Improving gut virome comparisons using predicted phage host information

Michael Shamash [‡ a], Anshul Sinha [‡ a], and Corinne F. Maurice [a, b *]

[‡] These authors contributed equally to this work

[a] Department of Microbiology & Immunology, McGill University, Montreal, QC, Canada

[b] McGill Centre for Microbiome Research, Montreal, QC, Canada

* Corresponding author: corinne.maurice@mcgill.ca

Competing interests statement

The authors have no conflicts of interest to disclose.

Abstract

The human gut virome is predominantly made up of bacteriophages (phages), viruses that infect bacteria. Metagenomic studies have revealed that phages in the gut are highly individual specific and dynamic. These features make it challenging to perform meaningful cross-study comparisons. While several taxonomy frameworks exist to group phages and improve these comparisons, these strategies provide little insight into the potential effects phages have on their bacterial hosts. Here, we propose the use of predicted phage host families (PHFs) as a functionally relevant, higher rank unit of phage taxonomy to improve these cross-study analyses. We first show that bioinformatic predictions of phage hosts are accurate at the host family level by measuring their concordance to Hi-C sequencing-based predictions in human and mouse fecal samples. Next, using phage host family predictions, we determined that PHFs reduce intra- and interindividual ecological distances compared to viral contigs in a previously published cohort of 10 healthy individuals, while simultaneously improving longitudinal virome stability. Lastly, by reanalyzing a previously published metagenomics dataset with > 1,000 samples, we determined that PHFs are prevalent across individuals and can aid in the detection of inflammatory bowel disease-specific virome signatures. Overall, our analyses support the use of predicted phage hosts in reducing between-sample distances and providing a biologically relevant framework for making between-sample virome comparisons.

Introduction

The human gut virome, the collection of viruses in the human gastrointestinal tract, is dominated by bacteria-infecting bacteriophages (phages). This community is highly diverse and individual-specific at the nucleotide level (1–5). This vast diversity makes it challenging to perform cross-individual or cross-cohort comparisons, as it is rare for all individuals in a cohort group to share a single viral OTU (vOTU). Most recently, a longitudinal analysis of 59 individuals further demonstrated that the individuality of the gut virome confounded disease signal detection in the context of inflammatory bowel diseases (IBDs), in part due to intraindividual fluctuations of viruses over time (6).

Viral clusters, such as those generated by vConTACT 2 (7), have been proposed as a potential solution to this, by grouping together viruses based on their shared protein content. While this approach is useful for complete genome sequences, it may not always be reliable in the context of virome datasets. Indeed, current short-read virome studies often generate several contigs for a single viral genome, raising the risk that each contig from a given virus would be placed into a different viral cluster, confounding ecological conclusions (7).

An important limitation of previous virome analyses was the inability to confidently link uncultured phages with their hosts. Recent experimental and bioinformatic advances aim to address this issue. For instance, proximity ligation sequencing is being used to assign phages to their hosts *in situ*. With this approach, phage DNA inside of host cells at the time of sampling is covalently crosslinked to the bacterial host DNA, leading to generation of chimeric reads during the sequencing process (8, 9). On the computational side, tools such as iPHoP enable the high-throughput prediction of hosts using phage sequence data alone (10). Using a combination of existing tools and machine learning models, iPHoP can consistently predict hosts down to the

genus level. These two approaches, proximity ligation and iPHoP, have yet to be formally compared with each other for assigning hosts to gut virome-derived sequences.

Here, we propose using predicted phage host range to allow for ecologically relevant comparisons of viromes across individuals, regardless of nucleotide-level diversity. These comparisons could provide broad insight on ecosystem function, as phages have the ability to alter bacterial abundances and metabolism (11, 12). We introduce the term Phage Host Family (PHF) as a term to describe the predicted bacterial host of a phage sequence, at the family level. This family-level cut-off was determined based on comparisons of predicted phage host range from iPHoP with experimental assignments via proximity ligation sequencing of human and mouse fecal samples, where high concordance was seen down to the family, but not genus level. Using this metric, we then re-evaluate two previously published large datasets. First, we apply PHF analysis to viromes from a cohort of 10 healthy individuals (1), sampled longitudinally for approximately 1 year, and conclude that incorporating PHFs reduced interindividual variation, while also increasing within-individual virome stability over time. Second, we analyze the phageome of a large cohort of individuals with IBDs (13), where we determine that aggregating vOTUs using PHFs allows for the detection of greater disease-specific differences in the virome, in addition to reducing interindividual variability. We propose that the use of PHFs as an ecologically informed unit of phage taxonomy is useful in allowing for cross-sample comparisons in gut virome studies.

Methods

Preparing fecal samples for proximity ligation sequencing

Human fecal samples were collected with the approval of protocol A04-M27-15B from the McGill University Institutional Review Board. Participants provided informed written consent for the utilization of their samples. Fresh fecal samples were collected, aliquoted in an anaerobic chamber, and kept at -70 °C until processing.

Adult female germ-free C57BL/6 mice were maintained in Tecniplast IsoCages at McGill University. Mice had unlimited access to irradiated diet (Research Diets, New Brunswick, NJ) and autoclaved water. Germ-free mice were humanized by oral gavage of 200 uL of resuspended human donor feces. Mouse fecal samples were collected with the approval of McGill University animal use protocol MCGL-7999.

A total of 10 mouse fecal samples (from 6 mice), and 2 human fecal samples, were collected for proximity ligation and bulk metagenome sequencing. Fresh fecal samples were collected and stored at -70 °C until processing. All mice had unlimited access to standard chow and water. Fecal samples were resuspended in 1 mL PBS (0.02 um filter-sterilized). After an initial centrifugation at 1,000xg for 1 minute to pellet large debris, the bacterial cell-containing supernatant was centrifuged again at 10,000xg for 10 minutes. Pelleted bacterial cells were resuspended in 1 mL of a PBS-formaldehyde solution (1% formaldehyde) and incubated at room temperature for 20 minutes to cross-link DNA. Glycine was added in excess to quench unused formaldehyde and incubated for 15 minutes at room temperature. The fixed bacterial cells were pelleted (10,000xg for 10 minutes) and washed twice with PBS. The final resuspended bacterial pellet was transferred to a BeadBug tube with 0.1mm silica glass beads (Benchmark Scientific, Sayreville, United States), and vortexed at maximum speed for 5 minutes. The sample was transferred to a DNA LoBind tube

and sent to Phase Genomics (Seattle, United States) for library preparation and sequencing using their ProxiMeta kit and analysed using the corresponding bioinformatic pipeline (14). In addition, a bulk metagenome was sequenced for each sample: immediately after resuspending the original fecal sample in PBS, 250 uL of sample was used for DNA extraction using the QIAGEN PowerFecal Pro DNA kit (QIAGEN, Hilden, Germany) following the manufacturer's instructions. All libraries (proximity ligation and bulk metagenome) were sequenced using the Illumina NovaSeq platform with 2x 150bp reads.

Comparing host predictions between iPHoP and proximity-ligation sequencing

Viral contigs identified by the ProxiMeta pipeline were used as input for iPHoP (v. 1.3.3) (10) to computationally predict hosts, the output was imported into R (v. 4.2.2) for analysis. The proximity-ligation linkage data was imported into R and filtered to keep only viral contigs who also had hosts predicted with iPHoP. These two datasets were then compared for concordance at the following taxonomic ranks: phylum, class, order, family, and genus. Concordance was calculated at each taxonomic rank using three distinct approaches to account for phages which have multiple assigned/predicted hosts: (1) consider the pairing concordant if the most confident iPHoP prediction matches the top Hi-C hit (most stringent); (2) consider the pairing concordant if the most confident iPHoP prediction matches any of the Hi-C hits; and (3) consider the pairing concordant if any of the iPHoP predictions match any of the Hi-C hits (least stringent). The percent concordance was calculated as the number of viral contigs with concordant hosts, divided by the total number of viral contigs, multiplied by 100%.

Re-analysis of Shkoporov *et al.* dataset

The published phyloseq object (1) was downloaded and imported into R (v. 4.2.2) using phyloseq (v. 1.42) (15). Published virome contigs were also downloaded and filtered to keep those > 1 kb using seqkit (v. 2.5.1), resulting in 57,721 contigs. iPHoP (v. 1.3.3) (10) was used to predict the bacterial hosts of these contigs, and the output was imported into R for analysis. The GTDB tree used by iPHoP (bac120_r202.tree) was also imported into R using phyloseq's read_tree command and combined with the downloaded phyloseq object. Viral contig relative abundance was calculated and added to the phyloseq object, replacing the existing otu_table object. Predicted host information was added to the phyloseq object as a tax_table object. In cases where a viral contig had more than 1 predicted host by iPHoP, the most confident host prediction was selected. Taxa bar plots were generated using microshades (v. 1.10) (16). Samples with less than 30% of the community consisting of contigs with unknown hosts were retained for subsequent analysis. Phyloseq's tax_glom function was used to agglomerate viral contigs that have the same predicted host at the family level. Vegan (v. 2.6-6.1) (17) was used to calculate distances between samples using the bray, wunifrac, and unifrac metrics. Distances were evaluated by comparison type, either inter- or intraindividual sample comparisons, and the Friedman test with post-hoc Wilcoxon signed-rank test (using Bonferroni correction for multiple comparisons) were used test for significance.

We define virome stability as the similarity between two sequential samples and it is calculated as follows: stability = (1 – distance from previous sample). We calculated virome stability using distances between consecutively collected samples from the same individual using Bray-Curtis distances at the contig and PHF levels and tested for significance using the Wilcoxon signed-rank test.

Re-analysis of the HMP2 dataset

In the human IBD cohort, originally analyzed by Lloyd-Price *et al.,* bulk metagenome reads were obtained from 1,595 samples belonging to 130 individuals (27 non-IBD, 65 CD, and 38 UC) sampled longitudinally over one year (13). Data was downloaded from: https://ibdmdb.org/results. Paired-end sequencing reads (101 bp) were generated using Illumina HiSeq 2000 or 2500. Raw reads were trimmed based on sequence quality using Trimmomatic (v 0.33) (18). Quality-controlled sequences that aligned to human and mouse genomes were removed by Bowtie2 (v. 2.2) (19). These steps were performed using the kneaddata workflow (20). Quality-controlled reads were then grouped by individual and co-assembled into 3,249,501 contigs > 1kb using MEGAHIT (v. 1.2.9) (21). The contigs within each co-assembly were classified as phages by VIBRANT (v. 1.2.1) (22). In total, there were 81,422 predicted phages across all co-assemblies. These contigs were then filtered for completeness using CheckV (v. 1.0.3) (23), keeping only the 6,741 contigs that were over 50% complete. A 50% completeness cutoff was used to balance the trade-offs of overestimating viral richness due to fragmentation during assembly and maintaining viral richness. These remaining contigs were then dereplicated using blastn, keeping contigs with an average nucleotide identity of 95% over 80% alignment fraction relative to the shorter sequence (24). Similar to the analyses of the Shkoporov *et al*. dataset, iPHoP (v. 1.3.3) (10) was used to predict the bacterial host of each phage contig, keeping the most confident host prediction if there were multiple predictions. Quality-controlled reads were mapped to the phage contigs library using Bowtie2 (19). Contigs were considered present in a given sample using mapping thresholds defined by Stockdale *et al*. (6), where a contig was present if Bowtie2 mapped reads covered 50% of contigs <5 kb, 30% of contigs ≥5 kb and <20 kb, or 10% of contigs ≥20 kb. After calculating Good's coverage and generating rarefaction curves for each sample, 502 samples were removed

which had below 1,500 length-normalized read counts (25). PCoAs and Bray-Curtis distances on the remaining 1,093 samples were generated using MicroViz (v. 0.12.1) (26), which uses Vegan as a wrapper. DESeq2 (v. 1.44) (27) was used to calculate differentially abundant PHFs based on dysbiosis status, using the simple formula: design = ~ Participant.ID + dysbiosis_binary. PHFs with an adjusted p value ≤ 0.05 and with a $\log_2$ fold-change ≥ 1 or with a $\log_2$ fold-change ≤ -1 were considered differentially abundant. For differential abundance analyses, only individuals which had both a dysbiotic and non-dysbiotic sample were included so that a paired analysis could be conducted. Only the 18 PHFs that were more than 50% prevalent across individuals were considered for analysed PHFs. This arbitrary threshold was used to consider only the features that were widely distributed and abundant across samples. Auxiliary metabolic genes (AMGs) were predicted from viral contigs using VIBRANT. Using KEGG annotations, VIBRANT categorizes these AMGs into metabolic categories (22). In some cases, a single AMG belonged to multiple metabolic categories.

Code and data availability

Code used for data analysis is available at: https://github.com/mshamash/PHF_manuscript. Whole genome and Hi-C sequencing reads are available on the NCBI SRA using accession number PRJNA1145458.

Results

Computational prediction of phage hosts is concordant with proximity ligation sequencing assignments to the family level

We conducted proximity ligation (Hi-C) sequencing on 10 fecal samples from human microbiota-associated mice, and 2 fecal samples from healthy human donors. After Hi-C host assignment, we identified 1,577 phage-host pairings consisting of 1,547 phages targeting 77 unique hosts at the genus level, with some phages being linked to more than one host. Using iPHoP, we then predicted hosts for the 1,547 phages with Hi-C-assigned hosts, yielding 1,587 phage-host pairings, comprising 1,243 phages targeting 108 unique hosts at the genus level. These 1,243 phages, which had hosts assigned by both Hi-C and iPHoP, were used for subsequent comparisons between approaches.

Concordance between the two approaches was calculated at each taxonomic rank from phylum to genus using three distinct approaches to account for some phages having multiple assigned/predicted hosts: (1) pairing is concordant if the most confident iPHoP prediction matches the top Hi-C hit (most stringent); (2) pairing is concordant if the most confident iPHoP prediction matches any of the Hi-C hits; and (3) pairing is concordant if any of the iPHoP predictions match any of the Hi-C hits (least stringent). Regardless of the comparison approach, there was consistently over 98% concordance at the phylum level, over 97% concordance at the class level, over 96% concordance at the order level, and over 92% concordance at the family level (**Figure 1**). Genus-level concordance was lower, with approximately 67% concordance using comparison metrics (1) and (2), and 73% concordance using metric (3) (**Figure 1**).

Based on these findings, we decided to use iPHoP predictions at the family-level for our subsequent analyses as this was the lowest taxonomic rank which still had high concordance

between the tool's predictions and Hi-C experimental assignments. These family-level host predictions will now be referred to as PHF.

Using predicted hosts as a functional measure of virome diversity reduces interindividual variation and increases intraindividual stability

We next wanted to use PHFs to evaluate functional virome diversity within and across individuals. Using a previously published dataset consisting of 140 total samples from 10 healthy individuals (1), we predicted hosts for the provided assembled phage contigs (n=57,721 contigs). In total, iPHoP yielded 197,994 host predictions for 49,852 (86.3%) of the viral contigs (an average of 3.97 bacterial hosts predicted per viral contig). The remaining 7,869 (13.7%) contigs had no host predicted. The most confident iPHoP prediction for each contig was retained and used in downstream analysis. Overall, there was large variation in the proportion of the virome made up of contigs with known hosts (**Figure 2A**). To ensure that the subsequent comparisons between samples are fair (i.e., by comparing samples with similar proportions of the community represented by contigs having assigned hosts), we filtered the dataset to retain only samples which had less than 30% of the community consisting of contigs with unknown hosts, resulting in a new dataset composed of 63 samples from 10 individuals. Keeping all samples, including those with a high proportion of contigs with unknown hosts, would introduce bias into our analyses by over-representing incomplete or ambiguous community structures, potentially leading to inaccurate conclusions about the relationships between samples. Using CheckV to filter for contigs >50% complete did not significantly change the proportion of the community consisting of contigs with unknown hosts (data not shown).

Phages with the same family-level host predictions were agglomerated into PHF groups and the resulting abundance matrix was used for subsequent analyses. Pairwise distances were calculated between all samples using traditional contig-level Bray-Curtis, PHF-level Bray-Curtis, and PHF-level weighted UniFrac metrics, as described in Methods. Intraindividual sample distances were consistently lower than interindividual sample distances (**Figure 2B**). Regardless of inter- or intraindividual sample comparison, PHF-level weighted UniFrac distances were significantly lower than PHF-level Bray-Curtis distances, which were themselves significantly lower than contig-level Bray-Curtis distances (**Figure 2B**). Finally, we evaluated the effects of using PHF-level distances on longitudinal intraindividual virome stability, calculated as the pairwise distance between all pairs of consecutive samples from each individual. Virome stability was consistently higher when incorporating PHF-level distances using the Bray-Curtis metric (**Figure 2C**).

PHFs are prevalent and can provide biological insight into the IBD virome

We next wanted to characterize PHFs in a larger dataset and determine whether agglomerating at the PHF-level could improve detection of disease-specific signatures. To do so, we re-analyzed the human microbiome project 2 (HMP2) dataset containing longitudinal bulk metagenome samples from IBD and non-IBD controls. After removing samples which contained low read counts (see methods), 1,093 samples from 57 CD, 31 UC, 27 non-IBD controls remained for further analyses (68.5% of total samples) (13). From these samples, we co-assembled contigs, dereplicated these contigs, predicted phages using VIBRANT, and filtered for phage completeness > 50% using CheckV. Using this approach, we obtained a total of 3,862 distinct virus operational taxonomic units (vOTUs) across the samples within the dataset. Of these 3,862 vOTUs, 87.1%

(3,365/3,862) had an iPHoP predicted host. In total, these vOTUs belonged to 75 distinct PHFs. Interestingly, the amount of vOTUs comprising each PHF varied greatly, with some PHFs comprised of hundreds of distinct vOTUs, whereas some rare PHFs were only comprised of a single vOTU (**Supplementary Figure 1**).

To further characterize PHFs, and to link their host associations with metabolic functionality, we searched for viral-encoded AMGs. These genes, which are expressed throughout the process of viral infection, are thought to provide phages with increased fitness via modulation of host metabolism (22, 28). In total, 45/75 PHFs carried at least one AMG and 12/75 PHFs carried at least 10 AMGs. In general, PHFs were enriched in amino acid metabolism, energy metabolism, and cofactor and vitamin metabolism genes (**Supplementary Figure 2A**), in line with previous surveys of AMGs in human microbiomes (22). Notably, compared to other PHFs, *Bifidobacteriaceae*-infecting phages were enriched in carbohydrate metabolism genes (**Supplementary Figure 2B**). *Enterobacteriaceae*-infecting phages on the other hand were enriched in protein folding, sorting, and degradation genes (**Supplementary Figure 2C**), and in particular *cysO*, which encodes a sulfur-carrier protein important in cysteine biosynthesis and resistance to oxidative stress (29). Fourteen distinct *Enterobacteriaceae*-infecting phage vOTUs carried *cysO* **(Supplementary Figure 2D).** Only 3/75 other PHFs (*Pasteurellaceae*, *Pseudomonadaceae*, *Burkholderiaceae*) carried *cysO* on 7 distinct contigs (**Supplementary Figure 2D**). Interestingly, all of these PHFs infect bacteria from the phylum Proteobacteria, potentially reflecting host-specific adaptation through the carriage of this AMG.

Consistent with the high levels of interindividuality at the vOTU level observed in the Shkoporov *et al.* dataset, we found that only 236/3,862 (6.11%) vOTUs were found in more than 50% of individuals in the HMP2 dataset (**Figure 3A**). In contrast, a higher proportion of PHFs

(18/75; 24%) were found in more than 50% of individuals (**Figure 3B**). Importantly, these prevalent features made up a significantly higher mean relative abundance in samples at the PHF level compared to the vOTU level (**Figure 3C**). Thus, prevalent PHFs represent a larger fraction of the total community in comparison to prevalent vOTUs. In line with these observations, intraindividual and interindividual Bray-Curtis distance between samples was significantly lower at the PHF level in comparison to the vOTU level (**Figure 3D**).

Given that PHFs reduced ecological distance between samples, we hypothesized that this would also allow for more biologically relevant comparisons between individuals, and ultimately a greater ability to detect disease-specific signatures in the human virome. We first generated PCoA plots using Bray-Curtis distance and found that the first two principal components explained more cumulative variance when agglomerating the virome at the PHF level in comparison to the vOTU level (**Figure 4A**; 39.3% vs. 11%). Importantly, the proportion of variance explained by diagnosis (non-IBD, CD, UC) was higher using PHFs than using vOTUs (**Supplementary Figure 3**; $R^2 = 0.0261$ vs. $R^2 = 0.0185$). Lloyd Price *et al*. defined dysbiotic samples within this HMP2 dataset as those with high microbiota divergence from non-IBD controls (13). Using this designation, we also found that dysbiosis status explained a higher proportion of variance using PHFs when compared to vOTUs (**Figure 4A;** $R^2 = 0.0394$ vs. $R^2 = 0.0157$). We also performed differential abundance analyses to determine whether certain PHFs were enriched or depleted depending on dysbiosis status. Including only prevalent PHFs (found in > 50% of individuals), we identified a single PHF enriched in dysbiotic samples (*Enterobacteriaceae*) and 4 significantly depleted PHFs (*CAG-74, Ruminococcaceae, Acidaminococcaceae*, *Acutalibacteraceae*) (**Figure 4B**). These observations suggest that predicted phage hosts can be used to identify certain IBD-specific virome signatures.

Discussion

In the past decade, the development of phage-specific bioinformatic tools, alongside large cohort viral metagenomic studies, has revealed key characteristics of the human gut virome. Notably, gut viromes exhibit high levels of interindividuality (1, 6) and temporal variation (6). While we can now appreciate the sheer genomic phage diversity that our collective guts harbor, it remains a challenge to understand how similar our viromes are over time and from one another. Here, we demonstrate that the use of predicted phage host families (PHFs) can improve virome comparisons between and within individuals, resulting in valuable functional information typically lost with current approaches.

Using PHFs as a unit of taxonomy in two independent published datasets, we showed that in comparison to vOTUs, intra- and inter-personal ecological distance is reduced, indicating that despite phages differing between samples at the contig/vOTU level, their functionality remains similar. These findings are reminiscent of the functional redundancy characteristic of gut bacterial communities, whereby phylogenetic differences between individuals exist despite conserved functional profiles (30, 31). The conserved functionality of both phage and bacterial communities over time likely contribute to the stability and resilience of both subsets.

The advantages of working with reduced between-sample virome distance were evident as we showed that the first two principal coordinate axes of PCoA plots explained more variance when using PHFs as the unit of taxonomy. We also showed that the proportion of variance explained by disease and dysbiosis status were greater when using PHFs. These findings are in line with those from Clooney *et al.* (2), who analyzed human IBD viromes. They showed that gene-sharing-based genus-level taxonomy, compared to contig-based analyses, better identified disease-associated compositional changes and increased the variance explained by the first two

322 principal coordinate axes. These observations together highlight the importance of using a higher
323 taxonomic rank when making cross-individual comparisons of gut viromes.

324 A key additional benefit of using predicted hosts lies in the biologically relevant
325 information they provide. This contrasts with existing gene-sharing and phage morphology-based
326 taxonomy approaches, where taxonomic groups are not necessarily informative of how phages
327 interact with their bacterial hosts or the ecosystem at-large (32). Phages in several ecosystems,
328 including the gut, have been shown to be strong regulators of bacterial abundance, diversity, and
329 metabolism (11, 12, 33). Therefore, grouping phages by their predicted hosts provides context for
330 the effects that they may have on the bacterial community and beyond. We showed that in the
331 context of IBD, dysbiotic samples were enriched in *Enterobacteriaceae* PHFs, and depleted in
332 *CAG-74, Ruminococcaceae*, *Acidaminococcaceae* and *Acutalibacteraceae* PHFs. Our analyses
333 provide a framework to identify interactions relevant to disease, although follow-up studies are
334 needed to understand the importance of these phage-host interactions. For instance, phage
335 enrichment in tandem with host depletion could be relevant to several diseases (34–36). While the
336 increased abundance of *Enterobacteriaceae* PHFs we observed in dysbiotic samples is likely a
337 consequence of increased host abundance, it is interesting to note that these phages were enriched
338 in *cys*O, a gene involved in cysteine biosynthesis. Notably, *cysO* has been directly tied to defence
339 against oxidative stress (37). As *Enterobacteriaceae* are known to proliferate in the inflamed gut
340 in the face of oxidative stress (38, 39), our data imply that phage-encoded AMGs could be a source
341 of this resistance. More broadly, the observation that different PHFs carry distinct AMGs suggests
342 that grouping at the phage host family level provides an additional layer of functional insight
343 beyond phage-host relationships. However, an important consideration is that potential

344 inaccuracies in defining prophage borders (40) could lead to an overestimation of AMGs (41).
345 Thus, while these findings merit further investigation, they should be interpreted with caution.

346 While we do propose the use of PHFs for between-sample virome comparisons, it should
347 be noted that this method is reliant on the sensitivity of iPHoP (or any other phage-host matching
348 bioinformatic tool used). In our analyses, between 12.9% (Lloyd-Price *et al*. dataset) and 13.7%
349 (Shkoporov *et al*. dataset) vOTUs did not have an iPHoP-assigned host. This may become an even
350 larger issue if this approach is applied to non-human associated microbiomes where iPHoP
351 performs with less sensitivity (10). Regardless, it is reasonable to assume that the sensitivity of
352 bioinformatic phage-host prediction tools will improve alongside recent improvements in phage
353 genome reconstruction approaches such as contig extension (42) and viral binning (43).

354 To assess the accuracy of bioinformatic phage-host predictions, we measured the
355 concordance at different taxonomic ranks between iPHoP, a bioinformatic tool, and Hi-C
356 sequencing, which relies on physical linkage between phage and host. Due to the prohibitive costs
357 associated with Hi-C sequencing, especially when applied to large volumes of sample, we suggest
358 using iPHoP for family-level host predictions as a suitable alternative. Still, this approach should
359 be interpreted with caution as the concordance between iPHoP and Hi-C sequencing was only
360 assessed using fecal samples. These trends could feasibly differ depending on the environment
361 sampled.

362 Lastly, as phage-host range is often not beyond the species and strain level, by grouping
363 phages at the host family level, this method lacks the sensitivity to detect trends in specific phage-
364 host pairs. Despite these limitations, as bioinformatics methods to detect phage-host pairs improve
365 their resolution, similar approaches to PHFs could be used at lower taxonomic ranks.
366

References

1. Shkoporov AN, Clooney AG, Sutton TDS, Ryan FJ, Daly KM, Nolan JA, McDonnell SA, Khokhlova EV, Draper LA, Forde A, Guerin E, Velayudhan V, Ross RP, Hill C. 2019. The human gut virome is highly diverse, stable, and individual specific. Cell Host Microbe 26:527–541.

2. Clooney AG, Sutton TDS, Shkoporov AN, Holohan RK, Daly KM, O'Regan O, Ryan FJ, Draper LA, Plevy SE, Ross RP, Hill C. 2019. Whole-virome analysis sheds light on viral dark matter in inflammatory bowel disease. Cell Host Microbe 26:764–778.

3. Nishijima S, Nagata N, Kiguchi Y, Kojima Y, Miyoshi-Akiyama T, Kimura M, Ohsugi M, Ueki K, Oka S, Mizokami M, Itoi T, Kawai T, Uemura N, Hattori M. 2022. Extensive gut virome variation and its associations with host and environmental factors in a population-level cohort. Nat Commun 13:5252.

4. Beller L, Deboutte W, Vieira-Silva S, Falony G, Tito RY, Rymenans L, Yinda CK, Vanmechelen B, Van Espen L, Jansen D, Shi C, Zeller M, Maes P, Faust K, Van Ranst M, Raes J, Matthijnssens J. 2022. The virota and its transkingdom interactions in the healthy infant gut. Proc Natl Acad Sci 119:e2114619119.

5. Zuo T, Sun Y, Wan Y, Chan FKL, Miao Y, Ng SC. 2020. Human-Gut-DNA Virome Variations across Geography, Ethnicity, and Urbanization. Cell Host Microbe 1–11.

6. Stockdale SR, Shkoporov AN, Khokhlova EV, Daly KM, McDonnell SA, O' Regan O, Nolan JA, Sutton TDS, Clooney AG, Ryan FJ, Sheehan D, Lavelle A, Draper LA, Shanahan F, Ross

RP, Hill C. 2023. Interpersonal variability of the human gut virome confounds disease signal detection in IBD. Commun Biol 6:221.

7. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR, Kropinski AM, Krupovic M, Lavigne R, Turner D, Sullivan MB. 2019. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol 37:632–639.

8. Marbouty M, Thierry A, Millot GA, Koszul R. 2021. MetaHiC phage-bacteria infection network reveals active cycling phages of the healthy human gut. eLife 10:e60608.

9. Shamash M, Maurice CF. 2021. Phages in the infant gut: a framework for virome development during early life. ISME J https://doi.org/10.1038/s41396-021-01090-x.

10. Roux S, Camargo AP, Coutinho FH, Dabdoub SM, Dutilh BE, Nayfach S, Tritt A. 2023. iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. PLOS Biol 21:e3002083.

11. Reyes A, Wu M, McNulty NP, Rohwer FL, Gordon JI. 2013. Gnotobiotic mouse model of phage-bacterial host dynamics in the human gut. Proc Natl Acad Sci U S A, 2013/11/20 ed. 110:20236–20241.

12. Hsu BB, Gibson TE, Yeliseyev V, Liu Q, Lyon L, Bry L, Silver PA, Gerber GK. 2019. Dynamic Modulation of the Gut Microbiota and Metabolome by Bacteriophages in a Mouse Model. Cell Host Microbe 25:803-814.e5.

13. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ, Casero D, Courtney H, Gonzalez A, Graeber TG, Hall AB, Lake K, Landers CJ, Mallick H, Plichta DR, Prasad M, Rahnavard G, Sauk J, Shungin D, Vázquez-Baeza Y, White RA, Bishai J, Bullock K, Deik A, Dennis C, Kaplan JL, Khalili H, McIver LJ, Moran CJ, Nguyen L, Pierce KA, Schwager R, Sirota-Madi A, Stevens BW, Tan W, ten Hoeve JJ, Weingart G, Wilson RG, Yajnik V, Braun J, Denson LA, Jansson JK, Knight R, Kugathasan S, McGovern DPB, Petrosino JF, Stappenbeck TS, Winter HS, Clish CB, Franzosa EA, Vlamakis H, Xavier RJ, Huttenhower C. 2019. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature 569:655–662.

14. Uritskiy G, Press M, Sun C, Huerta GD, Zayed AA, Wiser A, Grove J, Auch B, Eacker SM, Sullivan S, Bickhart DM, Smith TPL, Sullivan MB, Liachko I. 2021. Accurate viral genome reconstruction and host assignment with proximity-ligation sequencing. bioRxiv https://doi.org/10.1101/2021.06.14.448389.

15. McMurdie PJ, Holmes S. 2013. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE 8:e61217.

16. Dahl EM, Neer E, Bowie KR, Leung ET, Karstens L. 2022. *microshades* : An R Package for Improving Color Accessibility and Organization of Microbiome Data. Microbiol Resour Announc 11:e00795-22.

17. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. 2019. vegan: Community Ecology Package.

18. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics, 2014/04/01 ed. 30:2114–2120.

19. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.

20. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, Mailyan A, Manghi P, Scholz M, Thomas AM, Valles-Colomer M, Weingart G, Zhang Y, Zolfo M, Huttenhower C, Franzosa EA, Segata N. 2021. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. eLife 10:e65088.

21. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. Bioinformatics 31:1674–1676.

22. Kieft K, Zhou Z, Anantharaman K. 2020. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome 8:90.

23. Nayfach S, Pedro Camargo A, Eloe-Fadrosh E, Roux S. 2020. CheckV: assessing the quality of metagenome-assembled viral genomes. bioRxiv 1–20.

24. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, Kuhn JH, Lavigne R, Brister JR, Varsani A, Amid C, Aziz RK, Bordenstein SR, Bork P, Breitbart M, Cochrane GR, Daly RA, Desnues C, Duhaime MB, Emerson JB, Enault F, Fuhrman JA, Hingamp P, Hugenholtz P, Hurwitz BL, Ivanova NN, Labonté JM, Lee KB, Malmstrom RR, Martinez-Garcia M, Mizrachi IK, Ogata H, Páez-Espino D, Petit MA, Putonti C, Rattei T,

Reyes A, Rodriguez-Valera F, Rosario K, Schriml L, Schulz F, Steward GF, Sullivan MB, Sunagawa S, Suttle CA, Temperton B, Tringe SG, Thurber RV, Webster NS, Whiteson KL, Wilhelm SW, Wommack KE, Woyke T, Wrighton KC, Yilmaz P, Yoshida T, Young MJ, Yutin N, Allen LZ, Kyrpides NC, Eloe-Fadrosh EA. 2019. Minimum information about an uncultivated virus genome (MIUVIG). Nat Biotechnol 37:29–37.

25. Zhou Z, Yu M, Ding G, Gao G, He Y. 2020. Diversity and structural differences of bacterial microbial communities in rhizocompartments of desert leguminous plants. PLOS ONE 15:e0241057.

26. Barnett D, Arts I, Penders J. 2021. microViz: an R package for microbiome data visualization and statistics. J Open Source Softw 6:3201.

27. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550.

28. Hurwitz BL, U'Ren JM. 2016. Viral metabolic reprogramming in marine ecosystems. Curr Opin Microbiol 31:161–168.

29. Burns-Huang K, Mundhra S. 2019. Mycobacterium tuberculosis cysteine biosynthesis genes mec+-cysO-cysM confer resistance to clofazimine. Tuberculosis 115:63–66.

30. Allison SD, Martiny JBH. 2008. Resistance, resilience, and redundancy in microbial communities. Proc Natl Acad Sci 105:11512–11519.

31. Tian L, Wang X-W, Wu A-K, Fan Y, Friedman J, Dahlin A, Waldor MK, Weinstock GM, Weiss ST, Liu Y-Y. 2020. Deciphering functional redundancy in the human microbiome. Nat Commun 11:6217.

32. Turner D, Kropinski AM, Adriaenssens EM. 2021. A roadmap for genome-based phage taxonomy. Viruses 13:506.

33. Rasmussen TS, Mentzel CMJ, Kot W, Castro-Mejía JL, Zuffa S, Swann JR, Hansen LH, Vogensen FK, Hansen AK, Nielsen DS. 2020. Faecal virome transplantation decreases symptoms of type 2 diabetes and obesity in a murine model. Gut 69:2122–2130.

34. Cornuault JK, Petit M-A, Mariadassou M, Benevides L, Moncaut E, Langella P, Sokol H, De Paepe M. 2018. Phages infecting Faecalibacterium prausnitzii belong to novel viral genera that help to decipher intestinal viromes. Microbiome 6:65.

35. Tetz G, Brown SM, Hao Y, Tetz V. 2019. Type 1 Diabetes: an Association Between Autoimmunity, the Dynamics of Gut Amyloid-producing E. coli and Their Phages. Sci Rep 9:9685.

36. Madi N, Cato ET, Abu Sayeed Md, Creasy-Marrazzo A, Cuénod A, Islam K, Khabir MdIU, Bhuiyan MdTR, Begum YA, Freeman E, Vustepalli A, Brinkley L, Kamat M, Bailey LS, Basso KB, Qadri F, Khan AI, Shapiro BJ, Nelson EJ. 2024. Phage predation, disease severity, and pathogen genetic diversity in cholera patients. Science 384:eadj3166.

37. Tikhomirova A, Rahman MM, Kidd SP, Ferrero RL, Roujeinikova A. 2024. Cysteine and resistance to oxidative stress: implications for virulence and antibiotic resistance. Trends Microbiol 32:93–104.

38. Winter SE, Winter MG, Xavier MN, Thiennimitr P, Poon V, Keestra AM, Laughlin RC, Gomez G, Wu J, Lawhon SD, Popova IE, Parikh SJ, Adams LG, Tsolis RM, Stewart VJ, Bäumler AJ. 2013. Host-Derived Nitrate Boosts Growth of *E. coli* in the Inflamed Gut. Science 339:708–711.

39. Rivera-Chávez F, Lopez CA, Bäumler AJ. 2017. Oxygen as a driver of gut dysbiosis. Free Radic Biol Med 105:93–101.

40. Zünd M, Ruscheweyh H-J, Field CM, Meyer N, Cuenca M, Hoces D, Hardt W-D, Sunagawa S. 2021. High throughput sequencing provides exact genomic locations of inducible prophages and accurate phage-to-host ratios in gut microbial strains. Microbiome 9:77.

41. Enault F, Briet A, Bouteille L, Roux S, Sullivan MB, Petit MA. 2017. Phages rarely encode antibiotic resistance genes: A cautionary tale for virome analyses. ISME J 11:237–247.

42. Chen L, Banfield JF. 2024. COBRA improves the completeness and contiguity of viral genomes assembled from metagenomes. Nat Microbiol 9:737–750.

43. Kieft K, Adams A, Salamzade R, Kalan L, Anantharaman K. 2022. vRhyme enables binning of viral genomes from metagenomes. Nucleic Acids Res 50:e83–e83.

Figure Legends

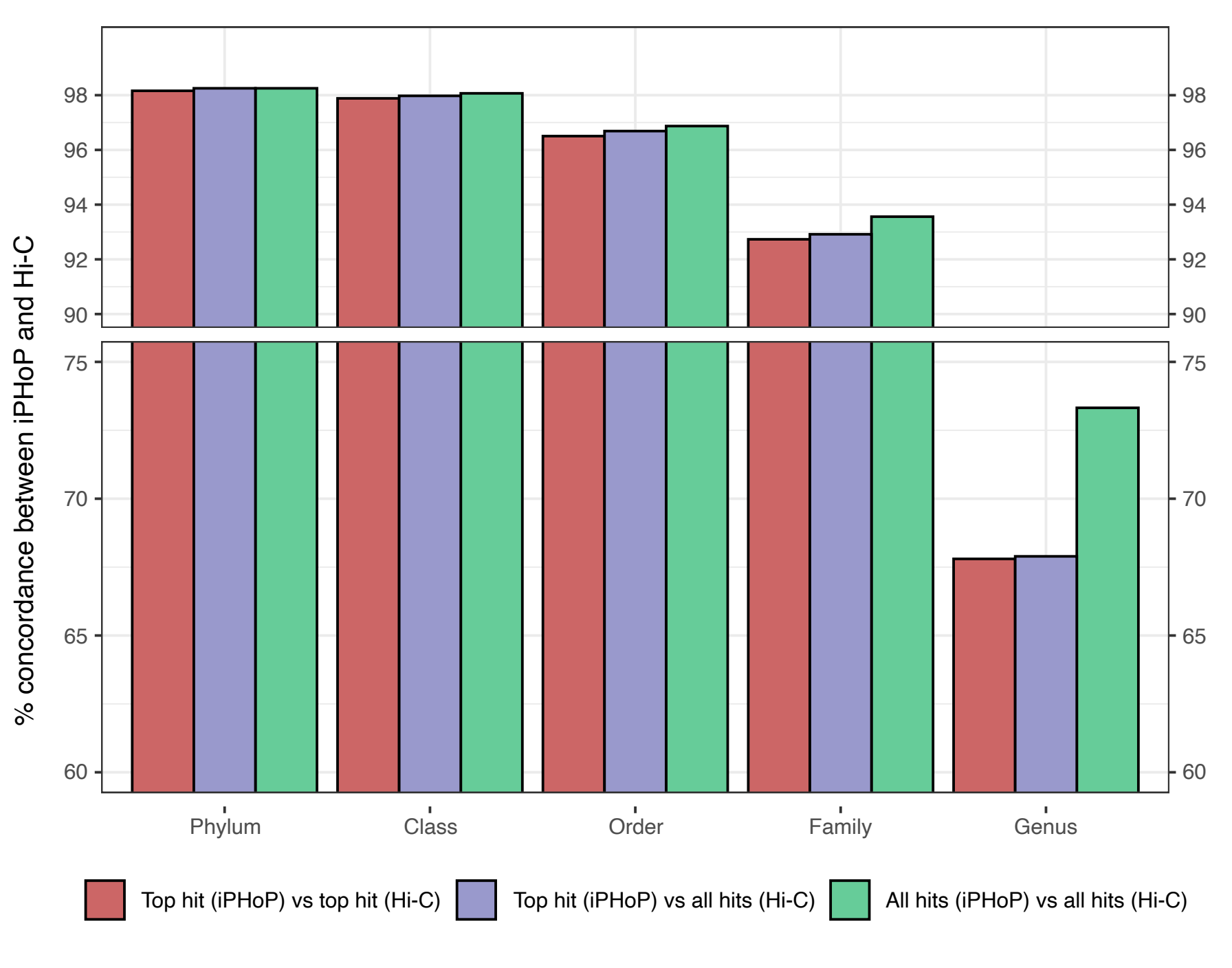
Figure 1. Computationally-predicted bacterial hosts for vOTUs are concordant with *in situ* associations to the bacterial family level. Agreement between iPHoP predicted host range and Hi-C assigned host range at various taxonomic ranks for 1,243 vOTUs. Additional comparisons were made when iPHoP predicted multiple hosts for a vOTU (see main text for details on the three comparisons).

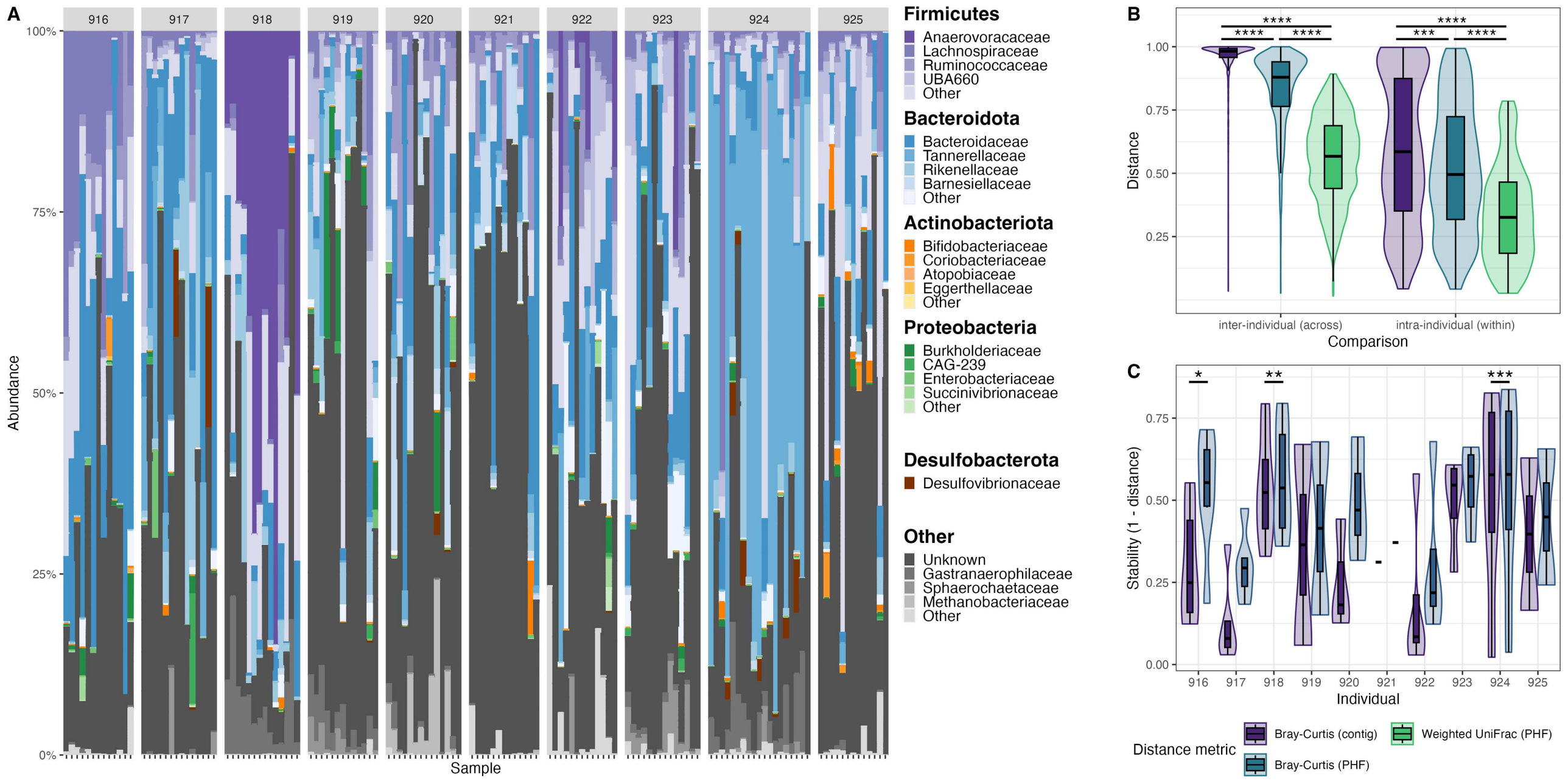
Figure 2. PHFs reduce interindividual variation and increase intraindividual virome stability in a cohort of 10 healthy individuals. Data were analyzed from a previously published study of 10 healthy individuals (1). **(A)** Taxonomic bar plots of virome composition at the PHF level for each individual over time. Facet labels above the bar plots correspond to the subject IDs from the original study. **(B)** Ecological distances between samples with Bray-Curtis at the contig level, Bray-Curtis at the PHF level, and Weighted UniFrac at the PHF level. Interindividual and intraindividual comparisons are both shown. Significance was assessed using the Friedman test with the post-hoc Wilcoxon signed-rank test, using Bonferroni correction for multiple comparisons ($p < 0.001$, ***; $p < 0.0001$, ****). **(C)** Virome stability, defined here as (1 - ecological distance from previous sample), was calculated for each individual using the Bray-Curtis distance metrics at the contig level and at the PHF level. Significance was assessed using the Wilcoxon signed-rank test ($p < 0.01$, **; $p < 0.001$, ***).

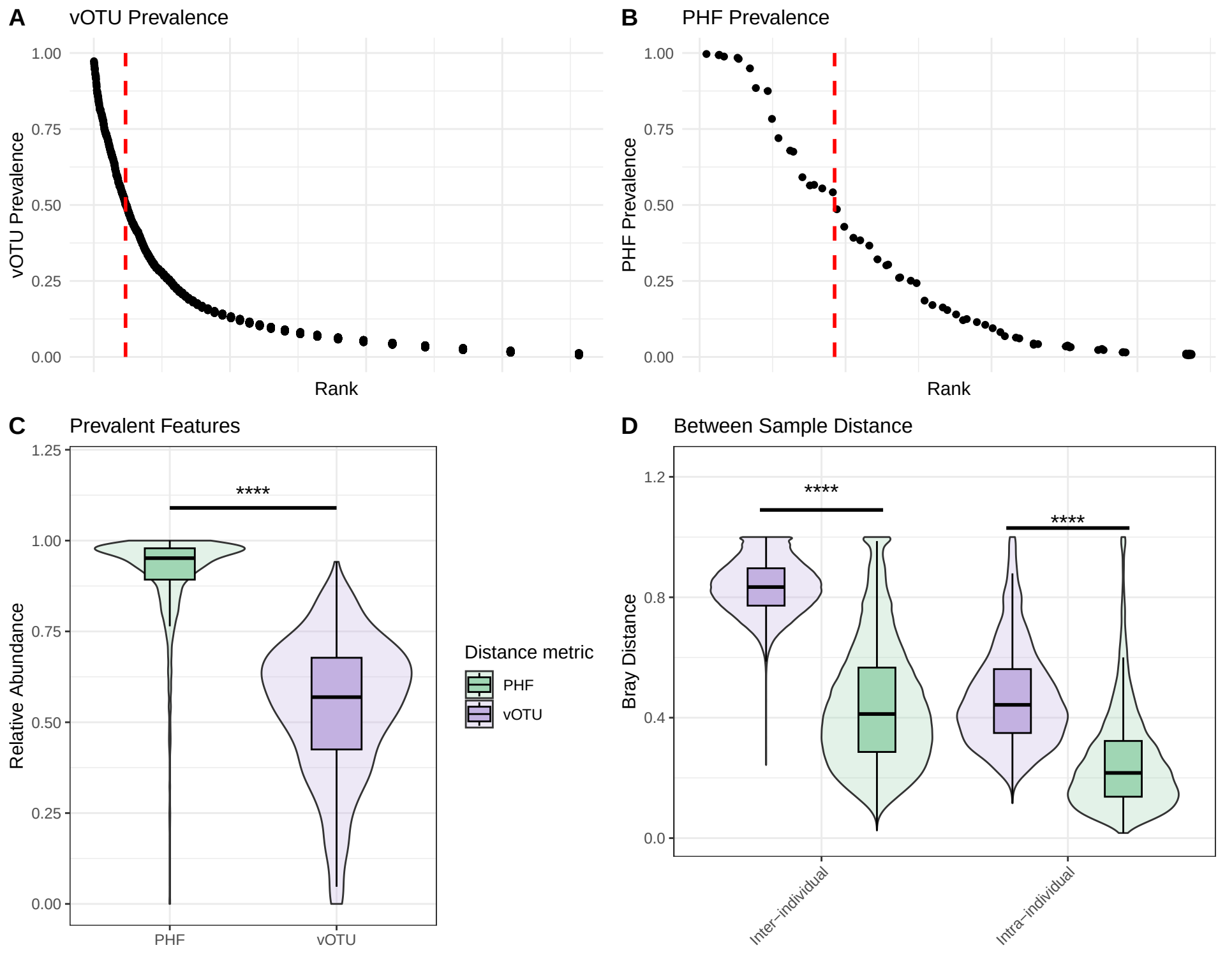
Figure 3. PHFs are prevalent and reduce intra- and interindividuality in a large human IBD cohort. Data were analyzed from the previously published HMP2 dataset (13). Samples with low read counts (< 1,500) were removed from analyses. In total, bulk metagenomes from 1,093 samples from 115 individuals (57 CD, 31 UC, 27 non-IBD controls) were included for downstream analyses. **(A, B)** Rank prevalence distributions of vOTUs **(A)** and PHFs **(B)** across individuals. In total, there were 3,886 distinct vOTUs and 75 distinct PHFs. The dotted red line indicates the rank at which features are more, or less than, 50% prevalent. **(C)** Mean relative abundance of features (PHFs vs. vOTUs) that were present in more than 50% of individuals in the dataset. **(D)** Bray-Curtis distance between samples according to interindividual or intraindividual comparisons. Significance was assessed using the Wilcoxon signed-rank test ($p \leq 0.0001$, ****).

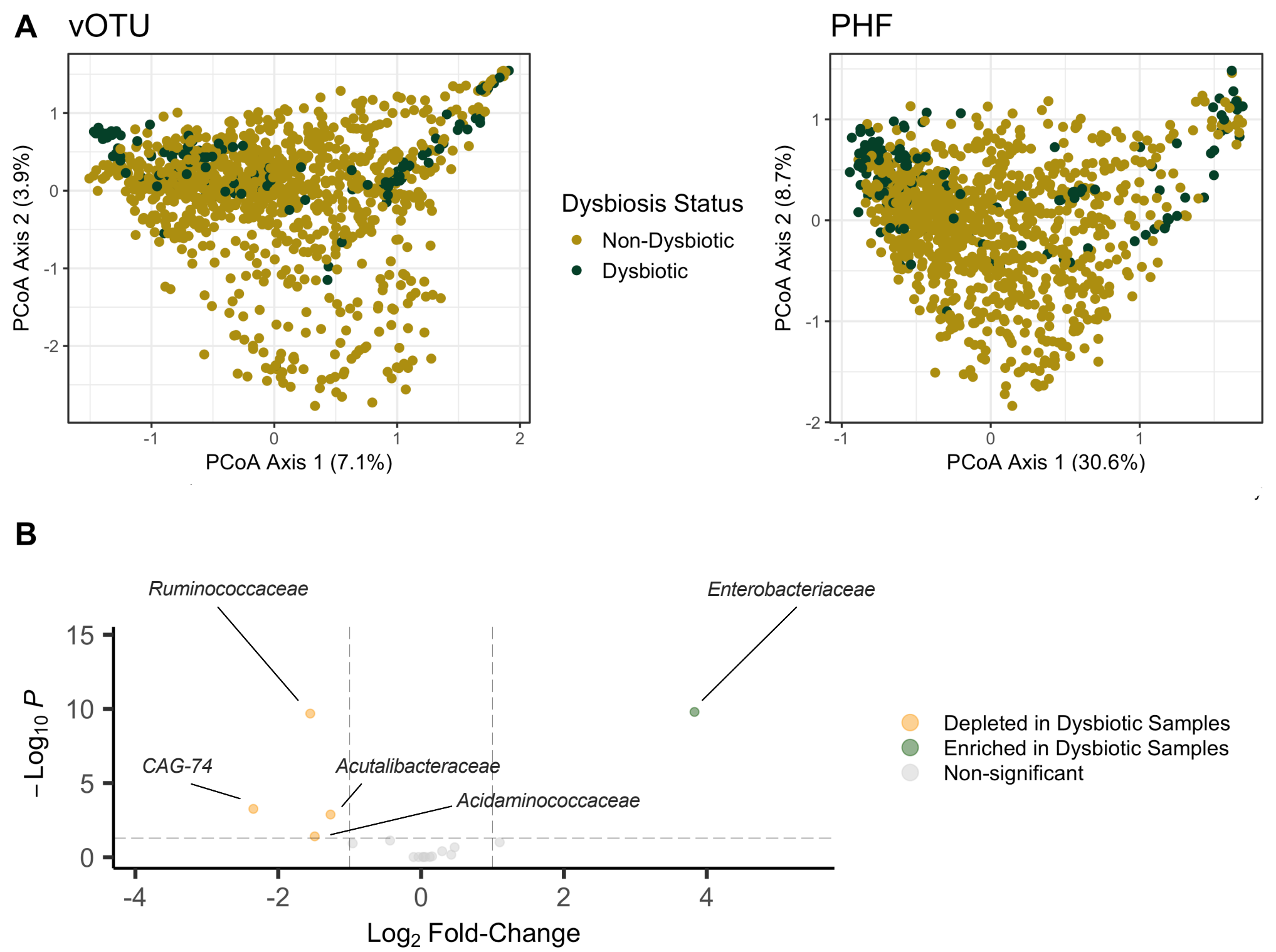
Figure 4. PHFs reveal disease-specific signatures of IBD. Data were analyzed from the previously published HMP2 dataset (13). Samples with low read counts (< 1,500) were removed from analyses. **(A)** PCoA plots generated from Bray-Curtis distances matrices using vOTUs (left) and PHFs (right). Samples are color-coded according to the dysbiotic status identified in (13). **(B)** Differentially abundant PHFs based on dysbiosis status. Only individuals which had both a dysbiotic and non-dysbiotic sample were included. Only PHFs that were more than 50% prevalent across individuals were considered for these analyses. PHFs with an adjusted p value ≤ 0.05 and with a $\log_2$ fold-change ≥ 1 or with a $\log_2$ fold-change ≤ -1 were considered differentially abundant.

548 **Supplementary Figure 1. vOTU membership of PHFs.** Data were analyzed from the previously
549 published HMP2 dataset (13). Samples with low read counts (< 1,500) were removed from
550 analyses. The distribution of PHFs is based on the number of distinct vOTUs that comprises them.
551 The 5 PHFs comprised of the most vOTUs are indicated on the plot.
552
553
554 **Supplementary Figure 2. AMG distribution across PHFs.** Data were analyzed from the
555 previously published HMP2 dataset (13). Only the 12 PHFs that contained > 10 AMGs are shown
556 here. AMGs were detected using VIBRANT, which uses KEGG annotations to assign metabolic
557 categories. **(A)** Distribution of the AMGs found within each PHF. We then determined the number
558 of AMGs per Mb of assembled vOTUs for each PHF, broken down by **(B)** carbohydrate
559 metabolism genes, and **(C)** folding, sorting and degradation genes. **(D)** Number of vOTUs
560 containing *cysO* across PHFs.
561
562
563 **Supplementary Figure 3. Ordination of samples based on patient diagnosis.** Data were
564 analyzed from the previously published HMP2 dataset (13). Samples with low read counts (<
565 1,500) were removed from analyses. PCoA plots were generated from Bray-Curtis distances
566 matrices using vOTUs (left) and PHFs (right). Samples are color-coded according to the diagnosis
567 status identified in (13).
568


% concordance between iPHoP and Hi-C
98
96
94
92
90
75
70
65
60
Phylum
Class
Order
Family
Genus
Top hit (iPHoP) vs top hit (Hi-C)
Top hit (iPHoP) vs all hits (Hi-C)
All hits (iPHoP) vs all hits (Hi-C)


A
916
917
918
919
920
921
922
923
924
925
100%
75%
50%
25%
0%
Abundance
Sample
Firmicutes
Anaerovoracaceae
Lachnospiraceae
Ruminococcaceae
UBA660
Other
Bacteroidota
Bacteroidaceae
Tannerellaceae
Rikenellaceae
Barnesiellaceae
Other
Actinobacteriota
Bifidobacteriaceae
Coriobacteriaceae
Atopobiaceae
Eggerthellaceae
Other
Proteobacteria
Burkholderiaceae
CAG-239
Enterobacteriaceae
Succinivibrionaceae
Other
Desulfobacterota
Desulfovibrionaceae
Other
Unknown
Gastranaerophilaceae
Sphaerochaetaceae
Methanobacteriaceae
Other
B
****
****
****
****
***
****
1.00
0.75
0.50
0.25
Distance
inter-individual (across)
intra-individual (within)
Comparison
C
*
**
***
0.75
0.50
0.25
0.00
Stability (1 - distance)
916
917
918
919
920
921
922
923
924
925
Individual
Distance metric
Bray-Curtis (contig)
Bray-Curtis (PHF)
Weighted UniFrac (PHF)


A
vOTU Prevalence
1.00
0.75
0.50
0.25
0.00
vOTU Prevalence
Rank
B
PHF Prevalence
1.00
0.75
0.50
0.25
0.00
PHF Prevalence
Rank
C
Prevalent Features
1.25
1.00
0.75
0.50
0.25
0.00
Relative Abundance
****
PHF
vOTU
Distance metric
PHF
vOTU
D
Between Sample Distance
1.2
0.8
0.4
0.0
Bray Distance
****
****
Inter−individual
Intra−individual


A
vOTU
PHF
PCoA Axis 2 (3.9%)
PCoA Axis 1 (7.1%)
PCoA Axis 2 (8.7%)
PCoA Axis 1 (30.6%)
Dysbiosis Status
Non-Dysbiotic
Dysbiotic
B
Ruminococcaceae
Enterobacteriaceae
CAG-74
Acutalibacteraceae
Acidaminococcaceae
−Log10 P
Log2 Fold-Change
Depleted in Dysbiotic Samples
Enriched in Dysbiotic Samples
Non-significant