Phagonaute: A web-based interface for phage synteny browsing and protein function prediction

Hadrien Delattre, Oussema Souiai, Khema Fagoonee, Raphaël Guérois, Marie-Agnès Petit

Virology 496 (2016) 42–50

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

Phagonaute: A web-based interface for phage synteny browsing and protein function prediction

Hadrien Delattre [a], Oussema Souiai [a], Khema Fagoonee [a], Raphaël Guerois [b,*], Marie-Agnès Petit [a,**]

[a] Micalis Institute, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[b] I2BC, CEA, Université Paris-Saclay, 91198 Gif-sur-Yvette, France

ARTICLE INFO

Article history:
Received 18 March 2016
Returned to author for revisions
4 May 2016
Accepted 9 May 2016

Keywords:
Bacteriophages
Archaeal viruses
HSV1
Recombination functions

ABSTRACT

Distant homology search tools are of great help to predict viral protein functions. However, due to the lack of profile databases dedicated to viruses, they can lack sensitivity. We constructed HMM profiles for more than 80,000 proteins from both phages and archaeal viruses, and performed all pairwise comparisons with HHsearch program. The whole resulting database can be explored through a user-friendly "Phagonaute" interface to help predict functions. Results are displayed together with their genetic context, to strengthen inferences based on remote homology. Beyond function prediction, this tool permits detections of co-occurrences, often indicative of proteins completing a task together, and observation of conserved patterns across large evolutionary distances. As a test, Herpes simplex virus I was added to Phagonaute, and 25% of its proteome matched to bacterial or archaeal viral protein counterparts. Phagonaute should therefore help virologists in their quest for protein functions and evolutionary relationships.

1. Introduction

Common ancestries among living organisms are usually considered, molecularly, at the gene/protein level, but relationships sometimes hold true for larger genetic loci, as seen in structural modules in bacteriophages, metabolic operons in bacteria, or even large segments of chromosomes in mammals. Accompanying these discoveries, bioinformatics tools have been developed to help visualize such conserved blocks, also termed regions of synteny. Genomicus is a site dedicated to synteny browsing across eukaryotic chromosomes (Louis et al., 2013), whereas on the bacterial side, the Integrated Microbial Genome (IMG) site of the Joint Genome Institute provides a tool for neighborhood browsing among bacterial chromosomes (Markowitz et al., 2012). In the vast realm of viruses, such a tool is presently missing.

This may not be so surprising, as a specific difficulty resides in the method used for homology detection between viral proteins. Indeed, BLAST performs poorly on such proteins, because of their remarkable degree of divergence. For instance, the context of the *bet* gene (a phage gene for homologous recombination) of several *Escherichia coli* bacteriophages, is contrasted in Fig. 1 to the synteny around the *recA* gene (a bacterial gene for homologous recombination) from a set of bacteria belonging to the *Bacillus* genus. *Bacillus* genomes exhibit local synteny around *recA*, and *E. coli* phages appear to have none around *bet*. Still, a careful observation indicates that all genes surrounding *bet* have similar lengths and orientations, suggesting that BLAST "missed" the homology signal, which could be detected with distant homology detection tools.

Several ways to circumvent the homology detection problem have been developed in the recent years, among which the "profile against profile" comparison approach has proven its efficiency for phage proteins (Hardies et al., 2015; Lopes et al., 2010, 2014; Sabri et al., 2011). In such comparison algorithms, rather than aligning two protein sequences, two protein alignments are being compared. These starting alignments are also called "profiles" because they can easily be converted into Hidden-Markov Model (HMM) profiles describing which amino-acids are likely to be found at each position. In addition, in a profile-profile comparison algorithm such as HHsearch (Soding, 2005), secondary structure predictions are also taken into account. One of the difficulties with such programs is that they are run against databases of protein profiles that have to be precomputed (such as PDB or Pfam), rather than against standard protein sequence databases. Another difficulty is that among the distant homologs displayed by HHsearch, some may be false-positives, so that matches in the absence of synteny may be meaningless with respect to predicting function.

* Corresponding author.
** Corresponding author
E-mail addresses: marie-agnes.petit@jouy.inra.fr (R. Guerois), raphael.guerois@cea.fr (M.-A. Petit).

http://dx.doi.org/10.1016/j.virol.2016.05.007
0042-6822/

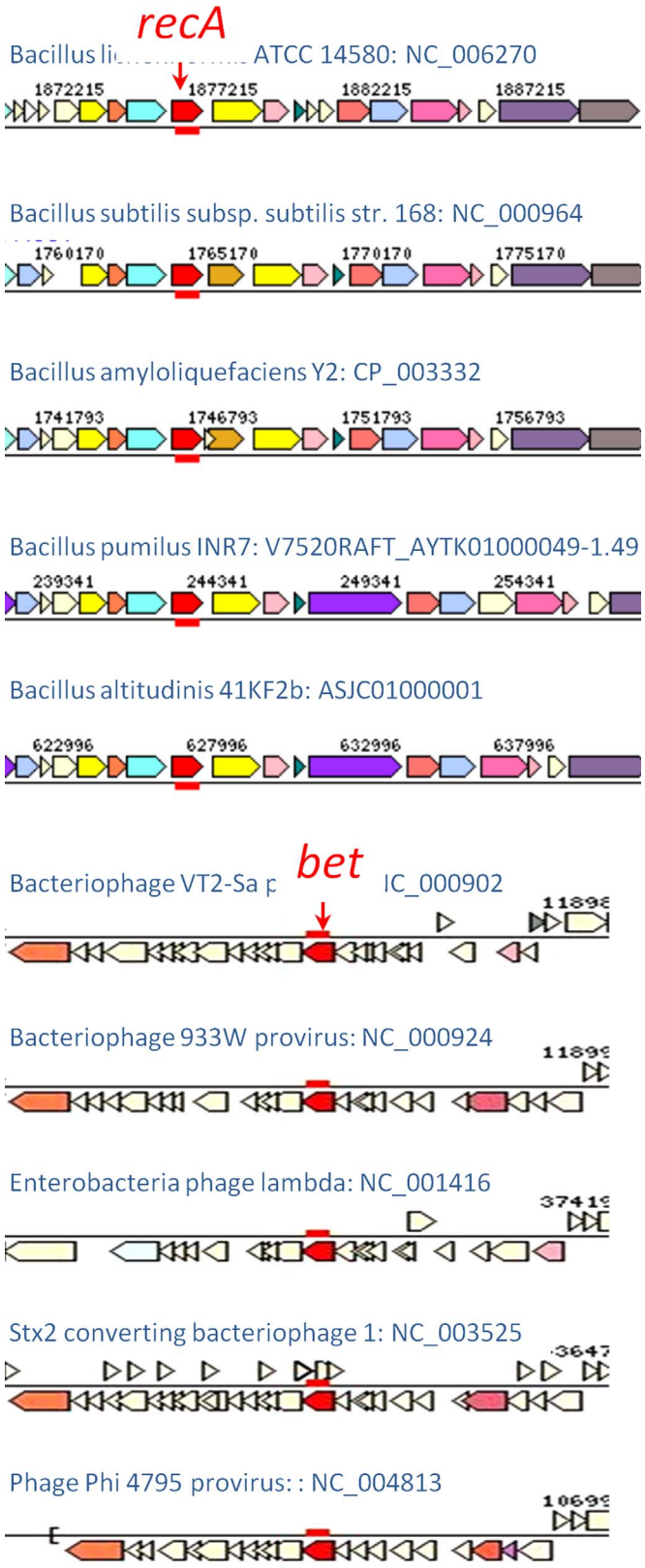

Fig. 1. A. Synteny around the *recA* gene in five *Bacillus* species, as recovered with the Integrated Microbial Genomes interface of JGI (https://img.jgi.doe.gov/). A region of the various genomes with a match to the query is shown horizontally, the query gene being centered and displayed in red. Other various colors are used to indicate, in the vicinity of the query, if conservation is present. Pale yellow genes are those for which no BLAST hit was recorded, among the genes drawn in the window. B. Absence of synteny around the phage gene *bet*, among five phages infecting *E. coli*, using the same IMG tool.

This motivated our attempt to combine in a single tool distant homology and synteny approaches, to reach predictive power.

No database of profiles dedicated to viral proteins was available at the onset of this work. To address this concern, we developed a web interface for synteny analyses among prokaryotic viral genomes, based on homology detection with HHsearch against a large collection of phage protein profiles. This web server is named Phagonaute, it allows navigation mainly across phage genomes (archaeal viruses represent 5% of the total), and we show that it permits to reach some unifying views among the realm of prokaryotic viral protein functions. As a test, we added the Herpes simplex virus I into Phagonaute, and report that 25% of its proteome match to bacterial or archaeal viral protein counterparts. We show that the Phagonaute tool, thanks to its possibility of interactive explorations, helps experimentalists decide in which direction to go for understanding a given phage protein function, and we illustrate its use with examples of successful function predictions.

2. Materials and methods

2.1. Phage and archaeal virus genomes

As a starting material, fully sequenced viral genomes available at the NCBI from viruses infecting prokaryotic hosts were collected (http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=10239&host=bacteria). They comprised 876 bacterial and 40 archaeal viruses (May 2014). This constituted the backbone of the PGC2 database (PGC stands for phage genome context). In addition a smaller database (hereafter PGC1), composed of the 441 phages, and 10 archaeal viruses present in the genome collection from the ACLAME database (Leplae et al., 2010), was also constructed, for comparison purposes. Indeed, this project was conceived as a continuation of the ACLAME effort to collect and organize available data on phages, prophages, and plasmids. We were particularly interested by the expert curated list of annotations performed at the protein family level (families constructed with both phages and prophages, named “fam:vir:proph” on the ACLAME website (http://aclame.ulb.ac.be/). Compared to ACLAME, the present work focused exclusively on active phages whose complete genome has received an NC number at the NCBI. The list of genomes present in PGC1 is available at http://genome.jouy.inra.fr/phagonaute/page4.html (and page5.html for PGC2). Both databases offer complementary interests: PGC2 subsumes PGC1 and is richer in protein content; it may therefore contain the protein the user is interested in. But it does not offer the ACLAME annotation level, and its outputs are sometimes so large that the eye no longer sees the useful informations. PGC1 has additional annotation content through ACLAME, and its limited size sometimes permits a better focus on critical information.

2.2. Profile constructions, intra-phage HHsearch comparisons and storage of the results in a relational database

Protein processing is summarized in Fig. 2. Each phage protein of the above mentioned genomes was compared with Psi-BLAST (Altschul et al., 1997) to the non-redundant database aggregated at 70% identity (as such aggregations are calculation intensive, the aggregation of December 2010 available at toolkit.tuebingen.mpg.de was used), with three iterations and an E-value cut-off at 10^{-4}. When an iteration retrieved more than 1000 matches, the search was stopped at the previous iteration, to prevent retrieval of aberrant matches. The resulting multiple sequence alignment was filtered to keep the 100 most diverse sequences (hhfilter with option –diff) and converted into a HMM profile using the HHsuite

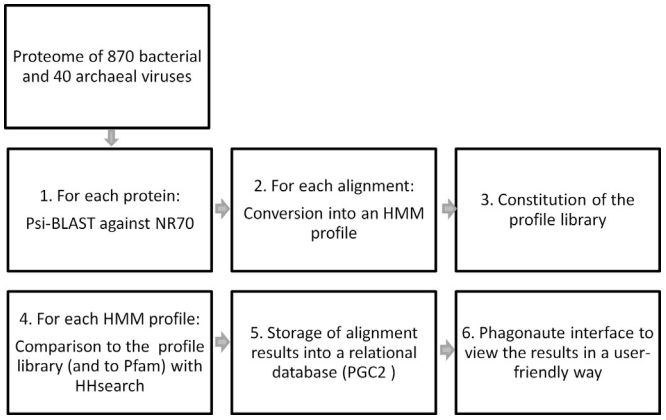

Fig. 2. Protein processing for Phagonaute. For each protein encoded by complete and active bacterial (870) and archaeal (41) viruses, a PSI-BLAST search against the Non-Redundant database aggregated at 70% identity (NR70) was performed (step 1). From the resulting alignment, an HMM profile was constructed (2), and all profiles were concatenated to constitute a viral profile library (3). Once this library was constructed, each individual profile was compared to all other profiles using HHsearch (4). In parallel, an HHsearch against Pfam (Bateman et al., 2000) was also conducted. Finally the coordinates of all matches, and the statistics associated with each match were stored in a database (5). The Phagonaute interface was next constructed to facilitate the view of the results (6). The strategy was identical for the smaller PGC1 database, except that 441 bacterial and 10 archaeal viruses were used as the starting proteome.

programs (Soding, 2005) integrating the secondary structure prediction from Psi-pred (Jones, 1999). Next, the profile built for each single protein (or the single protein itself for orphans) was compared to the set of all profiles with HHsearch (82,156 profiles altogether in PGC2, 28,230 in PGC1), and all match coordinates, with their associated confidence value, percentage of identity, E-value, and number of columns aligned, were stored in the PGC1 or PGC2 relational database. No cutoff was applied at this stage, so that even results with a very low confidence probability were stored. The confidence value indicates the probability that the match corresponds to *bona fide* homology: the closer to 100%, the better. Often, alignments with only 30% identical amino-acids reach a confidence probability of 100%, and alignments with 10–20% identical amino-acid may reach a confidence probability above 90%.

2.3. Comparisons of the phage protein profiles to Pfam

Phage genomes sometimes embark bacterial genes, it was therefore important to complement this "intra-phage" search by an orthogonal HHsearch comparison of all phage protein profiles against Pfam (as of May 2014), a generalist database organizing all available protein sequences into families (Bateman et al., 2000). To do so, the same outline as depicted in Fig. 2 was used, except that step 4 was replaced by a step 4′, where profiles were searched against the Pfam profile library. Family name and positions of all Pfam matches (with a confidence probability above 95%, so that only very confident Pfam assignments are considered) obtained for each phage protein were stored for the proteins of both PGC1 and PGC2 databases.

2.4. Result display, logic of the Phagonaute browser

The purpose of Phagonaute (http://genome.jouy.inra.fr/phagonaute) is to allow the user to perform a query on one of phage proteins present in the database, so as to retrieve its precomputed distant homologs within their genetic context. The query starts with a gene selected by the user ("query"), and collects related genes in the database. The process is iteratively repeated on the collected genes, and genes linked by a homology relationship (directly or transitively) will constitute a "family". In the output display, the color of a gene indicates its family. Once the family of the query has been collected, the same homology search is repeated among the neighboring genes, in order to look for co-occurrence and genetic context conservation around the query. Thus the family of the query serves as a backbone for the output visualization.

To permit a full exploration of HHsearch results, the visual representation of the matches to the selected query protein followed a double scope: (i) the first goal was to allow flexibility in the search of distant homologs. For this, the user can change the confidence cut-off of HHsearch results, and iteratively run HHsearch on the targets collected at each cycle, as is done with Psi-BLAST. Our earlier studies had shown that two iterations usually permit to saturate the search (Lopes et al., 2010; supplementary Fig. S1). Therefore, a default value of two iterations is given, but the user can change it. No limit to the number of displayed genomes is set. The list of all matches collected this way constitutes the backbone of the web page to be drawn, with all genes receiving the same color (or a combination of colors in cases of independent domain hits, or "fusion proteins", see below), and being placed vertically in the centre of the page. Genomes are ordered by confidence probability of the matches (shown in the pop-up window of the genome name), such that those with proteins closest to the query come first. (ii) The second scope was to consolidate the distant homology prediction by analysis of the synteny, i.e. the genetic organization on each phage, around the recovered matches. If HHsearch found a trace of homology between any two genes being displayed, this match would also be signaled by a new color.

This effort to connect proteins sharing potential homology was next combined to a display of annotation information, collected at three levels: Genbank file annotation, ACLAME family annotation (for PGC1 only, for PGC2, an update of ACLAME is first needed), and Pfam match. Above each gene, the three last letters of its locus or gene name are indicated. Hovering over any gene of the figure (or on the white area of the gene if partially colored) will display these three levels of information and highlight the family. To facilitate information compilation, the lower part of the result page contains a color legend, collecting and synthesizing information at three levels: for each colored gene, the three most frequent hits against Pfam, ACLAME families, and Genbank files (if different from unknown function) are indicated, allowing an 'at a glance' overview of potential functions. Finally, two types of outputs can be downloaded, (i) a table in tsv format (open with Excel or Libre Office) containing detailed information on all matches retrieved for the query gene, (ii) the figure displayed on the Phagonaute server. The figure format is svg (Scalable Vector Graphics), which can be imported into Libre Office and edited with Inkscape.

Since the confidence probability cut-off is not fixed in Phagonaute, the tool is not aimed at automatically annotating genomes. Phagonaute rather allows for interactive explorations, helping the experimentalist decide in which direction he should focus to understand a given phage protein function. In line with this philosophy, clicking on any gene of the page will place its information in the form, making Phagonaute ready for a next round of investigation by pushing again the "Visualize context" button.

2.5. Using the Phagonaute interface, options of the search

The user first selects in a drop down menu the phage genome of interest (phages are ordered by infected host, listed alphabetically). In case of a doubt concerning the host of the phage, a separate tab of the Phagonaute site gives access to the complete list of hosts and phages. Once a genome is selected, in the "gene" box

directly below, a drop down menu of all its genes, as recorded in the Genbank file (field "locus", fused with an hyphen to field "gene" when available), is displayed to help the user select one gene. Next to each gene, the HHsearch confidence probability associated with the best match that the encoded protein obtained during the distant homology search is given (and also displayed visually by shades of green). This confidence information will guide the user for selecting, in a next step, the confidence threshold for the homology search (99% by default). Finally, the user can change the number of neighbors to be displayed left and right of the target gene (3 by default). Clicking on the button "visualize context" will launch the computation needed to display the result page. The job takes usually 1–2 min, but sometimes more, for pages displaying hundreds of genomes. To reduce computing time, the confidence threshold can be increased, and the number of iterations/neighboring genes to be displayed can be decreased.

In line with the interactivity of the tool developed, Phagonaute offers some options. They are available upon further opening of two query windows. They help refine the search, and fine tune the final aspect of the figure. In the neighborhood section, a possibility is given to reduce the number of colors to be displayed on the final screen (useful when similar shades of color are present). The user chooses the minimum family size for color display. Another option in this window permits the display of the Pfam matches, rather than the "intra-phage" HHsearch matches set by default. Finally, in the HHsearch section, the number of iterations of HHsearch can be varied from 1 to 3. In the result page display, the number with a large font present above each central gene indicates at which stage of the homology search the genome was recovered: 1, query genome, 2, genome brought by HHsearch at first iteration, 3, genome brought by HHsearch at second iteration, etc…. Finally, the search can be restricted to some hosts only (they have to be spelled as in the scrolled-down phage list and separated by semicolons followed by a space), and cut-off parameters can be changed for the distant homology searches among neighboring genes.

Default parameters have been set so as to start with calculations of a short duration (only 3 neighbors displayed, probability set at 99%), while keeping sensitivity (2 iterations on central and neighboring genes), and increasing the chance to see synteny by setting probability cut-off at 95% on the neighbors.

2.6. The case of fusion proteins

Gene fusions occur in phages, and need to be dealt with for the visualization step. Otherwise, a query protein made of a fusion of a recombinase and nuclease, for instance (see Result and Discussion section), would lead to a representation where, among the matches, genes encoding both recombinases and nucleases would be displayed under the same color code. The concern was solved by declaring "domains" (i.e. sub-divisions) in a protein when the list of its matches (beside the full-length matches) constituted non-overlapping coordinate sets. To allow maximal sensitivity, even small matches of a size of 20 amino-acids are shown in the display (minimal size), and the pop-up window attached to the colored box indicates between brackets the boundaries of the match, in amino-acids. The domains displayed correspond to 'intra-phage' HHsearch matches, and they do not necessarily overlap with the matches obtained when the protein is compared to the Pfam database (to see these, tick the "display the Pfam matches" in the neighborhood parameters frame, as explained in the options paragraph above).

2.7. Quantification of orphans

To estimate improvement in homology detection brought by HHsearch analyses, the number of orphans (i.e. proteins not belonging to a family of at least two members), were compared between ACLAME version 0.4 (benefiting from PSI-BLAST searches and curated annotations) and our HHsearch-based PGC1 database, both containing the same set of viral proteins. For this comparison, a confidence threshold of 90% was chosen to define family membership in Phagonaute. A 21% increase in homology detection was obtained with HHsearch (26.8% of the 28,230 proteins were orphans in ACLAME, versus 21.2% in Phagonaute). It should be noted that whereas 90% is a safe confidence threshold for most distant homology searches with HHsearch (Lopes et al., 2010), lower thresholds can also be used and lead to interesting discoveries. We reported earlier that distant homologs of some constituents of the capsid connector, involved in the joining of head and tail parts of Caudoviridae, could be recovered at a lower, 70% confidence threshold, because the overall conservation of synteny allowed strengthening the homology prediction (Lopes et al., 2014). At this cut-off, orphans in PGC1 decrease to 16% of total proteins. In PGC2, the amount of orphans at the 90% cut-off was 13.9%, and decreased to 5.7%, upon lowering the cut-off to 70%. We conclude that using HHsearch improves homology detection among phage proteins, compared to ACLAME. Not surprisingly, the three-fold larger PGC2 database permits to reduce further the amounts of orphan proteins. Beyond incorporation of orphans into phage protein families, HHsearch is also a powerful tool to aggregate small families together, and to allow transfer of functional annotations, as shown earlier (Lopes et al., 2010, 2014) and developed in the Results section.

2.8. Extension of Phagonaute

Due to the heaviness of prior calculations, the Phagonaute website cannot add new phage genomes "on the fly". Three solutions are offered to users that do not find their phage genomes among the happy few. (1) Use BLAST against a generalist or dedicated database (such as ACLAME) to examine whether a close neighbor of the phage (or gene) of interest is not a member of Phagonaute, and then use this representative genome/gene as a starting point. (2) Use the Virfam site (http://biodev.extra.cea.fr/virfam) that we developed earlier to submit the phage of interest and find distant homologs of connector proteins and recombinase proteins. (3) Submit a request to the corresponding authors, to perform a unidirectional search for the phage of interest, as done for HSV1 in the Result section. These comparisons are unidirectional because a search starting from a gene submitted *a posteriori* is performed only against the fix profile library depicted Fig. 2 (box 3), at the exclusion of the genes from later phage additions. Furthermore, searches starting from one gene of the profile library will not display homologies to "late comer" genes, even if present. In addition to HSV1, some phages have already been added into Phagonaute using a unidirectional search: the CrassPhage (infecting an unknown host, (Dutilh et al., 2014)), the *E. coli* phi80 (Rotman et al., 2012), the jumbo *Bacillus megatherium* phage G (NC_023719, 497 kb) and several phages of *Streptococcus pyogenes* (phi1207.3, A25 and T12). The list of all unidirectional additions is displayed in the last section of the PGC2 genome list, on the Phagonaute web site.

3. Results and discussion

3.1. HHsearch favors detection of distant homology signals: illustration with recombinase genes

To test whether Phagonaute outputs (HHsearch-based homology detection) had increased sensitivity compared to IMG (BLAST-based homology detection), the *bet* gene of phage lambda

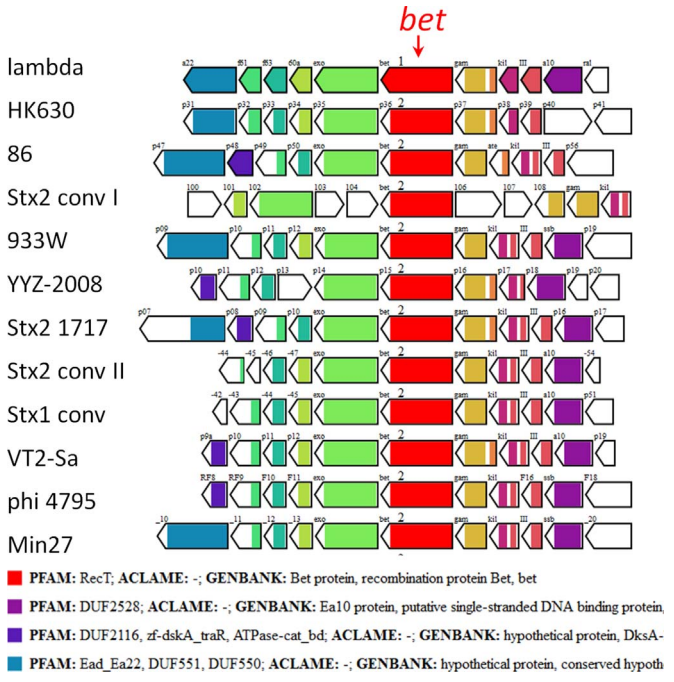

Fig. 3. Synteny around the *bet* gene (p84) of phage lambda, as detected with HHsearch and the Phagonaute interface, displaying 5 neighbors on each side of the query genes (and other parameters to their default value). All phages of Fig. 1 are present. For clarity, the figure is truncated and displays only 12 of the 72 genomes (names on the left side) of the result page, and the first 4 lanes of the color legend.

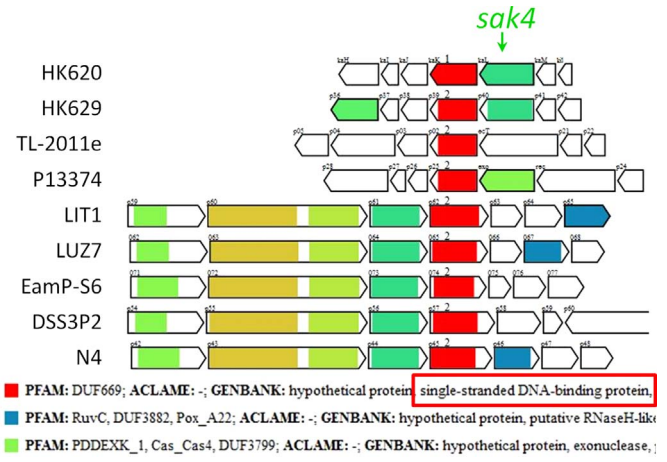

Fig. 4. Using co-occurrence to detect new functions. Starting from the *hkaK* gene (in red) with unknown function, encoded by phage HK620 (infecting *E. coli*), Phagonaute displays a large list of distant homologs, among which some are annotated as SSB (see legend). In addition, the *sak4* gene (also named *hkaL*, in green, right to *hkaK* in HK620) is often co-occurrent with this putative *ssb*. The Phagonaute search was run on PGC2 with a probability cut-off set at 95% for the central gene, and a single iteration on the central gene. The minimum number of members to color a family set to 5. Other parameters were by default. Image truncated, missing the last 32 genomes, and displaying only the first 3 color legends. The SSB annotation is framed in red.

infecting *Escherichia coli* was taken as a query, with cut-off parameters being left at their default value (confidence probability cut-off=99% for central gene, and 95% for neighbors, see Methods). A snapshot of the 12 first genomes and the first four legend lanes of the web page is reported in Fig. 3, and reveals that most of the neighboring genes around *bet* have homologs in the various phages. This conservation was not visible using IMG (see Fig. 1).

3.2. Co-occurrence as a path towards functional annotation

Localized co-occurrences can lead to the detection of "missing partners" for a biological function, for instance when a gene with known function is systematically placed next to a gene of unknown function. In a study focusing on the Sak4 recombinase of phage HK620 of *Escherichia coli* (*hkaL*, to be published elsewhere), we noticed using Phagonaute a frequent co-occurrence of a gene of unknown function (*hkaK*, in red, Fig. 4), placed one or two genes downstream of *sak4* (in green, Fig. 4). Among proteins related to this query gene of unknown function, the SSB annotation emerged (lower part of Fig. 4). SSB has important roles in protecting single strand DNA and facilitating the incoming of DNA repair proteins at replication forks (Shereda et al., 2008). We found that the HkaK protein encoded by this gene was indeed an SSB. Furthermore, its co-expression together with Sak4 increased recombineering efficiency.

While the Sak4/SSB example shows that co-occurrence can help detect accessory proteins, the reverse is also true. An accessory protein can lead to uncover a new protein family, or to extend it. The exonuclease with accessory function in DNA recombination is present in the vicinity of three different families of recombinases, Sak4, RecT and Erf (Lopes et al., 2010). Using Phagonaute and a query centered on the exonuclease (gp9) of *Lactococcus lactis* phage P335, these three types of neighbors could be observed (suppl. Fig. 1). Interestingly, a fourth "unknown" category of genes (with purple color) was visible on phages devoid of any of the three already characterized families, suggestive of functional replacement. A further analysis with Phagonaute of one of these proteins, the gp11 protein of phage Phi12, infecting *Staphylococcus aureus*, permitted to connect it to Gp2.5 (Suppl. Fig. 2), a protein cumulating SSB activity and homologous recombination activity in phage T7 of *E. coli* (Araki and Ogawa, 1981; Hollis et al., 2001; Kong and Richardson, 1996). One of these Gp2.5 distant homologs was tested in a recombineering assay, and found to be active as a recombinase (Lopes et al., 2010).

Sometimes accessory proteins fuse with their partner. Such a case of fusion between a phage recombinase of the RecT family and an exonuclease was observed in phage phiV10 of *E. coli* (probability cut-off=99%, suppl. Fig. 3). An exonuclease is often required to prepare the single strand DNA substrate onto which the RecT protein loads to perform homology search (Kuzminov, 1999), but proofs that the two proteins interact physically are scarce in the literature (Muyrers et al., 2000). Detection of such fusions is therefore useful for function prediction, as described some time ago under the "Rosetta stone" concept (Date, 2008).

3.3. Interest raised by abundance: new partners in capsid formation and putative homologs of an atypical nuclease

Some phage proteins have been neglected because they seemed to be orphans, based on BLAST searches. This is the case of Ne1, a protein likely involved in the connection between head and tail components of the capsid of many Caudoviridae (Lopes et al., 2014). In a previous work, we used distant homology and synteny to delineate membership in a classification based on these "neck" proteins. Within the neck module of phages belonging to Class I, a new Ne1 protein family was detected, encoded by a gene frequently positioned between genes coding for head closure (hc1) and tail completion (tc1) proteins (Fig. 5 illustrates this, starting from the ne1 (gp16.1) gene of SPP1). This important family of proteins might have escaped earlier scrutiny because of the large level of sequence divergence among its members. Its precise role in capsid morphogenesis remains to be determined.

Recently, the Ref (gp5) protein of phage P1 (infecting *E. coli*) was reported to behave as an endonuclease specific for DNA covered with RecA, whose role might be to boost the SOS response,

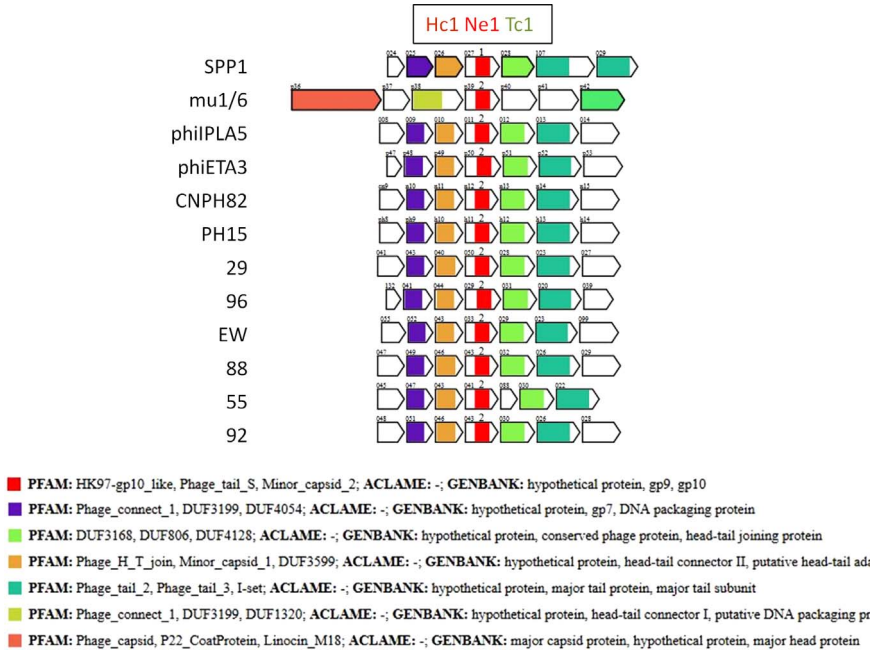

Fig. 5. The *ne1* gene (red) is frequently positioned between head-closure *hc1* gene (ochre) and tail completion *tc1* gene (green) in Caudoviridae. The Phagonaute query gene was *gp16.1* of phage SPP1 (corresponding to gene locus p027). The search was conducted on PGC2, using a cut-off of 95% for the central gene, minimum family size of 50 for coloring, and other parameters set by default. Only the top 13, among 415 genomes, are shown (phage name on the left side). The legend is also truncated.

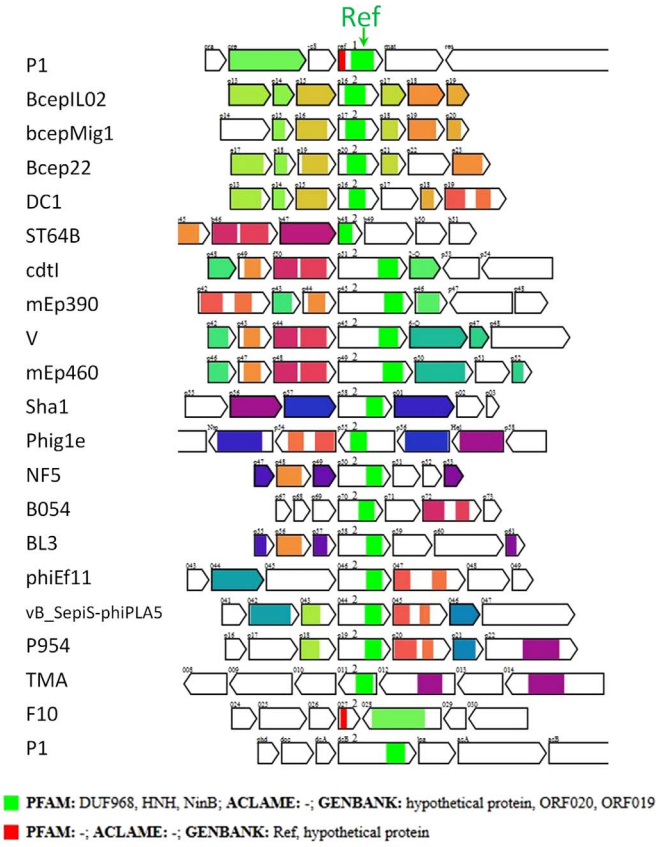

Fig. 6. The *ref* gene of P1 retrieves 20 distant homologs in Phagonaute. Query gene is *ref* (gp5) of P1 on PGC2; confidence cut-off on central gene is set to 90%, single HHsearch iteration on the central gene, other parameters to their default values. All genomes of the result page are shown, but the legend is restricted to the two central gene colors. In addition to the large green domain of Ref, a small red one is delineated, which matches a protein of F10 phage, but also several other small proteins retrieved at the next iteration (not on the figure). Of interest also, a second gene in P1 phage shares some homology with *ref* (last genome of the figure), namely *pdcB* (of unknown function).

and thereby increase the phage burst size (Gruber et al., 2015; Gruenig et al., 2011; Ronayne and Cox, 2014; Ronayne et al., 2016). This remarkable function is apparently limited in distribution to P1-like genomes. A Phagonaute query starting from this protein retrieves 20 more proteins at the first iteration (at 90% cut-off), none of which has a Genbank annotation so far (Fig. 6). However, the genetic context is not conserved on the various genomes. Moreover, the coverage of the match is less than 50% of the target protein, in most cases (starting from phage cdtI and below). Another protein of P1, PdcB of unknown function, also matches with Ref (Fig. 6, last lane). This case illustrates the possibility given by Phagonaute to enquire into domain functions (here possibly an endonuclease domain), rather than whole protein function. Experimental work is needed to examine to what extent these homologies relate to a broad, endonuclease-like activity, or narrow, Ref-like type of activity. Interestingly, the short N-terminal domain of Ref with a distinct, oligomerization and self-inhibition function (Gruber et al., 2015), matches to a distinct protein of phage F10 (in red, second-to- last genome of Fig. 6).

3.4. *Pushing the limits of Phagonaute: Herpes simplex virus 1 proteins have distant homologs among prokaryotic viruses*

Several proteins of Herpes simplex virus (HSV1) have been functionally related to distant homologs in bacteriophages, such as the UL6 portal protein (Newcomb et al., 2001; Trus et al., 2004), UL15 and UL28 sub-units of the terminase (Adelman et al., 2001; Davison, 1992), the floor domain of the major capsid protein VP5 (UL19) (Baker et al., 2005), the UL25 head-closure protein (Sae-Ueng et al., 2014), and the UL12 exonuclease (Reuven et al., 2003). This makes HSV1 (and its relatives) an interesting potential bridge between the bacterial and eukaryotic branches of the viral world. The Phagonaute tool being first dedicated to phages and archaeal viruses, eukaryotic viruses have not been included in the "all versus all" distant homology search depicted in Fig. 2. HSV1 was added *a posteriori,* and listed under the host name 'Eukaryots'. To do so, all steps presented in Fig. 2 were repeated, starting from the proteome of a single genome, except step 3: the profile library

Table 1.
HSV1 proteins with distant homologs in Phagonaute.

HSV1 gene	Annotation	Length (AA)	Match begin	Match end	Prob	% Coverage on target	Bacterial virus_gene or **Archaeal virus**_gene[a]	Annotation	Length (AA)	Match begin	Match end
DNA metabolism											
UL2	Uracil-DNA glycosylase	335	171	323	92.4	85	M6_gp048	Hypothetical protein	169	3	148
UL8	Helicase-primase subunit	751	454	726	91.15	35	**HVTV1**_91-91	DNA polymerase	914	547	867
UL9	DNA replication origin-binding helicase	852	83	402	100	93	CP21_194	Putative DNA replication origin-binding helicase	343	1	322
UL12	Deoxyribonuclease	627	218	370	99.37	54	ENT47670_gp26	Exonuclease	227	12	135
UL23	Involved in nucleotide metabolism	377	51	243	99.96	81	N4_35	Putative deoxyribonucleoside kinase	211	3	174
UL30	DNA polymerase catalytic subunit	1236	146	1080	100	89	**HVTV1**_91-91	DNA polymerase	914	5	822
UL39	Ribonucleotide reductase subunit 1	1138	470	1136	100	82	B4_0219	Ribonucleotide-diphosphate reductase su alpha	774	137	772
UL40	Ribonucleotide reductase subunit 2	341	28	336	100	95	pVp-1_gp036	Hypothetical protein	331	2	319
UL42	DNA polymerase processivity subunit	489	31	222	97.71	69	**PhiCh1**_p60	PCNA	248	2	175
UL50	Deoxyuridine triphosphatase	372	205	370	100	93	NCTC12673_gp006	Putative dUTP pyrophosphatase	149	8	147
Capsid											
UL15	DNA packaging terminase subunit 1	1931	147	731	99.97	83	KVP40.0355-17	Large terminase protein	601	48	552
UL26	Maturation protease	636	20	96	96.64	43	c5_gp07	Putative ClpP protease	202	9	97
Tegument											
UL13	Serine/threonine protein kinase	519	155	394	100	68	PaVLD_ORF037R	Protein kinase	273	13	199
UL24	Nuclear protein	270	63	107	93.43	26	RSL1_ORF212	Hypothetical protein	151	20	60
UL41	Host shutoff protein	490	173	276	98.94	40	LU11_gp153	DNA polymerase I	282	90	205
US3	Serine/threonine protein kinase	482	188	474	100	91	PaVLD_ORF037R	Protein kinase	273	6	255
US9	Membrane protein	91	28	84	90.36	74	PSS2_gp028	Hypothetical protein	63	7	54
Envelope											
UL27	Glycoprotein B	905	745	787	91.19	62	**AFV3**_gp44	Hypothetical protein	72	6	51
UL44	Glycoprotein C	512	227	446	98.95	49	WV8_gp041	Hypothetical protein	378	169	357
UL44	Glycoprotein C	512	275	353	96.71	15	RB49p168-hoc	Hoc head outer capsid protein	405	13	75

To get all results, see Phagonaute interface.

[a] Cut-off 90%, coverage of the target $>20\%$, first hit displayed unless a more "famous" hit is found lower in the list of all matches.

remained unchanged. This means that homology searches are unidirectional for this virus: starting from any HSV1 protein will lead to the display of potential distant homologs, but HSV1 proteins will not be present among the matches obtained with any other virus used as query, even if homology is present.

Table 1 lists the 19 HSV1 proteins (among 77 ORF, 25%) with matches to phage or archaeal virus proteins, grouped by function. All matches have a confidence probability above 90% and cover more than 20% of the length of the target protein. Among them, two of the six HSV1 proteins listed above were found, UL15 and UL12. The four others had a confidence probability below 90% (83.46% for UL6 portal, 69.58% for UL25 head closure, and 42% for UL28, and 38% for UL19).

Remarkably, most of the HSV1 proteins involved in DNA and nucleotide metabolism (10/13) have distant homologs among prokaryotic viruses, usually with similar annotated functions. Of the 3 that are missing, the UL5 helicase-primase has minor matches to phage helicases ($P > 90\%$ but coverage of the target protein below 20%, not shown in Table 1, see on Phagonaute), and UL29 and UL52 have no match at all. Interestingly, the UL29 (ICP8) protein, an SSB and recombination protein of 122 kDa (Darwish et al., 2015; Reuven and Weller, 2005; Reuven et al., 2004), has been reported to produce a 3D alignment with Gp2.5 of the *E. coli* phage T7, a 22.7 kDa protein with similar function (Kazlauskas and Venclovas, 2012). The Herpes UL29 protein being much larger in size (over 1000 residues) than Gp2.5, the homology stretches are probably too interspersed into the sequence to produce signal with HHsearch.

Among the 13 HSV1 virion proteins, 2 have a phage matching protein: the large subunit of terminase UL15 (as observed earlier, (Davison, 1992)), and the capsid maturation protease UL26. Unexpectedly, 5 of the 29 proteins of the tegument (a compartment of the virion outside the capsid) have some matches to prokaryotic viruses: two kinases UL13 and US3 have the best matches to several phage encoded kinases. One of these phage kinases (Stk of *E. coli* phage 933 W) functions as an exclusion factor, preventing super-infection by phage HK97 (Friedman et al., 2011). In addition, the UL41 host shut-off protein, which degrades host RNA, matches to several proteins sharing a $5'-3'$ exonuclease domain (also present in DNA polymerase I). The two remaining proteins have unknown functions on both sides of the matching pairs. Finally, even the "envelope" outer membrane of the virion (18 proteins) harbours two proteins with matches to archaeal or bacterial viruses: the first one is the glycoprotein B (gB) in its region 745–787, spanning the membrane proximal region and transmembrane domain (Cooper and Heldwein, 2015). It matches to a 72-AA long protein with unknown function of archaeal *Acidianus* filamentous viruses including AFV3 (Vestergaard et al., 2008). These viruses are not enveloped, therefore excluding a function similar to gB. This is a typical case where an HHsearch match, even with good confidence probability, should not be used to infer function. The second envelope protein matching to a phage protein is glycoprotein C, in a domain that corresponds to an immunoglobulin fold of RB49 Hoc protein (Fokine et al., 2011), and 31 other proteins. Here the HHsearch match corresponds to a broadly distributed protein fold, but cannot lead to function prediction.

Overall, using Phagonaute, some distant homologs between HSV1 proteins and prokaryotic viruses proteins can be tracked, especially for replication and DNA metabolism functions. However, synteny is usually absent in the comparisons displayed, except for the gene pair UL39-40 (see Table 1). This absence of substantial gene modules being shared between HSV1 and virus infecting other branches of the tree of life prevents making any guesses on ancestry among these.

4. Conclusion

To conclude, our goal with Phagonaute is to make accessible to a large community of virologists (with a special accent on bacteriophages) a new tool for homology detection among viral proteins. This should help get deeper into protein function identification, a real challenge of present-time biology. It may also open new avenues of reflection on the theme of virus origins and evolution.

Acknowledgments

We are thankful to Stéphane Roche and Paulo Tavares for their help in interpreting HSV1 results, and to Jessica Andreani and François Lecointe for their thorough reading of the manuscript. We also acknowledge the Migale Platform (INRA, Jouy) for the maintenance of the web site and databases associated with Phagonaute. This work was supported in part by the French ANR Grant Dynamophage ANR-10-BLAN-1328 to MAP, RG, OS and HD, and by the French ANR Grant Resisphage ANR-13-ASTR-011 to MAP and HD.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.virol.2016.05.007.

References

Adelman, K., Salmon, B., Baines, J.D., 2001. Herpes simplex virus DNA packaging sequences adopt novel structures that are specifically recognized by a component of the cleavage and packaging machinery. Proc. Natl. Acad. Sci. USA 98, 3086–3091.
Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.
Araki, H., Ogawa, H., 1981. The participation of T7 DNA-binding protein in T7 genetic recombination. Virology 111, 509–515.
Baker, M.L., Jiang, W., Rixon, F.J., Chiu, W., 2005. Common ancestry of herpesviruses and tailed DNA bacteriophages. J. Virol. 79, 14967–14970.
Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., Sonnhammer, E.L., 2000. The Pfam protein families database. Nucleic Acids Res. 28, 263–266.
Cooper, R.S., Heldwein, E.E., 2015. Herpesvirus gB: a finely tuned fusion machine. Viruses 7, 6552–6569.
Darwish, A.S., Grady, L.M., Bai, P., Weller, S.K., 2015. ICP8 filament formation is essential for replication compartment formation during herpes simplex virus infection. J. Virol. 90, 2561–2570.
Date, S.V., 2008. The Rosetta stone method. Methods Mol. Biol. 453, 169–180.
Davison, A.J., 1992. Channel catfish virus: a new type of herpesvirus. Virology 186, 9–14.
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., Felts, B., Dinsdale, E.A., Mokili, J.L., Edwards, R.A., 2014. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat. Commun. 5, 4498.
Fokine, A., Islam, M.Z., Zhang, Z., Bowman, V.D., Rao, V.B., Rossmann, M.G., 2011. Structure of the three N-terminal immunoglobulin domains of the highly immunogenic outer capsid protein from a T4-like bacteriophage. J. Virol. 85, 8141–8148.
Friedman, D.I., Mozola, C.C., Beeri, K., Ko, C.C., Reynolds, J.L., 2011. Activation of a prophage-encoded tyrosine kinase by a heterologous infecting phage results in a self-inflicted abortive infection. Mol. Microbiol. 82, 567–577.
Gruber, A.J., Olsen, T.M., Dvorak, R.H., Cox, M.M., 2015. Function of the N-terminal segment of the RecA-dependent nuclease Ref. Nucleic Acids Res. 43, 1795–1803.
Gruenig, M.C., Lu, D., Won, S.J., Dulberger, C.L., Manlick, A.J., Keck, J.L., Cox, M.M., 2011. Creating directed double-strand breaks with the Ref protein: a novel RecA-dependent nuclease from bacteriophage P1. J. Biol. Chem. 286, 8240–8251.
Hardies, S.C., Thomas, J.A., Black, L., Weintraub, S.T., Hwang, C.Y., Cho, B.C., 2015. Identification of structural and morphogenesis genes of Pseudoalteromonas phage phiRIO-1 and placement within the evolutionary history of Podoviridae. Virology 489, 116–127.
Hollis, T., Stattel, J.M., Walther, D.S., Richardson, C.C., Ellenberger, T., 2001. Structure of the gene 2.5 protein, a single-stranded DNA binding protein encoded by bacteriophage T7. Proc. Natl. Acad. Sci. USA 98, 9557–9562.

Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195–202.
Kazlauskas, D., Venclovas, C., 2012. Two distinct SSB protein families in nucleocytoplasmic large DNA viruses. Bioinformatics 28, 3186–3190.
Kong, D., Richardson, C.C., 1996. Single-stranded DNA binding protein and DNA helicase of bacteriophage T7 mediate homologous DNA strand exchange. EMBO J. 15, 2010–2019.
Kuzminov, A., 1999. Recombinational repair of DNA damage in Escherichia coli and bacteriophage lambda. Microbiol. Mol. Biol. Rev. 63, 751–813 (table of contents).
Leplae, R., Lima-Mendez, G., Toussaint, A., 2010. ACLAME: a classification of mobile genetic elements, update 2010. Nucleic Acids Res. 38, D57–D61.
Lopes, A., Amarir-Bouhram, J., Faure, G., Petit, M.A., Guerois, R., 2010. Detection of novel recombinases in bacteriophage genomes unveils Rad52, Rad51 and Gp2.5 remote homologs. Nucleic Acids Res. 38, 3952–3962.
Lopes, A., Tavares, P., Petit, M.A., Guerois, R., Zinn-Justin, S., 2014. Automated classification of tailed bacteriophages according to their neck organization. BMC Genom. 15, 1027.
Louis, A., Muffato, M., Roest Crollius, H., 2013. Genomicus: five genome browsers for comparative genomics in eukaryota. Nucleic Acids Res. 41, D700–D705.
Markowitz, V.M., Chen, I.M., Chu, K., Szeto, E., Palaniappan, K., Grechkin, Y., Ratner, A., Jacob, B., Pati, A., Huntemann, M., Liolios, K., Pagani, I., Anderson, I., Mavromatis, K., Ivanova, N.N., Kyrpides, N.C., 2012. IMG/M: the integrated metagenome data management and comparative analysis system. Nucleic Acids Res. 40, D123–D129.
Muyrers, J.P., Zhang, Y., Buchholz, F., Stewart, A.F., 2000. RecE/RecT and Redalpha/Redbeta initiate double-stranded break repair by specifically interacting with their respective partners. Genes Dev. 14, 1971–1982.
Newcomb, W.W., Juhas, R.M., Thomsen, D.R., Homa, F.L., Burch, A.D., Weller, S.K., Brown, J.C., 2001. The UL6 gene product forms the portal for entry of DNA into the herpes simplex virus capsid. J. Virol. 75, 10923–10932.
Reuven, N.B., Staire, A.E., Myers, R.S., Weller, S.K., 2003. The herpes simplex virus type 1 alkaline nuclease and single-stranded DNA binding protein mediate strand exchange in vitro. J. Virol. 77, 7425–7433.
Reuven, N.B., Weller, S.K., 2005. Herpes simplex virus type 1 single-strand DNA binding protein ICP8 enhances the nuclease activity of the UL12 alkaline nuclease by increasing its processivity. J. Virol. 79, 9356–9358.
Reuven, N.B., Willcox, S., Griffith, J.D., Weller, S.K., 2004. Catalysis of strand exchange by the HSV-1 μL12 and ICP8 proteins: potent ICP8 recombinase activity is revealed upon resection of dsDNA substrate by nuclease. J. Mol. Biol. 342, 57–71.
Ronayne, E.A., Cox, M.M., 2014. RecA-dependent programmable endonuclease Ref cleaves DNA in two distinct steps. Nucleic Acids Res. 42, 3871–3883.
Ronayne, E.A., Wan, Y.C., Boudreau, B.A., Landick, R., Cox, M.M., 2016. P1 Ref endonuclease: a molecular mechanism for phage-enhanced antibiotic lethality. PLoS. Genet. 12, e1005797.
Rotman, E., Kouzminova, E., Plunkett 3rd, G., Kuzminov, A., 2012. Genome of Enterobacteriophage Lula/phi80 and insights into its ability to spread in the laboratory environment. J. Bacteriol. 194, 6802–6817.
Sabri, M., Hauser, R., Ouellette, M., Liu, J., Dehbi, M., Moeck, G., Garcia, E., Titz, B., Uetz, P., Moineau, S., 2011. Genome annotation and intraviral interactome for the Streptococcus pneumoniae virulent phage Dp-1. J. Bacteriol. 193, 551–562.
Sae-Ueng, U., Liu, T., Catalano, C.E., Huffman, J.B., Homa, F.L., Evilevitch, A., 2014. Major capsid reinforcement by a minor protein in herpesviruses and phage. Nucleic Acids Res. 42, 9096–9107.
Shereda, R.D., Kozlov, A.G., Lohman, T.M., Cox, M.M., Keck, J.L., 2008. SSB as an organizer/mobilizer of genome maintenance complexes. Crit. Rev. Biochem. Mol. Biol. 43, 289–318.
Soding, J., 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics 21, 951–960.
Trus, B.L., Cheng, N., Newcomb, W.W., Homa, F.L., Brown, J.C., Steven, A.C., 2004. Structure and polymorphism of the UL6 portal protein of herpes simplex virus type 1. J. Virol. 78, 12668–12671.
Vestergaard, G., Aramayo, R., Basta, T., Haring, M., Peng, X., Brugger, K., Chen, L., Rachel, R., Boisset, N., Garrett, R.A., Prangishvili, D., 2008. Structure of the acidianus filamentous virus 3 and comparative genomics of related archaeal lipothrixviruses. J. Virol. 82, 371–381.