Genome Medicine

RESEARCH

Open Access

The Chinese gut virus catalogue reveals gut virome diversity and disease-related viral signatures

Qiulong Yan[1,2*†], Liansha Huang[3†], Shenghui Li[4†], Yue Zhang[4†], Ruochun Guo[1†], Pan Zhang[5†], Zhixin Lei[6*], Qingbo Lv[4], Fang Chen[1,2], Zhiming Li[7], Jinxin Meng[4], Jing Li[8], Guangyang Wang[1,9], Changming Chen[10], Hayan Ullah[2], Lin Cheng[2], Shao Fan[2], Wei You[11], Yan Zhang[12], Jie Ma[13], Shanshan Sha[2*] and Wen Sun[14,15*]

Abstract

Background The gut viral community has been increasingly recognized for its role in human physiology and health; however, our understanding of its genetic makeup, functional potential, and disease associations remains incomplete.

Methods In this study, we collected 11,286 bulk or viral metagenomes from fecal samples across large-scale Chinese populations to establish a Chinese Gut Virus Catalogue (cnGVC) using a de novo virus identification approach. We then examined the diversity and compositional patterns of the gut virome in relation to common diseases by analyzing 6311 bulk metagenomes representing 28 disease or unhealthy states.

Results The cnGVC contains 93,462 nonredundant viral genomes, with over 70% of these being novel viruses not included in existing gut viral databases. This resource enabled us to characterize the functional diversity and specificity of the gut virome. Using cnGVC, we profiled the gut virome in large-scale populations, assessed sex- and age-related variations, and identified 4238 universal viral signatures of diseases. A random forest classifier based on these signatures achieved high accuracy in distinguishing diseased individuals from controls (AUC = 0.698) and high-risk patients from controls (AUC = 0.761), and its predictive ability was also validated in external cohorts.

Conclusions Our resources and findings significantly expand the current understanding of the human gut virome and provide a comprehensive view of the associations between gut viruses and common diseases. This will pave the way for novel strategies in the treatment and prevention of these diseases.

Keywords Gut viral community, Gut virome, Viral metagenomics, Chinese Gut Virus Catalogue (cnGVC), Disease-associated viruses

†Qiulong Yan, Liansha Huang, Shenghui Li, Yue Zhang, Ruochun Guo, and Pan Zhang contributed equally to this work.

*Correspondence:
Qiulong Yan
qiulongy1988@163.com
Zhixin Lei
leizhixin@whut.edu.cn
Shanshan Sha
shanshan_s@126.com
Wen Sun
sunwen@bucm.edu.cn
Full list of author information is available at the end of the article

Background

The viral community in our gut, referred to as the human gut virome, is a key component of the gut microbial community with extensive unexplored genetic and functional diversity [1]. Traditional gut virology research usually has typically focused on some specific enteroviruses [2, 3], but its scope and applications have been limited due to the insufficient discovery of viruses. However, with the rapid advancement of high-throughput whole-metagenome (bulk) sequencing and virus-like particle (VLP)-based virome sequencing technologies, preliminary insights into the characteristics of the gut virome have been gained. In general, the normal gut viral composition is highly individualized and influenced by factors such as sex, age, geography, and lifestyle [4–7]. Longitudinal studies have shown that the gut virome in healthy adults is temporally stable, although it can fluctuate with changes in the external environment [4]. Significant alterations in the gut virome have been observed in a wide range of gastrointestinal and systemic disorders, including colorectal cancer (CRC) [8, 9], inflammatory bowel disease (IBD) [10], necrotizing enterocolitis [11], liver disease [12, 13], autoimmune diseases [14–17], metabolic syndrome [18], or even infectious diseases such as acquired immunodeficiency syndrome (AIDS) [19] and COVID-19 [20, 21]. These efforts underscore the critical role of the gut virome in human health and highlight the need for an in-depth exploration of its variation pattern and pathophysiological roles.

Reference viral genome catalogues are essential for quantitative and functional analyses of the human gut virome [22]. Recent studies have identified a vast array of viruses from publicly available fecal metagenomes, leading to the creation of databases such as the Gut Virome Database (GVD) [6], Gut Phage Database (GPD) [1], and Metagenomic Gut Virus (MGV) catalogue [23], each containing tens of thousands of viral genomes. An intriguing observation is that the viral sequences identified by different catalogues substantially differ, with less than 50% overlap (Additional file 1: Fig. S1a). Even in the same catalogue, only a small proportion of viruses are shared between samples from different regions (Additional file 1: Fig. S1b). Although technical variations may account for some differences, these findings strongly suggest that the gut virome is highly heterogeneous among populations, as supported by other studies [7, 24]. With the rapid expansion of bulk and VLP-based virome metagenomic samples, the representativeness of the gut viral catalogues can be significantly enhanced by utilizing large-scale samples from a single population and standardized, cutting-edge processing pipelines [22].

To expand the reference viral genome database and provide a more comprehensive understanding of the gut virome, we constructed a catalogue of gut viruses by processing over 11,000 fecal bulk or VLP-based viral metagenomes from Chinese populations. The catalogue, termed the Chinese Gut Virus Catalogue (cnGVC), comprises 93,462 nonredundant viral genomes (dereplicated from 426,496 viruses with > 95% nucleotide similarity) with a majority of which never found in existing gut viral databases. The cnGVC enables high-resolution functional characterization of the gut virome, significantly improving the recruitment of gut viruses in metagenomes (capturing 22.7% of viral reads in bulk metagenomes and 56.7% in viral metagenomes) and allowing for the assessment of the sex- and age-related variations in the virome. In terms of disease, we profiled the gut viromes of 6311 fecal metagenome samples representing 28 disease or unhealthy states and observed that most of the investigated diseases were associated with significant reductions in viral richness and diversity, as well as marked shifts in the overall gut viral composition. Furthermore, we identified 4238 differential viruses through meta-analysis across all diseased and healthy individuals, demonstrating the potential of these universal viral signatures to predict human health status.

Methods

Human fecal metagenomic datasets and public databases

We conducted a comprehensive review of studies based on human fecal metagenomic samples by searching relevant keywords (e.g., "gut metagenome," "fecal metagenome," "stool metagenome," "viral-like particle," and "VLP metagenome") in the PubMed database. From the search results, we manually selected 50 studies involving Chinese cohorts with publicly available metagenomic data, published until March 2022. These studies provided 11,327 fecal metagenomic samples, containing over 92 Tbp of high-throughput sequencing data. Detailed information about these 50 studies is presented in Additional file 2: Table S1. These studies were designated as the discovery cohorts for this investigation. Additionally, we incorporated data from six additional Chinese cohorts, which served as validation cohorts. These included five publicly available cohorts published after March 2022, covering CRC [25], chronic kidney disease (CKD) [26], rheumatoid arthritis (RA) [27], bipolar depression (BD) [28], and major depressive disorder (MDD) [29], as well as one newly recruited autoimmune cohort with metagenomic sequencing (described in the following section). Detailed information regarding these six studies is provided in Additional file 2: Table S2. This study also incorporated five public databases of human gut viral and microbial genomes: (1) GVD [6]; (2) GPD [1]; (3) MGV [23]; (4) Cenote Human Virome Database (CHVD)

[30]; and (5) Unified Human Gastrointestinal Genome (UHGG) [31].

Recruitment and metagenomic sequencing of the autoimmune cohort

Recruitment was conducted at the Second Affiliated Hospital of Dalian Medical University and the Second Affiliated Hospital of Guizhou University of Traditional Chinese Medicine. Patients with autoimmune diseases were enrolled based on confirmed diagnoses by licensed physicians, adhering to the 2019 European League Against Rheumatism/American College of Rheumatology (EULAR/ACR) classification criteria [32] for ankylosing spondylitis (AS), systemic lupus erythematosus (SLE), and primary Sjögren's syndrome (pSS). Exclusion criteria included diabetes, severe hypertension, severe obesity or metabolic syndrome, IBD, cancers, abnormal liver or kidney function, and recent use of antibiotics or probiotic products within the preceding 4 weeks. Based on these criteria, the study included 130 AS patients, 73 SLE patients, 66 pSS patients, and 118 healthy controls for further analysis. Fecal samples were collected from participants, temporarily stored on dry ice, and transported to the laboratory within 24 h, where they were stored at − 80 °C until further processing. DNA extraction was performed using the TIANamp Stool DNA Kit (TIANGEN, China), and DNA quality was evaluated using the Qubit 2.0 Fluorometer. Metagenomic sequencing libraries were prepared using the NEB Next Ultra DNA Library Prep Kit (NEB, USA) following the manufacturer's protocol, with unique index codes assigned to each sample. Library quality was verified using an Agilent 2100 Bioanalyzer. Indexed samples were clustered on a cBot Cluster Generation System using an Illumina PE Cluster Kit (Illumina, USA) according to the manufacturer's instructions. After cluster generation, DNA libraries were sequenced on the Illumina NovaSeq platform, producing 150 bp paired-end reads. Quality control and removal of human-derived contaminants were performed using the same pipeline applied to publicly available metagenomic datasets.

Sequence preprocessing and metagenome assembly

Raw reads were quality filtered and trimmed using fastp (v0.20.1) [33] with the parameters "-l 60 -q 20 -u 30 -y -trim_poly_g" for samples with read lengths ≤ 100 bp, or "-l 90 -q 20 -u 30 -y -trim_poly_g" for samples with read lengths > 100 bp, ensuring that the high-quality region accounted for at least 60% of the total read length. Human contamination was removed by mapping the quality-filtered reads to the human reference genome (GRCh38) using Bowtie2 (v2.4.1) [34]. The remaining clean reads were then de novo assembled into contigs using MEGAHIT v1.2.9 [35], with k-mer values selected based on the read length for each sample. Detailed information on all samples is shown in Additional file 2: Table S3.

Identification and decontamination of viral sequences

We performed an integrated homology- and feature-based pipeline to identify viral sequences based on our previously developed methodologies [7, 36–39]. Briefly, we first removed assembled contigs in which prokaryotic genes comprised more than half of the total gene content, and the number of prokaryotic genes exceeded viral genes by a factor of ten, as assessed by CheckV (v0.7.0) [40]. The remaining contigs were then assessed for viral content based on the following criteria: (1) the contig contained at least one viral gene, and the number of viral genes was greater than the number of prokaryotic genes, as determined by CheckV (v0.7.0); (2) the contig had a DeepVirFinder (v1.0) [41] score > 0.90 and a p value < 0.01; (3) the contig was identified as viral sequence by VIBRANT (v1.2.1) [42] using default options. Contigs meeting any of these criteria were identified as potential viral sequences. To further refine the analysis, we excluded viral sequences with a CheckV completeness score of < 50%, given their limited value for subsequent analyses. According to the previous study [6], we further performed a decontamination process for the remaining viral sequences based on the ratio of bacterial universal single-copy orthologs (BUSCO ratio) [43]. Using the hmmsearch program [44], we searched for BUSCO genes within each viral sequence with default parameters, calculating the BUSCO ratio as the number of BUSCO genes divided by the total number of genes in the sequence. Viral sequences with a BUSCO ratio > 5% were removed. After this decontamination step, a total of 426,496 highly credible viral sequences were retained for follow-up analysis.

Viral clustering and gene prediction

Following the clustering methodology described in our previous studies [36, 38], we clustered viral sequences at a 95% average nucleotide similarity threshold (≥ 85% coverage), resulting in a nonredundant gut virus catalogue consisting of 93,462 viral operational taxonomic units (vOTUs). For each vOTUs, the longest viral sequence was selected as the representative virus. Approximately 22.0 million putative protein sequences across all vOTUs were predicted via Prodigal (v2.6.3) [45] with the parameter "-p meta," and these sequences were clustered into nearly 1.6 million nonredundant protein sequences using MMseqs2 (v12.113e3) in *easy-linclust* mode [46] with the parameters "–min-seq-id 0.9 –cov-mode 1 -c 0.8 –kmer-per-seq 80."

Taxonomic classification, host assignment, and functional annotation

Taxonomic classification of vOTUs was performed using Diamond (v2.0.13.151) [47] with parameters "–id 30 –query-cover 50 –min-score 50 –max-target-seqs 10." Protein sequences were aligned against an integrated viral protein database derived from Virus-Host DB [48] (downloaded in November 2024). A vOTU was assigned into a known viral family when at least one of every five proteins matched the same family. Additionally, we also provided taxonomic classification results of all vOTUs based on the geNomad database [49], as shown in Additional file 2: Table S4.

Host assignment of vOTUs was performed by homology to genome sequences or CRISPR spacers of the UHGG database [31]. For homologous alignments, viral sequences were aligned against prokaryotic genomes in the UHGG using BLASTn with the parameters "-evalue 1e-2 -num_alignments 999999" [50], and a host as assigned when more than 30% region of the viral sequence matched to the corresponding host genome with > 90% nucleotide identity and an *e* value < 1e − 10. For CRISPR-spacer matches, host CRISPR-spacers were detected using MinCED v0.4.2 with the parameter "-minNR 2" [51]. Host assignment was confirmed when the viral sequences aligned with the host CRISPR spacer via BLASTn at a bit-score > 45 and an *e* value < 1e − 5.

To explore the functional properties of the viral sequences, we annotated viral proteins using several functional databases: Kyoto Encyclopedia of Genes and Genomes (KEGG) [52], Carbohydrate-Active enZymes (CAZy) [53], Virulence Factor Database (VFDB) [54], and a combined antimicrobial resistance databases (which includes CARD v3.1.0 [55], MEGARes v2.0 [56], ResFinder [57], and ARG-ANNOT [58]). For the KEGG and CAZy databases, viral proteins were assigned functional orthologs based on the best-hit gene in the respective database using Diamond with parameters "-e 1e-5 –query-cover 50 –subject-cover 50 –min-score 50." For the VFDB, proteins were assigned a virulence factor based on the best-hit gene using Diamond with parameters "–query-cover 50 –id 60." For antimicrobial resistance, viral proteins were aligned against the combined database using Diamond with the parameter "-e 1e-2." Distinct thresholds for protein identity were applied depending on the resistance type (> 70% for multidrug resistance, > 90% for beta-lactamases, and > 80% for other resistance types).

Phylogenetic analysis

We performed genome-based phylogenetic analyses for vOTUs with CheckV completeness > 90% using ViPTreeGen (v1.1.2) [59] with default parameters. The phylogenetic tree was visualized using iTOL (v6) [60]. In addition, we calculated the phylogenetic diversity (PD) of each viral family using the *pd* function in the R *picante* [61] package based on the proteomic tree generated by ViPTreeGen [59].

Taxonomic profiles

To characterize the gut virome composition, clean reads from each metagenome were mapped to the vOTUs in the cnGVC database using Bowtie2 with the parameters "–end-to-end –fast –no-unal -u 5,000,000." For each sample, the read count for each vOTU was normalized by dividing by its genomic size. The normalized read count was then divided by the total of all normalized read counts in the sample to define the relative abundance of each vOTU. The relative abundances of vOTUs within the same viral family were summed to generate family-level profiles.

Statistical analysis

Statistical analyses and data visualization were carried out via R language (v4.0.3).

Alpha diversity

Two alpha diversity estimates (i.e., the Shannon index and the observed number of vOTUs) were measured based on the relative abundance profiles at the vOTU level. The Shannon index was estimated using the *diversity* function within the *vegan* package [62]. The observed number of vOTUs was calculated as the count of vOTUs with a relative abundance greater than 0 in each metagenome.

Multivariate statistics

Bray–Curtis distances between samples were calculated based on relative abundance profiles at the vOTU level using the *vegdist* function from the *vegan* package. Principal coordinate analysis (PCoA) was carried out based on between-sample Bray–Curtis distances using the *pcoa* function within the *ape* package. Permutational multivariate analysis of variance (PERMANOVA) was implemented using the *adonis* function, and the resulting R^2 values were adjusted using the *RsquareAdj* function.

Correlation analyses

Correlation coefficients between metadata and viruses-associated variables were estimated using the *cor.test* function with the parameter "method = pearson." Smooth curves were plotted using the *geom_smooth* function with default parameters.

Statistical tests

Wilcoxon rank-sum tests were performed to compare virome diversity and viral relative abundances between controls and patients across studies using the *wilcox.test*

function. Additionally, we performed random-effects meta-analysis to identify universal viral signatures of common diseases following recent methodologies [63, 64]. Specifically, we first applied an arcsine square root transformation to the relative abundances of vOTUs, then used Hedges' *g* effect sizes to evaluate the difference in transformed values between controls and patients using the *escalc* function with the parameter "measure = SMD." Study heterogeneity was quantified based on I^2 statistics and tested by Cochran's *Q*-test via the *rma* function within the *metafor* package. For comparison analyses of occurrence rates of viral auxiliary metabolic genes (AMGs) between disease- and control-enriched vOTUs, the occurrence rate of AMG was calculated as the number of vOTUs with AMG divided by the total number of vOTUs within each group, and statistical significance was tested using the *fisher.test* function.

Classification models

We built the random forest model to assess the predictive ability of gut viral signatures in distinguishing human health status using the *train* function in the *caret* package. An example of the parameters used is as follows: train(Class ~ ., data = Data, method = "rf", metric = "ROC", trControl = trainControl(classProbs = TRUE, summaryFunction = twoClassSummary, method = "cv", number = 10, repeats = 10, sampling = "down", allowParallel = TRUE)). For discovery cohorts with fewer than 30 samples in any group, we applied threefold cross-validation repeated 10 times to account for the small sample size. For other discovery cohorts, standard tenfold cross-validation repeated 10 times was used. To address potential information leakage in independent validation cohorts, we ensured that, prior to testing each validation cohort, samples from the corresponding disease-type cohort in the discovery dataset were excluded from the training set. The model was then retrained using the remaining samples, and predictions were generated for the validation cohort. Receiver operating characteristic (ROC) analysis was performed using the *pROC* package, and the area under the ROC curve (AUC) was calculated accordingly.

Microbial signature-based disease risk stratification

Utilizing universal gut viral signatures, we categorized the common diseases into two groups based on the gross relative abundances of control-enriched vOTUs (tentatively termed "protective viruses") and disease-enriched vOTUs (proposed as "harmful viruses"). High-risk diseases were defined by a significant depletion of protective viruses coupled with an expansion of harmful viruses. Conversely, low-risk diseases exhibited no statistically significant alterations in either viral community.

Results

Construction of the gut virus catalogue in Chinese populations

To expand resources for gut virome research, we downloaded and reanalyzed raw data from a collection of 10,159 fecal bulk metagenomes and 1127 fecal viral metagenomes deriving from 50 previously published studies (Additional file 1: Fig. S2a; Additional file 2: Tables S1 and S3). This dataset contained samples spanning 18 provincial-level administrative regions of China (Fig. 1a), representing the current largest fecal metagenomic dataset available for Chinese populations. After processing with a unified pipeline, the dataset yielded 92.0 Tbp of high-quality non-human metagenomic data and produced a total of 290 million long contigs (≥ 5 kbp; total length 2.3 Tbp) via de novo metagenomic assembly for each sample. We identified 426,496 highly credible viral sequences (estimated completeness ≥ 50%) from the contigs using an integrated homology- and feature-based pipeline (see Methods), henceforth referred to as the cnGVC. The viral sequences ranged in length from 5000 to 504,568 bp, with an average length of 37,961 bp and an N50 length of 44,025 bp (Additional file 1: Fig. S3). We assessed the completeness and contamination of these viral genomes using the CheckV algorithm [40], revealing that 15.7% were complete, 31.8% had high completeness, and 52.5% had medium completeness (Fig. 1b). Notably, 99.3% of these viral genomes showed no contamination (Additional file 1: Fig. S3b), confirming the absence of microbial-specific genes at the genome termini.

Next, we clustered the viral sequences at a 95% average nucleotide similarity threshold (≥ 85% coverage), resulting in a nonredundant virus catalogue consisting of 93,462 vOTUs (Additional file 2: Table S4). Of these, 17.5% contained a complete representative virus as estimated by CheckV [40], while 39.6% and 42.9% had high- and medium-completeness representative viruses, respectively (Fig. 1b). Interestingly, only 37.6% (35,098/93,462) of vOTUs contained two or more viral members, while the remaining 62.4% were singletons. We also estimated the accumulation of vOTUs as a function of viral genome number to evaluate viral space coverage. The accumulation curve for vOTUs had not yet plateaued, while the curve for no-singleton vOTUs appeared to approach an asymptote (Fig. 1c). These findings suggest that, although a significant portion of common gut viruses has been uncovered in this sample, many rare virome members remain to be discovered.

Of the nonredundant vOTUs, 25.2% (23,594/93,462) could be robustly assigned to a known viral family. Several families, including *Microviridae*, *Aliceevansviridae*, *Herelleviridae*, and *Guelinviridae*, accounted for the majority of the taxonomically assigned vOTUs (Fig. 1d).

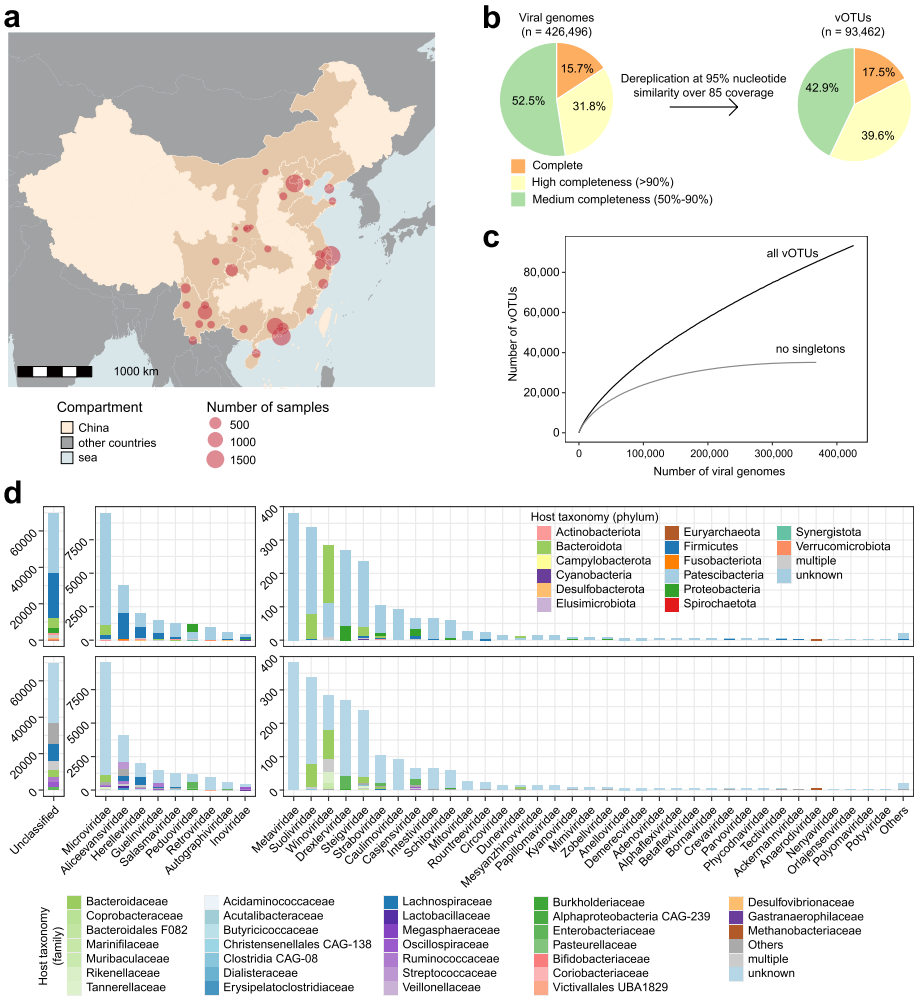

Fig. 1 Overview of the cnGVC. **a** Map of China showing the geographic distribution of metagenomic samples used to construct the gut virus catalogue. **b** Pie plots showing the estimated completeness of all viruses (left panel) and nonredundant vOTUs (right panel) in cnGVC. **c** Rarefaction curves of the number of vOTUs and no-singletons as a function of the number of all viral genomes. **d** Distribution of taxonomic annotation and host assignment of the cnGVC. The vOTUs are grouped at the family level, and the prokaryotic host taxa are shown at the phylum (upper panel) and family (bottom panel) levels. The number of vOTUs that had more than one predicted host is labeled by red color

Other viral families included *Salasmaviridae*, *Peduoviridae*, *Autographiviridae*, *Inoviridae*, and some eukaryotic viruses (e.g., *Retroviridae* and *Metaviridae*). Based on sequence similarity or CRISPR spacer matches in the comprehensive UHGG database [31], 46.2% of the 93,462 vOTUs could be assigned into one or more prokaryotic hosts. The most common identifiable hosts of *Aliceevansviridae* and *Herelleviridae* members were Firmicutes species, though their family-level hosts varied considerably (Fig. 1d). The most frequent hosts of *Microviridae* were Bacteroidota species, with a significant portion of viruses in this family predicted to infect *Bacteroidaceae*. Additionally, 2.8% (1228/43,188) of the annotated vOTUs had hosts from two or more bacterial phyla, and 11.2% (4816/43,188) vOTUs had hosts across different families, suggesting a relatively narrow host range of most gut viruses.

Comparison with existing gut viral databases

We compared the nonredundant viruses in the cnGVC with three available human gut viral catalogues (i.e., GVD [6], GPD [1], and MGV [23]) and the gut viruses from CHVD (CHVD-gut) [30]. All four existing catalogues were filtered to retain high- and medium-quality viruses (estimated completeness ≥ 50%) and dereplicated using the same thresholds as for the cnGVC. Following this processing, MGV and GPD contained the largest number of vOTUs, with 50,341 and 47,689 vOTUs, respectively, while CHVD-gut and GVD contained only 18,646 and 8882 vOTUs, respectively (Additional file 1: Fig. S4).

Pooling all databases revealed that 73.2% (68,435/93,462) of the vOTUs in the cnGVC were not found in other catalogues (Fig. 2a). These results indicate that, despite the presence of several existing gut viral catalogues, the cnGVC harbors a large number of novel viruses, likely due to the substantial representation of Chinese samples in the database. Furthermore, we found that the cnGVC covered 44.4% of the vOTUs from the Asia samples in MGV, but only 18.9% of the vOTUs from the non-Asia samples (Fig. 2b), suggesting that the cnGVC is more representative of Asian than Western populations.

To further illustrate the novelty of the cnGVC in phylogenomic terms, we compared all high completeness viruses (≥90%) between the cnGVC and the existing gut viral catalogues at the family level. For nearly all families, the cnGVC greatly expanded the content of known high-completeness viruses from the human gut (Fig. 2c). It increased the number of gut-derived vOTUs by 58.2–224.5% (average 111.2%) for the top five dominant families (*Microviridae*, *Aliceevansviridae*, *Herelleviridae*, *Guelinviridae*, and *Peduoviridae*). Notably, 385 high-completeness viruses belonging to *Retroviridae* were uniquely found in the cnGVC. Additionally, the cnGVC included 198 viruses belonging to *Metaviridae*, whereas only 1 high-quality *Metaviridae* virus existed in the other catalogues. In contrast, vOTUs of the small circular ssDNA viral family *Anelloviridae* were predominantly found in the existing catalogues but were rare in the cnGVC. Phylogenetic analyses based on genome sequences for five dominant families revealed that the newly found vOTUs in the cnGVC are broadly distributed across major taxonomic lineages within the phylogenetic trees (Additional file 1: Fig. S5), suggesting that they may help fill gaps in the viral tree of life in the human gut. Moreover, we calculated the phylogenetic diversity (PD) based on phylogenetic trees for each viral family and found that the cnGVC-specific viruses accounted for, on average, 43.5% of the PD in the trees for all families (Fig. 2d). Taken together, these findings underscore the comprehensiveness and novelty of the cnGVC and highlight its contribution to the global understanding of the human gut virome, particularly within large-scale Chinese populations.

Functional configuration of the gut viruses

Our extended viral catalogue may enable a high-resolution functional analysis of the gut virome. For this purpose, we predicted approximately 22.0 million protein-coding genes from the 426,496 viral genomes of cnGVC and clustered them into 1,595,487 nonredundant genes with an average amino acid identity (AAI) of 90% (Fig. 3a). The nonredundant gene catalogue contained 96.6% of complete genes, making it, to our knowledge, the largest and most comprehensive viral gene database for the human gut. Rarefaction analysis showed that the accumulation curve of no-singleton genes of the viral gene catalogue had approached an asymptote, indicating that further sampling would yield only minimal additions (Fig. 3b).

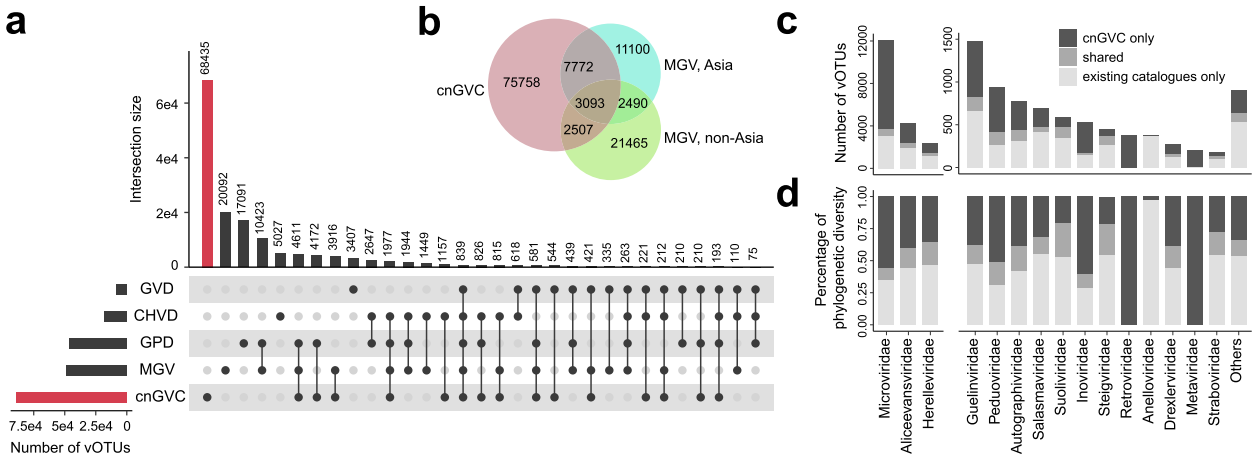

Fig. 2 Comparison between cnGVC and other existing databases. **a** UpSet plot showing the number of vOTUs shared by existing gut viral catalogues. The vOTUs that are uniquely found in cnGVC are labeled by red color. CHVD-gut, gut viruses from the Cenote Human Virome Database; GVD, Gut Virome Database; GPD, Gut Phage Database; MGV, Metagenomic Gut Virus. **b** Venn plot showing the sharing relationship of vOTUs in the cnGVC and MGV catalogues. Viruses in MGV are divided by their origin in Asia or non-Asia samples. **c** Comparison of vOTUs between cnGVC and other existing databases at the family level. **d** Contribution of phylogenetic diversity (PD) by the viruses from cnGVC and other existing databases. Phylogenetic trees are constructed for each viral family, and the PDs are calculated for cnGVC-specific vOTUs, existing-catalogue-specific vOTUs, and shared vOTUs accordingly

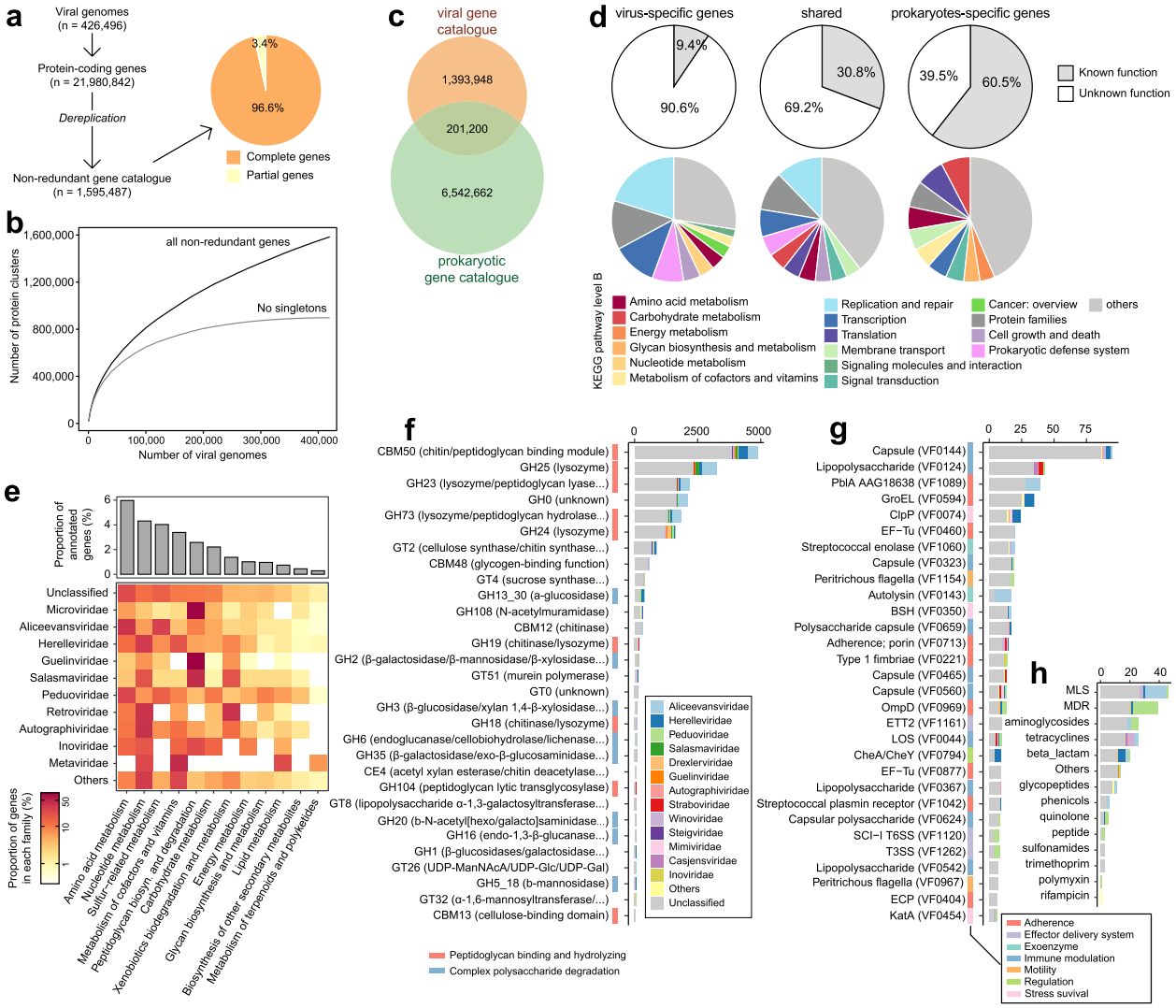

Fig. 3 Overview of the viral functions of cnGVC. **a** Construction of a nonredundant gene catalogue from the cnGVC viral genomes. **b** Rarefaction curves of the number of nonredundant viral genes and no-singletons as a function of the number of all viral genomes. **c** Venn plot showing the sharing relationship between gut viral genes and prokaryotic genes. **d** Functional composition of the virus-specific genes, prokaryotes-specific genes, and shared genes. Functions are categorized at the KEGG pathway level B. **e** Composition of viral auxiliary metabolic genes (AMGs) for each viral family. AMGs are grouped at the KEGG pathway level B and sulfur-related metabolism. The bar plot (upper panel) shows the overall proportions of AMGs versus the number of annotated genes for all viruses, and the heatmap (bottom panel) shows the proportions of AMGs versus the number of annotated genes for each family. Distribution of virus-encoded carbohydrate-active enzymes (CAZymes) (**f**), virulence factor genes (VFGs) (**g**), and antibiotic resistance genes (**h**) for each family. For CAZymes and VFGs, only the top 30 enzymes are shown

To further demonstrate the functional specificity of the gut virome, we first compared the viral genes with the gut prokaryotic gene catalogue UHGP-90 (the unified human gastrointestinal protein database clustered at 90% AAI) [31]. Although UHGP-90 was derived from an extensive collection of global gut prokaryotic genomes [31], it covered only 12.6% of the viral genes in cnGVC (Fig. 3c). This finding highlights a substantial disparity in gene contents between gut viruses and prokaryotes, underscoring that the gene/functional specificity of the gut virome has likely been underestimated in the past. However, only 9.4% of the virus-specific genes were functionally known under the KEGG database, which is significantly lower than the proportions for prokaryote-specific genes (60.5%) and shared genes (30.8%) (Fig. 3d). Consistent with the previous studies [23, 40], virus-specific genes were predominantly involved in replication and repair, transcription, prokaryotic defense system, and the metabolism of amino acids and nucleotides. We then focused on viral auxiliary metabolic genes (AMGs),

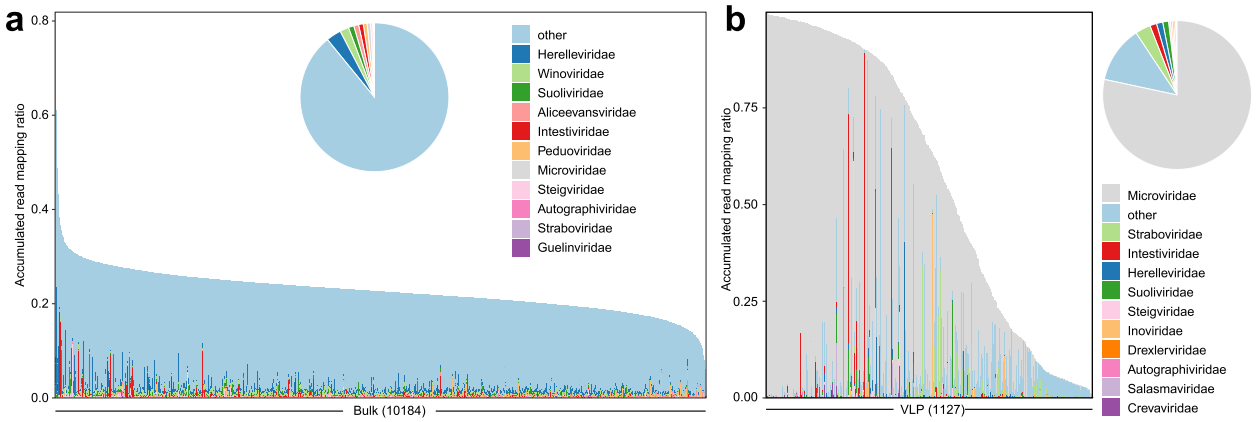

Fig. 4 Proportion of metagenomic reads mapped into the cnGVC. Bar plot showing the accumulated read mapping ratio of bulk metagenomes (**a**) and viral metagenomes (**b**) used in this study. Inset pie plots show the overall read proportions at the viral family level

as these genes play a key role in reprogramming host metabolic functions, directly influencing the gut ecosystem [65, 66]. Nearly one-fifth (19.8%) of the KEGG-annotated genes, corresponding to 2.4% of all nonredundant genes, were identified as AMGs based on a previously curated list [42]. Metabolism of amino acids, nucleotides, and sulfur compounds were the most popular auxiliary metabolic functions, and these were encoded by almost all viral families, with some exceptions: *Retroviridae* and *Inoviridae* lacked sulfur metabolism genes, and *Metaviridae* lacked genes for amino acid and sulfur metabolism (Fig. 3e). Viruses from *Microviridae*, *Guelinviridae*, and *Salasmaviridae* had a comparatively high proportion of genes involved in peptidoglycan biosynthesis and degradation.

Additionally, we identified 20,070 auxiliary carbohydrate-active enzymes (CAZymes), 694 virulence factor genes (VFGs), and 161 antibiotic resistance genes (ARGs) from the viral nonredundant gene catalogue (Fig. 3f–h). The majority of (>70%) the CAZymes were involved in the binding and hydrolyzing of bacterial peptidoglycans, likely associated with the degradation of bacterial cell wall during viral infection [67]. Many virus-encoded CAZymes belonged to glycoside hydrolase families that are involved in the decomposition of complex polysaccharides (Fig. 3f), and a large number of these genes were also validated through three-dimensional protein structural modeling, showing their role in the hydrolysis of bacterial polymers such as pectin, cellulose, and xylan. These results align with previous studies in environmental viral communities, highlighting the importance of polysaccharide decomposition in viral ecology [68, 69]. Notably, *Aliceevansviridae* and *Herelleviridae* viruses frequently encoded polysaccharide-degrading enzymes suggesting ecological specificity for these viruses. Regarding VFGs, the most common genes encoded enzymes involved in capsule synthesis (VF0144), lipopolysaccharide (LPS) synthesis (VF0124), and *PblA* (VF1089, a streptococcal phage-encoded protein that mediates binding to human platelets in the pathogenesis of infective endocarditis [70]) (Fig. 3g). In particular, VF0124, a LPS synthesis factor, was primarily encoded by *Straboviridae* and *Casjensviridae*, suggesting that these families may be high pathogenic.

Gut virome profiling and sex- and age-related variations

To demonstrate the utility of cnGVC in quantitative analyses of the gut viral community, we profiled the composition of 93,462 vOTUs across 11,286 bulk and viral metagenomic samples based on reads mapping. On average, 22.7% (interquartile range [IQR] = 20.2–25.2%) of bulk-metagenome reads were recruited by cnGVC, significantly higher than any other gut viral catalogues (Fig. 4a–b; Additional file 1: Fig. S6a). In viral metagenomic samples, cnGVC recruited an average of 56.7% (IQR = 17.9–92.6%) of the reads, which represented a substantial increase compared to existing catalogues (Fig. 4b; Additional file 1: Fig. S6b). For both bulk and viral samples, a large proportion of viral relative abundances (an average of 82.0% for bulk and 18.7% for viral samples) were composed of family-level unclassified viruses. Viruses belonging to *Herelleviridae*, *Winoviridae*, *Suoliviridae*, *Aliceevansviridae*, and *Intestiviridae* were the most dominated members in bulk metagenomic samples, with average relative abundances exceeding 1% (Fig. 4a). In viral metagenomic samples, viruses belonged to *Microviridae* were the most dominant members, followed by *Straboviridae*, *Intestiviridae*, *Herelleviridae*, and *Suoliviridae* (Fig. 4b). The substantial higher level of *Microviridae* viruses in

viral metagenomes is likely due to the preference the VLP technology for targeting free viral particles, which aligns with previous studies [6, 71].

Leveraging the large datasets, we sought to explore whether host sex and age influenced the gut virome. This analysis was based on viromes from bulk metagenomes of 4261 healthy subjects or subjects with low-risk diseases (see Methods for the definition), to void the confounding effects of high-risk diseases (Additional file 1: Fig. S2b). PCoA analysis based on Bray–Curtis distances of the vOTU-level community composition revealed that both host sex and age contributed modestly but significantly to the gut virome (Fig. 5a–b). Likewise, PERMANOVA analysis showed that sex and age explained 0.24% (*adonis* $p < 0.001$) and 0.40% (*adonis* $p < 0.001$) of the overall virome variability, respectively (Fig. 5c). These effect sizes remained significant after adjusting for study heterogeneity, host location, and body mass index (BMI).

Using the vOTU profiles, we evaluated the gut viral richness (estimated by the observed number of vOTUs) and diversity (Shannon index) across different age groups. Consistent with previous studies [71], a strong correlation between the virome and bacteriome was observed in both richness and diversity in the bulk metagenomic samples ($r > 0.80$), whereas the correlation was weaker in the VLP metagenomic samples (Additional file 1: Fig. S7). Overall, females exhibited higher viral richness and diversity than males (Wilcoxon rank-sum test $p < < 0.001$ for both indexes), with this difference primarily observed in the 30–39 and 40–49 are groups (Fig. 5d; Additional file 1: Fig. S8). Across the lifespan, we observed that female viral richness and diversity increased until age 40 (richness, $r = 0.135$, $p = 0.0038$), but decreased after 40 (richness, $r = 0.135$, $p = 4.5 \times 10^{-9}$; diversity, $r = 0.151$, $p = 7.3 \times 10^{-9}$) (Additional file 1: Fig. S9a). In contrast, for males, gut viral richness and diversity were relatively stable until age 30, after which they increased (richness, $r = 0.091$, $p = 7.6 \times 10^{-5}$; diversity, $r = 0.163$, $p = 1.3 \times 10^{-12}$; Additional file 1: Fig. S9b).

Sex- and age-related trajectories of the gut virome at the family level revealed several notable trends. For instance, in both females and males, two of the most dominant families, *Winoviridae* and *Aliceevansviridae*, exhibited distinct age-related patterns. *Winoviridae* showed an overall decline with increasing age, while *Aliceevansviridae* first decreased (reaching its lowest point around 20 years) before showing an upward trend in middle-aged and elderly individuals (Fig. 5e). Besides, *Microviridae* was significantly higher in males aged 30–39 compared to females (Wilcoxon rank-sum test $q < 0.05$), but no significant differences were found in other age groups (Additional file 1: Fig. S10). Conversely, *Aliceevansviridae* was higher in females aged 30–39 than in males, but significantly lower in elderly individuals (50–59, 60–69, and 70–79).

Diversity and compositional patterns of the gut virome across common diseases

After cataloguing and profiling the gut virome, we next wanted to explore the associations between the gut virome and common diseases based on 36 case–control studies (Additional file 1: Fig. S2b). For each surveyed study, samples were filtered according to exclusion criteria such as (1) non-standard disease definitions, (2) abnormal BMI (for samples with available phenotypic data), and (3) low metagenomic data amount or extreme virus proportion (see Methods). This process resulted in a total of 6311 fecal samples spanning 28 disease or unhealthy statuses across 40 case–control comparisons (Additional file 2: Table S1). Among these diseases, cardiometabolic (7 diseases from 10 studies) and immune (8 diseases from 9 studies) disorders were most common, followed by digestive (3 diseases from 4 studies), infectious (3 diseases from 3 studies), and psychiatric (2 diseases from 3 studies) disorders, and cancers (2 diseases from 4 studies).

Within-sample diversity analysis revealed a significant decrease in viral richness in the patient groups of 14 out of 40 case–control comparisons. Similarly, viral diversity significantly decreased in patients from 15 of the 40 case–control comparisons (Fig. 6a–b). Diseases such as Crohn's disease (CD), pulmonary tuberculosis (PT), COVID-19 infection, as well as several immune (i.e., AS, Graves' disease [GD], gout, and SLE) and cardiometabolic (i.e., hypertension, metabolic unhealthy obesity) diseases exhibited decreased in viral richness and diversity. Conversely, only 3 of 40 case–control comparisons, covering atrial fibrillation (AF) and Parkinson's disease, showed a significant increase in viral richness or diversity.

PERMANOVA analysis revealed that 28 of the 40 case–control comparisons, spanning 21 disease or unhealthy statuses, significantly altered the overall structure of the gut virome (*adonis* $p < 0.05$) (Fig. 6c). Patients with CD, AF, and polycystic ovarian syndrome (PCOS) showed the greatest variations in their gut virome, with effect sizes of 7.2%, 6.8%, and 5.1%, respectively. We further carried out random forest classification to distinguish cases from controls within each study based on their gut viral profiles. The classifiers achieved high discriminatory ability (AUC > 0.80) in 17 out of 40 case–control comparisons and moderate discriminatory ability (AUC = 0.70–0.80) in 11 comparisons (Fig. 6d). These findings underscore substantial shifts in the gut viral composition across diseases with varying clinical manifestations and pathogenesis. In contrast, combining the diversity, PERMANOVA, and random forest results, we identified 7

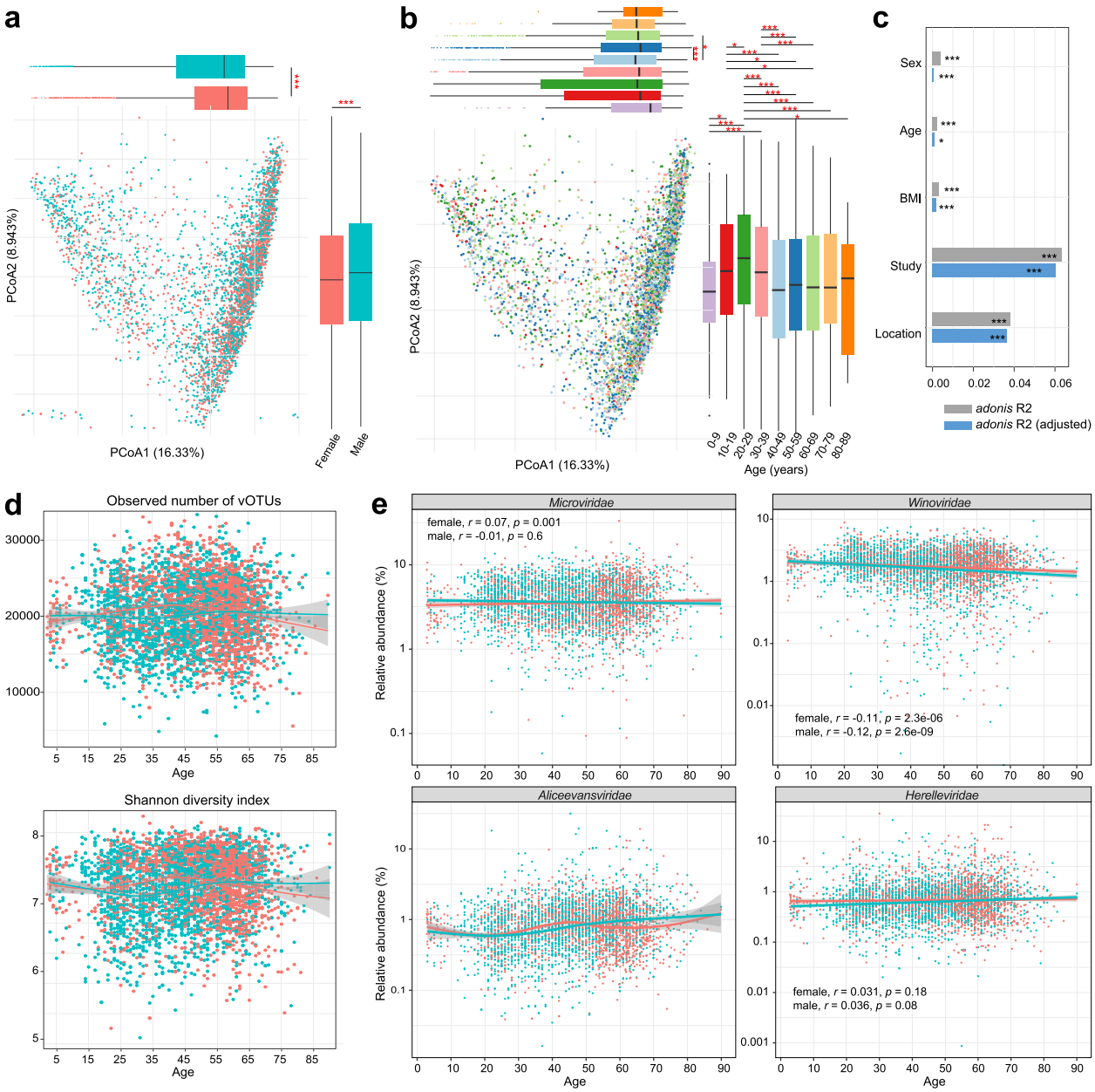

Fig. 5 Sex- and age-related variations of the gut virome. Principal coordinates analysis of the gut viromes of 4261 healthy subjects or subjects with low-risk diseases grouped by their sex (**a**) and age stages (**b**). Samples are shown at the first and second principal coordinates (PC1 and PC2), and the ratio of variance contributed by these two PCs is shown. The below and left boxplots show the sample scores in PC1 and PC2 (boxes show medians/quartiles; error bars extend to the most extreme values within 1.5 interquartile ranges). Wilcoxon rank-sum test: *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$. **c** Permutational multivariate analysis of variance showing the effect size of sex, age, and other confounding factors on the gut virome of all investigated samples. For each factor, the raw effect size (*adonis* R^2) and the effect size after adjusting for other factors (adjusted *adonis* R^2) are shown. *Adonis* analysis with 1000 permutations: *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$. **d** Scatter plots showing the sex-related trajectories of gut virome richness (upper panel) and diversity (bottom panel) at different ages. **e** Scatter plots showing the sex-related trajectories of the top 4 dominant viral families at different ages. For **d** and **e**, points indicate samples grouped by females (red) and males (green), and smooth curves are formed based on the diversity indexes and the ages of the samples using the *geom_smooth* function in the R platform

diseases, including carotid atherosclerosis (CA), bone mass loss (BL), irritable bowel syndrome (IBS), Behcet's disease, RA, Vogt-Koyanagi-Harada disease (VKH), and schizophrenia, that showed minimal change in both viral diversity and composition. These diseases are likely considered low-risk diseases within the gut virome scope.

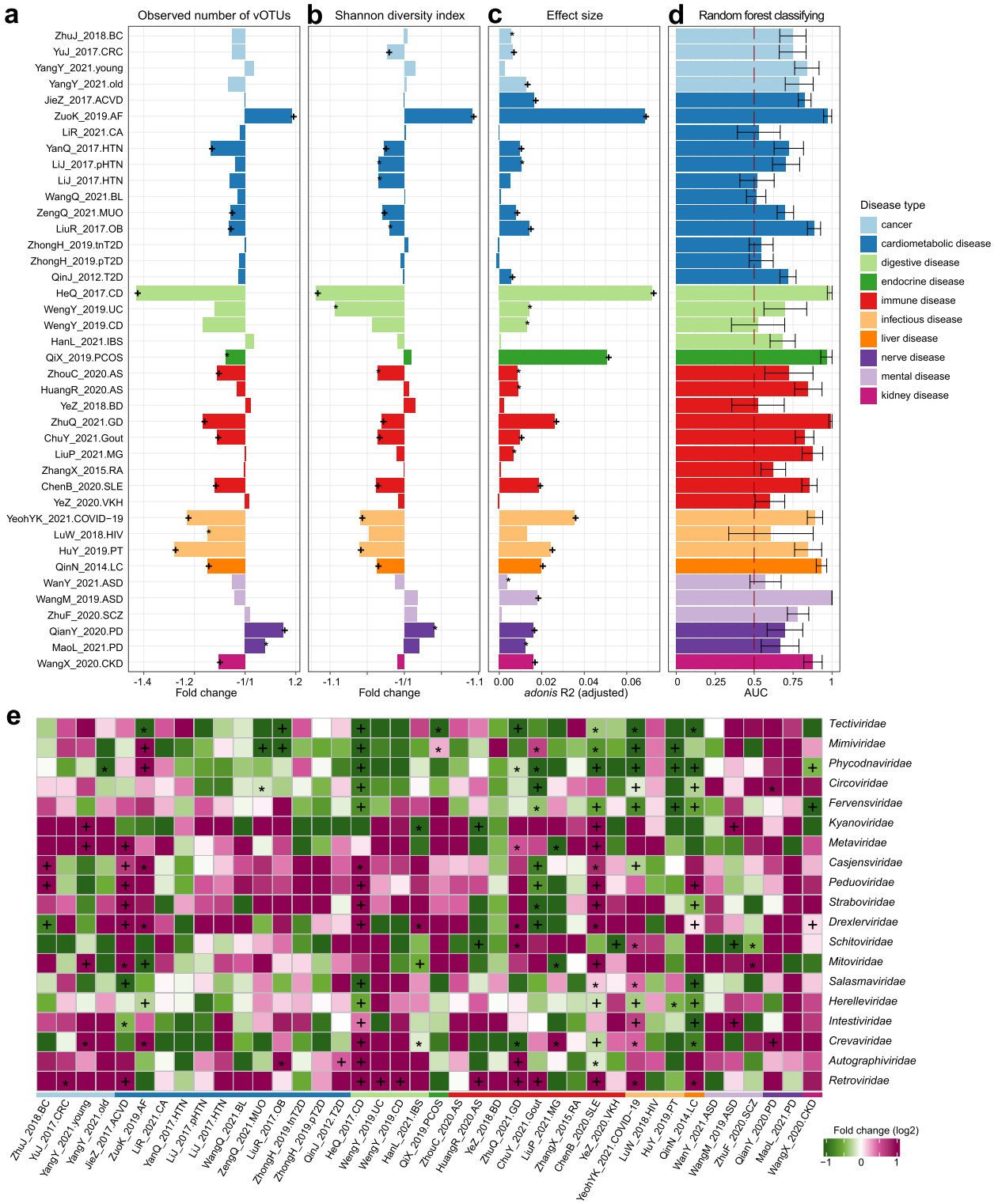

Fig. 6 Alterations of the gut virome across common diseases. Bar plot showing the fold changes of gut virome richness (**a**) and diversity (**b**), the disease-related effect size (**c**), and the within-study AUCs (**d**) of 40 case–control comparisons. Diseases are colored by the disease types. For **a** and **b**, Wilcoxon rank-sum test: *, $p < 0.05$; +, $p < 0.01$. For **c**, adonis analysis with 1000 permutations: *, $p < 0.05$; +, $p < 0.01$. For **d**, the dashed line shows an AUC of 0.50, and the error bars show the 95% confidence interval of the AUC values. **e** Heatmap showing the fold changes of each viral family within 40 case–control comparisons. Fold change > 0, enriched in cases; fold change < 0, enriched in controls. Wilcoxon rank-sum test: *, $q < 0.05$; +, $q < 0.01$. The disease types of each case–control comparison are shown by bottom colors (legend following **a**–**d**)

To further uncover gut viral signatures, we compared the gut viromes of patients and controls for each disease at the family level. Using the Wilcoxon rank-sum test, we found that the *Aliceevansviridae* was significantly enriched in patients with atherosclerotic cardiovascular disease (ACVD), AF, and liver cirrhosis (LC) ($q < 0.05$) compared to healthy controls, but reduced in patients with AS (only in the study ZhouC_2020, $q < 0.05$; Fig. 6e). *Winoviridae* was significantly decreased in patients with ACVD, AF, and obesity, but increased in COVID-19 patients. Similarly, *Microviridae* was notably reduced in ACVD and AF patients but increased in COVID-19 and gout patients. Interestingly, several viral families exhibited similar patterns across multiple diseases. For example, *Retroviridae* was significantly enriched in the patients from 11 case–control comparisons spanning 9 diseases, while *Tectiviridae* and *Fervensviridae* were depleted in patients from 8 and 7 different diseases, respectively. *Phycodnaviridae* was depleted in 9 diseases and enriched in 1. A random-effects meta-analysis supported that these 4 families showed significant differences in relative abundance in patients vs. controls across all diseases ($q < 0.05$; Additional file 1: Fig. S11), suggesting the existence of shared viral signatures for health.

Universal viral signatures of common diseases

Given that most of the investigated diseases exhibited significant alterations in the overall gut viral communities and in certain viral families, we aimed to identify the universal viral signatures of these diseases at the vOTU level. A total of 4238 vOTUs that differed in relative abundances across 36 case–control studies were identified through a combination of meta-analysis and direct comparison between all case and control subjects (random effects meta-analysis $q < 0.01$, $I^2 < 50\%$, and cases vs. controls $q < 0.01$; Fig. 7a; Additional file 2: Table S5). Among these, 1328 differential vOTUs were more abundant in subjects with diverse diseases, while 2910 were enriched in healthy individuals. Both disease-enriched and control-enriched vOTUs were mainly composed of family-level unclassified members, with small proportions of *Herelleviridae* and *Aliceevansviridae*. Consistent with our findings at the family level, members of *Peduoviridae* (disease-enriched vs. control-enriched vOTUs, 102 vs. 0) and *Retroviridae* (14 vs. 0) were frequently enriched in disease subjects, whereas *Herelleviridae* (45 vs. 266) predominantly appeared in control-enriched vOTUs (Fig. 7b; Additional file 1: Fig. S12a). When we assigned the differential vOTUs to their prokaryotic hosts, we found that the control-enriched vOTUs had a large proportion of members that were predicted to infect *Ruminococcaceae* (18.2% of 2910 control-enriched vOTUs) and *Butyricicoccaceae* (1.3%), whereas only 0.15% and 0% of disease-enriched vOTUs were members of these two families, respectively (Fig. 7b; Additional file 1: Fig. S12b). Notably, a large proportion of *Ruminococcaceae*-hosted vOTUs were predicted to infect bacteria from the *Faecalibacterium* genus ($n = 350$ vOTUs; Additional file 1: Fig. S12c), a well-known short-chain fatty acid (SCFA)-producing taxon with beneficial effects in multiple common disorders [72, 73], suggesting a potential link between *Faecalibacterium* phages and health. Conversely, disease-enriched vOTUs contained a high proportion of viruses predicted to infect *Enterobacteriaceae* (15.4% of 1328 disease-enriched vOTUs), *Tannerellaceae* (2.1%), *Erysipelatoclostridiaceae* (2.0%), *Fusobacteriaceae* (1.0%), and *Erysipelotrichaceae* (0.8%), while phages targeting these bacteria were rarely appeared in control-enriched vOTUs. At the genus level, *Escherichia* phages were most frequent in disease-enriched vOTUs (Additional file 1: Fig. S12c).

To further elucidate the functional and metabolic capabilities of the gut viral signatures associated with common diseases, we compared the profiles of AMGs between disease-enriched and control-enriched vOTUs at the enzyme level (Additional file 2: Table S6). Preliminary comparison revealed that control-enriched vOTUs encoded a higher frequency of AMGs compared to disease-enriched vOTUs (Additional file 1: Fig. S13a), suggesting that these viruses may play a role in the metabolism of more substances in the human gut. Specifically, 42 of the 50 most frequent AMGs differed in frequency between disease-enriched and control-enriched vOTUs (Fig. 7c). The control-enriched vOTUs exhibited a higher frequency of enzymes involved in biosynthesis of nicotinamide adenine dinucleotide (NAD+) ($n = 6$ enzymes, Additional file 1: Fig. S13b), cytosine/methionine metabolism (DNA cytosine-5-methyltransferase K00558/K17398 and S-adenosylmethionine synthetase K00789), folate metabolism (thymidylate synthase K03465/K00560 and methylenetetrahydrofolate dehydrogenase K01491), and assimilatory sulfate reduction (phosphoadenosine phosphosulfate reductase K00390 and sulfate adenylyltransferase K00957) compared to disease-enriched vOTUs. On the other hand, enzymes related to LPS biosynthesis, such as polyisoprenyl-phosphate glycosyltransferase K20534 and D-sedoheptulose 7-phosphate isomerase K03271, were more frequent in disease-enriched vOTUs. We also observed a higher level of viral-encoded NAD+ synthesis capacity in the gut virome of patients with chronic kidney disease (CKD) [39], probably linked to phage DNA replication and exploitation of host metabolic pathways and biochemical processes during infection [74], which needs to be validated by subsequent studies.

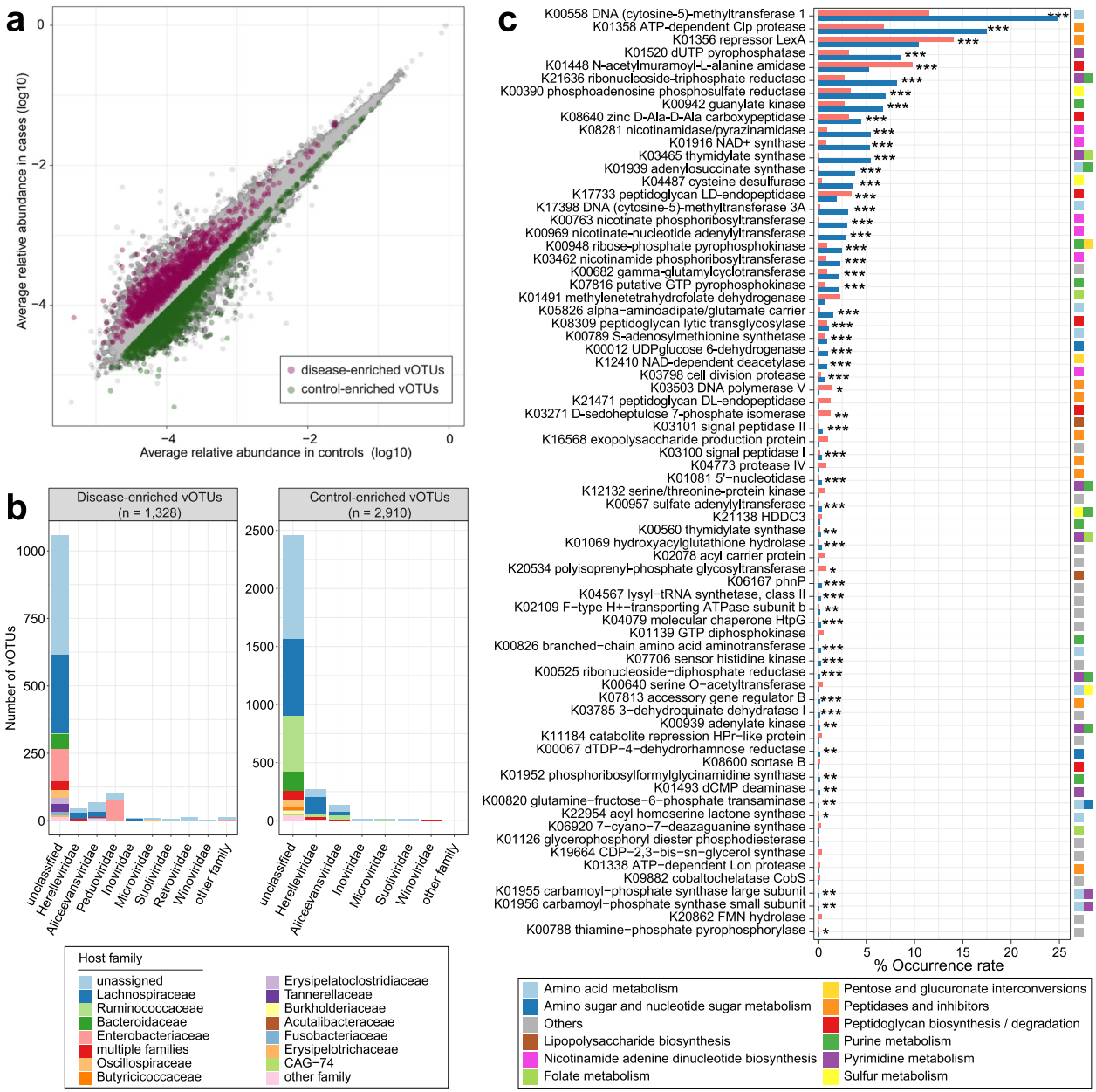

Fig. 7 The disease-associated viral signatures. **a** Scatter plot of median relative abundances of vOTUs in all investigated disease and control individuals. Gray points represent vOTUs not differentially abundant between two groups, and red and green points represent differentially abundant vOTUs. **b** Distribution of taxonomic annotation and host assignment of the disease-enriched and control-enriched vOTUs. The vOTUs are grouped at the family level, and the prokaryotic host taxa are also shown at the family level. The number of vOTUs that had more than one predicted host is labeled by red color. **c** Occurrence rate of most frequent AMGs in all disease-associated vOTUs. The functional categories of AMGs are shown by colored squares. Statistical test was performed using Fisher's exact test: *, $q < 0.05$; **, $q < 0.01$; ***, $q < 0.001$

Gut viral signatures as a predictor of health status

Finally, we tested the ability of gut viral signatures to predict human health status. A random forest classifier was trained using the abundances of 4238 universal viral signatures and tested with a tenfold cross-validation approach. This classifier achieved an AUC score of 0.682 (95% confidence interval [CI], 0.668–0.697) for classifying all case and control samples. It reached an AUC of 0.737 (95% CI, 0.718–0.755) in distinguishing patients with high-risk diseases from their corresponding controls, demonstrating the strong predictive power of the gut virome for these patients. Interestingly, a

new classifier trained by a subset of the most important vOTUs generated improved discriminatory power, with an AUC of 0.698 (95% CI, 0.684–0.712) for classifying all samples and 0.761 (0.742–0.778) for distinguishing high-risk diseases versus controls (Fig. 8a; Additional file 1: Fig. S14). This minimal set of gut viral signatures could serve as a more feasible indicator for health status.

To validate the reliability of the signatures in external data, we curated fecal metagenomes from several independent cohorts, including those involving CRC, CKD, RA, BD, and MDD, as well as a newly recruited cohort comprising 269 patients of autoimmune diseases (i.e., AS, SLE, and pSS) and 118 healthy controls from China (Additional file 2: Table S2). Using these large cohorts, we quantified the relative abundances of 4238 disease-associated vOTUs and compared them between cases and controls in the new cohorts. The results showed that most vOTUs exhibited a consistent trend in mean abundance between patients and controls within each disease compared with the observation in the original datasets. For example, in CKD patients versus controls, 70.7% (2998/4238) of vOTUs were more abundant in either patients or controls, consistent with the original datasets, and 34.7% (1444/4238) of vOTUs were significantly enriched (Additional file 1: Fig. S15). Moreover, the discriminatory power of the original random forest classifier on these new cohorts yielded AUC scores of 0.892, 0.700, 0730, 0.626, and 0.566 for CRC, CKD, RA, BD, and MDD patients versus controls, respectively (Fig. 8b). Similarly, for the newly sequenced autoimmune patient cohort versus controls, the AUC scores were 0.566, 0.586, and 0.621 for AS, SLE, and pSS, respectively. These findings suggest that the generalized disease-associated gut viral signatures identified in this study can accurately classify multiple diseases from health.

Discussion

Viruses represent in number the largest component of the human gut microbial community, yet much about their genome, function, and role in certain diseases remains unknown [75]. In this study, we report the creation of the cnGVC, which contains over 93,000 nonredundant viral sequences exhibiting high completeness (approximately 60% of viruses have ≥ 90% completion) and representativeness. To our knowledge, this is the current largest viral genome catalogue for fecal metagenomes from a single population. Notably, it is 95% and 85% larger than the GPD and MGV, respectively, even though these databases were compiled from large-scale populations across multiple countries. Remarkably, when comparing various gut virus catalogues, over 70% of the viruses in the cnGVC were newly discovered. This finding is not only due to technical differences (e.g., inclusion criteria and use of VLP datasets) but also highlights the specificity of the gut virome within the Chinese population. Overall, the construction of the cnGVC underscores the significant research value of viral identification within a single population.

We also constructed a gene catalogue from the cnGVC, which includes nearly 1.6 million nonredundant viral genes. The majority of these genes were not found in gut prokaryotic genomes, corroborating previous observations in other environments [76–78], and further highlighting the functional specificity of the gut virome in the context of future holistic microbiome research. In addition, we identified numerous AMGs, CAZymes, as well as some VFGs and ARGs in the viral gene catalogue.

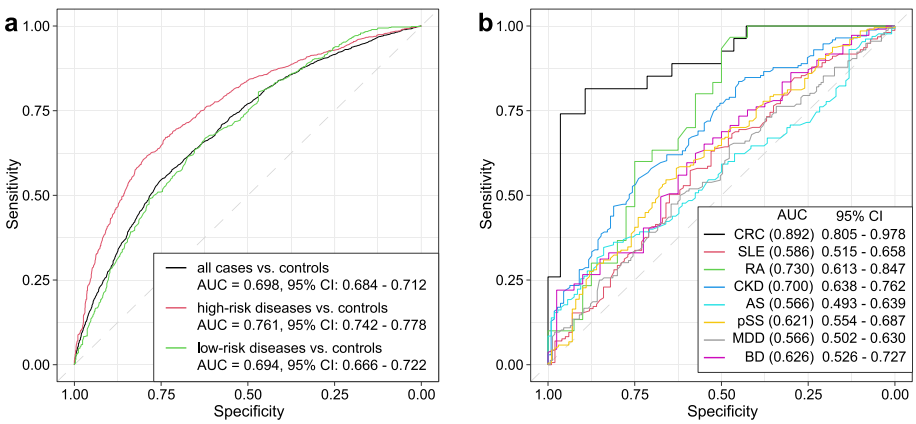

Fig. 8 Prediction of health status using the viral signatures. **a** Receiver operating characteristic (ROC) analysis of the classification of case/control status using the random forest model trained by 4238 universal viral signatures. **b** ROC analysis of the classification of case/control status in independent cohorts. The classification performance of the model was assessed by the area under the ROC curve (AUC). The AUC values and 95% confidence intervals (CIs) are shown

These resources may serve as valuable tools for future research aiming to decode important viral functions such as complex polysaccharide degradation and antibiotic resistance mechanisms.

Using massive fecal metagenome datasets, we investigated the impact of sex and age on the gut virome and uncovered significant sex- and age-related alterations. A previous study suggested a decline in gut viral diversity in older individuals compared to adults [6], but this pattern was not evident in our data. We observed that both gut viral diversity and the compositions of certain dominant viral families (e.g., *Microviridae*) in females have greatly changed at the age of 40–49. These findings are consistent with observations in the gut bacteriome and may be linked to the physiological changes that occur in women around menopause (e.g., changes in sex hormones or metabolism) [79, 80]. Based on our results, we hypothesize that the gut virome may play a role in human physiology, immunity, or metabolism, and further investigation into this idea is warranted.

The gut virome has been implicated in various human diseases; however, integrating and comparing viral signatures across different diseases remains challenging. From our large-scale datasets, we found that the viral richness and diversity were often reduced in several diseases, though they were increased in a few cases. Decreased gut viral diversity has also been reported in other single-disease studies, such as those involving T2D [81] and liver disease [12]. Similarly, a reduction in diversity is a typical feature of the gut bacteriome during disease states, possibly reflecting low resilience and dysbiosis within the microbiological ecosystem [82, 83]. Moreover, our analysis revealed that most of the investigated diseases are associated with substantial alterations in their viral communities, and the degree of virome alteration varies. These findings (1) align with previous studies showing substantial changes in the virome in conditions such as IBD, CRC, and other diseases, (2) reinforce the connection between the gut virome and immune/cardiometabolic diseases, and (3) propose new diseases, such as Parkinson's disease, autism spectrum disorder, and PCOS, that may also exhibit virome alterations. Follow-up exploration on these findings could generate testable hypotheses for disease-specific studies aimed at understanding the etiologies and developing therapeutic strategies.

Meta-analysis across all diseases investigated revealed extensive gut viral signatures at both the family and vOTU levels. The most significant disease-associated viral families were *Retroviridae*, which were significantly enriched in the gut of patients with 9 different diseases. Almost all gut retroviruses were newly discovered by the cnGVC, and their functions remain under exploration. Other universal viral signatures included numerous disease-enriched vOTUs predicted to infect *Enterobacteriaceae*, *Fusobacteriaceae*, *Erysipelotrichaceae*, and *Erysipelatoclostridiaceae*. Pathogenic *Enterobacteriaceae* bacteria are well-known opportunistic pathogens [84, 85], and their phages may interact with host bacteria to influence disease outcomes. The other bacterial taxa were found to have pathogenic roles in certain diseases (e.g., *Fusobacteriaceae* in CRC) [86, 87], suggesting that their phages may function somewhat independently of the bacteria, exerting roles in human diseases. Moreover, functional analyses revealed that some viral functions, such as NAD + synthesis, are widely associated with diseases, highlighting the relevance of viral functions in common diseases. We also found that gut viral signatures have high predictive power for disease status across all samples, with performance comparable to recent bacterial-level studies [64, 88]. On the test datasets, we observed that these signatures demonstrated high reproducibility across multiple external populations. Collectively, the broad and universal viral signatures identified in this study hold promise for future research into disease mechanisms, interventions, and phage therapy efforts.

Conclusions

In this study, we unveiled the vastness and specificity of the gut virome, with a particular focus on the Chinese population, by presenting the cnGVC—the largest viral genome catalogue to date. Our findings highlight the high functional specificity of the virome and its potential roles in human physiology, immunity, and metabolism. We also observed significant alterations in the gut virome across various diseases, emphasizing its potential involvement in disease etiologies. Notably, our meta-analysis identified broad and universal viral signatures, which could be pivotal for future disease mechanism studies, interventions, and phage therapy efforts. This research lays a solid foundation for further exploration into the gut virome's role in human health and disease.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-025-01460-6.

Additional file 1: Fig. S1. Substantial differences of the gut viruses from different sources. Fig. S2. Overview of the workflows in this study. Fig. S3. Statistics of length and contamination of the viral genomes in cnGVC. Fig. S4. Processing of the existing gut viral databases using a unified pipeline. Fig. S5. Genome-based phylogenetic trees of 5 dominant viral families. Fig. S6. Comparison of read mapping ratio between cnGVC and other gut viral databases. Fig. S7. Correlation analysis of alpha diversity indexes between the virome and bacteriome. Fig. S8. Sex-related variations in the richness and diversity of the gut virome. Fig. S9. Age-related variations in the richness and diversity of the gut virome. Fig. S10. Sex- and age-related variations of the gut virome at the family level. Fig. S11. Meta-analysis of

the abundances of viral families across common diseases. Fig. S12. Taxonomy and host assignment of the viral signatures. Fig. S13. Functions of the disease-associated viral signatures. Fig. S14. Random forest models for discriminating patients and controls using the universal viral signatures. Fig. S15. Alterations of universal viral signatures in patients with chronic kidney disease.

Additional file 2: Table S1. Detailed information of 50 studies used in this study. Table S2. Detailed information of the validation cohorts used in this study. Table S3. Detailed information of 11,286 metagenomic samples used in this study. Table S4. Detailed information of 93,462 vOTUs in the cnGVC. Table S5. Detailed information of 4238 disease-associated vOTUs identified by meta-analysis across common diseases. Table S6. Detailed information of the viral AMGs occurred in the disease-associated vOTUs.

Acknowledgements
Not applicable.

Authors' contributions
Q.Y., W.S., and Zhixin L. conceived the study; S.L. Yue Z., and L.H. performed data collection; S.L., Yue Z., R.G., F.C., Jinxing M., Q.L., and G.W. performed the gut virome analyses; Q.Y. and Pan Z. participated in development of analytical methods; S.S, Q.Y., L.C., S.F., R.L., W.Y., Yan Z., and Jie M. performed sample collection and experiments; S.L. drafted the manuscript. Zhiming L., J.L., C.C., and H.U. helped drafting the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by the National Natural Science Foundation of China (No. 82370563), Beijing University of Chinese Medicine (No. 5050071720001 and 2180072120049), 2024 High-quality Development Research Project of Shenzhen Bao'an Public Hospital (No. BAGZL2024138 and BAGZL2024130), Dalian Outstanding Young Scientific and Technological Talents Program (2024RJ018), and Young Top Talents Program of the Liaoning Revitalization Talents Initiative (XLYC2403207).

Data availability
The viral sequences and annotation files of cnGVC have been deposited in https://zenodo.org/records/14671177. The raw metagenomic sequencing data of the autoimmune cohort generated in this study have been deposited in the European Bioinformatics Institute (EBI) database under the project ID PRJEB87207 (https://www.ebi.ac.uk/ena/browser/view/PRJEB87207). The original codes used in the paper are provided in the GitHub website with URL: https://github.com/yexianingyue/GV_common_diseases/. The accession IDs and detailed information of all publicly available metagenomic datasets and samples are provided in Additional file 2: Table S1-S3.

Declarations

Ethics approval and consent to participate
Recruitment of autoimmune cohort in this study was conducted in strict accordance with the principles of the Declaration of Helsinki and the International Conference on Harmonization Good Clinical Practice (ICH-GCP) guidelines. This study received approval from the Ethics Committee of Dalian Medical University, and written informed consent was obtained from all participants.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]The Fifth Affiliated Hospital of Southern Medical University, Guangzhou 510900, China. [2]Department of Microbiology, Department of Biochemistry and Molecular Biology, College of Basic Medical Sciences, Dalian Medical University, Dalian 116044, China. [3]Department of Reproductive Health, Shenzhen Bao'an Chinese Medicine Hospital, Guangzhou University of Chinese Medicine, Shenzhen 518101, China. [4]Puensum Genetech Institute, Wuhan 430076, China. [5]Department of Gastroenterology, The Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710004, China. [6]School of Chemistry, Hubei Key Laboratory of Nanomedicine for Neurodegenerative Disease, Chemical Engineering and Life Science, Wuhan University of Technology, Wuhan 430070, China. [7]BGI-Shenzhen, Shenzhen 518083, China. [8]Department of Rheumatology and Immunology, Peking University People's Hospital, Beijing 100044, China. [9]Department Pathology, Dalian Municipal Central Hospital, Dalian 116033, China. [10]Department of Rheumatology and Immunology, The Second Affiliated Hospital of Guizhou University of Traditional Chinese Medicine, Guiyang 550025, China. [11]Department of Acupuncture and Moxibustion, Beijing Hospital of Traditional Chinese Medicine, Capital Medical University, Beijing 100010, China. [12]Department of Traditional Chinese Medicine, Beijing Friendship Hospital, Capital Medical University, Beijing 100050, China. [13]School of Traditional Chinese Medicine, Beijing University of Chinese Medicine, Beijing 100029, China. [14]Centre for Translational Medicine, Shenzhen Bao'an Chinese Medicine Hospital, Guangzhou University of Chinese Medicine, Shenzhen 518101, China. [15]Key Laboratory of Health Cultivation of the Ministry of Education, Beijing University of Chinese Medicine, Beijing 100029, China.

Received: 30 August 2024 Accepted: 18 March 2025
Published online: 26 March 2025

References

1. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. Massive expansion of human gut bacteriophage diversity. Cell. 2021;184(4):1098-1109 e1099.
2. Wang D, Urisman A, Liu Y-T, Springer M, Ksiazek TG, Erdman DD, Mardis ER, Hickenbotham M, Magrini V, Eldred J. Viral discovery and sequence recovery using DNA microarrays. PLoS Biol. 2003;1(2): e2.
3. Dang VT, Sullivan MB. Emerging methods to study bacteriophage infection at the single-cell level. Front Microbiol. 2014;5:724.
4. Shkoporov AN, Clooney AG, Sutton TDS, Ryan FJ, Daly KM, Nolan JA, McDonnell SA, Khokhlova EV, Draper LA, Forde A, et al. The human gut virome is highly diverse, stable, and individual specific. Cell Host Microbe. 2019;26(4):527-541 e525.
5. Zuo T, Sun Y, Wan Y, Yeoh YK, Zhang F, Cheung CP, Chen N, Luo J, Wang W, Sung JJY, et al. Human-gut-DNA virome variations across geography, ethnicity, and urbanization. Cell Host Microbe. 2020;28(5):741-751 e744.
6. Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B, Sullivan MB. The gut virome database reveals age-dependent patterns of virome diversity in the human gut. Cell Host Microbe. 2020;28(5):724-740 e728.
7. Yan Q, Wang Y, Chen X, Jin H, Wang G, Guan K, Zhang Y, Zhang P, Ayaz T, Liang Y, et al. Characterization of the gut DNA and RNA viromes in a cohort of Chinese residents and visiting Pakistanis. Virus Evol. 2021;7(1):veab022.
8. Nakatsu G, Zhou H, Wu WKK, Wong SH, Coker OO, Dai Z, Li X, Szeto CH, Sugimura N, Lam TY, et al. Alterations in enteric virome are associated with colorectal cancer and survival outcomes. Gastroenterology. 2018;155(2):529-541 e525.
9. Hannigan GD, Duhaime MB, Ruffin MTT, Koumpouras CC, Schloss PD. Diagnostic potential and interactive dynamics of the colorectal cancer virome. mBio. 2018;9(6):10.
10. Clooney AG, Sutton TDS, Shkoporov AN, Holohan RK, Daly KM, O'Regan O, Ryan FJ, Draper LA, Plevy SE, Ross RP, et al. Whole-virome analysis sheds light on viral dark matter in inflammatory bowel disease. Cell Host Microbe. 2019;26(6):764-778 e765.
11. Kaelin EA, Rodriguez C, Hall-Moore C, Hoffmann JA, Linneman LA, Ndao IM, Warner BB, Tarr PI, Holtz LR, Lim ES. Longitudinal gut virome analysis identifies specific viral signatures that precede necrotizing enterocolitis onset in preterm infants. Nat Microbiol. 2022;7(5):653–62.
12. Lang S, Demir M, Martin A, Jiang L, Zhang X, Duan Y, Gao B, Wisplinghoff H, Kasper P, Roderburg C, et al. Intestinal virome signature associated with severity of nonalcoholic fatty liver disease. Gastroenterology. 2020;159(5):1839–52.
13. Jiang L, Lang S, Duan Y, Zhang X, Gao B, Chopyk J, Schwanemann LK, Ventura-Cots M, Bataller R, Bosques-Padilla F, et al. Intestinal virome in patients with alcoholic hepatitis. Hepatology. 2020;72(6):2182–96.

14. Tomofuji Y, Kishikawa T, Maeda Y, Ogawa K, Nii T, Okuno T, Oguro-Igashira E, Kinoshita M, Yamamoto K, Sonehara K et al. Whole gut virome analysis of 476 Japanese revealed a link between phage and autoimmune disease. Ann Rheum Dis. 2022;81(2):278–88.
15. Chen C, Yan Q, Yao X, Li S, Lv Q, Wang G, Zhong Q, Tang F, Liu Z, Huang Y. Alterations of the gut virome in patients with systemic lupus erythematosus. Front Immunol. 2022;13:1050895.
16. Li C, Zhang Y, Yan Q, Guo R, Chen C, Li S, Zhang Y, Meng J, Ma J, You W, et al. Alterations in the gut virome in patients with ankylosing spondylitis. Front Immunol. 2023;14: 1154380.
17. Chen CM, Yan QL, Guo RC, Tang F, Wang MH, Yi HZ, Huang CX, Liu C, Wang QY, Lan WY, et al. Distinct characteristics of the gut virome in patients with osteoarthritis and gouty arthritis. J Transl Med. 2024;22(1):564.
18. de Jonge PA, Wortelboer K, Scheithauer TP, van den Born BJH, Zwinderman AH, Nobrega FL, Dutilh BE, Nieuwdorp M, Herrema H. Gut virome profiling identifies a widespread bacteriophage family associated with metabolic syndrome. Nat Commun. 2022;13(1):1–15.
19. Monaco CL, Gootenberg DB, Zhao G, Handley SA, Ghebremichael MS, Lim ES, Lankowski A, Baldridge MT, Wilen CB, Flagg M. Altered virome and bacterial microbiome in human immunodeficiency virus-associated acquired immunodeficiency syndrome. Cell Host Microbe. 2016;19(3):311–22.
20. Zuo T, Liu Q, Zhang F, Yeoh YK, Wan Y, Zhan H, Lui GC, Chen Z, Li AY, Cheung CP. Temporal landscape of human gut RNA and DNA virome in SARS-CoV-2 infection and severity. Microbiome. 2021;9(1):1–16.
21. Cao J, Wang C, Zhang Y, Lei G, Xu K, Zhao N, Lu J, Meng F, Yu L, Yan J, et al. Integrated gut virome and bacteriome dynamics in COVID-19 patients. Gut Microbes. 2021;13(1):1–21.
22. Li J, Yang F, Xiao M, Li A. Advances and challenges in cataloging the human gut virome. Cell Host Microbe. 2022;30(7):908–16.
23. Nayfach S, Paez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, Proal AD, Fischbach MA, Bhatt AS, Hugenholtz P, et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nat Microbiol. 2021;6(7):960–70.
24. Van Espen L, Bak EG, Beller L, Close L, Deboutte W, Juel HB, Nielsen T, Sinar D, De Coninck L, Frithioff-Bøjsøe C. A previously undescribed highly prevalent phage identified in a Danish Enteric Virome Catalog. Msystems. 2021;6(5):e00382-e321.
25. Chen F, Li S, Guo R, Song F, Zhang Y, Wang X, Huo X, Lv Q, Ullah H, Wang G. 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;49:103–14.
26. Zhang P, Wang X, Li S, Cao X, Zou J, Fang Y, Shi Y, Xiang F, Shen B, Li Y, et al. Metagenome-wide analysis uncovers gut microbial signatures and implicates taxon-specific functions in end-stage renal disease. Genome Biol. 2023;24(1):226.
27. Xing Y, Liu Y, Sha S, Zhang Y, Dou Y, Liu C, Xu M, Zhao L, Wang J, Wang Y, et al. Multikingdom characterization of gut microbiota in patients with rheumatoid arthritis and rheumatoid arthritis-associated interstitial lung disease. J Med Virol. 2024;96(7): e29781.
28. Li Z, Lai J, Zhang P, Ding J, Jiang J, Liu C, Huang H, Zhen H, Xi C, Sun Y, et al. Multi-omics analyses of serum metabolome, gut microbiome and brain function reveal dysregulated microbiota-gut-brain axis in bipolar depression. Mol Psychiatry. 2022;27(10):4123–35.
29. Yang J, Zheng P, Li Y, Wu J, Tan X, Zhou J, Sun Z, Chen X, Zhang G, Zhang H, et al. Landscapes of bacterial and metabolic signatures and their interaction in major depressive disorders. Sci Adv. 2020;6(49):eaba8555.
30. Tisza MJ, Buck CB. A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. Proc Natl Acad Sci U S A. 2021;118(23):e2023202118.
31. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol. 2021;39(1):105–14.
32. Aringer M, Costenbader K, Daikh D, Brinks R, Mosca M, Ramsey-Goldman R, Smolen JS, Wofsy D, Boumpas DT, Kamen DL, et al. 2019 European League Against Rheumatism/American College of Rheumatology classification criteria for systemic lupus erythematosus. Arthritis Rheumatol. 2019;71(9):1400–12.
33. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90.
34. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
35. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31(10):1674–6.
36. Li S, Guo R, Zhang Y, Li P, Chen F, Wang X, Li J, Jie Z, Lv Q, Jin H, et al. A catalog of 48,425 nonredundant viruses from oral metagenomes expands the horizon of the human oral virome. iScience. 2022;25(6):104418.
37. Wang G, Li S, Yan Q, Guo R, Zhang Y, Chen F, Tian X, Lv Q, Jin H, Ma X. et al. Optimization and evaluation of viral metagenomic amplification and sequencing procedures toward a genome-level resolution of the human fecal DNA virome. J Adv Res. 2023;48:75–86.
38. Huang L, Guo R, Li S, Wu X, Zhang Y, Guo S, Lv Y, Xiao Z, Kang J, Meng J, et al. A multi-kingdom collection of 33,804 reference genomes for the human vaginal microbiome. Nat Microbiol. 2024;9(8):2185–200.
39. Zhang P, Guo R, Ma S, Jiang H, Yan Q, Li S, Wang K, Deng J, Zhang Y, Zhang Y, et al. A metagenome-wide study of the gut virome in chronic kidney disease. Theranostics. 2025;15(5):1642–61.
40. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol. 2021;39(5):578–85.
41. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, Sun F. Identifying viruses from metagenomic data using deep learning. Quant Biol. 2020;8:1–14.
42. Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome. 2020;8(1):90.
43. Manni M, Berkeley MR, Seppey M, Simao FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38(10):4647–54.
44. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7(10): e1002195.
45. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
46. Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026–8.
47. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60.
48. Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H, Hingamp P, Goto S, Ogata H. Linking virus genomes with host taxonomy. Viruses. 2016;8(3): 66.
49. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, Chain PSG, Nayfach S, Kyrpides NC. Identification of mobile genetic elements with geNomad. Nat Biotechnol. 2024;42(8):1303–12.
50. Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, Madden TL, Matten WT, McGinnis SD, Merezhuk Y. BLAST: a more efficient report with usability improvements. Nucleic Acids Res. 2013;41(W1):W29–33.
51. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics. 2007;8: 209.
52. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44(D1):D457–62.
53. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014;42(D1):D490–5.
54. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. 2005;33(suppl_1):D325–8.
55. Alcock BP, Raphenya AR, Lau TT, Tsang KK, Bouchard M, Edalatmand A, Huynh W, Nguyen A-LV, Cheng AA, Liu S. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. 2020;48(D1):D517–25.
56. Lakin SM, Dean C, Noyes NR, Dettenwanger A, Ross AS, Doster E, Rovira P, Abdo Z, Jones KL, Ruiz J. MEGARes: an antimicrobial resistance database for high throughput sequencing. Nucleic Acids Res. 2017;45(D1):D574–80.

57. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, Philippon A, Allesoe RL, Rebelo AR, Florensa AF. ResFinder 4.0 for predictions of phenotypes from genotypes. Journal of Antimicrobial Chemotherapy. 2020;75(12):3491–500.
58. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, Rolain J-M. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother. 2014;58(1):212–20.
59. Nishimura Y, Yoshida T, Kuronishi M, Uehara H, Ogata H, Goto S. ViPTree: the viral proteomic tree server. Bioinformatics. 2017;33(15):2379–80.
60. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47(W1):W256–9.
61. Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. Picante: R tools for integrating phylogenies and ecology. Bioinformatics. 2010;26(11):1463–4.
62. Dixon P. VEGAN, a package of R functions for community ecology. J Veg Sci. 2003;14(6):927–30.
63. Yan Q, Li S, Yan Q, Huo X, Wang C, Wang X, Sun Y, Zhao W, Yu Z, Zhang Y, et al. A genomic compendium of cultivated human gut fungi characterizes the gut mycobiome and its relevance to common diseases. Cell. 2024;187(12):2969-2989 e2924.
64. Sun W, Zhang Y, Guo R, Sha S, Chen C, Ullah H, Zhang Y, Ma J, You W, Meng J, et al. A population-scale analysis of 36 gut microbiome studies reveals universal species signatures for common diseases. NPJ Biofilms Microbiomes. 2024;10(1):96.
65. Mangalea MR, Paez-Espino D, Kieft K, Chatterjee A, Chriswell ME, Seifert JA, Feser ML, Demoruelle MK, Sakatos A, Anantharaman K, et al. Individuals at risk for rheumatoid arthritis harbor differential intestinal bacteriophage communities with distinct metabolic potential. Cell Host Microbe. 2021;29(5):726-739 e725.
66. Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe J, Chisholm SW. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proc Natl Acad Sci U S A. 2011;108(39):E757-764.
67. Rodriguez-Rubio L, Martinez B, Donovan DM, Rodriguez A, Garcia P. Bacteriophage virion-associated peptidoglycan hydrolases: potential new enzybiotics. Crit Rev Microbiol. 2013;39(4):427–34.
68. Emerson JB, Roux S, Brum JR, Bolduc B, Woodcroft BJ, Jang HB, Singleton CM, Solden LM, Naas AE, Boyd JA, et al. Host-linked soil viral ecology along a permafrost thaw gradient. Nat Microbiol. 2018;3(8):870–80.
69. Jin M, Guo X, Zhang R, Qu W, Gao B, Zeng R. Diversities and potential biogeochemical impacts of mangrove soil viruses. Microbiome. 2019;7(1):1–15.
70. Bensing BA, Siboo IR, Sullam PM. Proteins PblA and PblB of Streptococcus mitis, which promote binding to human platelets, are encoded within a lysogenic bacteriophage. Infect Immun. 2001;69(10):6186–92.
71. Tian X, Li S, Wang C, Zhang Y, Feng X, Yan Q, Guo R, Wu F, Wu C, Wang Y, et al. Gut virome-wide association analysis identifies cross-population viral signatures for inflammatory bowel disease. Microbiome. 2024;12(1):130.
72. Zhou L, Zhang M, Wang Y, Dorfman RG, Liu H, Yu T, Chen X, Tang D, Xu L, Yin Y, et al. Faecalibacterium prausnitzii produces butyrate to maintain Th17/Treg balance and to ameliorate colorectal colitis by inhibiting histone deacetylase 1. Inflamm Bowel Dis. 2018;24(9):1926–40.
73. Sitkin S, Pokrotnieks J. Clinical potential of anti-inflammatory effects of Faecalibacterium prausnitzii and butyrate in inflammatory bowel disease. Inflamm Bowel Dis. 2019;25(4):e40–41.
74. Skliros D, Kalatzis PG, Katharios P, Flemetakis E. Comparative functional genomic analysis of two vibrio phages reveals complex metabolic interactions with the host cell. Front Microbiol. 2016;7:1807.
75. Liang G, Bushman FD. The human virome: assembly, composition and host interactions. Nat Rev Microbiol. 2021;19(8):514–27.
76. Gregory AC, Zayed AA, Conceicao-Neto N, Temperton B, Bolduc B, Alberti A, Ardyna M, Arkhipova K, Carmichael M, Cruaud C, et al. Marine DNA viral macro- and microdiversity from pole to pole. Cell. 2019;177(5):1109-1123 e1114.
77. Graham EB, Paez-Espino D, Brislawn C, Hofmockel KS, Wu R, Kyrpides NC, Jansson JK, McDermott JE: Untapped viral diversity in global soil metagenomes. BioRxiv. 2019:583997.
78. Li Z, Xia J, Jiang L, Tan Y, An Y, Zhu X, Ruan J, Chen Z, Zhen H, Ma Y. Characterization of the human skin resistome and identification of two microbiota cutotypes. Microbiome. 2021;9(1):1–18.
79. Santos-Marcos JA, Rangel-Zuñiga OA, Jimenez-Lucena R, Quintana-Navarro GM, Garcia-Carpintero S, Malagon MM, Landa BB, Tena-Sempere M, Perez-Martinez P, Lopez-Miranda J. Influence of gender and menopausal status on gut microbiota. Maturitas. 2018;116:43–53.
80. Zhang X, Zhong H, Li Y, Shi Z, Ren H, Zhang Z, Zhou X, Tang S, Han X, Lin Y. Sex-and age-related trajectories of the adult human gut microbiota shared across populations of different ethnicities. Nature Aging. 2021;1(1):87–100.
81. Yang K, Niu J, Zuo T, Sun Y, Xu Z, Tang W, Liu Q, Zhang J, Ng EK, Wong SK. Alterations in the gut virome in obesity and type 2 diabetes mellitus. Gastroenterology. 2021;161(4):1257–69 e1213.
82. Mosca A, Leclerc M, Hugot JP. Gut microbiota diversity and human diseases: should we reintroduce key predators in our ecosystem? Front Microbiol. 2016;7:455.
83. Larsen OF, Claassen E. The mechanistic link between health and gut microbiota diversity. Sci Rep. 2018;8(1):1–5.
84. Zeng M, Inohara N, Nuñez G. Mechanisms of inflammation-driven bacterial dysbiosis in the gut. Mucosal Immunol. 2017;10(1):18–26.
85. Serino M. Molecular paths linking metabolic diseases, gut microbiota dysbiosis and enterobacteria infections. J Mol Biol. 2018;430(5):581–90.
86. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, Guo J, Le Chatelier E, Yao J, Wu L, et al. Alterations of the human gut microbiome in liver cirrhosis. Nature. 2014;513(7516):59–64.
87. Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, Ojesina AI, Jung J, Bass AJ, Tabernero J, et al. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res. 2012;22(2):292–8.
88. Gupta VK, Kim M, Bakshi U, Cunningham KY, Davis JM 3rd, Lazaridis KN, Nelson H, Chia N, Sung J. A predictive index for health status using species-level gut microbiome profiling. Nat Commun. 2020;11(1):4635.

Publisher's Note