Journal of Basic Microbiology
Environment · Health · Techniques

Method Paper

A web-based restriction endonuclease tool for mycobacteriophage cluster prediction

Chris R. Gissendanner[1,†], Allison M. D. Wiedemeier[2,†], Paul D. Wiedemeier[3,†], Russell L. Minton[2], Swapan Bhuiyan[2,*], Jeremy S. Harmson[2] and Ann M. Findley[2]

[1] Department of Basic Pharmaceutical Sciences, College of Pharmacy, University of Louisiana at Monroe, Monroe, LA, USA
[2] Department of Biology, University of Louisiana at Monroe, Monroe, LA, USA
[3] Department of Computer Science and Computer Information Sciences, University of Louisiana at Monroe, Monroe, LA, USA

A recent explosion in the amount of genomic data has revealed a large genetic diversity in the bacteriophages that infect *Mycobacterium smegmatis*. In an effort to assess the novelty of newly described mycobacteriophage isolates and provide a preliminary determination of their probable cluster assignment prior to full genome sequencing, we have developed a systematic approach that relies on restriction endonuclease analysis. We demonstrate that a web-based tool, the Phage Enzyme Tool (or PET), is capable of rapidly facilitating this analysis and exhibits reliability in the putative placement of mycobacteriophages into specific clusters of previously sequenced phages. We propose that this tool represents a useful analytical step in the initial study of phage genomes and that this tool will increase the efficiency of phage genome characterization and enhance the educational activities involving mycobacteriophage discovery.

Keywords: Bioinformatics / Restriction endonuclease analysis / Phage enzyme tool (PET) / Mycobacteriophage / *Mycobacterium smegmatis*

Received: October 28, 2013; accepted: December 13, 2013

DOI 10.1002/jobm.201300860

Bacteriophages are biological entities that exhibit a high degree of genetic diversity [1]. Recently, insights into the evolutionary origins of bacteriophages have been greatly enhanced by the expansive sequencing of bacteriophage genomes. In particular, over 500 genomes of bacteriophage that infect *Mycobacterium smegmatis* mc^2 155 have been sequenced to date (data found at phagesdb.org), revealing tremendous diversity in this bacteriophage group despite their infecting a common host [2]. This explosion in the amount of genomic data for the mycobacteriophages has been facilitated by rapid next-generation sequencing technologies as well as the coordinated efforts of the SEA-PHAGES (Science Education Alliance Phage Hunters Advancing Genomics and Evolutionary Science) program. This program, which began in 2008, has involved over 70 universities and colleges and over 1400 students in the isolation of over 3000 mycobacteriophages as a component of inquiry-based laboratory courses [3]. The data from these projects have dramatically increased our knowledge of the structure and origins of mycobacteriophage genomes and have also yielded genetic tools that can be applied to *Mycobacterium* species that cause human disease [2].

Mycobacteriophages are clustered into distinct groups based on DNA sequence similarity [3, 4]. To date, there are 20 such clusters designated by letters (clusters "A" through "T"). In addition, there are "singleton" phages that do not fall into a particular cluster. Nine of the clusters (A, B, C, D, F, H, I, K, L) also contain multiple subclusters. Cluster designation can only be determined after the sequencing of a phage genome is complete. This dependence on DNA sequencing to establish the relatedness of different mycobacteriophages creates issues with respect to future research on mycobacteriophage genomes and their utility as an educational tool. For

[†]These authors contributed equally to this work.
[*]Current address: Department of Biological Sciences, University of North Texas, Denton, TX
Correspondence: Chris R. Gissendanner, University of Louisiana at Monroe, 700 University Avenue, Monroe, LA 71209, USA
E-mail: gissendanner@ulm.edu
Phone: 318 342 3314
Fax: 318 342 1790

the former, it may be desirable to expend resources to sequence phage genomes that will provide novel data and expand the diversity of the genome repertoire or that will increase the number of phages that belong to a specific cluster. Therefore, an accurate method that could predict phage clusters prior to whole genome sequencing would be highly desirable. For the latter, most laboratory courses that participate in the SEA-PHAGES program are only able to sequence one or two phage genomes even though many more phages are isolated by the students in the course. This hinders the depth of analysis that a student can perform in characterizing their phage.

We describe here a freely available online tool, the Phage Enzyme Tool, or PET (http://ec2-54-245-31-145.us-west-2.compute.amazonaws.com/) that can be effectively utilized to predict cluster designations using restriction endonuclease (REase) data. Since REase assays are relatively inexpensive and easy to perform, they represent an essential tool for the teaching and research laboratory engaged in mycobacteriophage discovery. Utilizing New England Biolabs NEBcutter® program [5], the PET provides a library of phage restriction data that is both searchable and diagnostic. PET has two broad functionalities (termed "Action 1" and "Action 2" in the program). The program allows the user to choose individual phages or groups of phages, clusters, subclusters, enzymes, and enzyme cut ranges for analysis ("Action 1"). A second function allows a user to input REase digestion data associated with an unknown, unsequenced, and unclustered phage ("Action 2"). The PET displays the number of cut sites of the unknown phage compared with the known phages within the database, facilitating comparisons and cluster/subcluster predictions.

For visualizations of known restriction sites (Action 1), the user inputs all necessary data through five drop down menus ("Phages", "Clusters", "Subclusters", "NEB Enzymes", and "Cut Ranges"). Known, or finished, phage genomes are those that have been sequenced and are documented by the PhagesDB website. The output, representing the number of restriction sites, is a 2D table displayed either horizontally or vertically. Each cell in the 2D table represents the number of times a specific restriction site is found in a specific phage genome (Fig. 1A). The background of each cell is shaded based on predetermined bin values representing a range of restriction sites or "cuts" (the selection of these bin values is described below). An empty white cell means that the enzyme site is not found in the phage genome. Other bin values are shaded from light gray to black depending on the number of times a restriction site is found in the phage genome, with black representing the bin value with the largest number of sites.

The user can also produce a PDF of a rooted or unrooted phylogeny tree with cut data from selected phages (Fig. 1B). The Phylip trees can be generated using either known phage cut data or trees can be generated that allow an unknown phage to be placed among a tree of known phages. This function uses the Phylip pars, consense, and drawgram programs [6] (Felsenstein, J., *PHYLIP (Phylogeny Inference Package) version 3.6*. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle, 2005; http://evolution.genetics.washington.edu/phylip.html) that are locally installed on the server that hosts the PET.

The second use of the tool allows the user to identify a possible cluster assignment of an isolated phage that is not yet sequenced. In order to use this portion of PET, the user must input data associated with the number of restriction fragments identified though REase digestion and electrophoresis of the unknown phage DNA. Upon selection of Action 2, the user may select any number of phages, clusters, and/or subclusters and choose the NEB enzymes for which the user has data. The user is then prompted to input the digestion data for the unknown phage cut with each chosen enzyme. The input data for the selected enzymes used to digest the phage DNA are specific bin designations-"None", "Few", "Some", "Many" and "Alot"-that represent a set of cut ranges based on the number restriction fragments generated by the digestion. The unknown phage's pattern of bin designations for the set of selected enzymes is then compared to the bin designations of phages in the database for the same enzymes. The data displayed show each phage and its cluster and subcluster (if applicable) designation. A similarity (SIM) score is generated for each known phage in relation to the unknown phage (Fig. 2). The SIM score corresponds to the percentage of enzymes that exactly match the specified bin ranges for the user-defined unknown phage. The numbers of cuts for each enzyme/phage are also provided, similar to the Action 1 operation, allowing the user to determine differences between clusters for the enzyme cut data.

The enzymes *Bam*HI, *Cla*I, *Eco*RI, and *Hin*dIII function as a useful first pass panel of enzymes for initial comparisons to known phage DNA sequences (Graham F. Hatfull, University of Pittsburgh, personal communication). However, since mycobacteriophage genome sizes range from ~40 to ~150 kb in size, an enzyme that cuts phage DNA multiple times may be difficult to accurately interpret when attempting to determine the number of restriction sites based on the number of restriction fragments separated by gel electrophoresis. To address

A

Phage	Cluster	Subcluster	BamHI	ClaI	EcoRI	HaeIII	HindIII
Airmid	A	A5	3		5	252	1
Benedict	A	A5	4		6	259	1
Chadwick	A	A5	1		5	247	1
Conspiracy	A	A5	1	1		309	2
Cuco	A	A5	5			301	2
ElTiger69	A	A5	5		5	264	2
George	A	A5	105			630	
Jovo	A	A5	1	2		305	2
LittleCherry	A	A5	2			288	3
Swirley	A	A5	1			296	3
Theia	A	A5	1			310	2
Tiger	A	A5	2			309	2
UnionJack	A	A5	1		4	260	4

B

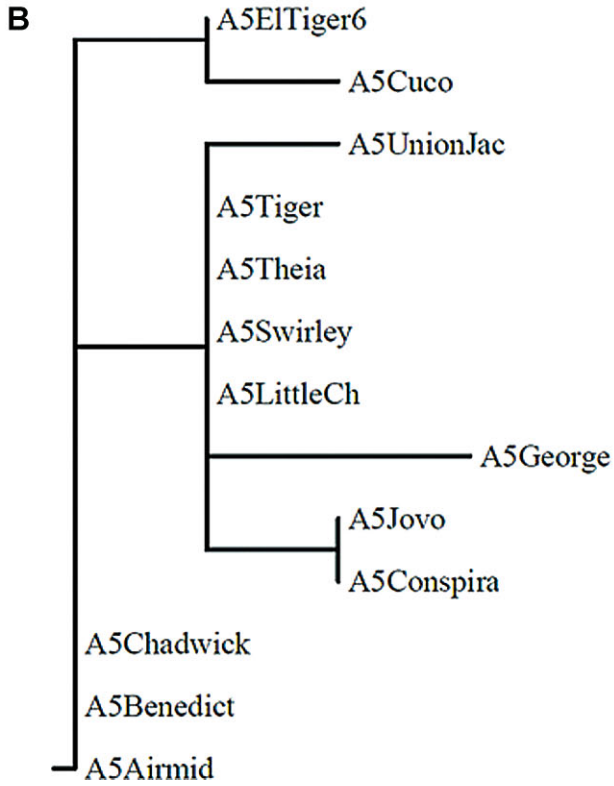


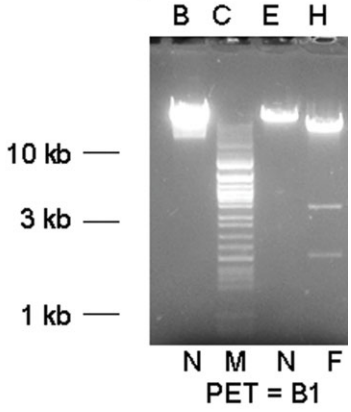
Figure 1. Tabular output for Action 1 by the PET (A) and corresponding PHYLIP tree (B).

this problem we based the bin ranges-"Few", "Some", "Many", "Alot"-on likely correlations between experimental restriction data and the number of restriction sites (Fig. 2A). These bin ranges were based on the restriction analysis of sequenced phages and serves as an entry point (default state) for the analysis. These range designations are intended to be fluid and may change as more genomes are analyzed by restriction analysis and subsequently sequenced. In an actual experiment, the user would count the number bands in a gel and then select the appropriate bin. For example, using these default bins, if a user counts between 6 and 10 bands in a gel for a given enzyme, the user would select the "5–15 cut sites" bin since the experimental data

A

# Fragments	# Restriction Sites	Bin
1	0	None
2-5	1-4	Few
6-10	5-15	Some
11-20	16-40	Many
>20	>40	A lot

B Sample #6



C Sample #6

Phage	Cluster	Subcluster	Sim Score
**	*Unknown*	*Unknown*	*None*
Trypo	B	B1	100.00%
Manad	B	B1	100.00%
Suffolk	B	B1	100.00%
UncleHowie	B	B1	100.00%
Oosterbaan	B	B1	100.00%
OliverWalter	B	B1	100.00%
Numberten	B	B1	100.00%
Murdoc	B	B1	100.00%
Crownjwl	B	B1	100.00%
Spartan300	B	B1	100.00%
DonSanchon	B	B1	100.00%
ShiVal	B	B1	100.00%
Hertubise	B	B1	100.00%
SDcharge11	B	B1	100.00%
Gyarad05	B	B1	100.00%
Pipsqueak	B	B1	100.00%
KingVeveve	B	B1	100.00%
KLucky39	B	B1	100.00%
OSmaximus	B	B1	75.00%
Cheetobro	K	K4	75.00%

Phage	BamHI	ClaI	EcoRI	HindIII
**	0 Cuts	[16-41) Cuts	0 Cuts	[1-5) Cuts
Trypo		28		3
Manad		29		3
Suffolk		32		3
UncleHowie		30		4
Oosterbaan		34		4
OliverWalter		34		3
Numberten		35		4
Murdoc		32		4
Crownjwl		30		3
Spartan300		32		4
DonSanchon		31		4
ShiVal		34		4
Hertubise		34		4
SDcharge11		29		4
Gyarad05		32		3
Pipsqueak		33		4
KingVeveve		29		3
KLucky39		34		2
OSmaximus	2	36		2

Figure 2. REase digestion and PET analysis of known phage DNA. (A) Selection of bins for experimental restriction data. (B) Ethidium bromide stained 1% agarose gel for sample #6. REases B, C, E, and H correspond to *Bam*HI, *Cla*I, *Eco*RI, and *Hin*dIII, respectively. "N", "M", and "F" correspond to "None", "Many", and "Few", respectively. (C) Action 2 PET tabular output for sample #6.

would likely correlate to an actual number of 5–15 restriction sites.

We assessed the accuracy of the PET with blind testing of known phage DNA by restriction digestion and then using the PET to predict the clusters/subclusters (Table 1 and Fig. 2B and C). This blind test was administered by the SEA-PHAGES program as part of a "Phage Grand Challenge" for institutions with a mycobacteriophage genomics course (see Table 1 legend for more information). We utilized the first pass panel of REases-*Bam*HI, *Cla*I, *Eco*RI, and *Hin*dIII. Of the 11 known phage we tested, PET correctly predicted a single cluster or subcluster at 100% SIM value for five phage. For five other samples, the PET generated information that allowed a correct deduction of the cluster/subcluster or produced a narrow list of clusters/subclusters including the correct cluster/subcluster assignment. Only one known phage was predicted incorrectly, but the experimental digestion produced anomalous data, possibly due to contamination or experimental error.

An important aspect of using the PET is the need to produce high quality restriction digests of unknown phage DNA. A poorly executed restriction analysis will hinder the ability of the PET to provide accurate predictions. Since undergraduate students will be using the PET in their phage hunting experiments, it is essential to instruct students on best practices for producing high quality analyses. Proper interpretation of restriction fragments is also essential. For linear DNA, the number of restriction fragments n corresponds to n-1

Table 1. REase and PET analysis of known phages.

Sample	B	C	E	H	PET	SIM	Prediction	Known	Notes
1	S	M	N	F	C1	100	C1	C1	
2	Not determined								
3	N	N	N	F	C2,K3,K4,A4	100	C2,K3,K4, or A4	K3	
4	S	S	S	F	A1,F1	100	A1 or F1	F1	F1: 3/5 SIM 100
5	Not determined								
6	N	M	N	F	B1	100	B1	B1	
	N	A	N	F	B1, multiple	75			
7	M	F	S	N	I1	100	Q or I1	I1	
	A	F	S	N	Q	100			
8	A	N	N	N	B3, A5	100	B3	B3	B3: 8/9 SIM 100
	M	N	N	N	Multiple	75			
9	M	S	N	N	A2	100	A2	A2	
10	A	N	F	F	G	100	G	Q	Anomalous digestion
11	Not determined								
12	N	F	M	F	E	100	E	E	
	N	S	M	F	E	100			
13	N	M	M	M	J	100	J	J	
14	Not determined								
15	N	N	N	N	Multiple	100	A4	A4	
	P	**SI**	**SII**	**St**					Additional digests: *Pfl*FI, *Sac*I, *Sac*II, *Stu*I
	M	M	N	N	Multiple A subcluster	87.5			
	A	A	N	N	A4, B1	100			B1 is not represented by the N,N,N,N pattern for B, C, E, H

B, C, E, H represent the REases *Bam*HI, *Cla*I, *Eco*RI, *Hin*dIII, respectively. Bins are represented by A ("Alot"), M ("Many"), S ("Some"), F ("Few"), N ("None"). Samples were provided by the Graham Hatfull lab at the University of Pittsburgh as part of the "Phage Grand Challenge". Samples 2, 5, 11, and 14 could not be analyzed by REase analysis due to questions regarding known phage identity, mixed phage sample, low DNA concentration, and DNA degradation, respectively. See http://phagesdb.org/blog/posts/10/ for more information. Alternate PET inputs were utilized for samples 6, 7, 8, 12, and15 which provided additional information for cluster prediction. Sample 15 was subjected to additional digests by *Pfl*FI (P), *Sac*I (SI), *Sac*II (SII), and *Stu*I (St).

restriction sites. Given the size of phage DNAs, it is likely that interpretations will be affected by the inability to identify small fragments, distinguish very large fragments, or distinguish fragments of similar size. Bands of similar size that run together on a gel can easily be identified as bands that are brighter than larger bands in the gel. Such bands should be counted as at least two fragments. Problems with interpretation escalate with increasing numbers of cuts. Running multiple different percentage gels may increase the accuracy of cut frequency estimates. Additionally, for enzyme digests that produce enough bands to be placed into the "Many" bin, the PET analysis should be performed with two inputs: one with the "Many" bin for that enzyme and another using "Alot". The same approach could also be used for the "Few" and "Some" bins. It is possible that one interpretation gives a clearer result, as was demonstrated for samples 6, 7, 8, and 15 in the PET test of known phages. Analysis of known restriction patterns of other phage in a potential cluster may also clarify gel interpretation. However, it should be noted that the PET can utilize far more enzymes than just these four and the more enzymes that are utilized, the higher the predictive power of the tool (see sample 15, Table 1).

The PET also has outstanding utility in the undergraduate teaching lab. First, it involves an experience of performing a bench experiment, assessing the data, and using quantitative and computational approaches to generate the best interpretation. It enhances the phage hunting experience of undergraduates by providing a genetic identity, although putative, to their phage. This provides greater satisfaction for the student as they get to draw a stronger conclusion from their efforts. Recently, it was shown that the sequence of the gene encoding the mycobacteriophage tape measure protein was an effective approach in placing mycobacteriophages into specific clusters in lieu of whole-genome sequencing [7]. While this approach is more cost- and time-effective compared to generating a complete genome sequence, the technique requires both PCR amplification and subsequent Sanger sequencing steps. Our approach is a simpler alternative since it only requires DNA isolation and single molecular biology technique that typically does not require optimization. An initial REase analysis

could also aid in the primer design for the tape measure gene experiments as it could identify a refined list of potential clusters/subclusters. The utilization of both approaches is capable of being incorporated into an undergraduate teaching lab and could represent a very robust cluster prediction protocol. We should also note that while we describe the PET in the context of mycobacteriophages, a similar database can be constructed for any type of phage once genomic sequences have been generated.

Acknowledgments

We thank Graham F. Hatfull, Deborah Jacobs-Sera, and Daniel A. Russell of the Department of Biological Sciences at the University of Pittsburgh for their scientific support of the University of Louisiana at Monroe mycobacteriophage genomics course. We also thank Lucia P. Barker, Kevin Bradley, Tuajuanda Jordan and the Howard Hughes Medical Institute SEA-PHAGES program for technical and administrative support. Daniel A. Russell and Michelle M. Boyle of the Department of Biological Sciences at the University of Pittsburgh prepared the DNA samples for the Phage Grand Challenge. This work was initiated through a HHMI-supported summer sabbatical program. Further support has been provided by the HHMI SEA-PHAGES program, an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103424, and an Amazon Web Services in Education Grant award.

Conflict of interest statement

All authors declare that there are no financial/commercial conflicts of interest.

References

[1] Hatfull, G.F., Hendrix, R.W., 2011. Bacteriophages and their genomes. Curr. Opin. Virol., **1**, 298–303.

[2] Hatfull, G.F., 2012. The secret lives of mycobacteriophages. Adv. Virus Res., **82**, 179–288.

[3] Pope, W.H., Jacobs-Sera, D., Russell, D.A., Peebles, C.L. et al., 2011. Expanding the diversity of mycobacteriophages: insights into genome architecture and evolution. PLoS One, **6**, e16329.

[4] Cresawn, S.G., Bogel, M., Day, N., Jacobs-Sera, D. et al., 2011. Phamerator: a bioinformatic tool for comparative bacteriophage genomics. BMC Bioinform., **12**, 395.

[5] Vincze, T., Posfai, J., Roberts, R.J., 2003. NEBcutter: a program to cleave DNA with restriction enzymes. Nucleic Acids Res., **31**, 3688–3691.

[6] Felsenstein, J., 1989. PHYLIP – Phylogeny Inference Package (Version 3.2). Cladistics, **5**, 164–166.

[7] Smith, K.C., Castro-Nallar, E., Fisher, J.N., Breakwell, D.P. et al., 2013. Phage cluster relationships identified through single gene analysis. BMC Genomics, **14**, 410.

