Skoltech

Identification of putative anti-defence genes in the genomes of *Autographiviridae* bacteriophages

Oksana Kotovskaya

Introduction

Bacteria and bacteriophages are in a constant arms race: a huge number of bacterial defence systems have been discovered in recent years, allowing them to avoid reproduction of phage progeny in many different ways (Hampton *et al.* 2020). The evolution of defence systems, in turn, leads to the evolution of anti-defence systems of bacteriophages. The first discovered bacterial defence system was a restriction-modification system (Luria and Human 1952), the effect of which is to modify the genome of a bacterium by methylation and recognition of an unmethylated, foreign DNA molecule, followed by restriction of this molecule. Accordingly, in order to avoid the restriction-modification system, at least two mechanisms (in fact, there are more) can be proposed by which the bacteriophage can avoid the defence system: an inhibition of modification protein and an inhibition of restriction protein (effector). Nevertheless, there are more types of defence systems and anti-defence proteins which are known and which are to be discovered.

Teseptimavirus T7 and Teetrevirus T3 are lytic bacteriophages (Demerec and Fano 1945) from the Autographiviridae family, which are known to have special anti-defence proteins. These two proteins have different mechanisms of action. Teseptimavirus T7 uses Overcome classical restriction (Ocr) as a DNA-mimic, which binds to defence system effectors, competitively inhibits the cleavage of phage DNA (Studier 1975; Walkinshaw *et al.* 2002). Teetrevirus T3 uses S-adenosyl-L-methionine lyase (SAMase) (Studier and Movva 1976), which hydrolyse S-adenosyl-L-methionine (Guo *et al.* 2021), substrate of the defence systems modification proteins, methyltransferases. Despite the differences in mechanisms of action, anti-defence proteins of T3 and T7 phages are coded by genes located similarly in genome (Fig. 1): both genes encode the first ORF of the viral genome, which allows them to be expressed in the first place (Dunn *et al.* 1983; Pajunen *et al.* 2002). Both ORFs are part of polycistronic mRNA, which also encode RNA polymerase, serine-threonine kinase and small gene products, playing a role in inhibition of cell growth (Molshanski-Mor *et al.* 2014; Kiro *et al.* 2013). Such a genome arrangement shows the presence of a synteny between T3 and T7 phages.

Since other Autographiviridae bacteriophages may also have shared synteny, we hypothesize that the ones with similar organization of RNA polymerase upstream region may have similar or different anti-defence proteins coded by early genes. Thus, our work is aimed at search and identification of putative anti-defence genes in the RNA polymerase upstreams of Autographiviridae bacteriophages.

Materials and methods

Data gathering

To obtain genomes with syntenic RNA polymerase upstream regions (hereinafter called “upstream regions” for simplicity) we searched for homologs of T7 RNA polymerases within Viruses superkingdom. To do so, we performed 3 rounds of online PSI-BLAST (Altschul 1997) against nr sequences among Viruses superfamily (taxid: 10239) with p-value threshold equal to 0.05, using T7 RNA polymerase sequence (NCBI Protein Identificator: NP_041960.1) as a query. We obtained 1626 hits with E-value < 10-4. Next we filtered hits based on an alignment length and removed hits with alignment length less than 750, corresponding to proteins with similar domains. As a result, 1051 proteins passed this filter.

We searched for bacteriophage genome assemblies, associated with found proteins with the help of efetch v. 16.2 from Entrez toolset (Maglott *et al.* 2010). 974 genomes were downloaded with NCBI Datasets v. 14.25.1 (Sayers *et al.* 2021). Next we removed short genomes (length is lower than 0.005 quantile). Thus, 968 genomes were used in further analysis.

Genomes annotation and manual curating

Since our goal was to analyze only upstream regions containing immediately early genes, one of the crucial steps was to define, does the particular genome have desired RNAP upstream organization. To check it, we split all genomes into 5 datasets, based on colocalization of RNAP, terminal direct repeats (TDRs) and intergenic regions.

To define RNAP position, we annotated genomes using prokka v. 1.13.4 (Seemann 2014). 23 genomes, in which

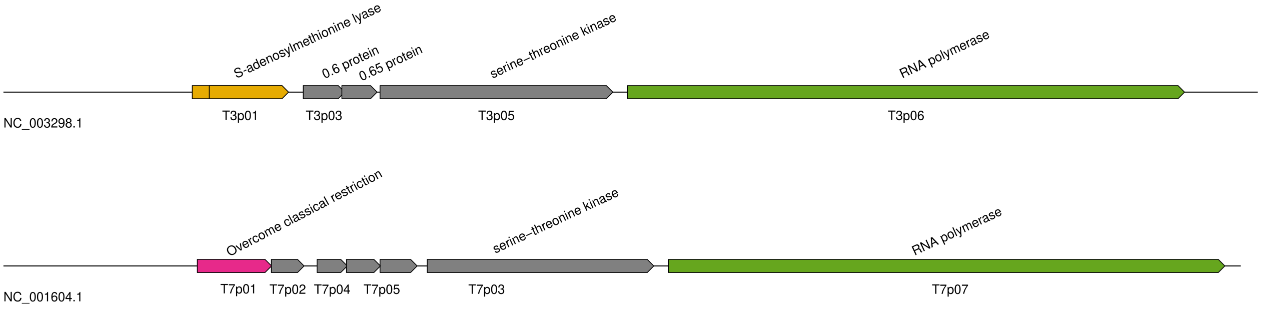

Figure 1 Schematic maps of T3 (top) and T7 (bottom) phages RNA polymerase upstream regions. The map was built using gggenomes v. 0.9.9.9000 package (R v.4.2.2 (R Core Team 2022)).

all proteins were annotated as hypothetical, were also excluded from analysis. Based on annotation, we also defined intergenic regions positions with the help of bedtools subtract v. 2.30.0 (Quinlan and Hall 2010). To find the TDRs, we performed alignment of genomes against themselves, using minimap2 v. 2.24 (Li 2021) with the following parameters: `-X -N 50 -p 0.1 -c`. Aligned sequences without mismatches and lengths more than 100 and less than 300 were considered as potential TDRs.

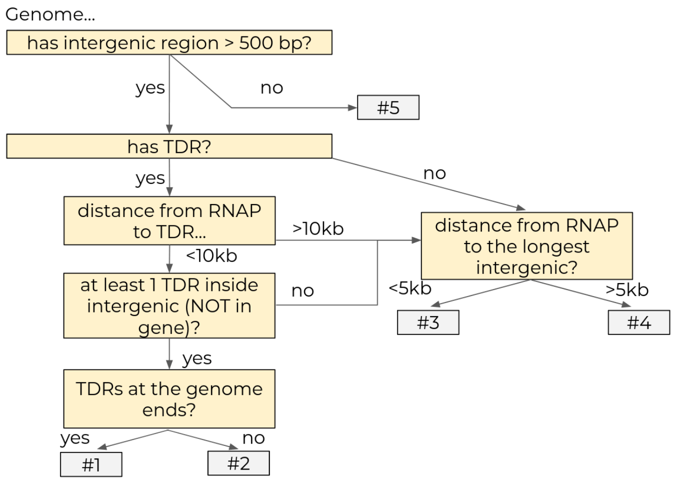
Genome...
has intergenic region > 500 bp?
yes
no
#5
has TDR?
yes
no
distance from RNAP to TDR...
>10kb
<10kb
distance from RNAP to the longest intergenic?
at least 1 TDR inside intergenic (NOT in gene)?
no
<5kb
>5kb
#3
#4
yes
TDRs at the genome ends?
yes
no
#1
#2

Figure 2 Decision rule to assign dataset. #1-5 – datasets.

Next we splitted genomes into datasets, based on following criteria: sequences without intergenic regions longer than 500 bp were assigned to a dataset with lower priority (Fig. 2). Other genomes were splitted into datasets, based on presence of TDRs in the sequence, its position with respect to ends of the genome and RNAP position. Accordingly, we considered the genome to be in the highest priority dataset, if its assembly contains TDRs at the ends of genome assembly and TDRs with upstream regions less than 10 kbp long (Table 1).

Then we build a phylogeny of sequences. To do so we aligned RNAP amino acids sequences within each dataset, using MAFFT v. 7.450 (global alignment mode, 1000 iterations) (Katoh 2002), then trimmed alignment with trimAL v. 1.4.1 (default parameters: `-automated1`) (Capella-Gutiérrez *et al.* 2009), and build trees with iqtree2 v. 2.2.2.3 (Nguyen *et al.* 2014), using model LG+I+R6, which was chosen by iqtree2 as the best during the launch of ModelFinder (Kalyaanamoorthy *et al.* 2017) with only dataset #1.

Table 1 Number of genomes in each dataset before and after curation.

dataset	Number of genomes	
	before curation	after curation
#1	160	144
#2	7	6
#3	311	289
#4	433	25
#5	34	1
total	945	465

Clusterization of genes in upstream regions

We selected genes in upstream regions, using RNAP gene end as a right border of an upstream region and either a TDR end, a closest intergenic region, or 10 kb distance from the start of RNA polymerase as a left border. Totally 3183 protein coding sequences were clustered with MMseqs2 v. 14.7 (Steinegger and Söding 2017) with connected component clustering mode (`-cluster-mode 1`), where sequences are connected, if the coverage of alignment was more that 80% alignment length (`-cov-mode 0 -c 0.7`) and minimal sequences identity was 40% (`-min-seq-id 0.4`). As a result, we obtained 685 protein clusters.

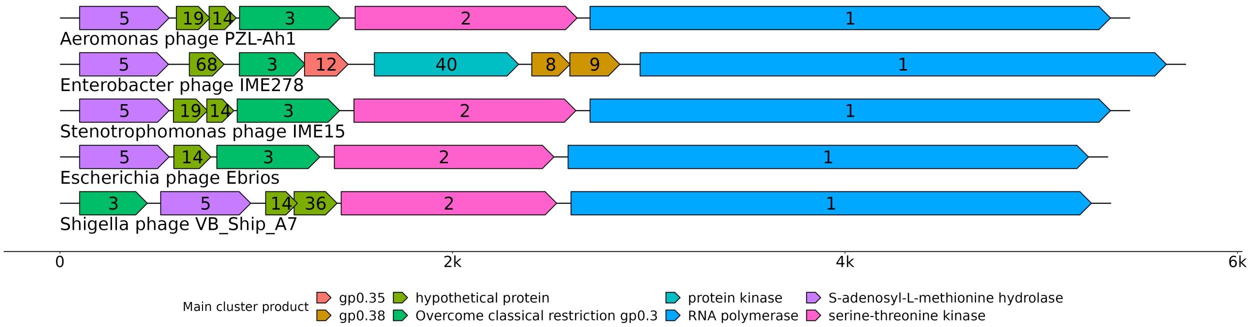

Figure 3 Genetic maps for upstream regions of bacteriophages, which have both Ocr and SAMase. Numbers show clusters. Visualized with R v. 4.2.3 R Core Team (2022), package gggenomes v. 0.9.9.9000

Table 2 Top-5 the most abundant clusters.

#	Cluster representative	Number of proteins in cluster
1	T7 RNA polymerase	354
2	Protein kinase 0.7	190
3	Ocr	150
4	gp0.6	149
5	SAMase	144

Results and discussions

We obtained 465 genomes, which RNA polymerase upstreams are similar to T7 and T3 bacteriophages. In these upstream regions we have identified 685 protein clusters, 159 of them consisting of more than 2 protein sequences. Examples of upstream regions shown on Fig. 3.

As we expected, the most abundant cluster is a cluster, containing RNA polymerases (354). Out of note, our approach resulted in 14 different clusters of RNA polymerases even with relaxed parameters of clusterization, which indicated the possibility of a more thorough restriction of genomes in the dataset. Most abundant clusters are given in Table 2. Out of 465 genomes in our dataset 150 genomes (32%) have Ocr protein homologs and 144 genomes (31%) have SAMase homologs, well-known anti-restriction proteins. Meanwhile, only 7 genomes have both Ocr and SAMase homologs in the upstream region. We also found a large number of clusters containing hypothetical proteins that have not been characterized before. Our future goal is to search for known domains in the proteins of these clusters and characterize them in general. It will allow us to find new anti-defence proteins.

Citations

Altschul S. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 25:3389–3402.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 25:1972–1973.

Demerec M, Fano U. 1945. BACTERIOPHAGE-RESISTANT MUTANTS IN ESCHERICHIA COLI. Genetics. 30:119–136.

Dunn JJ, Studier FW, Gottesman M. 1983. Complete nucleotide sequence of bacteriophage t7 DNA and the locations of t7 genetic elements. Journal of Molecular Biology. 166:477–535.

Guo X, Söderholm A, P SK, Isaksen GV, Warsi O, Eckhard U, Trigüis S, Gogoll A, Jerlström-Hultqvist J, Åqvist J *et al*. 2021. Structure and mechanism of a phage-encoded SAM lyase revises catalytic function of enzyme family. eLife. 10.

Hampton HG, Watson BNJ, Fineran PC. 2020. The arms race between bacteria and their phage foes. Nature. 577:327–336.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods. 14:587–589.

Katoh K. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Research. 30:3059–3066.

Kiro R, Molshanski-Mor S, Yosef I, Milam SL, Erickson HP, Qimron U. 2013. Gene product 0.4 increases bacteriophage t7 competitiveness by inhibiting host cell division. Proceedings of the National Academy of Sciences. 110:19549–19554.

Li H. 2021. New strategies to improve minimap2 alignment accuracy. Bioinformatics. 37:4572–4574.

Luria SE, Human ML. 1952. A NONHEREDITARY, HOST-INDUCED VARIATION OF BACTERIAL VIRUSES. Journal of Bacteriology. 64:557–569.

Maglott D, Ostell J, Pruitt KD, Tatusova T. 2010. Entrez gene: gene-centered information at NCBI. Nucleic Acids Research. 39:D52–D57.

Molshanski-Mor S, Yosef I, Kiro R, Edgar R, Manor M, Gershovits M, Laserson M, Pupko T, Qimron U. 2014. Revealing bacterial targets of growth inhibitors encoded by bacteriophage t7. Proceedings of the National Academy of Sciences. 111:18715–18720.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2014. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution. 32:268–274.

Pajunen MI, Elizondo MR, Skurnik M, Kieleczawa J, Molineux IJ. 2002. Complete nucleotide sequence and likely recombinatorial origin of bacteriophage t3. Journal of Molecular Biology. 319:1115–1132.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–842.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S *et al*. 2021. Database resources of the national center for biotechnology information. Nucleic Acids Research. 50:D20–D26.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 30:2068–2069.

Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology. 35:1026–1028.

Studier F. 1975. Gene 0.3 of bacteriophage t7 acts to overcome the DNA restriction system of the host. Journal of Molecular Biology. 94:283–295.

Studier FW, Movva NR. 1976. SAMase gene of bacteriophage t3 is responsible for overcoming host restriction. Journal of Virology. 19:136–145.

Walkinshaw M, Taylor P, Sturrock S, Atanasiu C, Berge T, Henderson R, Edwardson J, Dryden D. 2002. Structure of ocr from bacteriophage t7, a protein that mimics b-form DNA. Molecular Cell. 9:187–194.