Vast human gut virus diversity uncovered by combined short- and long-read sequencing

Authors and affiliations:

[Joint first authors] Wei-Hua Chen, Jingchao Chen, Chuqing Sun, Yanqi Dong, Menglu Jin

[Coauthors] Senying Lai, Longhao Jia, Xueyang Zhao, Na L Gao

[Corresponding authors] Zhi Liu, Peer Bork, Xing-Ming Zhao

Keywords:

gut phageome, doubled-stranded DNA phages, crAssphage, Gubaphage, terminase, PacBio Sequel II, virus-like particle

Abstract

Known gut virus diversity is currently skewed by challenges in fecal virus-like particles (VLPs) enrichment and towards active viruses detectable in shotgun metagenomic sequencing. Here, we apply a virus detection procedure, including vigorous enrichment to harvest large quantity of VLPs, and combined Illumina and PacBio sequencing, to fecal samples of 180 Chinese volunteers. Integrated assembly of the short- and long-reads generate more and longer viral genomes compared to existing methods. The resulting viral genome dataset, referred as to the Chinese Human Gut Virome collection (CHGV), covers the full spectrum of the gut virome, i.e., an CHGV-trained machine-learning algorithm recognizes most (81~97%) public gut viruses; meanwhile, it contains 71.50% novel genomes, including 20% that cannot be recognized by machine-learning models trained on public viruses. Further analysis of the CHGV reveals a substantially higher diversity of the human gut virome. For example, we identify thousands of viral genomes that are more prevalent than crAssphages and Gubaphages, the two most diverse phages in the human

gut, and several viral clades that are more diverse than the two. Tother, our results indicate a vastly enlarged gut viral diversity that significantly broadens our knowledge on the viral dark matter of the human gut microbial ecology.

Background

The gut viral community (also known as the gut virome), mainly bacteriophages and archaeal viruses, has been shown to be diverse in the human gut [1, 2]. Viruses play crucial roles in shaping the gut microbial composition, and hold great promise for precision manipulation of the gut bacteriome. Despite tremendous success in identifying human (gut) viral genomes [3-8], the gut virome has been far less characterized than the prokaryotic community [9, 10]. Most importantly, the diversity of the gut virome has been vastly underestimated because of biological and technical challenges [11]. In particular, viruses identified from whole microbial community sequencing (metagenomics) favor highly abundant ones, while virus-like particle (VLP) sequencing often uses whole genome amplification (WGA) methods such as multiple displacement amplification (MDA)[12] to obtain sufficient amounts of DNAs for next-generation sequencing (NGS) [13-18] and favors short and circular viruses [19, 20].

Here, we apply a rigorous VLP enrichment protocol on an increased amount of feces (ca. 500g per person) to obtain sufficient quantity of double-stranded viral DNAs for both short-(Illumina) and long-(PacBio) read sequencing without whole genome amplification (WGA). Applying it to fecal samples of 180 Chinese healthy volunteers, we construct a collection of 97,660 non-redundant viral genomes by integrated assembly of the short- and long-reads. We obtain more and longer viral genomes than short-read only methods, and overcome the biases of the MDA. This viral dataset, referred as to the Chinese Human Gut Virome collection (CHGV), covers the full spectrum of human gut virome. For example, a machine learning (ML) model built on it identifies most (81~97%) of the human (gut) viruses from public datasets [3-8]; meanwhile, it contains 71.50% (69,831) novel viruses that do not share an average nucleotide identity (ANI) over 95% with the public datasets [3-8] and ~20% that cannot be recognized by viral prediction models trained on any of the

public human (gut) viral datasets. We characterize the CHGV genomes, and identify thousands of viruses that are more prevalent than crAssphages and Gubaphages (the two most diverse gut viral clades known so far [4]) using experimentally validated thresholds. Furthermore, we cluster the CHGV viruses into viral clades (CL) using similar methods that the crAssphage and Gubaphage clades are defined, and identify multiple CLs with higher diversities. By comparing to the bacteriome of the same samples, we confirm that the gut virome is indeed individual-specific at both viral population (VP, equivalent to species) and viral cluster (VC, equivalent to genus) levels; however, the gut virome is significantly more prevalent than the gut bacteriome at the genus level, indicating the large distances between two individual virome profiles are mostly attributed by the more complex community compositions. Together, we reveal the hidden diversity of the gut virome using combined short- and long-read sequencing, and broaden our knowledge on the viral dark matter of the human gut microbial ecology.

Results

Combined short- and long-read sequencing of human fecal viral-like particles (VLPs) identify more and longer viral genomes

To survey the human gut virome without known limitations of previous methods, we applied a rigorous VLP enrichment protocol to an increased amount of feces (~500 grams per sample, corresponding to about three stools) from 180 healthy participants of Chinese residence to extract large quantity of high-quality and high-molecular-weight (HMW) DNAs from double-stranded DNA viruses (mostly bacteriophages and archaeal viruses; Methods). We submitted all samples to viral next-generation sequencing (vNGS; Illumina HiSeq2000) and those with sufficient amounts of DNAs to third-generation sequencing (vTGS; PacBio Sequel II). For comparison, we also submitted all fecal samples to regular whole microbial community sequencing (mNGS) (Fig. 1A). In total, we obtained 169 vNGS, 101 vTGS, and 167 mNGS sequencing datasets.

We evaluated the virome enrichment efficiency by directly comparing the identifiable bacterial and viral contents of the vNGS datasets (VLP-enriched) to the

matching mNGS counterparts (unenriched; Fig. 1A). After removing contaminations from human and common food genomes, we aligned ~32.54% of the clean vNGS reads and ~15.82% of the clean vTGS reads to the bacterial and archaeal genomes in the Unified Human Gastrointestinal Genome (UHGG) collection (excluding annotated prophages, also referred to as UHGG-Minus; Methods) [21], which were significantly less than the mNGS data (~84.92%, Fig. 1B; Methods). For comparison, we simulated viral NGS reads from several human (gut) viral genome collections [3-8] and were able to align ~26.65% of them to the UHGG genomes (red dashed line in Fig. 1B; Methods). These results suggest the gut virome and bacteriome can share significant amount of sequences, and only a small proportion of the vNGS reads in our data were true bacterial contaminations. Regardless, we removed the UHGG-aligned reads from our vNGS and vTGS datasets, and only used the "filtered" datasets for further analyses unless otherwise stated.

To quantify the virome enrichment, we compared the percentage of viral reads in the vNGS data of a vNGS data to that of its matching mNGS data, and defined a Fold-Enrichment for Virome (FEV) score as the ratio of the two percentages; here the "viral reads" are those can be mapped to "viral contigs" annotated by a customized bioinformatics pipeline adopted from previous studies [22-24] (Methods). Of note, because the existing tools rely on known viral/phage hallmark genes and recognize a fraction of gut viruses, the FEV scores can underestimate the virome enrichment. Regardless, 78.44% of the filtered vNGS samples had FEV scores ≥5 (median 11.85, mean 20.59; Fig. S1A, left panel). We further confirmed the virome enrichment using ViromeQC, a state-of-the-art tool for virome enrichment evaluation [25]; as shown in Fig. S1A, 76.04% of the filtered vNGS samples had ViromeQC scores≥5 (median 11.23, mean 21.44; Fig. S1B), although ViromeQC in general produces lower enrichment scores (Fig. S1B). Together, these results support the high efficiency of our VLP enrichment protocol. Regardless, to avoid any possible bacterial contaminations, we removed samples with FEV scores <5 and retained in total 135 vNGS, 83 vTGS, and 135 mNGS samples for further analysis (Table S2).

We obtained in total 279,133 non-redundant viral populations (VPs) using an integrated assembly strategy of the short- and long-reads (Methods, Fig. S2). Briefly, for each sample, we assembled the vNGS data using IDBA-UD [26], vTGS data using Canu [27] and Flye [28], and both (if available) using OPERA-MS [29] and metaSPAdes [30] (hybrid assemblies). We grouped contigs (or unitigs from Canu) into bins using MetaBAT2 [31], merged the bins from all assemblies, and dereplicated using CD-HIT with a global identity threshold of 95% [32] (Methods). The resulting viral bins were referred to as viral populations (VPs) that correspond to species-level taxa [3].

Among the non-redundant VPs, 97,253 (34.84%), 54,147 (19.40%) and 28,412 (10.18%) had lengths longer than 5, 10 and 20 kilobases (kbs) respectively (Fig. S3A). The three assembly strategies, namely vNGS, vTGS and hybrid assemblies contributed 37.80% (36,917), 16.39% (16,005) and 45.81% (44,737) VPs respectively (Fig. S3C). The TGS assembly generated the longest VPs with a median length of 22 kbs, followed by NGS assembly (10 kbs) and hybrid assembly (9 kbs; Fig. S3C). Rarefaction analyses suggested that the numbers of unique VPs were far from saturation (Fig. 1D), consistent with previous studies [33]. We obtained a similar rarefaction curve using a larger dataset consisting of 357 MDA-amplified samples (vMDA) from the Gut Virome Database [3] (GVD; Fig. S3D). These results suggest that more samples are required to uncover novel gut viruses at the VP/species level. Our method obtained significantly more unique VPs than public virome sequencing datasets including GVD (MDA amplified; Fig. S3D) and DEVoC (Danish Enteric Virome Catalog ; VLP enrichment followed by whole transcriptome amplification; see ref. [7]) with the same numbers of samples. As expected, the hybrid assembly recovered the highest number of VPs, followed by our vNGS data and then the GVD-vMDA (Fig. S3D). We found the same trends that our methods could recover more longer VPs (Fig. S3D).

We also applied an MDA amplification protocol (adapted from ref. [34]) to a subset of 20 samples from CHGV, and compared the sequencing results (vMDA data) with the vNGS data of the same samples. Overall, the vMDA data generated fewer and shorter contigs that were biased towards circular genomes (Fig. S4; $p = 3.61*10^{-11}$, Chi-squared test), further confirming that our protocol could overcome biases of existing methods.

As short contigs may represent fragmented phage genomes, we kept the 97,660 VPs that are either longer than 5kb or longer than 1.5kb but circular for further analyses , and referred to this dataset as the Human Gut Virome (CHGV) collection.

To check if the CHGV genomes can represent most viruses in our samples, we used them to recruit the filtered vNGS reads, and compared the mapping rates to those of the public viral databases. The CHGV genomes recruited most of the vNGS reads with the highest median mapping rates of 98.13%, which were significantly higher than those of the public databases ($p < 0.001$, paired Mann-Whitney test; Fig. 1E). The recently assembled gut virome datasets including GVD, GPD (the Gut Phage Database) [4], MGV (the Metagenomic Gut Virus catalog) [5] also recruited most of the vNGS reads (median of 76.84-90.13%), confirming that most of the reads (and hence the assembled genomes) were indeed viral. However, the overall mapping rate to the DEVoC dataset was significantly lower (median: ~50%), likely due to that the latter was enriched for RNA viruses [7]. We obtained the lowest mapping rates to the NCBI Viral RefSeq database (median: 0.18%).

Together, we show that our vigorous VLP enrichment protocol in combination with increased amount of feces and integrated assembly of short- and long-reads could obtain more and longer viral genomes, and overcome biases of existing methods.

CHGV captures essential sequence characteristics of the gut virome while contains a significant proportion of novel viruses

To check whether the CHGV genomes capture the essential sequence characteristics of the gut virome or viruses in general, we trained a virus detection machine learning (ML) model using the CHGV genomes as the true positives and a subset of the UHGG genomes as the negatives (Methods), and tested on a test dataset consisting of viruses from IMG/VR [35] as positives and the left-out UHGG-Minus [21] genomes as negatives. For comparison, we also trained ML models using the public virome datasets including GPD, GVD, GVD, DEVoC and CHVD-gut (the Cenote Human Virome Database; only gut viruses are used, see Methods) [6] as true positives and applied them to the same test dataset. The CHGV-model correctly

distinguished viral from bacterial genomes with an AUC (area under the receiver operating characteristic) of 0.93, and performed comparable to other models (AUC: 0.95-0.97; Fig. 1F). In contrast, a model trained on the NCBI Viral RefSeq as the true positives achieved a much lower AUC score of 0.78, while VirFinder [23], a popular tool for virus identification, achieved an AUC score of 0.80 (Fig. 1F).

To check which of the above models can better represent the human gut virome, we applied them to all the human gut virome databases and calculated the recognition rates as the percentage of correctly classified viral genomes in individual databases. As shown in the recognition matrix in Fig. 1G, our CHGV-model correctly recognized 81% to 97% of the gut viral genomes and ranked the first among all datasets with a Mean recognition rate Across Datasets (MAD, excluding self-prediction; the higher the better) of 91%. The GVD-model performed similarly with a MAD of 91% (Fig. 1G), followed by the three models trained on metagenome-derived gut virome catalogs (i.e., GPD, MGV, and CHVD-gut) with MADs of 84-86% (Fig. 1G). The DEVoC database ranked the last with a MAD of 84%, likely due to its focus on RNA viruses and low sequencing depth (~5GB per sample; ref [7]).

We also examined the novelty in each database using a Mean Across Models score (MAM) by averaging the non-self-recognition rate of a column in the recognition matrix (Fig. 1G); a lower MAM score indicates higher proportions of viruses not recognized by models trained on other databases, thus higher novelty. Our CHGV genomes had the lowest MAM score (58%), followed by GVD (76%) and CHVD-gut (92%). Interestingly, the three models trained on the metagenome-derived gut virome databases including GPD, MGV, and CHVD-gut had much higher recognition rates among themselves (93-99%; Fig. 1G) than to the VLP-sequencing assembled databases such as the CHGV and GVD (52-78%; Fig. 1G), consistent with our expectation that the VLP-sequencing rather than the bulk metagenomics better represent the gut virome. As expected, supplementing the three bulk-metagenomics virome databases with CHGV (GVD) genomes significantly improved their recognition rates on the GVD (CHGV) genomes (Fig. 1G, Fig. S5). Surprisingly, although the DEVoC database were also based on VLP-sequencing, it had the second highest MAM score of 96% (Fig. 1G), likely due to its unique methods [7]. Together, the machine learning

models trained on the public databases could recognize in total 72.46% (70,760 out of 97,660) of CHGV genomes.

To estimate the number of novel viral genomes in the CHGV, we directly compared the CHGV sequences against those in the public viral databases, and found that 76,776(78.61%) of which did not share an average nucleotide identity (ANI) over 95% with any of the public datasets (Table S3). Further comparison with relaxed criteria (E-value<1e-10 and coverage>50%) using BLASTn [36] identified additional 20.05% (19,581 out of 97,660) homologous sequences in the CHGV as compared with public databases. Interestingly, we found that the CHGV covered higher percentages of viral genomes of other databases than the other way round (Fig. 1H). For example, 58.04% of the GPD genomes had significant BLASTn hits (E-value<1e-10 and coverage>50%) in the CHGV database, but only 37.09% of the CHGV genomes had BLASTn hits in the GPD database (Fig. 1H). Thus, at least ~20% of the CHGV genomes represented completely novel viruses (Fig. S6).

The above results indicate that the CHGV genomes could be used to improve the coverage of human gut virome in public datasets. To prove this, we downloaded 24 samples of amplification-independent VLP sequencing data from the project PRJEB42612 [37, 38] (pub-vNGS). After the removal of human and bacterial contaminations, ~85.51% of the clean reads could be aligned to the combined dataset of the CHGV and public virome dataset (i.e., the combination of the GVD, GPD, CHVD-gut, MGV and DEVoC databases), significantly higher than the public virome datasets alone (Fig. 1I; 73.32%; $p = 1.19*10^{-7}$, paired Mann-Whitney test). We obtained similar results on the 300 MDA amplified VLP samples (referred to as pub-vMDA data) from PRJNA588514 [15], and found that ~80.30% of the clean reads from the pub-vMDA data could be mapped to the combined dataset compared with ~76.28% of the public virome databases alone (Fig. 1J; $p = 2.8*10^{-14}$, paired Mann-Whitney Test). The overall lower mapping rates of the vMDA reads to all the databases were also confirmed in our vMDA data (Fig. S7B), likely due to artifact synthesis that could account for up to 20~30% of the total reads [37].

Together, we showed that the Human Gut Virome (CHGV) collection not only well represented the human gut virome and viruses in public databases in general, but also contained a significant proportion of novel viral genomes.

CHGV contains viral clades that are more diverse than crAssphages and Gubaphages

We next investigated the diversity of the CHGV genomes by first grouping the VPs into viral clusters (VCs) using MCL (Markov Cluster), a graph-based clustering algorithm [39]. The VPs in a VC share at least 90% nucleotide identities with the aligned regions spanning more than 75% of their genome lengths, thus likely belonging to the same genus [4]. We obtained in total 5,011 VCs with two or more VPs, consisting of 13,150 (13.47% out of the total) VPs (Methods) and 84,510 singleton VCs. We ranked these VCs according to their sizes (i.e., the numbers of contained VPs; larger VCs are considered more diverse [4]) and identified 38 VCs with ten or more VPs, with the largest VC of the size of 39 (Fig. 2A; Table S4). Most of the genomes in these top VCs were recognized as viruses by our bioinformatics pipeline (Methods; Fig. 2A; Table S4).

We also annotated the VCs for crAssphages and Gubaphages, the two most diverse clades in the gut [4] by following the pipelines described previously [4, 40] (Methods). The top two VCs, i.e., VC_1 and VC_2, were identified as crAssphages (Fig. 2B), confirming that they were indeed more diverse than others at the VC/genus level. Together, we identified in total 146 crAssphage VCs, including VC_5 and VC_7 among the top ten VCs (Fig. 2B); they corresponded to 14 putative genera, consisting of in total 431 VPs; these results are consistent with previous studies that crAssphages can be classified into several genera and a few families [40]. 67 Gubaphage VCs were also identified, consisting of 232 VPs; among them, three (i.e., VC_6, VC_8, and VC_9) were among the top ten VCs. Surprisingly, the largest Gubaphage VC ranked only at the 6th in our database (Fig. 2B), suggesting that the CHGV database contained at least two VCs (i.e., VC_3 and VC_4) that were more diverse than the Gubaphages at the genus level.

Phylogenetic analysis of the genomes in the above-mentioned VCs using protein sequences of the terminase genes revealed that all crAssphages (i.e., those in the VC_1, VC_2, VC_5, and VC_7) and Gubaphages (i.e., those in the VC_6, VC_6, VC_9) were clustered

into their clades (Fig. 2B). The VC_3 viruses were close to the Gubaphages than to the crAssphages, while the VC_4 viruses formed their own clade in parallel to a sister clade that contained all Gubaphages and crAssphages (Fig. 2B). The long branch lengths between these VCs suggested that the two clusters were divergent from the Gubaphages and crAssphages, which was further supported by little overall protein similarities between the two clusters and other VCs (Fig. 2C; Methods). In addition, lifestyle analysis of the VPs using DeePhage [41] indicated that both VC_3 and VC_4 were "uncertain temperate" viruses, in contrast to the predicted "virulent" lifestyles of all the other VCs (Table S5; Methods). Further research is required to illustrate their high divergence and functions in the human gut.

The observation that the terminases could be used to correctly cluster multiple genera of the Gubaphages and crAssphages into their clades (Fig. 2B) had prompted us to use them to group all VCs into higher taxonomic levels. Terminases are essential for DNA recognition and initiation of viral DNA packaging [42], highly conserved in several viral families [4, 43], and thus suitable for such an analysis. We therefore first annotated putative terminase-encoding genes in the CHGV database and then clustered them using MCL with their protein sequence similarity scores as input (Methods). In total 8.05% (7,202 out of 89,521) of VCs were found to encode putative terminases. For VCs with at least one VP that encodes the similar terminase, these VCs will be considered to be in the same clade. We tested different protein similarity thresholds of the terminases for the subsequent MCL clustering and chose the BLASTp sequence identity of 70 (qid 70) as the final cutoff, which could cluster the highest numbers of crAssphage and Gubaphage VPs into their own clades that contained the least numbers of other VPs (Fig. S8; Methods). Of note, this cutoff was stricter than those used in the literature for crAssphage identification [40]; consequently, the crAssphages were grouped into several clades (Table S6).

We obtained in total 762 viral clades (CLs), consisting of 11,084 VPs (including 9,720 terminase-containing VPs and those in the same viral clusters, Table S6). Among these, the largest Gubaphage and crAssphage clades ranked the 2nd and 21st, and contained 232 (100% of 232 Gubaphages) and 98 (22.73% of 431 crAssphages) VPs respectively. The largest clade, CL_1, contained in total 318 VPs. Pair-wise analysis of shared protein clusters

(PCs) among clade members revealed that all the above-mentioned clades contained several sub-clades with higher fractions of shared PCs within sub-clades but lower fractions of shared PCs between sub-clades (Fig. 2C). The shared PC analysis measures function similarity between two genomes; a higher (lower) fraction of shared PCs indicates closely (distantly) related genomes. A sub-clade corresponds to one or more VCs (Fig. 2B). Similar results were found for crAssphages from a public dataset [40] (Fig. S9), and in Gubaphages as reported previously [4]. These results suggested that the terminase can be used to group a substantial fraction of gut viruses into higher clades.

Together, we identified two viral clusters (VCs) that were more diverse than Gubaphages at the VC/genus level, and one and 19 viral clades (CLs) that were more diverse than crAssphages and Gubaphages respectively at higher levels.

CHGV contains thousands of viral genomes more prevalent than crAssphages and Gubaphages

crAssphages and Gubaphages represent the most abundant and prevalent viral/phage clades in the human gut. A recent study based on short-read sequencing of enriched VLPs further confirmed this because the authors did not find more prevalent viruses than the two clades [7]. We revisited this issue by first determining the best relative abundance cutoff to define the presence of a virus in samples of interests, as it had strong effects on the viral prevalence calculation. A cutoff of 1e-3 (0.1%) is often used in bacteriome analysis [44, 45]. Here, we tested several lower cutoffs because the gut viral community is more complex than the bacteriome [14]. We randomly selected 21 CHGV viruses, including 5 crAssphages and 6 Gubaphages, and validated their presence using Polymerase Chain Reaction (PCR) in our enriched VLPs with calculated relative abundances ranging from 1e-5 (0.001%) to 1e-8 to non-detectable (ND) (Methods). We confirmed the presence of the selected viruses in 93% (25 out of 27 tested samples) at the 1e-5 threshold (Fig. 3A), suggesting that the latter could be used as the threshold for viral prevalence calculation. We found the validation rates decreased rapidly with decreasing thresholds. As expected, the PCR validation rates were the lowest in the ND samples (Fig. 3A). At the relative abundance threshold of 1e-5,

we estimated that there could be 3,748 and 1,403 CHGV genomes that were more prevalent than the most prevalent crAssphage and Gubaphage genomes respectively. We further selected 7 CHGV virus (software recognized, virulent) to validated their presenc, including 2 crAssphages and 2 Gubaphages. We found strong positive correlation between vNGS prevalence and PCR prevalence (Fig. 3B). Of the three phages that were selected but neither were crAssphage nor Gubaphages, one (A09_NODE_109) has both higher vNGS and PCR prevalence than these known phages with mean abundance of 1.5e-3. N-acetylmuramidase (PF11860.10, commonly known as lysozymes) and Phage Tail Collar Domain(PF07484.14) were found in A09_NODE_109, suggesting that it's a virulent, Caudovirales virus. The other 37 proteins encoded by this phage were hypothetical proteins without any annotation.

The human gut virome is more divergent and prevalent than gut bacteriome

The human gut virome is known to be highly divergent and individual-specific, as indicated by high pairwise dissimilarities in the virus abundance profiles [13]. However, the key parameters governing the dissimilarity matrixes of the gut virome and bacteriome have not been directly compared, including the complexity of the gut virome community and the prevalence (and also the relative abundances) of the community members, because of the lack of both types of data from the same samples. A greater dissimilarity could be observed in a community with complex structure (i.e., high Shannon index) and low prevalent/abundant members [46]. Here, we excluded VPs with relative abundance lower than 1e-5 from further analyses (the cutoff we experimentally validated; Methods), and calculated the two parameters for the remaining 21,652 VPs and their corresponding 20,572 VCs. For comparison, we determined the taxonomic classifications of the mNGS data and their relative abundances using Metaphlan2 [47] and also removed the low-abundance genomes with the same criteria. Since the VP and VC levels of the virome are equivalent to the species and genus levels of the bacteriome, we thus calculated the two parameters for both and compared them at the same levels accordingly.

We found significantly greater pairwise distances in the gut virome profiles than those of the gut bacteriome, as indicated by both the Bray-Curtis Dissimilarity (taking

abundance information into account) and Jaccard Distance scores (considering only presence/absence information); we found similar trends at both the VP/species and the VC/genus levels (Fig. 3C). Most Bray-Curtis dissimilarities between virome samples are close to 1, suggesting that most of the sample pairs had very different viral structures (Fig. S10A). Conversely, the differences in the gut bacteriome were significantly smaller (Fig. 3C; Fig. S10A&B). The co-occurrence analyses at both VP and VC levels further confirmed the dissimilarities (Fig. S10C).

We used the Shannon diversity (as known as alpha diversity) as an indicator for community complexity and noticed significantly higher Shannon indices in the gut virome than the bacteriome at both the VP/species and VC/genus levels (Fig. 3D), consistent with previous results [14]. Strikingly, at the VP/species level, we found a significantly higher prevalence of the gut virome than the bacteriome (Fig. 3D): we found that the VPs had a median prevalence of 12.59%, significantly higher than that of the bacterial species (median prevalence of 2.31%; $p = 1.2*10^{-12}$, *t*-test). We found a similar trend at the VC/genus level that the VCs had a median prevalence of 13.33%, higher than the median prevalence of 5.62% of the bacterial genera (Fig. 3D). These results indicate that the human gut virome is highly prevalent and divergent than the gut bacteria; the higher between-sample dissimilarities were attributed to the more complex community structure, rather than the low prevalence.

With the availability of unbiased virome abundance (i.e., VLP-enriched and amplification-independent) data, we next attempted to group our samples into clusters (also known as enterotypes) using the Partitioning Around Medoids (PAM) clustering algorithm on the VC level (Methods); the similar algorithm has been used for enterotype discovery in the gut bacteriome using genus-level relative abundances [48]. Finally, we obtained three clusters using the Cao index (or CYd index) distance matrix [49, 50], where the three groups respectively contained 122, 8, 5 samples (Fig. 4C), indicating a rather unbalanced distribution of the samples in the three viral "enterotypes". We identified driver VCs for Clusters B and C (Fig. 4D), where Cluster B was dominated by VC_2, a crAssphage sp. C0531BW4, while Cluster C is dominated by VC_8, phage FAKO05_000032F, annotated as Gubaphage (Table S7). The two VCs showed the highest abundances in the

samples of their respective enterotypes (Fig. 4C). However, we could not identify driver VCs for Cluster A, which harbored most of the samples. Further study is thus required for the discovery of enterotypes in gut virome and the stability of the obtained clusters.

Together, we confirmed that the gut virome was highly diverse and individual-specific. However, this was mostly driven by the high complexity in the gut viral community structure (i.e., alpha diversity) rather than the low prevalence; in fact, the human gut virome was both highly prevalent and highly diverse than the gut bacteriome.

Human gut virome is mostly temperate but dominated by virulent viruses

The observation that the virulent viruses accounted for eight of the top ten VCs (Fig. 2A; Table S5) has prompted us to further characterize the lifestyles of the human gut virome with Deephage [41], which uses a scoring system to classify viral genomes into four categories, including temperate (with scores <=0.3), uncertain temperate (0.3~0.5), uncertain virulent (0.5~0.7) and virulent (>0.7). Of all the CHGV genomes, 77% were classified as either temperate (12,097) or uncertain temperate (59,688); in contrast, only 23% were classified as either uncertain virulent (23,566) or virulent (2,309) (Fig. 4A). We found similar results in GVD and GPD, which consisted of 60% and 68% temperate or uncertain temperate phages respectively (Fig. 4A; Fig. S11A). Thus, the human gut virome is mostly temperate, consistent with previous results[3, 51].

However, we found that the virulent viruses were significantly more abundant than the temperate ones. For example, the total abundances of the virulent viruses accounted for 52.01% of the overall abundances, although they accounted for 26.5% of the viruses (including both virulent and uncertain virulent ones) (Fig. 4A; Methods). The mean abundances of the viruses increase with increasing Deephage virulence scores (Fig. 4B). We found similar trends in viral prevalence. For example, by using a relative abundance of 1e-5 as the cutoff for presence/absence of a virus in the samples, we found an increasing trend of the prevalence with increasing virulence scores (Fig. 4B; Methods). However, the trend can be strongly affected by the choice of the relative abundance cutoffs (Fig.4A).

Regardless, our results suggested that the virulent viruses represented the active part of the human gut virome and their virulent lifestyle contributed significantly to their expansion within and across samples.

Discussion

Effective and unbiased discovery of viral genomes is vital for exploring and understanding the human gut virome [33]. Here, we present the first comprehensive survey of the human gut virome using both long- and short-read sequencing techniques on feces-enriched virus-like particles (VLPs). To obtain sufficient amounts of viral DNA for long-read sequencing without genome-wide amplification, we collected 500 grams of feces from each of the 180 participants and submitted the samples to a rigorous VLP enrichment procedure. After the removal of putative contaminations from the host, food, bacterial and archaeal genomes, we obtained in total 279,133 unique viral genomes (VPs) longer than 1.5 kbs through a combined assembly using both the long- and short- reads. Our method, especially the long-read sequencing, is more efficient in recovering more and longer gut viral genomes than the amplification-dependent methods (e.g., MDA), short-read data alone (Fig. 1D, Fig. S3D,E) and other research using vNGS data [7], and overcomes biases of the MDA (Fig. S4). In the end, we excluded fragmented genomes from the dataset and constructed a Human Gut Virome (CHGV) collection consisting of 97,660 unique VPs that were either >=5kb or circular and >=1.5kb for further analyses.

We proved that the CHGV genomes captured the essential sequence characteristics of the human gut virome and supplemented the existing database with new viral genomes and sequence features. The machine learning (ML) model built on the CHGV identified most of the human gut phages in public databases including GPD, GVD, MGV, CHVD and DEVoC (81~97%; Fig. 1 F&G). However, the ML models trained on the public databases could only recognize on average 58% (52~68%) of the CHGV genomes. In addition, 69,831(71.50%) of the genomes in the CHGV database did not share an average nucleotide identity (ANI) over 95% with any of the above-mentioned databases. Together, 20% of the CHGV

genomes are novel. Overall, the CHGV genomes significantly improved the coverage of the existing databases on the public gut virome sequencing data (Fig. 1 I, J).

The CHGV genomes allowed us to unveil previously unappreciated features of the human gut virome and revisit established ones. For example, by clustering viral genomes into putative higher taxonomic levels using overall sequence similarities and that of a marker gene, we identified several viral clades (CLs) that contained more members (i.e., more diverse) than those of the crAssphages and Gubaphages, and several viral clusters (VCs) that were more diverse than the Gubaphages. We also identified thousands of gut viruses that were more prevalent than the crAssphages and Gubaphages, by using an experimentally established relative abundance threshold of 1e-5 for the presence of viruses in VLP-enriched samples, and further validated the prevalence of several selected viruses. We found that the human gut virome was mostly temperate, but dominated by the virulent viruses that were more diverse, prevalent, and abundant. We confirmed that the gut virome was highly diverse; however, in contrast to the previous conclusion that the greater dissimilarity between any pairs of viral samples was because of the low prevalence of most viruses, we found that the CHGV viruses were more prevalent than the bacterial clades at the same taxonomic levels (species and genus) in the same samples; thus, the greater dissimilarity in viral profiles was mostly contributed by the more complex community structure.

Together, our results revealed the vast diversity of the gut virome that was unappreciated using short-read sequencing data and significantly improved our understanding of its key characteristics.

Figure Legends

Figure.1 | Profiling of human gut virome by extensive viral-like particle (VLP) enrichment and combined assembly of short- and long-reads. **A**, Overall experiment design and the generation of CHGV (the Human Gut Virome) collection. **B**, Evaluation of bacterial contaminations by mapping the sequencing reads to the UHGG-Minus database (i.e., UHGG genomes excluding putative prophage fragments). The red dashed line is the mean mapping rate of the simulated reads from annotated human (gut) viral genomes in public

databases including GVD [3], GPD [4], MGV [5], CHVD [6], DEVoC [7] and mMGE [8] (Methods). **C**, UHGG-Minus mapping rates of the unfiltered vTGS and filtered vTGS datasets. **D**, Rarefaction curves of unique viral contigs obtained from short-(vNGS) and combined assemblies. Here the combined assemblies stand for union dataset of unique viral contigs assembled from short-(vNGS), long-(vTGS) reads or both (vNGS+vTGS, hybrid) datasets. **E**, Proportions of the reads from the filtered CHGV vNGS data mapped to the viral genomes in the CHGV, GVD, GPD, MGV, DEVoC and NCBI Viral RefSeq (NCBI_ref) genomes. **F**, Performance of machine learning (ML) models trained on viral contigs from the CHGV dataset from this study and several public datasets including GVD, GPD, CHVD, MGV and NCBI Viral RefSeq, and tested on a dataset from IMG/VR; in addition, VirFinder was performed directly on the test set. **G**, Cross-dataset validation results of the ML models built from different virome databases as training; shown here are the proportions of viral genomes correctly recognized (Test Set; X-axis) by the models (Model; Y-axis). MAD: Mean recognition rate Across Datasets, calculated as the mean value of the row excluding self-predictions (i.e., when the test and train datasets are the same); MAM (Mean Across Models, calculated as the mean value of the column excluding self-predictions (i.e., when the test and train datasets are the same). **H**, The BLASTn hit ratios of the CHGV genomes against those in the GVD, GPD, CHVD, MGV, and DEVoC databases, and the other way round. **I**, **J** The read mapping rates of public vNGS dataset (PRJEB42612, **I**) and public vMDA dataset (PRJNA588514, **J**) against the NCBI Viral RefSeq and the combination of the public viral databases including the GPV, GVD, CHVD, DEVoC and MGV databases (referred as to "Public Viral Dataset" in the plots), and the combination of the "Public Viral Dataset" and CHGV (referred as to Public dataset + CHGV).

Figure 2 | The viral clusters and clades identified from CHGV. **A**, Top viral clusters (VCs) ranked by their diversities, i.e., numbers of contained VPs. Each dot represents a VC; the filled colors indicate the proportions of the member VPs recognized as viruses by our bioinformatics pipeline (Methods); darker colors indicate higher proportions. The inset plot shows a blow-up of the top 20 VCs. **B**, Phylogenetic analysis of the top 10 VCs using terminase protein sequences. Different colors were used to indicate genomes from

different VCs, and randomly selected crAssphages (white) and Gubaphages (white) from public databases. The selected T7 terminases were used as an outgroup (blue). **C**, Shared protein cluster (PC) analysis within selected clades, where the fraction ranges from 0 to 1 and higher fraction indicates higher functional similarity and closer phylogenetic relationship.

Figure 3 | The prevalence and diversity analysis of the human gut virome. **A**, The PCR positive rate under different vNGS abundance cutoff. 1e-5 showed a overall higher positive rate, and was chosen to be the abundance cutoff in prevalence calculation. **B,** The Linear regression result of PCR prevalence and vNGS prevalence. There was a strong positive correlation between PCR and vNGS prevalence. Due to the high sensitivity of PCR, it showed a much higher prevalence. Suggesting that the current value of calculated prevalence might be underestimated. **C**, Distributions (presented as density plots) of the pair-wise Bray-Curtis Dissimilarities in the gut virome (yellow) and bacteria (blue) at VP/species (upper panel) and VC/genus (lower panel) levels. **D**, The comparison of prevalence (left panel, Violin plot) and divergence (Shannon Index; right panel, boxplot) between gut virome and bacteria. Level of significance: $p < 0.001$, *t*-Test.

Figure 4 | Lifestyle and bacterial host analysis of the human gut viruses. **A**, Lifestyle assignments of CHGV and GPD VPs. **B**, The prevalence (left panel) and abundances (right panel) of the CHGV-VPs. **C**, The three groups were obtained by Partitioning Around Medoids (PAM) clustering algorithm with cluster sizes of 122, 8, and 5, respectively. **D**, The relative abundances (presented as a stacked bar plot) of the VCs (here only present the VCs with top 10 abundance in each sample) in our vNGS samples grouped with PAM. Cluster B is dominated by VC_2 (crAssphage), while Cluster C is dominated by VC_8 (Gubaphage).

Supplementary Figure1 | **A,** Evaluation of virome enrichment over the unfiltered and filtered vNGS datasets using FEV score (left panel) and ViromeQC score (right panel). **B**,

ViromeQC score of GVD MDA samples and CHGV vNGS samples. Compared with GVD MDA samples, CHGV vNGS samples showed an overall higher ViromeQC score in both before and after bacterial reads removal datasets. **C,** Comparison of virome enrichment evaluations using FEV (X-axis) and ViromeQC (Y-axis) scores in the unfiltered vNGS dataset. Also shown are the linear regression results (the blue line and the equation) between the two sets of scores.

Supplementary Figure2 | An overview of our assembly strategy

Supplementary Figure 3 | **A**|**B**, bar plot showing the count and software annotation rate of CHGV contigs in different length range. The longer the contig is, the more likely it is to be recognized by virus detection software. **C**, Rarefaction curve of contigs from vNGS-assembly and contigs from GVD MDA samples. About 134,882 contigs(>=1.5kb) assembled from vNGS datasets (48,676 with length >= 5kb, 28,259 with length >= 10kb, in total 135 samples), while in datasets from GVD-vMDA samples(containing 357 samples) the numbers are 148,019 contigs(>=1.5kb), 35,399 contigs(>=5kb) and 14,725 contigs(>=10kb). C, The three assembly strategies, namely vNGS, vTGS and hybrid assemblies contributed 37.80% (36,917), 16.39% (16,005) and 45.81% (44,737) VPs respectively. TGS assemblies have significantly longer genome length ($p < 0.001$, T-Test). **D,E** |Rarefaction curves of unique viral contigs obtained from short-(vNGS) assemblies, public MDA assemblies (D), and recently published DEVoC assemblies (E).

Supplementary Figure 4 | MDA preference. Up left panel, Stack bar plot displaying the percentage of circular and linear/incomplete genomes in vMDA and vNGS datasets. The assembled contigs of the MDA datasets tend to have more circular genomes ($p < 0.01$, Chi-test). Up right panel, Genome count of vMDA and vNGS genomes assembled from samples with both vMDA and vNGS in different length range. vMDA assembled much less contigs than vNGS. Bottom left panel, prevalence distribution of vNGS and vMDA VPs of the same

amount of samples (20 samples). Bottom right panel, Shannon diversity distribution of vNGS and vMDA VPs of the same amount of samples (20 samples).

Supplementary Figure 5 | The representativeness of GVD, CHVD, GPD and MGV against each other. Together, GVD showed the highest representativeness, followed by CHVD, GPD and MGV.

Supplementary Figure 6 | Pie chart showing the novelty of CHGV collection. ~24% of CHGV genomes were not in any other published datasets or recognized by any built models. Identical stands for ≥ 95% ANI, partially stands for (E-value<1e-10 and coverage>50%).

Supplementary Figure 7 | **A,** ViromeQC scores of CHGV vMDA samples. CHGV vMDA samples contained much less bacterial contaminations in both filtered and unfiltered datasets. **B,** The mapping rate of CHGV vMDA samples to different datasets. **C,** The mapping rate of CHGV vMDA samples and public vMDA samples to filtered/unfiltered (length ≥ 5k) CHGV and MGV.

Supplementary Figure 8 | crAssphages and Gubaphages' diversity ranking and variation under different cutoff. Gubaphage reached the highest rank of CL diversity under the cutoff of 70, a relatively strict cutoff.

Supplementary Figure 9 | Shared PC analysis of public crAssphages. There are PC similarities between different VCs, but still, there are huge difference.

Supplementary Figure 10 | **A,** Heat map of Bray-Curtis Dissimilarity of vNGS sample and mNGS samples. vNGS samples showed Bray-Curtis Dissimilarity compared with mNGS samples. **B,** Heat map of Jaccard distance of vNGS sample and mNGS samples. vNGS samples showed an overall higher Jaccard distance compared with mNGS samples. **C,** Heat map showing the number of shared VPs/VCs across vNGS samples. vNGS samples shared little VPs/VCs.

Supplementary Figure 11 | **A,** GPD contigs' lifestyle. The temperate and uncertain temperate genomes accounted of 68% of GPD genomes. **B,** VC size of genomes with different lifestyle, all 3 datasets showed a similar tendency of that virulent VCs have bigger VC size, indicating that the virulence of the VC might contribute to its diversity. **C,** CHGV software annotated contigs' lifestyle distribution. **D,** CHGV software annotated contigs' VC size, prevalence, mean abundance distribution grouped with different lifestyle. Both the VC size, prevalence and abundance increased with the increasing of Deephage scores.

Supplementary Table 1 | Collected samples' metadata.

Supplementary Table 2 | Sample quality and enrichment scores.

Supplementary Table 3 | CHGV genomes' novelty and whether they were in published datasets or recognized by ML models.

Supplementary Table 4 | VC information and software recognition rate of CHGV genomes.

Supplementary Table 5 | Basic information of CHGV genomes including VC, lifestyle, genome length and whether they could be recognized by current viral annotation pipline.

Supplementary Table 6 | Clade information and their relationship with VC.

Supplementary Table 7 | NCBI annotation of VCs and their mean sequence similarity.

Methods

Resource table

Resource	Source	Description	Identifier/URL
Deposited data			
CHGV	This paper		
Software tools and algorithms			
Trimmomatic v0.38	[52]	Trimming low-quality bases and removal of short sequencing reads	http://www.usadellab.org/cms/index.php?page=trimmomatic
pbccs v2.0.0	-	Generate Highly Accurate Single-Molecule Consensus Reads (HiFi Reads)	https://github.com/nlhepler/pbccs
Bowtie2 v2.4.2	[53]	Aligning sequencing reads to long reference sequences.	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
PhageFinder v2.1	[54]	Identify prophage regions	http://phage-finder.sourceforge.net/
Wgsim v1.13	-	Simulating sequence reads	https://github.com/lh3/wgsim
ViromeQC v1.0	[25]	Virome enrichment evaluation	https://github.com/SegataLab/viromeqc
IDBA-UD v1.1.3	[26]	NGS assembly	https://github.com/loneknightpy/idba
Flye v2.8.2	[28]	TGS assembly	https://github.com/fenderglass/Flye/

Canu v2.0	[27]	TGS assembly	https://canu.readthedocs.io/en/latest/
MetaBAT2	[31]	Binning	https://bitbucket.org/berkeleylab/metabat
OPERA-MS v0.9.0	[29]	Hybrid assembly	https://github.com/CSB5/OPERA-MS
metaSpades v3.13.1	[30]	Hybrid assembly	https://cab.spbu.ru/software/meta-spades/
CD-HIT v4.8.1	[32]	Dereplication of genomes	https://www.bioinformatics.org/cd-hit/
Blast+v2.7.1	[36]	Sequence alignment	https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
Bedtools v2.29.1	[55]	Genomic interval files (e.g., BED, VCF, BAM) processing	https://bedtools.readthedocs.io/en/latest/index.html
MMseqs2 r13-45111	[56]	Sequence search and clustering	https://github.com/soedinglab/MMseqs2
eggNOG mapper v1.0.3-3	[57]	Genome annotation	http://eggnog-mapper.embl.de/
hmmscan v3.3.2	[58]	Biosequence analysis	http://hmmer.org/download.html
Prodigal v2.6.3	[59]	Gene prediction	https://github.com/hyattpd/Prodigal
VirSorter v2.0	[24]	Viral genome detection	https://github.com/jiarong/VirSorter2
VirFinder v1.1	[23]	Viral genome detection	https://github.com/jessieren/VirFinder
PPR-Meta v1.1	[22]	Viral genome detection	https://github.com/zhenchengfang/PPR-Meta
MCL v14-137	[39]	Clustering	http://www.micans.org/mcl/

MUSCLE v3.8.1551	[60]	Multiple sequence alignment	http://www.drive5.com/muscle/
FastTree v2.1.10	[61]	Phylogenetic tree building	http://www.microbesonline.org/fasttree/
iTol v6	[62]	Phylogenetic tree display	https://itol.embl.de/
EvolView v3	[63]	Phylogenetic tree display	https://www.evolgenius.info/evolview/
R v4.0.4	-	Statistical computing and graphics	https://www.r-project.org/
DeePhage v1.0	[41]	Viral lifestyle prediction	https://github.com/shufangwu/DeePhage
Piler-CR v1.06	[64]	CRISPR prediction	http://www.drive5.com/pilercr/
CRISPRCasFinder v4.2.20	[65]	CRISPR prediction	https://github.com/dcouvin/CRISPRCasFinder
GTDB-TK v1.4.1	[66]	Genome annotation	https://github.com/Ecogenomics/GtdbTk
tRNAscan-SE v2.0.9	[67]	tRNA prediction	http://lowelab.ucsc.edu/tRNAscan-SE/
Cytoscape v3.8.2	[68]	Network visualization	https://cytoscape.org/

Public databases

UHGG	[21]	Human gut microbial genomes; contains 4,644 bacterial genomes; downloaded on Jan 1, 2021	http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/human-gut/v1.0/uhgg_catalogue/
Pfam v34.0	[69]	A collection of protein families	http://pfam.xfam.org/

VOGdb v204	-	Virus Orthologous Groups	http://vogdb.org/
NCBI POG 2013	[70]	Virus Orthologous Groups	https://ftp.ncbi.nlm.nih.gov/pub/kristensen/thousandgenomespogs
NCBI Viral RefSeq v201	[71]	Reference viral genomes	https://www.ncbi.nlm.nih.gov/refseq/
GPD	[4]	The gut phage database; 142,000 viral genomes; February 2021	http://ftp.ebi.ac.uk/pub/databases/metagenomics/genome_sets/gut_phage_database
IMG/VR	[35]	A database gathering viral sequences identified from genomes and metagenomes; 18,373 cultivated and 2,314,329 uncultivated viral genomes; version 3	https://img.jgi.doe.gov/cgi-bin/vr/main.cgi
GVD	[3]	The gut virome database; 33,242 viral genomes, November 2020	https://de.cyverse.org/dl/d/E83EFBFF-2A23-4794-8819-ADD34160D018/FINAL_Gut_Viral_Database_GVD_1.7.2018.fna
CHVD	[6]	The Cenote Human Virome Database; 45,033 viral taxa; version 1	https://zenodo.org/record/4498884
MGV	[5]	The Metagenomic Gut Virus collection; 54,118 candidate viral species; July 2021	https://portal.nersc.gov/MGV
NCBI prokaryotic RefSeq database	[71]	Reference genomes	https://www.ncbi.nlm.nih.gov/refseq/
CRISPR Spacers Database	[72]	CRISPR Spacers Database	http://crispr.genome.ulaval.ca/

Sample collection

Human fecal samples were obtained from healthy volunteers recruited in Wuhan and Shanghai, China. All volunteers remained anonymous but were asked to complete a questionnaire to collect relevant information such as gender, age, height, weight, health statuses, recent antibiotic usage, and so on (Table S1). Exclusion criteria included: (1) those who used antibiotics or probiotic supplements up to one month before the study; (2) those who used drugs that are known to significantly affect gut microbiota compositions such as metformin [73, 74], statin [75] or proton-pump inhibitors [76, 77] in a month prior to the sample collection; (3) those who have chronic intestinal diseases or had a history of intestinal diseases; (4) females who are in menstruation. After collection, samples were immediately cooled with dry ice and transferred to a −80°C freezer within five hours. To obtain a large amount of feces for phage extraction, up to three stool samples were collected from each participant and mixed; those with totaling of 500 grams and more were processed further. In total 163 qualified samples were obtained (Table S1). This study was approved by the Ethics Committee of the Tongji Medical College of Huazhong University of Science and Technology (No, S1241) and the Human Ethics Committee of the School of Life Sciences of Fudan University (No, BE1940).

Virome enrichment, next-generation and third-generation sequencing

A virome enrichment protocol from the fecal sample was adapted from ref. [34] with modifications; see Fig. S2 for an overview. Briefly, 400~500g frozen feces taken from -80°C freezer was added with five liters of SM (200mM NaCl, 10mM MgSO4, 50mM Tris-Hcl (PH 7.5)) buffer and stirred by an automated stirrer (A200plus, OuHor, Shanghai, China) at low speed (120 rpm) at room temperature until all feces were dispersed. Then the suspended mixture was filtered by four layers of gauze (21sx32s/28x28) and centrifuged at 5000g for 45min at 4℃. The supernatant was transferred to fresh tubes and centrifuged at 8000g for 45min at 4 ℃. The supernatant was subsequently concentrated to ~300 ml by a 100 KD ultrafiltration membrane (Sartorius, VIVO FLOW 200). NaCl was then added to the filtrates to a final concentration of 0.5 mol/L and stored at 4℃ for an hour, then PEG 8000 was

added to a final concentration of 10% w/v and then incubated at 4 °C overnight. On the following day, phage particles were sedimented at 13000g for 35 min at 4 ℃.

The obtained pellets were then fully suspended by 18~36 mL TE buffer and then treated by gently shaking with an equal volume of chloroform. The mixture was centrifuged at 3500g for 10min at 4 ℃. The aqueous phase was then transferred to a sterile round-bottomed flask and evaporated for 15min using a rotary evaporator at room temperature to remove traces of chloroform which could affect the activity of DNase I in the subsequent step. The aqueous phase was transferred to a new centrifuge tube and added TE buffer to recover the volume before being treated by chloroform and then added DNase buffer to 1 × final concentration. Then for every 6 mL supernatant, 50 μL DNase I mixture (33.3U/μL, Biolab) and 25 μL RNase A mixture (0.5 U/μL, Biolab) were added and incubated in a thermostatic oscillator (THZ-C, Peiying, Suzhou, China) at 100 rpm for 30min at 37 ℃ before the enzymes were inactivated by EDTA buffer (final concentration 35 mM) and incubated at 70 °C for 10 min.

Then the nucleic acid extraction was performed by using HiPure HP DNA Maxi Kit (D6322, Magen, Guangzhou, China) according to the manufacturer's instructions. Briefly, proteinase K and SDS lysis buffer were added and the mixture was then incubated at 56℃ for an hour. Viral particles were further lysed by a CFL buffer provided by the kit and the lysates were subsequently treated with an equal volume of Phenol:Chloroform:Isoamyl alcohol (25:24:1, PH=8.0), followed by centrifugation at 12000g for 15min at room temperature. After centrifugation, the supernatant was transferred into a new centrifuge tube and treated with an equal volume of chloroform by gently shaking followed by centrifugation at 12000 g for 15 min at room temperature. The aqueous phase was transferred to a new tube, loaded to a DNA Mini Column provided by the kit, and centrifuged at 12000g for 1min. The DNA Mini Column was then washed by buffer GDP and GW2. DNA was eluted using DNA elution buffer and stored at -80℃ for further analysis. Note all buffers and columns used in this part are provided by the kit.

The purified virus-like particle (VLP) DNAs were quality-checked and subsequently sequenced by the Illumina and PacBio platforms. For Illumina sequencing, nucleic acids were sheared with g-TUBE (Covaris, USA) to generate a target size fragment of 400bp, followed by sequencing library construction using the Nextera XT DNA Library Preparation Kit (Cat. No. FC-131-1096, Illumina, USA) according to the manufacturer's instructions and sequencing using the Illumina HiSeq2000 sequencer (Novogen, Beijing, China) with pair-ended reads of 150 bases (bps). The generated dataset was then referred to as vNGS (viral next-generation sequencing) data. For PacBio sequencing, DNAs were sheared into about 5 kb by g-TUBE (Covaris, USA) and purified by AMPure PB magnetic beads, followed by a quality check using 0.7% agarose gel electrophoresis. The qualified samples were used to construct sequencing libraries using SMRTbellTM Express Template Prep Kit 2.0 (Pacific Biosciences, USA) according to the manufacturer's instructions. The quality of DNA libraries was checked by Agilent 2100 Bioanalyzer (Agilent Technologies, USA) and then sequenced by the PacBio RS II sequencer (Pacific Biosciences, Menlo Park, CA, USA) with Circular Consensus Sequencing (CCS) mode. The generated dataset was then referred to as vTGS (viral third-generation sequencing) data.

Virome-enrichment followed by multiple displacement amplification (MDA) and next-generation sequencing

To compare the recovery efficiency of our virome-enrichment protocol with that of widely used in the literature, we randomly selected 20 samples and submitted them to a protocol previously used by Zuo *et al* [15]. Briefly, 200mg of feces was taken from each sample and suspended in SM buffer by vortexing for 10 min at room temperature and centrifuged at 5000g for 15 min at4℃. The supernatant was then passed through a 0.45 μm filter followed by a 0.22 μm filter to remove bacteria and debris. Lysozyme was then added to the filtrates at the final concentration of 1mg/mL and the mixture was incubated at 100 rpm for 30min at 37℃, followed by treatment of 0.2 volume chloroform at room temperature for 10 min. Afterward, the chloroform was removed by centrifugation at 3500g for 10min at 4℃, and the supernatant was transferred to a new tube. The tube was

left stand on ice without a lid for 10 min to remove traces of chloroform. An enzyme cocktail containing 2μL DNase (5U/μL), 1μL Baseline Zero DNase (1U/μL) (Epicenter), 1μL RNase A (5U) was added to the sample to remove cell-free nucleic acids. The enzymes were inactivated by adding a 10 x stop solution (30 mM EDTA.) and heated for 10 min at 65℃. Then the sample was treated with SDS (final concentration 0.5%, Solarbio) and proteinase K (final concentration 20 μg/ml , Merck) at 56 ℃ for an hour. 1/4 volume of 10% CTAB and 1/10 volume of 5M NaCl were added to the sample and incubated at 65℃ for 10min. Then the nucleic acids were extracted by an equal volume of Phenol:Chloroform:Isoamyl alcohol (25:24:1, PH=8.0) and the residual phenol was removed by an equal volume of Chloroform. Finally, the supernatant was subjected to purification by DNA Clean & ConcentratorTM Kit (Zymo Research). Purified DNA was amplified using the REPLI-g whole genome amplification (WGA; a well-established multiple displacement amplification (MDA) method) kit (QIAGEN) and stored at -80°C for sequencing purposes.

Library construction and sequencing were then performed using a similar procedure to the vNGS dataset. The insert size for the libraries was 400bp; the same sequencer model (Illumina HiSeq2000, Novogen, Beijing, China) was used with pair-ended reads of 150 bases (bps). The resulting dataset was referred to as vMDA (viral multiple displacement amplification) data.

Fecal Bacterial DNA Extraction and DNA Sequencing

All fecal samples were also submitted to regular (no virome enrichment) metagenomic next-generation sequencing (mNGS). Briefly, total genomic DNA was extracted from 150mg of feces by using Stool DNA Kit (Omega, D4015) according to the manufacturer's instructions. The purified genomic DNAs were quality - checked by Agarose gel electrophoresis and Qubit Fluorometer (Life Technologies Qubit 2.0). Qualified DNAs were sheared with g-TUBE (Covaris, USA) to generate a target size fragment of 400bp. Sequencing libraries were then generated using the MGIEasy Universal DNA Library Prep Kit (MGI, Shenzhen, China), according to the manufacturer's instructions, and then

sequenced by the DNBSEQ-T7 sequencer (MGI, Shenzhen, China) with pair-ended reads of 150 bases (bps). This dataset was referred to as mNGS (metagenomic next-generation sequencing) in this study.

Raw data processing

Raw NGS reads were processed by Trimmomatic v0.38 [52] (with parameter LEADING:3 TRAILING:3 SLIDINGWINDOW:15:30 MINLEN:50) to remove adaptors and trim low-quality bases; reads with 50bps or less after trimming were discarded. TGS reads were corrected with Circular Consensus Sequencing (CCS) using pbccs (v4.0.0, https://github.com/nlhepler/pbccs) with default parameters.

Putative human reads were identified from the trimmed/CCSed samples by aligning the reads to the human reference genome (hg38; GCA_000001405.15) using Bowtie2 [53] (v2.4.2, --end-to-end) and removed from further analysis.

In total, we obtained 4.89 Terabyte clean data for the vNGS samples, 561 Gigabyte CCSed data for the vTGS samples, and 5.10 Gigabyte clean data for the mNGS samples.

Evaluation of bacterial contamination and removal of bacterial reads from virome sequencing datasets

To evaluate bacterial contaminations, Bowtie2 [53] (v2.4.2) was used with default parameters to map the clean reads from the vNGS, vTGS, and vMDA datasets to the UHGG [21] (Unified Human Gastrointestinal Genome) genomes. To prevent over-estimation of the contamination, possible prophage regions were identified using PhageFinder [54] (v2.1) and removed from UHGG genomes. The resulting UHGG dataset was referred to as UHGG-Minus in this study. The contamination rate was then calculated for each sample as the percentage of reads (read pairs for the vNGS and vMDA data, and CCS reads for the vTGS data) aligned to the UHGG-Minus genomes. The mapped reads were removed from further analyses to remove putative bacterial contaminations. The clean data before and after bacterial reads removal were referred to as “unfiltered” and “filtered” respectively in this study. The “filtered” data were used unless otherwise stated.

To estimate the mapping rates of sequencing reads from known human gut viruses to the UHGG-Minus genomes, a wgsim tool (v1.13, https://github.com/lh3/wgsim) was used with default parameters to generate PE150 (pair-ended, 150 bps) reads from four recently published human (gut) virome genome databases, GVD (the Gut Virome Database) [3], GPD (the Gut Phage Database) [4], CHVD (the Cenote Human Virome Database) [6] and MGV (the Metagenomic Gut Virus catalog) [5]; ~10 million paired reads, totaling ~800 million bases were generated from each dataset. These reads were also mapped to the UHGG-Minus genomes; the results were used to calculate the mapping rate for each of the simulated datasets.

Virome enrichment evaluation and exclusion of samples with low enrichment scores

To evaluate the ability of our experiment methods in virome enrichment, an FEV (fold-enrichment for virome) score was calculated for each sample, which was defined as the ratio between the percentage of viral reads in the enriched data (i.e., vNGS) and the percentage of viral reads in the corresponding unenriched data (i.e., mNGS). Here the "viral reads" were defined as those that could be mapped to software-annotated (SA for short; see below) viral contigs; thus, the percentage of "viral reads" in a sample referred to as the percentage of SA-viral reads out of total clean reads.

To identify putative viral reads, each vNGS and its corresponding mNGS sequencing datasets were assembled using methods mentioned in the next section; viral contigs were then identified using methods mentioned in the section "Annotation of CHGV contigs as viral using state-of-art software tools" below. Then Bowtie2 [53] (v2.4.2) was used with default parameters to align the clean reads in the vNGS and mNGS dataset to the viral contigs, and the mapping rate was calculated for each sample as the percentage of aligned reads out of total reads in a sample (referred to as SA-mapping rate).

For comparison, a viral enrichment evaluation tool, ViromeQC v1.0 [25], was also used on the vNGS data. ViromeQC is a state-of-art scoring system based on rRNA large and small subunit abundances and single-copy markers in virome-enriched samples versus pre-computed results in previous studies.

Samples with FEPscores lower than five and has SA-mapping rate lower than 20% were removed from further analysis (Table S2).

Generation of 279,133 unique viral populations (contigs) through a combined assembly of short- and long-reads

An overview of our assembly strategy can be found in Fig. S2; all assemblies were carried out on a per-sample basis. Briefly, IDBA-UD [26] (Release 1.1.3, parameters: --maxk 120 --step 10 –min_contig 1000) was used to assemble the filtered vNGS data. Canu [27] (v2.0-, parameters: genomeSize=20k corOutCoverage=1 -corrected) and Flye [28] (v2.8.2, parameters: --meta --genome-size 20k --min-overlap 1000) was used to assemble the filtered vTGS CCS reads. Because Canu does not have meta-assembly mode and tended to extend contigs by merging DNA sequences from different viral species to generate erroneous contigs, unitigs were used for subsequent analysis; unitigs are basic blocks of contigs that are shorter but more reliable than contigs ('unitigs' are derived from contigs. Wherever a contig end intersects the middle of another contig, the contig is split.) [78]. To further extend the sequences, MetaBAT2 [31] (version 2, default parameters) was used to group unitigs into bins. If all unitigs from one contig could be grouped into the same bin, contigs instead of the unitigs were used for further analysis. OPERA-MS [29] (v0.9.0, parameters: -contig-len-thr 1000 --polishing --no-strain-clustering --no-ref-clustering) and metaSpades [30] (v3.13.1, default parameters) were used for hybrid assemblies using both vTGS and vNGS datasets from the same samples.

Contigs/unitigs from all the above three strategies were merged; for samples that did not have vTGS data, contigs from the IDBA-UD assembler were used. Then the merged dataset was dereplicated using CD-HIT [32] (v4.8.1, parameters: -c 0.95 -n 8) using a global identity threshold of 95%. At the end, 279,133 unique putative viral contigs with lengths >=1.5 kilobases (kbs) were generated. Among which 97,253 (34.84%), 54,147 (19.39%) and 28,412 (10.18%) were longer than five, ten and twenty kbs respectively (Table S1, Fig. S3D,E). These viral contigs were also referred to as viral populations (VPs).

Rarefaction curves were generated by randomly re-sampling the pool of N samples 10 times and then plotting the number of VPs found on each set of samples.

Identification of circular viral populations

A customized pipeline was used to identify circular contigs. Briefly, the BLASTn program [36] was used to search for alignable regions within each contig; if the front and the tail parts of the contig are exact match over 30 basepairs (nucleotide identity=100, *E*-value<1e-5), it will be considered as a circular genome. A total of 4,003 contigs were identified as circular genomes, among which 3,596, 3,265, and 2,956 were longer than 5, 10, and 20 kbs, respectively (Table S1).

The Human Gut Virome database consists of 97,660 viral populations

As short contigs may only represent fragments of viral genomes, contigs that were longer than 5kb or circular ones longer than 1.5kb were selected for further analyses; this dataset was referred to as the Human Gut Virome (CHGV) database, which consisted of a total of 97,660 viral populations (Table S5).

Read recruitment analysis and comparisons among data sources

Read recruitment analysis was performed by aligning the sequenced reads to a set of assembled contigs/genomes such as the CHGV genomes using Bowtie2 [53] with default parameters, then a mapping rate for each sample was determined, calculated as the proportion of reads that could be aligned to the target dataset. Higher mapping rates (or recruitment rates) indicate better representation of the sequencing reads (and hence the genomes in the corresponding samples) by the assembled contigs. In this study, mapping rates were calculated for our vNGS/MDA samples against CHGV VPs as well as viral genomes from several public datasets, including GVD [3], GPD [4], CHVD [6], MGV[5] and NCBI Viral RefSeq [71]. Reads from several public VLP-sequencing projects were also obtained and analyzed [15, 37]. A two-sample Wilcoxon Rank Sum Test (also known as the Mann-Whitney test) was then used to compare the mapping rates between datasets.

Machine learning models for prediction of human gut virome and performance evaluation

To check the quality of CHGV and compare it with the above-mentioned public human virome databases (GVD [3], GPD [4], CHVD [6], DEVoC[7] and MGV[5]) and NCBI Viral RefSeq [71], we trained a series of neural network models with the same architecture as

DeepVirFinder [79] on them and their combinations. For CHGV, GVD and DEVoC, we kept all the sequences as the true positive datasets, while for GPD and MGV, we selected the longest sequence from each VC as representatives considering memory consumption and training time. The resulting two sub-datasets were referred to as GPD-rep and MGV-rep, respectively. As for CHVD, we constructed a gut-targeted dataset, CHVD-gut, by extracting sequences assigned with the intestine origin and longer than 1 kbp. We also collected 12,685 virus RefSeq genomes released before July 6, 2020, from NCBI [71]. Human gut bacterial sequences (negative samples) were collected from UHGG-Minus and those longer than 1.5kbp were kept. This collection was further partitioned into two parts randomly, 80% of which were used as the training set and the remaining 20% as the test set. All DNA sequences used for training were consecutively segmented into non-overlapping fragments of 1 kbp and encoded into numerical matrices with a one-hot encoding method. In the step of testing, we assigned each sequence the average score across all of its non-overlapping 1 kbp segments. Because the bacterial genomes yielded substantially more fragments, we down-sampled the bacterial training set to match the viral one for every model by random sampling.

We assessed the performance of the model trained on CHGV against those trained on other databases with two experiments. The performance of the models can be viewed as an indicator of the quality of each dataset because the same method was used. In the first experiment, we intended to prove that the sequences in CHGV were potential gut virome/phageome by comparing the performance of the models on an independent dataset, which consisted of viral sequences marked as the human intestinal origin from the IMG/VR database [35] (https://img.jgi.doe.gov/cgi-bin/vr/main.cgi) and bacterial sequences from the test set of UHGG-Minus. To ensure independence between the test set and the multiple training sets, we conducted pairwise comparisons by blasting IMG/VR human gut viral sequences against the union of CHGV, GVD, DEVoC, GPD-rep, CHVD-gut, MGV-rep, and virus RefSeq genomes and kept hits with *E*-value ≤ 1e-5. Then for every sequence in IMG/VR, we calculated the coverage by merging the aligned region that was≥ 500 bp and shared at least 90% nucleotide identity with bedtools v2.29.1 [55]. We only kept the sequences with coverage < 70% (n=1,720). For the bacterial genomes, we removed the sequences in the

test set covered by the training set following the same procedure as above. 1,720 sequences were then randomly sampled from the filtered test set to match the size of the viral one. Metrics such as accuracy, precision, recall, f1-score, true negative rate, and AUC (Area Under the receiver operating characteristic Curve) score were calculated to perform comprehensive evaluations among various models. In the second experiment, we managed to show that CHGV covers more diverse and novel viral sequences than any of the public human gut virome databases in a cross-dataset way, where the model trained on one dataset (e.g., CHGV) will be tested on another dataset (e.g., GPD). In the scenario where both the test set and training set were from the identical database, we used 80% of the sequences for training and the remaining 20% for testing, otherwise, we tested on all the sequences of each database.

We additionally constructed three new datasets, namely CHGV+GPD, CHGV+CHVD, and CHGV+MGV, CHGV+DEVoC by merging corresponding datasets, respectively. And when the test set and training set overlapped, we similarly combined the 80% sequences sampled from the overlapping dataset in the previous step with another corresponding dataset and left the remaining 20% for testing. Take CHGV+GPD as an example, we merged 80% of the sequences in CHGV and all the sequences in GPD-rep when testing on CHGV, all the sequences in CHGV and 80% of the sequences in GPD-rep when testing on GPD, and all the sequences in both datasets when testing on the other 3 databases. Finally, the nine newly generated datasets were dereplicated twice using MMseqs2 r13-45111 [56] easy-cluster module with the following parameters: '--min-seq-id 0.9 -c 0.9 --cov-mode 1 --cluster-mode 2 --max-seq-len 1100000'.

Functional annotation of CHGV proteins

The encoded protein sequences of the CHGV genomes were annotated using Prodigal [59] v2.6.3 with default parameters.

Proteins translated from the CDS sequences were then annotated with eggNOG mapper v1.0.3-3 [57] and hmmscan [58] v3.3.2 against Pfam [69] v34.0, VOGdb v204 (*E* value <1e-5, score >=50, http://vogdb.org/).

Eleven classes of phage parts were categorized based on their functions, including LYS (lysis), INT (integration), REP (replication), REG (regulation), PAC (packaging), ASB (assembly), INF (infection), EVA (immune evasion), HYP (hypothetical protein), UNS (unsorted), and tRNA according to a previous study [80].

Prediction of CHGV viral contigs with state-of-art tools

To check if contigs in our Human Gut Virome (CHGV) dataset could be correctly recognized by state-of-art software tools, the following tools were used, including VirSorter v2.0 [24] (--min-score 0.7), VirFinder v1.1 [23] (default parameters), PPR-Meta v1.1 [22] (default parameters). BLAST against the Viral RefSeq genomes was also performed using BLASTn v.2.7.1 [36] with default parameters and an *E*-value cutoff of <1e-10; the Release 201(Jul 06, 2020) of Viral RefSeq contained 13,148 viral genomes. In addition, the annotated protein sequences were used to BLAST against the NCBI POG (Phage Orthologous Groups) database 2013 [70].

A contig was annotated as a virus if it was circular or met at least two of the following criteria; the same criteria have been adopted by the Gut Virome Database (GVD) [3]:

- VirSorter score >= 0.7
- VirFinder score > 0.6
- PPR-Meta phage score > 0.7
- Hits to Viral RefSeq with > 50% identity & > 90% coverage
- Have a minimum of three ORFs, producing BLAST-hits to NCBI POG database 2013 with *E*-value of <= 1e-5, with at least two per 10 kb of contig length

In total, 19,195 of the CHGV contigs were annotated as viral, accounting for 19.65% of all contigs (Table S5).

Clustering viral populations into viral clusters (VCs)

Clustering gut viral contigs into viral clusters (VCs) was performed using a strategy adopted from GPD, the Gut Phage Database [4]. For comparison, gut viral genomes/contigs from several recent publications, including GVD [3, 14], GPD [4] and MGV [5] were also downloaded; the three datasets included 33,242, 143,882 and 189,680 genomes/contigs respectively. Briefly, all four datasets were merged; a BLASTn algorithm with default parameters was used to search the merged dataset against itself for homologous sequences. An *E*-value threshold of 1E-10 was first used to filter the BLASTn results; the BLASTn query-hit pairs were further filtered to keep those with a coverage > 70% on bigger genomes and coverage >90% on smaller genomes. Here the coverage was calculated by merging the aligned fraction length of sequence that shared at least 90% sequence similarity. Finally, a Markov Clustering Algorithm [39] (MCL v14-137) was used with an inflation value of 4.0, which took the filtered BLASTn results as input, carried out a graph-based clustering and clustered the viral contigs into VCs.

Identification of crAssphage and Gubaphage in CHGV VPs

crAss-like phage VPs were annotated by following the method reported in a previous study (ref. [81]). First, the nucleotide sequences of all CHGV VPs were BLASTed against the protein sequence of the polymerase (UGP_018) and the terminase (UGP_092) of the prototypical crAssphage (p-crAssphage, NC_024711.1) using BLASTx. Second, the nucleotide sequence similarities between the CHGV VPs and the p-crAssphage genome were assessed using BLASTn. A VP was then labeled as a putative crAssphage when it was longer than 70kb and met at least one of the following criteria:

- had a BLASTx hit with an *E*-value <1e-10 against either p-crAssphage polymerase or terminase
- showed ≥95% nucleotide identity over 80% of the contig length with the p-crAssphage genome

Gubaphage VPs were annotated by clustering viral populations with the Gubaphage genomes obtained from the GPD database [4] into viral clusters. Viral populations that were in the same VC with Gubaphage were annotated as Gubaphage. The clustering of viral clusters is further explained in the section below.

In addition, crAssphage and Gubaphage VPs can also be identified using phylogenetic analysis of the terminase protein sequences; see the section below for details.

Phylogenetic analysis of phages of interest

Phylogenetic analysis was performed for selected phages using the terminase protein sequences. Viral populations with similar terminase were considered as the same viral clade. Briefly, for each group of phages of interest, their terminase protein sequences were aligned using MUSCLE [60] v3.8.1551 with default parameters. Phylogenetic trees were built with FastTree [61] v2.1.10 with default parameters. Phylogenetic trees were then visualized and annotated using iTol [62] and EvolView [63].

Clustering viral clusters into viral clades (CLs)

Clustering viral clusters into viral clades was performed based on protein sequence similarities of the terminase genes. Similar strategies have been used to cluster Gubaphages and crAssphages [4]. However, after an all-against-all BLASTp comparison of the terminase protein sequences of the crAssphages and Gubaphages, we found that the terminase from the same viral clade has a rather low protein identity (~50%-75%). Thus, we tested different protein identity cutoffs of the terminases to find the best threshold for the clustering. For each of the BLASTp identity cutoff from 0 to 90 (step size 10), viral clades were built by feeding the MCL with the terminases having BLASTp hit length >=300 amino acids and *E*-value < 1e-10 (Fig. S8); an inflation value of 1.5 was used by MCL, which took the filtered BLASTp result as input, and 1 as the edge weight, carried out a graph-based clustering and clustered the viral contigs into viral clades. In the end, the protein identity of 70% was chosen because it could cluster the highest numbers of crAssphage and Gubaphage VPs into their own clades that contained the least numbers of other VPs. For VCs with at least one VP that encodes the similar terminase, these VCs will be considered to

be in the same clade. If a VC contains two types of terminases, the VC will appear in both CLs. In total, 762 CLs were obtained, consisting of in total 7,202 VCs; the rest of VCs were classified into "Unclassified clade" (Table S6).

Estimation of relative abundances of human gut phages at VP and VC levels

To estimate the abundance of viral populations, the vNGS clean reads were mapped to the CHGV database using Bowtie2. Then we calculated RPKM (Reads Per Kilobase Million) of each viral population. Relative species abundance is calculated by dividing the RPKM of a specific viral population by the total RPKM of all viral populations.

VC abundance was generated through viral populations' abundance by dividing the sum of RPKM of the viral population from the same viral cluster by the total RPKM.

To avoid the noise of low abundant taxa, VPs and VCs with relative abundance lower than 0.001% were considered as background noise and removed from the sample. After recalculation, VPs and VCs that did not meet a mean relative abundance cutoff of 0.001% were further removed (same remove criteria to bacterial species and genus)

Identification of the best relative abundance cutoff

To further identify the best relative abundance cutoff, we randomly selected several Gubaphage contigs, CrAssphage contigs, and nomal contigs and checked their existence by PCR. The primers were designed by primer/primer3 online software and their specificity were checked by BLASTn. Then the PCR reactions were carried out using Hieff PCR Master Mix (YEASEN) and the residual virome DNA extractions of NGS and Pacbio were used as templates for PCR. All PCR amplifications were performed for 35 cycles. And the PCR products were detected by agarose gel electrophoresis.

Enterotype analysis of human gut virome using abundance profiles of CHGV VCs

Samples were grouped into "enterotypes" using methods reported in a previous study [48]. However, due to the high between-sample diversity of the human virome, we failed to

obtain any meaningful sample groupings using distance-matrixes based on Jensen-Shannon Divergence Distance, Bray-Curtis dissimilarity, or other methods provided by the R package 'vegan' (ver 2.5), including "manhattan", "euclidean", "canberra", "clark", "bray", "kulczynski", "jaccard", "gower", "altGower", "morisita", "horn", "mountford", "raup", "binomial", "chao", "mahalanobis", "chisq" or "chord". In the end, we used the Cao index [49, 50] to calculate the similarities between samples using VC/genus-level relative abundances as input. The Cao index is often used on data with high beta diversity and variable sample intensity [49, 50]. We found that the generated groups using the Cao index had the evenest number and a higher silhouette score.

The optimal number of clusters was estimated to be 3 (k=3). The grouping results were further validated using Silhouette Coefficient (S(i)), as suggested by the previous study [48].

Prediction of viral lifestyles

The lifestyle classification of all the CHGV genomes was analyzed using DeePhage [41] v1.0 with default parameters**.** DeePhage uses a scoring system to classify viral genomes into four categories, including temperate (with scores <=0.3), uncertain temperate (0.3∼0.5), uncertain virulent (0.5∼0.7), and virulent (>0.7). Higher scores indicate higher virulence.

References

1. Ogilvie, L.A., et al., *Genome signature-based dissection of human gut metagenomes to extract subliminal viral sequences.* Nat Commun, 2013. **4**: p. 2420.
2. Lynch, S.V. and O. Pedersen, *The Human Intestinal Microbiome in Health and Disease.* N Engl J Med, 2016. **375**(24): p. 2369-2379.
3. Gregory, A.C., et al., *The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut.* Cell Host & Microbe, 2020. **28**(5): p. 724-+.
4. Camarillo-Guerrero, L.F., et al., *Massive expansion of human gut bacteriophage diversity.* Cell, 2021. **184**(4): p. 1098-+.
5. Nayfach, S., et al., *Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome.* Nature Microbiology, 2021. **6**(7): p. 960-+.
6. Tisza, M.J. and C.B. Buck, *A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases.* Proceedings of the National Academy of Sciences, 2021. **118**(23): p. e2023202118.

7. Espen, L.V., et al., *A Previously Undescribed Highly Prevalent Phage Identified in a Danish Enteric Virome Catalog.* mSystems, 2021. **6**(5): p. e00382-21.
8. Lai, S., et al., *mMGE: a database for human metagenomic extrachromosomal mobile genetic elements.* Nucleic Acids Res, 2021. **49**(D1): p. D783-D791.
9. Schmidt, T.S.B., J. Raes, and P. Bork, *The Human Gut Microbiome: From Association to Modulation.* Cell, 2018. **172**(6): p. 1198-1215.
10. Qin, J., et al., *A metagenome-wide association study of gut microbiota in type 2 diabetes.* Nature, 2012. **490**(7418): p. 55-60.
11. Roux, S., et al., *Viral dark matter and virus-host interactions resolved from publicly available microbial genomes.* Elife, 2015. **4**.
12. Yilmaz, S., M. Allgaier, and P. Hugenholtz, *Multiple displacement amplification compromises quantitative analysis of metagenomes.* Nat Methods, 2010. **7**(12): p. 943-4.
13. Moreno-Gallego, J.L., et al., *Virome Diversity Correlates with Intestinal Microbiome Diversity in Adult Monozygotic Twins.* Cell Host & Microbe, 2019. **25**(2): p. 261-272.e5.
14. Shkoporov, A.N., et al., *The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific.* Cell Host & Microbe, 2019. **26**(4): p. 527-541.e5.
15. Zuo, T., et al., *Human-Gut-DNA Virome Variations across Geography, Ethnicity, and Urbanization.* Cell Host & Microbe, 2020. **28**(5): p. 741-751.e4.
16. Gogokhia, L., et al., *Expansion of Bacteriophages Is Linked to Aggravated Intestinal Inflammation and Colitis.* Cell Host & Microbe, 2019. **25**(2): p. 285-+.
17. Clooney, A.G., et al., *Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease.* Cell Host & Microbe, 2019. **26**(6): p. 764-+.
18. Tetz, G., et al., *Type 1 Diabetes: an Association Between Autoimmunity, the Dynamics of Gut Amyloid-producing E. coli and Their Phages.* Scientific Reports, 2019. **9**.
19. Pinard, R., et al., *Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing.* BMC Genomics, 2006. **7**: p. 216.
20. Yilmaz, S. and A.K. Singh, *Single cell genome sequencing.* Curr Opin Biotechnol, 2012. **23**(3): p. 437-43.
21. Almeida, A., et al., *A unified catalog of 204,938 reference genomes from the human gut microbiome.* Nat Biotechnol, 2021. **39**(1): p. 105-114.
22. Fang, Z.C., et al., *PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning.* Gigascience, 2019. **8**(6).
23. Ren, J., et al., *VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data.* Microbiome, 2017. **5**.
24. Roux, S., et al., *VirSorter: mining viral signal from microbial genomic data.* Peerj, 2015. **3**.
25. Zolfo, M., et al., *Detecting contamination in viromes using ViromeQC.* Nature Biotechnology, 2019. **37**(12): p. 1408-1412.
26. Peng, Y., et al., *IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth.* Bioinformatics, 2012. **28**(11): p. 1420-8.
27. Koren, S., et al., *Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation.* Genome Res, 2017. **27**(5): p. 722-736.

28. Kolmogorov, M., et al., *Assembly of long, error-prone reads using repeat graphs.* Nat Biotechnol, 2019. **37**(5): p. 540-546.
29. Bertrand, D., et al., *Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes.* Nat Biotechnol, 2019. **37**(8): p. 937-944.
30. Nurk, S., et al., *metaSPAdes: a new versatile metagenomic assembler.* Genome Res, 2017. **27**(5): p. 824-834.
31. Kang, D.D., et al., *MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies.* PeerJ, 2019. **7**: p. e7359.
32. Fu, L., et al., *CD-HIT: accelerated for clustering the next-generation sequencing data.* Bioinformatics, 2012. **28**(23): p. 3150-2.
33. Shkoporov, A.N. and C. Hill, *Bacteriophages of the Human Gut: The "Known Unknown'' of the Microbiome.* Cell Host & Microbe, 2019. **25**(2): p. 195-209.
34. Shkoporov, A.N. and C. Hill, *Reproducible protocols for metagenomic analysis of human faecal phageomes.* Microbiome, 2018. **6**(1): p. 68.
35. Roux, S., et al., *IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses.* Nucleic Acids Research, 2021. **49**(D1): p. D764-D775.
36. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
37. Mangalea, M.R., et al., *Individuals at risk for rheumatoid arthritis harbor differential intestinal bacteriophage communities with distinct metabolic potential.* Cell Host & Microbe, 2021.
38. Mangalea, M.R., et al., *Individuals at risk for rheumatoid arthritis harbor differential intestinal bacteriophage communities with distinct metabolic potential.* Cell Host & Microbe, 2021. **29**(5): p. 726-+.
39. Enright, A.J., S. Van Dongen, and C.A. Ouzounis, *An efficient algorithm for large-scale detection of protein families.* Nucleic Acids Research, 2002. **30**(7): p. 1575-1584.
40. Guerin, E., et al., *Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut.* Cell Host & Microbe, 2018. **24**(5): p. 653-+.
41. Wu, S., et al., *Distinguish virulent and temperate phage-derived sequences in metavirome data with a deep learning approach.* bioRxiv, 2020: p. 2020.12.25.424404.
42. Garneau, J.R., et al., *PhageTerm: a tool for fast and accurate determination of phage termini and packaging mechanism using next-generation sequencing data.* Scientific Reports, 2017. **7**.
43. Gentry, B.G., E. Bogner, and J.C. Drach, *Targeting the terminase: An important step forward in the treatment and prophylaxis of human cytomegalovirus infections.* Antiviral Res, 2019. **161**: p. 116-124.
44. Wirbel, J., et al., *Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox.* Genome Biol, 2021. **22**(1): p. 93.
45. Jiang, P., et al., *Metagenomic Analysis of Common Intestinal Diseases Reveals Relationships among Microbial Signatures and Powers Multidisease Diagnostic Models.* mSystems, 2021. **6**(3).
46. Goodrich, J.K., et al., *Conducting a microbiome study.* Cell, 2014. **158**(2): p. 250-262.

47. Segata, N., et al., *Metagenomic microbial community profiling using unique clade-specific marker genes.* Nature Methods, 2012. **9**(8): p. 811-+.
48. Arumugam, M., et al., *Enterotypes of the human gut microbiome.* Nature, 2011. **473**(7346): p. 174-80.
49. Cao, Y., W.P. Williams, and A.W. Bark, *Similarity measure bias in river benthic Aufwuchs community analysis.* Water Environment Research, 1997. **69**(1): p. 95-106.
50. Anderson, M.J. and A.A. Thompson, *Multivariate control charts for ecological and environmental monitoring.* Ecological Applications, 2004. **14**(6): p. 1921-1935.
51. Gregory, A.C., et al., *The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut.* Cell Host Microbe, 2020. **28**(5): p. 724-740 e8.
52. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-20.
53. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.
54. Fouts, D.E., *Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences.* Nucleic Acids Res, 2006. **34**(20): p. 5839-51.
55. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-842.
56. Steinegger, M. and J. Söding, *MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets.* Nature Biotechnology, 2017. **35**(11): p. 1026-1028.
57. Huerta-Cepas, J., et al., *Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper.* Mol Biol Evol, 2017. **34**(8): p. 2115-2122.
58. Mistry, J., et al., *Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions.* Nucleic Acids Res, 2013. **41**(12): p. e121.
59. Hyatt, D., et al., *Prodigal: prokaryotic gene recognition and translation initiation site identification.* BMC Bioinformatics, 2010. **11**: p. 119.
60. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
61. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree 2--approximately maximum-likelihood trees for large alignments.* PLoS One, 2010. **5**(3): p. e9490.
62. Letunic, I. and P. Bork, *Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation.* Nucleic Acids Research, 2021. **49**(W1): p. gkab301-.
63. Subramanian, B., et al., *Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees.* Nucleic Acids Res, 2019. **47**(W1): p. W270-W275.
64. Edgar, R.C., *PILER-CR: fast and accurate identification of CRISPR repeats.* BMC Bioinformatics, 2007. **8**: p. 18.
65. Couvin, D., et al., *CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins.* Nucleic Acids Res, 2018. **46**(W1): p. W246-W251.
66. Parks, D.H., et al., *A complete domain-to-species taxonomy for Bacteria and Archaea (vol 58, pg 561, 2020).* Nature Biotechnology, 2020. **38**(9): p. 1098-1098.
67. Chan, P.P. and T.M. Lowe, *tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences.* Methods Mol Biol, 2019. **1962**: p. 1-14.

68. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.
69. Mistry, J., et al., *Pfam: The protein families database in 2021.* Nucleic Acids Res, 2021. **49**(D1): p. D412-D419.
70. Kristensen, D.M., et al., *Orthologous gene clusters and taxon signature genes for viruses of prokaryotes.* J Bacteriol, 2013. **195**(5): p. 941-50.
71. Sayers, E.W., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2021. **49**(D1): p. D10-D17.
72. Dion, M.B., et al., *Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter.* Nucleic Acids Res, 2021. **49**(6): p. 3127-3138.
73. Forslund, K., et al., *Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota.* Nature, 2015. **528**(7581): p. 262-266.
74. Wu, H., et al., *Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug.* Nature Medicine, 2017. **23**(7): p. 850-858.
75. Vieira-Silva, S., et al., *Statin therapy is associated with lower prevalence of gut microbiota dysbiosis.* Nature, 2020. **581**(7808): p. 310-315.
76. Jackson, M.A., et al., *Proton pump inhibitors alter the composition of the gut microbiota.* Gut, 2015. **65**(5): p. 749-756.
77. Wu, S., et al., *Treatment regimens may compromise gut-microbiome-derived signatures for liver cirrhosis.* Cell Metabolism, 2021. **33**(3): p. 455-456.
78. Suzuki, Y., et al., *Long-read metagenomic exploration of extrachromosomal mobile genetic elements in the human gut.* Microbiome, 2019. **7**(1).
79. Ren, J., et al., *Identifying viruses from metagenomic data using deep learning.* Quant Biol, 2020. **8**(1): p. 64-77.
80. Song, W., et al., *Prophage Hunter: an integrative hunting tool for active prophages.* Nucleic Acids Res, 2019. **47**(W1): p. W74-W80.
81. Fujimoto, K., et al., *Metagenome Data on Intestinal Phage-Bacteria Associations Aids the Development of Phage Therapy against Pathobionts.* Cell Host Microbe, 2020. **28**(3): p. 380-389 e9.

Funding

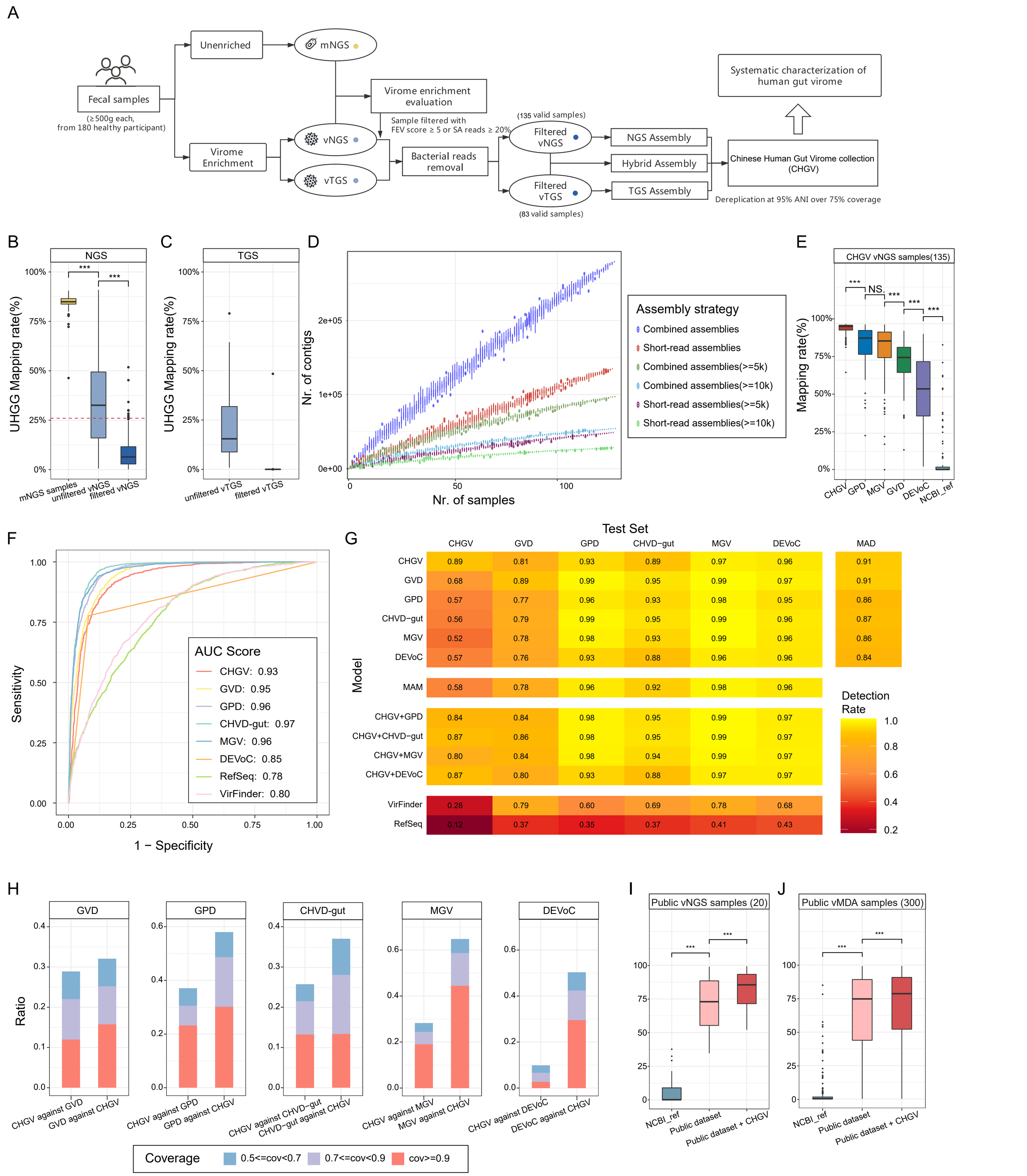
This research is supported by National Key Research and Development Program of China (2020YFA0712403 to X.M.Z;2019YFA0905600 to W.H.C and Z. L), National Natural Science Foundation of China (61932008, 61772368 to X.M.Z; 32070660 to W.H.C; 31770132, 81873969 to Z. L), and Shanghai Municipal Science and Technology Major Project (2018SHZDZX01 to X.M.Z),

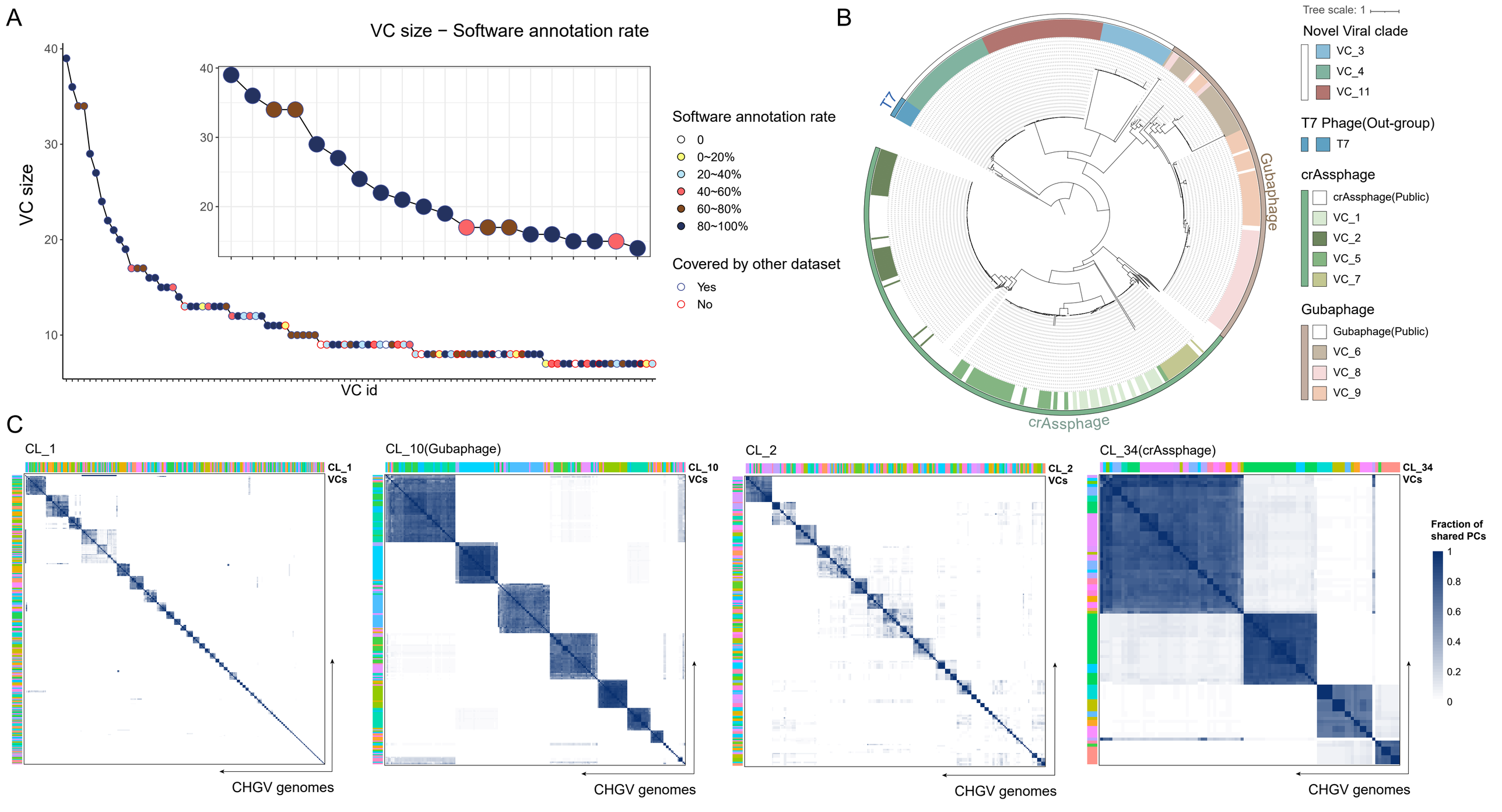
Ethics approval

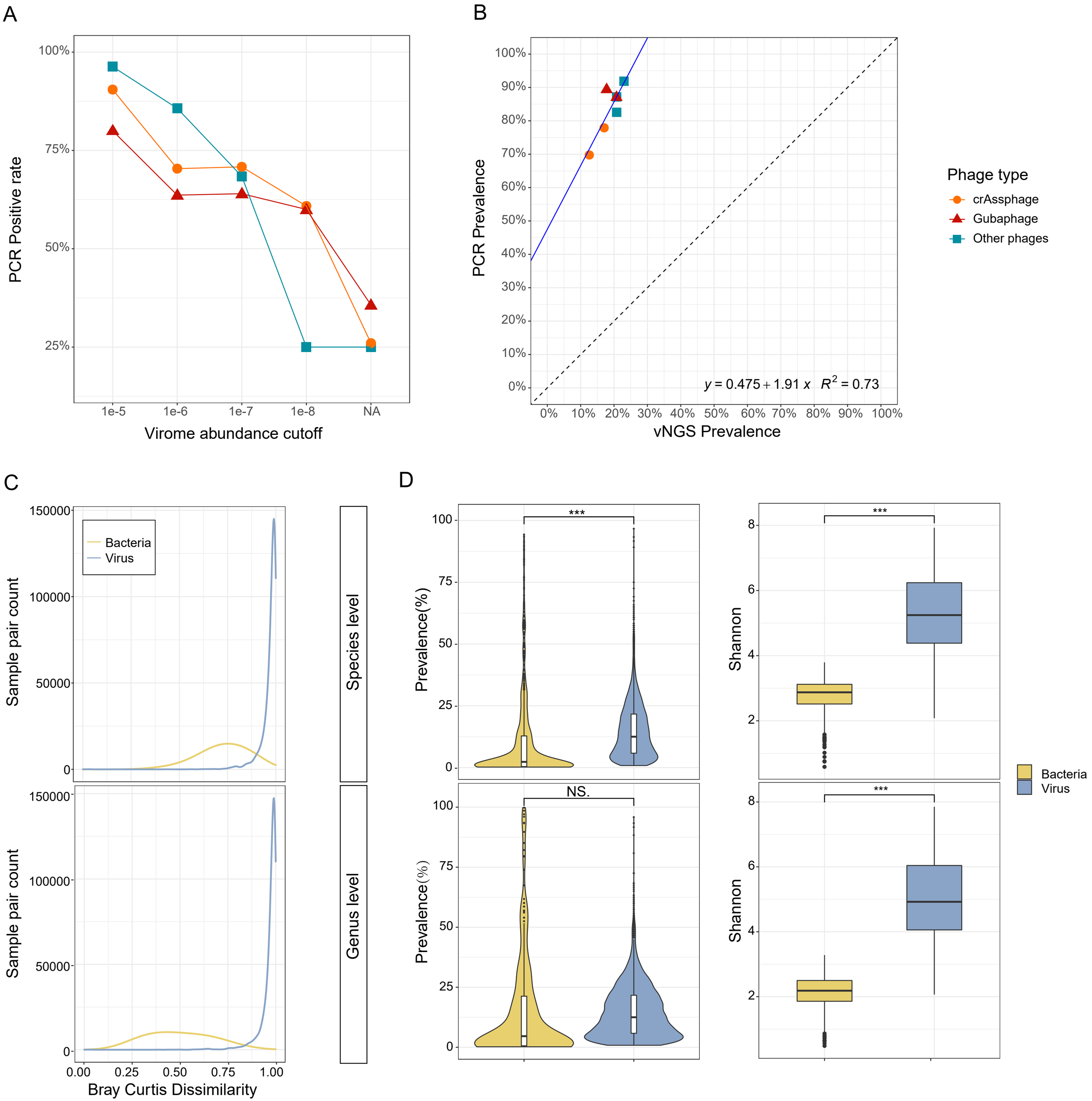
This study was approved by the Ethics Committee of the Tongji Medical College of Huazhong University of Science and Technology (No, S1241) and the Human Ethics Committee of the School of Life Sciences of Fudan University (No, BE1940).

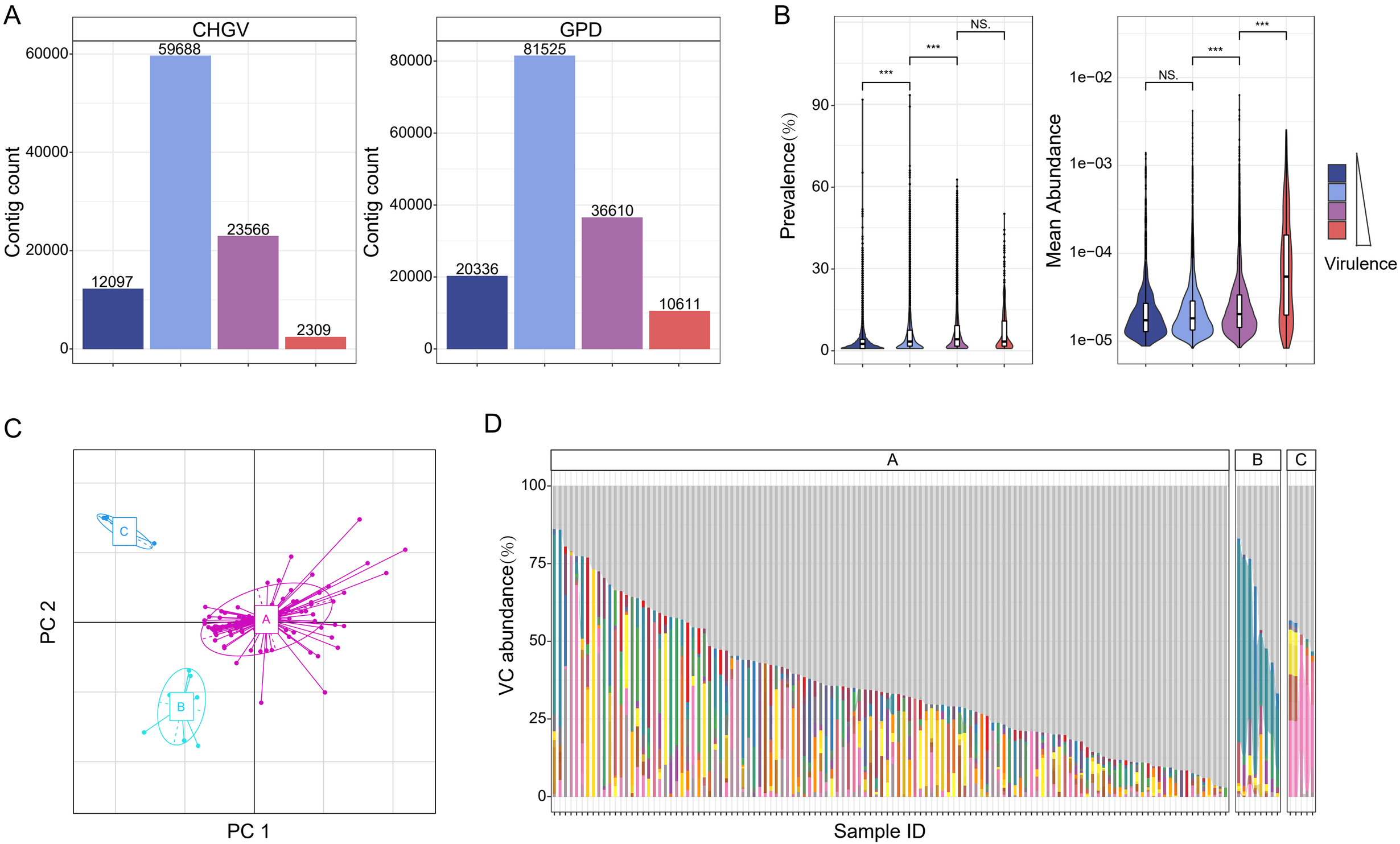
Authors contributions

WHC, XMZ, ZL and PB designed and directed the research; JC managed the sampling and did most of the experiments; CS did most of the analysis; XZ, and MJ also helped with the sample collection and phage enrichment experiments; YD did the machine learning analysis. CS and JC wrote the paper with results from all authors; WHC, XMZ, ZL and PB polished the manuscript through multiple iterations of discussions with all authors. All authors have read and approved the final manuscript.


A
Fecal samples
(≥500g each, from 180 healthy participant)
Unenriched
Virome Enrichment
mNGS
vNGS
vTGS
Virome enrichment evaluation
Sample filtered with FEV score ≥ 5 or SA reads ≥ 20%
Bacterial reads removal
(135 valid samples)
Filtered vNGS
Filtered vTGS
(83 valid samples)
NGS Assembly
Hybrid Assembly
TGS Assembly
Chinese Human Gut Virome collection (CHGV)
Dereplication at 95% ANI over 75% coverage
Systematic characterization of human gut virome
B
NGS
UHGG Mapping rate(%)
mNGS samples
unfiltered vNGS
filtered vNGS
C
TGS
unfiltered vTGS
filtered vTGS
D
Nr. of contigs
Nr. of samples
Assembly strategy
Combined assemblies
Short-read assemblies
Combined assemblies(>=5k)
Combined assemblies(>=10k)
Short-read assemblies(>=5k)
Short-read assemblies(>=10k)
E
CHGV vNGS samples(135)
Mapping rate(%)
CHGV
GPD
MGV
GVD
DEVoC
NCBI_ref
F
Sensitivity
1 − Specificity
AUC Score
CHGV: 0.93
GVD: 0.95
GPD: 0.96
CHVD-gut: 0.97
MGV: 0.96
DEVoC: 0.85
RefSeq: 0.78
VirFinder: 0.80
G
Test Set
Model
Detection Rate
H
GVD
GPD
CHVD-gut
MGV
DEVoC
Ratio
CHGV against GVD
GVD against CHGV
CHGV against GPD
GPD against CHGV
CHGV against CHVD−gut
CHVD−gut against CHGV
CHGV against MGV
MGV against CHGV
CHGV against DEVoC
DEVoC against CHGV
Coverage
0.5<=cov<0.7
0.7<=cov<0.9
cov>=0.9
I
Public vNGS samples (20)
NCBI_ref
Public dataset
Public dataset + CHGV
J
Public vMDA samples (300)


A
VC size − Software annotation rate
VC size
VC id
40
30
20
10
Software annotation rate
0
0~20%
20~40%
40~60%
60~80%
80~100%
Covered by other dataset
Yes
No
B
Tree scale: 1
Novel Viral clade
VC_3
VC_4
VC_11
T7 Phage(Out-group)
T7
crAssphage
crAssphage(Public)
VC_1
VC_2
VC_5
VC_7
Gubaphage
Gubaphage(Public)
VC_6
VC_8
VC_9
T7
Gubaphage
crAssphage
C
CL_1
CL_1 VCs
CHGV genomes
CL_10(Gubaphage)
CL_10 VCs
CHGV genomes
CL_2
CL_2 VCs
CHGV genomes
CL_34(crAssphage)
CL_34 VCs
CHGV genomes
Fraction of shared PCs
1
0.8
0.6
0.4
0.2
0


A
PCR Positive rate
100%
75%
50%
25%
1e-5
1e-6
1e-7
1e-8
NA
Virome abundance cutoff
B
PCR Prevalence
vNGS Prevalence
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
y = 0.475 + 1.91 x R2 = 0.73
Phage type
crAssphage
Gubaphage
Other phages
C
Sample pair count
150000
100000
50000
0
Bacteria
Virus
Species level
Genus level
0.00
0.25
0.50
0.75
1.00
Bray Curtis Dissimilarity
D
Prevalence(%)
Shannon
***
NS.
Bacteria
Virus


A
CHGV
GPD
Contig count
60000
40000
20000
0
80000
12097
59688
23566
2309
20336
81525
36610
10611
B
Prevalence(%)
90
60
30
0
***
NS.
Mean Abundance
1e−02
1e−03
1e−04
1e−05
Virulence
C
PC 1
PC 2
A
B
C
D
VC abundance(%)
100
75
50
25
0
Sample ID