bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.090266; this version posted May 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SpacePHARER: Sensitive identification of phages from CRISPR spacers in prokaryotic hosts

Zhang R.,[1] Mirdita M.,[1] Levy Karin E.,[1] Norroy C.,[1] Galiez C.,[1,2] and Söding J.[1,3]

[1] *Quantitative and Computational Biology, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany*
[2] *Univ. Grenoble Alpes, CNRS, Grenoble INP/Institute of Engineering Univ. Grenoble Alpes, Grenoble, France*
[3] *soeding@mpibpc.mpg.de*

Summary: SpacePHARER (CRISPR Spacer Phage-Host Pair Finder) is a sensitive and fast tool for *de novo* prediction of phage-host relationships via identifying phage genomes that match CRISPR spacers in genomic or metagenomic data. SpacePHARER gains sensitivity by comparing spacers and phages at the protein-level, optimizing its scores for matching very short sequences, and combining evidences from multiple matches, while controlling for false positives. We demonstrate SpacePHARER by searching a comprehensive spacer list against all complete phage genomes.
Availability and implementation: SpacePHARER is available as an open-source (GPLv3), user-friendly command-line software for Linux and macOS at **spacepharer.soedinglab.org**.

I. INTRODUCTION

Viruses of bacteria and archaea (phages) are the most abundant biological entities in nature. However, little is known about their roles in the microbial ecosystem and how they interact with their hosts, as cultivating most phages and hosts in the lab is challenging. Many prokaryotes possess an adaptive immune system against phages, the Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) system. After surviving a phage infection, they incorporate a short DNA fragment (32-38 nt) as a spacer in a CRISPR array. The transcribed spacer will be used with other Cas components for a targeted destruction of future invaders. Some CRISPR-Cas systems require a 2-6 nucleotide long, highly conserved protospacer-adjacent motif (PAM) flanking the viral target to prevent autoimmunity. Multiple spacers targeting the same invader are not uncommon, due to either multiple infection events or the primed spacer acquisition mechanism identified in some CRISPR subtypes. CRISPR spacers have been previously exploited to identify phage-host relationship [7, 8, 11]. These methods compare individual CRISPR spacers with phage genomes using BLASTN [1] and apply stringent filtering criteria, e.g. allowing only up to two mismatches. They are thus limited to identifying very close matches. However, a higher sensitivity is crucial because phage reference databases are very incomplete and often will not contain phages highly similar to those to be identified. To increase sensitivity, (1) we compare protein instead of nucleotide sequences because phage genomes are mostly coding, and, to evade the CRISPR immune response, they are under pressure to mutate their nucleotides with minimal change on the amino acid level. (2) We optimized the substitution matrix and gap penalties for short, highly similar protein fragments. (3) We combine evidence from multiple spacers matching to the same phage genome.

II. METHODS

SpacePHARER accepts spacer sequences as multiple FASTA files each containing spacers from a single prokaryotic genome or as multiple output files from the CRISPR detection tools PILER-CR [5], CRISPR Recognition tool (CRT) [4], MinCED [9] or CRISPRDetect [3]. Phage genomes are supplied as separate FASTA files or can be downloaded by SpacePHARER from NCBI GenBank [2].

Algorithm. SpacePHARER is divided into five steps (**Figure 1A**, **Supp. Materials**). (0) Preprocess input: scan the phage genome and CRISPR spacers in six reading frames, extract and translate all putative coding fragments of at least 27 nt. Each query set Q consists of the translated ORFs q of CRISPR spacers extracted from one prokaryotic genome, and each target set T comprises the putative protein sequences t from a single phage. We refer to similar q and t as *hit*, and an identified host-phage relationship $Q-T$ as *match*. (1) Search all q's against all t's using the fast, sensitive MMseqs2 [10], with VTML40 substitution matrix [6], gap open cost of 16 and extension cost of 2. We optimized a short, spaced k-mer pattern for the prefilter stage (10111011) with six informative ('1') positions. (2) For each $q-T$ pair, compute the P-value for the best hit p_{bh} from first-order statistics. (3) Compute a combined score S_{comb} from best-hit P-values of multiple hits between Q and T using a modified truncated-product method (**Supp. Materials**). (4) Compute the false discovery rate (FDR = FP /(TP + FP)) and only retain matches with FDR < 0.05. For that purpose, SpacePHARER is run on a null model database and the fraction of null matches with S_{comb} below a cutoff (empirical P-value) is used to estimate the FDR. (5) Scan 10 nt upstream and downstream of the phage's protospacer for a putative PAM.

Output is a tab-separated text file. Each host-phage match spans two or more lines. The first starts with '#': prokaryote accession, phage accession, S_{comb}, number of hits in the match. Each following line describes an individual hit: spacer accession, phage accession, p_{bh}, spacer start and end, phage start and end, putative 5' PAM|putative 3' PAM. Optionally, the spacer–phage sequence alignment can be included.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.090266; this version posted May 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

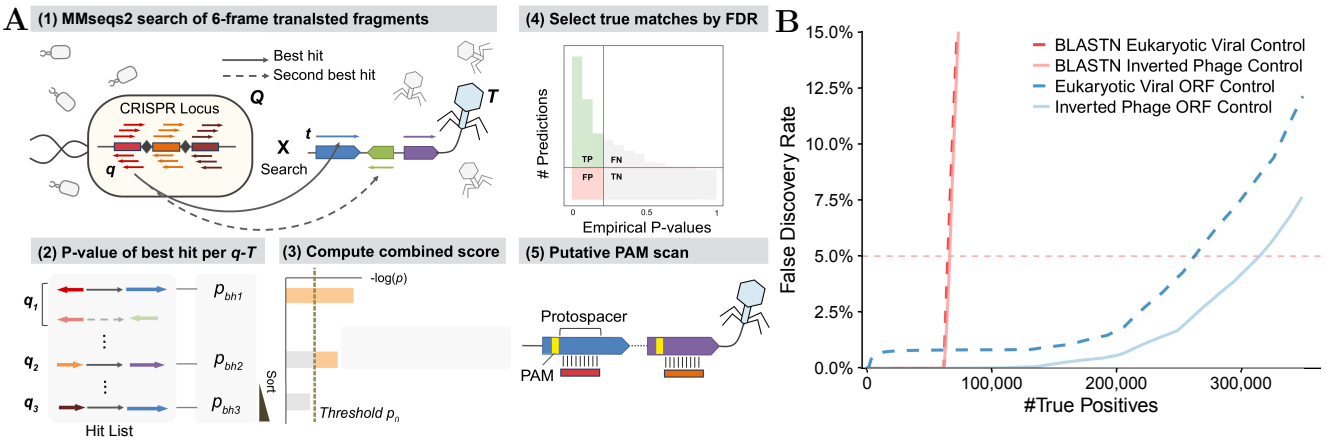


FIG. 1. (**A**) SpacePHARER algorithm. A query set Q consists of 6-frame translated ORFs (q) from CRISPR spacers, and a target set T consists of 6-frame translated ORFs (t) of phage proteins. (1) Search all qs against all ts using MMseqs2. (2) For each $q-T$ pair, compute the P-value for the best hit from first-order statistics. (3) Compute score S_{comb} by combining the best-hit P-values from multiple hits between Q and T using a modified truncated-product method. (4) Estimate the FDR by searching a null database. (5) Scan for putative protospacer adjacent motif (PAM). (**B**) Performance comparison between SpacePHARER (blue) and BLASTN (red) using inverted phage sequences (solid lines) or eukaryotic viral ORFs as null set (dashed lines) demonstrated by expected number of true positive (TP) predictions at different false discovery rates (FDRs). The inflated FDR when using the eukaryotic null database (dashed blue line) is caused by prokaryotic viruses mis-annotated as eukaryotic (Suppl. Material, section IV).

III. RESULTS

Datasets. We split a previously published spacer dataset [8] of 363,460 unique spacers from 30,389 prokaryotic genomes randomly into an optimization set (20%, 6,067 genomes) and a test set (80%, 24,322 genomes). The performance of SpacePHARER was evaluated on the spacer test set against a target database of 7,824 phage genomes. We used two null databases: 11,304 eukaryotic viral genomes and the inverted translated sequences of the target database. Viral genomes were downloaded from GenBank in 09/2018.

Prediction quality. At FDR = 0.05, SpacePHARER predicted 319,029 prokaryote-phage matches using the inverted phage sequences as null model and 253,419 using the eukaryotic viruses (**Figure 1B**), 4 to 5 times more than BLASTN (65,712 matches using inverted phage sequences and 62,804 using eukaryotic viruses).

Run time. SpacePHARER took 12 minutes to process the test dataset on 2×6-core 2.40 GHz CPUs, 13 times faster than BLASTN (160 minutes).

IV. CONCLUSION

SpacePHARER is over 4× more sensitive than BLASTN in detecting phage-host pairs, thanks to searching with protein sequences, optimizing short sequence comparisons, and combining statistical evidence. SpacePHARER is also fast enough to analyze large-scale genomic and metagenomic datasets.

FUNDING

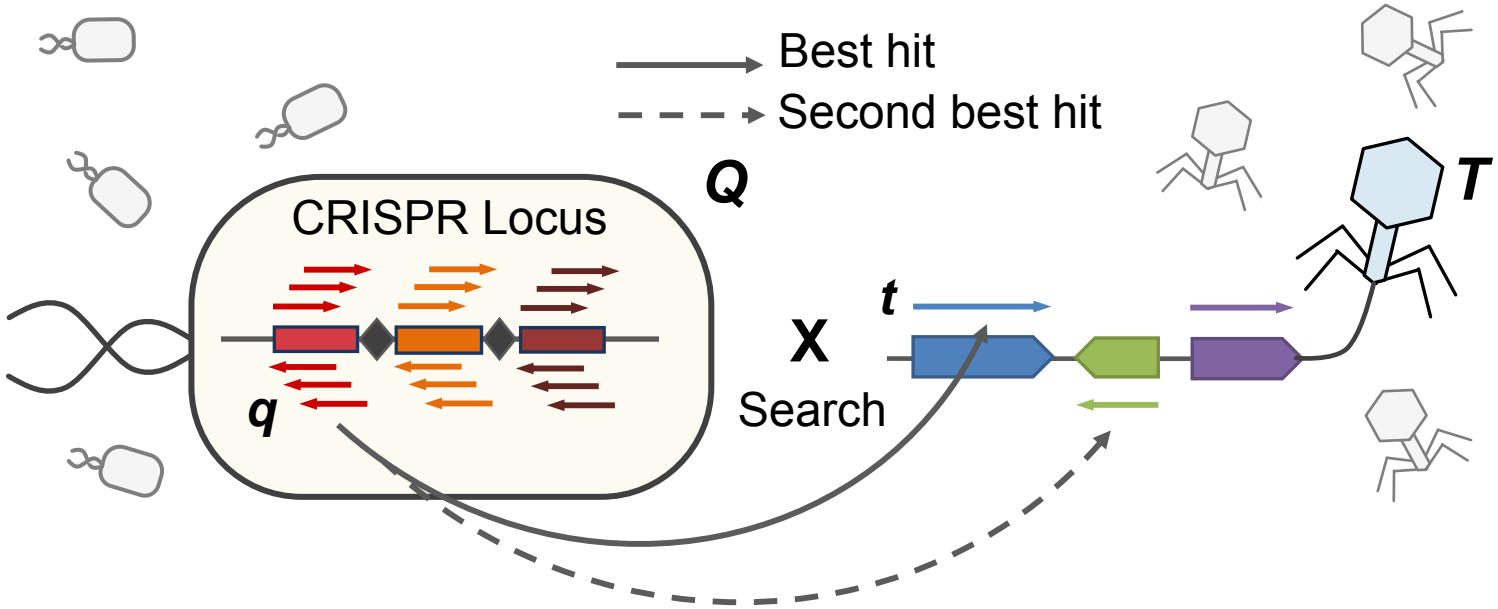
ELK is a FEBS long-term fellowship recipient. The work was supported by the ERC's Horizon 2020 Framework Programme (grant 'Virus-X', project no. 685778) and the BMBF CompLS project horizontal4meta.

Conflict of Interest: none declared

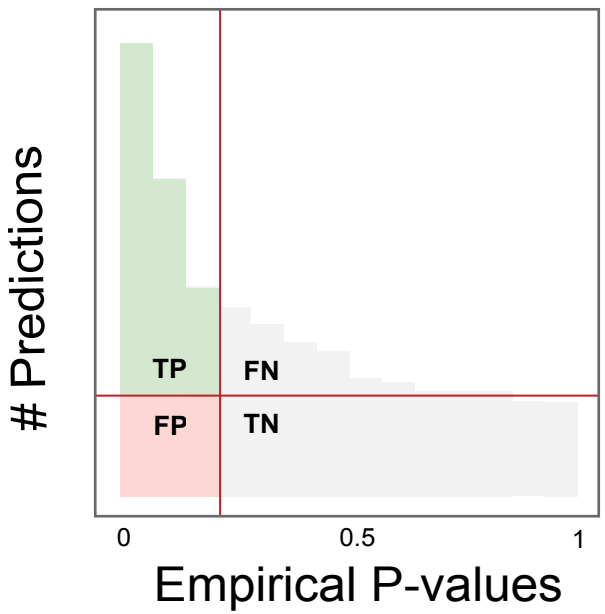
REFERENCES

[1] Altschul, S.F. et al (1990). Basic local alignment search tool. *J. Mol. Biol.*, **215**(3), 403–410.
[2] Benson, D.A. et al (2013). GenBank. *Nucleic Acids Res.*, **41**(D1), D36–D42.
[3] Biswas, A. et al (2016). CRSIPRdetect: A flexible algorithm to define CRSIPR arrays. *BMC Genom.*, **17**(1), 356.
[4] Bland, C. et al (2007). CRSIPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinform.*, **8**(1), 209.
[5] Edgar, R.C. (2007). PILER-CR: Fast and accurate identification of crispr repeats. *BMC Bioinform.*, **8**(1), 18.
[6] Müller, T. et al (2002). Estimating amino acid substitution models: A comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Mol. Biol. Evol.*, **19**(1), 8–13.
[7] Paez-Espino, D. et al (2016). Uncovering Earth's virome. *Nature*, **536**(7617), 425–430.
[8] Shmakov, S.A. et al (2017). The CRSIPR spacer space is dominated by sequences from species-specific mobilomes. *mBio*, **8**(5), e01397–17.
[9] Skennerton, C. (2016). Minced - mining CRSIPRs in environmental datasets. `https://github.com/ctSkennerton/minced`.
[10] Steinegger, M. and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**(11), 1026–1028.
[11] Stern, A. et al (2012). CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res.*, **22**(10), 1985–1994.

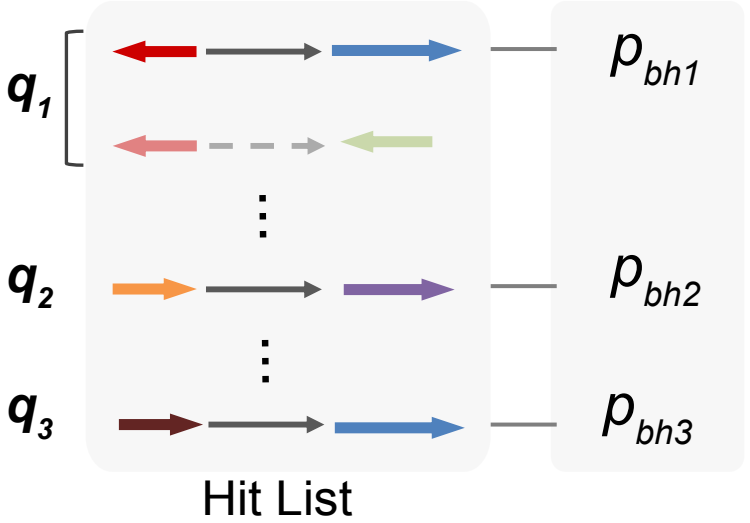
(1) MMseqs2 search of 6-frame tranalsted fragments



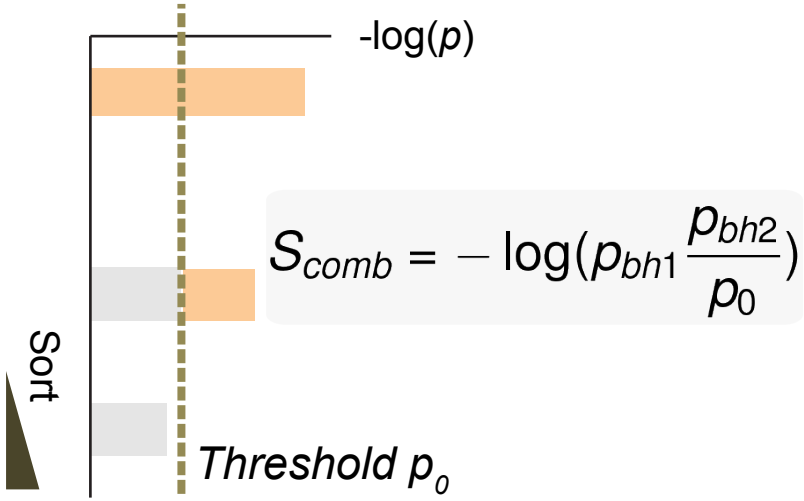
(4) Select true matches by FDR



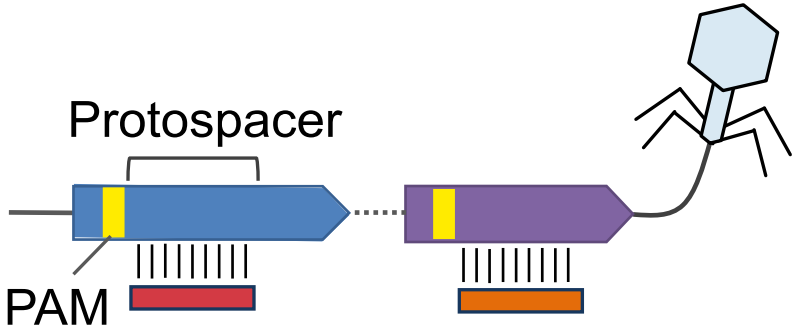
(2) P-value of best hit per *q*-*T*

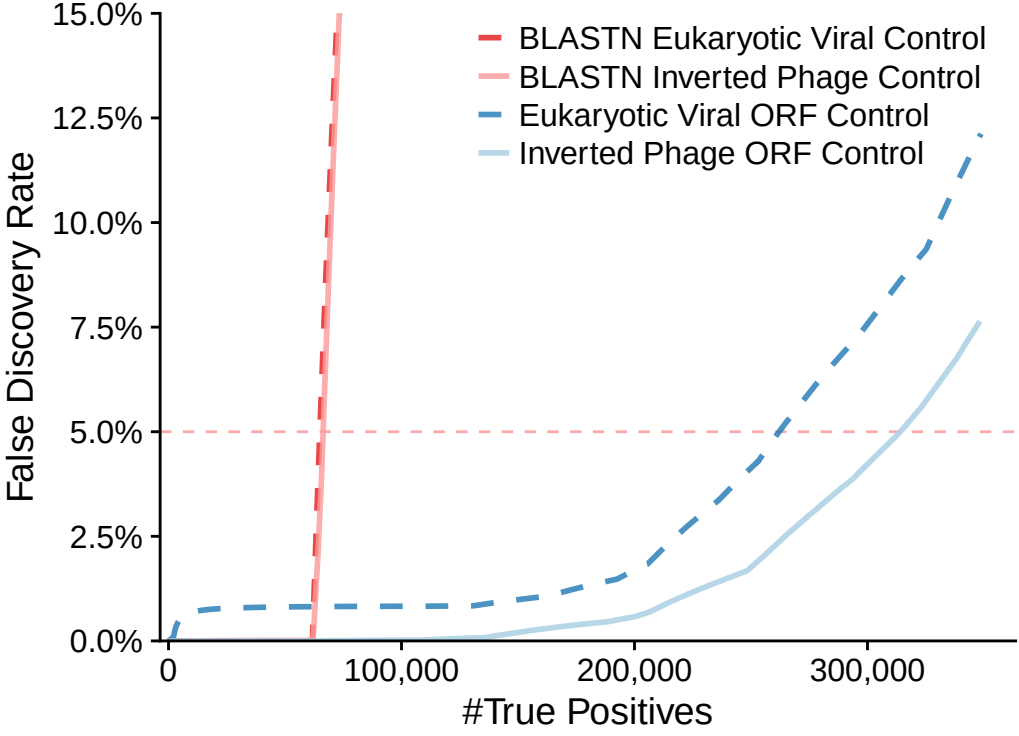


(3) Compute combined score



(5) Putative PAM scan




BLASTN Eukaryotic Viral Control
BLASTN Inverted Phage Control
Eukaryotic Viral ORF Control
Inverted Phage ORF Control
False Discovery Rate
15.0%
12.5%
10.0%
7.5%
5.0%
2.5%
0.0%
0
100,000
200,000
300,000
#True Positives