantly upregulated metabolites shown in red, significantly downregulated metabolites shown in blue, and non-significant metabolites in grey. (A) Volcano plot for group UC-NAFLD (MASLD) vs only-NAFLD (MASLD); (B) Volcano plot for group CD-NAFLD (MASLD) vs only-NAFLD (MASLD); (C) Volcano plot for group UC-NAFLD (MASLD) vs CD-NAFLD (MASLD). D-E: The horizontal axis represents the log2FC value, the vertical axis represents differential metabolite abundance, red represents upregulation of differential metabolite abundance, and green represents downregulation of differential metabolite abundance. (D) bar chart of differential metabolite fold for group UC-NAFLD (MASLD) vs only-NAFLD (MASLD); (E) bar chart of differential metabolite fold for group CD-NAFLD (MASLD) vs only-NAFLD (MASLD); (F) bar chart of differential metabolite fold for group UC-NAFLD (MASLD) vs CD-NAFLD (MASLD).

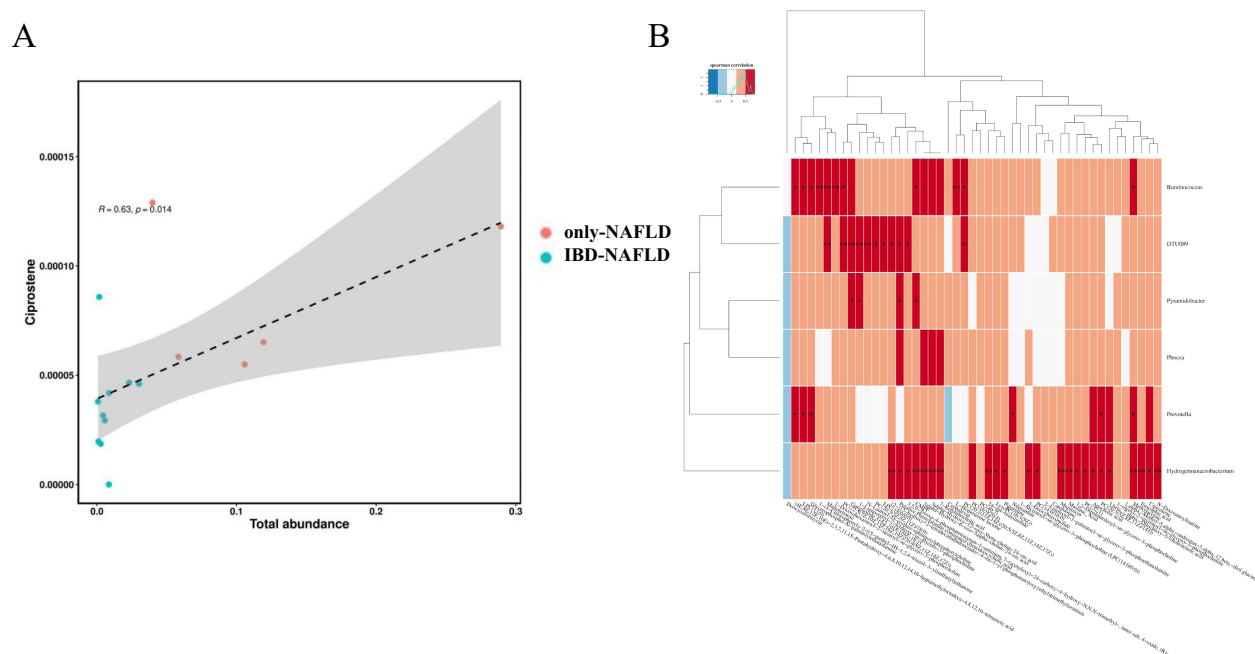


Figure 3 Correlation analysis of differential metabolites. (A) Scatter plot analysis. (B) heatmap for group IBD-NAFLD (MASLD) vs only-NAFLD (MASLD). (A): The x-axis represents the total abundance value of differential bacterial species relative abundance, and the y-axis represents the relative abundance of differential metabolites; (B) Below are differential metabolites, and on the right are differential species. The color depth indicates the magnitude of correlation, blue indicates negative correlation, red indicates positive correlation, and p-value is the correlation test result. In the figure, * represents $p < 0.05$, ** represents $p < 0.01$.

The differential metabolites obtained from the above analysis often have similar/biological results and functions or are complementary, or they are regulated by the same metabolic pathways, either positively or negatively, manifesting as similar or opposite expression patterns across different experimental groups. We further performed hierarchical cluster analysis on these characteristics, grouping metabolites with the same features together, and identified the variation patterns of metabolites among groups. For each comparison, we calculated the Euclidean distance matrix for the quantitative values of differential metabolites, clustered the differential metabolites using the complete linkage method, and displayed them in a heatmap. As shown in [Figure 3B](#). Concurrently, the correlation heatmap analysis indicates that there are significant differences in differential metabolites among various differential species between IBD-NAFLD (MASLD) and only-NAFLD (MASLD). Most show a positive correlation, while the metabolic difference of Deoxycorymycin shows a negative correlation trend.

Metabolic Pathway Analysis of Differential Metabolites

Through KEGG annotation analysis, we identified all the pathways involved in the differential metabolites, and further performed enrichment analysis and topological analysis to screen the key pathways most relevant to the metabolite differences in IBD-NAFLD (MASLD) patients: the galactose metabolism pathway ($P<0.001$, enrichment score=6.23). The key differential metabolites enriched in this pathway include galactose-1-phosphate, UDP-galactose, and galactitol, which were all significantly down-regulated in the IBD-NAFLD (MASLD) group ($P<0.05$). The results of the metabolic pathway analysis between IBD-NAFLD (MASLD) and only-NAFLD (MASLD) are presented in a bubble plot and a treemap plot, as shown in [Figure 4](#).

Differences in Fecal Virome Composition Between IBD-NAFLD (MASLD) and Only-NAFLD (MASLD) Patients

Identification and Taxonomic Annotation of Viral Sequences

Viral sequences identified by CheckV and Virsorter2 software were classified into different taxonomic labels based on identification methods and completeness, and all viral sequences were extracted. For proviral sequences, the host parts of the sequences were removed. The statistical information of the viral sequences is presented in [Table 2](#). CheckV confirms

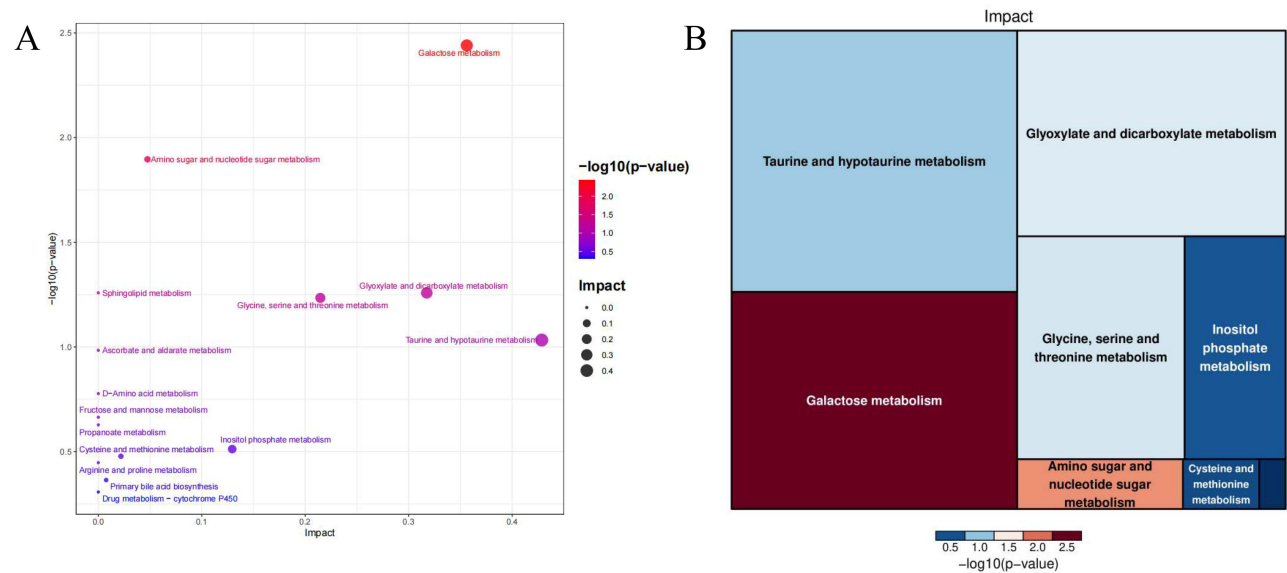


Figure 4 Pathway analysis comparing IBD-NAFLD (MASLD) and only-NAFLD (MASLD) groups. **(A)** Bubble plot. Each bubble in the chart represents a metabolic pathway, with the horizontal axis and the size of the bubble indicating the magnitude of the influencing factor of the pathway in topological analysis. The larger the size, the greater the influencing factor; the vertical axis on which the bubble is positioned and the color of the bubble represent the P-value of the enrichment analysis (using the negative logarithm with base 10). The redder the color, the smaller the P-value, indicating a more significant enrichment. **(B)** Treemap plot. Each block in the diagram represents a metabolic pathway, with the size of the block indicating the magnitude of the influencing factor of the pathway in topological analysis. The larger the size, the greater the influencing factor; the color of the block represents the P-value of the enrichment analysis (using the negative logarithm with base 10), and the darker the color, the smaller the P-value and the more significant the enrichment degree.

Table 2 Virus Sequence Statistics

Sample	Total_Bases	Total_num	Max_len	Min_len	N50	GC_Content (%)
MergedSample	59,656,035	27,933	134,702	300	4,843	46.32

Notes: (1) Sample: The sample name is usually the set of sequences that have been merged to remove redundancy; (2) Total_bases: The total number of contigs bases in the virus; (3) Total_num: Total number of virus contigs; (4) Max_len: The maximum length of virus contigs; (5) Min_len: The minimum length of virus contigs; (6) N50: N50 length of virus contigs; (7) GC_content (%): The average GC content of viral contigs.

the integrity of viral sequences by recognizing three viral characteristic structures, including direct terminal repeats (DTR), inverted terminal repeats (ITR), and AAI method-based identification. The statistical results of sequences with complete viral genome structures are shown in Table 3. The annotation information determines whether the viral sequences are phages, and the statistical results are provided in Table 4. Additionally, based on the taxonomic information, the genome type of the virus can be confirmed (dsDNA, ssDNA, dsRNA, ssRNA, RT, or unable to determine the genome type), as shown in Table 5. The statistics at each taxonomic level (up to the genus level) are presented in Table 6.

Table 3 Complete Structure Statistics of Viruses

Sample	Total	AAI_based_num	DTR_based_num	ITR_based_num
MergedSample	177	79	94	4

Notes: (1) Sample: The name of the sample, usually refers to the set of sequences obtained after deduplication of merged samples; (2) Total: The sum of sequences with complete structures; (3) AAI_based_num: The number of contigs determined based on the AAI method; (4) DTR_based_num: The number of contigs with DTR (direct terminal repeats) structures; (5) ITR_based_num: The number of contigs with ITR (inverted terminal repeats) structures.

Table 4 Phage Identification Statistics

Sample	Total	Phage (%)	Other_Viruses (%)	Unassigned (%)
MergedSample	27933	24,498(87.7%)	1197(4.29%)	2238(8.01%)

Notes: (1) Sample: The name of the sample; (2) Total: The total number of viral contigs in the sample; (3) Phage: The number of contigs identified as phages in the sample; (4) Other_virus: The number of contigs identified as non-phage viruses in the sample; (5) Unassigned: The number of viral contigs in the sample that could not be annotated to any taxonomic classification.

Table 5 Statistical Analysis of the Types of Virus Genomes Identified

Sample	Total	dsDNA (%)	ssDNA (%)	dsRNA (%)	ssRNA (%)	RT (%)	Unassigned (%)
MergedSample	27933	24,736(88.55%)	918(3.29%)	1(0%)	24(0.09%)	16(0.06%)	2238(8.01%)

Notes: (1) Sample: The name of the sample; (2) Total: The total number of viral sequences; (3) dsDNA: The number of contigs identified as double-stranded DNA viruses; (4) ssDNA: The number of contigs identified as single-stranded DNA viruses; (5) dsRNA: The number of contigs identified as double-stranded RNA viruses; (6) ssRNA: The number of contigs identified as single-stranded RNA viruses; (7) RT: The number of contigs identified as retroviruses; (8) Unassigned: The number of viral sequences that cannot confirm the type of genome (including parts with non-unique genome types).

Table 6 Statistics of Different Classification Levels for Virus Sequence Identification

Sample	Total	Level_D	Level_P	Level_C	Level_O	Level_F	Level_G	Unknown
MergedSample	27933	27,933	25,695	25,689	2725	2536	320	0

Notes: (1) Sample: The name of the sample; (2) Total: The total number of viral contigs in the sample; (3) Level D: The number of viral contigs classified at the domain level; (4) Level P: The number of viral contigs classified at the phylum level; (5) Level C: The number of viral contigs classified at the class level; (6) Level O: The number of viral contigs classified at the order level; (7) Level F: The number of viral contigs classified at the family level; (8) Level G: The number of viral contigs classified at the genus level; (9) Unknown: The number of viral contigs that could not be annotated with a species classification.

Viral Abundance and Gene Prediction

Based on the aforementioned results of viral sequence identification and taxonomic annotation, viral sequences with genome types of RNA or RT were filtered out. Using the filtered viral sequences, BWA (v0.7.17, default parameters: mem-k 30) software was employed to map the cleaned reads, free from contamination, against the obtained viral contigs. Mapping results with lengths below 80% of the total read length were filtered out, and the proportion of viral reads was statistically analyzed. The RPKM values for each viral contig were calculated; based on the annotation results of viral contigs, the distribution of viral reads was statistically analyzed to obtain a heatmap of viral distribution and a percentage chart of RPKM values. There were significant differences in viral sequence abundance between the Only-NAFLD (MASLD), CD-NAFLD (MASLD), and UC-NAFLD (MASLD) groups, as shown in Figure 5. At the phylum level, the abundance of Uroviricota (tailed bacteriophages) in the IBD-NAFLD (MASLD) group was significantly higher than in the Only-NAFLD (MASLD) group; at the class level, the abundance of Caudovirales (tailed bacteriophages) in the IBD-NAFLD (MASLD) group was significantly higher than in the Only-NAFLD (MASLD) group, as shown in Figure 6.

The Prokka (v1.13) software was used for gene prediction on viral contigs, filtering out contigs sequences with gene nucleotide lengths less than 200bp. The statistical results are presented in Table 7.

Viral KEGG Functional Annotation and Diversity Analysis

Kyoto Encyclopedia of Genes and Genomes(KEGG) is a practical database resource for understanding high-level functions and biological systems, particularly from molecular-level information such as large-scale molecular datasets generated by genomic sequencing and other high-throughput experimental technologies. According to the KEGG annotation results, bar charts of the number of genes annotated at Level 1, Level 2, and Level 3 are drawn for each group of samples. The results are shown in Figure 7.

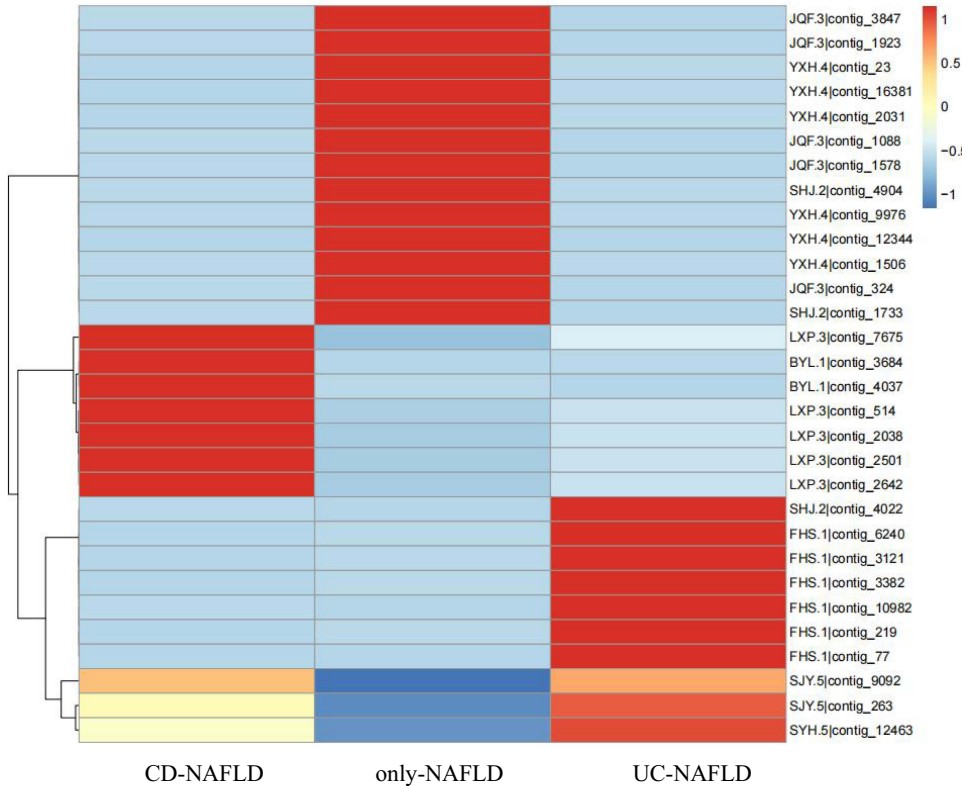


Figure 5 Heatmap is used to represent virus abundance. Note: Select the viral sequences with the top 30 abundances; The horizontal axis represents the groups, and the vertical axis represents the sequence numbers.

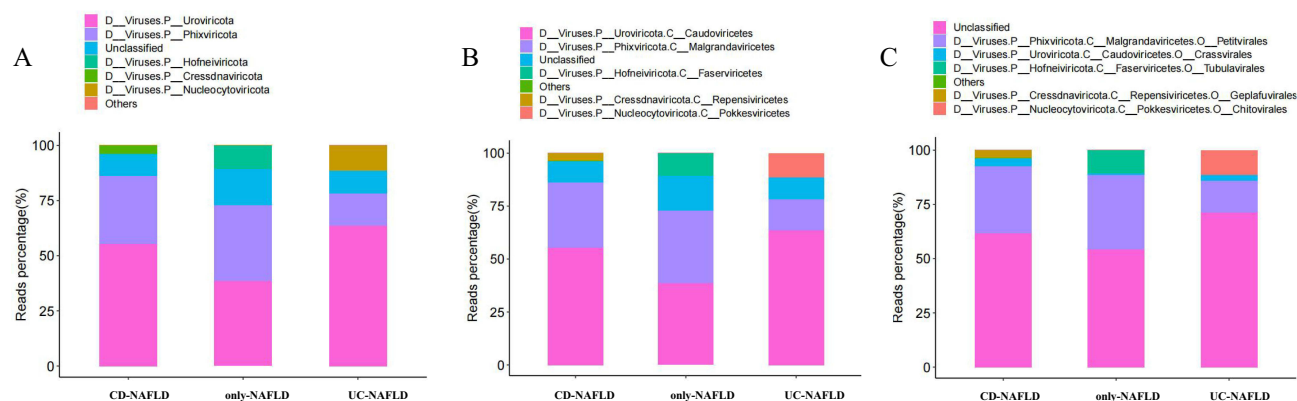


Figure 6 Virus abundance statistics (RPKM Value Percentage Statistical Chart). **(A)** Phylum Level. **(B)** Class Level. **(C)** Order Level.

We used alpha diversity analysis, which reflects the abundance and diversity of environmental viruses, to show no significant differences between IBD-NAFLD (MASLD) and Only-NAFLD (MASLD). As shown in Figure 8A. We employed a two-dimensional Principal Component Analysis (PCA) plot to illustrate the beta diversity among samples. As shown in Figure 8B.

Statistical Analysis

Linear discriminant analysis Effect Size (LEfSe) was utilized for comparative analysis between groups to identify viral contigs (ie., biomarkers) that exhibit significant differences in abundance between groups. This analysis is based on the RPKM values of viral sequences from individual samples. Viruses with significantly different abundances between the CD-NAFLD (MASLD) and only-NAFLD (MASLD) groups are presented in Figure 9.

Virus Host Prediction

Based on the host prediction results from the CHERRY software, species information at the domain, phylum, class, order, family, genus, and species taxonomic levels were statistically analyzed. The sequences were sorted according to the host species information annotated at the family and genus levels, and the top 10 host species at the species level are presented. The first on the list is *Bacteroides fragilis*; followed by *Parabacteroides merdae*. As shown in Figure 10. To further verify the reliability of the host prediction, we also used the PHP software for host prediction validation; species information at six taxonomic levels was statistically analyzed. The top-ranked genus is *Bacteroides*, which is consistent with the previous results. As shown in Figure 10.

Discussion

IBD is a global chronic inflammatory disease with high rates of extraintestinal manifestations, among which hepatobiliary manifestations are the most common, and NAFLD (MASLD) accounts for about 40% of liver abnormalities in IBD patients.³⁷ The global prevalence of NAFLD (MASLD) is about 30.2%, and the prevalence in IBD patients is as high as 24.4%, nearly twice that of the healthy population.^{37,38} Previous studies have confirmed that gut bacterial dysbiosis is involved in the pathogenesis of both IBD and NAFLD (MASLD), but the role of the gut virome—an important

Table 7 Statistics of Gene Prediction Results

#Sample	Total_base (Mb)	Total_num	Max_len	Min_len	N50	GC
Viral.contig.ffn.n50	43.43	74,532	16,611	90	792	47.26%

Notes: (1) Total_base: Total number of sequence bases; (2) Total_num: Total number of sequences; (3) Max_len: Maximum length; (4) Min_len: Minimum length; (5) N50: N50 value; (6) GC: Average GC content.

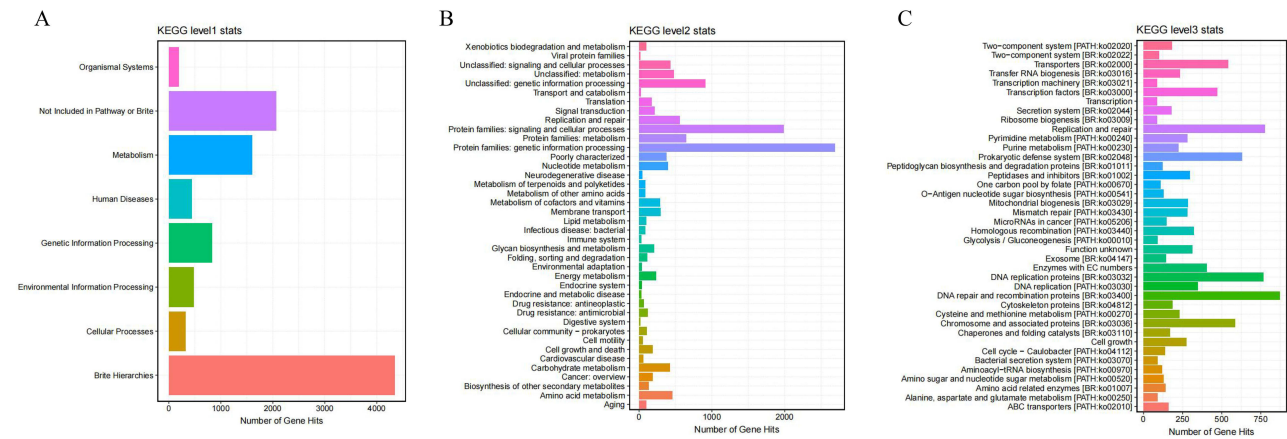


Figure 7 Column charts of Unigene gene numbers at different KEGG annotation levels. (A) Number of genes annotated at KEGG Level 1 for each sample; (B) Number of genes annotated at KEGG Level 2 for each sample; (C) Number of genes annotated at KEGG Level 3 for each sample. All charts were generated based on the original KEGG annotation results of Unigenes.

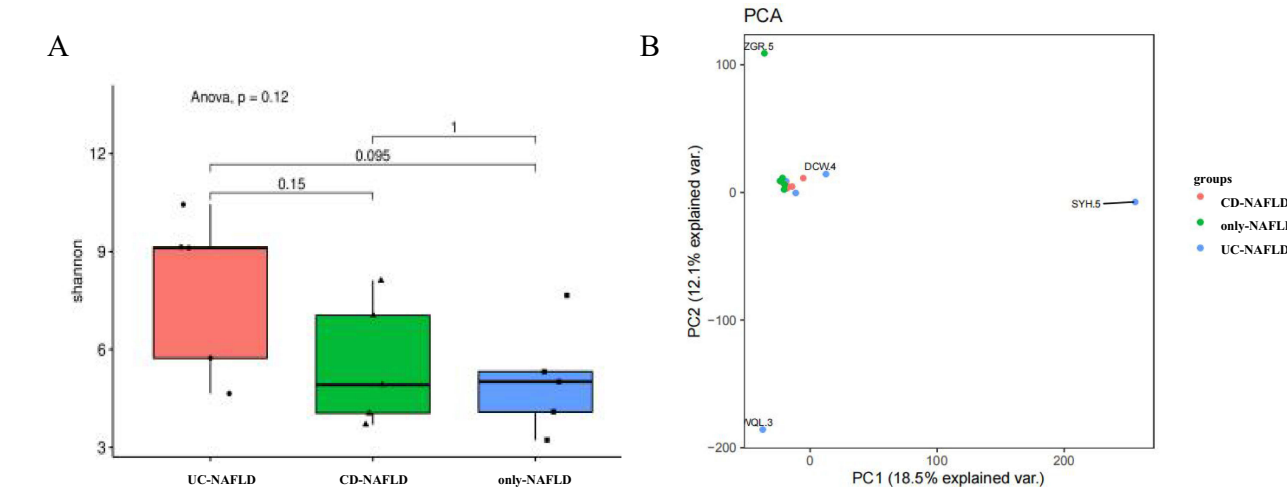


Figure 8 Diversity analysis. (A) Shannon exponential grouping boxplot. (B) PCA analytic result. (1) a point represents a sample, and the spatial distance between the point and the point indicates the difference in the species composition structure; (2) the principal component 1 (PC1) and principal component 2 (PC2) are the first and second largest eigenvalues that cause the difference between samples, and the percentage represents the variance contribution of the two principal components to the difference in the sample).

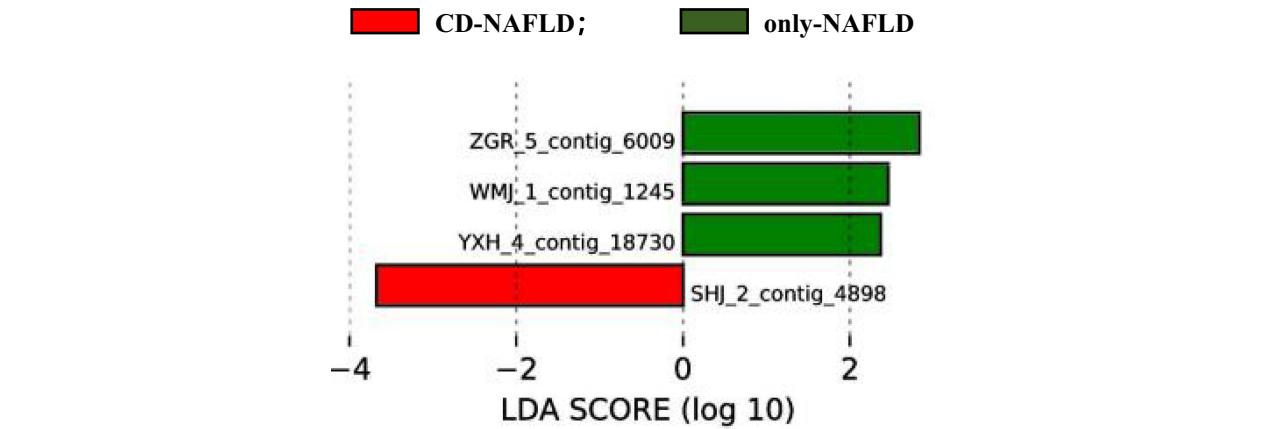


Figure 9 LDA (Linear discriminant analysis) graph. The figure represents viruses with significantly different abundances across different groups under the condition of an LDA value greater than the set threshold (default set to 2), with the length of the bars representing the magnitude of the effect of the differential viruses (i.e., the LDA Score).

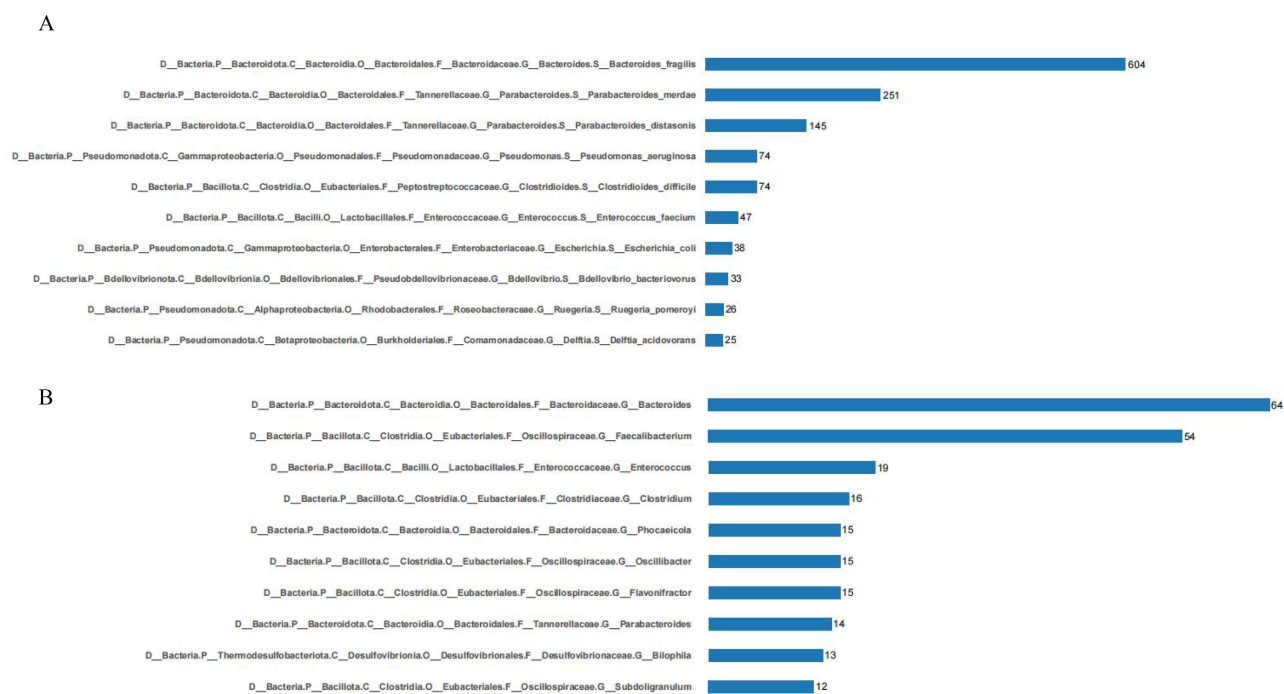


Figure 10 Classification and statistical analysis of the top 10 host species in the IBD-NAFLD (MASLD) and only-NAFLD (MASLD) groups. **(A)** Classification and statistical analysis at the species level; **(B)** Classification and statistical analysis at the genus level.

component of the gut microbiota—remains unclear. The main technical challenges in gut virome research include the lack of systematic cross-validation of detection technologies and the incompleteness of viral reference databases.²⁰ Currently, the limitations faced in exploring the role of the gut virome in specific diseases mainly stem from two technical challenges. On the one hand, techniques for detecting viral communities can mostly only use VLP or large-scale metagenomics methods, which lack systematic cross-validation. On the other hand, despite the establishment of several large viral genomic reference databases, many unknown viral sequence repositories in the human gut²⁰ remain undiscovered. The analysis of the gut virome heavily relies on these reference databases, significantly narrowing the scope of viral research.

In this study, we attempted to overcome these challenges by combining VLP and large-scale metagenomic sequencing technologies to conduct an in-depth study of the “full gut virome” associated with IBD-NAFLD (MASLD) in a set of samples. It is noteworthy that these two sequencing technologies demonstrated different capabilities in capturing viruses. Consistent with previous assumptions,^{39,40} VLP sequencing tends to capture free viral particles, especially those from the Microviridae family, while metagenomic sequencing is better at reconstructing viral sequences that are partially integrated into bacterial hosts. Additionally, we constructed a reference gut virome catalog based on high-throughput sequencing samples, representing over 20,000 non-redundant vOTUs. In this study, 8.01% of the vOTUs were not found in existing gut virome catalogs (ie., the Gut Virome Database (GVD),⁴¹ the Gut Phage Database (GPD),⁴² or the Metagenomic Gut Viruses (MGV)⁴³) (Table 4). These results indicate that our catalog complements the deficiencies of previous viral reference literature, enabling a more comprehensive examination of the IBD-NAFLD (MASLD) gut virome. Overall, our study emphasizes the importance of utilizing complementary sequencing techniques and viral reference databases to fully explore the diversity of viruses within the human gut. This research provides a model for future virome studies in other related diseases.

We characterized the gut virome in patients with comorbid IBD and NAFLD (MASLD) and examined its relationship with serum metabolic profiles. The IBD-NAFLD (MASLD) group showed enrichment of Caudovirales, with Bacteroides identified as the core predicted viral host. Galactose metabolism emerged as the most significantly altered metabolic pathway. These findings point to a potential connection between gut virome changes and gut-liver axis dysfunction in this patient population.

Our results showed that the gut virome of IBD-NAFLD (MASLD) patients was significantly different from that of NAFLD (MASLD)-only patients, with the most prominent feature being the significant increase in the abundance of Uroviricota (phylum) and Caudovirales (order). Caudovirales are the most abundant tailed bacteriophages in the human gut, and their abundance changes are closely related to intestinal inflammation.^{44,45} Previous studies have found that the abundance of Caudovirales is increased in IBD patients,⁵ and our study further confirmed that this characteristic is more significant in IBD patients with comorbid NAFLD (MASLD), suggesting that the enrichment of Caudovirales may be a specific viral signature of the IBD-NAFLD (MASLD) comorbidity. In contrast, the NAFLD (MASLD)-only group had higher abundances of Microviridae and Inoviridae, which are typical non-tailed phages and are considered to be associated with intestinal microbial homeostasis.⁴⁵ The imbalance of phage communities may lead to the disruption of intestinal bacterial homeostasis, which in turn affects the gut-liver axis and promotes the occurrence of NAFLD (MASLD) in IBD patients.

Bacteroides was identified as the top predicted viral host in this study, especially *Bacteroides fragilis*. *Bacteroides* is the core genus of the human gut microbiota, and its metabolic activity and interaction with phages are key to maintaining intestinal homeostasis.⁴² *Bacteroides fragilis* can be divided into non-toxigenic (NTBF) and enterotoxigenic (ETBF) strains; ETBF is associated with intestinal inflammation, while NTBF plays an anti-inflammatory role by promoting the differentiation of Treg cells and secreting IL-10.^{46,47} Phages can directly regulate the abundance of their host bacteria—the enrichment of Caudovirales in IBD-NAFLD (MASLD) patients may lead to the imbalance of *Bacteroides* strains (eg., the increase of ETBF/NTBF ratio), which not only exacerbates intestinal inflammation but also affects the metabolism of bile acids and carbohydrates by *Bacteroides*,^{48,49} and further induces hepatic fat accumulation through the gut-liver axis. This is the first report of *Bacteroides* as the core viral host in IBD-NAFLD (MASLD) patients, which provides a new target for the regulation of the gut virome-bacteriome interaction in this comorbid population.

KEGG pathway analysis showed that the differential metabolites between the two groups were most significantly enriched in the galactose metabolism pathway, and key metabolites (galactose-1-phosphate, UDP-galactose, galactitol) in this pathway were significantly down-regulated in IBD-NAFLD (MASLD) patients. Galactose metabolism is an important part of carbohydrate metabolism, and its disorder is closely related to hepatic steatosis and intestinal barrier damage.⁵⁰ The down-regulation of galactose metabolism-related metabolites may be caused by the imbalance of *Bacteroides* (the main galactose-metabolizing bacteria in the gut) induced by virome alterations. In addition, we found that 17 differential metabolites were significantly correlated with the relative abundance of key viral species, which directly confirmed the correlation between gut virome alterations and serum metabolic disorders in IBD-NAFLD (MASLD) patients. This finding reveals the potential mechanism of the gut virome regulating the gut-liver axis through metabolic pathways, and provides a novel research direction for the intervention of IBD-NAFLD (MASLD) from the perspective of “virome-metabolism”.

This study has several limitations. The sample size of our study is relatively small, but it is the first study on the gut virome in IBD-NAFLD (MASLD), which helps provide a reference basis for future related virome research. The lack of an IBD-only control group makes it impossible to further distinguish the virome characteristics specific to the comorbidity from those of IBD itself. We only analyzed fecal and blood samples; however, the comparison of tissue samples might allow for a more sensitive detection of some viral lineages. Although our subjects were included in the study at the time of their initial diagnosis of IBD-NAFLD (MASLD), there was a time lapse between the onset of symptoms and sample collection, so any initial environmental triggers may have been significantly reduced or obscured by subsequent treatments. Inevitably, the patient population is heterogeneous, with interference from environmental exposure factors. While the study results support the notion of significant differences in the gut virome between only-NAFLD (MASLD) and IBD-NAFLD (MASLD), offering new insights into the mechanisms of different diseases, the development of health and disease markers, and more effective diagnostic and treatment methods, it is still unclear whether changes in the gut virome are a result of the disease or a cause of the disease. Perhaps further animal experiments could help strengthen the interpretation of this issue, and the complex relationships between diseases still require further exploration.

In future studies, we will expand the sample size, add multi-center and multi-ethnic subjects, set up an IBD-only control group, and conduct a prospective cohort study to verify the causal relationship between the gut virome and IBD-NAFLD (MASLD). In addition, animal experiments will be used to explore the mechanistic role of Caudovirales and

Bacteroides in the pathogenesis of IBD-NAFLD (MASLD), and to provide experimental basis for the development of phage-based targeted therapy.

Conclusion

This study characterized the gut virome profile specific to patients with both IBD and NAFLD (MASLD). We found elevated abundance of Caudovirales in the fecal virome of these patients and identified Bacteroides as the core predicted viral host. The analysis also revealed a significant association between gut virome alterations and galactose metabolism disturbances, suggesting a potential connection between the gut virome and the gut-liver axis in this comorbid condition. These findings offer viral and metabolic signatures that may inform future research into the pathophysiology of IBD-NAFLD (MASLD) overlap, though large-scale prospective and multi-center studies are required to validate these observations and establish causal relationships.

Data Sharing Statement

The data are available from the corresponding author at hj042153@hotmail.com.

Ethics Statement and Isolate Source

All clinical isolates used in this study were additionally collected—not residual or discarded—from patients attending The Second People's Hospital of Changzhou, The Third Affiliated Hospital of Nanjing Medical University, Jiangsu, China. The study protocol was reviewed and approved by the ethics committees of The Second People's Hospital of Changzhou, The Third Affiliated Hospital of Nanjing Medical University (approval no. [2024]KY204-01). Written informed consent was obtained from every participant before sample collection. All procedures were performed in accordance with the Declaration of Helsinki, relevant Chinese regulations, and the STROBE guidelines.

The patients agreed to all management. Written informed consent for examination and operation was obtained from the patient or close relatives.

Informed Consent

All relevant data were obtained during the period of hospitalization. They provided informed consent for themselves or through an authorized representative.

Acknowledgments

We thank the Endoscopy Centre of The Second People's Hospital of Changzhou, the Third Affiliated Hospital of Nanjing Medical University for technical and logistical support.

Author Contributions

Conception and design: SS Lu, WJ Liu, J Huang

Data acquisition: Y Xia, RY Chen, HS Jin, JX Zhang

Data analysis and interpretation: SS Lu, Y Xia, QY Sun, YH Sun, RY Chen, HS Jin, JX Zhang

Software and statistical support: SS Lu, Y Xia, Y Xia, QY Sun, YH Sun, RY Chen, HS Jin, JX Zhang

Drafting of the manuscript: SS Lu, Y Xia

Critical revision of the manuscript for important intellectual content: WJ Liu, J Huang

Supervision and project administration: SS Lu, WJ Liu, J Huang

Final approval of the version to be published: all authors

Accountability for all aspects of the work: all authors

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

Supported by The Key Project of Changzhou Medical Center of Nanjing Medical University, No. CMCM202310; and Science and Technology Development Fund of Nanjing Medical University, No. NMUB20220196; and Changzhou Municipal Health Commission Science and Technology Project, No. QN202447.

Disclosure

All the authors report no relevant conflicts of interest for this article.

References

- Gros B, Kaplan GG. Ulcerative colitis in adults: a review. *JAMA*. 2023;330(10):951–965. doi:10.1001/jama.2023.15389
- Kaplan GG. The global burden of IBD: from 2015 to 2025. *Nat Rev Gastroenterol Hepatol*. 2015;12(12):720–727. doi:10.1038/nrgastro.2015.150
- Yan XX, Wu D. Intestinal microecology-based treatment for inflammatory bowel disease: progress and prospects. *World J Clin Cases*. 2023;11(1):47–56. doi:10.12998/wjcc.v11.i1.47
- Qiu P, Ishimoto T, Fu L, et al. The gut microbiota in inflammatory bowel disease. *Front Cell Infect Microbiol*. 2022;12:733992. doi:10.3389/fcimb.2022.733992
- Daan J, Jelle M. The emerging role of the gut virome in health and inflammatory bowel disease: challenges, covariates and a viral imbalance. *Viruses*. 2023;15(1). doi:10.3390/v1501017
- Reyes A, Semekovich NP, Whiteson K, et al. Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat Rev Microbiol*. 2012;10(9):607–617. doi:10.1038/nrmicro2853
- Deng L, Wojciech L, Png CW, et al. Colonization with two different *Blastocystis* subtypes in DSS-induced colitis mice is associated with strikingly different microbiome and pathological features. *Theranostics*. 2023;13(3):1165–1179. doi:10.7150/thno.81583
- Eslam M, Newsome Philip N, Sarin Shiv K, et al. A new definition for metabolic dysfunction-associated fatty liver disease: an international expert consensus statement. *J Hepatol*. 2020;73:202–209. doi:10.1016/j.jhep.2020.03.039
- Kiarash R, Hassan A, Jacob H, et al. The prevalence and incidence of NAFLD worldwide: a systematic review and meta-analysis. *Lancet Gastroenterol Hepatol*. 2022;7.
- Sonja L, Münevver D, Anna M, et al. Intestinal virome signature associated with severity of nonalcoholic fatty liver disease. *Gastroenterology*. 2020;159:1839–1852.
- Alireza BM, Amin SM, Bahareh S, et al. Prevalence of hepatobiliary manifestations in inflammatory bowel disease: a GRADE assessed systematic review and meta-analysis of more than 1.7 million patients. *J Crohns Colitis*. 2024;18:360–374. doi:10.1093/ecco-jcc/jjad157
- Pilar N, Lucia G-R, Antonio T-M, et al. Systematic review and meta-analysis: prevalence of non-alcoholic fatty liver disease and liver fibrosis in patients with inflammatory bowel disease. *Nutrients*. 2023;15:undefined. doi:10.3390/nu15214507
- Mancina RM, De Bonis D, Pagnotta R, et al. Ulcerative colitis as an independent risk factor for hepatic steatosis. *Gastroenterol Nurs*. 2020;43:292–297. doi:10.1097/SGA.0000000000000461
- Gizard E, C A, Ford J-P, et al. Systematic review: the epidemiology of the hepatobiliary manifestations in patients with inflammatory bowel disease. *Aliment Pharmacol Ther*. 2014;40. doi:10.1111/apt.12794
- Kerri G, Malaty BP, Abraham BP. Epidemiology and risk factors of nonalcoholic fatty liver disease among patients with inflammatory bowel disease. *Inflamm Bowel Dis*. 2017;23. doi:10.1097/MIB.0000000000001085
- Manik A, Rajat G, Gopandhan P, et al. Crohn's disease is associated with liver fibrosis in patients with nonalcoholic fatty liver disease. *Dig Dis Sci*. 2022;68. doi:10.1007/s10620-022-07562-0
- Wei-Lin W, Shao-Yan X, Zhi-Gang R, et al. Application of metagenomics in the human gut microbiome. *World J Gastroenterol*. 2015;21:803–814. doi:10.3748/wjg.v21.i3.803
- Xiangge T, Shenghui L, Chao W, et al. Gut virome-wide association analysis identifies cross-population viral signatures for inflammatory bowel disease. *Microbiome*. 2024;12:130. doi:10.1186/s40168-024-01832-x
- Norman Jason M, Handley Scott A, Baldrige Megan T, et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell*. 2015;160:447–460. doi:10.1016/j.cell.2015.01.002
- Junhua L, Fangming Y, Minfeng X, et al. Advances and challenges in cataloging the human gut virome. *Cell Host Microbe*. 2022;30:908–916. doi:10.1016/j.chom.2022.06.003
- Christian M, Andreas S, Vavricka Stephan R, et al. ECCO-ESGAR guideline for diagnostic assessment in IBD part 1: initial diagnosis, monitoring of known IBD, detection of complications. *J Crohns Colitis*. 2019;13:144–164. doi:10.1093/ecco-jcc/jjy113
- Eslam M, Sarin SK, Wong VWS, et al. The Asian Pacific Association for the Study of the Liver clinical practice guidelines for the diagnosis and management of metabolic associated fatty liver disease. *Hepatol Int*. 2020;14(6):889–919.
- Bolger AM, Marc L, Bjoern U. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*. 2014;30:2114–2120. doi:10.1093/bioinformatics/btu170
- Ruiqiang L, Yingrui L, Karsten K, et al. SOAP: short oligonucleotide alignment program. *Bioinformatics*. 2008;24:713–714. doi:10.1093/bioinformatics/btn025
- Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010;26:589–595. doi:10.1093/bioinformatics/btp698
- Peng Y, Leung Henry CM, Yiu SM, et al. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*. 2012;28:1420–1428. doi:10.1093/bioinformatics/bts174
- Anton B, Sergey N, Dmitry A, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19:455–477. doi:10.1089/cmb.2012.0021

28. Dinghua L, Ruibang L, Chi-Man L, et al. MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*. 2016;102:3–11. doi:10.1016/j.ymeth.2016.02.020
29. Grabherr Manfred G, Haas Brian J, Moran Y, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29:644–652. doi:10.1038/nbt.1883
30. Stephen N, Pedro CA, Frederik S, et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol*. 2021;39:578–585. doi:10.1038/s41587-020-00774-7
31. Jiarong G, Ben B, Zayed Ahmed A, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome*. 2021;9:37. doi:10.1186/s40168-020-00990-y
32. Jing-Zhe J, Wen-Guang Y, Jiayu S, et al. Virus classification for viral genomic fragments using PhaGCN2. *Brief Bioinform*. 2023;24:undefined. doi:10.1093/bib/bbac505
33. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30:2068–2069. doi:10.1093/bioinformatics/btu153
34. Chantal H, de Castro E, Patrick M, et al. ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res*. 2011;39:D576–82. doi:10.1093/nar/gkq901
35. Shang J, Sun Y. CHERRY: a Computational metHod for accuratE pRediction of virus-pRokaryotic interactions using a graph encoder-decoder model. *Brief Bioinform*. 2022;23. doi:10.1093/bib/bbac182
36. Congyu L, Zheng Z, Zena C, et al. Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics. *BMC Biol*. 2021;19:5. doi:10.1186/s12915-020-00938-6
37. Mohammad Z, Shaghayegh A-T, Siddharth S, et al. Meta-analysis: prevalence of, and risk factors for, non-alcoholic fatty liver disease in patients with inflammatory bowel disease. *Aliment Pharmacol Ther*. 2022;55:894–907. doi:10.1111/apt.16879
38. Ehsan A-S, Negin L, Naeim N, et al. Global prevalence of nonalcoholic fatty liver disease: an updated review meta-analysis comprising a population of 78 million from 38 countries. *Arch Med Res*. 2024;55. doi:10.1016/j.arcmed.2024.103043
39. Clooney Adam G, S STD, Shkoporov Andrey N, et al. Whole-Virome analysis sheds light on viral dark matter in inflammatory bowel disease. *Cell Host Microbe*. 2019;26:764–778.e5. doi:10.1016/j.chom.2019.10.009
40. Gregory Ann C, Olivier Z, Zayed Ahmed A, et al. The gut virome database reveals age-dependent patterns of virome diversity in the human gut. *Cell Host Microbe*. 2020;28:724–740.e8. doi:10.1016/j.chom.2020.08.003
41. Fengting S, Qingsong Z, Jianxin Z, et al. A potential species of next-generation probiotics? The dark and light sides of *Bacteroides fragilis* in health. *Food Res Int*. 2019;126:108590. doi:10.1016/j.foodres.2019.108590
42. Camarillo-Guerrero Luis F, Almeida A, Rangel-Pineros G, et al. Massive expansion of human gut bacteriophage diversity. *Cell*. 2021;184:1098–1109.e9. doi:10.1016/j.cell.2021.01.029
43. Stephen N, David P-E, Lee C, et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat Microbiol*. 2021;6. doi:10.1038/s41564-021-00928-6
44. Krishnamurthy Siddharth R, Janowski Andrew B, Zhao G, et al. Hyperexpansion of RNA Bacteriophage Diversity. *PLoS Biol*. 2016;14:e1002409. doi:10.1371/journal.pbio.1002409
45. Samuel M, Rohini S, Jun C, et al. The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res*. 2011;21:1616–1625. doi:10.1101/gr.122705.111
46. Sears Cynthia L, Geis Abby L, Housseau F. *Bacteroides fragilis* subverts mucosal biology: from symbiont to colon carcinogenesis. *J Clin Invest*. 2014;124:4166–4172. doi:10.1172/JCI72334
47. Kaakoush Nadeem O. Sutterella species, IgA-degrading bacteria in ulcerative colitis. *Trends Microbiol*. 2020;28:519–522. doi:10.1016/j.tim.2020.02.018
48. Wei L, Saiyu H, Yuan F, et al. A bacterial bile acid metabolite modulates T activity through the nuclear hormone receptor NR4A1. *Cell Host Microbe*. 2021;29:1366–1377.e9. doi:10.1016/j.chom.2021.07.013
49. Koji A, Takeshi T, Kenshiro O, et al. Treg induction by a rationally selected mixture of Clostridia strains from the human microbiota. *Nature*. 2013;500:232–236. doi:10.1038/nature12331
50. Hsu BB, Gibson TE, Yeliseyev V, et al. Dynamic modulation of the gut microbiota and metabolome by bacteriophages in a mouse model. *Cell Host Microbe*. 2019;25. doi:10.1016/j.chom.2019.05.001