Tentaclins – a Novel Family of Phage Receptor-Binding Proteins That Can Be Hypermutated by DGR Systems

Ivan K. Baykov *, Artem Y. Tikunov, Igor V. Babkin, Valeria A. Fedorets, Elena V. Zhirakovskaia, Nina V. Tikunova *

Posted Date: 15 November 2023

doi: 10.20944/preprints202311.0981.v1

Keywords: Bacteriophage; Genome sequence; Diversity-generating retroelement; C-type lectin; Tentaclin; Adhesin; Ig-like domain; Receptor-binding protein

Article

Tentaclins – A Novel Family of Phage Receptor-Binding Proteins That Can Be Hypermutated by DGR Systems

Ivan K. Baykov [1,2,*], Artem Y. Tikunov [1], Igor V. Babkin [1], Valeria A. Fedorets [1], Elena V. Zhirakovskaia [1] and Nina V. Tikunova [1,*]

[1] Federal State Public Scientific Institution «Institute of Chemical Biology and Fundamental Medicine», Siberian Branch of the Russian Academy of Sciences, 630090, Novosibirsk, Russia
[2] Shared Research Facility "Siberian Circular Photon Source" (SRF "SKIF") of Boreskov Institute of Catalysis SB RAS , Novosibirsk, Russia
* Correspondence: tikunova@niboch.nsc.ru, ivan_baykov@mail.ru

Abstract: Diversity-generating retroelements (DGRs) are prokaryotic systems providing rapid modification and adaptation of target proteins. In phages, the main targets of DGRs are receptor-binding proteins that are usually parts of tail structures and the variability of such host-recognizing structures enables phage adaptation to changes on the bacterial host surface. Sometimes, more than one target gene containing a hypermutated variable repeat (VR) can be found in phage DGRs. The role of mutagenesis of two functionally different genes is unclear. In this study, several phage genomes that contain DGRs with two target genes were found in the gut virome of healthy volunteer. Bioinformatics analysis of these genes indicated that they encode proteins with different topology; however, both proteins contain the C-type lectin (C-lec) domain with a hypermutated beta-hairpin on its surface. One of the target proteins belongs to a new family of proteins with a specific topology: N-terminal C-lec domain followed by one or more immunoglobulin domains. Proteins from the new family were named tentaclins after TENTACLe+proteIN. The genes encoding such proteins were found in the genomes of prophages and phages from the gut metagenomes. We hypothesized that tentaclins are involved in binding either to bacterial receptors or intestinal/immune cells.

Keywords: bacteriophage; genome sequence; diversity-generating retroelement; C-type lectin; Ig-like domain; receptor-binding protein; tentaclin; adhesin

1. Introduction

Diversity generating retroelement (DGR) is a prokaryotic molecular system that provides hypermutation in a certain variable region of the target gene, which is part of the DGR cassette [1]. Both bacteria and phages use this mechanism for rapid adaptation to permanent changes in the environment [2,3]. Probably, bacteria can also use DGR cassettes to increase the diversity of proteins that perform protective or immune functions [4]. Phages mainly use this mechanism to modify their receptor-binding proteins for maintaining the ability to infect host bacteria when certain components change on the cell surface [5–9]. In particular, this is observed for phages infecting bacteria that inhabit the intestine [10–13]. The habitat of such bacteria often changes depending on nutrition, the health of the macroorganism, and some external factors that can lead to modification of the state of bacteria including their surface molecules.

The signature components of the DGR cassette are the reverse transcriptase (RT) gene, template repeat (TR) with a length of 100-150 bp, and target gene containing a variable locus similar to the TR that is called a variable repeat (VR). At the 3'-end of the VR, the initiating of mutagenic homing (IMH) sequence is located, whereas IMH* sequence that is not identical to IMH is at the 3'-end of TR [1,2]. In addition, DGR cassette usually contains the accessory gene that encodes the accessory variability determinant (Avd) protein or its analog, which is essential for the DGR activity [6,14]. The molecular mechanism of DGR machinery is not completely clear. It has been established that the key stage is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2023 doi:10.20944/preprints202311.0981.v1

the reverse transcription of the RNA copy of TR by the RT, which substitutes only the adenine nucleotides during the process of hypermutagenesis [1,7]. Then, the mutated TR copy replaces VR in the target gene in the process of retrohoming. In the DGR cassette of the *Bordetella* phage BPP-1 that is the first described DGR the target gene *mtd* encodes the major tropism-determinant protein. This protein is connected to the distal end of tail fibers and is responsible for the recognition of receptor structures on the surface of the host cell [5–7,15]. So, modification of some receptor-binding proteins facilitates the adaptation of phage structures to possible changes in the surface structures of the cell [7–16].

It has been shown that approximately 13% of the analyzed phage DGR cassettes contain two target genes [1]. In this study, we found that several phage genomes that were assembled from human gut metagenomes also maintain DGR cassettes with two target genes. One of the target genes contained VR in the 5′-part of the gene, whereas in most DGR cassettes VR is located at the 3′-end of the target gene. Bioinformatic analysis indicated that both target proteins have different topology and the protein with VR located in its N-terminal part is a member of a new large family of proteins that were named tentaclins. Notably, hundreds of genes encoding tentaclins were found in the human gut microbiomes and bacterial genomes.

2. Results

2.1. Search for Phage DGRs with Two Target Genes

In order to find phage DGR cassettes with two target genes, several gut viromes of healthy people were sequenced and all assembled contigs were screened for the presence of essential phage genes. Contigs that contained the genes encoding both phage large terminase subunit and portal protein were selected. Then, sequences containing DGR cassettes were identified among the selected phage and prophage genomes using myDGR service [17]. Finally, complete DGR cassettes containing two target genes (target1 and target2) were found in two phage genomes named nd4 and nd12 (Figure 1). In both DGR cassettes from the nd4 and nd12 genomes, the target1 gene contained VR close to the 5′-end of the gene, whereas target2 had VR at its 3′-end like most of the known phage target genes [1].

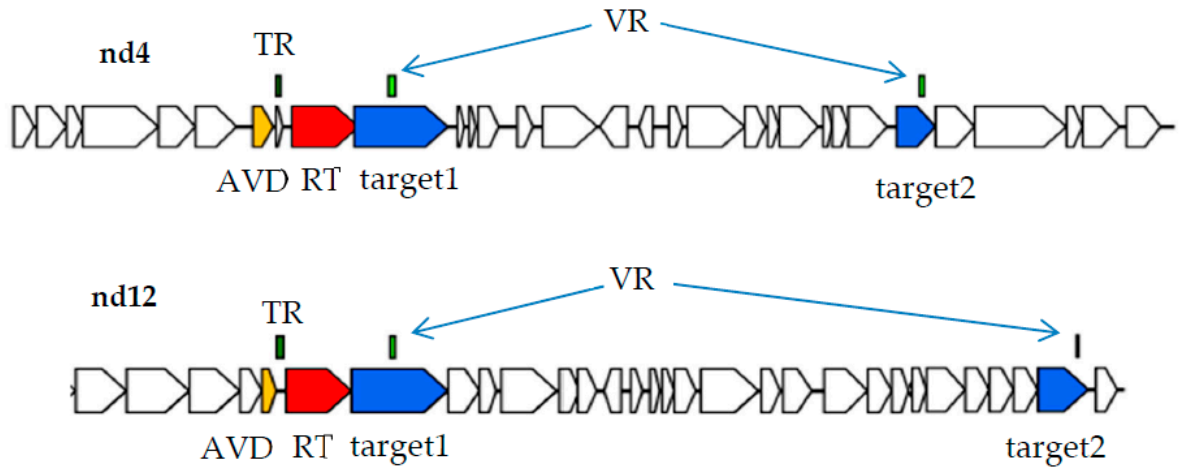


Figure 1. DGR cassettes from the nd4 and nd12 genomes generated using myDGR.

Phage genomic sequences nd4 and nd12 (46,542 bp and 32,765 bp, respectively) contained the gene encoding the tail sheath protein (Supplementary Data S1 and S2) that is a signature protein of phages with myovirus morphology. Notably, nd4 and nd12 were quite distant; their nucleotide identity (NI) was calculated as 31.2%. A search for sequences related to the nd4 and nd12 phage genomes in the GenBank databases revealed similar phage sequences from the human gut viromes for both studied phages (Figure 2). The size of the genomes similar to nd4 did not exceed 51 kbp (GenBank OP074837.1; query coverage 93%; NI 99%) and varied from 48.2 kbp to 49.3 kbp for the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2023 doi:10.20944/preprints202311.0981.v1

closest relatives (GenBank OP074460.1, BK055766.1, and OP074962.1; query coverage 96-97%; NI 98-99%). Therefore, the sequence of the nd4 genome can be considered almost complete. Similarly, the size of the genome that was most similar to nd12 was 36,318 bp (GenBank OP075616.1; query coverage 94%, NI 98.7%); so, nd12 genome contained ~90% of the complete sequence. Given the small size of the nd4 and nd12 genomes, the phages can be attributed to small myoviruses. Notably, only a single nd4-like prophage genome was found among the bacterial genomes. This prophage was found in the genome of the Flavonifractor plautii strain VE303-08 (query coverage 84%, NI 85%). No prophages similar to nd12 were found.

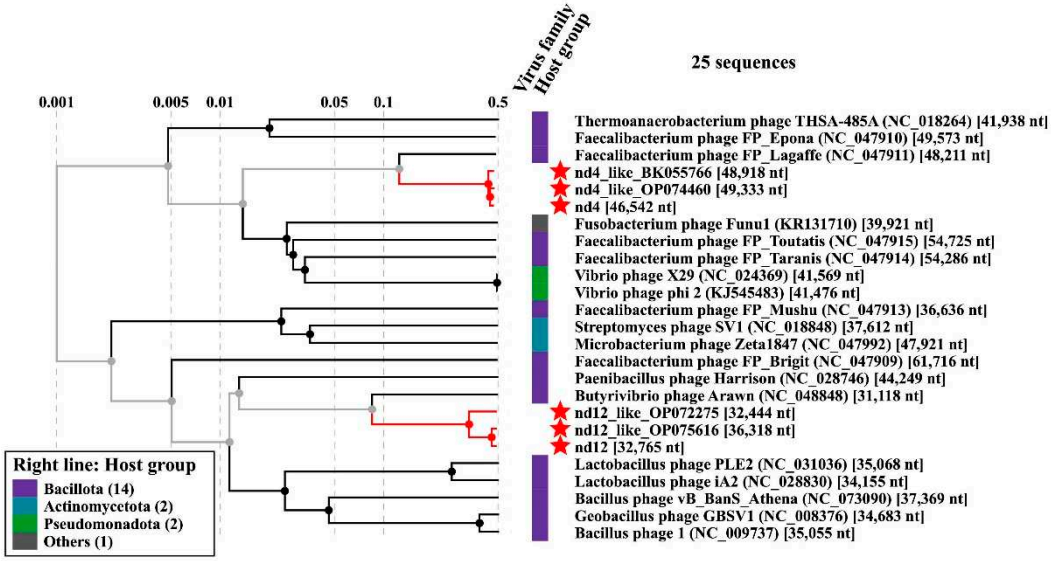


Figure 2. ViPTree generated proteomic dendrogram indicating the position of the nd4, nd12 and several nd4-like and nd12-like genomes (marked with red asterisks).

2.2. Proteins Encoded by the Target1 Genes from the nd4 and nd12 Genomes

Target genes containing VR close to the 5′-end of the gene are rare among phage DGR systems [1]. Analysis of the proteins encoded by the nd4 and nd12 target1 genes (569 and 579 aa, respectively) using AlignX (Vector NTI suite 8.0) indicated that the identity of these proteins is 47%. NCBI Conserved domain (CD) search tool revealed that the nd4_tgt1 and nd12_tgt1 proteins contain the N-terminal DUF6273 domain of unknown function followed by the fibronectin type 3 family (Fn3) domain belonging to the Ig-like domain superfamily. Analysis using HHpred did not provide additional information. To clarify the possible function of these proteins, Alphafold2 was applied to obtain putative three-dimensional (3D) structures of the proteins with high confidence (high pLDDT score) (Figures 3 and S1). According to the models, the nd4_tgt1 and nd12_tgt1 proteins have similar 3D structures: they contain the N-terminal globular domain followed by three beta-sandwich domains that belong to the immunoglobulin (Ig) superfamily. The search for similar structures among experimentally determined structures using DALI indicated that the N-terminal DUF6273-like domain resembles the C-type lectin (C-lec) domain. Notably, VR forms a characteristic beta-hairpin type structure flanked by two additional loops located on the surface of this domain in both nd4_tgt1 and nd12_tgt1 (Figures 3 and 4). Molecular dynamics relaxation for 50 ns did not reveal any significant deviations in the conformation of the hairpin during the simulation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2023 doi:10.20944/preprints202311.0981.v1

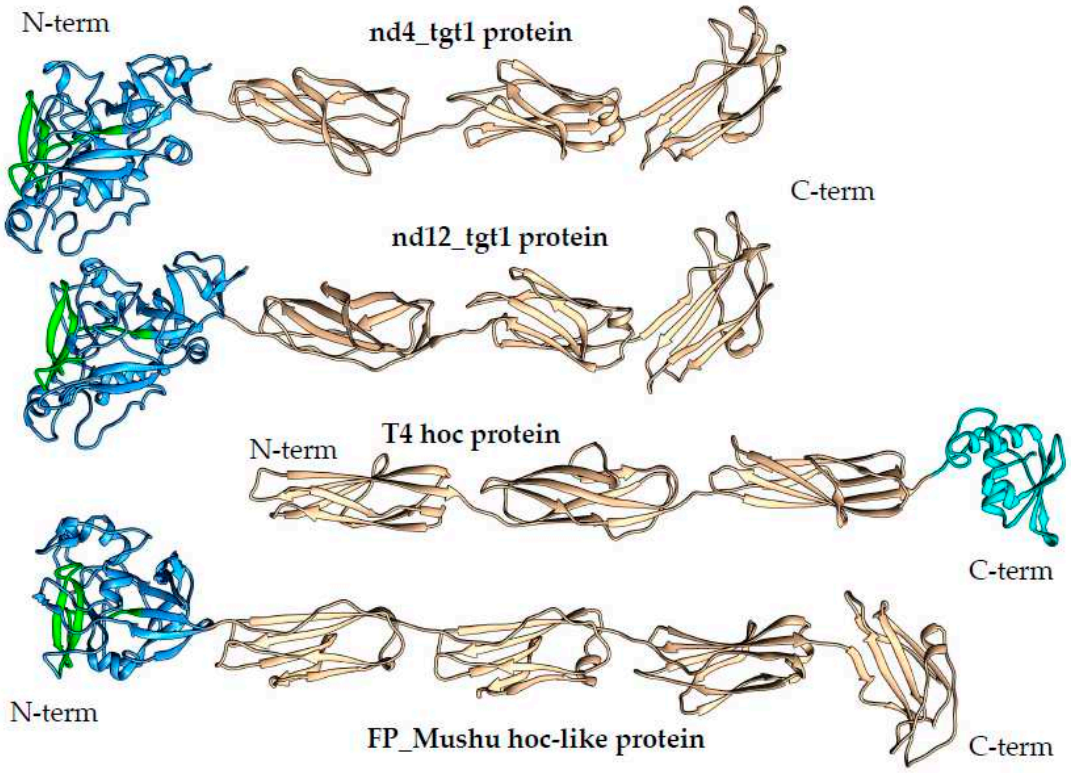


Figure 3. Ribbon representation of 3D structures of the nd4_tgt1, nd12_tgt1, T4 Hoc proteins and FP_Mushu phage hoc-like protein. C-type lectin domains are in blue, Ig domains are in tan, anchoring domain of the T4 Hoc protein is cyan. VR-encoded regions are marked with green. 3D models were predicted using AlphaFold2 and rendered using UCSF Chimera, version 1.13. Also see Figure S1.

The presence of Ig-like domains indicated a possible structural similarity of the nd4_tgt1 and nd12_tgt1 proteins with the capsid-embedded Hoc-protein of the phage T4 (Figure 3). Moreover, similar proteins have been also mentioned as hoc-like targets in DGR cassettes of the prophages FP_Mushu and FP_Brigit found in Faecalibacterium prausnitzii [18]. However, nd4_tgt1 and nd12_tgt1 proteins have some differences compared to the T4 Hoc-protein. According to the alphafold model [19], the Hoc protein of the T4 phage does not contain a globular N-terminal lectin domain and the structure of the C-terminal capsid-anchoring domain is also different. Modeling using AlphaFold2 showed that the hoc-like proteins of the FP_Mushu and FP_Brigit phages have a topology more similar to that of the nd4_tgt1 and nd12_tgt1 proteins than the T4 Hoc. The hoc-like proteins of the FP_Mushu and FP_Brigit phages also have a C-lec domain at their N-terminus; however, it is followed by four Ig-like domains instead of three ones (Figures 3 and 6). Importantly, the aa sequence similarity of the nd4_tgt1/nd12_tgt1 proteins and the hoc-like proteins of the FP_Mushu and FP_Brigit was low despite the similar topology. The overall protein identity for these four proteins was ~ 7%; however, last ~100 aa residues forming the C-terminal Ig-like domain (Cterm_Ig domain) showed identity of ~17%. Thus, we assume that the Cterm_Ig-like domain is involved in embedding of the tgt1 protein into the phage virion, whereas N-terminal C-lec domain is used for binding to some receptors.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2023 doi:10.20944/preprints202311.0981.v1

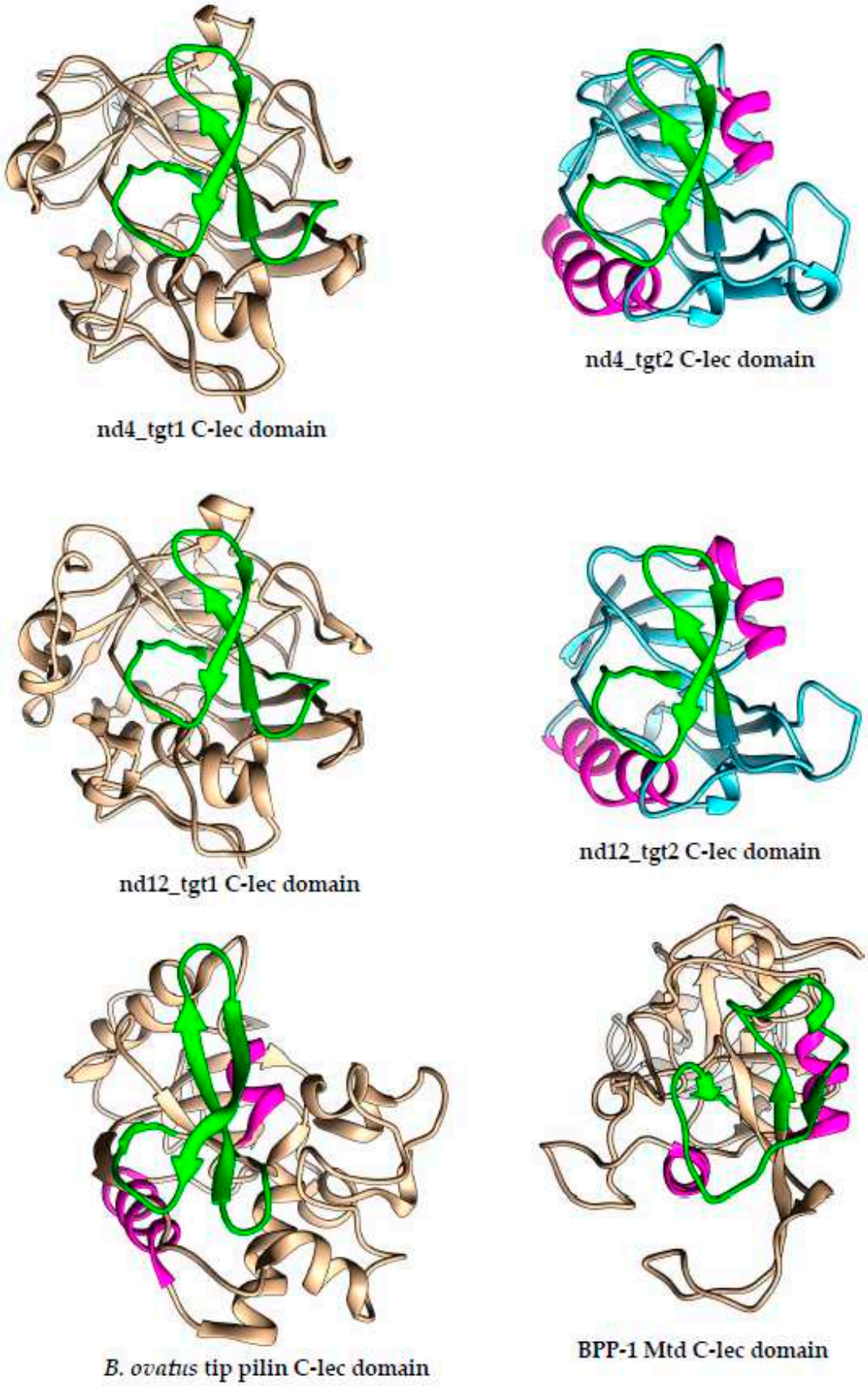


Figure 4. Ribbon representation of 3D structures of C-lec domains of the nd4_tgt1, nd4_tgt2, nd12_tgt1 and nd12_tgt2 proteins. Structures of C-lec domains of the BPP-1 Mtd protein (pdb 1YU0) and tip pilin of Bacteroides ovatus (pdb 4EPS) are shown for comparison. VR-encoded regions are marked with green. Highlighted alpha helices indicate similar orientation of the molecules. 3D

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2023 doi:10.20944/preprints202311.0981.v1

models were predicted using AlphaFold2, structure relaxation performed using GROMACS, and final models were rendered using UCSF Chimera, version 1.13.

2.3. Comparative Analysis of the nd4_tgt1 and nd12_tgt1 Proteins

Given the high similarity of Cterm_Ig domains of the nd4_tgt1, nd12_tgt1, and the hoc-like proteins of the FP_Mushu and FP_Brigit phages, sequences of their Cterm_Ig domains were used to find similar proteins using BLASTp search. As a result, 912 heterogeneous sequences with various degree of similarity were extracted. Notably, a clear consensus of seven conservative aa residues was found in the Cterm-Ig regions, despite the low similarity of these regions (Figure 5A). Most of the conservative residues were located close to each other and formed a specific structure (Figure 5B).

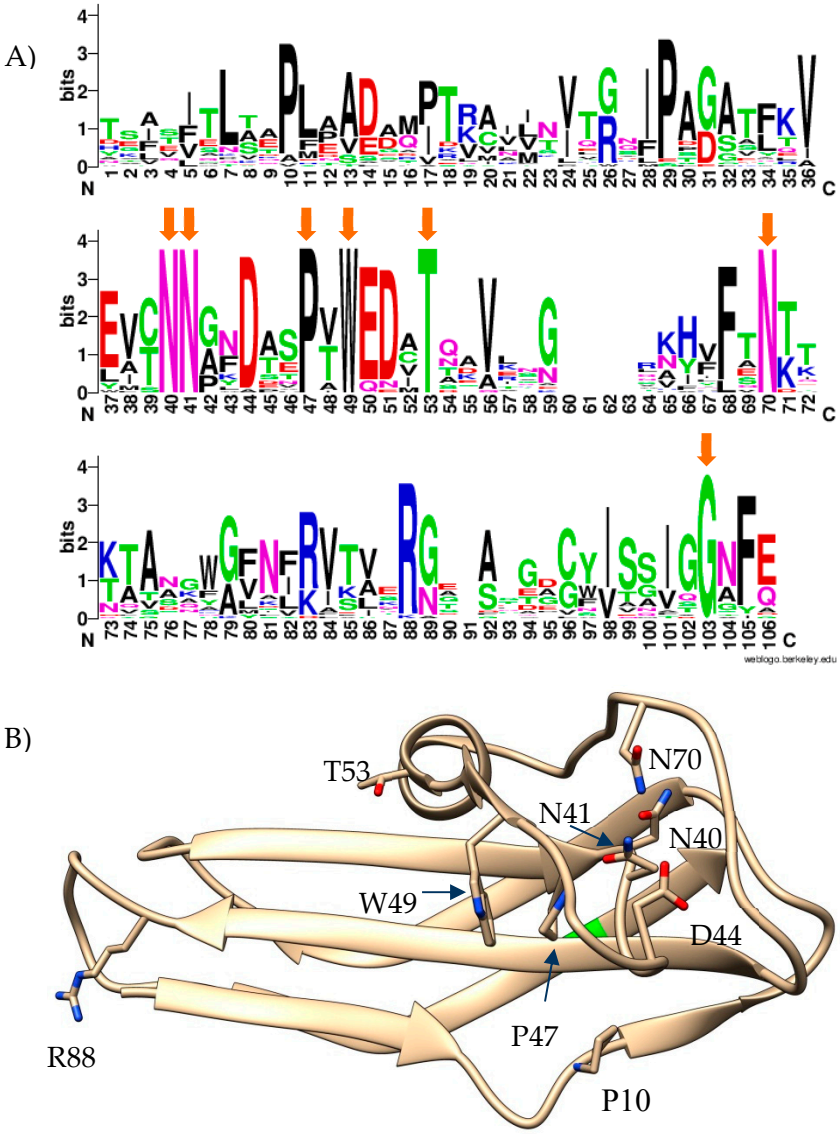


Figure 5. Conserved residues of the Cterm-Ig domain. A) Weblogo diagram representing consensus sequence of the Cterm-Ig domain of 29 selected proteins containing this domain. Orange arrows indicate 100% conserved aa residues. B) Ribbon view of nd12_tgt1 Cterm-Ig domain. Residues are numbered according to Weblogo diagram.

Some of these 912 Cterm_Ig domain-containing sequences were analyzed using AlphaFold2. The obtained results indicated that even proteins with the lowest aa identity (~ 25%) and a low expectation value (~ 0.05) had a topology similar to the nd4_tgt1 and nd12_tgt1 proteins. Importantly, all analyzed proteins had the identified aa consensus in their last 100 aa sequences. Of these 912 sequences, 329 were phage sequences (up to 649 aa) encoded mainly by the metagenome assembled

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2023 doi:10.20944/preprints202311.0981.v1

genomes (MAGs); the rest sequences (up to 1445 aa) were found in bacteria (mainly in Brevibacillus spp. and Bacillus badius).

So, a large group of phage proteins was discovered. Since all of them contained a C-lec domain at the N-terminus, and the vast majority of C-lec domains bind polysaccharides or proteins [20], these proteins are probably receptor-binding ones. In addition, these proteins contain the Ig-like domains that are connected by unstructured regions; so, the proteins are possibly flexible like the muscle protein titin [21] or bacterial adhesins – invasin and intimin [22]. Taking into consideration the prevalence of such proteins and their possible flexibility, these proteins were names tentaclins after (TENTACLe + proteIN). Despite a certain size variability, the specific features of tentaclins are the presence of the N-terminal C-lec domain with the characteristic beta-hairpin structure and several Ig-like domains with the Cterm domain containing a particular consensus (Figures 5 and 6).

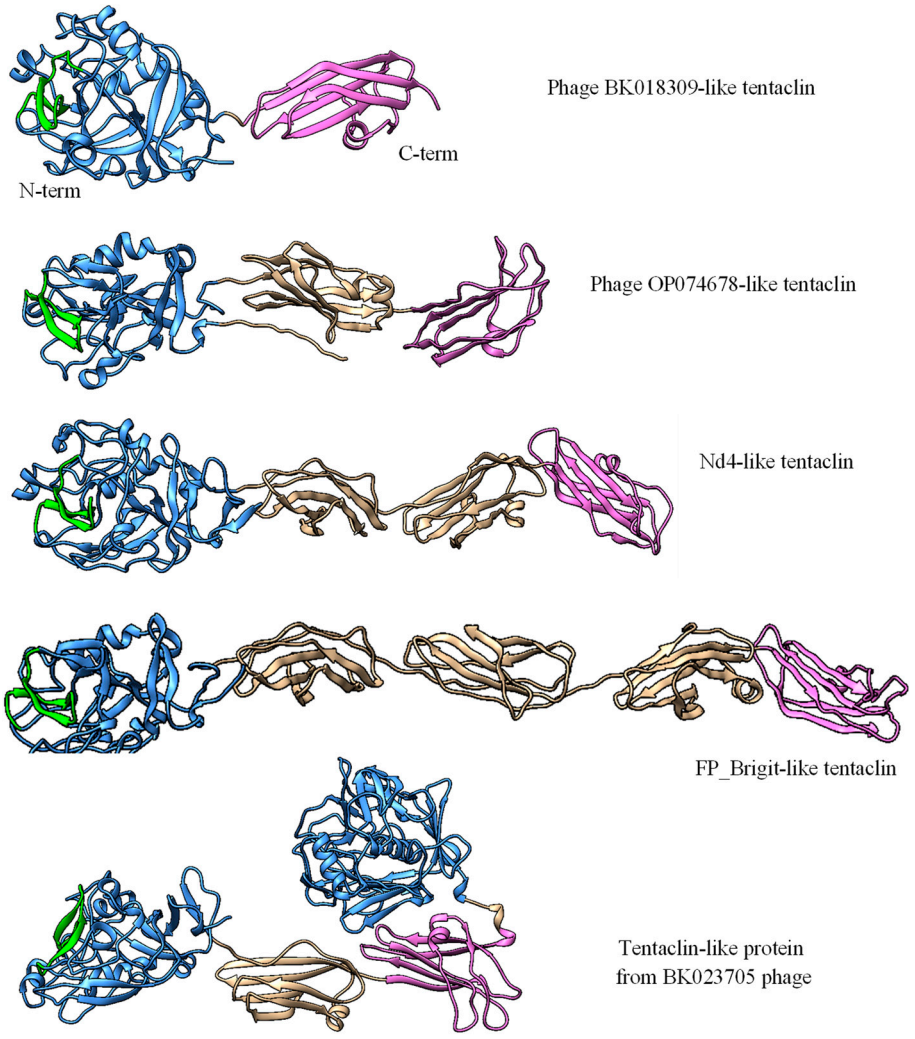


Figure 6. Tentaclins from various gut phage metagenome sequences. C-type lectin domains are in blue, beta-hairpins are in green, C-terminal anchoring Ig domains are in pink, other Ig domains are in tan. (Also see Figure S1 for pLDDT-colored models)

The number of Ig-like domains that occur in the tentaclins of phages and bacteria varied. All phage tentaclins contained from one to four Ig domains (Figure 7) and had a size from 313 aa to 649 aa. In addition, the phage genome (BK023705) was found that encoded a tentaclin-like protein with a size of 830 aa. This protein contained an additional C-lec domain at the C-terminus. This C-lec domain was similar (~60% identity) to one of the tail collar domains of myoviruses. The anchoring Cterm-Ig domain presumably required for tentaclin attachment to the virion was also found in this

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2023 doi:10.20944/preprints202311.0981.v1

tentaclin-like protein (Figure 7). As for bacteria, their genomes encoded both phage-like tentaclins ranging from 313 aa to 650 aa and more complex tentaclin-like proteins (Figure S3). In Brevibacillus spp., the genes encoding tentaclin-like proteins up to 1,445 aa were identified in addition to genes encoding "ordinary" tentaclins with a size of ≤ 650 aa.

2.4. Diversity of Phages Containing the Tentaclin Genes

To analyze the diversity of phages containing the tentaclin genes, 373 tentaclin aa sequences annotated as phage proteins were extracted from the GenBank non-redundant protein database (nr) using BLASTp. For these sequences, the corresponding phage genomes were selected and grouped using VIRIDIC (Figures 7 and S2).

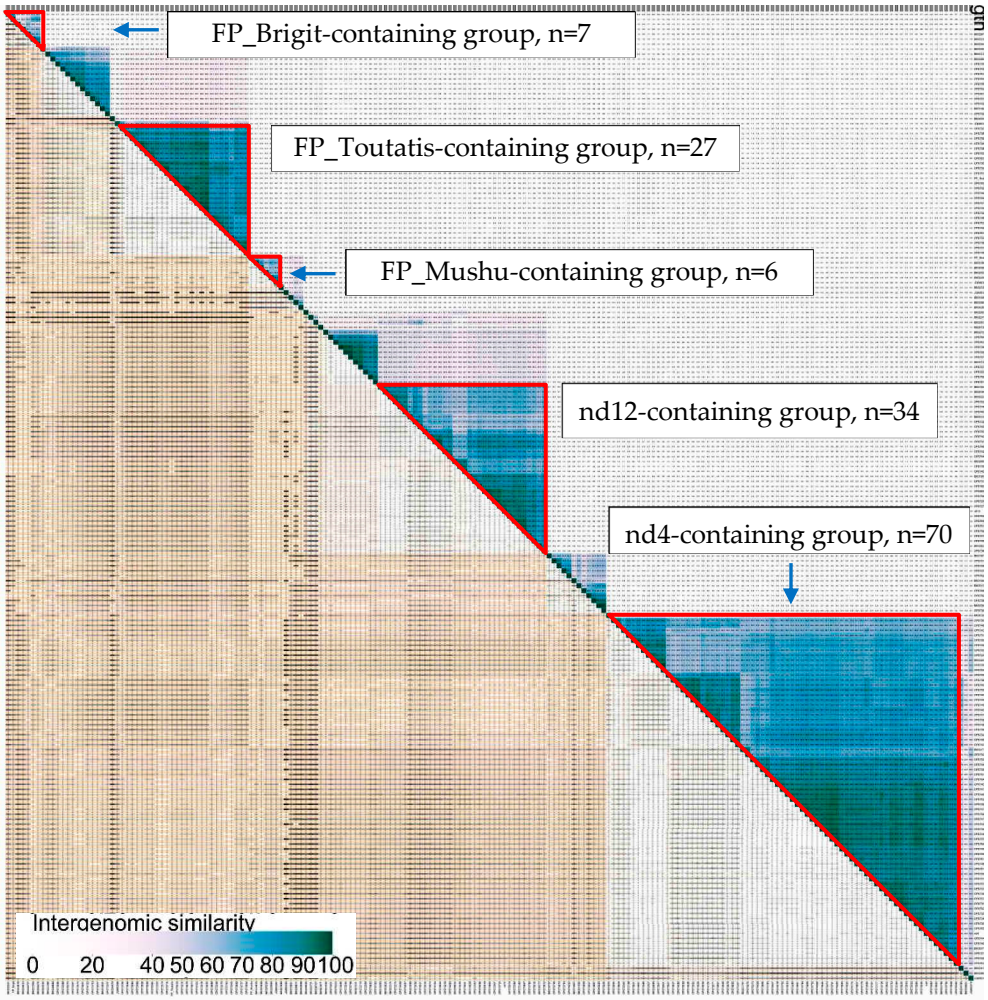


Figure 7. VIRIDIC heatmap indicating intergenomic similarity between phage sequences containing tentaclin genes. In this figure, 195 sequences out of 373 are given. Also see Figure S2.

The majority of these sequences were MAGs, with the exception of the previously described prophages from Faecalibacterium prausnitzii [18]. Grouping criteria were chosen as: at least 40% intergenomic similarity (IS) with any member of the group, at least 60% IS with at least one member of the group. The nd4 phage was part of the largest group that included 70 sequences (~19% of all sequences). The nd12 phage was part of the second largest group containing 34 sequences (~9% of all sequences). The FP_Toutatis phage formed the third group of 27 sequences. The FP_Mushu and FP_Brigit phages grouped with five and six MAGs, respectively (Figure 7). The remaining sequences were unique or they formed small groups (Figure S2). Since the intergenomic similarity between the

 doi:10.20944/preprints202311.0981.v1

phage genomes from different groups in most cases was less than 10%, it can be concluded that the tentaclin genes occur in phages that are distant from each other and belong to at least different sub-families.

From five groups containing at least ten phage genomes, several sequences were randomly selected and analyzed for the presence of DGR cassette (Figure S4). It was shown that each analyzed phage genome contained the tentaclin gene as part of its DGR cassette.

2.5. Proteins Encoded by the Target2 Genes from the nd4 and nd12 Genomes

As for the target2 gene that contains VR at the 3'-end, we expected that this gene would encode a protein resembling the Mtd protein of the phage BPP-1. This would be in good agreement with the fact that small myoviruses FP_Lagaffe and FP_Epona found in the genomes of the Faecalibacterium prausnitzii strains contain mtd-like genes as part of their DGR cassettes [18]. However, the analysis using HHpred did not reveal a significant similarity of the secondary structure of the nd4_tgt2 and nd12_tgt2 proteins with proteins from the PBD database, including the Mtd protein. According to NCBI CD-search, both nd4_tgt2 and nd12_tgt2 proteins contain only DUF6273 domain, whereas the identity between the proteins was only 24%.

AlphaFold2 modelling indicated that nd4_tgt2 and nd12_tgt2 proteins consist of a single globular domain flanked by short alpha-helix regions (Figure 8). Analysis of the alphafold models of the nd4_tgt2 and nd12_tgt2 proteins using DALI indicated that the globular domains exhibited similarity with the C-type lectin domain, like it was found for the N-terminal domains of the tentaclins from nd4 and nd12 (Figure 4). Notably, Mtd_BPP-1 protein that mediates binding to the bacterial receptor pertactin also contains a C-lec domain with VR at the C-terminus [16]. However, nd4_tgt2 and nd12_tgt2 proteins show substantial differences from the Mtd_BPP-1 protein: they are shorter (234 aa for nd4_tgt2 and 294 aa for nd12_tgt2 versus 381 aa for Mtd_BPP-1); they do not have a beta-sandwich domain and N-terminal beta prism, by which the Mtd_BPP-1 trimer presumably attaches to the tail fiber protein [15,23]. Nevertheless, we suppose that the nd4_tgt2 and nd12_tgt2 proteins can perform a function similar to the Mtd_BPP-1 protein despite the differences.

The nd12_tgt2 protein contains an additional alpha-helical motif at the N-terminus, which is present in some related phages but absent in the orthologous nd4_tgt2 protein (Figure 8, yellow part). In this alpha-helical motif, InterproScan recognized a DUF3310-like motif that is found in phage and bacterial proteins. It is not yet clear, whether this motif is involved in the formation of multimeric complexes or it forms an interface for interaction with other phage proteins in the same way as the beta-prism domain of Mtd_BPP-1 interacts with the tail fiber protein.

It should be noted that VRs in the nd4_tgt2 and nd12_tgt2 proteins also form characteristic beta-hairpin structures on the surface of the C-lec domain as in the nd4 and nd12 (Figure 4). Strikingly, the shape of the hairpin structure in both target proteins (nd4 and nd12 tentaclins and nd4_tgt2/nd12_tgt2) is similar despite the different aa sequences, domain sizes and folding details. In all studied C-lec domains, the hairpins occupy a considerable part of the surface (Figure S5) and possibly form the receptor-binding region of these proteins. This can explain how one TR can be used as a template for hypermutagenesis of two different proteins despite the differences in their topology. Notably, VR in the Mtd_BPP-1 protein that is also subjected to hypermutagenesis has a different conformation – a loop containing a short beta–strand (Figure 4). However, the structure of the C-lec domain of the terminal pilin of Bacteroides ovatus (pdb 4EPS) shows a beta-hairpin that is similar to that of the tentaclins and tgt2 proteins. This fact indicates that C-lec domains with the beta-hairpin motif, which were found in phage and bacterial proteins, can perform similar functions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2023 doi:10.20944/preprints202311.0981.v1

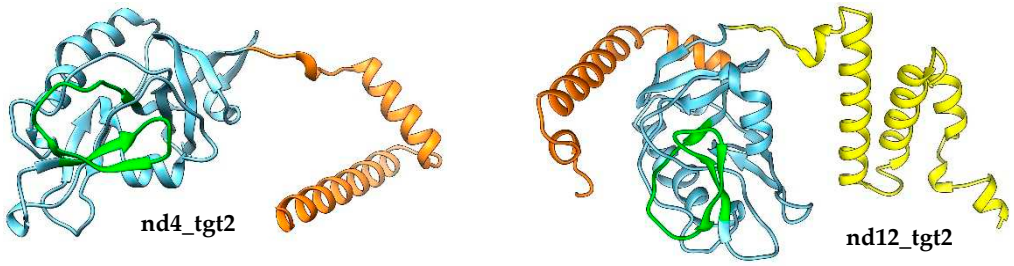


Figure 8. 3D models of the nd4_tgt2 and nd12_tgt2 proteins. C-type lectin domains are in blue, N-terminal alpha helices are yellow, C-terminal alpha helices are orange. VR-encoded regions are marked with green. 3D models were predicted using AlphaFold2 and rendered using UCSF Chimera, version 1.13. Also see Figure S1.

2.6. Analysis of the Hypermutagenic Potential of TRs from nd4-like and nd12-like Phages

It was analyzed whether there is a similarity of beta-hairpins sequences encoded by VR1 and VR2 in two target genes from the same phage genome. These VR sequences originated from the same TR during hypermutagenesis and the analysis of aa substitution in both VR1 and VR2 was of particular interest. If such a similarity could be detected, it would suggest that both target proteins bind to the same receptor. To test this hypothesis, TR and VR sequences from the nd4, nd12, and related genomes were involved in the analysis.

A total of 54 putative phage genomes with TR sequences identical to that in the nd4 genome were selected from the GenBank Nucleotide collection (nt) database. Analysis using VIRIDIC indicated that these phages probably belong to the same genus (intergenomic similarity > 87%). Only 39 from 54 genomes contained DGR cassettes with two target genes [17] (Figure 9). TR in the nd4-like genomes contains 25 adenines (Figure 9). Of them, 24 positions were mutated in VR1 (in the target1 gene) in at least one of the nd4-like genomes (Figure 9). As a result, 13 aa were substituted in the tentaclin of nd4. In other nd4-like phages, from 7 aa to 19 aa substitutions in their tentaclins were identified (Figure 10). As for VRs of the target2 genes, 21 adenines from 21 ones could be mutated as VR2 is shorter. So, eight aa were substituted in the nd4_tgt2 protein and from 6 to 13 substitutions were found in the orthologous proteins of nd4-like phages (Figure 10).

As for the nd12-like genomes, 13 genome sequences were found that contained TR identical to that in the nd12 genome and two target genes in their DGR cassettes (Figure 9). A total of 28 adenines were found in TR in these genomes; all of them were mutated in VR1 and 20 adenines could be mutated in VR2. In the nd12 phage, 12 aa were substituted in the tentaclin (from 9 aa to 15 aa in the nd12-like phages) and 9 aa were mutated in the nd12_tgt2 protein (from 6 aa to 9 aa in other relative phages) (Figure 10).

In addition, along with substitutions A→N, there were 14 and 5 substitutions B→B (B = T, C or G) found in both VR1 and VR2 sequences in the nd4-like and nd12-like genomes, respectively. Probably, these mutations appeared independently of retrohoming mediated by DGR.

It is noteworthy that in each examined phage genome, VR1 and VR2 sequences differed between themselves. Only sometimes, mutations coincided in corresponding positions in both VRs of the same phage genome.

Notably, most adenines are grouped in pairs in TRs of both nd4-like and nd12-like phages (Figure 9) and adenine in the second position of the AAC codons rarely changed to cytosine in contrast to adenine in the first position. Such an imbalance between the A→C mutations in the first and second positions can be caused both by a feature of hypermutagenesis for double adenines and the result of selection of preferred aa residues in target proteins. As for the AAT codons, which were found only in TRs of the nd12-like phages, both adenines infrequently changed to cytosine (Figure 9).

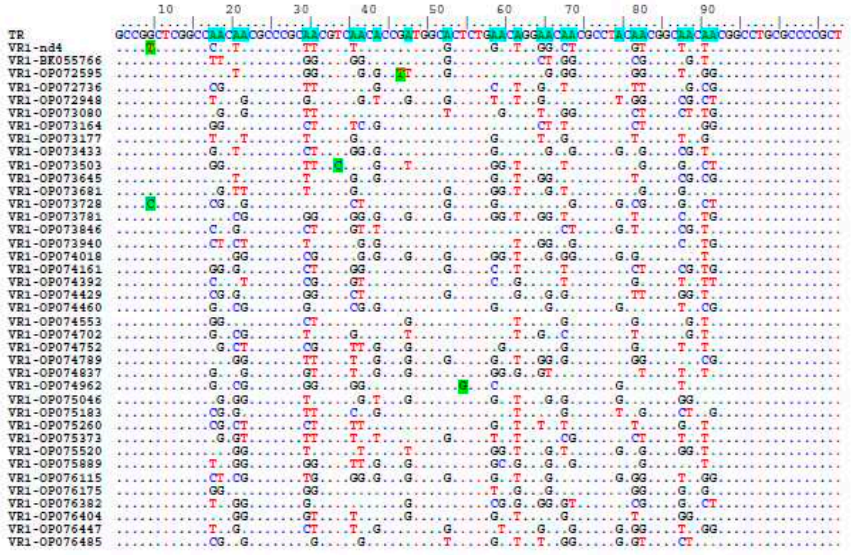
Since most adenines in the studied TRs are grouped into AAC and AAT codons (both encode Asn), 15 aa residues could appear as a result of hypermutagenesis (with the exception of *Gln*, *Met*, *Lys*, *Glu*, and *Trp*). However, substitutions for aromatic and charged (*Arg* and *Asp*) aa residues in the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2023 doi:10.20944/preprints202311.0981.v1

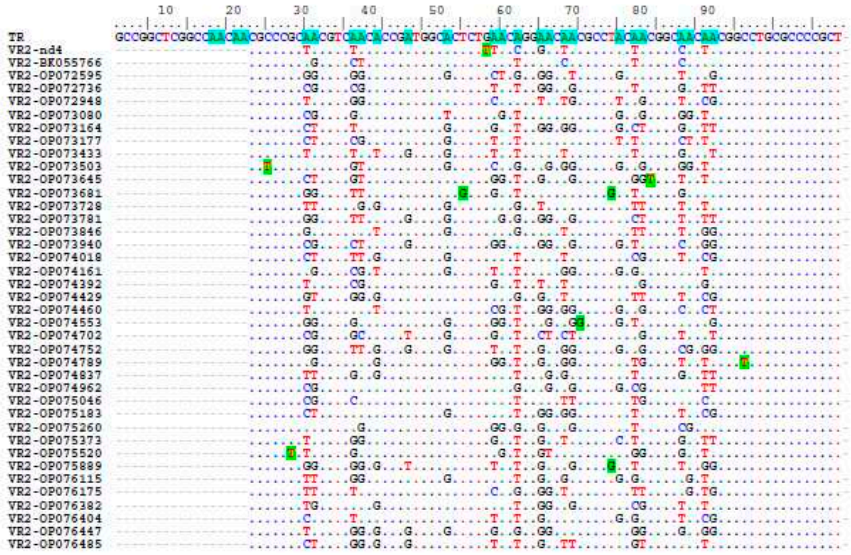
nd4-like phages and aromatic aa residues in the nd12-like phages are predominantly found. In addition, *Asn* is also often substituted by *Ser* in both groups of phages, unlike rare replacements for *Thr* (Figure 10).

Notably, there is a tendency to replace aa residue with *Cys* within the *TyrAsnGlyAsnAsn* motif of TR. According to 3D models of C-lec domains, this *Cys* appears close to another *Cys* residue outside the beta-hairpin. We suppose that such mutations lead to the formation of a disulfide bridge that stabilizes the C-lec domain (Figures 10 and S6).

A) TR vs. VR1 alignment (nd4-like phages):

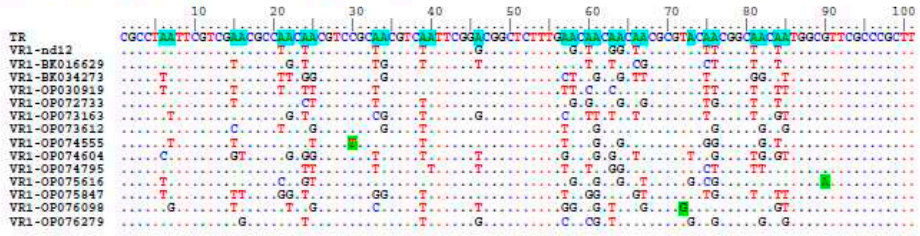


TR vs. VR2 alignment (nd4-like phages):



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2023 doi:10.20944/preprints202311.0981.v1

B) TR vs. VR1 alignment (nd12-like phages):



TR vs. VR2 alignment (nd12-like phages):

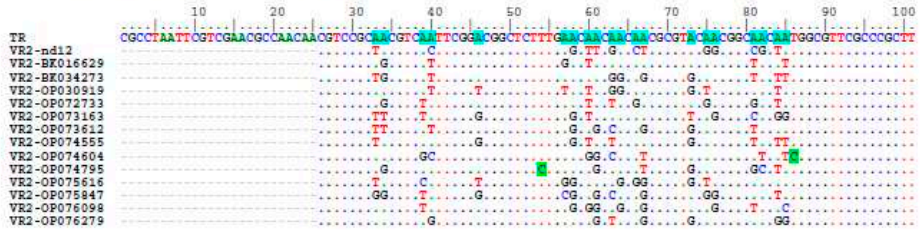


Figure 9. Alignment of nucleotide sequences of TR vs.VR1 and TR vs. VR2 for nd4-like (A) and nd12-like phages. Substitutions arising from nucleotides other than A are highlighted in green.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2023 doi:10.20944/preprints202311.0981.v1

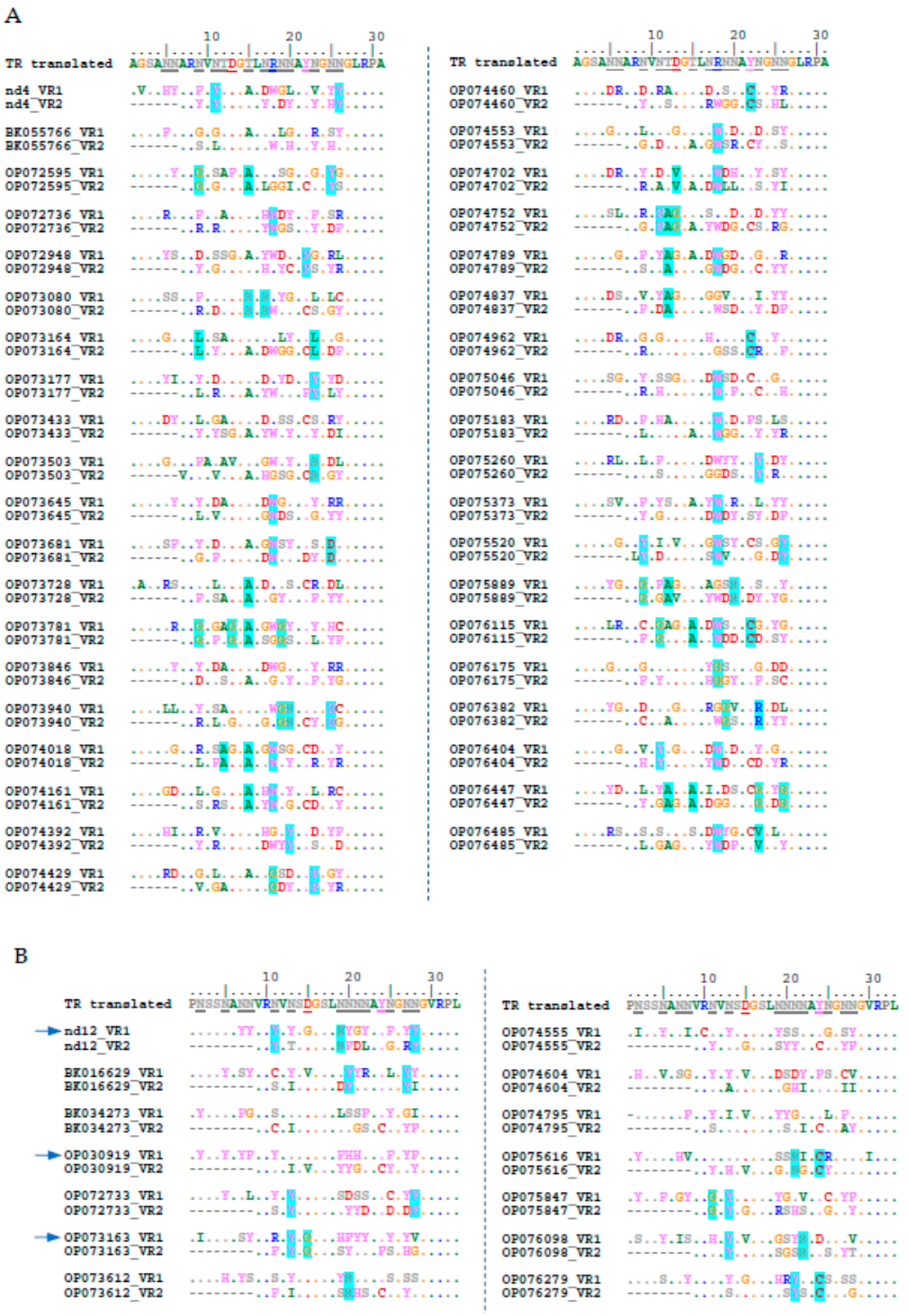


Figure 10. Alignment of aa sequences of TR, VR1 and VR2 in the genomes of the nd4-like (**A**) and nd12-like (**B**) phages. In TR, aa residues that can mutate as a result of retrohoming are underlined. Negatively charged aa are shown in red, positively charged aa in blue, aromatic aa in lilac, hydrophobic non-aromatic aa (except cysteine) in green, and hydrophilic uncharged aa in gray. The blue background marks the positions in which the same aa appeared as a result of mutagenesis and selection.

3. Discussion

In this study, we addressed the question of how one TR sequence can simultaneously be a template for two different VRs in the target genes found in the DGR cassettes of metagenomic phages nd4 and nd12. It is noteworthy that in both phages, the VR1 sequence is located in the 5′-terminal

 doi:10.20944/preprints202311.0981.v1

part of the target1 gene, whereas VR2 is found at the 3'-end of the target2 gene. In addition, no sequence similarity was observed between the target1 and target2 genes from each phage and 3D structure prediction indicated that the proteins encoded by the genes have different topology. However, both nd4_tgt1/nd12_tgt1 and nd4_tgt2/nd12_tgt2 proteins contain the C-lec domain, which is known to be involved in binding to certain proteins or oligosaccharides [24]. Importantly, the hypermutated site (VR) in the studied target genes encodes a beta hairpin located on the surface of the C-lec domain (Figures 4 and S5). Probably, hypermutagenesis of VRs in both target proteins is required for these phages to adapt to the changing environment, including modification of the bacterial receptors profile.

Proteins encoded by the target1 genes of the nd4 and nd12 phages, in addition to the N-terminal C-lec domain, contain three the Ig-like domains. Genes encoding proteins with a similar topology (N-terminal C-lec domain followed by Ig/Fn3-like domains) have been found in other MAGs (n > 350) and bacterial genomes (probably in prophages) and may contain from one to four Ig-like domains (Figure 6). The presence of the C-lec domain with the hypermutated beta hairpin indicates possible involvement of these proteins in receptor binding. A chain of several Ig-like domains gives overall flexibility to such molecules, similar to bacterial adhesins [22]. Given the prevalence of such proteins and their possible flexibility, we propose to call these proteins "tentaclins" (TENTACLe + proteIN). Importantly, the C-terminal anchor Ig-like domains of tentaclins have a clear consensus motif, despite the high diversity of the aa sequences of the domains (Figure 5A). Along with the specific topology, this motif can serve as a distinctive feature of tentaclins. As for the remaining Ig-like domains, it is not clear whether they are involved in additional binding to any molecules. So, a novel family of proteins with specific structure was discovered. These proteins contain the N-terminal C-lec domain with a specific hairpin structure on its surface, followed by one to four Ig-like domains and the C-terminal Ig-like domain has a consensus motif. The Tentaclin family is divergent and quite numerous and tentaclins occur in at least hundreds of phages.

Apparently, the topology of the tentaclins is favorable and has been repeatedly used during evolution. In addition, more complex proteins from phages and bacteria that have "tentaclin"-like organization were found. These molecules contain some additional elements besides C-lec and Ig-like domains. Examples of such molecules are a protein from *Brevibacillus* sp. (GenBank id: NRS19645) containing an additional beta-propeller domain and a phage protein (BK023705) with the second C-lec domain at the C-terminus (Figures 6 and S3). It should be noted that in the nd4-like and nd12-like MAGs containing only one target gene within DGR cassette, this gene encoded the tentaclin an all cases. This fact confirms the importance of tentaclins.

As for the target2 genes of the nd4, nd12 and relative phages, they encode proteins that differ from tentaclins. Since nd4_tgt2 and nd12_tgt2 have C-lec pattern, they probably specifically recognize some bacterial structures. We hypothesize that the nd4_tgt2 and nd12_tgt2 proteins are involved in binding and infecting host cells, similar to the Mtd protein of the BPP-1 phage.

Comparison of VR sequences between tentaclins and tgt2 proteins showed that these sequences differ both within the same phage and between related phages. The profile of selected mutations in both target proteins has a clear shift towards aromatic residues and *Ser* (for nd4-like and nd12-like phages) and also charged residues including *Arg* (for nd4-like phages). Apparently, this type of aa residues in hypermutated sites provides the best binding to yet unknown receptors recognized by these proteins.

It is unclear whether phages use their tentaclins and tgt2 proteins to bind to the host bacterium, or these proteins perform different functions. The first hypothesis is that only tgt2 proteins bind bacterial receptors, whereas tentaclins, like bacterial adhesins, are used for interaction with receptors on the surface of the intestinal epithelium, which allows phages to remain in the intestine. This hypothesis is supported by the fact that no similar pattern of mutated aa residues was found among VR sequences within the same phage. Moreover, some bacterial adhesins, such as invasin and intimin, have a similar organization – one C-lec domain and several Ig-like domains, although the C-lec domain is located at the C-terminus [25]. However, the profile of proteins and polysaccharides on the surface of intestinal cells is relatively constant, and hypermutagenesis of VR sequences in

 doi:10.20944/preprints202311.0981.v1

tentaclins is not required. The second hypothesis is that phages use tentaclins for interaction with immune cells that present in the intestine. These cells have a wide range of receptors and phages that interact with them can affect their immune response and thereby participate in the interection between bacteria and macroorganism. In this case, hypermutagenesis of the VR sequences helps phages to adapt to the dynamic profile of immune cell receptors. The third hypothesis is that both tentaclins and tgt2 proteins bind to different receptors of one bacterial host or recognize different epitopes within the same receptor. Thus, it has been shown that the *Bordetella* BPP-1 and *Bordetella* BPP-6 phages recognize the same bacterial receptor pertactin, despite they have different VR sequences in the Mtd protein [16].

In conclusion, the organization and role of two different target proteins from the same DGR cassette of metagenomic phages were investigated using bioinformatic methods. It was shown that one of the target proteins can be a member of a novel family of proteins - tentaclins Tentaclins have a specific topology and the genes encoding tentaclins are relatively common in phage and bacterial genomes. The obtained data can be useful for further study of the mechanism of retrohoming and the molecular organization of phages that affect bacteria inhabiting the intestine.

4. Materials and Methods

4.1. Virome Sequencing

Viral DNA isolation and sequencing from a fecal sample was described previously [26]. Briefly, the sample from a healthy volunteer was re-suspended in sterile phosphate-buffered saline and clarified by several consecutive centrifugations. DNase I, 5 U (Thermo Fisher Scientific, MA, USA) was added to the final supernatant and the mixture was incubated for 4 h at 55 °C. Then, the mixture was treated with Proteinase K, 100 μg/ml (Thermo Fisher Scientific, MA, USA) supplemented with 20 mM EDTA and 0.5% SDS for 3 h at 55 °C. Phenol-chloroform extraction with subsequent ethanol precipitation was used for DNA purification. The obtained DNA was diluted in 50 μl of 10% TE-buffer and, after measuring the concentration by Qubit 4.0 (Thermo Fisher Scientific, MA, USA), used in the standard procedure for constructing a virome shot-gun library using the NEB Next Ultra DNA library prep kit (New England Biolabs, MA, USA). A MiSeq Benchtop Sequencer (Illumina Inc, CA, USA) and a MiSeq Reagent Kit 2 × 250 v.2 (Illumina Inc, CA, USA) were used for sequencing. The obtained sequences were assembled *de novo* using both the CLC Genomics Workbench software v.6.0 and SPAdes 3.15.

4.2. Genome Analysis

All contigs longer than 10 kb obtained after assembly in SPAdes were used to search for similar sequences in the NCBI GenBank Nucleotide collection (nt) database using BLASTn. Sequences found to be similar to phage sequences were analyzed for the presence of portal protein and large terminase subunit genes. “Positive” sequences were analyzed for the presence of the DGR cassette using the myDGR online service (https://omics.informatics.indiana.edu/myDGR/) [17]. A search for nd4- and nd12-related sequences was performed using BLASTn and the NCBI GenBank nt database. Genome annotation was carried out using RAST server v. 2.0 (https://rast.nmpdr.org/) [27]. In addition, manual verification of the annotation results was carried out using the NCBI GenBank nr protein database. Comparative analysis of nd4 and nd12 genomic sequences was performed using the ViPTree version 3.7 web server (https://www.genome.jp/viptree) with default parameters [28].

4.3. Analysis of target1 and target2 gene functions

BLASTp search, NCBI Conserved domain search (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) and HHPred search (https://toolkit.tuebingen.mpg.de/tools/hhpred) were used to predict putative functions of proteins encoded by target1 and target2 genes. DALI server (http://ekhidna2.biocenter.helsinki.fi/dali/) was used to find structural similarity between alphafold2-generated models and experimental structures [29].

 doi:10.20944/preprints202311.0981.v1

4.4. Modeling of Protein 3D Structure and Molecular Dynamics Simulation

3D models of proteins were predicted using ColabFold v1.5.3 implementation of AlphaFold2 program available at https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb [30]. Only models with a high degree of confidence (average pLDDT > 75) were used for the study. The models were visualized using UCSF Chimera, version 1.13 [31].

Protein relaxation was performed using GROMACS v2020.3 [32] on Nvidia V100-equipped GPU nodes of the High Performance Computing Center of Novosibirsk State University ("NUSC NSU"). Molecular dynamics simulations were performed for 50 ns at 310 K and 1 bar pressure using the amber99SB force field and tip3p water molecules. Molecular dynamics trajectories were analyzed using VMD v 1.9.3.

4.5. Analysis of Diversity of Tentaclin Genes

Sequences of C-terminal IgG domains (last 100 aa residues) of tentaclins of the phages nd4, nd12, FP_Mushu, FP_Brigit, and FP_Toutatis were used to perform PSI-BLAST search using NCBI Genbank non-redundant protein sequences (nr) database. The number of target sequences was chosen to be 1000, and the expectation threshold was 0.05. Non-redundant RefSeq proteins (records starting with "WP_") were excluded from results due to the inability to reference the original nucleotide sequence for such entries. Three consecutive iterations of PSI-BLAST search were performed for each Cterm-IgG sequence. Results were downloaded as single file; phage-related records were extracted using home-written python scripts. Then, all the records were combined, duplicates were removed, and corresponding phage nucleotide sequences were downloaded. Finally, a set of 383 phage sequences was divided into two parts (due to limitations of online version of VIRIDIC), which were used for intergenomic similarity calculation using VIRIDIC (https://rhea.icbm.uni-oldenburg.de/viridic/) [33]. Finally these two parts were reorganized so that the largest groups were in the first heatmap, whereas the smaller groups and individual sequences were in the second heatmap.

4.6. Analysis of Diversity of VR sequences

BioEdit 7.2.5 [34] and AlignX (a tool from Vector NTI suite 8.0) were used for performing nucleotide and amino acid sequence alignment as well as for calculation of the sequence identity.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1. Figure S1: Alphafold2-generated models of nd4_tgt1, nd4_tgt2, nd12_tgt2, nd12_tgt2 and related proteins colored based on pLDDT value, indicating high overall confidence of the models. Figure S2: VIRIDIC heatmap indicating intergenomic similarity between tentaclin gene-containing phage-like sequences from second part of dataset (178 sequences of 373). Figure S3: Alphafold2 model of unusual 1180aa-long tentaclin-related protein from *Brevibacillus* sp., protein accession number NRS19465. Figure S4: Schematic view of DGR cassettes for phage sequences randomly selected from the five largest groups obtained using VIRIDIC analysis Figure S5: Surface representation of C-lec domains of nd4_tgt1 and nd4_tgt2 proteins and nd4_tgt1 tentaclin molecule. Figure S6: Models of the C-lec domains of the proteins nd4_tgt1, nd4_tgt2 and nd12_tgt1, showing the close location of cysteine residues in the beta hairpin to cysteine residues from the lectin core. Data S1: Annotation of nd4 genome. Data S2: Annotation of the nd12 genome.

Author Contributions: Conceptualization, I.K.B., and N.V.T.; sequencing – A.Y.T. and V.A.F., formal analysis, I.K.B., I.V.B. and E.V.Z.; investigation, I.K.B., A.Y.T. and V.A.F.; Software: I.K.B; data curation, I.K.B. and A.Y.T.; writing—original draft preparation, I.K.B. and N.V.T.; writing—review and editing, I.K.B. and N.V.T.; supervision, N.V.T.; project administration, N.V.T.; funding acquisition, N.V.T. (from Russian Science Foundation) and I.K.B. (from Ministry of Education and Science) All authors have read and agreed to the published version of the manuscript.

Funding: Virome sequencing, genome assembly, annotation, and analysis were funded by the Russian Science Foundation; Project No. 21-14-00360. 3D structure modeling was supported by the Ministry of Science and

Higher Education of the Russian Federation within the governmental order for Boreskov Institute of Catalysis SB RAS allocated to SRF "SKIF" (project FWUR-2023-0003).

Institutional Review Board Statement: This work was approved by the Local Ethics Committee of the Center for personalized medicine, Novosibirsk (protocol #2, 12.02.2019), where this sample was obtained. Written consent of the healthy volunteer was obtained according to guidelines of the Helsinki ethics committee.

Informed Consent Statement: Informed consent was obtained from all healthy volunteers involved in the study.

Data Availability Statement:

Raw NGS data containing nd4 and nd12 sequences is available at Genbank (Bioproject PRJNA1027629). The nd4 and nd12 sequences were deposited to GenBank, accession numbers: XXXXXXXX and XXXXXXXX.

Acknowledgments: The authors would like to thank Anton V. Chechushkov and Vera V. Morozova for valuable advice and suggestions. The authors would also thank High Performance Computing Center of Novosibirsk State University ("NUSC NSU") for providing resources to perform molecular dynamics simulation experiments. The authors would also thank Google Colab team and authors of Colabfold for providing free access to perform Alhafold2 modelling.

Conflicts of Interest: All co-authors have seen and agree with the contents of the manuscript and the order of authors, and there is no financial interest to report. All co-authors declare that they have no conflict of interest.

References

1. Wu, L.; Gingery, M.; Abebe, M.; Arambula, D.; Czornyj, E.; Handa, S.; Khan, H.; Liu, M.; Pohlschroder, M.; Shaw, K.L.; et al. Diversity-Generating Retroelements: Natural Variation, Classification and Evolution Inferred from a Large-Scale Genomic Survey. *Nucleic Acids Res* **2018**, *46*, 11–24, DOI: 10.1093/nar/gkx1150.
2. Guo, H.; Arambula, D.; Ghosh, P.; Miller, J.F. Diversity-Generating Retroelements in Phage and Bacterial Genomes. *Microbiol Spectr* **2014**, *2*, DOI: 10.1128/9781555819217.ch53.
3. Roux, S.; Paul, B.G.; Bagby, S.C.; Nayfach, S.; Allen, M.A.; Attwood, G.; Cavicchioli, R.; Chistoserdova, L.; Gruninger, R.J.; Hallam, S.J.; et al. Ecology and Molecular Targets of Hypermutation in the Global Microbiome. *Nature Communications 2021 12:1* **2021**, *12*, 1–12, DOI: 10.1038/s41467-021-23402-7.
4. Belalov, I.S.; Sokolov, A.A.; Letarov, A. V. Diversity-Generating Retroelements in Prokaryotic Immunity. *Int J Mol Sci* **2023**, *24*, doi:10.3390/IJMS24065614/S1.
5. Liu, M.; Deora, R.; Doulatov, S.R.; Gingery, M.; Eiserling, F.A.; Preston, A.; Maskell, D.J.; Simons, R.W.; Cotter, P.A.; Parkhill, J.; et al. Reverse Transcriptase-Mediated Tropism Switching in Bordetella Bacteriophage. *Science (1979)* **2002**, *295*, 2091–2094, DOI: 10.1126/science.1067467.
6. Liu, M.; Gingery, M.; Doulatov, S.R.; Liu, Y.; Hodes, A.; Baker, S.; Davis, P.; Simmonds, M.; Churcher, C.; Mungall, K.; et al. Genomic and Genetic Analysis of Bordetella Bacteriophages Encoding Reverse Transcriptase-Mediated Tropism-Switching Cassettes. *J Bacteriol* **2004**, *186*, 1503–1517, DOI: 10.1128/jb.186.5.1503-1517.2004.
7. Doulatov, S.; Hodes, A.; Dal, L.; Mandhana, N.; Liu, M.; Deora, R.; Simons, R.W.; Zimmerly, S.; Miller, J.F. Tropism Switching in Bordetella Bacteriophage Defines a Family of Diversity-Generating Retroelements. *Nature 2004 431:7007* **2004**, *431*, 476–481, DOI: 10.1038/nature02833.
8. Arambula, D.; Wong, W.; Medhekar, B.A.; Guo, H.; Gingery, M.; Czornyj, E.; Liu, M.; Dey, S.; Ghosh, P.; Miller, J.F. Surface Display of a Massively Variable Lipoprotein by a Legionella Diversity-Generating Retroelement. *Proc Natl Acad Sci U S A* **2013**, *110*, 8212–8217, DOI: 10.1073/pnas.1301366110.
9. Benler, S.; Cobián-Güemes, A.G.; McNair, K.; Hung, S.H.; Levi, K.; Edwards, R.; Rohwer, F. A Diversity-Generating Retroelement Encoded by a Globally Ubiquitous Bacteroides Phage 06 Biological Sciences 0605 Microbiology. *Microbiome* **2018**, *6*, 1–10, DOI: 10.1186/s40168-018-0573-6.
10. Minot, S.; Grunberg, S.; Wu, G.D.; Lewis, J.D.; Bushman, F.D. Hypervariable Loci in the Human Gut Virome. *Proc Natl Acad Sci U S A* **2012**, *109*, 3962–3966, DOI: 10.1073/pnas.1119061109.
11. Ye, Y. Identification of Diversity-Generating Retroelements in Human Microbiomes. *International Journal of Molecular Sciences 2014, Vol. 15, Pages 14234-14246* **2014**, *15*, 14234–14246, DOI: 10.3390/ijms150814234.
12. Morozova, V.; Fofanov, M.; Tikunova, N.; Babkin, I.; Morozov, V. V.; Tikunov, A. First CrAss-Like Phage Genome Encoding the Diversity-Generating Retroelement (DGR). *Viruses 2020, Vol. 12, Page 573* **2020**, *12*, 573, DOI: 10.3390/v12050573.
13. Yutin, N.; Benler, S.; Shmakov, S.A.; Wolf, Y.I.; Tolstoy, I.; Rayko, M.; Antipov, D.; Pevzner, P.A.; Koonin, E. V. Analysis of Metagenome-Assembled Viral Genomes from the Human Gut Reveals Diverse Putative CrAss-like Phages with Unique Genomic Features. *Nature Communications 2021 12:1* **2021**, *12*, 1–11, DOI: 10.1038/s41467-021-21350-w.

14. Alayyoubi, M.; Guo, H.; Dey, S.; Golnazarian, T.; Brooks, G.A.; Rong, A.; Miller, J.F.; Ghosh, P. Article Structure of the Essential Diversity-Generating Retroelement Protein BAvd and Its Functionally Important Interaction with Reverse Transcriptase. *Structure* **2013**, DOI: 10.1016/j.str.2012.11.016.
15. Dai, W.; Hodes, A.; Hui, W.H.; Gingery, M.; Miller, J.F.; Zhou, Z.H. Three-Dimensional Structure of Tropism-Switching Bordetella Bacteriophage. *Proc Natl Acad Sci USA* **2010**, *107*, 4347–4352, DOI: 10.1073/pnas.0915008107.
16. Miller, J.L.; Le Coq, J.; Hodes, A.; Barbalat, R.; Miller, J.F.; Ghosh, P. Selective Ligand Recognition by a Diversity-Generating Retroelement Variable Protein. *PLoS Biol* **2008**, *6*, e131, DOI: 10.1371/journal.pbio.0060131.
17. Sharifi, F.; Ye, Y. MyDGR: A Server for Identification and Characterization of Diversity-Generating Retroelements. *Nucleic Acids Res* **2019**, *47*, W289–W294, DOI: 10.1093/nar/gkz329.
18. Cornuault, J.K.; Petit, M.A.; Mariadassou, M.; Benevides, L.; Moncaut, E.; Langella, P.; Sokol, H.; De Paepe, M. Phages Infecting Faecalibacterium Prausnitzii Belong to Novel Viral Genera That Help to Decipher Intestinal Viromes. *Microbiome* **2018**, *6*, 65, DOI: 10.1186/s40168-018-0452-1.
19. Fokine, A.; Islam, M.Z.; Fang, Q.; Chen, Z.; Sun, L.; Rao, V.B. Structure and Function of Hoc—A Novel Environment Sensing Device Encoded by T4 and Other Bacteriophages. *Viruses* **2023**, *15*, 1517, DOI: 10.3390/v15071517.
20. Brown GD, Willment JA, Whitehead L. C-type lectins in immunity and homeostasis. *Nat Rev Immunol*. 2018;18(6):374-389. doi:10.1038/s41577-018-0004-8
21. Tskhovrebova, L.; Trinick, J. Flexibility and Extensibility in the Titin Molecule: Analysis of Electron Microscope Data. *J Mol Biol* **2001**, *310*, 755–771, DOI: 10.1006/jmbi.2001.4700.
22. Weikum J, Kulakova A, Tesei G, et al. The extracellular juncture domains in the intimin passenger adopt a constitutively extended conformation inducing restraints to its sphere of action. Sci Rep. 2020;10(1):21249. Published 2020 Dec 4. doi:10.1038/s41598-020-77706-7
23. McMahon, S.A.; Miller, J.L.; Lawton, J.A.; Kerkow, D.E.; Hodes, A.; Marti-Renom, M.A.; Doulatov, S.; Narayanan, E.; Sali, A.; Miller, J.F.; et al. The C-Type Lectin Fold as an Evolutionary Solution for Massive Sequence Variation. *Nature Structural & Molecular Biology 2005 12:10* **2005**, *12*, 886–892, DOI: 10.1038/nsmb992.
24. Zelensky, A.N.; Gready, J.E. The C-Type Lectin-like Domain Superfamily. *FEBS J* **2005**, *272*, 6179–6217, DOI: 10.1111/j.1742-4658.2005.05031.x.
25. Niemann HH, Schubert WD, Heinz DW. Adhesins and invasins of pathogenic bacteria: a structural view. *Microbes Infect*. 2004;6(1):101-112. doi:10.1016/j.micinf.2003.11.001
26. Morozova, V.; Kozlova, Y.; Shedko, E.; Kurilshikov, A.; Babkin, I.; Tupikin, A.; Yunusova, A.; Chernonosov, A.; Baykov, I.; Kondratov, I.; Kabilov, M.; Ryabchikova, E.; Vlassov, V.; Tikunova, N. Lytic bacteriophage PM16 specific for Proteus mirabilis: a novel member of the genus Phikmvvirus. *Arch. Virol.* **2016**, *161*(9), 2457-2472. https://doi.org/10.1007/s00705-016-2944-2
27. Brettin T, Davis JJ, Disz T, et al. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci Rep*. 2015;5:8365. Published 2015 Feb 10. doi:10.1038/srep08365
28. Nishimura, Y.; Yoshida, T.; Kuronishi, M.; Uehara, H.; Ogata, H.; Goto, S. Viptree: the viral proteomic tree server. *Bioinformatics* **2017**, *33*, 2379–2380. https://doi.org/10.1093/bioinformatics/btx157
29. Holm L, Laiho A, Törönen P, Salgado M. DALI shines a light on remote homologs: One hundred discoveries. *Protein Sci*. 2023;32(1):e4519. doi:10.1002/pro.4519
30. Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: making protein folding accessible to all. *Nat. Methods* **2022**, *19(6)*, 679–682. https://doi.org/10.1038/s41592-022-01488-1
31. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem*. **2004**, *25(13)*, 1605–1612. https://doi.org/10.1002/jcc.20084
32. Pronk S, Páll S, Schulz R, et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*. 2013;29(7):845-854. doi:10.1093/bioinformatics/btt055
33. Moraru, C.; Varsani, A.; Kropinski, A.M. VIRIDIC – a novel tool to calculate the intergenomic similarities of prokaryote-infecting viruses. *Viruses* **2020**, *12(11)*, 1268. https://doi.org/10.3390/v12111268
34. Hall, T.A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids Symp. Ser.* **1999**, *41*, 95-98.