# DBSCAN-SWA: an integrated tool for rapid prophage detection and annotation

Rui Gan[#], Fengxia Zhou[#], Yu Si[#], Han Yang, Chuangeng Chen, Jiqiu Wu, Fan Zhang[*], Zhiwei Huang[*]

[1]School of Life Science and Technology, Harbin Institute of Technology, Harbin 150080, China

*To whom correspondence should be addressed.

## Abstract

**Summary:** As an intracellular form of a bacteriophage in the bacterial host genome, a prophage is usually integrated into bacterial DNA with high specificity and contributes to horizontal gene transfer (HGT). Phage therapy has been widely applied, for example, using phages to kill bacteria to treat pathogenic and resistant bacterial infections. Therefore, it is necessary to develop effective tools for the fast and accurate identification of prophages. Here, we introduce DBSCAN-SWA, a command line software tool developed to predict prophage regions of bacterial genomes. DBSCAN-SWA runs faster than any previous tool. Importantly, it has great detection power based on analysis using 184 manually curated prophages, with a recall of 85% compared with Phage_Finder (63%), VirSorter (74%) and PHASTER (82%) for raw DNA sequences. DBSCAN-SWA also provides user-friendly visualizations including a circular prophage viewer and interactive DataTables.

**Availability and implementation:** DBSCAN-SWA is implemented in Python3 and is freely available under an open source GPLv2 license from https://github.com/HIT-ImmunologyLab/DBSCAN-SWA/.

## 1 Introduction

Bacteriophages are viruses that specifically infect their bacterial hosts. Passive replication of the phage genome relies on integration into the host's chromosome and becoming a prophage (Panis et al., 2010). Prophages coexist and evolve with bacteria, influencing the entire ecological environment. Recently, phage therapy, defined as using phages to treat bacterial infections, has also been greatly emphasized. Therefore, the identification of prophages in their host genomes become critical not only for understanding their biological mechanisms but also for developing therapeutic strategies.

Several tools have been developed to predict putative prophage regions. Phage_Finder (Fouts et al., 2006) is a standalone software program based on a heuristic algorithm to identify prophage regions in completely sequenced bacterial genomes. VirSorter (Roux et al., 2015) supports the detection of viral segments in microbiome sequencing data. PHASTER is a popular webserver for the identification and annotation of prophage sequences in prokaryotic genomes and plasmids (Wishart et al., 2016). Prophage Hunter (Xiao et al., 2019) provides a one-stop web service to identify prophage regions in bacterial genomes and evaluate the activity of the prophages. All these tools have substantially revolutionized the prediction of prophages in bacterial genomes. However, despite supporting prophage detection from massive bacterial genomes, Phage_Finder and VirSorter have limitations in speed and predictive power. PHASTER and Prophage Hunter support predictions using the webserver but cannot perform large-scale predictions of high-throughput microbiome sequencing data. Here, we introduce DBSCAN-SWA, a tool inspired by previous algorithms and tools, to detect prophages. DBSCAN-SWA has the fastest running speed and outperforms previous tools in detection rate and applicability for the prediction of complete and incomplete sequencing data (**Table 1**).

## 2 Methods and Implementation

### 2.1 Prophage detection

DBSCAN-SWA implements an algorithm combining density-based spatial clustering of applications with noise (DBSCAN) and a sliding window algorithm (SWA) to detect putative prophages with reference to the theory underlying PHASTER, which is the most popular prophage detection tool but has no standalone version or source code available. Furthermore, we made some improvements. DBSCAN-SWA accepts two types of bacterial sequence files (multi-FASTA, GBK) for input. If a multi-FASTA input file is received, gene prediction and annotation will be performed by Prokka (Torsten, 2014) to obtain a standard GenBank format file with tRNA sites annotated using ARAGORN (Laslett et al., 2004). If a GenBank annotated file is submitted, gene annotations including protein sequences, descriptions and tRNA sites will be extracted for subsequent analysis. Considering prophage regions are composed of phage or phage-like genes clustered in a local region in the bacterial genome (Zhou et al., 2011), phage or phage-like proteins are first identified using Diamond BLASTP (e-values < 1e-7) (Buchfink et al., 2015) by searching against DBSCAN-SWA's local viral UniProt TrEML reference database (UniProt C, 2014), which greatly improves the efficiency of DBSCAN-SWA compared to tools such as PHASTER using BLASTP. Then, DBSCAN is performed to detect phage protein clusters with the minimal number of phage-like genes required to form a prophage cluster (set to 6 proteins as the default parameter value) and the maximal spatial distance between two neighbor genes within the same cluster, which reflects the protein density within the prophage region (set to 3000 bps as the default parameter value). These two parameters are learned using a gradient method based on 184 manually curated prophage regions (Casjens et al., 2003) by trying the minimal prophage size from 6 to 10 proteins (step=1) and the protein density from 3000 to 10000 bp (step=1000 bp). Additionally, considering the biological features of prophages from different bacterial species may vary case by case, DBSCAN-SWA supports users to flexibly modify the two key parameters of DBSCAN, while PHASTER provides only fixed parameters. DBSCAN-SWA also uses SWA to scan specific key phage-related proteins in the GenBank file, such as ‘protease’, ‘integrase’, 'transposase', 'terminase', 'lysis', 'bacteriocin' and other key phage structural genes. Regions with at least six key proteins within a moving window of 60 proteins are considered putative prophage regions. The borders of the prophage region are determined as the positions of the first and the last occurred key proteins. Because the integrase enzyme typically encoded within temperate phages usually determines the specificity of integration sites (Williams KP, 2002), putative attachment sites will be examined for putative prophage clusters containing integrase. Using the integrase protein as an anchor, as each cluster contains an integrase, the sequences of 10 upstream and downstream proteins in the cluster will be extracted to detect the putative attL-attR pairs (mismatch=0) using BLASTN with the parameters ‘-task blastn-short –evalue 1000’. The attL-attR pair with the highest bit score and length>=12 bp is considered the putative att sites of the prophage region. Finally, each prophage region is assigned a taxonomy by a majority vote based on the detected phage-like genes with annotated taxonomic information within the prophage.

## 2.2 Phage annotation for prophage

DBSCAN-SWA provides two ways to annotate infecting phages for the predicted prophage regions. If a candidate phage genome is given, DBSCAN-SWA will evaluate the similarity between the integrated prophage(s) and the phage genome(s) based on three prophage-related features, defined to evaluate the similarity between the integrated prophage(s) and the phage genomes based on homologous protein alignment by Diamond BLASTP and nucleotide alignment by BLASTN (Supplementary Table S2). Alternatively, users can choose the local phage genome and protein database (PGPD) containing 10,463 complete phage genome sequences and 684,292 nonredundant phage proteins collected from millardlab (http://millardlab.org/bioinformatics/bacteriophage-genomes/) to predict the infecting phages by a Diamond BLASTP and a BLASTN search against PGPD.

# 3 Results

DBSCAN-SWA is an integrated tool for the detection of prophages that combines ORF prediction and gene function annotation, phage-like gene cluster detection, attachment site identification, and infecting phage annotation **(**Fig. 1A) with user-friendly outputs **(**Fig. 1B, 1C). To evaluate the performance of DBSCAN-SWA, 184 manually curated prophages from 50 complete bacterial genomes (Casjens et al., 2003) are collected to examine the prophage prediction capability based on recall **(**the number of correctly predicted prophages detected divided by the number of total prophages) and precision (the number of correctly predicted prophages divided by the number of total predicted prophages). The results show that DBSCAN-SWA performs the best with recalls of 84% or 85% for GenBank annotated sequences or raw DNA sequences, compared to PHASTER with recall of 86% or 82% and Phage_Finder and VirSorter with recall of 63% and 74% (Supplementary Table S1). With the best recall, DBSCAN-SWA shows a highly nonconservative predictive power with a precision of ~0.45. This means that DBSCAN-SWA fully considers the enrichment of viral genes in a putative prophage and will probably identify additional putative active prophages that can be induced by physiological cues. Moreover, DBSCAN-SWA presents better predictive power in NGS data than Phage_Finder and VirSorter based on the analysis of 19,989 contigs of 400 bacterial genomes in the human gastrointestinal tract collected from HMP (https://www.hmpdacc.org/hmp). DBSCAN-SWA predicts 2253 prophages on 1469 contigs from 389 bacterial genomes in approximately 13 hours with a detection rate (the percentage of bacterial genomes with putative prophages detected) of (389/400) 97%, while Phage_Finder predicts 580 prophages from 261 bacterial genomes in approximately 14 hours with a detection rate of (320/400) 80%. Compared to VirSorter, DBSCAN-SWA runs 6 times faster than VirSorter, which takes approximately 63 hours to predict 3016 prophages from 384 bacterial genomes (Supplementary Table S3, S4, S5). Simultaneously, DBSCAN-SWA also has a good degree of agreement with the prediction results of Phage_Finder, sharing 433 prophages with Phage_Finder (433/580=74.7%) and 1186 prophages with VirSorter (1186/3016=39.3%). It proves that DBSCN-SWA can predict putative prophages for increasingly high-throughput sequencing data and outperforms the existing standalone prophage prediction tools in terms of either efficiency or predictability. In summary, the highlights of DBSCAN-SWA are as follows:

- **High efficiency**. DBSCAN-SWA takes approximately 1.35 min~6.8 min to detect prophages in complete bacterial genomes (1.2 Mbp~7 Mbp)**.**
- **High recall.** DBSCAN-SWA achieves an excellent recall of 100% for *Xylella fastidiosa* Temecula1 using raw DNA sequences, while PHASTER obtains only 71.4% (Fig. 1B).
- **Suitable for high-throughput sequencing data.** DBSCAN-SWA was compared to VirSorter and Phage_Finder, the only tools with standalone versions. DBSCAN-SWA showed better performance in efficiency, installation and usage than VirSorter, which is difficult to install based on a complex configuration environment, whereas DBSCAN-SWA is well packaged and easy to install. In addition, DBSCAN-SWA is suitable for both complete and incomplete sequenced genomes, while Phage_Finder is fit only for complete ones.
- **Provide phage annotation**. DBSCAN-SWA provides a custom phage database to facilitate the annotation of prophage regions. In addition, the similarity between the integrated prophage(s) and the phage genome(s) can be evaluated based on three prophage-related features (Fig. 1C).
- **User-friendly visualizations.** DBSCAN-SWA provides a user-friendly interactive HTML page to browse prophages in a genome viewer and detailed prophage information and bacterium-phage interactions in tables (Fig. 1 B, 1C).
- **Freely modified parameters.** DBSCAN-SWA enables users to adjust the parameters for phage-like protein identification, att site identification and phage annotation to meet their requirements for the prediction results based on their knowledge of prophages and phage-host interactions.

**Table 1.** Comparison of DBSCAN-SWA with other prophage detection tools

| | DBSCAN-SWA | Prophage Hunter | PHASTER | Phage_Finder | VirSorter |
|---|---|---|---|---|---|
| Last updated | 2020 | 2019 | 2016 | 2006 | 2015 |
| Input type | FASTA/GBK | FASTA | FASTA/GBK | Special format | FASTA |

| Timing | ~1 min | N/A | Slow | ~2 min | Slow |
|---|---|---|---|---|---|
| Standalone | YES | NO | NO | YES | YES |
| Interactive | YES | YES | YES | NO | YES |
| att site prediction | YES | YES | YES | NO | NO |
| Gene annotation | YES | YES | NO | YES | NO |
| Recall | 100% | N/A | ~71% | ~57% | ~57% |

N/A means more tests are needed. Timing was tested on a Linux platform for *Xylella fastidiosa* Temecula1, which has a genome of approximately 2.5 Mbp (NC_004556). Slow depends on the queueing time. No in 'standalone' means only a webserver is provided The detection rates were calculated for *Xylella fastidiosa* Temecula1 using raw DNA sequences.

## References

Panis G, Duverger Y, Courvoisier-Dezord E, Champ S, Talla E, Ansaldi M (2010) Tight Regulation of the intS Gene of the KplE1 Prophage: A New Paradigm for Integrase Gene Regulation. PLoS Genet 6(10): e1001149.

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., and Wishart, D. H. (2011). Phast: a fast phage search tool. Nucleic Acids Res. 39, 347–352. doi: 10.1093/nar/gkr485

Fouts,D.E. (2006) Phage_Finder: automated identifification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Res., 34, 5839–5851.

Laslett D, Canback B (2004), ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic acids research, 32(1):11-16.

Roux et al. (2015), VirSorter: mining viral signal from microbial genomic data. PeerJ 3:e985; DOI 10.7717/peerj.985

Wishart,D.S. (2016) PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res., 44, W16–W21.

Wenchen Song, Hai-Xi Sun, Carolyn Zhang, Li Cheng, Ye Peng, Ziqing Deng, Dan Wang, Yun Wang, Ming Hu, Wenen Liu, Huanming Yang, Yue Shen, Junhua Li, Lingchong You, Minfeng Xiao, Prophage Hunter: an integrative hunting tool for active prophages, Nucleic Acids Research, Volume 47, Issue W1, 02 July 2019, Pages W74–W80, https://doi.org/10.1093/nar/gkz380.

Torsten Seemann (2014), Prokka: rapid prokaryotic genome annotation, Bioinformatics, Volume 30, Issue 14, Pages 2068–2069.

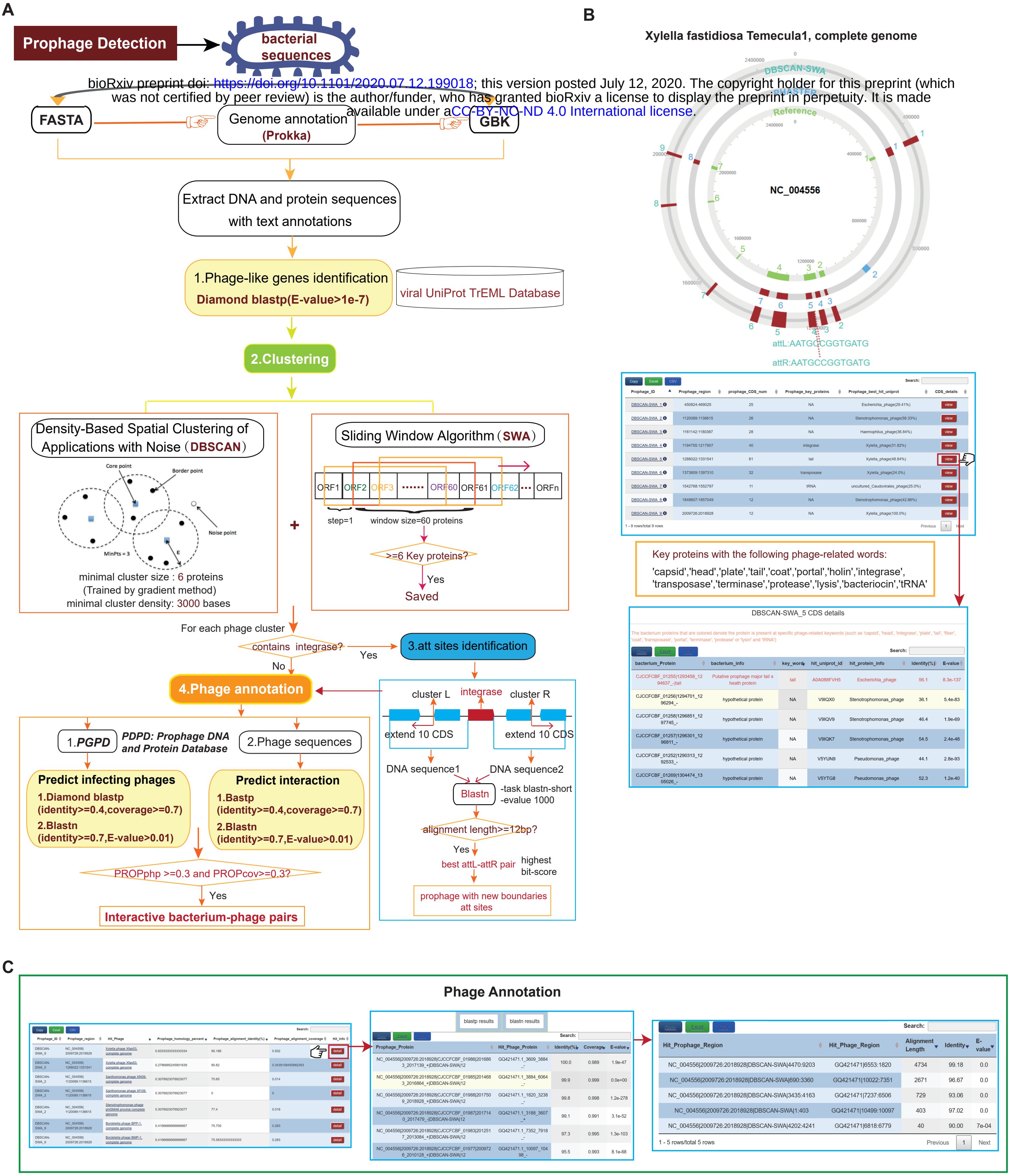
Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nature methods, 12(1):59-60.

UniProt C. 2014. Activities at the universal protein resource (UniProt). Nucleic Acids Res 42:D191–D198. doi:10.1093/nar/gkt1140.

Williams KP (2002) Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. Nucleic Acids Res 30: 866–875.

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.

Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far? molecular microbi- ology. Nucleic Acids Res. 49, 277–300.


A
Prophage Detection
bacterial sequences
FASTA
Genome annotation (Prokka)
GBK
Extract DNA and protein sequences with text annotations
1.Phage-like genes identification
Diamond blastp(E-value>1e-7)
viral UniProt TrEML Database
2.Clustering
Density-Based Spatial Clustering of Applications with Noise（DBSCAN）
Core point
Border point
Noise point
MinPts = 3
ε
minimal cluster size : 6 proteins
(Trained by gradient method)
minimal cluster density: 3000 bases
+
Sliding Window Algorithm（SWA）
ORF1
ORF2
ORF3
ORF60
ORF61
ORF62
ORFn
step=1
window size=60 proteins
>=6 Key proteins?
Yes
Saved
For each phage cluster
contains integrase?
Yes
No
3.att sites identification
cluster L
integrase
cluster R
extend 10 CDS
extend 10 CDS
DNA sequence1
DNA sequence2
Blastn
-task blastn-short
-evalue 1000
alignment length>=12bp?
Yes
best attL-attR pair
highest bit-score
prophage with new boundaries att sites
4.Phage annotation
1.PGPD
PDPD: Prophage DNA and Protein Database
2.Phage sequences
Predict infecting phages
1.Diamond blastp (identity>=0.4,coverage>=0.7)
2.Blastn (identity>=0.7,E-value>0.01)
Predict interaction
1.Bastp (identity>=0.4,coverage>=0.7)
2.Blastn (identity>=0.7,E-value>0.01)
PROPphp >=0.3 and PROPcov>=0.3?
Yes
Interactive bacterium-phage pairs
B
Xylella fastidiosa Temecula1, complete genome
DBSCAN-SWA
PHASTER
Reference
NC_004556
0
400000
800000
1200000
1600000
2000000
2400000
1
2
3
4
5
6
7
8
9
attL:AATGCCGGTGATG
attR:AATGCCGGTGATG
C
Phage Annotation

Copy Excel CSV Search:

| Prophage_ID | Prophage_region | prophage_CDS_num | Prophage_key_proteins | Prophage_best_hit_uniprot | CDS_details |
|---|---|---|---|---|---|
| DBSCAN-SWA_1 | 450824:469025 | 25 | NA | Escherichia_phage(29.41%) | view |
| DBSCAN-SWA_2 | 1120089:1136615 | 26 | NA | Stenotrophomonas_phage(58.33%) | view |
| DBSCAN-SWA_3 | 1161142:1180387 | 28 | NA | Haemophilus_phage(36.84%) | view |
| DBSCAN-SWA_4 | 1194755:1217907 | 40 | integrase | Xylella_phage(31.82%) | view |
| DBSCAN-SWA_5 | 1288022:1331541 | 61 | tail | Xylella_phage(48.84%) | view |
| DBSCAN-SWA_6 | 1373809:1397310 | 32 | transposase | Xylella_phage(24.0%) | view |
| DBSCAN-SWA_7 | 1542768:1552797 | 11 | tRNA | uncultured_Caudovirales_phage(25.0%) | view |
| DBSCAN-SWA_8 | 1848607:1857049 | 12 | NA | Stenotrophomonas_phage(42.86%) | view |
| DBSCAN-SWA_9 | 2009726:2018928 | 12 | NA | Xylella_phage(100.0%) | view |

1 - 9 rows/total 9 rows Previous 1 Next

Key proteins with the following phage-related words:

'capsid','head','plate','tail','coat','portal','holin','integrase', 'transposase','terminase','protease','lysis','bacteriocin','tRNA'

DBSCAN-SWA_5 CDS details

The bacterium proteins that are colored denote the protein is present at specific phage-related keywords (such as 'capsid', 'head', 'integrase', 'plate', 'tail', 'fiber', 'coat', 'transposase', 'portal', 'terminase', 'protease' or 'lysin' and 'tRNA')

Copy Excel CSV Search:

| bacterium_Protein | bacterium_info | key_word | hit_uniprot_id | hit_protein_info | Identity(%) | E-value |
|---|---|---|---|---|---|---|
| CJCCFCBF_01255\|1293458_1294637_-\|tail | Putative prophage major tail sheath protein | tail | A0A088FVH5 | Escherichia_phage | 56.1 | 8.3e-137 |
| CJCCFCBF_01256\|1294701_1296294_- | hypothetical protein | NA | V9IQX0 | Stenotrophomonas_phage | 36.1 | 5.4e-83 |
| CJCCFCBF_01258\|1296851_1297745_- | hypothetical protein | NA | V9IQV9 | Stenotrophomonas_phage | 46.4 | 1.9e-69 |
| CJCCFCBF_01257\|1296301_1296811_- | hypothetical protein | NA | V9IQK7 | Stenotrophomonas_phage | 54.5 | 2.4e-48 |
| CJCCFCBF_01252\|1290313_1292533_- | hypothetical protein | NA | V5YUN9 | Pseudomonas_phage | 44.1 | 2.8e-93 |
| CJCCFCBF_01269\|1304474_1305026_- | hypothetical protein | NA | V5YTG8 | Pseudomonas_phage | 52.3 | 1.2e-40 |

Copy Excel CSV Search:

| Prophage_ID | Prophage_region | Hit_Phage | Prophage_homology_percent | Prophage_alignment_identity(%) | Prophage_alignment_coverage | Hit_info |
|---|---|---|---|---|---|---|
| DBSCAN-SWA_9 | NC_004556\|2009726:2018928 | Xylella phage Xfas53, complete genome | 0.8333333333333334 | 95.186 | 0.932 | detail |
| DBSCAN-SWA_5 | NC_004556\|1288022:1331541 | Xylella phage Xfas53, complete genome | 0.2786885245901639 | 90.82 | 0.3439108455882353 | detail |
| DBSCAN-SWA_2 | NC_004556\|1120089:1136615 | Xanthomonas phage Xf409, complete genome | 0.3076923076923077 | 75.65 | 0.014 | detail |
| DBSCAN-SWA_2 | NC_004556\|1120089:1136615 | Xanthomonas phage Xf109, complete genome | 0.3076923076923077 | 0 | 0 | detail |
| DBSCAN-SWA_2 | NC_004556\|1120089:1136615 | Stenotrophomonas phage phiSMA6 proviral complete genome | 0.3076923076923077 | 77.4 | 0.018 | detail |
| DBSCAN-SWA_9 | NC_004556\|2009726:2018928 | Bordetella phage BPP-1, complete genome | 0.4166666666666667 | 75.705 | 0.283 | detail |
| DBSCAN-SWA_9 | NC_004556\|2009726:2018928 | Bordetella phage BMP-1, complete genome | 0.4166666666666667 | 75.58333333333333 | 0.283 | detail |

blastp results blastn results

Copy Excel CSV Search:

| Prophage_Protein | Hit_Phage_Protein | Identity(%) | Coverage | E-value |
|---|---|---|---|---|
| NC_004556\|2009726:2018928\|CJCCFCBF_01986\|2016863_2017139_+\|DBSCAN-SWA\|12 | GQ421471.1_3609_3884_- | 100.0 | 0.989 | 1.9e-47 |
| NC_004556\|2009726:2018928\|CJCCFCBF_01985\|2014683_2016864_+\|DBSCAN-SWA\|12 | GQ421471.1_3884_6064_- | 99.9 | 0.999 | 0.0e+00 |
| NC_004556\|2009726:2018928\|CJCCFCBF_01988\|2017509_2018928_+\|DBSCAN-SWA\|12 | GQ421471.1_1820_3238_- | 99.8 | 0.998 | 1.2e-278 |
| NC_004556\|2009726:2018928\|CJCCFCBF_01987\|2017140_2017479_-\|DBSCAN-SWA\|12 | GQ421471.1_3188_3607_+ | 99.1 | 0.991 | 3.1e-52 |
| NC_004556\|2009726:2018928\|CJCCFCBF_01983\|2012517_2013084_+\|DBSCAN-SWA\|12 | GQ421471.1_7352_7918_- | 97.3 | 0.995 | 1.3e-103 |
| NC_004556\|2009726:2018928\|CJCCFCBF_01977\|2009726_2010128_+\|DBSCAN-SWA\|12 | GQ421471.1_10097_10498_- | 95.5 | 0.993 | 8.1e-68 |

Copy Excel CSV Search:

| Hit_Prophage_Region | Hit_Phage_Region | Alignment Length | Identity | E-value |
|---|---|---|---|---|
| NC_004556\|2009726:2018928\|DBSCAN-SWA\|4470:9203 | GQ421471\|6553:1820 | 4734 | 99.18 | 0.0 |
| NC_004556\|2009726:2018928\|DBSCAN-SWA\|690:3360 | GQ421471\|10022:7351 | 2671 | 96.67 | 0.0 |
| NC_004556\|2009726:2018928\|DBSCAN-SWA\|3435:4163 | GQ421471\|7237:6506 | 729 | 93.06 | 0.0 |
| NC_004556\|2009726:2018928\|DBSCAN-SWA\|1:403 | GQ421471\|10499:10097 | 403 | 97.02 | 0.0 |
| NC_004556\|2009726:2018928\|DBSCAN-SWA\|4202:4241 | GQ421471\|6818:6779 | 40 | 90.00 | 7e-04 |

1 - 5 rows/total 5 rows Previous 1 Next