Am J Cancer Res 2025;15(11):4905-4922
www.ajcr.us /ISSN:2156-6976/ajcr0167098

Original Article

Identification of colorectal cancer using enteroviral and bacterial subtypes

Shuwen Han[1,2,3,4,5], Jing Zhuang[2,3,4], Jian Chu[2,3,4], Zheng Wu[2,3,4], Yin Jin[2,3,4], Jin Liu[2,3,4], Yinhang Wu[2,3,4]

[1]School of Medicine, Tarim University, Alaer 843300, Xinjiang, The People's Republic of China; [2]Huzhou Central Hospital, Affiliated Central Hospital Huzhou University, Huzhou 313000, Zhejiang, The People's Republic of China; [3]Huzhou Central Hospital, Fifth School of Clinical Medicine of Zhejiang Chinese Medical University, Huzhou 313000, Zhejiang, The People's Republic of China; [4]Zhejiang-France United Laboratory of Integrated Traditional Chinese and Modern Medicine in Colorectal Cancer, Huzhou 313000, Zhejiang, The People's Republic of China; [5]ASIR (Institute - Association of Intelligent Systems and Robotics), 14B Rue Henri Sainte Claire Deville, Rueil-Malmaison 92500, France

Received June 28, 2025; Accepted November 17, 2025; Epub November 25, 2025; Published November 30, 2025

Abstract: Colorectal cancer (CRC) is influenced by both enteroviruses and bacteria, yet current microbial typing schemes rely primarily on bacterial data. To construct a more comprehensive microbial typing scheme for CRC, this study integrated enteroviral and bacterial profiling data from fecal samples of 414 healthy controls (NCs), 151 advanced adenoma (AAs) patients, and 255 CRC patients using Illumina sequencing. Metabolites and trace elements were analyzed via liquid chromatography and ICP-MS, respectively. Microbial subtyping was performed with ConsensusClusterPlus based on combined viral and bacterial sequencing data, leading to the identification of two initial viral subtypes (V1, n=309; V2, n=511). The V2 group was further split into two bacterial subtypes (V2B1, V2B2), yielding three distinct microbial subtypes. Disease ratios (CRCs&AAs/NCs) were 1.06 (V1), 0.67 (V2B1), and 1.29 (V2B2). V1 showed increased *Streptococcus agalactiae*, *Peduvirus*, and Imidazopyrimidines; V2B1 had higher *CAG_127sp900553925*, nickel (Ni), and benzene derivatives; V2B2 exhibited elevated *Citrobacter farmeri*, *Svunavirus*, arsenic, and organic sulfonic acids. Gut disease prediction model was more accurate after virus typing (86.54% of accuracy in V2B2 subtype; 73.33% of accuracy without typing). These results suggest that integrating enteroviral and bacterial subtypes offers a more precise framework for CRC identification than bacterial-based typing alone.

Keywords: Colorectal cancer, enteroviruses, bacteria, subtype, unsupervised clustering

Introduction

Colorectal cancer is a malignant tumor originating from the epithelial cells of the mucosa in the colon or rectum, accounting for more than one-fourth of all gastrointestinal cancers and one-fifth of cancer-related mortality [1]. Inflammatory bowel disease (IBD) and familial adenomatous polyposis are both genetic factors of CRC [2]. In addition, low fiber and high red meat diets, staying up late, smoking, excessive alcohol consumption, and obesity are all considered as high-risk factors of CRC [3-5]. No matter what the initiating factor of CRC is, the destruction of homeostasis immune balance and intestinal microbiome derived signals usually involve, thus stimulating the overrepair response of epithelial cells, a series of gene mutations in intestinal stem cells, such as APC gene, K-RAS gene or p53 gene, and the occurrence of tumors [6, 7].

One characteristic associated with the pathogenesis of CRC is the dysregulation of the gut microbiome. Gut microbiome, including a series of microbial genomes, such as bacteria, viruses, fungi and so on, is a complex microecosystem. At present, with the progress and development of metagenomic sequencing methods, intestinal metagenomic databases of 4,644 gut bacterial genomes [8] and 54,118 viral populations (vOTUs) [9] have been constructed. Host and gut microbes have co-evolved and formed a robust immune system defense against potentially harmful pathogens. At the

https://doi.org/10.62347/QAYZ9788

same time, gut flora closely interacts with host gut epithelial cells and affects the occurrence and development of CRC by participating in host immune regulation, food metabolism, or gene toxin production. It has been found that pks plasmid containing *Escherichia coli (pks+ E. coli)* [10] and enterotoxigenic *Bacteroides fragilis* (ETBF) [11] induces intestinal mucosal cancer by inducing DNA damage. *Fusobacterium nucleatum* can amplify E-cadherin/β-catenin signaling and lead to tumorigenesis leading to CRC [12]. In addition, enterovirus coexists with its host for a long time. In rare cases, due to the high level of mutations in the genome during replication, it may also encode some viral products, such as *Epstein-Barr Virus* that promotes CRC by intervening in cell metabolic reprogramming [13]. However, these microorganisms have only been proved to promote the development of CRC, but the presence and number of certain microorganisms in the human body alone are not enough to cause cancer.

Microbiome interactions are not limited to microbes and their hosts, but also exist between microbial communities. Gut microbiome coexistence refers to the relationship between different types of microorganisms in the gut. In the gut, the relationship between microbes can be symbiosis, competition, coexistence, reciprocity, and other different ways. In addition to bacteria, gut microbes also contain many viruses. Bacteria can act as hosts for some viruses, coexist with viruses, and maintain their balance in the ecosystem. Viruses, such as specific bacteriophages, can enter the interior of bacteria by interfering with the membrane complex of bacteria or using the receptors of certain cells, and then destroy the cell structure, change the metabolic process, and eventually lead to cell death or the activation of immune response [14]. On the other hand, bacteria use a common core genome shared across various bacterial species, along with mobile genetic elements, to facilitate swift bacterial evolution through genome exchange, ultimately achieving defensive effects [15, 16]. The microbiome of microbes that inhabit the human gut is very individual diverse and identifies individual disease states from only a single or a few species that lack the mindset to analyze gut microbes. The classification of enteroviruses combined with gut bacteria is helpful to distinguish the disease status of individuals in diagnosis and can be used as a risk or susceptibility indicator for specific human conditions. In 2011, researchers used different pairs of bacterial genes to identify the different makeup of each person's microbiome. The population was roughly divided into three types (also known as enterotypes): *Bacteroides, Prevotella*, and *Ruminococcus* [17]. However, until now, the human microbial typing is still mainly based on bacteria, and there is a lack of enterovirus-based typing.

In addition, the microbes present in the human gut pose impacts on metabolites and metal ions in the gut microecological environment. Metabolites are diverse, including organic acids, amino acids, polybrain, fatty acids, growth factors, polysaccharides, aromatic compounds, etc. [18]. Gut bacteria affect host metabolic processes of energy production, lipid production, glucose production and cholesterol synthesis through the production of short-chain fatty acids (SCFA) [19, 20]. Meanwhile, gut microbes also affect the absorption of metal ions such as Fe in the gut [21].

In this study, 820 subjects' fecal samples were sequenced by metagenomic sequencing. Using viral and bacterial annotation data, the subjects were categorized into gut microbial types. The differences of gut microbes among different subtypes, the differences of metabolites and ions in feces, and the characteristics of clinical features were analyzed. This study was expected to type the human gut microecology from the perspective of bacteria and viruses, establish a classification method of gut microbes, and clarify the characteristics of these subtypes of gut microbes and the relationship between microbial subtypes and intestinal diseases (advanced adenoma and CRC), to provide a basis for describing intestinal health and disease status based on intestinal microbial typing.

Methods

The flow chart of this study is shown in Figure S1.

Subjects and samples

This study recruited the subjects from Huzhou Central Hospital from March 2020 to December 2022, with 414 normal controls (NCs), 151

advanced adenoma patients (AAs), and 255 colorectal cancer patients (CRCs). The clinical information was shown as Table S1. All subjects signed informed consent in accordance with the guidelines approved by the Ethics Committee of Huzhou Central Hospital. The patients' clinical protocol and informed consent have been approved by the Ethics Committee of Huzhou Central Hospital (No. 2019-1101-01 and No. 202202005-01) and the Chinese Clinical Trial Registry (http://www.chictr.org.cn, ChiCTR2100050167).

Inclusion criteria: patients with CRC confirmed by pathological examination and patients with AA volunteered to participate in the study. The pathological diagnosis of CRC and AA was confirmed, and the clinical stage of CRCs was determined according to American Joint Committee on Cancer-Staging Manual 8th edition. The NCs were endoscopically negative.

Exclusion criteria: 1) Coexistence of other malignant tumors; 2) Severe cardiopulmonary diseases; 3) Presence of other intestinal diseases, such as ulcerative colitis, Crohn's disease, etc.; 4) Recent use of antibiotics within 3 months prior to admission; 5) History of oral microbiological and lipid-modulating drug use within the last 2 months; 6) Known primary organ failure. Informed consent forms were obtained from all study participants.

The basic information and clinical detection indicators of subjects and pathological data of CRCs were obtained from the medical record management system of Huzhou Central Hospital. The stool samples were collected before breakfast without the use of laxatives or lubricants, about 5-10 grams and stored in the ultra-low temperature refrigerator within half an hour (The total storage time was less than 1 month).

Metagenomic sequencing

Whole-genome shotgun sequencing of stool sample was carried on an Illumina HiSeq X instrument and the specific steps are as follows.

Microbial DNA was extracted from stool samples using the E.Z.N.A.® stool DNA Kit (Omega Bio-tek, Norcross, GA, U.S.) according to manufacturer's protocols. The DNA integrity was determined by electrophoresis in 0.8% agarose gels, and the concentration and quality were determined using a Nanodrop ND-1000 (Thermo Scientific). For samples yielding less than 500 ng of DNA per extraction, the process was repeated, and extracts were pooled to obtain a minimum of 1 μg of DNA for downstream applications. High-quality DNA sample (OD260/280=1.8-2.2, OD260/230 ≥ 2.0) was used to construct sequencing library. Metagenomic shotgun sequencing libraries were constructed and sequenced at Shanghai Biozeron Biological Technology Co., Ltd. In briefly, for each sample, 1 μg of genomic DNA was sheared by Covaris S220 Focused-ultrasonicator (Woburn, MA, USA), and sequencing libraries were prepared with a fragment length of approximately 450 bp. All samples were sequenced in the Illumina HiSeq X instrument with pair-end 150 bp (PE150) mode. Raw sequence reads underwent quality trimming using Trimmomatic (http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic) to remove adaptor contaminants and low-quality reads [22]. The reads underwent quality control and were subsequently aligned to the human genome (version: hg19) using the BWA mem algorithm with the following parameters: -M -k 32 -t 16 (source: http://bio-bwa.sourceforge.net/bwa.shtml). The reads removing host-genome contaminations and low-quality data were named as clean reads and used for further analysis.

The human gut microbial genome data were sourced from the Unified Human Gastrointestinal Genome (UHGG) repository (http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/human-gut/) [23], which included both assembled genomes and their functionally annotated gene models. For viral genome data, we utilized the Metagenomic Gut Virus catalog (MGV, https://portal.nersc.gov/MGV/) [24], a comprehensive collection of curated human gut virome sequences.

Taxonomy of clean reads for each sample was determined by Kraken2 [25] using the customized kraken database. The customized kraken database included all bacteria and virus genome sequences in NCBI RefSeq database (release number: 90). All reads were classified to seven phylogenetic levels (domain, phylum, class, order, family, genus, and species) or unclassified. The abundances of taxonomy were estimated by Bracken (https://ccb.jhu.edu/

software/bracken/) that can produce accurate species- and genus-level abundance even in multiple near-identical species. The beta diversity analysis was performed using the community ecology package and R-vegan package.

Untargeted metabolomics detecting

LC-MS/MS technique was used for untargeted detection of metabolites. The chromatographic peaks detected in the samples were integrated using CD3.1 data processing software, where the peak area of each characteristic peak represented the relative quantitative value of a metabolite, while the background ions were removed with blank samples, and the original quantitative results were normalized using total peak area. Afterwards, metabolites with less than 30% of Coefficient of Variance (CV) in QC samples were retained, and finally the identification and relative quantitative results of metabolites were obtained.

Ions detecting

Pretreatment: After drying the samples in the freeze dryer, a certain amount of samples were weighed in a graphite digestion tube. First, 15 mL of nitric acid was add at 80°C for 20 min, then 3 mL of perchloric acid was added at 80°C of digestion for 10 min, after which it was heat up to 130°C and maintain 15 min, and continue to 180°C for 120 min. Until the digestion solution was colorless and clear, perchloric acid evaporated to be almost dry, and the digestion was complete. After cooling to room temperature, the digestion solution was transferred to a 50 mL volumetric flask and adjusted to the scale line with distilled water to be tested. The relevant reagents and instruments were listed below.

Nitroric acid (GR): Sinopharm Chemical Reagent Co., LTD. Shanghai 10014508.

Perchloric acid: Tianjin Zhengcheng Chemical Products Co., LTD.

Standard material: Tan ink quality inspection BWT30121-100-100 B22120033. Instrument model number: ICP-MS7800.

Cluster analysis

Package "ConsensusClusterPlus" [26] in R studio was used to perform cluster analysis to identify gut microbial subtypes using metagenomic sequencing data.

Statistical analysis

For identification of biomarkers for highly dimensional virus, linear discriminant analysis effect size (LEfSe) analysis was conducted [27]. Kruskal-Wallis sum-rank test was performed to examine the changes and dissimilarities among classes, followed by LDA analysis to determine the size effect of each distinctively abundant taxa [28]. A threshold of 4.0 on the logarithmic LDA score was used for identifying discriminative features.

Wilcoxon rank-sum test (Mann-Whitney U test) was used to analyze the species, metabolites, or ions of the two groups of samples for significant differences (p-value < 0.05). After difference analysis, false discovery rate (FDR) method was further adopted to calibrate the obtained P-values, and species, metabolites or ions with significant abundance differences in different groups were obtained (Q-value < 0.05). Kruskal-Wallis sum-rank analyzed the species, metabolites or ions of the three groups of samples for significant differences. After the difference analysis, bonferroni correction method was used for multiple correction of p-value obtained.

Spearman analysis was made to calculate correlations between different species and groups.

Bioinformatic analysis was performed using the OmicStudio tools at https://www.omicstudio.cn/tool.

Disease prediction model

The entire dataset were randomly assigned to the discovery (80%) and test (20%) sets. No specific grouping or stratification factors were applied during this process to ensure an unbiased split. The methods of model construction were the same as previously described [29]. Catboost package (version 0.16.5) uses relevant functions in R language rminer Package (version 1.4.5) for modeling analysis and importance calculation of variables by using multinom from nnet package. In classification, the class with the highest probability is selected, while regression analysis is performed using probability averaging. Key biomarkers are identified for sample categorization. By applying varying cutoff values to continuous variables, sensitivity and specificity are computed, fol-

lowed by generating an ROC curve (sensitivity vs. 1 - specificity). Model accuracy is assessed via a cross-validation matrix (CV matrix).

Results

Gut microbial typing combined enteroviruses and bacteria

A total of 820 subjects were recruited, including 414 normal controls (NCs), 151 advanced adenoma patients (AAs), and 255 colorectal cancer patients (CRCs). Metagenomic sequencing was performed on stool samples of the study subjects, and 820 viromic and bacteromic data were obtained. Unsupervised clustering of 820 cases of viromic data was performed, and 820 cases were classified into two categories (k=2): V1 (309 samples) and V2 (511 samples) (**Figures 1A**, S2A, S2B; Table S2). Subsequently, unsupervised clustering of bacteriological data was performed on samples V1 and V2 respectively, and the clustering effect of V1 was poor (Figure S2C-E). The 511 samples of V2 were regrouped into two classes (k=2), named V2B1 (252 samples) and V2B2 (259 samples) (**Figures 1B**, S2E-G). Finally, through unsupervised clustering, three subtypes of gut microbes were obtained: V1, V2B1, and V2B2.

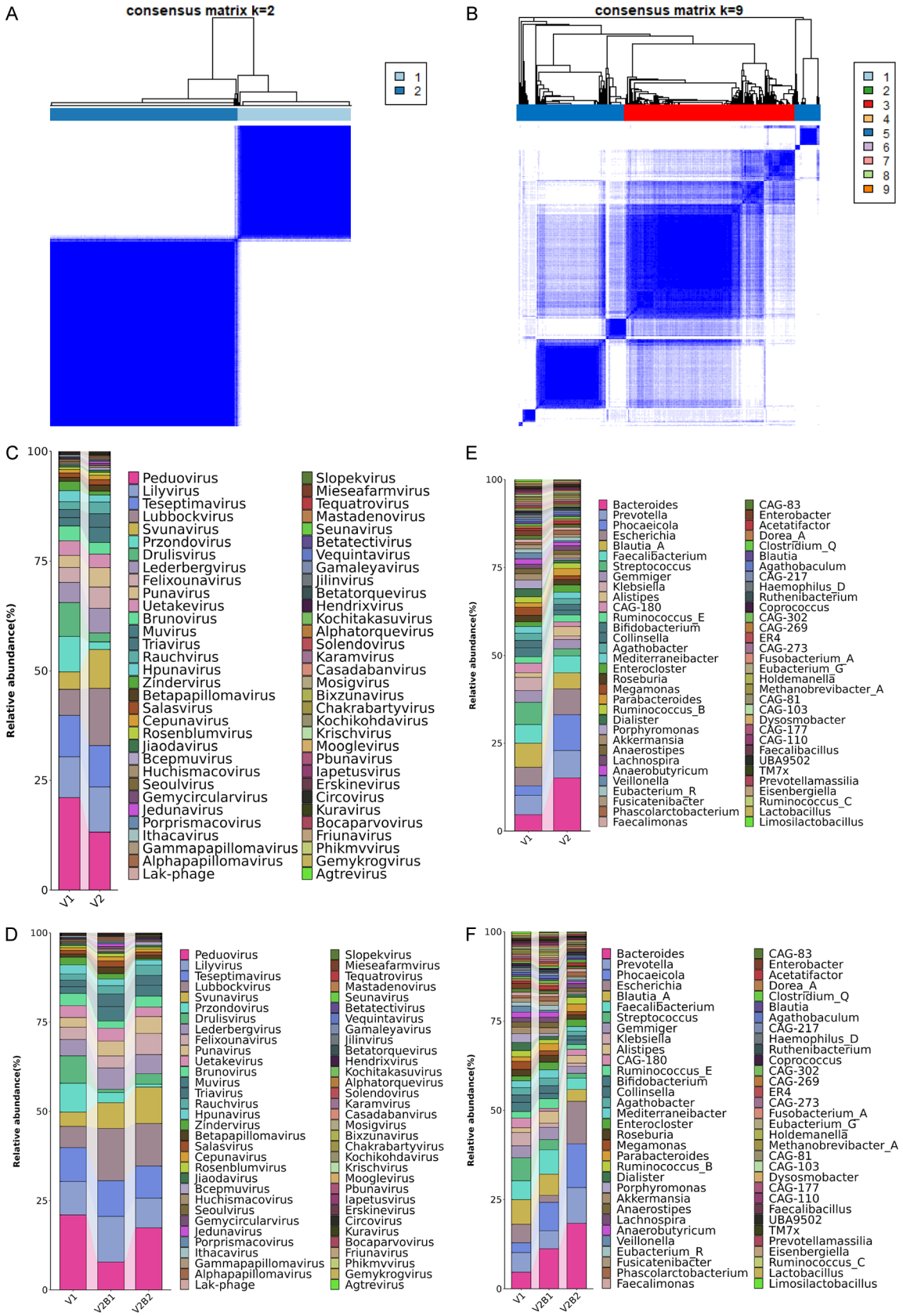
The community composition and structure of the three subtypes of enterovirus and gut bacteria were analyzed respectively. *Peduovirus* was the dominant virus in the three subtypes, but the abundance of *Peduovirus* in V1 subtype was significantly higher than that of V2 subtype, while the abundance of *Lubbockvirus* and *Svunavirus* in V2 subtype was significantly higher than that of V1 subtype (**Figure 1C**, **1D**). In terms of bacteria, *Bacteroides* had the highest abundance of the three subtypes. In addition, it was found that the abundances of *Bacteroides*, *Prevotella* and *Phocaeicola* gradually increased in V1, V2B1 and V2B2. The abundances of *Escherichia* significantly increased in the community structure proportion of V2B2 (**Figure 1E**, **1F**).

To verify the representativeness of the three subtypes, subtypes V1 and V2 in the first cluster and subtypes B1 and B2 in the second cluster were verified, respectively. Kruskal-Wallis test screened out 20 different viruses, including *Peduovirus*, *Iphppillomvirus* and *Porprimcovirus* between V1 and V2 subtypes (Figure S3A). LDA analysis found 11 different bacteria at the species level, containing *Bifidobacteriumscardovii*, *Collinsellasp900541125, Eubacterium_Msp900540015* and so on (Figure S3B). Four differential viruses such as *Krischvirus* (Figure S3C) and ten differential bacteria such as *Phacaeicoiasp002493165* at the species level (Figure S3D) were screened out between subtypes V2B1 and V2B2. Microbial typing prediction models of virus and bacterial subtypes were constructed based on 20 different viruses and 20 different bacteria of V1 and V2, respectively. The results revealed that the accuracy of the training set of the virus typing model (**Figure 1G**) was 96.18% (sensitivity: 98.04%; specificity: 93.12%), and the accuracy of the verification set was 84.24% (sensitivity: 83.50%; specificity: 85.48%). The accuracy of the training set of the bacterial typing model (**Figure 1H**) was 86.55% (sensitivity: 80.75%; specificity: 94.71%), and the accuracy of the verification set was 81.19% (sensitivity: 76.27%; specificity: 88.10%). The results of intestinal microbial typing are reliable.

Characterization and difference analysis of gut microbes and metabolites based on microbial typing

For subtypes V1, V2B1, and V2B2, NMDS analysis illustrated that the bacterial and viral community structures of the three subtypes were similar (Figure S4A, S4B). Among the three subtypes, 17 different viruses, including *Peduovirus*, *Porprimcovirus*, and *Salasvirus*, were screened out (**Figure 2A**); 26 species of differential bacteria, including *Collinsellasp900547125*, *CAG_617sp000438115*, and *Prevotellasp900540415*, were screened at species level by LDA analysis (**Figure 2B**).

Moreover, the clinical information of the three subtypes were analyzed, and it was found that the average age of V1 was high (61.11±9.64), and the positive rate of fecal occult blood test (FOBT) of V2B2 was the highest. The proportions of CRC&AA/NC in V1, V2B1, and V2B2 were 1.06, 0.67, and 1.29, respectively (**Table 1**). In addition, pathological information of CRC was analyzed, and it was discovered that the ratio of stage III + V/I + II cancer stages in CRC patients with V2B2 subtype was higher than that of the other two subtypes (**Table 2**).


A
consensus matrix k=2
1
2
B
consensus matrix k=9
1
2
3
4
5
6
7
8
9
C
Relative abundance(%)
100
75
50
25
0
V1
V2
Peduovirus
Lilyvirus
Teseptimavirus
Lubbockvirus
Svunavirus
Przondovirus
Drulisvirus
Lederbergvirus
Felixounavirus
Punavirus
Uetakevirus
Brunovirus
Muvirus
Triavirus
Rauchvirus
Hpunavirus
Zindervirus
Betapapillomavirus
Salasvirus
Cepunavirus
Rosenblumvirus
Jiaodavirus
Bcepmuvirus
Huchismacovirus
Seoulvirus
Gemycircularvirus
Jedunavirus
Porprismacovirus
Ithacavirus
Gammapapillomavirus
Alphapapillomavirus
Lak-phage
Slopekvirus
Mieseafarmvirus
Tequatrovirus
Mastadenovirus
Seunavirus
Betatectivirus
Vequintavirus
Gamaleyavirus
Jilinvirus
Betatorquevirus
Hendrixvirus
Kochitakasuvirus
Alphatorquevirus
Solendovirus
Karamvirus
Casadabanvirus
Mosigvirus
Bixzunavirus
Chakrabartyvirus
Kochikohdavirus
Krischvirus
Mooglevirus
Pbunavirus
Iapetusvirus
Erskinevirus
Circovirus
Kuravirus
Bocaparvovirus
Friunavirus
Phikmvvirus
Gemykrogvirus
Agtrevirus
D
Relative abundance(%)
100
75
50
25
0
V1
V2B1
V2B2
Peduovirus
Lilyvirus
Teseptimavirus
Lubbockvirus
Svunavirus
Przondovirus
Drulisvirus
Lederbergvirus
Felixounavirus
Punavirus
Uetakevirus
Brunovirus
Muvirus
Triavirus
Rauchvirus
Hpunavirus
Zindervirus
Betapapillomavirus
Salasvirus
Cepunavirus
Rosenblumvirus
Jiaodavirus
Bcepmuvirus
Huchismacovirus
Seoulvirus
Gemycircularvirus
Jedunavirus
Porprismacovirus
Ithacavirus
Gammapapillomavirus
Alphapapillomavirus
Lak-phage
Slopekvirus
Mieseafarmvirus
Tequatrovirus
Mastadenovirus
Seunavirus
Betatectivirus
Vequintavirus
Gamaleyavirus
Jilinvirus
Betatorquevirus
Hendrixvirus
Kochitakasuvirus
Alphatorquevirus
Solendovirus
Karamvirus
Casadabanvirus
Mosigvirus
Bixzunavirus
Chakrabartyvirus
Kochikohdavirus
Krischvirus
Mooglevirus
Pbunavirus
Iapetusvirus
Erskinevirus
Circovirus
Kuravirus
Bocaparvovirus
Friunavirus
Phikmvvirus
Gemykrogvirus
Agtrevirus
E
Relative abundance(%)
100
75
50
25
0
V1
V2
Bacteroides
Prevotella
Phocaeicola
Escherichia
Blautia_A
Faecalibacterium
Streptococcus
Gemmiger
Klebsiella
Alistipes
CAG-180
Ruminococcus_E
Bifidobacterium
Collinsella
Agathobacter
Mediterraneibacter
Enterocloster
Roseburia
Megamonas
Parabacteroides
Ruminococcus_B
Dialister
Porphyromonas
Akkermansia
Anaerostipes
Lachnospira
Anaerobutyricum
Veillonella
Eubacterium_R
Fusicatenibacter
Phascolarctobacterium
Faecalimonas
CAG-83
Enterobacter
Acetatifactor
Dorea_A
Clostridium_Q
Blautia
Agathobaculum
CAG-217
Haemophilus_D
Ruthenibacterium
Coprococcus
CAG-302
CAG-269
ER4
CAG-273
Fusobacterium_A
Eubacterium_G
Holdemanella
Methanobrevibacter_A
CAG-81
CAG-103
Dysosmobacter
CAG-177
CAG-110
Faecalibacillus
UBA9502
TM7x
Prevotellamassilia
Eisenbergiella
Ruminococcus_C
Lactobacillus
Limosilactobacillus
F
Relative abundance(%)
100
75
50
25
0
V1
V2B1
V2B2
Bacteroides
Prevotella
Phocaeicola
Escherichia
Blautia_A
Faecalibacterium
Streptococcus
Gemmiger
Klebsiella
Alistipes
CAG-180
Ruminococcus_E
Bifidobacterium
Collinsella
Agathobacter
Mediterraneibacter
Enterocloster
Roseburia
Megamonas
Parabacteroides
Ruminococcus_B
Dialister
Porphyromonas
Akkermansia
Anaerostipes
Lachnospira
Anaerobutyricum
Veillonella
Eubacterium_R
Fusicatenibacter
Phascolarctobacterium
Faecalimonas
CAG-83
Enterobacter
Acetatifactor
Dorea_A
Clostridium_Q
Blautia
Agathobaculum
CAG-217
Haemophilus_D
Ruthenibacterium
Coprococcus
CAG-302
CAG-269
ER4
CAG-273
Fusobacterium_A
Eubacterium_G
Holdemanella
Methanobrevibacter_A
CAG-81
CAG-103
Dysosmobacter
CAG-177
CAG-110
Faecalibacillus
UBA9502
TM7x
Prevotellamassilia
Eisenbergiella
Ruminococcus_C
Lactobacillus
Limosilactobacillus

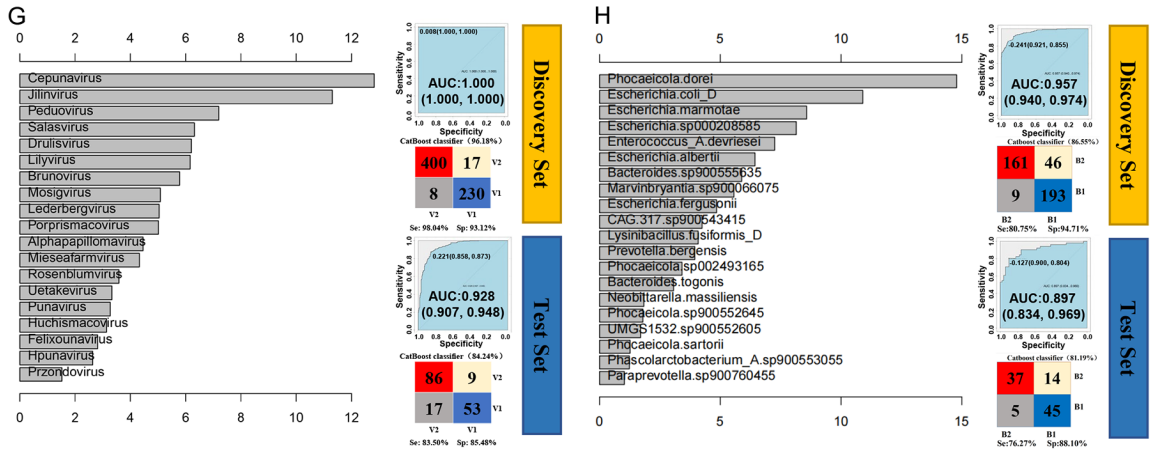


Figure 1. Gut microbial typing combined enteroviruses and bacteria. A, B: Unsupervised cluster analysis of enteroviruses and bacteria, respectively. C, E: Percentile chart of enterovirus community and bacterial community of V1 and V2. D, F: Percentile chart of enterovirus community and bacterial community of V1, V2B1, and V2B2. G, H: Microbial typing prediction models of virus and bacterial subtypes.

Furthermore, fecal metabolites were detected on 276 samples from 820 subjects. The content of ions in fecal samples was detected on 283 samples from 820 subjects, and 32 ions were detected. Among the three subtypes, the content of Mn ion (**Figure 2C**) and carboxylic acids and derivatives (**Figure 2D**) was the most abundant. There were 12 differential metabolites among the three subtypes. The content of Imidazopyrimidines increased in V1, benzene and substituted derivatives increased in V2B1, and organic sulfonic acids and derivatives increased in V2B2 (**Figure 2E**). According to the difference analysis, there was no statistical - difference between the ion groups of V1, V2B1, V2B2. However, there are different ions Ni (**Figure 2F**) between V1 and V2B1 and different ions As (**Figure 2G**) between V2B1 and V2B2.

Characterization of gut microbes based on clinical differences in microbial typing

NMDS analysis demonstrated that there was no significant difference in the community composition of gut microbes in CRC&AA (disease group) and NCs (**Figure 3A**, **3B**). In terms of enteroviruses, *Peduovirus* is the most abundant virus in the intestinal tract of all populations, while *Przondovirus* and *Drulisvirus* are highly abundant in the intestinal tract of the disease group, and *Teseptimavirus* is highly abundant in the intestinal tract of the NCs group (**Figure 3C**). In terms of gut bacteria, there was no significant difference in bacterial composition between the disease and NCs groups (**Figure 3D**).

In addition, differential viruses, including *Muvirus*, *Solendovirus*, and *Chakrabartyvirus* (**Figure 3E**), differential bacteria, such as *CAG-312 sp900546565*, *UBA1259 sp900770685*, and *UBA4636 sp900764595* (**Figure 3F**), were screened out between the disease and NCs groups. Intra-group and inter-group correlation analyses were performed for different bacteria and viruses, respectively. The results demonstrated stronger associations between bacteria and viruses in the NCs group (**Figures 3G**, S5B), whereas weaker correlations were observed in the CRCs and AAs groups. Linear regression analysis revealed a correlation coefficient of 0.0017 in the NCs group, which was significantly higher than that in the CRCs and AAs groups (0.00022) (Figure S5A, S5B). Inter-group correlation analysis indicated that *Muvirus* was more closely related to NCs (**Figure 3H**). The association between *CAG-312 sp900546565*, *UBA1259 sp900770685*, *UBA4636 sp900764595* and NCs was closer (**Figure 3I**).

According to different clinical information, the population was divided into age > 50 years old and age ≤ 50 years old. There were 5 different viruses, including *Cepunavirus*, *Friunavirus*, *Muvirus*, *Peduovirus* and *Svunavirus* (**Figure 4A**), and 56 different bacteria, including *Olsenella_E sp002160255*, *Adlercreutzia celatus_A*, and *CAAEEV01 sp900754955* (**Figure 4B**). The

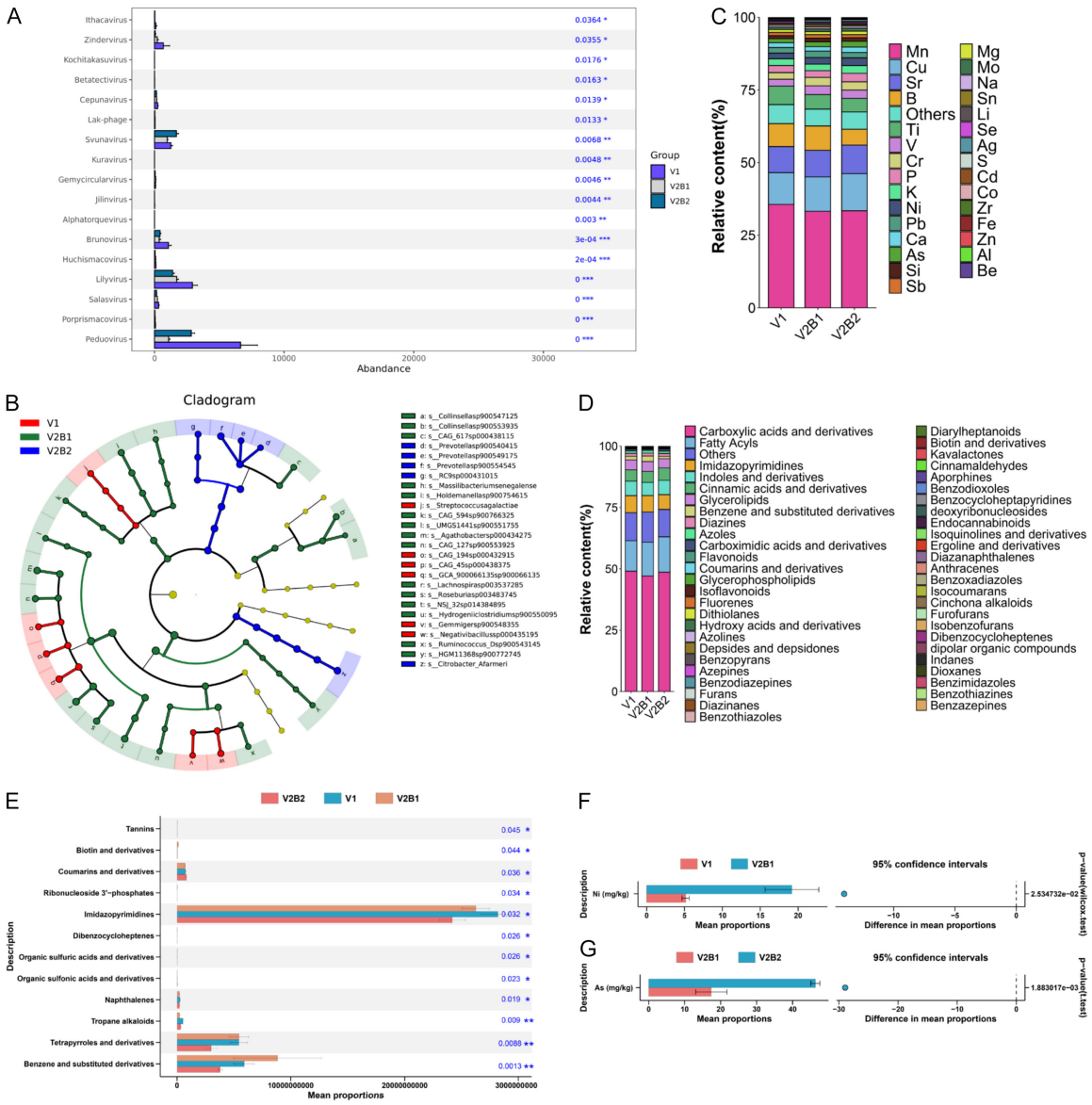


Figure 2. Characterization of gut microbes, metabolites, and ions of three subtypes. A: Kruskal-Wallis sum-rank test was used to detect the differential viruses of the three subtypes. B: LDA analysis was used to detect the differential bacteria of the three subtypes (LDA score > 4). C, D: Percentile chart of content of metabolites and ions among the three subtypes. E: Kruskal-Wallis sum-rank test was used to detect the differential metabolites of the three subtypes. F, G: Wilcoxon rank-sum test was used to detect the differential ions between V1 and V2B1, and V2B1 and V2B2, respectively.

human population was divided into fecal OB (+) and OB (-), and 41 differential viruses, including *Betatectivirus*, *Bixzunavirus*, and *Cepunavirus* (**Figure 4C**), and 2,189 differential bacteria, including *Absicoccus porci*, *Acinetobacter baumannii*, and *Agathobacter sp000434275* (**Figure 4D**) were screened out. Among these different strains, the viruses found simultaneously were *Svunavirus* and *Cepunavirus* (**Figure 4E**). Meanwhile, 37 species of bacteria were found, including *Adlercreutzia celatus_A*, *Allisonella pneumosintes*, etc. (**Figure 4F**).

Construct CRC&AA disease prediction model based on microbial typing

CRC&AA disease prediction models were constructed based on different groups (**Figure 5**; **Table 3**). Catboost model had the best performance, and CRC&AA disease prediction model was more accurate after virus typing. In V2B2,

Table 1. Clinical information of three gut microbial subtypes

	V1 (n=309)	V2B1 (n=252)	V2B2 (n=259)	*p*-value
Sex				0.740
Male	157	120	126	
Female	152	132	133	
Age	61.11±9.64	58.77±9.66	60.48±10.03	0.016
≤ 50	40	53	40	0.033
> 50	269	199	218	
BMI	24.04±6.18	23.75±5.36	24.60±6.65	0.346
FOBT				0.005
OB (+)	90	62	92	
OB (-)	181	154	120	
NA	38	36	47	
Disease				
NC	150	151	113	
AA	59	36	56	0.006
CRC	100	65	90	
AA&CRC/NC	1.06	0.67	1.29	0.001
White blood count, WBC	6.17±4.00	6.08±3.61	6.12±2.26	0.962
Albumin	39.30±4.37	39.58±4.06	39.29±5.40	0.829
Triglyceride, TG	1.43±1.17	1.47±1.64	1.35±0.83	0.719
Total cholesterol, TC	4.72±3.86	5.28±6.06	4.64±1.13	0.384
High density lipoprotein, HDL	45.68±12.48	48.31±14.55	49.78±19.09	0.062
Low density lipoprotein, LDL	96.86±31.58	94.03±27.87	95.71±31.26	0.748

the accuracy of Catboost model was 93.24%, with 85.24% of sensitivity and 91.87% of specificity in the discovery set. In the test set, the accuracy was 86.54%, the sensitivity was 80.77% and the specificity was 92.31%.

Discussion

Through unsupervised clustering, three subtypes of gut microbes were obtained in the present study, namely V1, V2B1, and V2B2. The gut microbe is considered as a "new organ" with an important role, and many species of bacteria have evolved and adapted to live and grow in human gut. The structure and organization of the gut microbes reflect the exchange of information at the microbial and host levels, which promotes cooperation and functional stability within this complex ecosystem. The normal flora is the microbe of the human body as the host, while the microecosystem is composed of the normal flora and its host microenvironment (tissues, cells, metabolites). According to the dominant flora, the gut microbes are basically divided into 6 gates in terms of phylogenetic status, including *Firmicutes*, *Bacteroidetes*, *Proteobacteria*, *Actinobacteria*, *Micrococcus verrucosa* and *Fusobacteria*. At present, the classification of gut microbes is mainly based on the bacteria, thus ignoring the role of enterovirus and the correlation between enterovirus and bacteria.

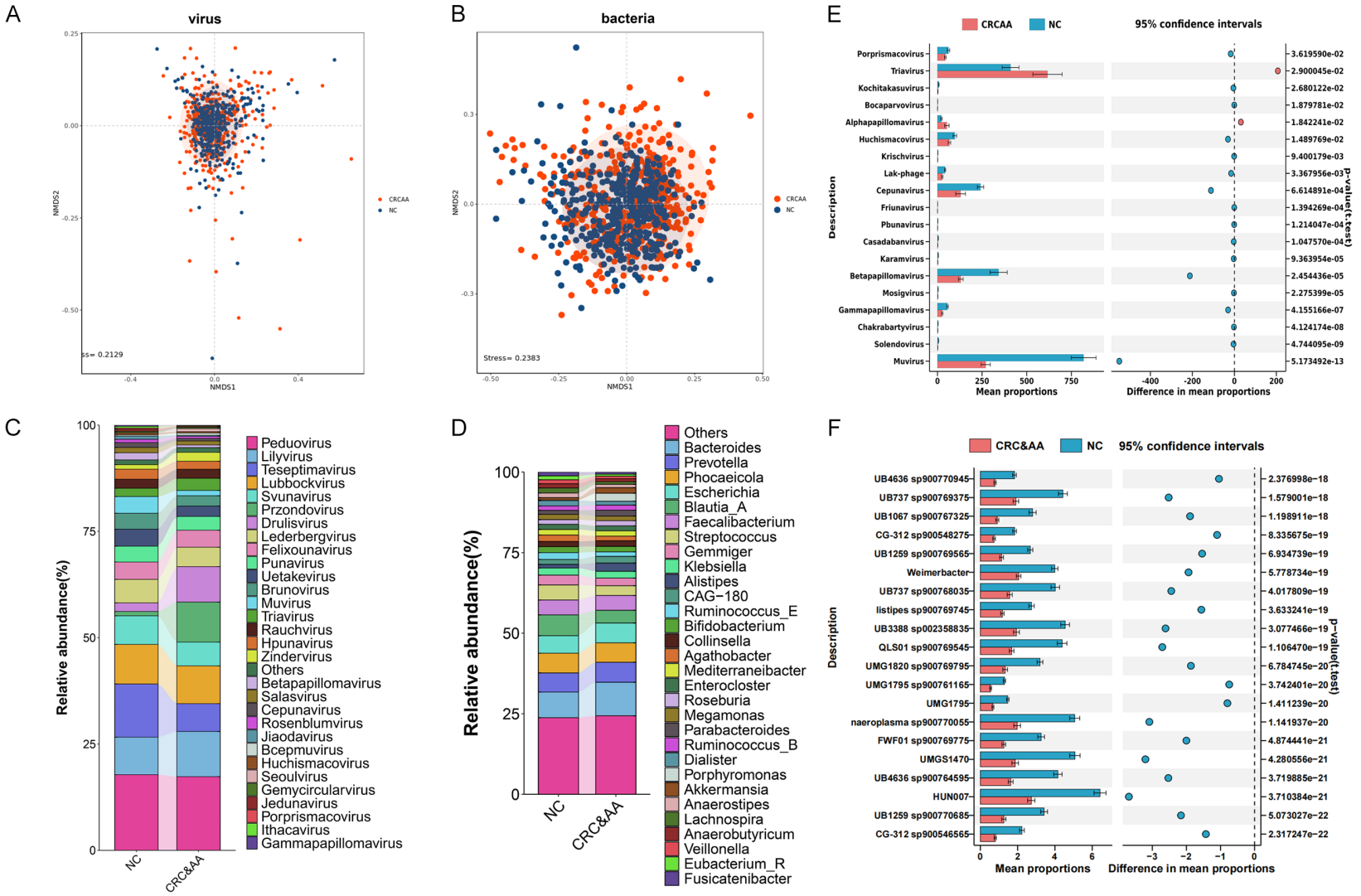
The human gut virome exhibits complex composition and remarkable interindividual variability, often referred to as the "dark matter" of the gut microbiome. Significant differences have been observed in the gut virome (particularly bacteriophage communities) between colorectal cancer (CRC)/adenoma patients and healthy individuals [30]. Altered virome diversity shows positive correlation with disease severity in both IBD and CRC patients [31, 32]. Current evidence suggests three potential mechanisms by which gut viruses may promote CRC pathogenesis: ① direct infection of intestinal epithelium triggering chronic inflammation and driving tissue dysplasia [33, 34]; ② indirect modulation of host physiology through regulation of bacterial community stability and composition [2]; and ③ interaction with host immune system to induce specific immune

Table 2. CRC clinical information of three gut microbial subtypes

CRC clinical information	V1 (n=100)	V2B1 (n=65)	V2B2 (n=90)	*p*-value
Male	56	43	54	0.554
Age	66.06±10.64	65.92±9.97	64.88±10.35	0.745
BMI	24.15±7.87	25.98±8.39	25.41±9.10	0.350
FOBT				0.239
OB (+)	59	40	65	
OB (-)	32	20	24	
NA	9	5	1	
TNM Stage				0.047
I	36	16	19	
II	23	18	27	
III	31	21	30	
IV	7	6	11	
NA	3	4	3	
Tumor location				0.329
colon	42	31	42	
rectum	56	32	42	
NA	2	2	6	
Pathological type				0.753
Protrude	34	26	30	
Ulcerative	20	13	18	
Infiltrating	8	2	10	
NA	38	24	32	
Differentiated degree				0.618
Poorly	29	11	22	
Moderately	57	44	54	
Highly	1	0	0	
NA	13	10	14	
Mismatch Repair, MMR				
pMMR	75	53	67	0.453
dMMR	1	0	2	
NA	24	12	21	

responses [35]. Nevertheless, the extreme diversity of viruses and technical limitations in sequencing have hindered the establishment of a comprehensive human gut virome reference database, resulting in scarce research on virome-based gut microbial subtyping [36]. Our study integrating virome data from metagenomic sequencing successfully identified three novel gut microbial subtypes through unsupervised clustering, providing valuable complementation to existing classification systems.

By analyzing the clinical information of the 3 subtypes, it was found that the disease ratios (CRCs&AAs/NCs) of the three subtypes were 1.06, 0.67 and 1.29, respectively, which indicated that enteroviruses and bacteria are indeed involved in the incidence of CRC and its precancerous lesions. Moreover, the characteristics of gut microbes of the 3 subtypes were analyzed. The abundance of *Peduovirus* in V1 subtype was significantly higher than that of V2 subtype. The abundances of *Bacteroides*, *Prevotella*, and *Phocaeicola* increased gradually in V1, V2B1 and V2B2, while the abundances of *Escherichia* increased significantly in the community structure proportion of V2B2. The current study proved a link between enteroviruses and inflammatory bowel disease (IBD), one of the precancerous lesions of CRC, with IBD patients showing a significant increase in *Caudovirales* and a decrease in bacterial rich-


A
virus
NMDS1
NMDS2
Stress= 0.2129
CRCAA
NC
B
bacteria
NMDS1
NMDS2
Stress= 0.2383
CRCAA
NC
C
Relative abundance(%)
NC
CRC&AA
Peduovirus
Lilyvirus
Teseptimavirus
Lubbockvirus
Svunavirus
Przondovirus
Drulisvirus
Lederbergvirus
Felixounavirus
Punavirus
Uetakevirus
Brunovirus
Muvirus
Triavirus
Rauchvirus
Hpunavirus
Zindervirus
Others
Betapapillomavirus
Salasvirus
Cepunavirus
Rosenblumvirus
Jiaodavirus
Bcepmuvirus
Huchismacovirus
Seoulvirus
Gemycircularvirus
Jedunavirus
Porprismacovirus
Ithacavirus
Gammapapillomavirus
D
Relative abundance(%)
NC
CRC&AA
Others
Bacteroides
Prevotella
Phocaeicola
Escherichia
Blautia_A
Faecalibacterium
Streptococcus
Gemmiger
Klebsiella
Alistipes
CAG−180
Ruminococcus_E
Bifidobacterium
Collinsella
Agathobacter
Mediterraneibacter
Enterocloster
Roseburia
Megamonas
Parabacteroides
Ruminococcus_B
Dialister
Porphyromonas
Akkermansia
Anaerostipes
Lachnospira
Anaerobutyricum
Veillonella
Eubacterium_R
Fusicatenibacter
E
CRCAA
NC
95% confidence intervals
Description
p-value(t.test)
Porprismacovirus 3.619590e-02
Triavirus 2.900045e-02
Kochitakasuvirus 2.680122e-02
Bocaparvovirus 1.879781e-02
Alphapapillomavirus 1.842241e-02
Huchismacovirus 1.489769e-02
Krischvirus 9.400179e-03
Lak-phage 3.367956e-03
Cepunavirus 6.614891e-04
Friunavirus 1.394269e-04
Pbunavirus 1.214047e-04
Casadabanvirus 1.047570e-04
Karamvirus 9.363954e-05
Betapapillomavirus 2.454436e-05
Mosigvirus 2.275399e-05
Gammapapillomavirus 4.155166e-07
Chakrabartyvirus 4.124174e-08
Solendovirus 4.744095e-09
Muvirus 5.173492e-13
Mean proportions
Difference in mean proportions
F
CRC&AA
NC
95% confidence intervals
Description
p-value(t.test)
UB4636 sp900770945 2.376998e−18
UB737 sp900769375 1.579001e−18
UB1067 sp900767325 1.198911e−18
CG-312 sp900548275 8.335675e−19
UB1259 sp900769565 6.934739e−19
Weimerbacter 5.778734e−19
UB737 sp900768035 4.017809e−19
listipes sp900769745 3.633241e−19
UB3388 sp002358835 3.077466e−19
QLS01 sp900769545 1.106470e−19
UMG1820 sp900769795 6.784745e−20
UMG1795 sp900761165 3.742401e−20
UMG1795 1.411239e−20
naeroplasma sp900770055 1.141937e−20
FWF01 sp900769775 4.874441e−21
UMGS1470 4.280556e−21
UB4636 sp900764595 3.719885e−21
HUN007 3.710384e−21
UB1259 sp900770685 5.073027e−22
CG-312 sp900546565 2.317247e−22
Mean proportions
Difference in mean proportions

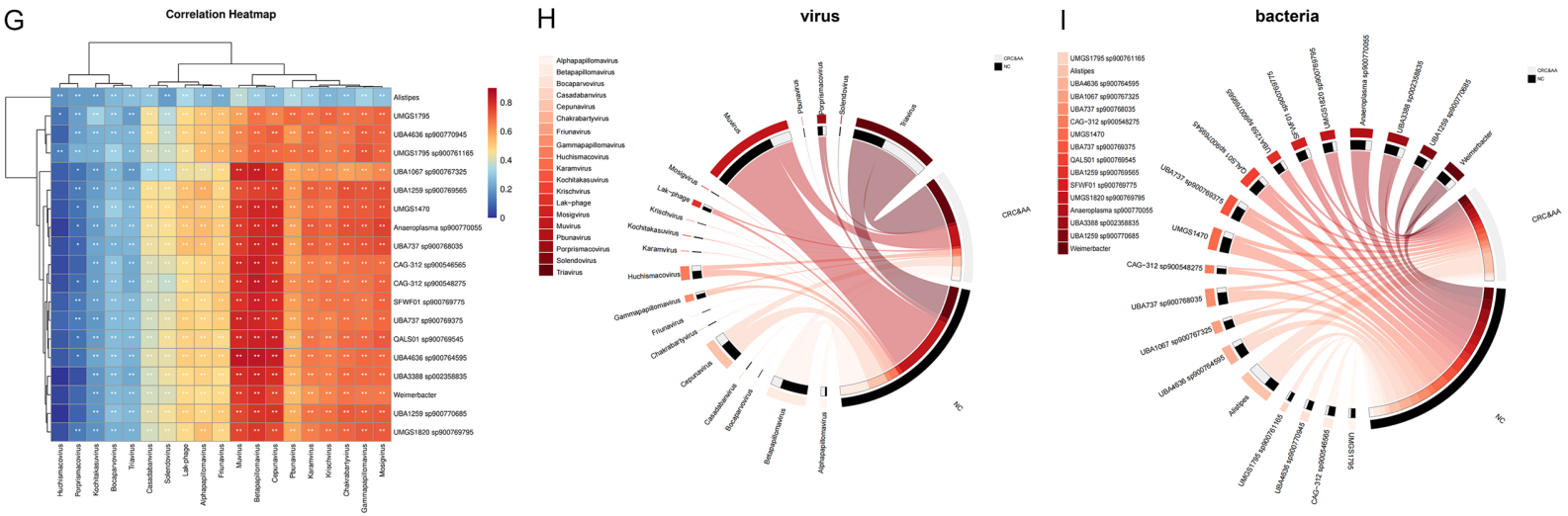


Figure 3. CRC and its precancerous lesions related gut microbes. A, B: NMDS analysis of enteroviruses and bacteria beta diversity, respectively. C, D: Percentile chart of enterovirus and bacterial community. E, F: Wilcoxon rank-sum test was used to detect the differential enteroviruses and bacteria, respectively. G: Spearman correlation analysis between differential enteroviruses and differential bacteria in NCs. H: Intergroup correlation analysis of differential enterovirus. I: Intergroup correlation analysis of differential bacteria.

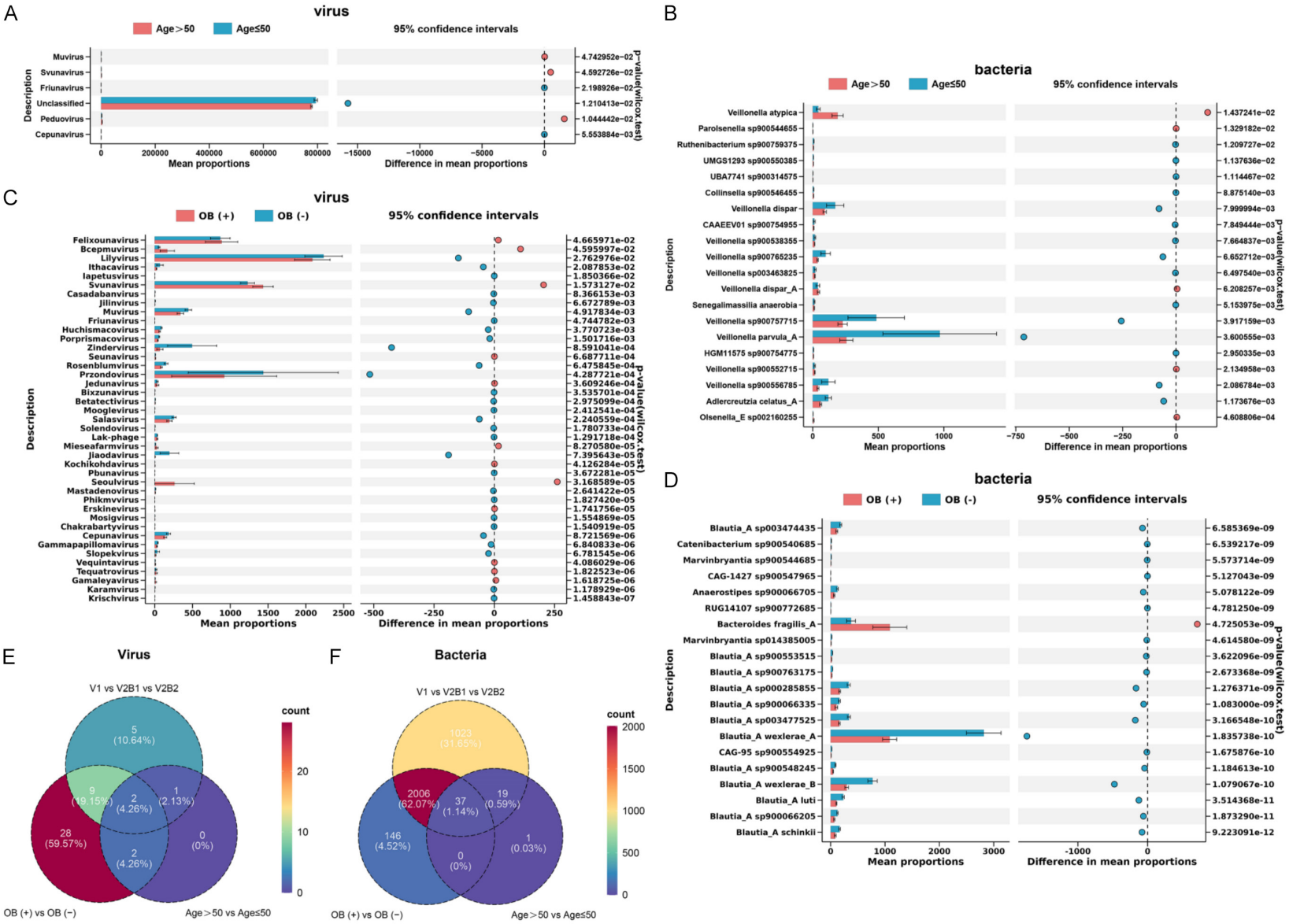


Figure 4. Gut microbial characteristics based on differential clinical features. A, B: Differential enteroviruses and differential bacteria between age > 50 years old and age ≤ 50 years old. C, D: Differential enteroviruses and differential bacteria between FOBT positive and negative. E, F: Venn diagram showed different viruses and bacteria based on grouping.

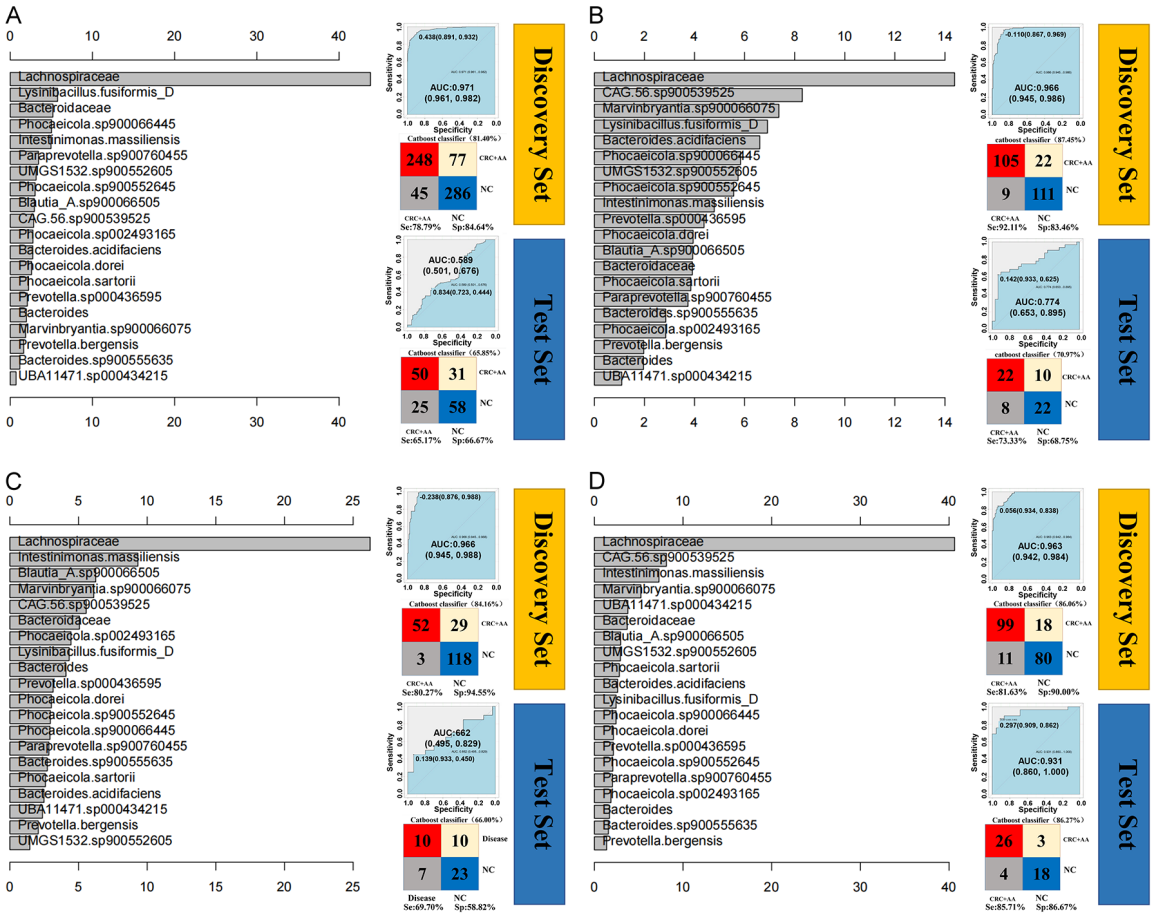


Figure 5. CRC&AA disease prediction model. A: Disease prediction Catboost model based on differential gut microbes of CRC&AA and NC. B: Disease prediction models of V1 subtype. C: Disease prediction models of V2B1 subtype. D: Disease prediction models of V2B2 subtype.

ness and diversity [37]. As a member of the *Polyomaviridae family*, *John Cunningham virus* (JCV) has been found to be associated with CRC chromosomal instability [38]. There are few clinical or mechanistic studies on the relationship between viruses and CRC. The results of our study will open the research idea of the relationship between enterovirus and CRC. Furthermore, the higher positive rate of the fecal occult blood test (FOBT) in the V2B2 subtype is consistent with our finding that this subtype harbors the highest proportion of patients with CRC/adenoma. As FOBT positivity is a well-established indicator for colorectal bleeding and CRC risk, this result validates the clinical relevance of the V2B2 subtype. Furthermore, the co-enrichment of the high-risk bacterium *Citrobacter farreri* and *Svunavirus* within V2B2 suggests a potential microbial synergy that may contribute to a more aggressive pathogenic environment, potentially explaining the elevated CRC risk. This observation underscores the potential of our integrated virus-bacteria typing approach to improve CRC risk stratification.

At the same time, the relationship between gut microbes and human metabolites and metal ions was linked, and the differences of fecal metabolites and ions of different subtypes of gut microbes were analyzed. Differential metabolites (Benzene and substituted derivatives, etc.) were screened among the three subtypes. The content of Imidazopyrimidines increased in V1, benzene and substituted derivatives increased in V1B1, and organic sulfonic acids and derivatives increased in V2B2. Coker et al. compared gut metabolites between 118 CRC patients, 140 colorectal adenomas patients, and 128 normal controls, and it was found that n-valine and myristate increased from the NC group to the CRA group and to the CRC group.

Table 3. CRC&AA Disease risk prediction model

	Discovery Set			Test Set		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Normal	77.86%	78.41%	77.35%	73.33%	75%	71.91%
V1	82.49%	80.95%	93.46%	74.19%	69.44%	80.77%
V2B1	86.14%	84.44%	89.55%	70%	70.27%	69.23%
V2B2	93.24%	85.24%	91.87%	86.54%	80.77%	92.31%

The metabolites associated with CRC were enriched in branched chain amino acids, aromatic amino acids, and aminoacyl-tRNA biosynthesis pathways [39]. Different ions Ni between V1 and V2B1 and different ions As between V2B1 and V2B2 were found. It has been reported that Arsenic exposure increases the incidence of malignant tumors [40]. The correlation between CRC-associated metabolites and gut metal ions and gut microbes changed at different CRC stages. This study indicates that there may be a potentially certain relationship or interference between altered metabolites and ions in the stage of CRC development and gut microbial changes. The changes in gut ion content found in this study based on microbial typing are a new exploration for CRC research.

In addition, gut disease prediction model was more accurate after virus typing. CRC is a heterogeneous disease, and individualized treatment is needed to optimize treatment and reduce CRC-related morbidity and mortality. Focusing on tumor heterogeneity and genetic mutations, it has been reported that different CRC subtypes better understand the biological characteristics that distinguish patients with CRC [41]. Considering enterovirus based on the results of metagenomic sequencing, three new subtypes of gut microbes were constructed based on unsupervised clustering, which is a supplement to gut microbes typing. Our research results indicate that viral typing significantly improves the accuracy of intestinal disease prediction models, achieving an accuracy rate of 93.24% in the training set and 86.54% in the test set, which is superior to many reported microbiome prediction models. For instance, the "Four-kingdom" model (bacteria, fungi, archaea, and viruses) constructed by Liu et al. is currently the CRC microbiological diagnosis study with the largest sample size (n=1368). However, the AUROC of its single-kingdom bacterial model was only 0.80, while the AUROC of the four-kingdom combined model increased to 0.83 [42]. In contrast, our model demonstrated higher predictive performance by integrating virus-bacterial characteristics. It indicates that it has a stronger generalization ability. Furthermore, although the research of Qin's team developed a complex microbial-metabolite diagnostic detection group (AUROC= 0.912 for adenoma and AUROC=0.994 for CRC) [43], it did not perform stratified optimization for CRC molecular subtypes 4. Our model further enhances the subtype-specific prediction ability through viral typing (V1/V2/V2B2), compensating for the limitations of traditional microbial models in differentiating precancerous lesions (such as adenomas) from CRC. Our research has for the first time confirmed that virus typing can significantly improve the accuracy of CRC prediction, and has optimized the processing capacity of high-dimensional microbial data through the machine learning algorithm (CatBoost), making it more advantageous in clinical transformation. These findings provide new strategies for the precise classification screening and individualized intervention of CRC, especially having significant application value in the early identification of high-risk populations. The combined enterovirus and bacteria for microbiological typing provide a new strategy for CRC risk prediction, which would more accurately locate CRC high-risk groups based on gut microbial typing.

However, there are still some shortcomings in this study. First, it is still necessary to expand the sample size to strengthen the feasibility of enterovirus combined with bacteria for gut microbial typing. Future research will delve into the correlation between subtypes and microorganisms, subtypes and metabolites, and subtypes and ions. This analysis aims to define characteristic gut microbes, as well as representative metabolites and ions for different subtypes. These findings will furnish a more robust foundation for investigating CRC and its precancerous lesions through the lens of gut microbial typing.

Conclusion

Unsupervised clustering was used to identify three subtypes of gut microbes (V1, V2B1 and V2B2 subtypes) based on the combination of enteroviruses and gut bacteria. These subtypes were associated with the incidence of CRC and its precancerous lesions. Most importantly, microbial subtypes were associated with corresponding changes in fecal metabolites and ion content. Compared with bacterial subtypes or without unsupervised clustering, the combination of enteroviral and bacterial subtypes is expected to better identify CRC.

Acknowledgements

This work was supported by Medical and Health Research Project of Zhejiang Province (No. 2025KY1531) and Public Welfare Technology Application Research Program of Huzhou (No. 2024GY22).

All patients volunteered to participate in the study and signed a written informed consent.

Disclosure of conflict of interest

None.

Address correspondence to: Yinhang Wu, Huzhou Central Hospital, Affiliated Central Hospital Huzhou University, No. 1558, Sanhuan North Road, Wuxing District, Huzhou 313000, Zhejiang, The People's Republic of China. E-mail: wuyinhang2482@hzhospital.com; bawnywuyinhang@163.com

References

[1] Huang J, Lucero-Prisno DE 3rd, Zhang L, Xu W, Wong SH, Ng SC and Wong MCS. Updated epidemiology of gastrointestinal cancers in East Asia. Nat Rev Gastroenterol Hepatol 2023; 20: 271-287.

[2] Janney A, Powrie F and Mann EH. Host-microbiota maladaptation in colorectal cancer. Nature 2020; 585: 509-517.

[3] Yang J, Wei H, Zhou Y, Szeto CH, Li C, Lin Y, Coker OO, Lau HCH, Chan AWH, Sung JJY and Yu J. High-fat diet promotes colorectal tumorigenesis through modulating gut microbiota and metabolites. Gastroenterology 2022; 162: 135-149, e2.

[4] Chapelle N, Martel M, Toes-Zoutendijk E, Barkun AN and Bardou M. Recent advances in clinical practice: colorectal cancer chemoprevention in the average-risk population. Gut 2020; 69: 2244-2255.

[5] Sánchez-Alcoholado L, Ordóñez R, Otero A, Plaza-Andrade I, Laborda-Illanes A, Medina JA, Ramos-Molina B, Gómez-Millán J and Queipo-Ortuño MI. Gut microbiota-mediated inflammation and gut permeability in patients with obesity and colorectal cancer. Int J Mol Sci 2020; 21: 6782.

[6] Grivennikov SI, Greten FR and Karin M. Immunity, inflammation, and cancer. Cell 2010; 140: 883-899.

[7] Peterson LW and Artis D. Intestinal epithelial cells: regulators of barrier function and immune homeostasis. Nat Rev Immunol 2014; 14: 141-153.

[8] Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P, Segata N, Kyrpides NC and Finn RD. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol 2021; 39: 105-114.

[9] Li J, Yang F, Xiao M and Li A. Advances and challenges in cataloging the human gut virome. Cell Host Microbe 2022; 30: 908-916.

[10] Iftekhar A, Berger H, Bouznad N, Heuberger J, Boccellato F, Dobrindt U, Hermeking H, Sigal M and Meyer TF. Genomic aberrations after short-term exposure to colibactin-producing E. coli transform primary colon epithelial cells. Nat Commun 2021; 12: 1003.

[11] Cao Y, Wang Z, Yan Y, Ji L, He J, Xuan B, Shen C, Ma Y, Jiang S, Ma D, Tong T, Zhang X, Gao Z, Zhu X, Fang JY, Chen H and Hong J. Enterotoxigenic bacteroidesfragilis promotes intestinal inflammation and malignancy by inhibiting exosome-packaged miR-149-3p. Gastroenterology 2021; 161: 1552-1566, e12.

[12] Rubinstein MR, Wang X, Liu W, Hao Y, Cai G and Han YW. Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. Cell Host Microbe 2013; 14: 195-206.

[13] Meng Q, Sun H, Wu S, Familiari G, Relucenti M, Aschner M, Li X and Chen R. Epstein-Barr Virus-Encoded MicroRNA-BART18-3p promotes colorectal cancer progression by targeting de novo lipogenesis. Adv Sci (Weinh) 2022; 9: e2202116.

[14] Wahida A, Tang F and Barr JJ. Rethinking phage-bacteria-eukaryotic relationships and their influence on human health. Cell Host Microbe 2021; 29: 681-688.

[15] Erickson AK, Jesudhasan PR, Mayer MJ, Narbad A, Winter SE and Pfeiffer JK. Bacteria facilitate enteric virus co-infection of mammalian cells and promote genetic recombination. Cell Host Microbe 2018; 23: 77-88, e5.

[16] Hussain FA, Dubert J, Elsherbini J, Murphy M, VanInsberghe D, Arevalo P, Kauffman K, Rodino-Janeiro BK, Gavin H, Gomez A, Lopatina A,

Le Roux F and Polz MF. Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. Science 2021; 374: 488-492.

[17] Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen HB, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, Torrents D, Ugarte E, Zoetendal EG, Wang J, Guarner F, Pedersen O, de Vos WM, Brunak S, Doré J; MetaHIT Consortium; Antolín M, Artiguenave F, Blottiere HM, Almeida M, Brechot C, Cara C, Chervaux C, Cultrone A, Delorme C, Denariaz G, Dervyn R, Foerstner KU, Friss C, van de Guchte M, Guedon E, Haimet F, Huber W, van Hylckama-Vlieg J, Jamet A, Juste C, Kaci G, Knol J, Lakhdari O, Layec S, Le Roux K, Maguin E, Mérieux A, Melo Minardi R, M'rini C, Muller J, Oozeer R, Parkhill J, Renault P, Rescigno M, Sanchez N, Sunagawa S, Torrejon A, Turner K, Vandemeulebrouck G, Varela E, Winogradsky Y, Zeller G, Weissenbach J, Ehrlich SD and Bork P. Enterotypes of the human gut microbiome. Nature 2011; 473: 174-180.

[18] Krautkramer KA, Fan J and Bäckhed F. Gut microbial metabolites as multi-kingdom intermediates. Nat Rev Microbiol 2021; 19: 77-94.

[19] Cani PD, Van Hul M, Lefort C, Depommier C, Rastelli M and Everard A. Microbial regulation of organismal energy homeostasis. Nat Metab 2019; 1: 34-46.

[20] Aron-Wisnewsky J, Warmbrunn MV, Nieuwdorp M and Clément K. Metabolism and metabolic disorders and the microbiome: the intestinal microbiota associated with obesity, lipid metabolism, and metabolic health-pathophysiology and therapeutic strategies. Gastroenterology 2021; 160: 573-599.

[21] Das NK, Schwartz AJ, Barthel G, Inohara N, Liu Q, Sankar A, Hill DR, Ma X, Lamberg O, Schnizlein MK, Arqués JL, Spence JR, Nunez G, Patterson AD, Sun D, Young VB and Shah YM. Microbial metabolite signaling is required for systemic iron homeostasis. Cell Metab 2020; 31: 115-130, e6.

[22] Bolger AM, Lohse M and Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014; 30: 2114-2120.

[23] Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P, Segata N, Kyrpides NC and Finn RD. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol 2021; 39: 105-114.

[24] Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, Proal AD, Fischbach MA, Bhatt AS, Hugenholtz P and Kyrpides NC. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nat Microbiol 2021; 6: 960-970.

[25] Wood DE and Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 2014; 15: R46.

[26] Wilkerson MD and Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 2010; 26: 1572-1573.

[27] Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS and Huttenhower C. Metagenomic biomarker discovery and explanation. Genome Biol 2011; 12: R60.

[28] Ijaz MU, Ahmed MI, Zou X, Hussain M, Zhang M, Zhao F, Xu X, Zhou G and Li C. Beef, casein, and soy proteins differentially affect lipid metabolism, triglycerides accumulation and gut microbiota of high-fat diet-fed C57BL/6J mice. Front Microbiol 2018; 9: 2200.

[29] Yinhang W, Jing Z, Jie Z, Yin J, Xinyue W, Yifei S, Zhiqing F, Wei W and Shuwen H. Prediction model of colorectal cancer (CRC) lymph node metastasis based on intestinal bacteria. Clin Transl Oncol 2023; 25: 1661-1672.

[30] Chen F, Li S, Guo R, Song F, Zhang Y, Wang X, Huo X, Lv Q, Ullah H, Wang G, Ma Y, Yan Q and Ma X. Meta-analysis of fecal viromes demonstrates high diagnostic potential of the gut viral signatures for colorectal cancer and adenoma risk assessment. J Adv Res 2023; 49: 103-114.

[31] Duerkop BA, Kleiner M, Paez-Espino D, Zhu W, Bushnell B, Hassell B, Winter SE, Kyrpides NC and Hooper LV. Murine colitis reveals a disease-associated bacteriophage community. Nat Microbiol 2018; 3: 1023-1031.

[32] Gogokhia L, Buhrke K, Bell R, Hoffman B, Brown DG, Hanke-Gogokhia C, Ajami NJ, Wong MC, Ghazaryan A, Valentine JF, Porter N, Martens E, O'Connell R, Jacob V, Scherl E, Crawford C, Stephens WZ, Casjens SR, Longman RS and Round JL. Expansion of bacteriophages is linked to aggravated intestinal inflammation and colitis. Cell Host Microbe 2019; 25: 285-299, e8.

[33] Liu W, Lau HCH, Ding X, Yin X, Wu WKK, Wong SH, Sung JJY, Zhang T and Yu J. Transmission of antimicrobial resistance genes from the environment to human gut is more pronounced in colorectal cancer patients than in healthy subjects. Imeta 2025; 4: e70008.

[34] Wang Z, Guo K, Liu Y, Huang C and Wu M. Dynamic impact of virome on colitis and colorec-

tal cancer: immunity, inflammation, prevention and treatment. Semin Cancer Biol 2022; 86: 943-954.
[35] Duerkop BA and Hooper LV. Resident viruses and their interactions with the immune system. Nat Immunol 2013; 14: 654-659.
[36] Li J, Yang F, Xiao M and Li A. Advances and challenges in cataloging the human gut virome. Cell Host Microbe 2022; 30: 908-916.
[37] Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, Kambal A, Monaco CL, Zhao G, Fleshner P, Stappenbeck TS, McGovern DP, Keshavarzian A, Mutlu EA, Sauk J, Gevers D, Xavier RJ, Wang D, Parkes M and Virgin HW. Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell 2015; 160: 447-460.
[38] Goel A, Li MS, Nagasaka T, Shin SK, Fuerst F, Ricciardiello L, Wasserman L and Boland CR. Association of JC virus T-antigen expression with the methylator phenotype in sporadic colorectal cancers. Gastroenterology 2006; 130: 1950-1961.
[39] Coker OO, Liu C, Wu WKK, Wong SH, Jia W, Sung JJY and Yu J. Altered gut metabolites and microbiota interactions are implicated in colorectal carcinogenesis and can be non-invasive diagnostic biomarkers. Microbiome 2022; 10: 35.
[40] Smith AH, Marshall G, Roh T, Ferreccio C, Liaw J and Steinmaus C. Lung, bladder, and kidney cancer mortality 40 years after arsenic exposure reduction. J Natl Cancer Inst 2018; 110: 241-249.
[41] Wu Y, Zhuang J, Qu Z, Yang X and Han S. Advances in immunotyping of colorectal cancer. Front Immunol 2023; 14: 1259461.
[42] Liu NN, Jiao N, Tan JC, Wang Z, Wu D, Wang AJ, Chen J, Tao L, Zhou C, Fang W, Cheong IH, Pan W, Liao W, Kozlakidis Z, Heeschen C, Moore GG, Zhu L, Chen X, Zhang G, Zhu R and Wang H. Multi-kingdom microbiota analyses identify bacterial-fungal interactions and biomarkers of colorectal cancer across cohorts. Nat Microbiol 2022; 7: 238-250.
[43] Gao R, Wu C, Zhu Y, Kong C, Zhu Y, Gao Y, Zhang X, Yang R, Zhong H, Xiong X, Chen C, Xu Q and Qin H. Integrated analysis of colorectal cancer reveals cross-cohort gut microbial signatures and associated serum metabolites. Gastroenterology 2022; 163: 1024-1037, e9.

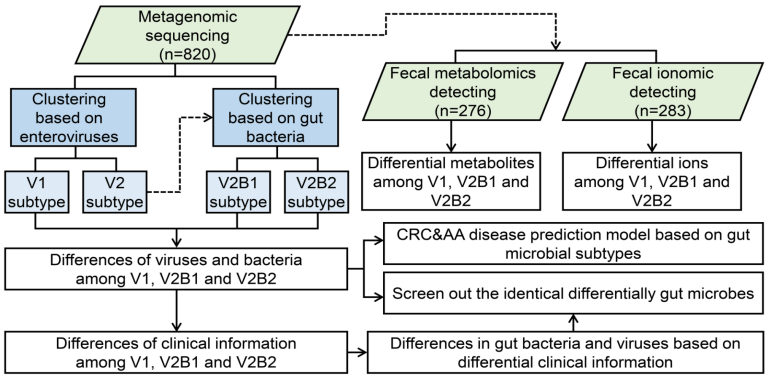
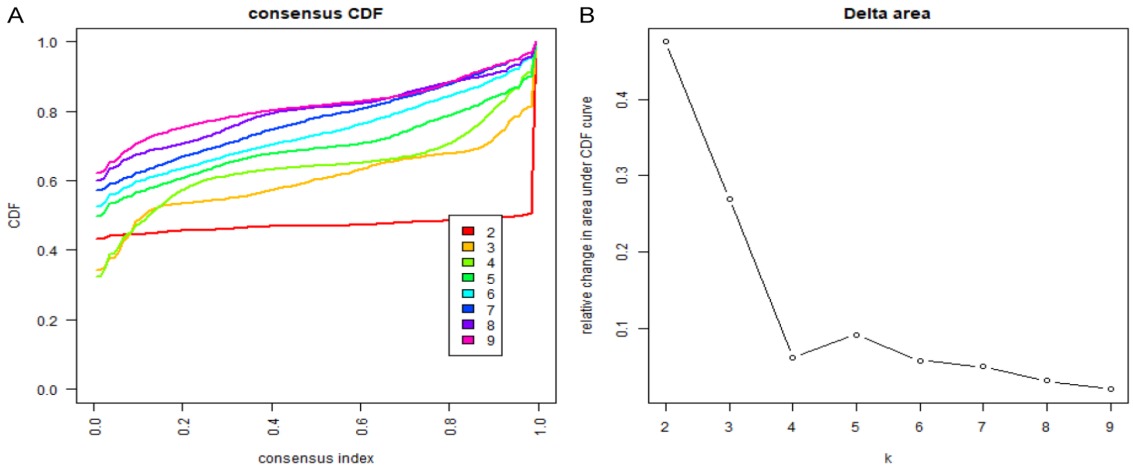


Figure S1. The flow chart of this study. Stool samples were collected from 820 patients for metagenomic sequencing. The analytical procedure was as follows: Initially, unsupervised clustering based on virome data categorized patients into two clusters, V1 and V2. Subsequently, using bacterial metagenomic data, a secondary clustering analysis was conducted specifically on patients within the V2 cluster (as the bacterial clustering within V1 was ineffective, it was not subdivided), leading to the identification of two subgroups, V2B1 and V2B2. Thus, all patients were classified into three final subtypes: V1, V2B1, and V2B2. The study comprehensively compared the differences in gut microorganisms (viruses and bacteria), metabolites, ions, and clinical information across these subtypes. Finally, a disease risk prediction model for colorectal cancer (CRC) was constructed based on this typing system.

Table S1. Clinical information of three groups

	CRC (n=255)	AA (n=151)	NC (n=414)	*p*-value
Sex				< 0.001
Male	153	96	154	
Female	102	55	260	
Age	65.61±10.34	59.82±8.03	56.98±8.03	< 0.001
BMI	24.74±7.74	24.71±5.84	23.42±4.54	0.021
FOBT				< 0.001
OB (+)	164	33	47	
OB (-)	76	99	280	
NA	15	19	87	
White blood count, WBC	6.09±3.18	6.06±3.30	6.23±3.76	0.881
Albumin	37.59±4.61	39.61±4.22	42.16±3.70	< 0.001
Triglyceride, TG	1.30±0.82	1.57±1.89	1.43±0.88	0.178
Total cholesterol, TC	4.82±4.65	4.97±4.96	4.84±4.05	0.923
High density lipoprotein, HDL	45.13±15.89	48.08±17.24	47.80±15.61	0.002
Low density lipoprotein, LDL	91.78±31.03	94.52±30.40	95.71±31.26	0.008



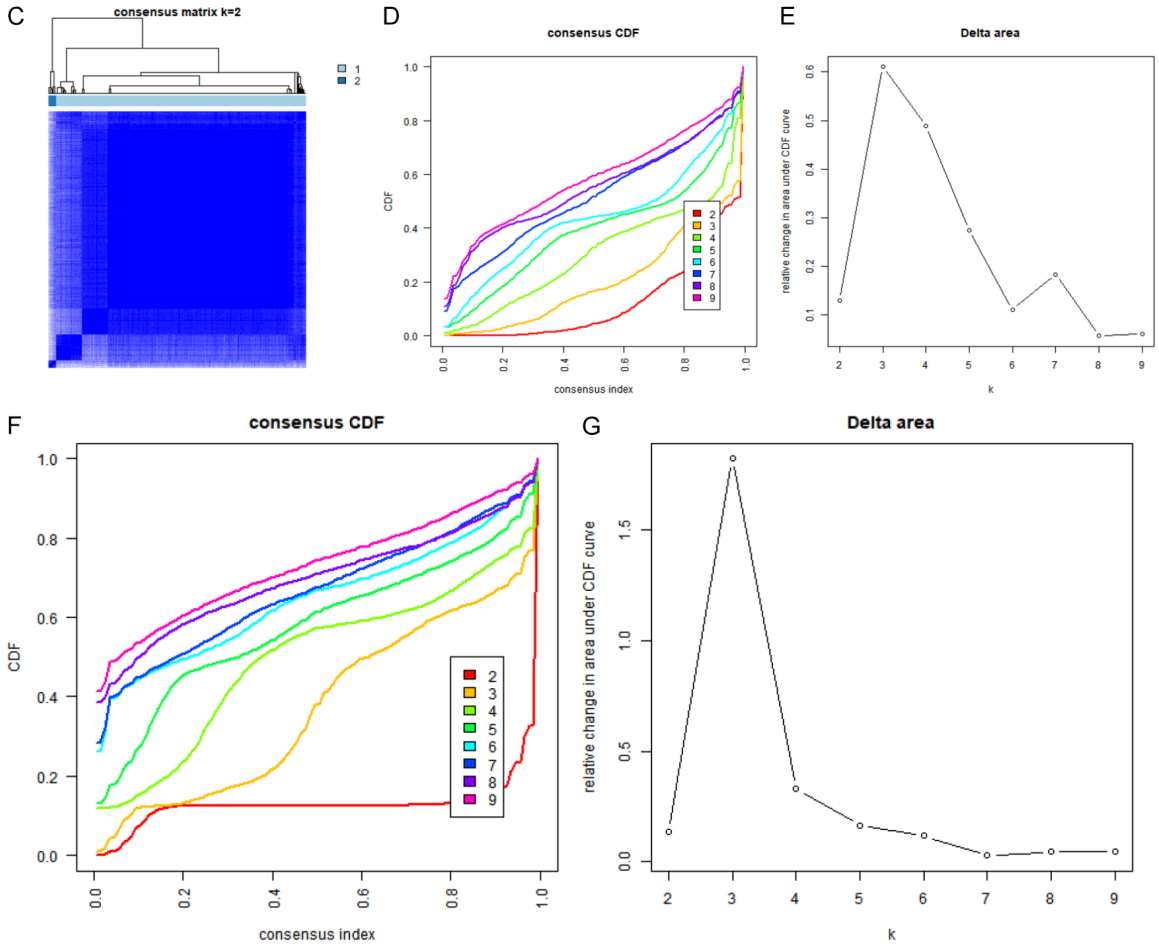


Figure S2. Unsupervised clustering process using ConsensusClusterPlus.

Table S2. Quantitative analysis for selecting the optimal k-value in ConsensusClusterPlus

k	PAC (Virus: V1 and V2)	PAC (Bacteria: V2B1 and V2B2)
2	0.0476601142159434	0.230986483162484
3	0.234092303315344	0.550574784178099
4	0.327666989695671	0.646007628226215
5	0.27284392957375	0.671672134451191
6	0.292107127086723	0.674194131540967
7	0.303353870865451	0.675474393283368
8	0.230915357510128	0.582050848914296
9	0.220359125273019	0.521964490518061

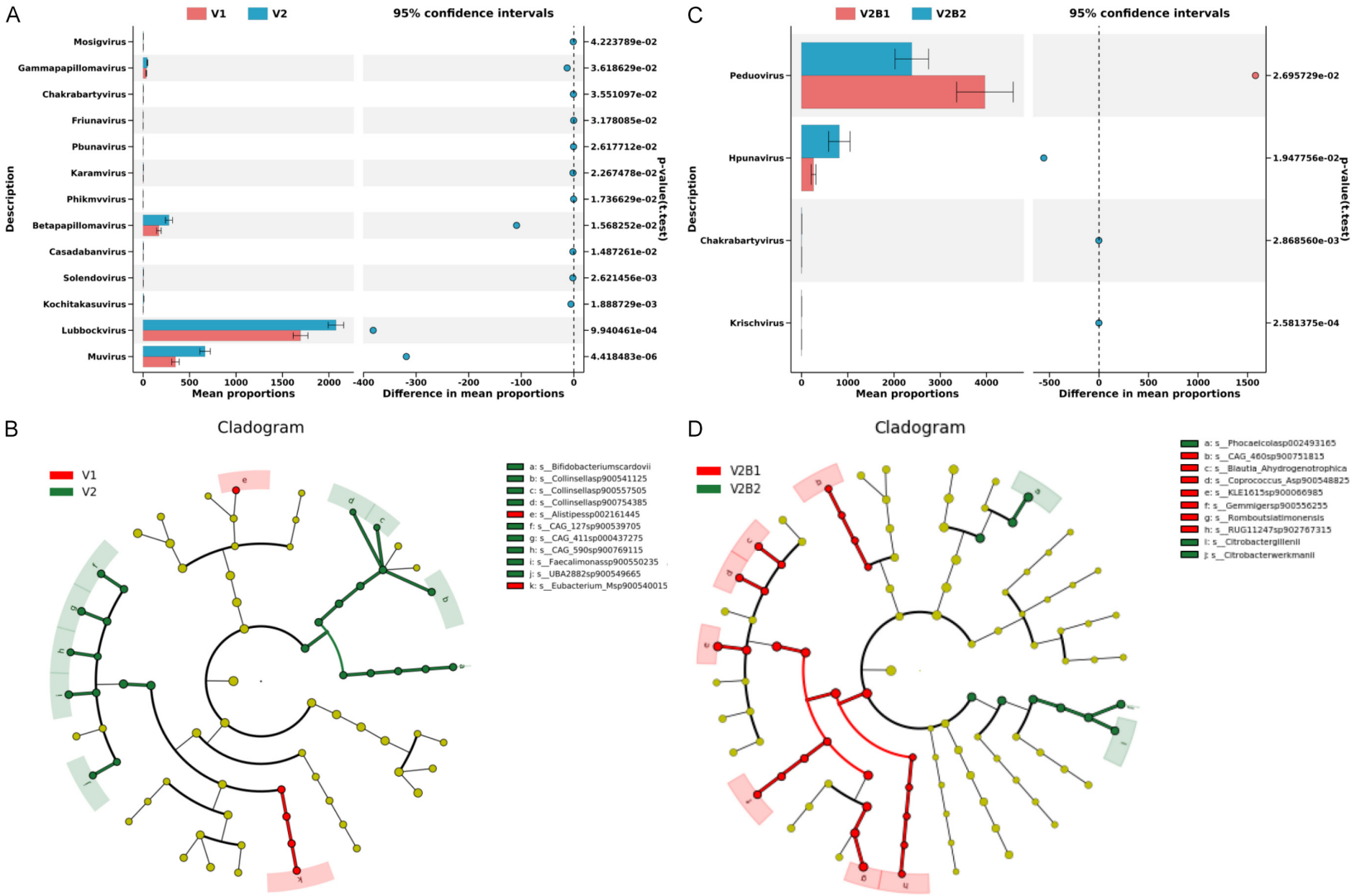


Figure S3. Differential gut microbes based on enterovirus and bacterial typing.

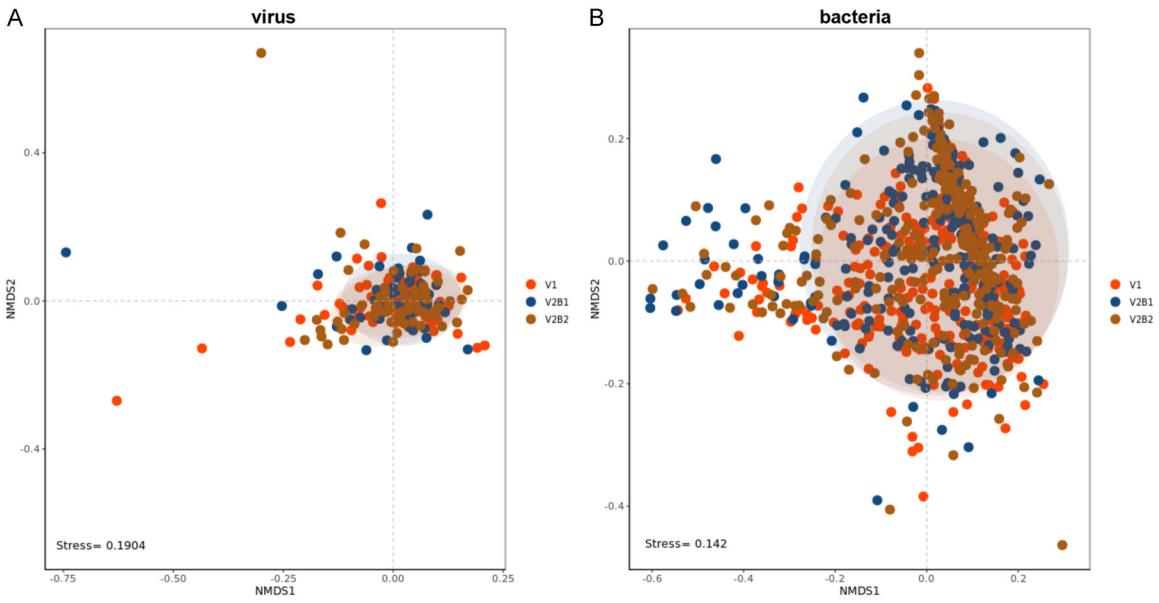


Figure S4. NMDS analysis of enterovirus and bacteria in three subtypes.

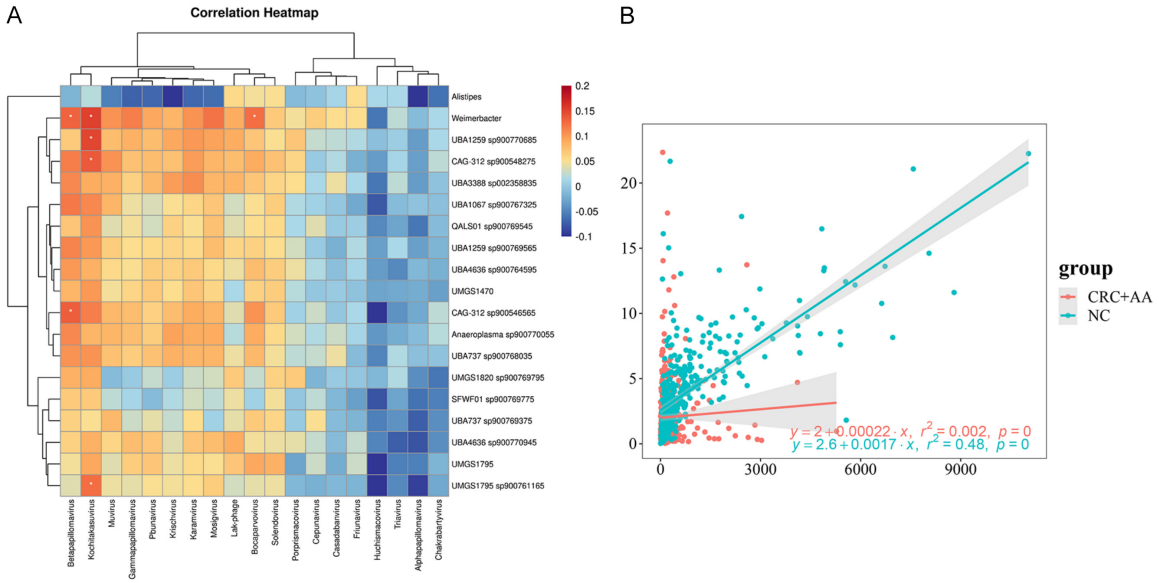


Figure S5. Correlation between bacteria and viruses in the CRCs and AAs.