iPHoP: an integrated machine-learning framework to maximize host prediction for metagenome-assembled virus genomes

Simon Roux[1]*, Antonio Pedro Camargo[1], Felipe H. Coutinho[2], Shareef M. Dabdoub[3], Bas E. Dutilh[4,5], Stephen Nayfach[1], Andrew Tritt[6]

[1] DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[2] Instituto de Ciencias del Mar (ICM-CSIC), Barcelona, Spain
[3] Division of Biostatistics and Computational Biology, University of Iowa College of Dentistry, Iowa City, IA, USA
[4] Institute of Biodiversity, Faculty of Biological Sciences, Cluster of Excellence Balance of the Microverse, Friedrich Schiller University, Rosalind Franklin Strasse 1, 07743, Jena, Germany
[6] Theoretical Biology and Bioinformatics, Science for Life, Utrecht University, Padualaan 8, 3584 CH, Utrecht, The Netherlands
[6] Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

* Correspondence to: sroux@lbl.gov

Abstract

The extraordinary diversity of viruses infecting bacteria and archaea is now primarily studied through metagenomics. While metagenomes enable high-throughput exploration of the viral sequence space, metagenome-derived genomes lack key information compared to isolated viruses, in particular host association. Different computational approaches are available to predict the host(s) of uncultivated viruses based on their genome sequences, but thus far individual approaches are limited either in precision or in recall, i.e. for a number of viruses they yield erroneous predictions or no prediction at all. Here we describe iPHoP, a two-step framework that integrates multiple methods to provide host predictions for a broad range of viruses while retaining a low (<10%) false-discovery rate. Based on a large database of metagenome-derived virus genomes, we illustrate how iPHoP can provide extensive host prediction and guide further characterization of uncultivated viruses. iPHoP is available at https://bitbucket.org/srouxjgi/iphop, through a Bioconda recipe, and a Docker container.

Introduction

Viruses are widespread and influential throughout all ecosystems. In microbial communities, viral infections can shift community composition and structure via viral lysis, and alter biogeochemical processes and metabolic outputs through reprogramming of host cells during infection[1–3]. Given current challenges for cultivating many environmental microbes and their viruses, the extensive viral diversity is now primarily explored via metagenomics, i.e., by assembling genomes of uncultivated viruses directly from whole community shotgun sequencing data[4–6]. Over the last decade, metagenomic studies have been incredibly powerful in revealing viral diversity on Earth, investigating eco-evolutionary drivers of viral biogeography, and connecting viruses to ecological and metabolic processes[7–9]. A major limitation of these approaches is that metagenome-derived viral genomes have no inherent link with a host, as is the case for isolates[4]. This lack of host association remains a critical hurdle when attempting to study virus-host interactions and dynamics in natural communities, in particular for the highly diverse bacteriophages ("phages"), viruses infecting bacteria[4].

Given the importance of phage-host interactions in microbiome processes, computational methods to predict the host(s) of a phage based on its genome sequence are highly desirable, and the subject of active research[10,11]. Existing host prediction tools leverage either various levels and patterns of sequence similarity between phage and host genomes ("host-based" tools hereafter), or use a "guilt-by-association" approach by comparing the query phage to a database of viruses with known host(s) ("phage-based" tools).

In host-based tools, sequence similarity between phages and hosts can be based on sequence alignment, reflecting for instance prophages integrated in the host genome or similarity between phage genomes and host CRISPR spacers[10,12], or can rely on alignment-free approaches, e.g., comparison of nucleotide k-mer frequencies, in which case these typically reflect the overall adaptation of virus genomes to their host cell machinery[13,14,23,15–22]. Because they rely on different signals, these host-based tools display varying levels of recall and specificity, and are likely to be each relevant for different types of samples and viruses[10]. In previous benchmarks, alignment-based methods could reach high specificity when using strict cutoffs, for instance >75% of predictions correct at the species level, but only for a limited subset of the input phages due to limitations of the host reference database[10]. Meanwhile, in the same benchmark, alignment-free methods appeared to contain a genuine and strong phage-host signal for a broader range of phages, but more complex to parse as the highest scoring host was often (>50% of the time) yielding an incorrect prediction at the species, genus, and family level.

Complementarily, "phage-based" tools rely not on phage-host similarity, but extract information from a database of reference phages and archaeoviruses with known host(s)[24–27]. The most recent tools in this category have been the most promising overall, with benchmarks suggesting both high recall and high specificity. For instance, RaFAH achieved a 33% improvement in F1 score (combination of recall and precision) at the genus level compared to host-based methods[25]. While phage-based approaches are particularly suitable if related phages exist with known hosts, RaFAH also predicted hundreds of archaeal viruses, i.e. domain-level host predictions, despite archaeoviruses being under-represented in the database[25]. However, it remains unclear to what extent phage-based tools can provide reliable host prediction at lower ranks such as genus or species for entirely novel phages, and how to best complement these phage-based predictions with host-based signals[11].

With multiple tools available for host prediction, several studies have attempted to integrate the results from several approaches into a single prediction for each virus. This integration step was originally performed via empirical "rule sets" prioritizing methods based on empirical accuracy or error rate estimations[28,29]. Recently, several automated tools were developed that instead leverage machine learning to obtain an integrated host prediction. PhisDetector[30] combines multiple host-based methods, both alignment-based and alignment-free, and uses an ensemble of machine-learning approaches to

evaluate the confidence of each potential phage-host pair. VirHostMatcherNet[31] proposes to integrate both virus-virus and virus-host signal in a modeled virus-host network, from which potential virus-host pairs are evaluated using a logistic regression. While both tools showed potential improvements compared to single methods, none of the benchmarks provided suggested that they could reach a low (<10%) false-discovery rate (FDR) at the host genus level, even with the strictest cutoffs. In addition, no benchmark was carried out across different degrees of phage "novelty", i.e., different degrees of similarity to the most closely related reference, so it remains unclear how these approaches perform on "known" and "novel" phages.

Here we present iPHoP, a tool for **i**ntegrated **P**hage-**Ho**st **P**rediction, enabling high recall and low FDR at the host genus level for both known and novel phages. We first demonstrate the complementarity of phage-based and host-based approaches, and describe a new modular machine-learning framework that yields highly accurate predictions at the genus level. Using a diverse set of 216,015 metagenome-derived phage genomes, we further show that iPHoP enables high-confidence host genus prediction (estimated <10% FDR) for phages across a broad range of ecosystems and novelty compared to isolated references.

Results

To design an integrated framework for host prediction, we first evaluated the performance and complementarity of 10 existing methods on a common benchmark dataset[10,12–14,16,24–27]. We especially focused on comparing tool performances across a range of "novelty", i.e., using a test set that included both viruses closely related to references and viruses entirely novel.

Limitations and complementarity of individual host prediction methods

A set of published alignment-based and alignment-free methods, either phage-based or host-based, was selected for benchmarking (Table S1). These tools were evaluated on a common test dataset including bacteriophage and archaeovirus genomes available in NCBI GenBank but not included in NCBI RefSeq[32], and thus typically not used to train any of these tools (see Methods and Table S2). This test dataset contained 1,870 genomes, spanning across 170 host genera, including both temperate and virulent phages, and with both "known" and "novel" genomes (>90% and <5% genome-wide average amino acid identity, or AAI[33], to the closest reference, respectively, see Supplementary Fig. S1). As host references, we opted to use all genomes included in the GTDB database[34], supplemented by additional publicly available genomes from the IMG isolate database[35] and the GEM catalog[36]. For each tool, we assessed host predictions at the host genus rank based on a naive "best hit" approach and using relaxed cutoffs (see Methods).

First, we evaluated the recall of each tool, i.e., the total number of correct predictions obtained (Fig. 1A). The recall differed across the tool categories, with the lowest observed for host-based alignment-based tools such as blast and CRISPR, and the highest observed for phage-based tools. The only exception was a very high recall observed for blast-based predictions of temperate phages, which is due to the detection of integrated copies of these phages, or closely related ones, in host genomes. A similar trend, i.e., a higher recall for temperate phages compared to virulent phages, was observed for most approaches albeit to a much lower degree (Supplementary Fig. S2). For CRISPR-based predictions, the low recall compared to other approaches is likely due to limitations of the host database as CRISPR arrays can be absent from large clades of bacteria[37] and, when present, CRISPR spacers are typically highly variable even between closely related strains[10,38].

Next, we evaluated the precision of each tool, i.e., its ability to distinguish correct from incorrect hosts among all its predictions. As previously noted[10], host-based alignment-free tools struggled to achieve a high Positive Predictive Value (PPV), i.e., a low False-Discovery Rate (FDR), even when

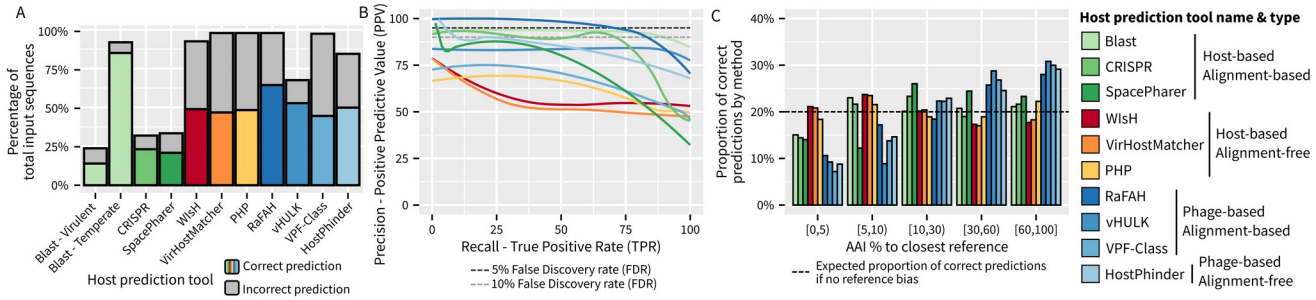

Figure 1. Comparison of different host prediction approaches on a single test dataset. *A. Total number of predictions and number of correct predictions (y-axis) obtained for each tool (x-axis) using a "best hit" approach and relaxed cutoffs (see Methods) on sequences from the test dataset (Table S2). For each tool, the number of correct predictions is indicated by the colored bar, while the total number of predictions is indicated by the gray bar. Similar plots including the whole test dataset, virulent phages only, and temperate phages only are available in Fig. S2. B. Precision-Recall curves for the different tools, using the same color code as in panels A and C. Two standard thresholds, 5% and 10% false discovery rates, are indicate by horizontal dashed lines. C. Relationship between "novelty" of input virus, represented as AAI (Average Amino Acid Identity) percentage to the closest reference on the x-axis, and the number of correct host predictions obtained with each tool. To evenly represent both "known" and "novel" input viruses, 300 sequences were randomly subsampled from each AAI percentage category (x-axis).*

using strict cutoffs (Fig. 1B). In contrast, alignment-based tools, both phage-based and host-based, were able to reach high (>80%) PPV when filtering hits based on score(s). Pragmatically, this means that the scores provided by alignment-based tools are able to distinguish correct from incorrect predictions, while the scores provided by alignment-free tools are usually not sufficient to identify correct predictions.

Phage-based tools thus seemingly present an ideal combination of high recall and high precision, with RaFAH[25] in particular able to maintain a very low FDR (<5%) while providing the highest recall of all tools (Fig. 1, Supplementary Figure S2). However, phage-based tools depend on the availability of a related phage with a known host in the reference database. Specifically, phage-based tools mostly provide predictions for phages that are related to reference sequences, and much less frequently for "novel" phages (<5% AAI to closest reference, Fig. 1C). A similar trend, although less pronounced, can be observed for host-based tools relying on sequence alignment. Meanwhile, alignment-free host-based tools show little to no bias for phages with closely related references, suggesting that these methods would be well suited for dealing with the most "novel" phages. This bias is important to consider because the vast majority (57-80%) of viral genomes identified from metagenomes have <5% AAI to their closest reference (Supplementary Fig. S3), so that phage-based approaches alone would thus not yield reliable host predictions.

Ultimately, these simplified benchmarks suggest that to tackle diverse "known" and "novel" phages, as typically obtained through metagenomics, host prediction tools will need to combine phage-based and host-based approaches. For instance, based on the benchmark in Fig. 1A, the tool with the highest recall (RaFAH) provided host prediction for 52% of phages, while 83% of phages overall were associated with a correct host prediction across all tools. For phage-based approaches, several tools such as RaFAH already provide both high recall and high precision. Conversely, all current host-based methods suffer from either limited recall (alignment-based methods) or limited specificity (alignment-free methods), at least when used individually and in a simple "best hit" approach. In addition, the predictions from different host-based methods partially overlap, suggesting that multiple methods could be considered together to either reinforce or correct each other (Supplementary Fig. S4). For our integrated host prediction tool, we thus decided to first optimize host-based predictions by integrating multiple hits per method and several methods together, and then combine these host-based predictions with an established phage-based method to derive a single host prediction.

Increasing host prediction accuracy by robustly integrating multiple hits for each virus

Elevated false discovery rates with host-based methods have been highlighted previously[10,13,14]. Traditionally, these have been addressed by applying relatively strict cutoffs on the prediction score, and by considering an arbitrary number of hits passing these cutoffs, e.g., the 5 or 10 best hits. These hits might be further integrated using a lowest common ancestor (LCA) approach. Intuitively, this will allow to distinguish reliable cases, where the top hits all point to the same host taxon, from unreliable cases where the top hits correspond to different taxa. Alternatives to LCA approaches have been proposed including the taxonomy-aware sequence similarity ranking framework from PHIRBO[39]. Here, we explored whether machine learning approaches could help improve these predictions by integrating all hits obtained for a virus using a given method.

To consider an ensemble of hits in a taxonomy-aware context, we opted to treat each hit as a separate classification problem, i.e., "is this host hit reliable or not considering the context of other hits obtained for this same virus with the same approach?". For a given input genome, each hit is thus considered as a candidate host, and an ensemble of hits for a virus is provided as input to different classifiers with information on the hits quality as well as phylogenetic distances between each hit and the candidate host based on the GTDB[34] framework (Fig. 2A, Supplementary Fig. S5). The task asked

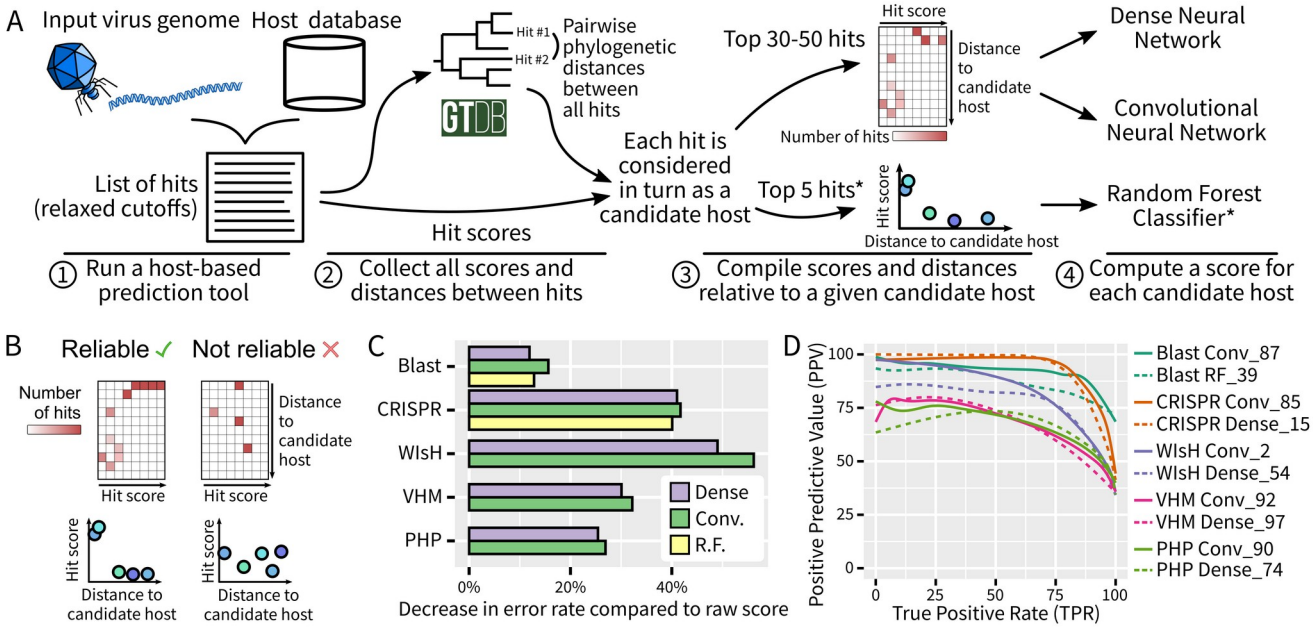

Figure 2. Overview of the single-tool classifiers used in iPHoP. *A. Schematic representation of the process used to score individual hits from host-based tools. Briefly, each hit was scored by a neural network or random forest classifier, which also considered other top hits for the same virus and the same tool. This process was applied to the 5 host-based tools selected ("Blast", "CRISPR", "WIsH", "VHM", "PHP"), except for the random forest classifiers (highlighted with a *) which were only used for "Blast" and "CRISPR". When considering multiple hits, their similarity or difference in terms of host prediction was estimated from the GTDB phylogenies[34] B. Illustration of how multiple hits are represented in neural networks input matrices (top) or random forest classifier inputs (bottom). Two examples are provided, one "reliable" in which the hits with high scores are all consistent and at a small distance to the candidate host considered (left), and the other "unreliable" in which a few hits with medium-to-high scores are scattered across hosts with variable distance to the candidate host considered. C. Estimated improvement in classification provided by the automated classifiers compared to "naive" raw scores. These estimations are based on smoothed ROC curves obtained from the test dataset (see Supplementary Fig. S6) and calculated as the average decrease in false-discovery rate for 17 true positive rates ranging from 10 to 90%. Random forest classifiers were only evaluated for Blast and CRISPR approaches. D. Precision Recall curves for the two classifiers selected for each host-based tool (see Supplementary Table S3). VHM: "VirHostMatcher". Conv: "Convolutional Neural Network". "RF": "Random Forest classifier".*

of the classifiers is to predict whether the candidate host belongs to the correct host genus, and the underlying assumption is that classifiers would learn to recognize reliable series of hits, e.g., cases where most of the top hits are close to the candidate host, from unreliable series of hits, e.g., cases where hits are distributed across diverse hosts and/or distant from the candidate host, without having to resort to arbitrary cutoffs (Fig. 2B).

To evaluate this approach, we applied it separately to 5 host-based methods (Blast, CRISPR, WIsH[13], VirHostMatcher[14,31], PHP[16], see Supplementary Table S1), used RefSeq Virus sequences to train and optimize 3 types of classifier, namely dense neural networks, convolutional neural networks, and random forest classifiers, and compared the results obtained on the test dataset (see above) to a standard best hit approach (Supplementary Fig. S5). Overall, considering multiple hits with automated classifiers reduced the error rate (average FDR) for all methods and all types of classifiers, with the highest reductions obtained with convolutional neural networks (Fig. 2C, Supplementary Fig. S6). This reduction in average error rate was especially important for WIsH and CRISPR-based predictions (>40%), and smaller for BLAST, for which standard scores already seem to perform well.

Finally, we verified whether different variants of each classifier could be complementary, i.e., provide reliable scores for different types of sequence. In all cases, a set of two variants appeared to be the best combination to maximize the number of correct predictions while minimizing the false-discovery rate (see Methods). The 10 classifiers that were ultimately selected (2 for each of the 5 host-based methods) showed improved positive predictive value, often >75%, at most true positive rates, confirming their improved ability to distinguish likely and unlikely candidate hosts compared to the raw score of each method (Fig. 2D).

Integrating host- and phage-based predictions for a comprehensive coverage of phage diversity

After optimizing scoring systems for each host-based method, the next step was to integrate predictions across different methods to obtain a single prediction score taking into account all different approaches for each potential phage-host pair. Traditionally, this has been done using fixed “rule sets” informed by estimation of false-discovery rate for each approach, e.g., prioritizing alignment-based approaches over alignment-free approaches[28,29]. Here, we instead used a 2-step integration process to robustly consider all hits for each input sequence.

First, to leverage the high sensitivity of alignment-free approaches but reduce their error rate, we trained and optimized a random forest classifier based on the scores from the 10 individual host-based classifiers described in the previous section (“iPHoP-RF classifier”, Fig. 3A). This iPHoP-RF score yielded low FDR (≤10%) even at high TPR (≥75%), and was comparable in that regard to the scores obtained from the phage-based tool RaFAH[25], as well as host-based aligment-based tools (Blast and CRISPR, Fig. 3B).

Next, we designed a composite confidence score for each phage-host pair to summarize results from both phage- and host-based methods (Fig. 3A, see Methods). Because blast- and CRISPR-based predictions can also be reliable on their own without the need for any other approach (Fig. 2D), we included the best score for each of these approaches along with the iPHoP-RF score and the score from RaFAH[25], the most reliable phage-based tool in our benchmark (Fig. 1A & B). As expected based on our initial benchmarks (Fig. 1C), different methods provided correct host predictions for different input phages, and combining them led to high rates of phages with correct predictions (≥50%) for both “known” and “novel” phages (Supplementary Fig. S7).

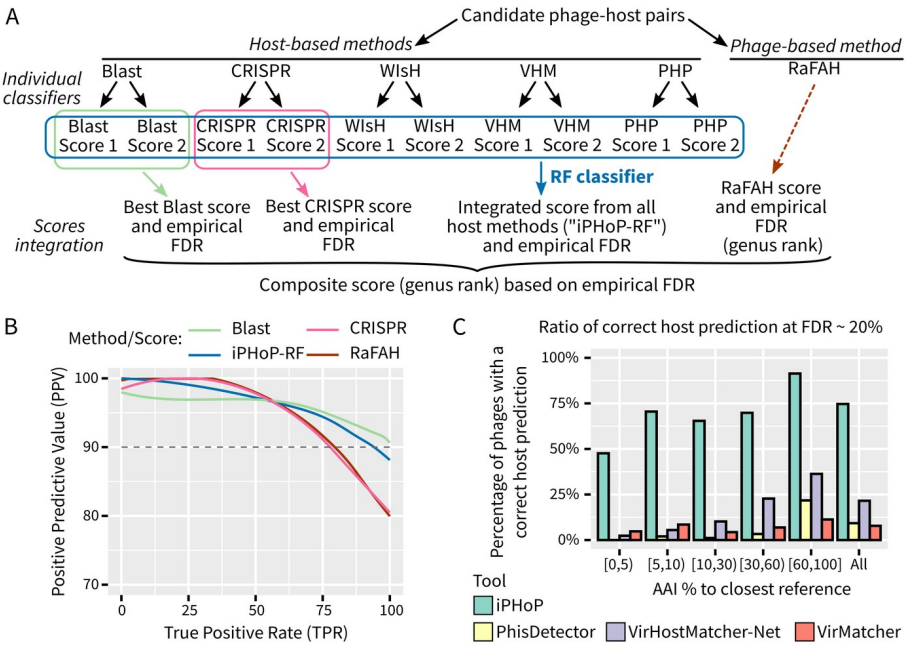

Figure 3. Schematic and performance of iPHoP host genus predictions. *A. Schematic representation of the integration process. "Individual classifiers" refer to single-tools scores calculated for each virus-candidate host pair (see Fig. 2). B. Precision Recall curve for each of the four scores considered in iPHoP composite score, based on the test dataset. C. Comparison of the percentage of input sequences from the test dataset for which a correct host genus prediction was obtained, when using cutoffs limiting false-discovery rate to 20% maximum. This percentage is given for all sequences in the test dataset ("All"), and for subsets of sequences defined based on their amino-acid similarity to the closest reference phage genome. RF: Random Forest. FDR: False-Discovery Rate. AAI: Amino Acid Identity.*

To illustrate the unique features and performance improvements provided by iPHoP, we compared it to other automated tools integrating multiple approaches for host prediction, namely VirMatcher[29], PhisDetector[30], and VirHostMatcher-Net[31]. Based on Receiver Operating Characteristic and Precision Recall curves, iPHoP performed as well as, or better than all other integrated tools (Supplementary Fig. S8). However, the major improvement of iPHoP comes from the number of phages with a host prediction: for a given FDR, iPHoP typically provides ~3 to 5 times more predictions than the next best tool, especially for "novel" phages (Fig. 3C). This is likely due to the fact that (i) iPHoP uniquely leverages both phage-based and host-based approaches, (ii) iPHoP integrates more approaches than any other tool, (iii) the iPHoP host database is larger and more diverse than those used by other tools, and (iv) iPHoP was specifically optimized for predictions at the host genus rank. In contrast, VirHostMatcher-Net relies on a network architecture to represent virus-host interactions and derive host predictions at multiple taxonomic ranks, while PhisDetector was designed to provide host predictions down to the species rank[30,31].

Expanding host predictions in a large database of metagenome-derived viruses

To further evaluate the improvements provided by iPHoP and the remaining challenges when analyzing diverse metagenome-derived phage genomes, we applied iPHoP to 216,015 high-quality (i.e., predicted to be ≥90% complete by CheckV) IMG/VR sequences (Supplementary Fig. S3). We then compared the iPHoP predictions to the current host predictions available in the IMG/VR database, which were primarily based on blast hits to host genomes and CRISPR spacers[8] (Fig. 4A). Overall, iPHoP predictions at an estimated FDR ≤10%, i.e., score ≥90, represented a 1.5- to 13-fold increase compared to the original number of host prediction in the IMG/VR v3 database, however these numbers vary greatly depending on the ecosystem (Fig. 4A). For human-associated microbiomes, about 89% of the high-quality genomes had a host predicted using iPHoP, including 57% with very high confidence predictions (iPHoP score ≥95). For all other ecosystems, the total number of phages with predictions was lower, ranging from ~40–50%, including ~15–22% with medium or high confidence (score ≥90). Across ecosystems, host predictions originated primarily from host-based methods,

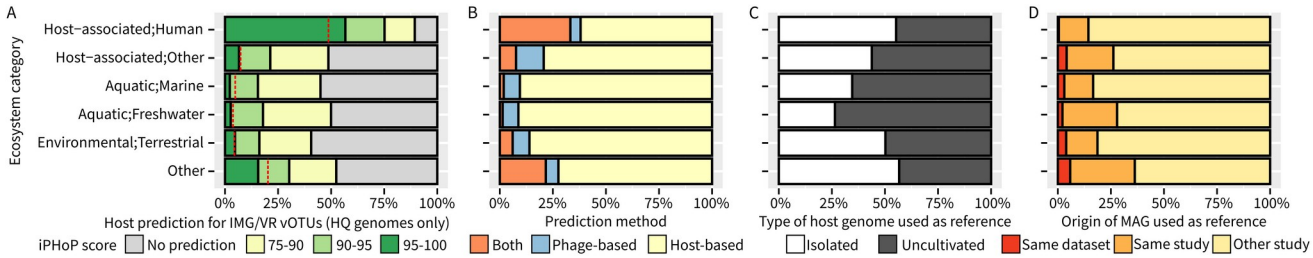

Figure 4. Overview of iPHoP Host prediction for high-quality IMG/VR v3 genomes. *A. Distribution of the best score provided by iPHoP for high-quality genomes from the IMG/VR v3 database by ecosystem. For each IMG/VR vOTU, the best score from iPHoP was considered if ≥ 75, or the vOTU was considered as not having a predicted host. The proportion of sequences for which a host prediction was available in the original IMG/VR database is indicated with a dashed red line. B. Distribution of the type of signal used to achieve host prediction with a score ≥ 90 in iPHoP. "Host-based" includes all 5 host-based tools, while "Phage-based" includes predictions obtained with RaFAH. "Both" includes consistent predictions obtained with RaFAH and at least one host-based tool. C. Percentage of hits from isolated or uncultivated host genomes used in host-based predictions with final scores ≥ 90. These are based on the individual genome hits underlying iPHoP genus-level predictions. D. Origin of the uncultivated host genomes used in host-based predictions with final scores ≥ 90. The original dataset and study ID for the query virus and the uncultivated host genome were obtained from the Gold database, and when both were available, these were compared to evaluate whether the uncultivated host genome originated from the same dataset, a different dataset from the same study, or another study from the query virus.*

consistent with the high number of metagenome-derived sequences unrelated to those in the reference databases (Fig. 4B, Fig. S3, Supplementary Fig. S9). Human microbiomes again stand out with >25% of host predictions confirmed by both phage- and host-based methods, which explains the high number of high-confidence predictions (Fig. 4A). For all ecosystems, iPHoP provided host prediction for both temperate and virulent phages, although a higher percentage of predictions was obtained for temperate ones (Supplementary Fig. S10). While these results reflect the inherent bias in current microbial and phage reference databases, they suggest that iPHoP is already useful across different biomes and for different virus types, and may be expected to improve as more of the global microbial and viral diversity is characterized.

Within host-based approaches, nearly half (mean: 45%) of the predictions were based on genomes of uncultivated bacteria and archaea, highlighting the value of using large databases including single-cell amplified genomes (SAGs) and/or metagenome-assembled genomes (MAGs)[34,36] (Fig. 4C). These genomes from uncultivated microbes were particularly important for predicting hosts of environmental phages, especially freshwater and marine phages (Fig. 4C). We next wondered what proportion of these hosts were "local", i.e., assembled from the same sample as the query phage or another sample in the same study. Overall, in several ecosystems, a substantial (>25%) proportion of MAGs used for host predictions were obtained from metagenomes generated in the same study from which the input phage was derived (Fig. 4D). Hence, for comprehensive host prediction of a new phage dataset, it may be valuable to also integrate into the host genome database additional bacterial and archaeal MAGs obtained from the same sample or experiment, if available. To facilitate this, we included an automated database building module in iPHoP, enabling users to add their own MAGs in a host database based on phylogenies and taxonomies generated through GTDB-tk[40].

Estimating host diversity coverage by metagenome-derived viruses

Finally, we evaluated these IMG/VR host predictions from the host perspective, specifically assessing which host taxa were most frequently associated with viruses, and how much of the bacterial and archaeal diversity remained without any known or predicted virus. Overall, across the 5,711 bacteria and archaea genera with at least 2 genomes in the host database, 205 (3.6%) were associated with at least one reference virus in NCBI RefSeqVirus, while 1,700 (31.5%) were exclusively associated with metagenome-derived virus(es) through iPHoP (score ≥90, Fig. 5A). These host genera

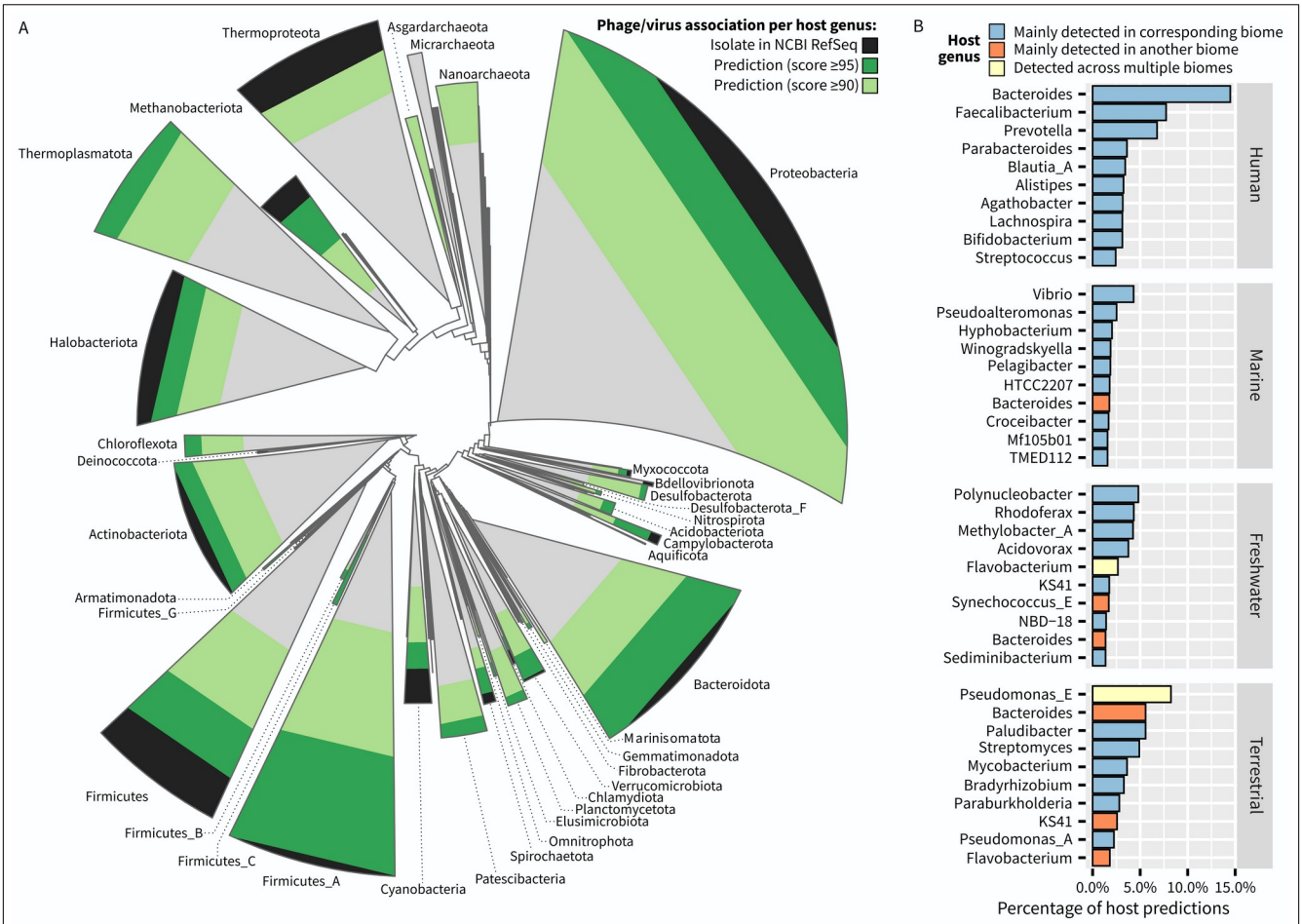

Figure 5. Taxonomic and environmental distribution of hosts predicted using iPHoP from the IMG/VR v3 genomes. *A. Archaeal (top left) and bacterial (bottom right) genome diversity from the GTDB database r202[34]. The GTDB phylogenetic trees were collapsed at the phylum level. The status of virus association, i.e., isolated virus, predicted virus only at score ≥ 95 or ≥ 90, or no prediction, was evaluated for each host genus, and the phyla shapes are colored according to the number of genera in each category within this phylum. B. For each major biome type, the 10 host genera with the highest number of predicted IMG/VR high-quality virus genomes are included in the plot. Each host genus was also determined to be mainly detected in a biome type or detected across multiple biomes based on the distribution of MAGs assigned to this genus across ecosystems in the GEM catalog (see Methods).*

only associated with viruses through iPHoP predictions were found across various bacterial and archaeal phyla, from *Firmicutes* and *Bacteroidota* to *Methanobacteriota* (31-48% of genera with host prediction only; Fig. 5A), and were not systematically associated with the largest genera, i.e. the ones with the highest number of species (Supplementary Fig. S11). Meanwhile, other phyla such as *Patescibacteria*, *Planctomycetota*, *Acidobacteriota*, and *Chloroflexota*, still displayed a majority of genera without any associated virus, either isolated or predicted (79-83%), highlighting the large diversity of viruses likely still to be identified and characterized.

We also evaluated which host taxa were associated with the largest number of predicted viruses for each biome reasoning that, if the predictions were mostly correct, these should correspond to taxa that are frequently observed in these ecosystems. Overall, the 10 genera most frequently predicted as hosts in each ecosystem did indeed correspond to taxa primarily detected in these same biomes, e.g., *Bacteroides* and *Faecalibacterium* for human microbiome, *Vibrio* and *Pseudoaltermomonas* for marine samples, and *Streptomyces* and *Mycobacterium* for terrestrial samples (Fig. 5B). The main exception to this pattern was the unexpectedly high number of host predictions to the *Bacteroides* genus for marine,

freshwater, and terrestrial viruses. As the Bacteroides-infecting *Crassvirales* phages[41–43] have been used as markers for fecal contamination[44,45], these predictions might reflect pervasive contamination of these environments, although these may also reflect a bias in the current phage and host isolate databases, skewing predictions towards this host genus. Overall, while these results illustrate how *in silico* host predictions must always be considered critically and in light of the current limitations of databases and tools, increasing the diversity of isolated phage-host pairs from various environments will likely help refine these predictions in the future.

Discussion

Viral metagenomics has profoundly transformed our understanding of global viral diversity and viral impacts on microbial communities. One critical piece of information missing compared to isolated viruses is the host connection, which significantly limits the inference and biological knowledge extracted from viromics data[4]. Accordingly, different methods have been developed to address this critical challenge, each with their specific limitation. Here, we present the iPHoP framework as a way to automatically integrate results from multiple host prediction approaches, which enables reliable prediction of host genus for a larger diversity of phages than any previous tool. The iPHoP tool and database are available as a stand-alone tool (bitbucket.org/srouxjgi/iphop/), a Bioconda recipe (https://bioconda.github.io/recipes/iphop/README.html), and a Docker container (https://hub.docker.com/r/simroux/iphop).

While iPHoP substantially improved host predictions on viruses from real metagenomic datasets, several limitations remain. First, because it relies on a suite of different tools, iPHoP remains relatively slow compared to other tools: a full iPHoP host prediction takes ~12 minutes for a test set of 5 complete phage genomes using the Sept_2021_pub database and 6 CPUs. This running time may not be problematic for viromics studies which typically run host prediction only once on a large set of metagenome-derived virus genomes, but it makes iPHoP suboptimal for time-sensitive analyses. Second, while iPHoP scores are designed to reflect false-discovery rates, these estimations depend on the composition of the test dataset used. Even though we tried to use a balanced set as much as possible by ensuring that we included viruses with a broad range of relatedness to reference sequences, iPHoP scores should only be interpreted as approximated FDRs at best. Third, since iPHoP was designed with a viral ecology framework in mind, our goal was to provide reliable host predictions at the genus rank, i.e., with FDRs ideally <10%, from diverse input phages. Arguably, in other contexts such as phage therapy applications, host predictions will need to be more specific and reach the host species or strain level. Such a high-resolution host prediction will likely require the reconstruction of detailed virus-host networks, as attempted by VirHostMatcher-Net[31], or detailed analysis of receptor-binding proteins[18]. In the near future however, we anticipate that genus-level approaches like iPHoP will be broadly applicable and provide host predictions for a large range of viruses, while higher resolution approaches such as VirHostMatcher-Net will likely be more limited in scope, so that both types of tools will be useful for different applications. Fourth, several potential improvements to iPHoP can already be envisioned, including for instance the addition of complementary approaches such as the detection of shared tRNA between phages and hosts, or the consideration of additional features such as whether the input virus is temperate or virulent. Finally, iPHoP remains limited by host, virus, and host-virus databases, as illustrated by the difference in the number of phages with host prediction between the human microbiome and other ecosystems. Achieving similar performance across all biomes will require in particular expanding the catalog of potential host genomes, with a particular attention paid to CRISPR arrays which are often not fully assembled from metagenomes[28,46], and expanding the diversity of viruses associated with a host, either from isolation or using *in vitro* host linkage[47–50]. In that context, to accommodate future expansions of the tool set and databases, iPHoP was intentionally designed as a

modular framework, and we envision the current tool as only the first step towards a comprehensive automated *in silico* host prediction toolkit.

Acknowledgments

BED was supported by the European Research Council (ERC) Consolidator grant 865694: DiversiPHI, and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2051 – Project-ID 390713860. FHC was supported by a Juan de la Cierva - Incoporación fellowship (Grant IJC2019-039859-I). This work was supported by the U.S. Department of Energy, Office of Science, Biological and Environmental Research, Early Career Research Program (SR) awarded under UC-DOE Prime Contract DE-AC02-05CH11231. The work conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231.

Online Methods

Virus-host training sets and host databases

To evaluate different host prediction approaches and train new classifiers, a curated dataset of known virus genomes with corresponding host taxonomy was established based on genomes available in the NCBI databases up to January 2021. For training new classifiers, sequences from bacteriophages and archaeoviruses were obtained from NCBI RefSeq release 201 (released in July 2020), and the host genus of each virus was obtained from the corresponding genome annotation and/or publication[51]. This dataset was used to train new classifiers (see below), but not to evaluate any tool since these virus-host pairs were likely to have been used for training in previously published tools as well.

To complement this training set, a distinct test set was established based on NCBI GenBank. Specifically, INPHARED[32] was used to download a collection of bacteriophages and archaeoviruses from NCBI in January 2021, and all genomes already present in NCBI RefSeq release 201 were removed. For the remaining ones, host taxonomic information was obtained from the corresponding annotation and/or publication, and genomes for which host taxonomy was uncertain were removed, leading to a final dataset of 1,870 viruses with host taxonomy (Table S2). These genomes were compared to the NCBI RefSeq references (see above) as well as the phage reference database used in RaFAH v0.1[25] using diamond blastp v0.9.24 (default parameters, [52]) after de novo prediction of cds using Prodigal v2.6.3 (option "-p meta", [53]), and the AAI estimation script provided with the Metagenomic Gut Virus catalogue (https://github.com/snayfach/MGV/blob/master/aai_cluster/README.md, [33]). Temperate phages were identified in the test set based on the annotation provided with each genome by searching for the keywords "prophage", "provirus", "lysogen", and "integrated", and based on BACPHLIP v0.9.6[54] with a minimum score of ≥0.8. When annotation and BACPHLIP prediction were conflicting, the information from the genome annotation was prioritized. Virulent phages were identified based on BACPHLIP v0.9.6[54] with a minimum score of ≥0.8.

Host database consolidation

The host genome database currently used in iPHoP, named "iPHoP_db_Sept21", was built from three publicly available genome sets, namely the GTDB database (release 202, [34]), published genomes from the IMG database (as of July 7, 2021, [35]), and the Genomes from Earth's Microbiomes (GEM) catalog[36], as follows. First, the 47,894 representative genomes from each GTDB species cluster were obtained from the GTDB database itself. Next, bacteria and archaea genomes from the IMG database that were not already included in GTDB release 202 and with a total length ≥100kb (n = 22,188), and medium- and high-quality metagenome-assembled genomes from the GEM catalogue (n = 52,515), were compared to the GTDB species representatives using the ani_rep function from GTDB-tk v1.5.0 (default parameters, [40]), based on Mash version 2.3[55] and FastANI v1.32[56]. All genomes with a similarity of ≥99% ANI over ≥99% AF were considered as identical to one of the GTDB representatives and discarded (n = 1,724). Non-identical genomes with a similarity of ≥95% ANI to one of the GTDB representatives were retained in the database as members of the corresponding species cluster (n = 32,735). Finally, the remaining genomes (n = 27,279) were considered as potential representatives of additional species clusters. To include these in a GTDB-compatible phylogenomic framework, these genomes were first checked for quality using CheckM v1.1.3[57], discarding all genomes with <50% completeness or >10% contamination, and then dereplicated with dRep v3.2.2[58] with cutoffs of 90% ANI and 60% coverage. The non-redundant genomes (n = 13,658) were then integrated in updated bacteria and archaea phylogenomic trees using the function de_novo_wf from GTDB-tk v1.5.0 (default parameters, [40]). The resulting trees are then used in iPHoP for taxonomic assignation of and phylogenetic distance estimation between all representatives.

The representative genomes included in the GTDB-tk-generated trees are used in iPHoP for all prediction methods (n = 60,000, i.e. 47,894 existing GTDB representatives and 12,106 additional ones from IMG and GEM). For blast-based prediction, these representative genomes are supplemented with additional genomes clustered into one of these species clusters, after removing duplicate genomes (n = 43,022, for a total of 103,022 genomes used). Finally, for CRISPR-based predictions, CRISPR spacers were predicted *de novo* in all 121,781 genomes (i.e., representatives, clustered, and duplicates), with CRT 1.2[59] and PilerCR[60] using custom python scripts from https://github.com/snayfach/MGV/tree/master/crispr_spacers [33]. All spacer sequences from arrays with ≥3 spacers were collected and dereplicated (100% identity), and spacers with a sequence length <10 or >100 were excluded. Ultimately, the spacer collection used in the iPHoP_db_Sept21 database includes 1,398,130 spacers from 40,036 distinct genomes.

Evaluation and benchmarking of selected host prediction methods

A set of published tools performing host predictions based on a single approach were selected for benchmarking and potential inclusion in the iPHoP integrated framework (Table S1). All these tools were benchmarked against the same test dataset (see above) established from virus sequences from NCBI GenBank (January 2021). Blast-based predictions were based on a blastn comparison (v2.12.0+, maximum e-value 1e-3, minimum identity percentage 80, maximum target sequence 25,000, minimum hit length 500nt, [61]) between the input virus genomes and the iPHoP_db_Sept21 blast database (see above). Metrics considered for each pair of input virus and host contigs were total number of matches and average identity percentage. CRISPR-based predictions were based on a blastn comparison (v2.12.0+, word size 7, no filtering of hits based on low complexity, i.e., "-dust no", maximum target sequence 10,000,000, [61]) between the input virus genomes and the iPHoP_db_Sept21 spacer database (see above), considering only hits to spacers 25 nucleotides or longer, with less than 8 mismatches overall, and with a custom complexity score <0.6. The custom spacer complexity score was calculated based on sequence AT skew content and the complexity estimation by Wootton–Federhen (CWF)[62] as follow: the complexity score is set as (CWF score – 2) * 2, except if (AT skew) > 0.65, in which case the complexity score is set to (AT skew) + 0.1. For Fig. 1A and Fig. S2, only hits with 2 or less mismatches over the entire spacer were considered. The metric considered to rank hits for individual input virus genomes was the total number of mismatches when considering the entire spacer. SpacePHARER predictions were based on the predictmatch function from SpacePHARER v2.fc5e668[12] applied to the input virus genomes with a sensitivity of 7.5 ("-s 7.5") and a maximum number of results per query sequence of 10,000, using the IMG/VR v3 CRISPR database[8]. SpacePHARER "Combined score" metric was used to rank predicted hosts for each input, with a minimum score cutoff of 20 applied for Fig. 1A and Fig. S2.

For WIsH predictions, input virus genomes were compared to the iPHoP_db_Sept21 WIsH database with WIsH v1.0[13] and a maximum p-value of 0.2. The WIsH p-value was also used to rank predictions for each input virus. For predictions based on the s_2* similarity, the corresponding code from VirHostMatcher-Net (July 2021 version[31]) was used to compare input virus genomes to the VirHostMatcher database in iPHoP_db_Sept21 (see above). The s_2* similarity score is the only metrics considered for each hit. For PHP, input virus genomes were compared to the iPHoP_db_Sept21 PHP database (see above) using PHP (July 2021 version[16]), and the PHP score was used as a metric for each hit. The upset plot comparing the predictions obtained for different host-based tools was generated with the UpSetR package[63].

RaFAH[25] predictions were obtained by running the "predict" function from RaFAH v0.3 on the input virus genomes with default parameters, and using the Predicted_Host_Score as metric. vHULK[26] predictions were obtained by running vHULK v1.0.0 with default parameters, and using

score_genus_relu as the metric. VPF-Class predictions were obtained by running the vpf-class function from the vpf-tools 0.1.0.0 toolkit[27] with default parameters, using the host taxon with the highest membership_ratio as the prediction for each input virus and the confidence_score as the metric. Finally, HostPhinder predictions were obtained by running the "latest" version of HostPhinder docker container (December 2015) with default parameters, and the main reported score as metric.

For all the tools, the best prediction was taken for each input virus based on the relevant metric, and considered as correct if the genus of the predicted host genome or the predicted genus for tools predicting host taxonomy was consistent with the information collected from the reference database (Table S2).

Establishment of balanced training sets for single-tool iPHoP classifiers

For the 5 host-based approaches selected to be included in the iPHoP framework ("Blast", "CRISPR", "WIsH", "VirHostMatcher", and "PHP"), individual machine-learning classifiers taking into consideration multiple top hits for each input virus were optimized as follows. A training set was built from the hits obtained from NCBI Virus RefSeq release 201[51] against the iPHoP_db_Sept21 database, using similar cutoffs as for the benchmarks (see above) but considering for each input virus the 50 best hits (blast) or 30 best hits (all other methods). All hits were associated with the corresponding host genome representative (see "Virus-host training sets and host databases" above), and for each input virus, all host genome representatives with one hit were considered as a candidate host.

For each pair of input virus-candidate host, the different hits obtained for this virus were gathered as follows. First, the phylogenetic distance between the host genome representative of each hit and the candidate host was obtained from the GTDB-tk-generated trees (see above), so that hits can be ordered by distance to the candidate host. Next, depending on the tools, one to three scores were used to describe the strength of the hit, and all the hits for a given input virus are tallied, i.e., the number of hits observed for a given distance and set of scores is tabulated. The resulting matrices then serve as input to either neural network or random forest classifiers. For more detailed information about the cutoffs, score selection, and transformation used for each tool, please see Supplementary Fig. S5.

For classifier training, a subset of 20,000 to 60,000 virus-host instances were randomly selected for each tool, with the following constraints: (i) between 60 to 85% of incorrect virus-host pairs, i.e., instances where the candidate host was assigned to a genus different from the host genus listed for this virus in the database, and (ii) between 45 to 70% of instances with a "known" virus, i.e., for which the virus had an AAI percentage of 70% or higher to the closest reference. These constraints were included to ensure that the training set was not too unbalanced in favor of (i) incorrect predictions, since most hits are to genomes from a different genus than the host, and (ii) "known" viruses, which typically represent the majority of databases and could bias the classifiers. A subset (10%) of these training data were set aside and used as a common validation set when comparing different versions of each classifier (see below). Further, for each instance, 3 different sets of hits were used: one including all the hits obtained for the virus, one including only a random subset (from 0 to 100%) of hits, and one including only one randomly selected hit with a distance ≤4 to the candidate host (if any) and all hits with a distance >4, or one randomly selected hit among all hits if all display a distance >4. This random subsampling of hits was included to simulate different levels of representation of host diversity in the database, since current bacteria and archaea genome databases do not provide an even coverage of the global diversity, and isolated viruses used here for training are likely to be biased towards well-represented hosts.

Optimization and evaluation of single-tool classifiers for iPHoP

All dense and convolution networks were built using TensorFlow 2.7.0[64], and all random forest classifiers were built with the TensorFlow Decision Forests v0.2.1, both within the Keras 2.7.0 Python library[65]. Classifiers were trained on the corresponding training set, using 80% of the data for training and 20% for validation ("validation_split=0.2" for the neural networks). The Adam optimizer was used to train all the neural networks. Classifier parameters including the number of layers, kernel size, and dilation rate for convolution networks, number of dense layers for dense networks, and number of trees and maximum tree depth for random forest classifiers, were optimized for each individual classifier using the Optuna v2.5.0 framework[66], by running 100 training trials (see Supplementary Fig. S5). For each type of classifier (convolution network, dense network, random forest classifier), the 5 best versions based on minimum Binary Cross Entropy loss (for networks) or maximum accuracy (for random forests) on the common validation set (see above) were selected as potential candidates.

To select the optimal combination of classifiers, these candidates were then applied to the test set, and the results obtained on the non-ambiguous cases were observed (i.e.: blast hit ≥10kb, CRISPR match with 0 mismatches, WIsH p-value ≤1E-05, VHM score ≥0.8, PHP score ≥1450). For each classifier, the tenth percentile of scores for these non-ambiguous cases where the classifier prediction was correct was used as an estimate of a "high-confidence" score for this classifier, and the number of incorrect predictions with a score higher than this cutoff was used as an estimate of the error rate, i.e., incorrect prediction with a score comparable to non-ambiguous correct predictions. This error rate was then used to iteratively select classifiers by first selecting the one with the lower error rate, then selecting additional classifiers if they provided ≥5% additional correct prediction among non-ambiguous cases, or if they "corrected" ≥10% of the previous false-positive errors. If no classifier fulfilled these conditions, the selection process was stopped. Ultimately, all selected classifiers (See Table S3) were run on the full test set to derive Precision-Recall curves and False Discovery Rate estimations.

Training, optimization, and evaluation of iPHoP main Random Forest classifier (iPHoP-RF)

To integrate signal from multiple approaches, a random forest classifier ("iPHoP-RF") was trained to obtain a single confidence score for a given virus-candidate host pair based on the score obtained for all individual classifiers selected (see Supplementary Table S3). Specifically, for each virus-candidate host pair used in the training set (see above), the following information were included for each selected classifier: the score obtained for the virus-candidate host pair, the rank of this pair among all candidate hosts considered for this given virus, and the difference between the score of the pair and the highest score obtained for the given virus. This led to a final input matrix with 30 columns, i.e., 3 features (score, rank, distance to best score) for each of the 10 selected classifiers. A balanced training set was built from the training sets created for each individual classifier (see above), including 700 randomly sampled viruses with at least 1 blast hit and 1 CRISPR hit, 700 each from viruses with either at least 1 blast hit or 1 CRISPR hit, and 700 viruses with neither blast or CRISPR hits. For each selected virus, up to 10 correct and up to 5 incorrect predictions (i.e., candidate virus-host pairs) were randomly selected. Eventually, the balanced training set included 17,105 correct and 13,960 incorrect virus-candidate host pairs.

Random Forest Classifiers were built using the TensorFlow Decision Forests v0.2.1[64] package within the Keras 2.7.0 python library[65], with parameters optimized with the Optuna v2.5.0 framework[66]. Parameters to be optimized included maximum tree depth (between 4 and 32), minimum number of examples in a node (between 2 and 10) and number of trees (between 100 and 1,000). A total of 100 trials were performed, each was evaluated on the test dataset, the 5 classifiers with the highest accuracy were selected as the best candidates, and the candidate with the highest recall at 5% FDR was then selected as the final iPHoP-RF classifier.

Integrating iPHoP classifiers and RaFAH into a final host prediction

In order to rank host predictions for individual viruses obtained with different methods, and since the scores from different classifiers are not directly comparable, the test dataset was used to transform raw scores into empirical false-discovery rates (FDRs). Specifically, the positive predictive value (PPV), i.e., the number of correct predictions divided by the total number of predictions, which corresponds to 1 minus the false-discovery rate, was computed on sliding windows of each tool score from 0 to 1, with window size 0.05 (for Blast Conv_87, Blast RF_39, CRISPR Conv_85, CRISPR Dense_15) or 0.01 (for iPHoP-RF, and RaFAH). For each tool, a generalized linear model was then fitted on these values using the mgcv v1.8-36 library[67] in R v 4.0.5[68] with REML estimation, and an empirical PPV and FDR was then calculated for scores ranging from 0 to 1 by steps of 0.001.

These empirical positive predictive values are then used in the iPHoP framework to derive a single composite score for each virus-candidate host genus pair as follows. For each pair, all methods with PPV <0.5 are first discarded. Next, the method with the highest PPV, i.e., the lowest FDR, for this pair is selected as the source of the main FDR. To take into account prediction from other methods which passed the PPV threshold, i.e. were ≥0.5, the FDR from these additional predictions are then multiplied by 2 (to rescale between 0 and 1), and the final composite score is then calculated as 1 minus the product of the main FDR and the additional "rescaled" FDRs, if any. This means that additional methods pointing to the same virus-host genus pair can only improve the composite score, as they will multiply the main FDR by factors always ≤1. Finally, a similar empirical approach based on the test dataset was used to transform these composite scores in PPVs (see above), and these empirically estimated PPVs are provided to iPHoP users as "Confidence score" in the result files. By default, only predictions with a confidence score ≥90, i.e. an estimated FDR <10%, are included in the summary output file, however users can select any confidence score ranging from 75 to 100.

To enable this integration of results from host-based tools and RaFAH, the predictions from RaFAH had to be converted into GTDB-compatible taxa. To this end, each genus listed in the RaFAH output file was searched for in the GTDB metadata files, and the list of genomes associated with this RaFAH genus along with their GTDB genus-level taxon was tallied. Each RaFAH genus was then associated to all GTDB genus-level taxa representing ≥50% of the genome list if the list included <10 genomes, ≥20% of the genome list if the list included 10 to 100 genomes, or ≥10% if the list included ≥100 genomes. This approach provided GTDB genus-level taxa for 595 RaFAH genera, with 492 linked to a single taxa, and 90 linked to 2 taxa, often closely related (e.g., "Thioalkalivibrio" and "Thioalkalivibrio_B", "Pseudothermotoga_A" and "Pseudothermotoga_B", etc).

Comparison to other integrated host prediction approaches

Three other tools providing host prediction based on multiple approaches were benchmarked on the same test dataset (see above) as iPHoP. VirHostMatcher-Net (July 2021 version[31]) was run on the test dataset with default parameters, requesting the top 100 predictions to be included in the output files, and using the default host database provided with the tool. PhisDetector (February 2021 version[30]) was run on the test dataset with the following parameters: "--min_mis_crispr 2 --min_cov_crispr 70 --min_per_prophage 30 --min_id_prophage 70 --min_cov_prophage 30 --min_PPI 1 --min_DDI 5 --min_per_blast 10 --min_id_blast 70 --min_cov_blast 10", and using the default database provided with the tool. Finally, VirMatcher v0.3.2[29] was run on the test dataset via it KBase App[69]. Since no host database was provided with VirMatcher, a custom host genome database was built based on the RefSeq genomes that displayed at least one hit to any of the test dataset virus with blast, CRISPR, or WIsH.

For each tool, the prediction with the highest score was considered as the host genus predicted for a given virus, excluding predictions to hosts with unknown genera. These "best hit" predictions were

then used to evaluate the recall of each tool, i.e., the number of correct host genus predictions, at different false discovery rates level, either on the complete test dataset or when restricting to specific ranges of "novelty", i.e. AAI to the closest reference ranging from 0 to 5%, 5 to 10%, 10 to 30%, 30 to 60%, or 60 to 100%.

Evaluation of iPHoP host predictions on high-quality genomes from the IMG/VR database

To evaluate iPHoP on real metagenome-derived virus genomes, 216,015 high-quality genomes from the IMG/VR v3 database[8], i.e. metagenome-derived viral genomes estimated to be ≥90% complete based on CheckV v0.4.0[70], were processed with iPHoP v1.0, using the iPHoP_db_Sept21 database. Host genus prediction was based on the host genus with the best iPHoP composite score for each input sequence, with a minimum score cutoff of 75. Metadata for IMG/VR sequences, including corresponding study and dataset if available, were obtained from the IMG/VR database (2020-10-12_5.1 version)[8]. Temperate and virulent phages were identified based on BACPHLIP v0.9.6[54] with a minimum score of ≥0.8. Metadata for the host genome, including the corresponding study and dataset if available, were obtained from the IMG and Gold databases (information downloaded in Jan. 2022[35,71]).

To represent the diversity of hosts included in these IMG/VR-derived host predictions, the GTDB bacteria and archaea trees were plotted using the ggtree v2.4.1 package[72], with clades collapsed at the phylum level. Each phylum was then colored according to the status of its member genera, i.e., whether each host genus is associated with an isolated virus in RefSeq, a host prediction with a score ≥95, a host prediction with a score ≥90, or no isolate or host prediction. To verify whether iPHoP host predictions linked viruses from each main biome to host taxa consistently found in the same biomes, the GEM dataset[36] was used to evaluate the biome distribution of individual host genera. Specifically, each GEM MAG was associated to its corresponding genus and original sample biome, if this information was available (n = 38,556). Each genus was then associated with a given biome if ≥50% of the corresponding MAGs originated from a sample of this biome (n = 3,500 genera), or was considered as "Detected across multiple biomes" if the majority biome represented <50% of the genus MAGs (n = 90 genera).

Supplementary Figure

Supplementary Figure S1. Characteristics of the test dataset. A. Distribution of the host genera for the test dataset. Note: only genera associated with ≥5 viruses are included, another 125 host genera were associated with <5 viruses and are not displayed. B. Distribution of AAI to the closest reference in NCBI RefSeq for the test dataset. The corresponding list of viral genomes included in the test dataset is provided in Table S2.

Supplementary Figure S2. Comparison of different host prediction approaches on different subsets of the test dataset. Total number of predictions and number of correct predictions (y-axis) obtained at any rank for each tool (x-axis) on sequences from the test dataset (Table S2). For each tool, the number of correct predictions is indicated by the colored bar, while the total number of predictions is indicated by the gray bar. The top panel displays the results obtained on the entire test dataset (n = 1,870). The middle panel includes results obtained for all phages predicted as temperate, either via BacPhlip or based on the genome annotation (n = 949). The middle panel includes results obtained for all phages predicted as virulent by BacPhlip (n = 663).

Supplementary Figure S3. Characteristics of the high-quality IMG/VR genomes. A. Number of high-quality viral genomes from IMG/VR v3 identified across the 5 major biomes in the database. Genomes sampled from other biomes of lacking a biome information are gathered in the "Other" category. B. Distribution of the average amino-acid identity between IMG/VR v3 viral genomes and the NCBI Viral RefSeq v203.

Supplementary Figure S4. Overlap between host-based tools for individual viruses. For each host-based tool included in the benchmark (see Fig. 1), the overlap in terms of input sequence for which a correct prediction was obtained is presented here as an upset plot. The intersection size represents the number of phages with correct prediction using the combination of methods indicated at the bottom. This number is also indicated above each bar, and the bar color indicates the number of tools included in the combination.

Supplementary Figure S5. Schematic of the data transformation and classifier architectures used in iPHoP. A. Summary of the cutoff and metrics used for each host-based tool considered in iPHoP (see Table S1). B. Overview of the three different types of classifiers evaluated in iPHoP. The different parameters optimized using the Optuna framework are highlighted in blue. For varying numbers of layers, the same parameters were optimized for each layer, but each was optimized separately, i.e., the parameters values were independent between the different layers.

Supplementary Figure S6. ROC and Precision-Recall curves for single-tool classifiers. For each host-based tool the ROC curves (left) and Precision-Recall curves (right) based on the test dataset are presented for the 5 best classifiers of each type, and compared to the "naive" approach, i.e. best hit based on the raw score. TPR: True Positive Rate. FPR: False Positive Rate. PPV: Positive Predictive Value. The 1-to-1 line is indicated as a dashed black line on the ROC curves. Random Forest Classifiers were only evaluated for Blast and CRISPR approaches.

Supplementary Figure S7. Percentage of correct host predictions obtained for viruses with different degrees of "novelty". The number of correct host predictions was evaluated for 3 different score cutoffs corresponding to 20%, 10%, and 5% estimated FDR (False Discovery Rate). Input viruses

were classified into 5 categories (x-axis) based on their AAI (Average Amino Acid Identity) to the closest reference phage genome. The number of correct host predictions is indicated for each iPHoP classifier (see Fig. 3A), and for the composite score considering all classifiers ("combined").

Supplementary Figure S8. Comparison of different integrated host prediction tools, including iPHoP, on the test dataset. Standard Receiver Operating Characteristic (left) and Precision Recall (middle) curves for the 4 integrated host prediction approaches compared. To take into account the number of predictions provided by each tool, a third plot (right panel) indicates the positive predictive value (y-axis) when considering an increasing number of predictions (x-axis). To obtain this, cutoffs were progressively lowered to include an increasing number of predictions for each tool, and prioritize the highest confidence ones, i.e. starting with the highest PPV possible. For the ROC curve, a 1-to-1 line is indicated with a dashed black line. For the Precision Recall and PPV curves (middle and right panels), the red and purple dashed lines indicate 5% and 10% False Discovery rates, respectively.

Supplementary Figure S9. Type of host prediction obtained for high-quality IMG/VR v3 genomes with different degrees of "novelty". High-quality genomes from the IMG/VR v3 database for which a host prediction was obtained with iPHoP (score ≥90) were binned based on the average amino acid identity (AAI) to the closest reference in NCBI RefSeq Virus r203 (x-axis). Predictions entirely based on host-based tools are indicated as "Host only", predictions exclusively based on RaFAH are indicated as "Phage only", and predictions where both types of tools were consistent and with score ≥90 are listed as "Both".

Supplementary Figure S10. Breakdown of iPHoP host predictions for high-quality IMG/VR v3 genomes assigned as virulent (top) or temperate (bottom). Similar as Fig. 4A and 4B, the left panel shows the distribution of the best score provided by iPHoP for the corresponding subset of IMG/VR v3 quality genome (top: virulent, bottom: temperate), organized by ecosystem. For each IMG/VR vOTU, the best score from iPHoP was considered if ≥75, or the vOTU was considered as not having a predicted host. The right panel shows the the type of signal used to achieve host prediction with a score ≥90. "Host-based" includes all 5 host-based tools, while "Phage-based" includes predictions obtained with RaFAH. "Both" includes consistent predictions obtained with RaFAH and at least one host-based tool. Temperate and virulent phages were identified via BACPHLIP[54] with a minimum score of 0.8 and based on genome annotation (see Methods).

Supplementary Figure S11. Number of species and iPHoP prediction per host genus. Each dot represents a host genus with at least 2 species, with the x-axis reflecting the total number of species in the genus, and the y-axis reflecting the total number of IMG/VR v3 HQ sequences predicted to infect this host genus with a score ≥ 90. Host genera and species were obtained from the GTDB database[34]. The right panel presents a zoomed-in version of the area highlighted with dashed black lines in the left panel.

Supplementary Tables

Supplementary Table S1. List of individual tools benchmarked, included in, and/or compared to iPHoP.

Supplementary Table S2. List of viral genomes included in the test dataset, obtained from NCBI GenBank, and used to evaluate the performance of individual and integrated tools.

Supplementary Table S3. Characteristics of the single-tool classifiers considered for inclusion in iPHoP. The classifiers eventually included in iPHoP v1.0 are indicated with a "x" symbol in the column "Classifiers selected for inclusion in iPHoP".

References

1. Fernández, L., Rodríguez, A. & García, P. Phage or foe: An insight into the impact of viral predation on microbial communities. *ISME J.* **12**, 1171–1179 (2018).

2. Correa, A. M. S. *et al.* Revisiting the rules of life for viruses of microorganisms. *Nat. Rev. Microbiol.* **0123456789**, 1–13 (2021).

3. Abeles, S. R. & Pride, D. T. Molecular bases and role of viruses in the human microbiome. *J. Mol. Biol.* **426**, 3892–3906 (2014).

4. Roux, S. *et al.* Minimum information about an uncultivated virus genome (MIUVIG). *Nat. Biotechnol.* **37**, 29–37 (2019).

5. Taş, N. *et al.* Metagenomic tools in microbial ecology research. *Curr. Opin. Biotechnol.* **67**, 184–191 (2021).

6. Sommers, P., Chatterjee, A., Varsani, A. & Trubl, G. Integrating Viral Metagenomics into an Ecological Framework. *Annu. Rev. Virol.* **8**, 133–158 (2021).

7. Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465–70 (2016).

8. Roux, S. *et al.* IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res.* **49**, D764–D775 (2020).

9. ter Horst, A. M. *et al.* Minnesota peat viromes reveal terrestrial and aquatic niche partitioning for local and global viral populations. *Microbiome* **9**, 1–18 (2021).

10. Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol. Rev.* **40**, 258–272 (2016).

11. Coclet, C. & Roux, S. Global overview and major challenges of host prediction methods for uncultivated phages. *Curr. Opin. Virol.* **49**, 117–126 (2021).

12. Zhang, R. *et al.* SpacePHARER: sensitive identification of phages from CRISPR spacers in prokaryotic hosts. *Bioinformatics* **37**, 3364–3366 (2021).

13. Galiez, C. *et al.* WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* **33**, 3113–14 (2017).

14. Ahlgren, N., Ren, J., Lu, Y. Y., Fuhrman, J. A. & Sun, F. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res.* **45**, 39–53 (2016).

15. Liu, D., Ma, Y., Jiang, X. & He, T. Predicting virus-host association by Kernelized logistic matrix factorization and similarity network fusion. *BMC Bioinformatics* **20**, 1–10 (2019).

16. Lu, C. *et al.* Prokaryotic virus Host Predictor: A Gaussian model for host prediction of prokaryotic viruses in metagenomics. *BMC Biol.* **19**, 1–11 (2021).

17. Leite, D. M. C. *et al.* Computational prediction of inter-species relationships through omics data analysis and machine learning. *BMC Bioinformatics* **19**, (2018).

18. Boeckaerts, D. *et al.* Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins. *Sci. Rep.* **11**, 1–14 (2021).

19. Tan, J. *et al.* HoPhage: an ab initio tool for identifying hosts of phage fragments from metaviromes. *Bioinformatics* **38**, 543–545 (2022).

20. Li, M. & Zhang, W. PHIAF: Prediction of phage-host interactions with GAN-based data augmentation and sequence-based feature fusion. *Brief. Bioinform.* **23**, 1–10 (2022).

21. Zielezinski, A., Deorowicz, S. & Gudyś, A. PHIST: Fast and accurate prediction of prokaryotic hosts from metagenomic viral sequences. *Bioinformatics* **38**, 1447–1449 (2022).

22. Ruohan, W., Xianglilan, Z., Jianping, W. & Shuai Cheng, L. I. DeepHost: Phage host prediction with convolutional neural network. *Brief. Bioinform.* **23**, 1–10 (2022).

23. Shang, J. & Sun, Y. CHERRY: a Computational metHod for accuratE pRediction of virus–pRokarYotic interactions using a graph encoder–decoder model. *Brief. Bioinform.* 1–16 (2022) doi:10.1093/bib/bbac182.

24. Villarroel, J. *et al.* HostPhinder: A phage host prediction tool. *Viruses* **8**, 116 (2016).

25. Coutinho, F. H. *et al.* RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content. *Patterns* **2**, (2021).

26. Amgarten, D., Vázquez Iha, B. K., Piroupo, C. M., da Silva, A. M. & Setubal, J. C. vHULK, A new tool for bacteriophage host prediction based on annotated genomic features and deep neural networks. *bioRxiv* 0–15 (2020) doi:10.1101/2020.12.06.413476.

27. Pons, J. C. *et al.* VPF-Class : taxonomic assignment and host prediction of uncultivated viruses based on viral protein families. *Bioinformatics* 1–9 (2021) doi:10.1093/bioinformatics/btab026.

28. Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* **3**, 870–880 (2018).

29. Gregory, A. C. *et al.* The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host Microbe* **28**, 724-740.e8 (2020).

30. Zhang, F. *et al.* PHISDetector: a tool to detect diverse in silico phage-host interaction signals for virome studies. *bioRxiv* 1–20 (2020) doi:10.1101/661074.

31. Wang, W. *et al.* A network-based integrated framework for predicting virus–prokaryote interactions. *NAR Genomics Bioinforma.* **2**, 1–19 (2020).

32. Cook, R. *et al.* INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes. *Phage* **2**, 214–223 (2021).

33. Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* **6**, 960–970 (2021).

34. Parks, D. H. *et al.* GTDB: An ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).

35. Chen, I. M. A. *et al.* The IMG/M data management and analysis system v.6.0: New tools and advanced capabilities. *Nucleic Acids Res.* **49**, D751–D763 (2021).

36. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0718-6.

37. Burstein, D. *et al.* Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nat. Commun.* **7**, 10613 (2016).

38. Shmakov, S. A., Wolf, Y. I., Savitskaya, E., Severinov, K. V. & Koonin, E. V. Mapping CRISPR spaceromes reveals vast host-specific viromes of prokaryotes. *Commun. Biol.* **3**, 1–9 (2020).

39. Zielezinski, A., Barylski, J. & Karlowski, W. M. Taxonomy-aware, sequence similarity ranking reliably predicts phage–host relationships. *BMC Biol.* **19**, 1–14 (2021).

40. Chaumeil, P., Mussig, A. J., Hugenholtz, P., Parks, D. H. & Hugenholtz, P. GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* **36**, 1925–1927 (2019).

41. Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **5**, 4498 (2014).

42. Shkoporov, A. N. *et al.* ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nat. Commun.* **9**, 1–8 (2018).

43. Yutin, N. *et al.* Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat. Microbiol.* **3**, 38–46 (2018).

44. Stachler, E. & Bibby, K. Metagenomic Evaluation of the Highly Abundant Human Gut Bacteriophage CrAssphage for Source Tracking of Human Fecal Pollution. *Environ. Sci. Technol. Lett.* **1**, 405–409 (2014).

45. Ahmed, W. *et al.* Evaluation of the novel crAssphage marker for sewage pollution tracking in storm drain outfalls in Tampa, Florida. *Water Res.* **131**, 142–150 (2018).

46. Skennerton, C. T., Imelfort, M. & Tyson, G. W. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Res.* **41**, e105 (2013).

47. Sakowski, E. G. *et al.* Interaction dynamics and virus–host range for estuarine actinophages captured by epicPCR. *Nat. Microbiol.* **6**, 630–642 (2021).

48. Tadmor, A. D. *et al.* Probing Individual Environmental Bacteria for Viruses by Using Microfluidic Digital PCR. *Science (80-.).* **333**, 58–62 (2011).

49. Ignacio-Espinoza, J. C. *et al.* Ribosome-linked mRNA-rRNA chimeras reveal active novel virus host associations. *bioRxiv* (2020) doi:10.1101/2020.10.30.332502.

50. Uritskiy, G. *et al.* Accurate viral genome reconstruction and host assignment with proximity-ligation sequencing. *bioRxiv* 2021.06.14.448389 (2021).

51. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).

52. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2014).

53. Hyatt, D. *et al.* Prodigal : prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, (2010).

54. Hockenberry, A. J. & Wilke, C. O. BACPHLIP: Predicting bacteriophage lifestyle from conserved protein domains. *PeerJ* **9**, (2021).

55. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).

56. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 1–8 (2018).

57. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

58. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* 1–5 (2017) doi:10.1038/ismej.2017.126.

59. Bland, C. *et al.* CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209 (2007).

60. Edgar, R. C. PILER-CR: Fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* **8**, 1–6 (2007).

61. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

62. Wootton, J. C. & Federhen, S. Analysis of Compositionally Biased Regions in Sequence Databases. *Methods Enzymol.* **266**, 554–571 (1006).

63. Gehlenborg, N. UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets. (2019).

64. Abadi, M. *et al.* {TensorFlow}: Large-Scale Machine Learning on Heterogeneous Systems. (2015).

65. Chollet, F. & others. Keras. (2015).

66. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* 2623–2631 (2019) doi:10.1145/3292500.3330701.

67. Wood, S. N., Pya, N. & Säfken, B. Smoothing parameter and model selection for general smooth models (with discussion). *J. Am. Stat. Assoc.* **111**, 1548–1575 (2016).

68. R Core Team. R: A Language and Environment for Statistical Computing. (2022).

69. Arkin, A. P. *et al.* KBase: The United States department of energy systems biology knowledgebase. *Nat. Biotechnol.* **36**, 566–569 (2018).

70. Nayfach, S. *et al.* CheckV: assessing the quality of metagenome-assembled viral genomes. *Nat. Biotechnol.* **in press**, 1–20 (2020).

71. Mukherjee, S. *et al.* Genomes OnLine database (GOLD) v.7: Updates and new features. *Nucleic Acids Res.* **47**, D649–D659 (2019).

72. Yu, G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr. Protoc. Bioinforma.* **69**, 1–18 (2020).