W74–W80 *Nucleic Acids Research, 2019, Vol. 47, Web Server issue* *Published online 22 May 2019*
doi: 10.1093/nar/gkz380

Prophage Hunter: an integrative hunting tool for active prophages

Wenchen Song[1,2,8,†], Hai-Xi Sun[1,2,†], Carolyn Zhang[3], Li Cheng[1,2,7], Ye Peng[1,2,7], Ziqing Deng[1,2], Dan Wang[1,2], Yun Wang[1,2], Ming Hu[8], Wenen Liu[6], Huanming Yang[1,2], Yue Shen[1,2], Junhua Li[1,2,7], Lingchong You[3,4,5] and Minfeng Xiao[1,2,*]

[1]BGI-Shenzhen, Shenzhen 518083, China, [2]China National Genebank, BGI-Shenzhen, Shenzhen 518120, China, [3]Department of Biomedical Engineering, Duke University, Durham, NC 27708, USA, [4]Center for Genomic and Computational Biology, Duke University, Durham, NC 27708, USA, [5]Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, NC 27708, USA, [6]Department of Clinical Laboratory, Xiangya Hospital, Central South University, Changsha 410008, China, [7]South China University of Technology, Guangzhou, Guangdong 510640, China and [8]Department of Basic Medical Sciences, Qingdao University, Qingdao 266071, China

Received March 01, 2019; Revised April 03, 2019; Editorial Decision April 19, 2019; Accepted May 17, 2019

ABSTRACT

Identifying active prophages is critical for studying coevolution of phage and bacteria, investigating phage physiology and biochemistry, and engineering designer phages for diverse applications. We present Prophage Hunter, a tool aimed at hunting for active prophages from whole genome assembly of bacteria. Combining sequence similarity-based matching and genetic features-based machine learning classification, we developed a novel scoring system that exhibits higher accuracy than current tools in predicting active prophages on the validation datasets. The option of skipping similarity matching is also available so that there's higher chance for novel phages to be discovered. Prophage Hunter provides a one-stop web service to extract prophage genomes from bacterial genomes, evaluate the activity of the prophages, identify phylogenetically related phages, and annotate the function of phage proteins. Prophage Hunter is freely available at https://pro-hunter.bgi.com/.

INTRODUCTION

Compared with over 199 000 bacterial genomes, fewer than 11 000 bacteriophage genomes are deposited in NCBI Genome as of 26 April 2019. The traditional source of phage information mainly depends on searching for phages in nature, which is stochastic, and sometimes difficult as proven for anaerobic bacteria and fastidious bacteria that grow only in specific nutrients and growth conditions (1,2). Advances in next-generation sequencing (NGS) technologies support the easy access, analysis, and identification of temperate phages. This is because nearly half of the sequenced bacteria are lysogens, representing a tremendous and previously under-explored source of prophages (3). Prophages are temperate phages integrated in bacterial genomes. While some prophages are active—they can be induced by stresses like UV or antibiotics, others are defective due to bacterial defence systems or mutational decay (4). Prophages can participate in a number of bacterial cellular processes, including antibiotic resistance, stress response, and virulence (5). With advances in synthetic biology, prophages are also considered as potential therapeutics for infectious or chronic diseases caused by bacteria (6–11).

Several tools have been developed to predict the existence of prophage sequences from bacterial NGS data (Table 1). MARVEL predicts phage sequences in metagenomics bins based on random forest machine learning approach (12). VirFinder is the first *k*-mer based program for identifying prokaryotic viral sequences from metagenomic data (13). PHASTER is a web server for the rapid identification and annotation of prophage sequences within bacterial genomes and plasmids (14,15). MetaPhinder identifies assembled genomic fragments (i.e. contigs) of phage origin in metagenomic data sets by comparison with a database of whole genome bacteriophage sequences (16). VirSorter detects prophages in complete microbial genomes or in fragmented genomic datasets, including incomplete genomes, SAGs, or metagenomic assemblies (17). PhiSpy identifies prophages by focusing on the characteristics of prophages that exhibit no similarity to sequenced genomes (18). Among these tools, only PHASTER considers the completeness of a putative prophage region. The evalua-

*To whom correspondence should be addressed. Tel: +86 755 33945504; Fax: +86 755 32960023; Email: xiaominfeng@genomics.cn
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

tion system of PHASTER, however, simply adds up scores representing the number of nucleotides, total genes, cornerstone genes, and phage-like genes while neglecting common mutational events that prophages experience. Also, like all other database-driven annotation systems, PHASTER only recognizes phages whose genes/proteins are close enough to the record in its database (14,15).

We developed Prophage Hunter, a novel integrative tool that employs sequence similarity-based searches within our customized phage parts library and prophage genetic features-based machine learning classification, to score the probability of a prophage being active. While incorporating similarity searches increases prediction accuracy, skipping it might raise the possibility of finding novel phages. Prophage Hunter provides both options, and allows the users to choose freely. Prophage Hunter systematically locates prophage regions within bacterial genomes, and predicts the activity of prophages. It also identifies the phylogenetically most related phage to the target prophages, as well as annotate the function of proteins throughout the phage genome. Distinguishing active prophages from inactive ones contributes to the study of coevolution between phage and bacteria (19). Also, obtaining active prophages facilitates further studies in phage physiology. Thus, Prophage Hunter may attract a wide range of users from and facilitate research in genomics, microbiology, synthetic biology, and other related areas.

MATERIALS AND METHODS

Features and datasets

A disrupted phage genome could exhibit changes in genetic features including transcription orientation, protein length, amino acid composition, Watson-Crick ratio, and transcription strand switch, etc. We used 24 features in our analysis: (i) Transcriptional orientation, the ratio of the number of prophage/bacterial genes in the longest stretch of consecutive genes in the same direction to the total number of genes in the prophage/bacterial genome, (ii) protein length, the average length of prophage or bacterial genes, 3–22) Composition of 20 amino acids, the frequency of each amino acid in all protein sequences in prophage or bacterial genome, 23) Watson-Crick ratio, the ratio of the number of genes transcribed from Watson strand to that of genes transcribed from Crick strand in the prophage or bacterial genomes, and 24) transcription strand switch, the ratio of total gene number to the number of transcription strand switches in the prophage or bacterial genomes.

Of the 3540 phage genomes extracted from the NCBI Genome database (https://www.ncbi.nlm.nih.gov/genome), 1031 temperate phages were identified using integrase as a marker (20), and considered as the positive training set for the machine learning approach of Prophage Hunter. The remaining 2509 phage genomic sequences were considered as the positive validation set. By aligning the 1031 genomic sequences to NCBI nt database using BLASTN with parameters '-task dc-megablast -dust no' (alignment length $\geq$ 10 kb and *e* value $\leq$ 1e–5 as cutoffs), 718 bacterial hosts were identified and their genomic sequences were used as the reference dataset. After removing all possible phage sequences from the host by BLASTN against the 3540 available phage genomes, 21 979 clean fragments from the host were generated whose lengths were equal to the prophages they contained and were used as the negative training set, and 5495 clean fragments were used as negative validation set. The 24 features of the positive/negative training set were calculated and sent for modelling.

After a temperate phage enters a bacterial cell, the phage genome could experience mutational events including insertions, deletions or inversions of genome fragments, which can compromise genome integrity and render the prophage inactive (4). Thus, based on the reference dataset, prophages were randomly disturbed *in silico* with insertions (sequences inserted from bacterial hosts), deletions (sequences deleted from prophages) or inversions (sequences inverted within prophages) equal to half of the size of the prophage region generating 2154 (718*3) bacterial genomes as the semi-synthetic dataset. 2154 temperate phage genomes disturbed as described above are used as the synthetic dataset.

The genomic sequences of *K. pneumoniae* KP6512, *A. baumannii* AB8929, and induced prophages have been deposited in the CNSA (https://db.cngb.org/cnsa/) of CNGBdb with accession code CNP0000380.

General workflow

Prophage Hunter combines sequence comparisons to known phage parts (fuzzy matching), scanning for attachment (*att*) site (boundary locating), machine learning of genetic features (activity scoring), to predict the probability of a prophage being active (Figure 1).

Fuzzy matching. To create the phage parts library, 2101 annotated phage genomes were downloaded from RefSeq database (21) (retrieved 18 January 2018). All CDS were extracted, and those with 75% nucleotide similarity were de-duplicated with Cd-hit (22) resulting in 99 465 parts in total. In the future, non-coding elements such as promoters and *loxP*-like sites could be used in the same process. Eleven classes of phage parts were categorized based on their function, LYS (lysis), INT (integration), REP (replication), REG (regulation), PAC (packaging), ASB (assembly), INF (infection), EVA (immune evasion), HYP (hypothetical protein), UNS (unsorted), and tRNA (Supplementary Table S1). These protein sequences were then searched against Pfam database using InterProScan (23) to obtain their domains with parameters '-dp -f tsv -goterms -appl Pfam -iprlookup'. Among the above 11 classes, four classes (ASB, HYP, INF and UNS) were found to be highly correlated with the existence of active prophages, and thus were used to find the initial prophage region. To do this, assembled bacterial genomic sequences (sequences with length $\leq$ 10 kb were removed) were used as query sequences, and they were searched against protein sequences of ASB, HYP, INF and UNS libraries using BLASTX with parameter '-seg no'. The translated sequences with premature stop codon were removed. Then the filtered sequences were searched against Pfam database using InterProScan with parameters '-dp -f tsv -goterms -appl Pfam -iprlookup', and only the sequences that contain domains of protein sequences of ASB, HYP, INF and UNS libraries were kept. For the option of 'skip similarity matching', the above procedures

Table 1. General characteristics of Prophage Hunter relative to other phage finding tools

	Prophage Hunter	MARVEL	VirFinder	PHASTER	MetaPhinder	VirSorter	PhiSpy
Last updated	2019	2018	2017	2016	2016	2015	2012
Target	Prophage	Phage	Virus	Prophage	Phage	Virus	Prophage
Similarity matching with databases	Yes	Yes	No	Yes	Yes	Yes	Yes
Genome features-based	Yes	Yes	Yes	No	No	Yes	Yes
Machine learning classifier	Logistic regression	Random forest	Logistic regression	No	No	No	Random forest
Prediction depth	active/ambiguous /inactive	phage/negative	phage/negative	intact/ questionable /incomplete	Phage/negative	phage/negative	phage/negative
att site prediction	Yes	No	No	Yes	No	No	Yes
Protein annotation	Yes	No	No	Yes	No	No	Yes
Local analysis	No	Yes	Yes	No	Yes	Yes	Yes
Programing skills required	No	Yes	Yes	No	Yes	No	Yes
Interactive	Yes	N/A	N/A	Yes	Yes	Yes	N/A

were omitted. Instead, sliding windows of 10 kb was generated across the genome and the 24 features mentioned above were calculated for each window. Only the windows scoring >0.8 were kept, clustered and proceeded to the subsequent 'boundary locating'.

Boundary locating. Bacterial genomic DNA sequences that contain the above four classes of genes were used to form initial clusters if the intergenic distance was less than 5 kb. Clusters that do not contain any of INF, ASB, PAC or integrase were removed. The 20 kb upstream and downstream sequences of each cluster were extracted to examine putative *att* sites using BLASTN with parameters '-task blastn-short -evalue 1000'. For all clusters, we used integrases as anchors because they were located downstream of the genuine *att* site. Only the putative *att* sites with length ≥12 bp were kept, and the *attL-attR* pair with the best sequence alignment (the highest bit score) was considered as the boundary of prophage region.

Activity scoring. A method based on logistic regression was defined to differentiate between host and prophage sequences. Specifically, we design an ensemble bagging approach which averages a set of linear regression models with lasso regularization to predict the probability of a sequence being phage, and a score closer to 1 indicates a higher probability of the prophage being active. Here, we train on a dataset containing 1031 active prophage sequences and 21 979 host sequences with the above 24 features. To account for the highly imbalanced nature of this dataset, we undersample (sample without replacement) from the majority class, such that the training sets have a 50–50 split between host and phage observations. To implement this model, we train 50 logistic regression models on 50 randomly selected training sets. The training sets each contain all phage observations and a subset of 1031 host observations. This allows us to take into account variation among the host observations while ensuring a balanced dataset.

Program and web implementation

The above scripts/pipelines were written using a combination of Perl, Python and R. In addition, public bioinformatics tools such as BLAST (24), InterProScan (23), GeneMark (25), StringTie (26) and Cuffcompare (27) were also embedded. The Prophage Hunter web server was constructed using the classical LAMP framework (Linux, Apache, MySQL and PHP). For further exploration of the results of active prophage, an interactive light weight genome browser based on Tnt Board (http://tntvis.github.io/tnt.genome/) was built. In the genome browser, one can zoom in or zoom out to easily overview the distribution of predicted phages and their activity scores (shown in color scales), and predicted genes (both strands). The genome browser and the result table is interlocked for fast information focusing. Clicking phage candidates in the genome browser will trigger the table filtering of detailed information and clicking table items will enlarge its position in the genome browser. This web server is compatible with most modern web browsers like Mozilla Firefox, Google Chrome, Safari and Microsoft edge. For more details about usages and results explanations, please check the 'Tutorial' section in the Prophage Hunter web server at https://pro-hunter.bgi.com/.

RESULTS

Validating the machine learning approach of Prophage Hunter

For each of the 50 models, we first used cross validation on the training set to optimize λ, the lasso penalty term. We used the 'one-standard-error' rule to select λ. We then applied the algorithm to the positive and negative validation set composed of the observations independent from the training set. With this implementation, we attained an average training set accuracy of 0.99, sensitivity of 0.99, specificity of 0.99, negative set accuracy of 0.98, and positive set accuracy of 0.97 (Supplementary Table S2).

We also tested the performance of our modelling in discriminating mutational events using the synthetic dataset containing 718 disturbed prophage genomic sequences, and

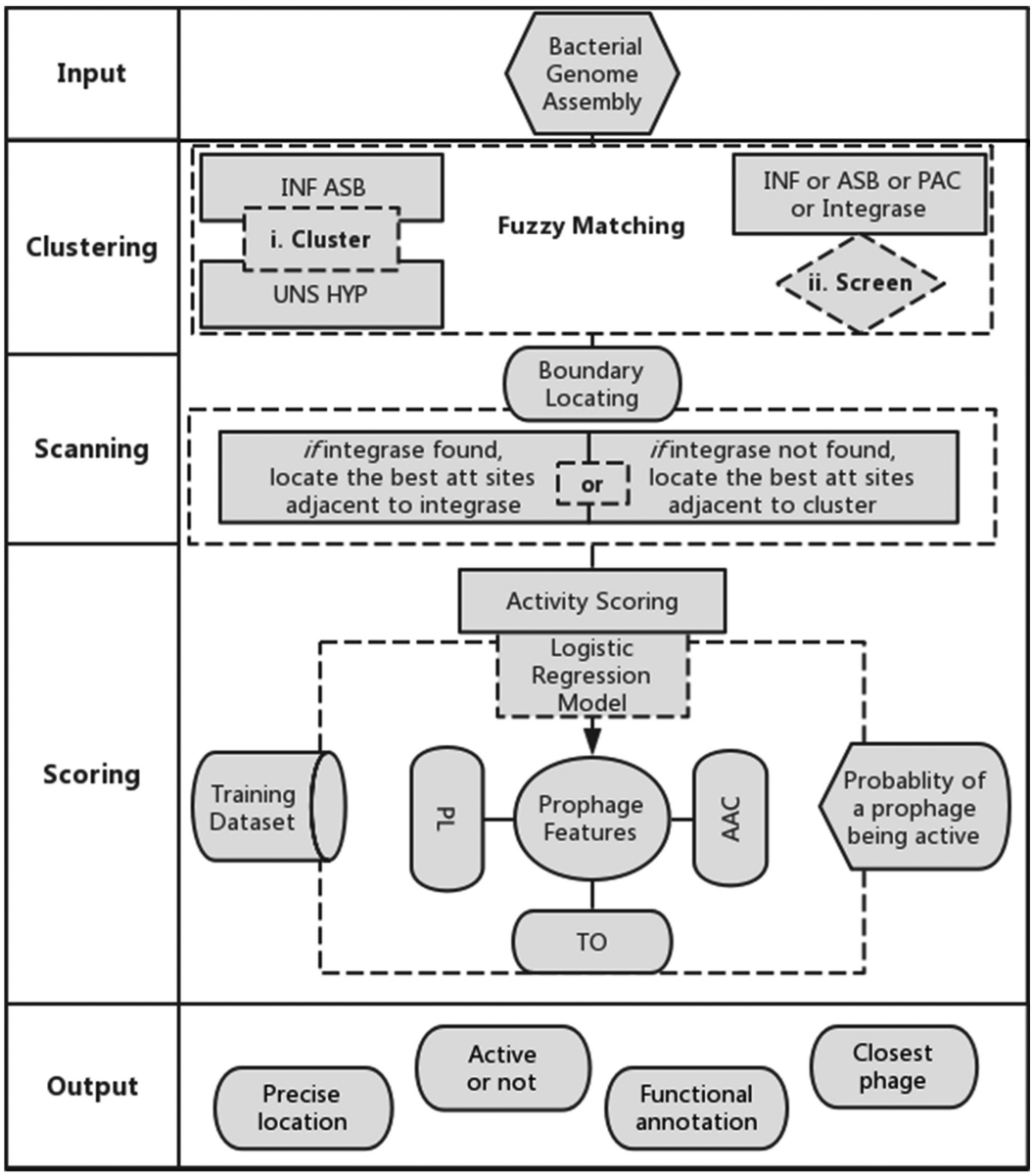

Figure 1. Overview of Prophage Hunter workflow Four classes of phage parts—ASB, HYP, INF and UNS—were used to form initial clusters within bacterial genomic sequences if the intergenic distance was less than 5 kb. Clusters that contain none of INF, ASB, PAC or integrase were discarded. The 20 kb upstream and downstream sequences of each cluster were scanned to locate putative *att* sites. Only the *att* sites with length ≥12 bp were kept, and the *attL-attR* pair with the highest bit score was considered as the boundary of the prophage region. A scoring system based on machine learning of prophage genetic features was defined to potentially differentiate between inactive and active prophage sequences. PL indicates protein length—the average length of prophage or bacterial genes. TO indicates transcriptional orientation—the ratio of the number of prophage/bacterial genes in the longest stretch of consecutive genes in the same direction to the total number of genes in the prophage/bacterial genome, the ratio of the number of genes transcribed from Watson strand to that of genes transcribed from Crick strand in the prophage or bacterial genomes, and the ratio of total gene number to the number of transcription strand switches in the prophage or bacterial genomes. AAC indicates amino acids composition—the frequency of each amino acid in all protein sequences in prophage or bacterial genome. A score closer to 1 indicates a higher probability of the prophage being active, we set a putative prophage region scoring >0.8 as active, 0.5–0.8 as ambiguous, and <0.5 as inactive. The start and end position of the prophage, the probability of the prophage being active, the phylogenetically most related phage, and the functional annotations of phage proteins are output.

compared it with the three most recently updated tools. MetaPhinder (16) is based on similarity matching with a phage database, while VirFinder (13) is based on machine learning of k-mer features to indicate whether a contig is phage or not. In this study, we set a putative prophage region scoring >0.8 as active prophage, 0.5–0.8 as ambiguous, and <0.5 as inactive for Prophage Hunter. TN (true negative) represents 'not phage' for MetaPhinder and VirFinder, and 'ambiguous/inactive' for Prophage Hunter. FP (false positive) represents 'phage' for MetaPhinder and VirFinder, and 'active' for Prophage Hunter. The prediction accuracy is defined as TN/(TN+FP), which indicates the proportion of mutational events successfully discriminated by the tools. Prophage Hunter modelling demonstrated

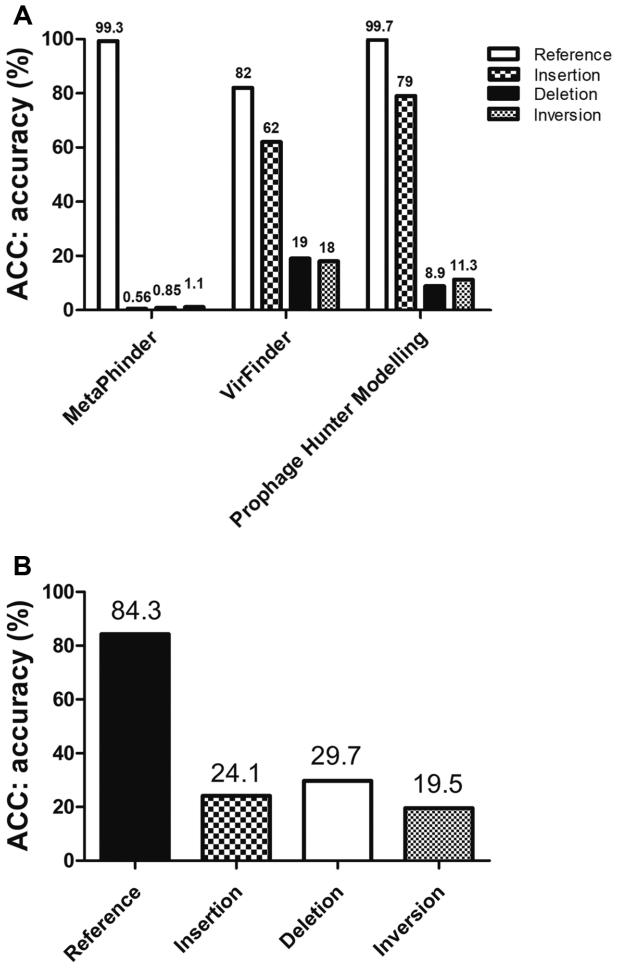

Figure 2. Evaluating the performance of Prophage Hunter. (**A**) The prediction accuracy of Prophage Hunter modelling was compared to MetaPhinder and VirFinder using the synthetic dataset, comprising 2154 (718*3) temperate phage genomes randomly disturbed by insertions, deletions, or inversions respectively. Reference indicates non-disturbed phage genomes. (**B**) The prediction accuracy of Prophage Hunter integrative pipeline was tested on reference dataset comprising 718 bacterial genomes carrying prophages, and on semi-synthetic dataset comprising bacterial genomes carrying 2154 (718*3) disturbed prophages. The prediction accuracy is defined as TN/(TN + FP) for the synthetic and semi-synthetic datasets, and TP/(TP + FN) for the reference dataset.

an accuracy of 79% in discriminating insertions which is higher than VirFinder and MetaPhinder, ∼10-fold higher accuracy in discriminating deletions and inversions than MetaPhinder, and slightly lower accuracy in discriminating deletions and inversions than VirFinder (Figure 2A). Both MetaPhinder and VirFinder displayed reasonably high accuracy in predicting non-disturbed reference genomes, suggesting the performance is not biased towards the datasets tested (Figure 2A).

We randomly chose 30 disturbed prophages from the synthetic dataset with insertions, deletions, and inversions respectively, and submitted them to PHASTER web server. To the extent of the tested synthetic dataset, Prophage Hunter modelling was comparable to PHASTER in discriminating inversions and deletions, and showed higher accuracy in discriminating insertions than PHASTER (Supplementary Table S3). Altogether, these results suggest the modelling of Prophage Hunter alone outperforms the three tools in differentiating prophages inactivated by the mutational events tested, except VirFinder displayed moderately better performance in discriminating deletions and inversions than Prophage Hunter modelling.

Evaluating the performance of Prophage Hunter integrative pipeline

We validated the performance of Prophage Hunter integrative pipeline in predicting active prophages using the reference dataset, and in discriminating inactive prophages using the semi-synthetic dataset. Prophage Hunter showed an accuracy of 84.3% with the reference dataset, 24.1% with insertions, 29.7% with deletions, and 19.5% with inversions, respectively (Figure 2B).

Because VirFinder and MetaPhinder—built for relatively short sequences e.g. metagenomics contigs—did not identify the start and end of a prophage region (13,16), the two tools were not suitable to test the semi-synthetic dataset comprising bacterial whole genome assemblies. We then compared Prophage Hunter's ability in discriminating inactive prophages with PHASTER. To the extent of the tested semi-synthetic dataset, Prophage Hunter integrative pipeline was comparable to PHASTER in discriminating deletions, and showed higher accuracy in discriminating insertions and inversions than PHASTER (Supplementary Table S4).

Finally, we tested Prophage Hunter on *B. licheniformis* DSM13, and compared the results with experimental evidence published previously (28). Indeed, Prophage Hunter predicted *B. licheniformis* DSM13 contain three active prophages and one ambiguous prophage, and the regions are consistent with the experimental data, while PHASTER's prediction included one false negative and one false positive (Figure 3A).

Case Study: Hunting for active prophages in the clinical isolates of *Klebsiella pneumoniae* and *Acinetobacter baumannii*

To hunt for active prophages with clinical significance, we applied Prophage Hunter on clinical isolates of *K. pneumoniae* and *A. baumannii*. Gram-negative bacteria, including multidrug resistant *A. baumannii* and extended-spectrum beta-lactamases (ESBL) producing *Enterobacteriaceae*, have been associated to severe healthcare-associated infections, and *K. pneumoniae* and *A. baumanni* are among the most commonly isolated microorganisms in intensive care unit (ICU)-acquired infections in different parts of the world (29–32). Prophage Hunter suggested both *K. pneumoniae* KP6512 and *A. baumannii* AB8929 strains contained active prophages (Figure 3B, C). We then performed mitomycin C induction experiments using KP6512 and AB8929, and sequenced the induced prophages. For KP6512, four active prophages were predicted by Prophage Hunter, three were predicted by PHASTER, and one was successfully induced (Figure 3B). For AB8929, four active prophages were predicted by Prophage Hunter, none were predicted by PHASTER, and two were successfully induced (Figure 3C). Overall, although possible false positives—due

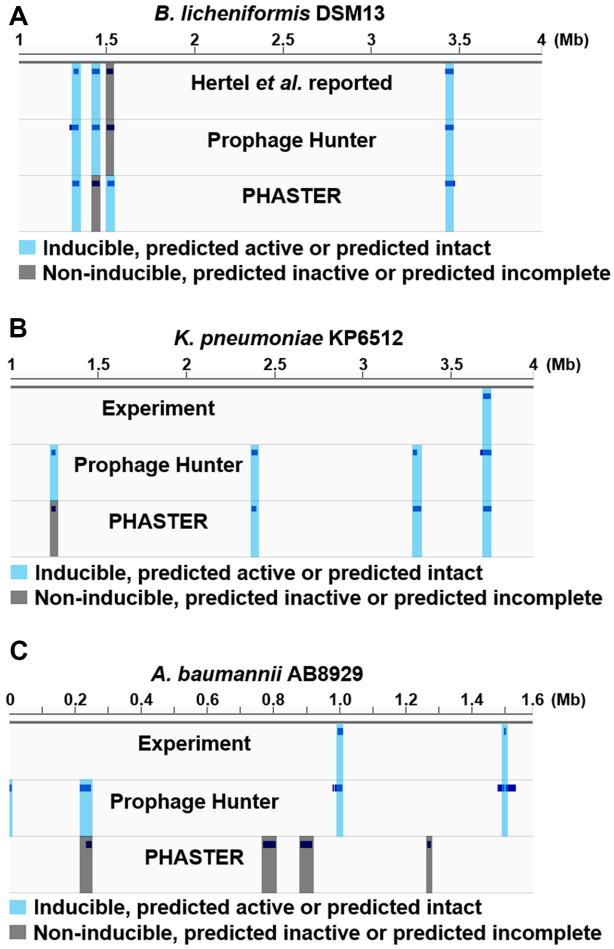

Figure 3. Identifying prophages in (**A**) *B. licheniformis* DSM13, (**B**) *K. pneumoniae* KP6512, and (**C**) *A. baumannii* AB8929. Prophages identified by experiments, Prophage Hunter, and PHASTER are indicated by blue rectangles on the top, medium and bottom of each panel respectively. Inducible, active and intact prophages are highlighted by azure strips, whereas non-inducible, inactive and incomplete prophages are highlighted by gray strips. The results were visualized using Integrative Genomics Viewer (IGV, version 2.4.19) (34).

to inaccurate prediction or unsuccessful prophage induction by mitomycin C—existed, Prophage Hunter outperformed PHASTER and managed to hunt all the inducible prophages in the two cases.

DISCUSSION

Prophages are highly abundant in bacterial genomes and are more readily available than lytic phages. With Prophage Hunter, we can take advantage of the growing bacterial NGS data to mine prophage information. Predicting prophage activity may lead to the finding of CRISPR-like systems or other novel defence systems of bacteria against phages (19,33). Prophage Hunter will also guide the induction and isolation of temperate phages from bacteria, thus enabling further investigation of phage physiology, as well as the development of designer phages for therapeutic purposes (6–11).

Prophage Hunter was not meant for replacing any other similar tools, instead, it was created to meet certain research needs that may have been neglected by previous tools. Prophage Hunter can be further improved in several aspects. First, phage proteins may not show high homology to the phage parts in Prophage Hunter library and thus novel phages can be overlooked. To overcome this, the users can choose to omit the similarity searches but might at the same time attenuate the prediction accuracy (Supplementary Tables S3 and S4). Second, the positive training set used in modelling is limited—1031 prophages and highly skewed towards *Caudovirales*, and the prediction accuracy was not tested for phages other than *Caudovirales*. Third, the activity of a prophage can be affected by other events and thus correlated with other genetic features not considered by Prophage Hunter. Therefore, continuous efforts on tools like this could be to augment the active prophage database as more researchers combine prophage prediction and experimental validation across a broader range of hosts, search for more signatures of phages compared with bacteria, and study further in the processes leading to prophage inactivation. Since predicting prophage activity is largely affected by the integrity and quality of input genome sequences, with future progress in metagenomics sequencing and assembling technology, it will be exciting to look into the status of prophages in the microbiome.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank China National GeneBank for sequencing the genomic DNA of bacterial isolates and phages, and Dr Tong Chen from EHBIO gene technology (http://www.ehbio.com/) and his colleagues, Pu Xue and Yu Liu, for providing technical support on the web implementation. Our special gratitude goes to Yiran—M.X.'s baby who was born around the same time this project started—for kindly sharing her mommy with the study.

FUNDINGS

National Key R&D Program of China [2018YFC1200501]; National Natural Science Foundation of China [31601073]; Shenzhen Municipal Government of China [JCYJ20160531194327655]. Funding for open access charge: National Key R&D Program of China [2018YFC1200501]; National Natural Science Foundation of China [31601073].
Conflict of interest statement. None declared.

REFERENCES

1. Hargreaves,K.R. and Clokie,M.R. (2014) Clostridium difficile phages: still difficult? *Front. Microbiol.*, **5**, 184.
2. 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.

3. Touchon,M., Bernheim,A. and Rocha,E.P. (2016) Genetic and life-history traits associated with the distribution of prophages in bacteria. *ISME J.*, **10**, 2744–2754.
4. Canchaya,C., Proux,C., Fournous,G., Bruttin,A. and Brussow,H. (2003) Prophage genomics. *Microbiol. Mol. Biol. Rev.*, **67**, 238–276.
5. Argov,T., Azulay,G., Pasechnek,A., Stadnyuk,O., Ran-Sapir,S., Borovok,I., Sigal,N. and Herskovits,A.A. (2017) Temperate bacteriophages as regulators of host behavior. *Curr. Opin. Microbiol.*, **38**, 81–87.
6. Monteiro,R., Pires,D.P., Costa,A.R. and Azeredo,J. (2018) Phage therapy: going temperate? *Trends Microbiol.*, **27**, 368–378.
7. Kilcher,S. and Loessner,M.J. (2018) Engineering bacteriophages as versatile biologics. *Trends Microbiol.*, **27**, 355–367.
8. Park,J.Y., Moon,B.Y., Park,J.W., Thornton,J.A., Park,Y.H. and Seo,K.S. (2017) Genetic engineering of a temperate phage-based delivery system for CRISPR/Cas9 antimicrobials against Staphylococcus aureus. *Sci. Rep.*, **7**, 44929.
9. Yosef,I., Manor,M., Kiro,R. and Qimron,U. (2015) Temperate and lytic bacteriophages programmed to sensitize and kill antibiotic-resistant bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 7267–7272.
10. Kilcher,S., Studer,P., Muessner,C., Klumpp,J. and Loessner,M.J. (2018) Cross-genus rebooting of custom-made, synthetic bacteriophage genomes in L-form bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 567–572.
11. Zhang,H., Fouts,D.E., DePew,J. and Stevens,R.H. (2013) Genetic modifications to temperate Enterococcus faecalis phage Ef11 that abolish the establishment of lysogeny and sensitivity to repressor, and increase host range and productivity of lytic infection. *Microbiology*, **159**, 1023–1035.
12. Amgarten,D., Braga,L.P.P., da Silva,A.M. and Setubal,J.C. (2018) MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. *Front. Genet.*, **9**, 304.
13. Ren,J., Ahlgren,N.A., Lu,Y.Y., Fuhrman,J.A. and Sun,F. (2017) VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, **5**, 69.
14. 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–W21.
15. Zhou,Y., Liang,Y., Lynch,K.H., Dennis,J.J. and Wishart,D.S. (2011) PHAST: a fast phage search tool. *Nucleic Acids Res.*, **39**, W347–W352.
16. Jurtz,V.I., Villarroel,J., Lund,O., Voldby Larsen,M. and Nielsen,M. (2016) MetaPhinder-Identifying bacteriophage sequences in metagenomic data sets. *PLoS One*, **11**, e0163111.
17. Roux,S., Enault,F., Hurwitz,B.L. and Sullivan,M.B. (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ.*, **3**, e985.
18. Akhter,S., Aziz,R.K. and Edwards,R.A. (2012) PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res.*, **40**, e126.
19. Ofir,G. and Sorek,R. (2018) Contemporary phage biology: from classic models to new insights. *Cell*, **172**, 1260–1270.
20. Howard-Varona,C., Hargreaves,K.R., Abedon,S.T. and Sullivan,M.B. (2017) Lysogeny in nature: mechanisms, impact and ecology of temperate phages. *ISME J.*, **11**, 1511–1520.
21. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
22. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, **22**, 1658–1659.
23. Jones,P., Binns,D., Chang,H.Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)*, **30**, 1236–1240.
24. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
25. Besemer,J., Lomsadze,A. and Borodovsky,M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, **29**, 2607–2618.
26. Pertea,M., Pertea,G.M., Antonescu,C.M., Chang,T.C., Mendell,J.T. and Salzberg,S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.
27. Trapnell,C., Hendrickson,D.G., Sauvageau,M., Goff,L., Rinn,J.L. and Pachter,L. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.*, **31**, 46–53.
28. Hertel,R., Rodriguez,D.P., Hollensteiner,J., Dietrich,S., Leimbach,A., Hoppert,M., Liesegang,H. and Volland,S. (2015) Genome-based identification of active prophage regions by next generation sequencing in Bacillus licheniformis DSM13. *PLoS One*, **10**, e0120759.
29. Robenshtok,E., Paul,M., Leibovici,L., Fraser,A., Pitlik,S., Ostfeld,I., Samra,Z., Perez,S., Lev,B. and Weinberger,M. (2006) The significance of Acinetobacter baumannii bacteraemia compared with Klebsiella pneumoniae bacteraemia: risk factors and outcomes. *J. Hosp. Infect.*, **64**, 282–287.
30. Perez,F., Endimiani,A., Ray,A.J., Decker,B.K., Wallace,C.J., Hujer,K.M., Ecker,D.J., Adams,M.D., Toltzis,P., Dul,M.J. *et al.* (2010) Carbapenem-resistant Acinetobacter baumannii and Klebsiella pneumoniae across a hospital system: impact of post-acute care facilities on dissemination. *J. Antimicrob. Chemother.*, **65**, 1807–1818.
31. Azimi,L., Lari,A.R., Talebi,M., Owlia,P., Alaghehbandan,R., Asghari,B. and Lari,E.R. (2015) Inhibitory-based method for detection of Klebsiella pneumoniae carbapenemase Acinetobacter baumannii isolated from burn patients. *Indian J. Pathol. Microbiol.*, **58**, 192–195.
32. Agodi,A., Barchitta,M., Quattrocchi,A., Maugeri,A., Aldisio,E., Marchese,A.E., Mattaliano,A.R. and Tsakris,A. (2015) Antibiotic trends of Klebsiella pneumoniae and Acinetobacter baumannii resistance indicators in an intensive care unit of Southern Italy, 2008–2013. *Antimicrob. Resist. Infect. Control.*, **4**, 43.
33. Dedrick,R.M., Jacobs-Sera,D., Bustamante,C.A., Garlena,R.A., Mavrich,T.N., Pope,W.H., Reyes,J.C., Russell,D.A., Adair,T., Alvey,R. *et al.* (2017) Prophage-mediated defence against viral attack and viral counter-defence. *Nat. Microbiol.*, **2**, 16251.
34. Thorvaldsdottir,H., Robinson,J.T. and Mesirov,J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.