viruses

MDPI

Article

First crAss-Like Phage Genome Encoding the Diversity-Generating Retroelement (DGR)

Vera Morozova *,†, Mikhail Fofanov †, Nina Tikunova *, Igor Babkin, Vitaliy V. Morozov and Artem Tikunov

Institute of Chemical Biology and Fundamental Medicine SB RAS, 630090 Novosibirsk, Russia; mvfofanov@mail.ru (M.F.); i_babkin@mail.ru (I.B.); doctor.morozov@mail.ru (V.V.M.); arttik@ngs.ru (A.T.)

* Correspondence: morozova@niboch.nsc.ru (V.M.); tikunova@niboch.nsc.ru (N.T.)

† These authors contributed equally to this work.

Received: 10 April 2020; Accepted: 19 May 2020; Published: 22 May 2020

Abstract: A new crAss-like genome encoding diversity-generating retroelement (DGR) was found in the fecal virome of a healthy volunteer. The genome of the phage referred to as the crAssphage LMMB, belonged to the candidate *genus I* of the AlphacrAssvirinae subfamily. The DGR-cassette of the crAssphage LMMB contained all the essential elements: the gene encoding reverse transcriptase (RT), the target gene (TG) encoding the tail-collar fiber protein, and variable and template repeats (VR and TR) with IMH (initiation of mutagenic homing) and IMH* sequences at the 3′-end of the VR and TR, respectively. Architecture of the DGR-cassette was TG-VR(IMH)-TR(IMH*)-RT and an accessory variable determinant (avd) was absent from the cassette. Analysis of 91 genomes and genome fragments from *genus I* of the AlphacrAssvirinae showed that 15 (16%) of the genomes had DGRs with the same architecture as the crAssphage LMMB, while 66 of the genomes contained incomplete DGR-cassettes or some elements of the DGR.

Keywords: crAssphage; genome; diversity-generating retroelement (DGR); variable repeat; template repeat; initiation of mutagenic homing

1. Introduction

CrAss-like phages have been discovered by computational analysis of human fecal metagenome data and their genomes have been shown to be the most abundant group of sequences (up to 90%) in the human gut virome [1]. Subsequently, they were classified as the crAss-like group [2], and then it was proposed to allocate them into a separate family, consisting of four subfamilies and ten genera, namely AlphacrAssvirinae (genera *I, III, IV* and *IX*), BetacrAssvirinae (*VI*), GammacrAssvirinae (*II, V*) and DeltacrAssvirinae (*VII, VIII, X*) [3]. CrAss-like phages were predicted to infect bacteria of the phylum Bacteroidetes, which are the most widely represented bacteria in the intestinal tract of humans [1]. Bacteroidetes are difficult to cultivate and thus limited data are available about these viruses; however, the CrAss001 phage has been isolated and confirmed to have podoviral morphology and infect *Bacteroides intestinalis* [4].

Intensive studies on the biology, taxonomy, and role of crAss-like phages have shown that they have high levels of genetic diversity. Members of this group of phages have been found in a range of environments, including the human gut and feces, termite gut, terrestrial/groundwater environments, soda lakes (hypersaline brine), marine sediments, and plant roots [2,3,5–7]. CrAss-like phages likely infect a variety of bacterial hosts; however, the mechanisms generating this variability are unknown.

One of the remarkable systems responsible for the variability of prokaryotic microorganisms are the diversity generating retroelements (DGRs), which use reverse transcription to introduce

huge numbers of nucleotide substitutions in specific target genes [8,9]. The DGR has been initially discovered in the genome of the temperate *Bordetella* phage BPP-1 [10], and found to provide changes in the host-recognizing structures of *Bordetella* phages, hence enabling phage adaptation to dynamic changes on the surface of the *Bordetella* host [10,11]. Subsequent genetic and metagenomics studies have shown that DGRs contain several essential elements. A crucial element of each DGR is the gene encoding reverse transcriptase (RT). This enzyme plays an important role in exchanging between two repeats, which have similar nucleotide sequences, during a process called mutagenic retrohoming [8,9,12]. One repeat is a template repeat (TR) while the other is a variable repeat (VR), the latter of which is often located at the 3′-end of the target gene. During mutagenic retrohoming, an RNA-transcript from the TR is reverse transcribed by the RT and almost all the adenines in the cDNA sequence could be subjected to A-to-N mutagenesis. This leads to changes in the VR sequence and, hence, in the corresponding *C*-terminal amino acid (aa) positions in the protein encoded by the target gene [9]. The initiation of the mutagenic homing (IMH) sequence is located at the 3′-end of the VR, while a non-identical IMH* repeat is usually found at the 3′-end of the TR. Additionally, the accessory variability determinant (avd) gene or, sometimes, a gene encoding a component of the bacterial efflux pump may be part of the DGR-cassette. The target gene, RT, VR, TR, IMH, and IMH* are essential elements of the DGR-cassette, while approximately one quarter of DGRs have no homologs to the accessory gene [8,9]. DGRs have been classified based on their architectural variations and the phylogeny of their specific elements [9].

DGRs have been found in the genomes of a wide range of microorganisms belonging to the Bacteroidetes, Cyanobacteria, Firmicutes, Proteobacteria, and Archaea phyla [13–18]. Most DGRs are associated with bacterial chromosomes, including the genomes of putative prophages [8,9,14]. Only a few DGRs containing the RT gene have been found in the genomes of free phages including the first finding in the *Bordetella* phage BPP-1 [10,12,19,20]. Here we describe the first complete genome of a crAss-like phage containing the DGR-cassette with the RT and other essential elements, and also analyze the occurrence of such cassettes in other relative crAss-like phages.

2. Materials and Methods

2.1. Viral DNA Isolation and Sequencing

A fecal sample (0.3 g) from a healthy volunteer was re-suspended in 1.2 mL of sterile phosphate-buffered saline (PBS, pH 7.5) and clarified by centrifugation 4 times at 20,000× *g* for 5 min at 4 °C. After every centrifugation, supernatant was transferred to a new sterile centrifugation tube. The final supernatant was treated with 5 U of DNase I (Thermo Fisher Scientific, Waltham, MA, USA) followed by treatment with 100 μg/mL of Proteinase K (Thermo Fisher Scientific) supplemented with EDTA and SDS to final concentrations of 20 mM and 0.5%, respectively. The mixture was incubated for 3 h at 55 °C. DNA was extracted from the whole volume of supernatant using a phenol-chloroform method with subsequent ethanol precipitation. Purified DNA was diluted in 50 μL of 0.1 × TE-buffer. The DNA was further used to construct a virome shotgun library using the NEB Next Ultra DNA library prep kit (New England Biolabs, Ipswich, MA, USA). Sequencing was carried out using a MiSeq Benchtop Sequencer (Illumina Inc., Foster City, CA, USA) in the SB RAS Genomics Core Facility, ICBFM SB RAS, Novosibirsk, Russia, and a MiSeq Reagent Kit 2 × 250 v.2 (Illumina Inc.). The obtained sequences were assembled *de novo* using the CLC Genomics Workbench software v.6.0. This study was approved by a local Ethics committee of the Center for Personalized Medicine in Novosibirsk; protocol #2, 12.02.2019.

2.2. Genome Analysis

The most abundant contig was compared with sequences deposited in the GenBank database using the BLASTN algorithm. The putative open reading frames (ORFs) were revealed using Vector NTI [21] and their functions were predicted by the BLASTX algorithm according to their similarity with annotated protein sequences from the GenBank database. CrAssphage genomes were downloaded from GenBank using the crAssphage LMMB nucleotide sequences of genes encoding

the major capsid protein and RT as a query in a BLASTN search against nucleotide collection (nr/nt) and whole-genome shotgun contig (WGS) databases. The DGR sequences were identified manually using multiple alignments of putative DGR-containing genome sequences and, additionally, were verified using myDGR software [22]. MAFFT software (https://mafft.cbrc.jp/alignment/server) was used to align genome sequences and make a dot-plot analysis of the genomes. Alignment of RT sequences was performed using algorithm M-Coffee from the T-Coffee software package [23], and MEGA X software [24] was used for phylogenetic analysis of aligned sequences. A phylogenetic analysis of crAss-like phages was performed using the Viral Proteomic Tree server (ViPTree) [25]. Sequence logos for VRs were generated using WebLogo [26]. The CGView server was used for comparative analysis of the genome of crAssphage LMMB and the genomes of related phages [27]. The investigated nucleotide sequence was deposited to the GenBank database under accession number [MT006214].

3. Results

3.1. Analysis of the crAssphage LMMB Genome

Following the *de novo* assembly of the fecal virome of a healthy volunteer, several contigs with a length >40,000 bp were identified. One of the contigs had a length of 98,458 bp with an average coverage of 185, and this contig was found to have high similarity (identity level ~97% with the query cover of 95%) to the genome of the prototype p-crAssphage [NC_024711.1], which has been previously analyzed in detail [1,2]. The sequence similarity analysis performed using the CGView server revealed that the studied genome of a phage, referred to as the crAssphage LMMB, showed gene synteny typical of other crAss-like genomes (Figure 1) from the candidate *genus I* of the previously suggested subfamily AlphacrAssvirinae, which contains p-crAssphage [3].

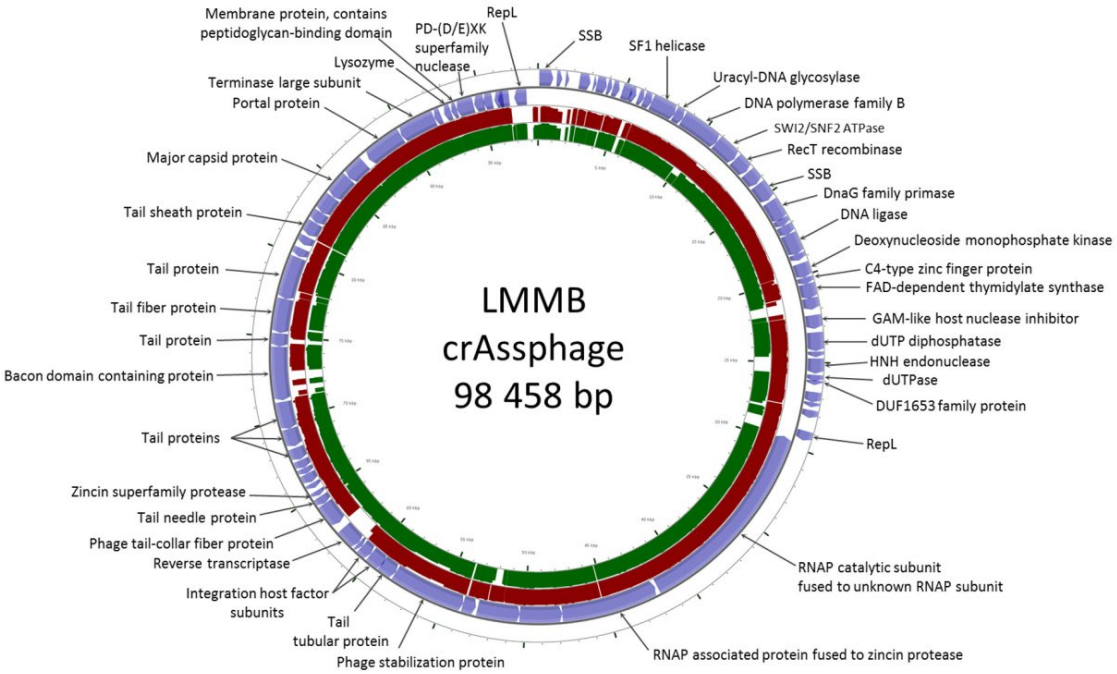

Figure 1. CrAssphage LMMB genome map visualized using the CGView server. Open reading frames (ORFs) are denoted with blue in the outer circle. The TBLASTX algorithm was used for sequence similarity comparison of the genomes of crAssphage LMMB, p-crAssphage [NC_024711.1] (similar genomic regions are marked with red), and NODE_105 [ODVQ01000105.1] (similar genomic regions are marked with dark green). Genomes of p-crAssphage and phage NODE_105 belong to *genus I* of the proposed subfamily AlphacrAssvirinae.

The genome of the crAssphage LMMB comprised 81 putative ORFs with 38 ORFs located on the forward strand and 43 ORFs found on the reverse strand (Figure 1). Previously it was reported that

all of the crAss-like phages of candidate genera *I*, *II*, and *IV* do not contain tRNA genes [3]. Among the annotated ORFs, no tRNA genes were found in the genome of the crAssphage LMMB.

The taxonomy of the crAssphage LMMB was confirmed using phylogenetic analysis of the putative proteomes of related crAss-like phages (Figure 2), and the phage was identified as a member of *genus I* of the subfamily AlphacrAssvirinae.

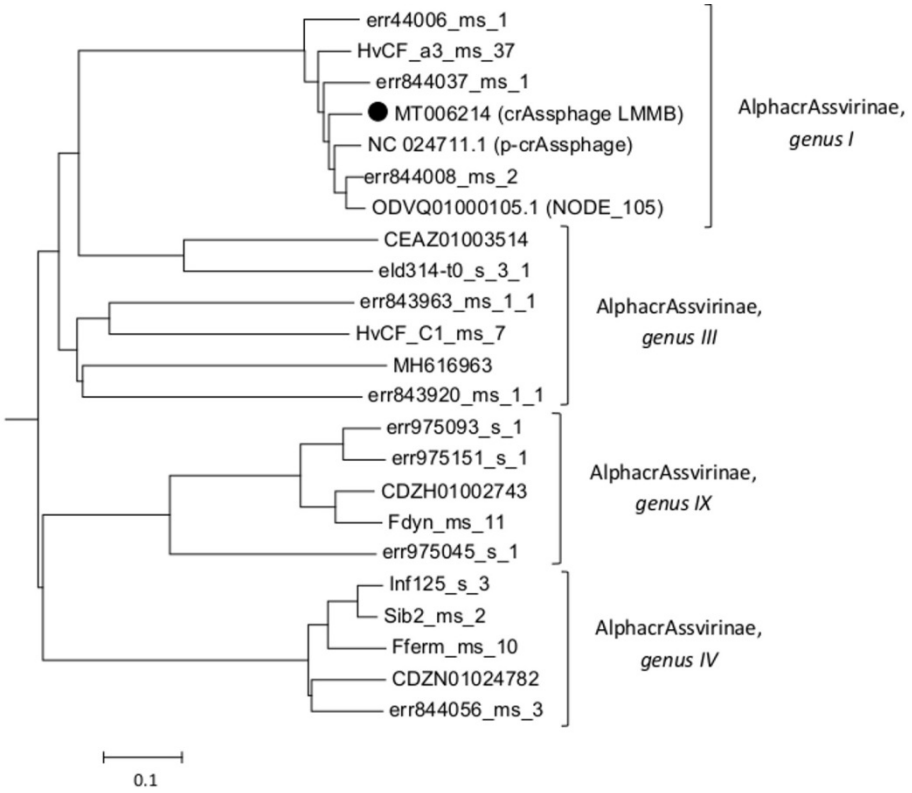

Figure 2. A phylogenetic analysis of the crAssphage LMMB and a number of crAss-like phages of the candidate subfamily AlphacrAssvirinae was performed using the Viral Proteomic Tree server. The investigated sequence is marked with a black circle. Genomes IDs are given. Genomes ODVQ01000105.1, CEAZ01003514, CDZH01002743, CDZN01024782, MH616963, and NC_024711 were downloaded from the GenBank databases, other sequences were extracted from the Supplementary data [3].

3.2. Identification and Characterization of the DGR in the Genome of crAssphage LMMB

A peculiarity of the crAssphage LMMB genome, which distinguished it from the reference genome of p-crAssphage [NC_024711.1], was the presence of the gene encoding the putative RT (protein ID QIN93296). This gene (ORF49) was located at the 3′-end of a cluster of structural genes, downstream the gene (ORF50), which encodes the tail-collar fiber protein (T-CFP, protein ID QIN93297) (Figures 1 and 3). Detailed characterization of the genome regions upstream and downstream the RT showed that a cassette of essential DGR elements was located upstream the RT (Figure 3). This DGR-cassette contained the target gene encoding T-CFP, the VR at the 3′-end of the target gene, TR, and RT. Both the VR and TR had identical IMH and IMH* at their 3′-ends, respectively (Figure 3).

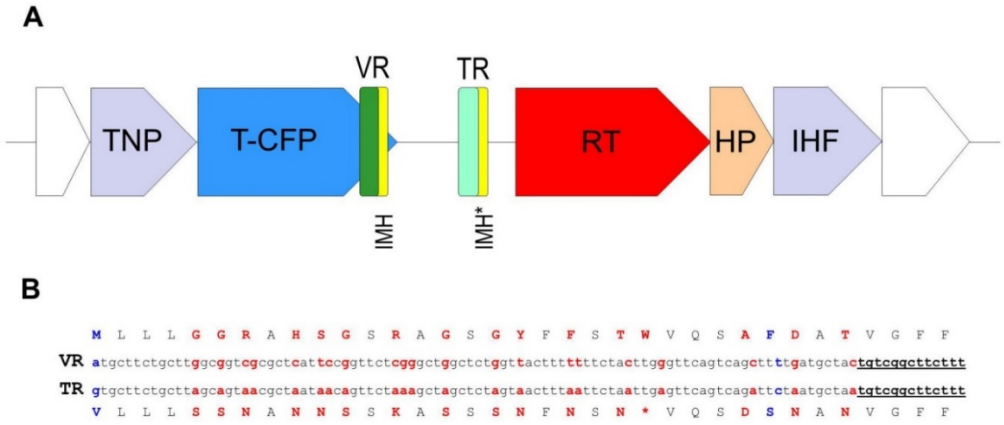

Figure 3. Schematic structure of the crAssphage LMMB diversity generating retroelement (DGR). (**A**) DGR in the genome of the crAssphage LMMB was verified using myDGR software (https://omics.informatics.indiana.edu/myDGR). The target gene (T-CFP) is marked with blue, the gene encoding reverse transcriptase (RT) is marked with red, the variable repeat (VR) is marked with green, the target repeat (TR) is marked with light green, both IMH and IMH* are marked with yellow. The flanking genes, encoding a tail-needle protein (TNP) and integration host factor (IHF) are marked with purple, the hypothetical protein (HP) is marked with orange. (**B**) Comparison of the VR and TR of the crAssphage LMMB shown for the DNA and aa sequences (The TR is not translated in vivo.). Variable positions in the VR and the corresponding adenine residues in the TR are marked with red letters. Non A-to-N mutations are indicated with blue letters. The IMH and IMH* sequences are underlined.

To clarify the type and taxonomy of the identified DGR, phylogenetic analysis of the RT from the crAssphage LMMB versus RT sequences extracted from the study of Wu et al. [9] was performed, and this RT was subsequently classified as a member of the lineage 3 (Figure 4). The major architecture of the core elements in DGRs with RT from lineage 3 is known to be "target gene-TR-RT" [9], which corresponded to the DGR-cassette found in the genome of the crAssphage LMMB. No avd gene sequence was revealed in the crAssphage LMMB DGR, which is similar to other DGRs from lineage 3. It has been shown that all the revealed VRs, which most directly correlated with RTs of lineage 3, corresponded to C-type lectin folds of major class 3 (CLec3) [9].

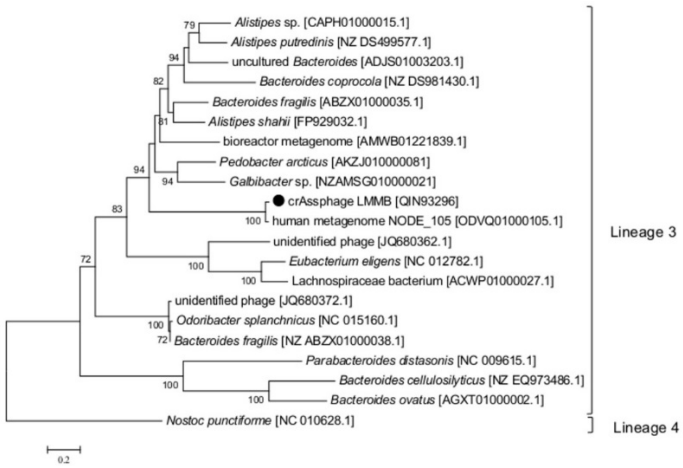

Figure 4. Phylogenetic analysis of the reverse transcriptase (RT) of the crAssphage LMMB and a number of similar RTs. The investigated sequence is marked with a black circle. The IDs of the genomes, which were used to extract gene sequences encoding RTs and translate them into aa sequences, are indicated in square brackets. Nucleotide sequences were extracted from the Supplementary data [9]. Phylogenetic analysis was performed using the maximum-likelihood method. The RT of *Nostoc punctiforme* (the Cyanobacteria phylum) was used as an out-group. Bootstrap values >70% are given at nodes.

In the crAssphage LMMB, the target protein (T-CFP) consists of 486 aa, 35 of which can be defined as the VR. There are 35 codons in the TR (Figure 3), containing 28 adenines that can theoretically generate ~10^17 (428) DNA sequences in the VR that correspond to 10^16 aa sequences in the target protein. When the VR and TR were compared, 21 adenines in the structure of the VR were found to undergo A-to-N mutagenesis. Additionally, two non A-to-N mutations were detected in the VR (Figure 3).

3.3. Comparative Analysis of Putative DGRs in the Genomes of crAss-Like Phages from Genus I of the AlphacrAssvirinae

To investigate the prevalence of DGRs in crAss-like phages from the candidate *genus I* of the subfamily AlphacrAssvirinae, a total of 90 genomes were selected. The selected genomes included 63 genomes of *genus I* that were downloaded from Supplementary data [3], and 27 crAss-like genomes that were extracted from the GenBank database using the sequences of genes encoding major capsid protein and RT of the crAssphage LMMB as queries in a BLASTN search against the human WGS database. Dot-plot analysis of the genomes and phylogeny of the major capsid proteins confirmed that they belonged to the candidate *genus I* of the subfamily AlphacrAssvirinae (Figure S1).

To identify the presence of putative DGRs, 90 genomes were screened for the presence of the gene encoding T-CFP using BLASTN; however, it was found that only 58 genomes contained the gene and 22 genomes contained its fragment. Then, the ORF encoding the tail needle protein (protein ID QIN93298) was found in 88 screened genomes (two genomes did not contain the ORF), and genomic regions of approximately 7500 bp adjacent to this ORF were then compared. The analysis showed substantial variability in these investigated regions (Figure 5, Table S1, Data S1). DGRs containing all the essential elements of the retrohoming system (target gene with VR, TR, and RT) were found in 16% (15/91, including crAssphage LMMB) of the examined genomes (Figure 5, Table S1). Architectures of these DGRs were identical and corresponded to the DGR-cassette in the genome of the crAssphage LMMB (Figure 5). The presence of core elements of the DGR in these genomes was confirmed using myDGR software. Ten (9%) investigated genomes had no elements of a DGR system in the genome region (Figure 5). The remaining 66 (73%) genomes demonstrated incomplete DGR-cassettes possessing truncated genes encoding T-CFP and RT with either the presence or absence of VR and TR (Figure 5, Table S1). Eighty of the analyzed genomes contained VRs; however, only 30 of these genomes contained TRs. Despite the substantial variability of the genomic region containing DGR elements, ORFs encoding the tail needle protein, hypothetical protein, and integration host factor subunit were conserved between the investigated genomes, if the ORFs were present (Figure 5).

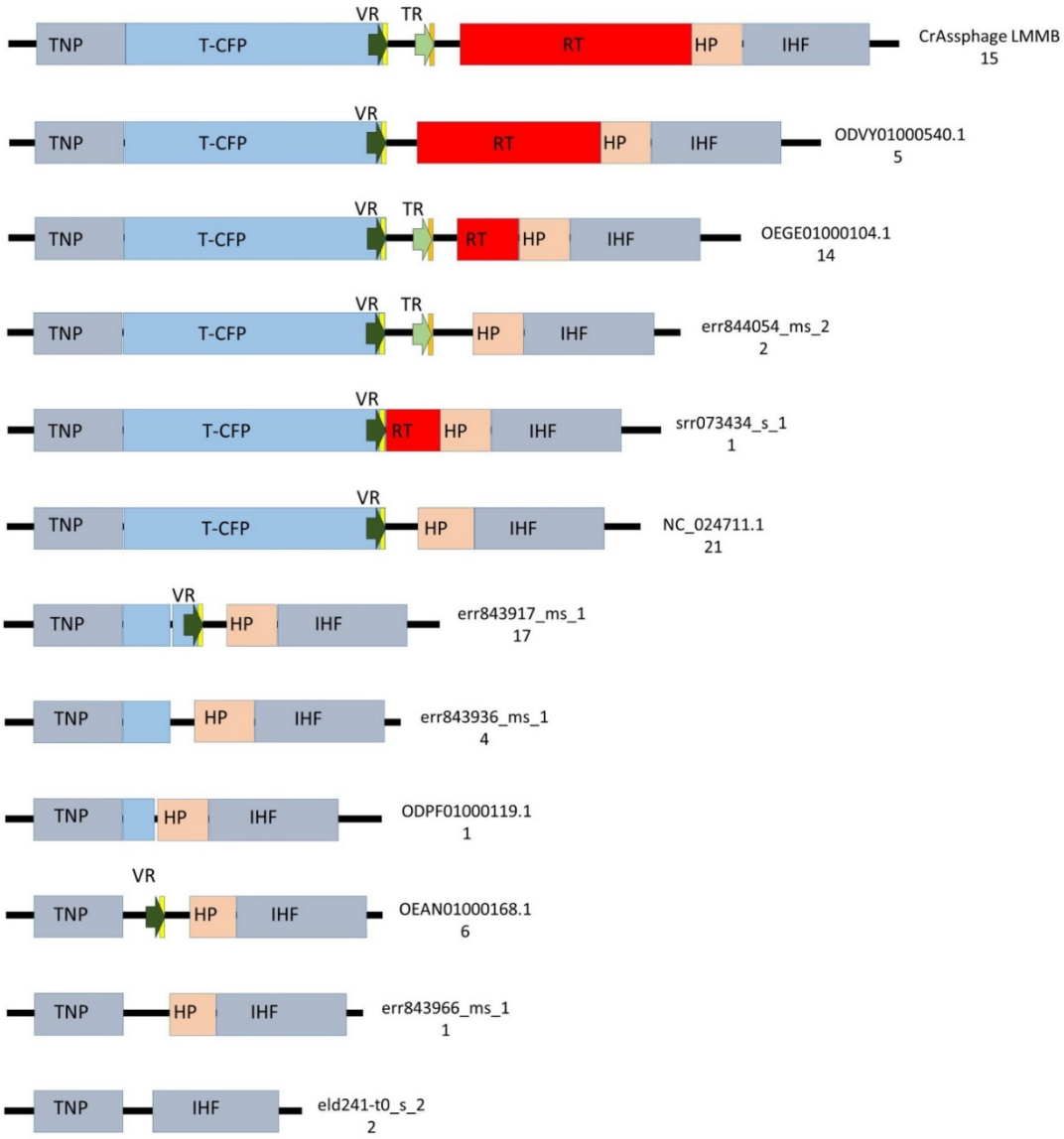

Figure 5. Schematic structure of the regions between the genes encoding tail needle protein (TNP) and integration host factor subunit (IHF) in 89 genomes of crAss-like phages from the candidate *genus I* of subfamily AlphacrAssvirinae. T-CFP — tail-collar fiber protein (marked with blue), RT—the gene encoding reverse transcriptase (marked with red), VR—variable repeat (marked with a dark green arrow), TR—template repeat (marked with a light green arrow), IMH—initiation of mutagenic homing sequence at the 3′-end of the VR (marked with yellow), IMH*— a non-identical repeat of IMH at the 3′-end of TR (marked with orange), HP—the gene encoding the hypothetical protein (marked with beige), flanking genes IHF and TNP (marked with grey). ID number of the reference genome and a number of found genomes for each architecture are shown on the right side.

3.4. Analysis of VRs

Complete target genes encoding T-CFP were obtained from a number of genomes, including the crAssphage LMMB and reference p-crAssphage (NC_024711.1) (Table S1), and the corresponding aa sequences were aligned. Then, similar VRs were identified in 57 complete T-CFPs (Figure 5, Table S1). Additionally, the nucleotide sequences of TRs were extracted from 30 genomes and the appropriate aa sequences were aligned. This analysis revealed that the flanking regions of TRs aligned well with the same regions of VRs (Figure S2).

The consensus VR structure of crAssphages from *genus I* was calculated using WebLogo software and compared with the consensus structures of VRs from DGRs with RT of lineage 3 and VRs from CLec1, CLec2, and CLec3 classes (Figure 6). The calculated consensus VR structure was

more conservative than the other structures, as it represented only sequences from *genus I* of the subfamily AlphacrAssvirinae, while the other consensus sequences comprised sequences from different members of large taxonomic groups such as Actinobacteria, Bacteroidetes, Firmicutes, and various prophages. It was revealed that the consensus VR structure of crAss-like phages of *genus I* had a similar motif to VRs from DGRs of lineage 3 and CLec3 VRs. The similarity was higher when comparing the consensus VR to VRs from DGRs of lineage 3; however, this similarity was still lower than the similarity between VRs from DGRs of lineage 3 and VRs of the CLec3 class (Figure 6).

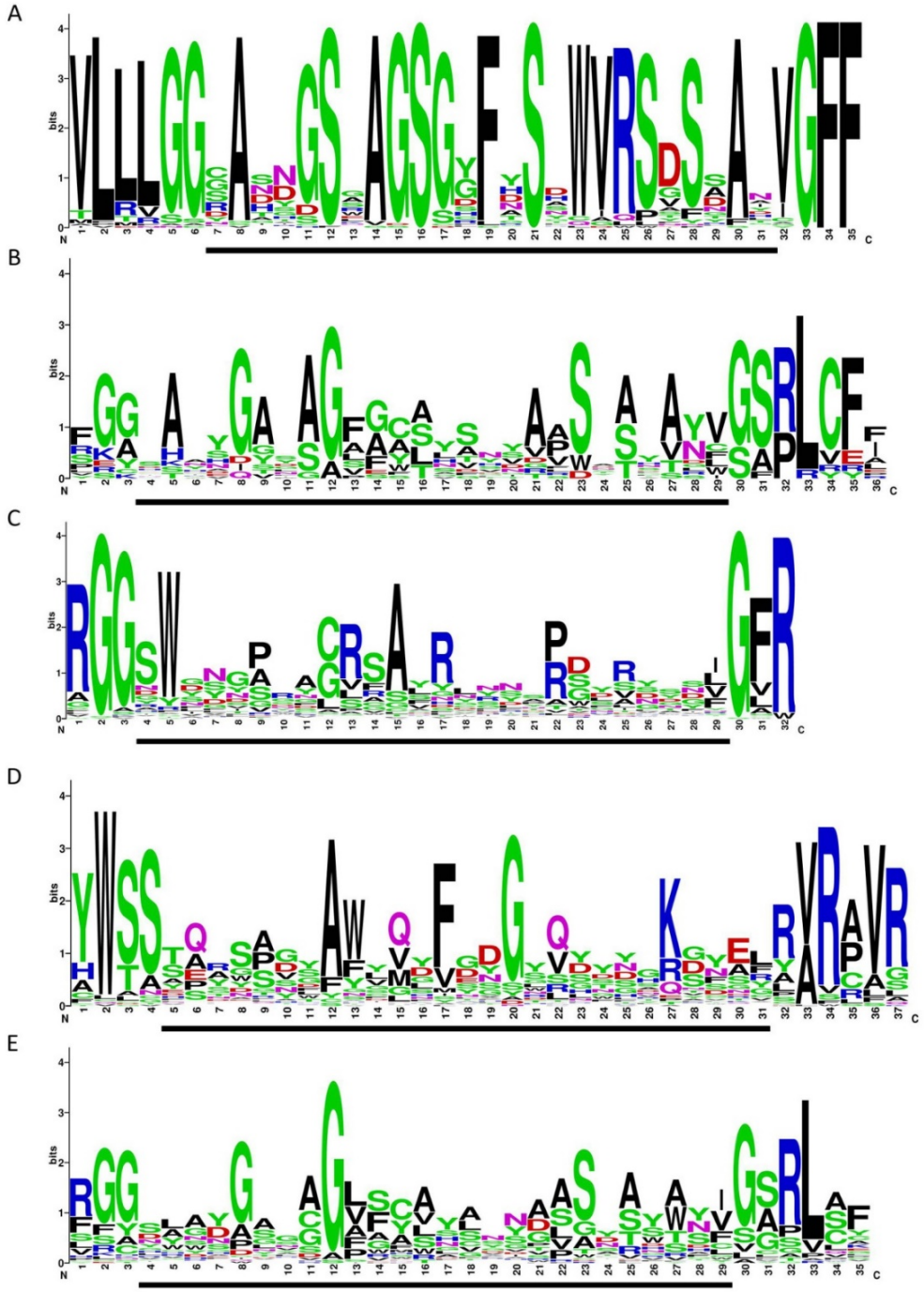

Figure 6. Variable repeat (VR) sequences shown in WebLogo format. (**A**) crAssphages from the candidate *genus I* of the subfamily AlphacrAssvirinae (extracted from 77 genomes). (**B**) VRs from diversity-generating retroelements (DGRs) with the reverse transcriptase (RT) of lineage 3. (**C**) CLec1 VRs. (**D**) CLec2 VRs (**E**) CLec3 VRs. Data used in the B–E were extracted from the Supplementary Material of Wu et al. [9] The regions involved in mutagenic retrohoming are indicated by a black bar under each WebLogo image.

4. Discussion

In this study, we describe the first finding of a DGR-cassette in the genome of the crAss-like phage. Gene synteny, phylogeny of the major capsid protein, and the absence of tRNA genes reliably confirmed that the studied crAssphage LMMB is a member of the candidate *genus I* of the subfamily AlphacrAssvirinae. The DGR-cassette containing all the essential elements (target gene-VR(IHF)-TR(IHF*)-RT) was found at the 3′-end of a cluster of structural genes, between the ORF51, which encodes the tail needle protein, and ORF48, which encodes the hypothetical protein. Based upon the DGR architecture, the fact that its RT belongs to lineage 3, and the absence of the avd gene, it was shown that the DGR of the crAssphage LMMB was a member of lineage 3.

Importantly, the genome of the crAssphage LMMB is not a unique crAss-like genome containing the DGR. Probably, this genetic element has not been identified in other crAss-like phages as the genome of the prototype p-crAssphage [1], which is usually used for annotation of genomes of the phages, does not contain the DGR. In fact, 14 other genomes from *genus I* of the subfamily AlphacrAssvirinae had orthologous DGRs with all essential elements. Usually, VRs from this type of DGR correspond to CLec3 [9]; however, the VRs in studied DGRs had some peculiarities. Although the consensus structure of VRs generated from the genomes of *genus I* showed a consensus motif with the consensus structures of VRs from DGRs with RTs of lineage 3 and CLec3 class VRs, this similarity was not high (Figure 6). The IMH sequences of the examined crAss-like phages also differed from the IMH sequences of DGRs of lineage 3 and CLec3 VRs. Through manual analysis, we showed that VRs of the phages have lost four aa (SRLC/A) in the IMH region compared to the VRs of DGR lineage 3 and CLec3. Previously, it has been reported that each VR within the class contains several conserved aa in their central regions, which are involved in the formation of the protein's structural scaffold, and at their N- and C-ends, and these conserved aa are not subjected to A-to-N mutagenesis [9,27]. When we examined the VR sequences in the crAssphage LMMB and related phages to them, each adenine was found to undergo A-to-N mutagenesis (Figure S2). This leads to a higher theoretical level of VR mutagenesis in the target gene of crAss-like phages from *genus I* than in BPP-1 (10^16 aa vs. 10^13 aa). This level of mutagenesis is one potential reason why crAss-like phages are known as the most abundant viruses in the human gut.

However, our analysis showed that most of the genomes of crAss-like phages of *genus I* did not contain the DGR-cassette with all the essential elements. This fact indicates that, on the one hand, the DGR is not vital for the existence of crAss-like phages, and damage to the DGR does not lead to the elimination of such phages co-existing with a certain host in the relatively constant conditions of a healthy human gut. On the other hand, the instability of this region may be due to imperfect recombination occurring during mutagenic retrohoming.

In conclusion, the first DGR-cassette was discovered in the genomes of the crAssphage LMMB and several relative crAss-like phages. The DGRs had all the essential elements for retrohoming. However, complete DGR-cassettes were not found in most of the examined genomes of crAss-like phages. Therefore, we cannot conclude whether the DGR-cassette is an evolutionary advantage of crAss-like phages, or these DGRs were simply taken from the host genome. The functionality of DGRs in the crAss-like phages can only be proven in further experimental studies.

Supplementary Materials: The following are available online at: www.mdpi.com/1999-4915/12/5/573/s1, Figure S1: Dot-plot analysis of the crAss-like genomes of *genus I* extracted from the GenBank database. Figure S2: TR and VR sequence alignments. The TR and VR sequences were extracted from the genomes containing both repeats. Table S1: List of genomes analyzed for DGR elements in this study. Data S1: DGR, TR, and VR sequences from the genomes of *genus 1* crAss-like phages.

Author Contributions: Conceptualization, V.M., M.F., and A.T.; data curation, A.T. and V.V.M.; formal analysis, V.M., M.F., and I.B.; funding acquisition, A.T.; investigation, A.T.; visualization, V.M., M.F., and N.T.; Writing—Original draft, V.M., M.F., A.T., and N.T.; Writing—Review and editing, N.T.; supervision, N.T. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Russian Science Foundation under grant #18-74-00082. VM and IB were funded by the Russian State-funded budget project of ICBFM SB RAS # AAAA-A17-117020210027-9.

Conflicts of Interest: The authors declare no conflicts of interest. The funder had no role in the design of the study; in the collection, analysis, and interpretation of data; in writing of the manuscript, or in decision to publish the results.

References

1. 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.; et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **2014**, *5*, 4498, doi:10.1038/ncomms5498.
2. Yutin, N.; Makarova, K.S.; Gussow, A.B.; Krupovic, M.; Segall, A.; Edwards, R.A.; Koonin, E.V. Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat. Microbiol.* **2018**, *3*, 38–46, doi:10.1038/s41564-017-0053-y.
3. Guerin, E.; Shkoporov, A.; Stockdale, S.R.; Clooney, A.G.; Ryan, F.J.; Sutton, T.D.S.; Draper, L.A.; Gonzalez-Tortuero, E.; Ross, R.P.; Hill, C. Biology and taxonomy of crAss-like Bacteriophages, the most abundant virus in the human gut. *Cell Host Microbe* **2018**, *24*, 653–664.e6, doi:10.1016/j.chom.2018.10.002.
4. Shkoporov, A.N.; Khokhlova, E.V.; Fitzgerald, C.B.; Stockdale, S.R.; Draper, L.A.; Ross, R.P.; Hill, C. ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects *Bacteroides intestinalis*. *Nat. Commun.* **2018**, *9*, 4781, doi:10.1038/s41467-018-07225-7.
5. Liang, Y.; Jin, X.; Huang, Y.; Chen, S. Development and application of a real-time polymerase chain reaction assay for detection of a novel gut bacteriophage (crAssphage). *J. Med. Virol.* **2018**, *90*, 464–468, doi:10.1002/jmv.24974.
6. Tikhe, C.V.; Husseneder, C. Metavirome sequencing of the termite gut reveals the presence of an unexplored bacteriophage community. *Front. Microbiol.* **2018**, *8*, 2548, doi:10.3389/fmicb.2017.02548.
7. Pramono, A.K.; Kuwahara, H.; Itoh, T.; Toyoda, A.; Yamada, A.; Hongoh, Y. Discovery and complete genome sequence of a bacteriophage from an obligate intracellular symbiont of a cellulolytic protist in the termite gut. *Microbes Environ.* **2017**, *32*, 112–117, doi:10.1264/jsme2.ME16175.
8. Guo, H.; Arambula, D.; Ghosh, P.; Miller, J.F. Diversity-generating retroelements in phage and bacterial genomes. *Microbiol. Spectrum* **2014**, *2*, MDNA3-0029-2014, doi:10.1128/microbiolspec.MDNA3-0029-2014.
9. Wu, L.; Gingery, M.; Abebe, M.; Arambula, D.; Czornyj, E.; Handa, S.; Khan, H.; Liu, M.; Pohlschroder, M.; Shaw, K.; 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.
10. 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* **2002**, *295*, 2091–2094, doi:10.1126/science.1067467.
11. Liu, M.; Gingery, M.; Doulatov, S.; 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.
12. Doulatov, S.; Hodes, A.; Dai, 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*, 476–481, doi:10.1038/nature02833.
13. Arambula, D.; Wong, W.; Medhekar, B.; Guo, H.; Gingery, M.; Czornyj, E.; Liu, M. Surface display of a massively variable lipoprotein by a Legionella diversity-generating retroelement. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 1–6, doi:10.1073/pnas.1301366110.
14. Benler, S.; Cobián-Güemes, A.; McNair, K.; Hung, S.; Levi, K.; Edwards, R.; Rohwer, F. A Diversity-generating retroelement encoded by a globally ubiquitous *Bacteroides* phage. *Microbiome* **2018**, *6*, 191, doi:10.1186/s40168-018-0573-6.
15. Coq, J.L.; Ghosh, P. Conservation of the C-Type lectin fold for massive sequence variation in a Treponema diversity-generating retroelement. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 14649–14653, doi:10.1073/pnas.1105613108.
16. Paul, B.G.; Bagby, S.C.; Czornyj, E.; Arambula, D.; Handa, S.; Sczyrba, A.; Ghosh, P.; Miller, J.F.; Valentine, D.L. Targeted diversity generation by intraterrestrial Archaea and Archaeal Viruses. *Nat. Commun.* **2015**, *6*, 1–8, doi:10.1038/ncomms7585.

17. Paul, B.G.; Burstein, D.; Castelle, C.J.; Handa, S.; Arambula, D.; Czornyj, E.; Thomas, B.C.; Ghosh, P.; Miller, J.F.; Banfield, J.F.; et al. Retroelement-guided protein diversification abounds in vast lineages of Bacteria and Archaea. *Nat. Microbiol.* **2017**, *2*, 17045, doi:10.1038/nmicrobiol.2017.45.
18. Ye, Y. Identification of diversity-generating retroelements in human microbiomes. *Int. J. Mol. Sci.* **2014**, *15*, 14234–14246, doi:10.3390/ijms150814234.
19. Medhekar, B.; Miller, J.F. Diversity-generating retroelements. *Curr. Opin. Microbiol.* **2007**, *10*, 388–395.
20. Minot, S.; Grunberg, S.; Wu, G.D.; Lewis, J.D.; Bushman, F.D. Hypervariable loci in the human gut virome. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 3962–3966, doi:10.1073/pnas.1119061109.
21. Lu, G.; Moriyama, E.N. Vector NTI, a balanced all-in-one sequence analysis suite. *Brief. Bioinform.* **2004**, *5*, 378–388.
22. 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.
23. Notredame, C.; Higgins, D.G.; Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **2000**, *302*, 205–217, doi:10.1006/jmbi.2000.4042.
24. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549, doi:10.1093/molbev/msy096.
25. Rohwer, F.; Edwards, R. The Phage Proteomic Tree: A genome-based taxonomy for phage. *J. Bacteriol.* **2002**, *184*, 4529–4535, doi:10.1128/jb.184.16.4529-4535.2002.
26. Crooks, G.E.; Hon, G.; Chandonia, J.M.; Brenner, S.E. WebLogo: A sequence logo generator. *Genome Res.* **2004**, *14*, 1188–1190, doi:10.1101/gr.849004.
27. Grant, J.R.; Stothard, P. The CGView Server: A comparative genomics tool for circular genomes. *Nucleic Acids Res.* **2008**, *36*, W181–W184, doi:10.1093/nar/gkn179.