PHISDetector: a web tool to detect diverse *in silico* phage-host interaction signals

Fan Zhang[1, #], Fengxia Zhou[1, #], Rui Gan[1, #], Chunyan Ren[3], Yuqiang Jia[2], Ling Yu[1] and Zhiwei Huang[1,*]

[1] HIT Center for Life Sciences, School of Life Science and Technology, Harbin Institute of Technology, 150080 Harbin, China
[2] School of Computer Science and Technology, Harbin Institute of Technology, 150080 Harbin, China.
[3] Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, USA
* To whom correspondence should be addressed. Tel: +86-451-86403163; Fax: +86-451-86403163; Email: huangzhiwei@hit.edu.cn

* These authors contributed equally to this work.

ABSTRACT Phage-host interactions are appealing systems to study co-evolution. Their roles in human health and diseases as well as novel therapeutics development also have been increasingly emphasized. Meanwhile, such interactions leave signals in bacterial and phage genomic sequences, defined as phage-host interaction signals (PHIS), allowing us to predict novel phage-host interactions. Due to the intrinsic complexity and recent emerging of metagenomics sequencing data, there is an urgent requirement to develop computational tools to analyze massive data and extract meaningful information. Here, we seize comprehensive *in silico* PHIS and utilize sophisticated bioinformatics to develop PHISDetector, a web tool to detect and systematically study diverse *in silico* PHIS, including analyses for co-occurrence/co-abundance patterns, oligonucleotide profile/sequence composition, CRISPR-targeting, prophages, phage genome similarity, protein-protein interactions, and special gene check. PHISDetector accepts various genomic and metagenomic data as input and provides well-designed visualizations and detailed data tables to download. Prediction tasks are processed remotely by the server using custom python scripts and a series of public tools. PHISDetector can be accessed at http://www.microbiome-bigdata.com/PHISDetector/index/ .

INTRODUCTION

Phages not only play key roles in shaping the community structure of human and environmental microbiota, but also provide potential tools for precision manipulation of specific microbes. Recent studies further reveal that phage-microbe interactions likely impact aspects of mammalian health and disease (1). Therefore, it is critical to identify and fully understand these interactions, which may contribute to novel therapeutics, such as phage therapy, a promising strategy to combat multi-drug resistant infections. Molecular and ecological co-evolutionary processes of phages and bacteria leave various signals in their genomic sequences for the tracing of phage-host interactions (2) . In addition to experimental methods, recent advances in large-scale genomic- and metagenomic sequencing efforts and computational approaches have profoundly deepened our knowledge about microbe–phage interactions and advanced new challenges about investigating such interaction signals.

Phage-**h**ost **i**nteraction **s**ignals (PHIS) can be detected by identifying putative prophage regions from bacterial genomes, defined as integrated phages that insert their genomes into their hosts. Several *in silico* tools for prophage detection in sequenced genomes have been developed such as VirSorter (3), PHASTER (4), Prophinder (5), Phage_Finder (6), and PhageWeb (7). Oligonucleotide

profile analysis is a common used alignment-free method for PHIS detection. It is based on the observation that phages share highly similar genomic signatures (such as *k*-mer) with their hosts due to the fact that virus replication is dependent on translational machinery of its host (8). Software VirHostMatcher (9) and WIsH (10) were developed to predict hosts for virus genomes or even short viral contigs using alignment-free (dis)similarity measures. The third group of host prediction methods could be called similarity analysis, which is based on the observation that similar viruses often share the same host range. HostPhinder predicts the host of a query virus using virus-virus similarity measure defined as the proportion of the shared *k*-mers between the query and the reference virus genomes (11). Another program called GeneNet used the gene-based virus-virus similarity network to predict the host species or genus for a query virus (12). Co-abundance-based methods have also been used to detect the correlation of phage and bacterial abundance patterns in metagenomes to identify their association (13). CoNet (14) has been developed to carry out microbial network inference from sequencing data and used to infer a virus and bacterial co-occurrence network in Human Skin metagenomics data (15). In addition, CRISPR spacer sequences can also be used to infer host–phage interactions based on the principle that hosts incorporate spacer sequences from the phages that infect them (16-18).

Though various methods have been proposed for predicting phage-host interactions, accuracy is the limit when using a single type of *in silico* signal (2). With the exponentially increasing number of viruses uncovered, there is a huge demand for a tool capable to incorporate all types of PHIS and conveniently predict the hosts of viruses. However, to our knowledge, all these tools are limited to certain interacting features and there is no published web-server available for comprehensive prediction of global phage-host interactions based on diverse *in silico* PHIS. Here, we developed PHISDetector, which utilizes sophisticated bioinformatics methods to seize comprehensive *in silico* PHIS including co-occurrence/co-abundance patterns based on metagenome data, oligonucleotide profile/sequence composition, CRISPR-targeting, prophage, similarity between the phages sharing the same host range, specialty gene transferring and protein-protein interactions.

TOOL DESCRIPTION

PHISDetector workflow

PHISDetector receives bacterial or virus genomic sequences in GenBank or FASTA format as input. If a bacterial sequence has been submitted (Figure 1, upper left), prophage analysis, oligonucleotide profile analysis against a phage reference database, and CRISPR-spacer targeting analysis against the RefSeq Viral genome database will be carried out to predict the potential infecting phage. If a phage sequence has been submitted (Figure 1, upper right), similarity analysis, oligonucleotide profile analysis upon a bacterial reference database, and CRISPR analysis using a spacer reference database will be executed to predict the host. If a pair of bacteria-phage genome sequences has been submitted, diverse *in silico* PHIS will be detected to characterize the interaction, including k-mer usage (dis)similarity, spacer-protospacer matching, specialty genes check, and protein-protein interactions. Finally, a consensus analysis is performed to indicate the possible integrity of the

predicted interactions. Generically, for a FASTA input file, open reading frames (ORFs) will be firstly predicted on the input genome using FragGeneScan (19), while for a GenBank (GBK) file, DNA sequence and ORF amino acid sequences of the genome will be extracted directly from the input GBK file (Figure 1).

Analysis modules

PHISDetector is composed of three types of analysis modules that allow (i) identifying diverse *in silico* PHIS including co-occurrence/co-abundance analysis, oligonucleotide profile analysis, CRISPR analysis, prophage analysis, and similarity analysis; (ii) checking specialty genes including virulence factors (VFs) and antibiotic resistance genes (ARGs); and (iii) detecting protein-protein interactions between a pair of phage and bacteria genomes.

Oligonucleotide profile analysis. This module is used to predict the bacterial host of phages by examining various oligonucleotide frequency (ONF) based distance/dissimilarity using VirHostMatcher. For the prediction of the prokaryotic host of short viral contigs, an extra WIsH approach is provided. Note that when using the VirHostMatcher approach, an extra taxonomy file is required, so we also provide a tool for users to generate the taxonomy file by providing the NCBI accession IDs for their input bacterial genomes.

CRISPR analysis. The CRISPR spacer sequences are computationally identifiable sequence signature of previous phage-host infections. In this module three scenarios of analysis are supported: (i) Users can provide their input either as spacer sequences in (multi-)FASTA format, or as CRISPRFinder (20), PILER-CR (21) or Seq2CRISPR (22) output files, or a bacterial genome sequence for which the CRISPR spacers will be automatically identified using PILER-CR. Next, putative protospacer targets will be identified by a BLASTn search of the spacer input against the viral reference database. (ii) Users can upload viral sequences which will undergo BLASTn search against a spacer reference database. Two spacer reference databases have been built in our pipeline including spacers predicted from complete and/or draft bacterial genomes in NCBI. The bacterial sources of the hitting spacers are predicted as the potential hosts of the viral sequences. (iii) Users can check the phage–host links by CRISPR spacer-protospacer matching between the uploaded bacterial and phage sequences in (multi-)FASTA format. The spacer sequences will be predicted on the bacterial sequence using PILER-CR first, and be aligned to the phage sequences.

Prophage analysis. The prophage analysis module accepts both raw DNA sequence in FASTA format and annotated genomes in GenBank format, and provides three prophage detection programs including Phage_Finder, VirSorter, and DBSCAN-SWA. DBSCAN-SWA implements an algorithm combining DBSCAN algorithm (23) and Sliding Window Algorithm (SWA), referring to the theory of PHASTER, a widely used web tool for prophage prediction with no source code available (4). First, if a FASTA genomic sequence is given, all ORFs were first identified using FragGeneScan, and then putative phage proteins on the query genome were identified against the viral UniProt TrEML reference database via DIAMOND (24). Next, DBSCAN was used to cluster phage or phage-like genes into prophage regions. DBSCAN was implemented using scikit-learn (a free machine learning library for Python) with the minimal number of phage-like genes required to form a prophage cluster

and the protein density within the prophage region set to 11 and 0.015, respectively. If an annotated GenBank file is given, SWA is also performed to scan specific key phage-related proteins in the GenBank file, such as ‘protease’, ‘integrase’, 'transposase', 'terminase', 'lysis', 'bacteriocin' and key phage structural genes. Those regions with more than six key proteins within a moving window of 60 proteins are considered as putative prophage regions. The borders of the prophage region are determined as the positions of the first and last occurred key protein. Besides, tRNA and tmRNA sites are also annotated using tRNAscan-SE (25) and ARAGORN (26). Finally, the characterization of the predictive prophage region is performed using BLASTn against the Uniprot viral genome DNA sequences and the best hitting phage organism is returned. We also used the viral UniProt TrEML reference database to annotate the predicted ORFs in the prophage region. Annotated ORFs with taxonomy information were then subjected to a voting system and the prophage region was assigned a taxonomy based on the most abundant ORF taxonomy annotated within the prophage. Distribution of prophage-like elements detected by different methods and their size relative to the genome of their host are shown on an interactive circular genome viewer encoded using AngularPlasmid (http://angularplasmid.vixis.com). The corresponding prophage annotation will be shown on the right panel when clicking on the regions.

Similarity analysis. In this module, the similarity between the query phage genome and the genomes of 2,196 (or 1,871) reference phages with known host genus (or species) will be calculated using HostPhinder and the corresponding bacterial host species of the similar phages will be returned using a tree viewer and a table to illustrate the prediction process. GeneNet program is also provided to predict the phage host range based on a built-in gene-based virus-host reference network.

Co-occurrence analysis. This module receives relative abundance profiles in text file format as input, and uses CoNet implementation with Java to calculate the co-occurrence or co-exclusion relationships between the abundance of bacterial and phage organisms across samples. The co-occurrence analysis is mainly divided into initial network computation and assessment of significance. To score the association strength between bacteria and phages, by default, five metrics were calculated including correlation metrics (Pearson, Spearman), similarity metrics (mutual information), and distance metrics (Kullback-Leibler, Bray Curtis). In the next step, the significance of the associations was assessed with a permutation test and bootstraps, and multiple testing correction can be performed with Benjamini-Hochberg procedure by default. Finally, networks obtained from diverse measures were combined through voting systems using the Simes method. We also incorporated Cytoscape.js, an open-source graph theory library written in JavaScript for network visualisation so that the differences among the networks constructed using distinct metrics could easily be observed and compared.

Specialty genes check. As accessory genetic elements, bacteriophages play a crucial role in disseminating genes and promoting genetic diversity within bacterial populations. They can transfer genes encoding virulence factors such as toxins, adhesins and agressins, to promote the virulence of the host bacteria. Also, antibiotic resistance genes in bacterial chromosomes or plasmids can be mobilized by phages during the infection cycle, to increase antibiotic resistance. To identify specialty

genes for a pair of bacteria-phage genomes, ORFs were first predicted using FragGeneScan, then ShortBRED (27) and Resistance Gene Identifier (RGI) v3.1.1 (https://github.com/arpcard/rgi) were used to search predicted ORFs against VFDB (Virulence Factors of Pathogenic Bacteria) database (28) and the Comprehensive Antibiotic Resistance Database (CARD) (29), respectively. This kind of analysis facilitates our understanding about how specialty genes are transferred between bacteria and phages.

Protein-protein interaction analysis. Interactions between bacteriophage proteins and bacterial proteins are important for efficient infection of the host cell. We assigned bacterial and phage genes to homologs in UniProtKB protein database based on amino-acid sequence homology via DIAMOND searches, then the interactions between bacteriophage and bacterial proteins were inferred through checking the PPIs of their homologs in IntAct Molecular Interaction Database (https://www.ebi.ac.uk/intact/). The interactions between bacteriophage proteins and bacterial proteins may contribute to understand the infectious interactions between bacteria and phages.

Implementation of the PHISDetector server

The PHISDetector server is developed using Django, a high-level Python Web framework, on a Linux platform with an Apache web server. The web interface typically consists of an input page and a result page, which are generated with HTML, CSS, JavaScript and jQuery. The analysis modules were implemented using a combination of Python, Perl, R, Java, C++ and Shell scripts, and a series of public tools. The result visualization is mainly implemented by Cytoscape (https://cytoscape.org) for co-occurrence/co-abundance analysis and protein-protein interaction analysis, ECharts (https://echarts.baidu.com) for oligonucleotide profile analysis (Heatmap) and similarity analysis (Tree), AngularPlasmid (http://angularplasmid.vixis.com) for Prophage analysis, and DataTables (https://datatables.net) for displaying the results in the form of interactive tables. The web application is platform independent and has been tested successfully on Internet Explorer (version 9, 10, 11), Mozilla Firefox, Safari, and Google Chrome. Google Chrome is the recommended Web browser to run PHISDetector.

Evaluation of predictive power of diverse *in silico* signals

To assess the discriminatory power of each type of PHIS scores calculated using different bioinformatics tools, a set of 820 known interaction pairs (between 820 phages and 153 different bacterial hosts) (2) were used as the positive set, and a negative dataset of equal size was built artificially by matching phages with bacteria from a different species (within the 2,698 bacteria) other than their known host. We designed multiple types of features to represent diverse *in silico* PHISs that contribute to the prediction of phage-host interactions. One sided t-test was then used to examine if the signal scores are significantly different between positive and negative phage-host pairs (Figure 2). Receiver operating characteristic (ROC) curves were used to assess the power of predictive signals by plotting false positive rate (100-specifity) versus true positive rate (sensitively) according to the change of threshold for each signal feature (Figure 3), and the values of Sensitivity (Sn) and Positive

Predictive Value (PPV) were used as evaluation metrics for user to better assess the prediction results (Table 1).

Table 1. Comparative analysis of diverse PHIS values obtained for Sensitivity (Sn), Positive Predictive Value (PPV) and threshold for evaluation.

	Prophage analysis			Oligonucleotide analysis		CRISPR analysis		Similarity analysis	
Method	**Phage_Finder**	**DBSCAN-SWA**	**VirSorter**	**VirHost Matcher**	**WIsH**	**Percent identity**	**Spacer number**	**HostPhinder**	**GeneNet**
Sn	68.7%	79.1%	86.0%	88.4%	68.6%	92.3%	83.9%	75.9%	44.1%
PPV	77.6%	75.4%	69.7%	71.0%	71.2%	56.1%	42.9%	62.6%	72.5%
Threshold	46.6%	41.985%	1088.000	0.314	-1.395	60.606%	3.5	1806.5	71

The d_2* dissimilarity score provided by VirHostMatcher was used to compare two sequences based on the normalized oligonucleotide frequency (ONF). Positive phage-host pairs have significantly lower (*p*-value<2.2e-16 for one sided t-test) d_2* dissimilarity score than the negative pairs (Figure 2D). The WIsH score computes the log-likelihood of a phage genome coming from a host based on Markov chain model and has significantly different medians (p-value< 2.2e-16) between the positive and the negative pairs (Figure 2C). We predicted a phage-host pair as interacting if the VirHostMatcher d_2*<0.314 or log-likelihood>-1.395 (Table 1). For prophage analysis (Figure 3G-L), DBSCAN-SWA, our in-house developed tool, presents the best classification power, followed by Phage_Finder and VirSorter. We also found that protein searches using BLASTp are more accurate than nucleotide searches using BLASTn to annotate the integrated phage (Figure 3J and K). The area under the ROC, which measures the discriminative ability, was 0.83 for DBSCAN-SWA+BLASTp (Figure 3K) and 0.7 for DBSCAN-SWA+BLASTn (Figure 3H) at species level. Two CRISPR scores are defined either as the accumulated aligned scores between the host spacers and the viral genome, or as the number of host spacers matching the phage genome. Both CRISPR scores were significantly higher for the true interacting virus-host pairs than the non-interacting pairs with *p*-value<2.2e-16 and *p*-value<4e-12 for one sided (Figure 2A and B) and the area under the ROC was 0.77 and 0.65 respectively (Figure 3A and B) at species level. For virus-virus similarity measure, *K*-mer-based method (HostPhinder) performed better than gene-based method (GeneNet) (Figure 3E and F). In addition, to quantify the performance of the prophage detection module, we used a controlled dataset including 50 complete bacterial genomes containing 183 manually annotated prophages (7,30). Among these validated prophages, 72% were detected by DBSCAN-SWA that outperformed Phage_Finder (57%) and VirSorter (48%) (Supplementary Table S1).

Furthermore, to compare the prediction power of diverse PHIS using various methods, for each of the 820 known phage-host pairs, we statistic that at the species or genus level which PHIS could support or verify the interaction. We observed that single prediction approach could only partially identify the interactions (11.59%~74.024%), while combing multiple types of prediction algorithms could correctly identify about 88.293% of the known interactions at the species level (Figure 4A). HostPhinder was not counted in the statistics due to the overfitting of the test set. We attribute such

phenomenon to the possibility that different types of PHIS may require distinct methodology to explore and a universal or single method could hardly accommodate all situations. We illustrate the output using the results obtained from the prediction for phage of *Staphylococcus aureus subsp. aureus* JH1 (NC_009632) (Figure 4B) and for host of *Staphylococcus* phage 47 (NC_007054) (Figure 4C), and the characterization of this phage-host interaction (Figure 4D).

CONCLUSION AND DISCUSSION

Previous studies have indicated that molecular and ecological co-evolutionary processes of phages and bacteria leave signals in their genomic sequences for tracing their interactions. Several computational tools have been developed to detect various signals, such as integrated prophage, oligonucleotide frequency patterns, and CRISPR spacer sequences. Compared with previous webservers or softwares for phage-host interaction prediction, PHISDetector pipeline is uniquely comprehensive by integrating all previously published analysis types into one tool and adds valuable novel functionalities. Our prophage analysis module combined two popular programs, Phage_Finder and VirSorter, and our in-house developed tool DBSCAN-SWA. DBSCAN-SWA presents the best detection power based on the analysis using a controlled dataset including 183 manually annotated prophages, with a detection rate of 72% compared with Phage_Finder (57%) and VirSorter (48%). If combing all the three methods (provided as a “merge” function in prophage analysis module), 80% of the reference prophages could be detected. In addition, we added a prophage annotation step to indicate the possible integrated phage of the predicted regions. Our CRISPR analysis module facilitate two-way analysis. If a phage genome is submitted, it will be compared with our in-house collected spacer database (including CRISPR arrays from 63,182 bacterial and archaeal genomes) to quickly detect CRISPR targeting association between the input phage sequence and microbe genomes in NCBI. If a bacterial genome is received, PHISDetector will detect CRISPR-spacer automatically and compare against NCBI RefSeq Viral genome database to find the target phage. Oligonucleotide profile analysis module, supports both VirHostMatcher and WIsH, which are complementary to each other because VirHostMatcher may be more suitable for complete genomes while WIsH for virus contigs shorter than 10 kb. In co-abundance analysis module, we used CoNet program to infer a virus and bacterial co-occurrence network. As a plug-in of Cytoscape, we adapted CoNet to a web-version to better facilitate biologists without computational background to use and adjust parameters. We also provide a function module for the detection of protein-protein interactions between a pair of phage-host genomes, to better understand their interplay on protein level. In addition, to assist the characterization of phage genomes for therapeutic applications, we also introduce specialty genes check module to detect VFs and ARGs. Finally, a consensus analysis is performed to indicate the possible integrity of the predicted interactions and interplay among different PHIS. Based on our case study of PHIS detections for 820 known phage-host interactions, more than 88.293 % of the hosts were correctly identified at the species level combining various approaches. In summary PHISDetector is a tool that can detect diverse PHIS and could reflect various interaction patterns or mechanisms, so that users can easily compare the results from different methods and better understand their co-evolution. In addition, we provide various well-designed, interactively

visualization outputs for better result interpretation and illustration. PHISDetector will continue to evolve, to incorporate more *in silico* phage-host signals and evaluate the consistency or association of various signals upon extensive analysis of large amounts of datasets. We hope that PHISDetector can facilitate our understanding of the phage-host co-evolution, their roles in human health and disease, and provide novel therapeutic strategies.

DATA AVAILABILITY

WIsH is an open source collaborative initiative available in the GitHub repository (https://github.com/soedinglab/wish). VirHostMatcher is an open source collaborative initiative available in the GitHub repository (https://github.com/jessieren/VirHostMatcher). HostPhinder is an open source collaborative initiative available in the GitHub repository (https://github.com/julvi/HostPhinder). GeneNet is an open source collaborative initiative available in the GitHub repository (https://github.com/coevoeco/GeneNet). PILER-CR is an open source collaborative initiative available in the PILER website (http://www.drive5.com/pilercr/). CRISPRCasFinder is an open source collaborative initiative available in the CRISPR-Cas++ website (https://crisprcas.i2bc.paris-saclay.fr/Home/Download). VirSorter is an open source collaborative initiative available in the GitHub repository (https://github.com/simroux/VirSorter). Phage_Finder is an open source collaborative initiative available in the SOURCEFORGE (https://sourceforge.net/projects/phage-finder/files/). ShortBRED is an open source collaborative initiative available in the Bitbucket (https://bitbucket.org/biobakery/shortbred/src). RGI is an open source collaborative initiative available in the GitHub repository (https://github.com/arpcard/rgi). DIAMOND is an open source collaborative initiative available in the GitHub repository (https://github.com/bbuchfink/diamond). BLAST is an open source collaborative initiative available in NCBI website (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/). FragGeneScan is an open source collaborative initiative available in the SOURCEFORGE (https://sourceforge.net/projects/fraggenescan/)

ACKNOWLEDGEMENT

We thank Yunfei Ji, Jiale Zhang, and Yongkui Lai for their help in the developments of PHISDetector. We thank Zeguo Sun, Weijia Zhang from Icahn School of Medicine at Mount Sinai, Jiqiu Wu from Imperial College, and Fang Wang from MD Anderson Cancer Center for helpful comments and for testing the webserver software.

FUNDING

This work was supported by the National Natural Science Foundation of China [31825008 to Z.H., 61872117 to F.Z., 31800630 to Y.Z.]; and the Natural Scientific Research Innovation Foundation in Harbin Institute of Technology [GFEQ5750006918] to F.Z.

REFERENCES

1. Chatterjee, A. and Duerkop, B.A. (2018) Beyond Bacteria: Bacteriophage-Eukaryotic Host Interactions Reveal Emerging Paradigms of Health and Disease. *Front Microbiol*, **9**, 1394.
2. Edwards, R.A., McNair, K., Faust, K., Raes, J. and Dutilh, B.E. (2016) Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol Rev*, **40**, 258-272.
3. Roux, S., Enault, F., Hurwitz, B.L. and Sullivan, M.B. (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ*, **3**, e985.
4. Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y. and Wishart, D.S. (2016) PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res*, **44**, W16-21.
5. Lima-Mendez, G., Van Helden, J., Toussaint, A. and Leplae, R. (2008) Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics*, **24**, 863-865.
6. Fouts, D.E. (2006) Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res*, **34**, 5839-5851.
7. de Sousa, A.L., Maues, D., Lobato, A., Franco, E.F., Pinheiro, K., Araujo, F., Pantoja, Y., da Costa da Silva, A.L., Morais, J. and Ramos, R.T.J. (2018) PhageWeb - Web Interface for Rapid Identification and Characterization of Prophages in Bacterial Genomes. *Front Genet*, **9**, 644.
8. Pride, D.T., Wassenaar, T.M., Ghose, C. and Blaser, M.J. (2006) Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. *BMC Genomics*, **7**, 8.
9. Ahlgren, N.A., Ren, J., Lu, Y.Y., Fuhrman, J.A. and Sun, F. (2017) Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res*, **45**, 39-53.
10. Galiez, C., Siebert, M., Enault, F., Vincent, J. and Soding, J. (2017) WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics*, **33**, 3113-3114.
11. Villarroel, J., Kleinheinz, K.A., Jurtz, V.I., Zschach, H., Lund, O., Nielsen, M. and Larsen, M.V. (2016) HostPhinder: A Phage Host Prediction Tool. *Viruses*, **8**.
12. Shapiro, J.W. and Putonti, C. (2018) Gene Co-occurrence Networks Reflect Bacteriophage Ecology and Evolution. *MBio*, **9**.
13. Dutilh, B.E., Cassman, N., McNair, K., Sanchez, S.E., Silva, G.G., Boling, L., Barr, J.J., Speth, D.R., Seguritan, V., Aziz, R.K. *et al.* (2014) A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun*, **5**, 4498.
14. Faust, K. and Raes, J. (2016) CoNet app: inference of biological association networks using Cytoscape. *F1000Res*, **5**, 1519.
15. Hannigan, G.D., Meisel, J.S., Tyldsley, A.S., Zheng, Q., Hodkinson, B.P., SanMiguel, A.J., Minot, S., Bushman, F.D. and Grice, E.A. (2015) The human skin double-stranded DNA virome: topographical and temporal diversity, genetic enrichment, and dynamic associations with the host microbiome. *MBio*, **6**, e01578-01515.
16. Stern, A., Mick, E., Tirosh, I., Sagy, O. and Sorek, R. (2012) CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res*, **22**, 1985-1994.

17. Wang, J., Gao, Y. and Zhao, F. (2016) Phage-bacteria interaction network in human oral microbiome. *Environ Microbiol*, **18**, 2143-2158.
18. Biswas, A., Gagnon, J.N., Brouns, S.J., Fineran, P.C. and Brown, C.M. (2013) CRISPRTarget: bioinformatic prediction and analysis of crRNA targets. *RNA Biol*, **10**, 817-827.
19. Rho, M., Tang, H. and Ye, Y. (2010) FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res*, **38**, e191.
20. Grissa, I., Vergnaud, G. and Pourcel, C. (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res*, **35**, W52-57.
21. Edgar, R.C. (2007) PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics*, **8**, 18.
22. Ye, Y. and Zhang, Q. (2016) Characterization of CRISPR RNA transcription by exploiting stranded metatranscriptomic data. *RNA*, **22**, 945-956.
23. Ester, M., Kriegel, H.-P., #246, Sander, r. and Xu, X. (1996), *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. AAAI Press, Portland, Oregon, pp. 226-231.
24. Buchfink, B., Xie, C. and Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods*, **12**, 59-60.
25. Lowe, T.M. and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*, **25**, 955-964.
26. Laslett, D. and Canback, B. (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res*, **32**, 11-16.
27. Kaminski, J., Gibson, M.K., Franzosa, E.A., Segata, N., Dantas, G. and Huttenhower, C. (2015) High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED. *PLoS Comput Biol*, **11**, e1004557.
28. Chen, L., Zheng, D., Liu, B., Yang, J. and Jin, Q. (2016) VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. *Nucleic Acids Res*, **44**, D694-697.
29. Jia, B., Raphenya, A.R., Alcock, B., Waglechner, N., Guo, P., Tsang, K.K., Lago, B.A., Dave, B.M., Pereira, S., Sharma, A.N. *et al.* (2017) CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res*, **45**, D566-D573.
30. Casjens, S. (2003) Prophages and bacterial genomics: what have we learned so far? *Mol Microbiol*, **49**, 277-300.

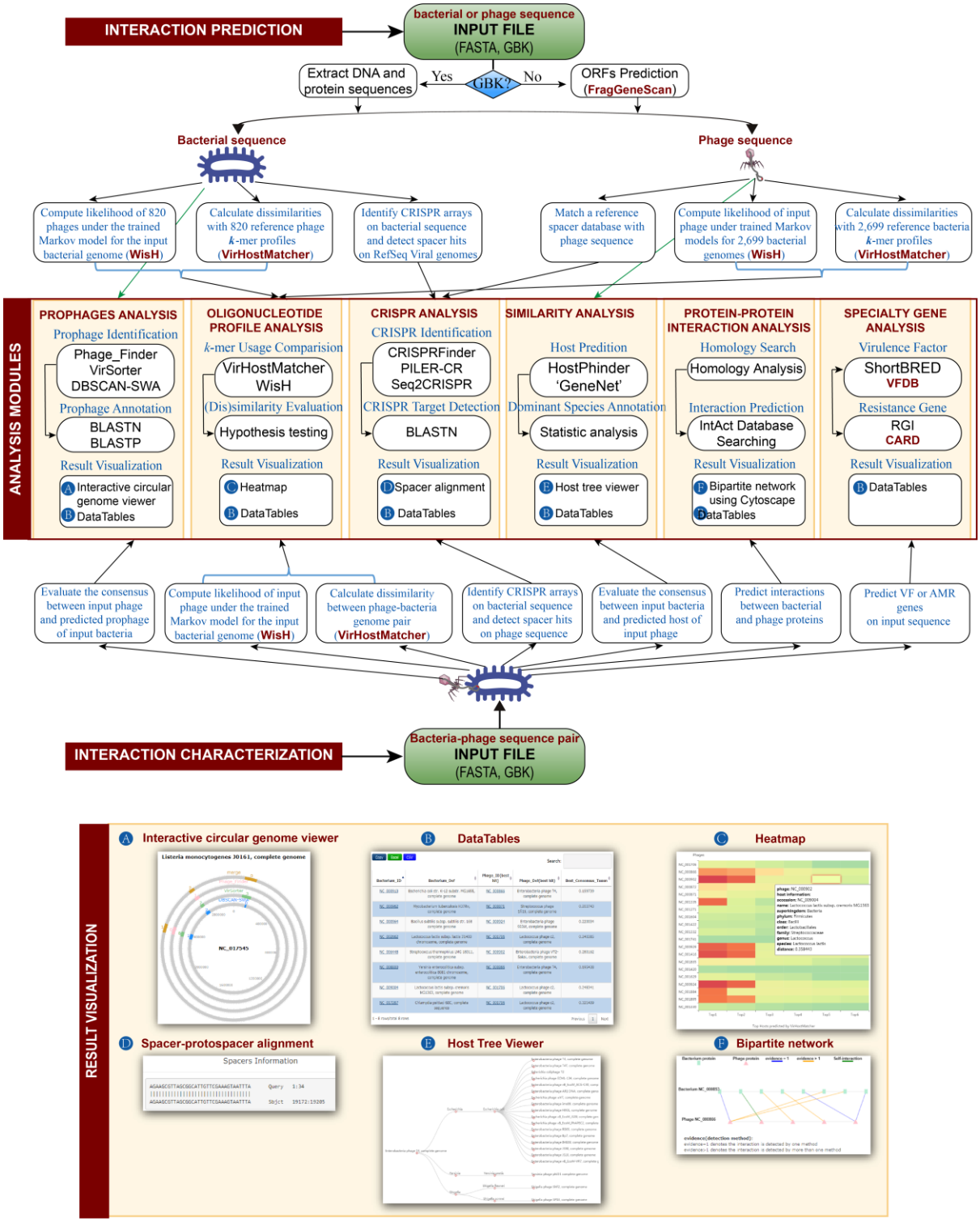

Figure 1. Pipeline for prediction and evaluation of microbe-phage interactions by PHISDetector. The pipeline receives the FASTA sequence file and the annotation file (GenBank) of the microbe or phage genome as input for further evaluation. PHISDetector is composed of three main analysis components: (i) interaction prediction that allows predicting microbe-phage interactions using diverse methods depending on the input sequence (microbe or phage) (Upper panel); (ii) interaction characterization of a pair of phage and bacteria genomes that will help to understand their co-evolution (Lower panel); and (iii) six underlying analysis modules (Middle panel) support diverse *in*

silico signals detection. In addition, PHISDetector uses more modern JavaScript tools for visualization of diverse prediction outputs.

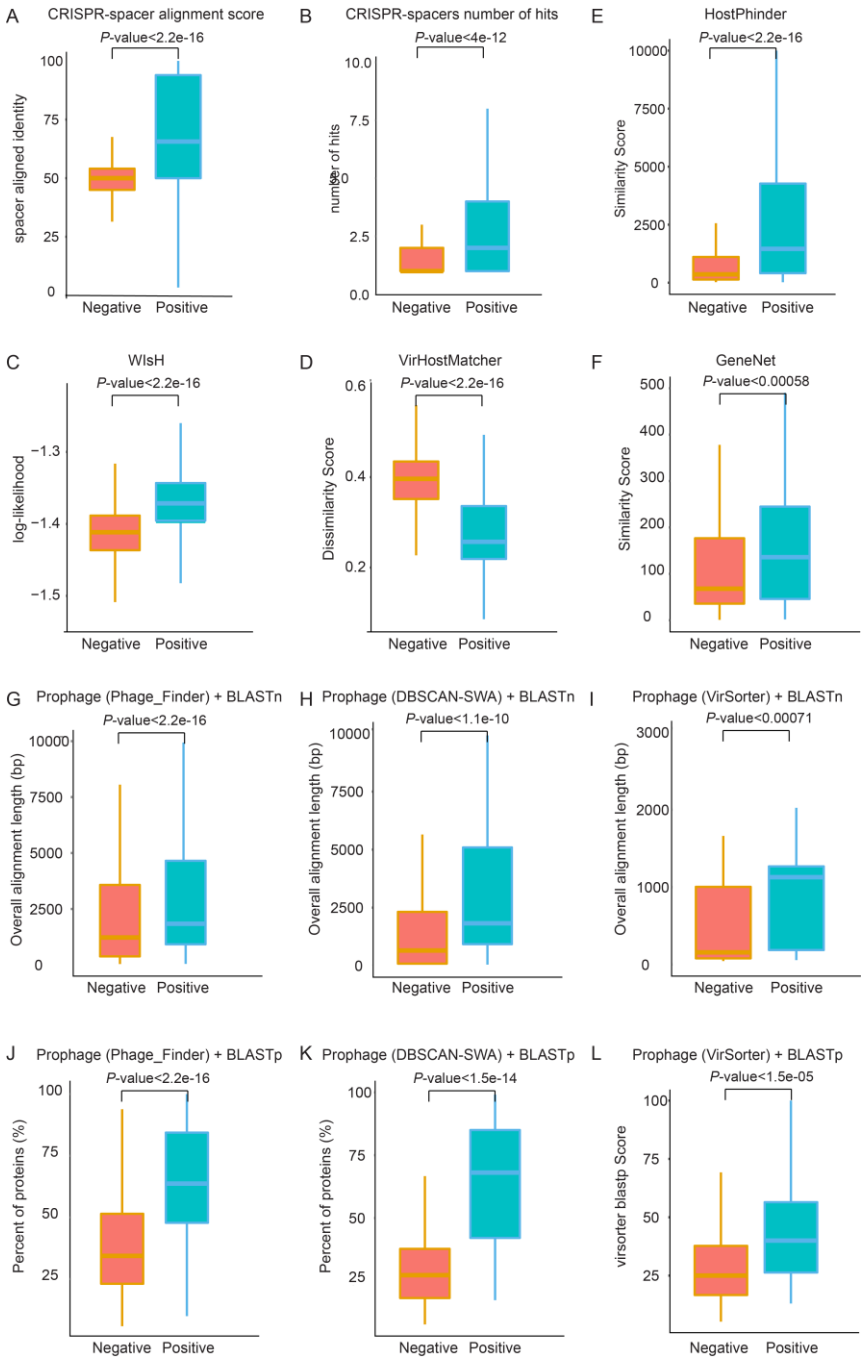

Figure 2 Distributions of the different feature values in the 820 interacting virus-host pairs (positive set) and the same number of non-interacting virus-host pairs (negative set) in the training data. (A-B) Boxplots of CRISPR scores defined as accumulated aligned identity between the host spacers and the viral genome (A), or as the number of host spacers matching the phage genome (B). (C) Boxplots of the log-likelihood scores calculated using WIsH. (D) Boxplots of the d_2* dissimilarity score provided by VirHostMatcher. (E) Boxplots of the *K*-mer-based virus-virus similarity scores given by HostPhinder.

(F) Boxplots of the gene-based virus-virus similarity scores given by GeneNet. (G-I) Overall alignment length of BLASTn hits between the phage and the prophage region predicted by Phage_Finder (G), DBSCAN-SWA (H), or VirSorter (I). (J-L) Boxplots of homologous protein percentage between the prophage regions predicted by Phage_Finder (J), DBSCAN-SWA (K), or VirSorter (L) and the phage genome. The horizontal bar displays the median, boxes display the first and third quartiles, whiskers depict minimum and maximum values.

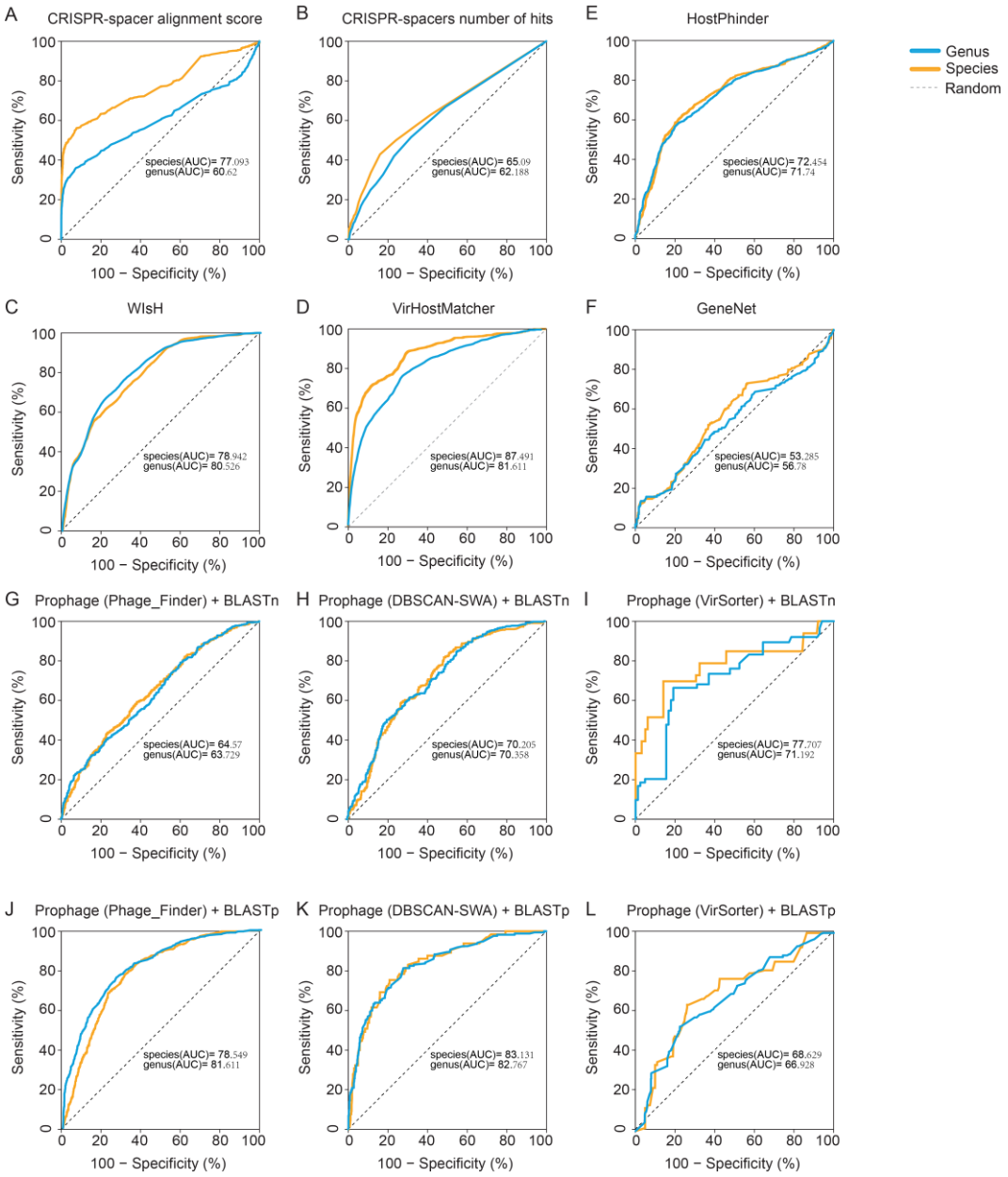

Figure 3. ROC curves displaying the classification accuracy of diverse *in silico* signals. (A) Accumulated aligned identity between the host spacers and the viral genome. (B) Number of CRISPR spacers (identified by either PILER-CR or CRISPRFinder) matching in phage genomes. (C) Log-likelihood scores given by WIsH of a phage contig against the Markov model trained from the bacterial genome. (D) Distance/dissimilarity measure of oligonucleotide usage profile between the phage and bacterial genomes estimated using VirHostMatcher. (E) Virus-virus similarity evaluated

using k-mer-based resemblance measure given by HostPhinder. (F) Gene-based virus-virus similarity given by GeneNet. (G-I) Overall alignment length of BLASTn hits between the phage and the prophage region predicted by Phage_Finder (G), DBSCAN-SWA (H), or VirSorter (I) on the bacterial genome. (J-L) Homologous protein percentage between the prophage regions predicted by Phage_Finder (J), DBSCAN-SWA (K), or VirSorter (L) and the phage genome. In a ROC curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points of a signal score.

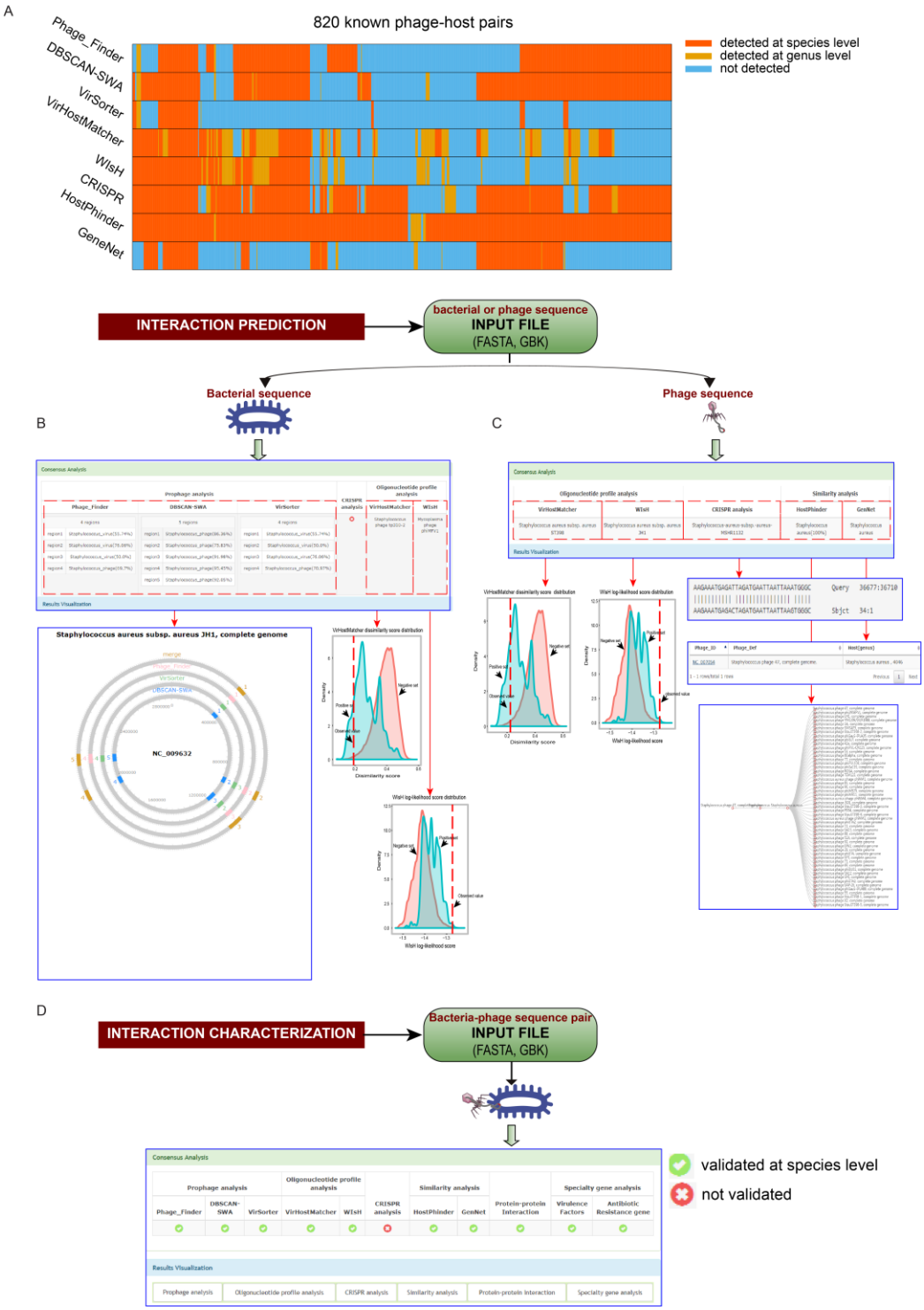

Figure 4. PHIS detected for 820 known phage-host pairs using diverse approaches. (A) Heatmap showing if a known phage-host pair is validated by each type of *in silico* PHIS detected using different tools. Orange or yellow denotes that a phage-host pair is validated by the approach at species level, or at genus level. Blue denotes not validated by an approach. (B) Prediction output for submitting the genome sequence of bacteria *Staphylococcus aureus subsp. aureus* JH1 (NC_009632). (C) Prediction output for submitting the genome sequence of phage *Staphylococcus* phage 47 (NC_007054). (D) Output for the characterization of this phage-host pair.

Supplementary figure S1. Comparative analysis for degree of matching between computational tools and reference coordinates in literature.