Genome Replication of Bacterial and Archaeal Viruses

Česlovas Venclovas, Vilnius University, Vilnius, Lithuania

Glossary

Negative sense (−) strand A negative-sense DNA or RNA strand has a nucleotide sequence complementary to the messenger RNA and cannot be directly translated into protein.

Positive sense (+) strand A positive sense DNA or RNA strand has a nucleotide sequence, which is the same as that of the messenger RNA, and the RNA version of this sequence is directly translatable into protein.

Protein-primed DNA replication DNA replication whereby a DNA polymerase uses the 3′-OH group provided by the specialized protein as a primer to synthesize a new DNA strand.

RNA-primed DNA replication Conventional DNA replication used by all cellular organisms whereby a primase synthesizes a short RNA primer with a free 3′-OH group which is subsequently elongated by a DNA polymerase.

Rolling-circle DNA replication DNA replication whereby the replication initiation protein creates a nick in the circular double-stranded DNA and becomes covalently attached to the 5′ end of the nicked strand. The free 3′-OH group at the nick site is then used by the DNA polymerase to synthesize the new strand.

Genomes of Prokaryotic Viruses

At present, all identified archaeal viruses have either double-stranded (ds) or single-stranded (ss) DNA genomes. Although metagenomic analyzes suggested the existence of archaeal viruses with RNA genomes, this finding remains to be substantiated. Bacterial viruses, also refered to as bacteriophages or phages for short, have either DNA or RNA genomes, including circular ssDNA, circular or linear dsDNA, linear positive-sense (+)ssRNA or segmented dsRNA (**Table 1**). So far, no bacteriophages with negative sense (−)ssRNA genomes have been identified. Both archaeal and bacterial viruses with dsDNA genomes are most abundant, whereas those with ssDNA genomes represent a smaller group. The genome size of prokaryotic viruses is clearly linked with the structure of the genome. Thus, genomes approximately up to 25 kb in size are represented by both types of nucleic acids (DNA or RNA) and various topologies (circular, linear, segmented). All genomes larger than that are represented only by circular or linear dsDNA (**Fig. 1**). This size-dependent choice of the carrier of genomic information likely reflects physical constraints related to the genome stability, which is the highest for dsDNA.

Genome Replication of Prokaryotic dsDNA Viruses

Genomes of prokaryotic dsDNA viruses range from ∼5 to ∼500–600 kb, they can be either circular or linear. At present, dsDNA bacteriophages comprise the largest fraction of prokaryotic viruses that have their genomes sequenced. According to the data on complete viral genomes available at the NCBI database (see "Relevant Websites section") currently the smallest dsDNA genome (10079 nt) of a bacterial virus is represented by the tailless *Pseudoalteromonas* virus PM2. Phage PM2 is a representative of family *Corticoviridae* and the first lipid-containing phage to be isolated. The largest genome among officially classified bacterial dsDNA viruses belongs to the tailed *Bacillus* virus G (Order *Caudovirales*, Family *Myoviridae*) reaching nearly 500 kb (497513 nt). Recent metagenomics studies have identified phages with even larger dsDNA genomes that are in the 540–600 kb range. This points to the hidden diversity and abundance of very large phages and suggests that phages carrying even larger genomes might still be discovered. Notably, largest bacteriophage genomes are comparable in size with genomes of small bacteria capable of autonomous growth such as *Mycoplasma genitalium* with the 580 kb genome and exceed those of many symbiotic and parasitic bacteria. Archaeal dsDNA viruses range in size from 5.3 kb in the *Aeropyrum pernix* bacilliform virus 1 (APBV1), a member of the family *Clavaviridae*, to 144 kb for *Halogranum* tailed virus 1 (HGTV-1), a representative of the *Caudovirales*.

DNA replication has been studied in detail for only a handful of prokaryotic dsDNA viruses such as model bacteriophages phi29, T7 and T4. For most prokaryotic dsDNA viruses, however, there is little or no experimental data on their genome replication mechanisms. Therefore, whatever is inferred about DNA replication machineries for majority of prokaryotic dsDNA viruses typically comes from *in silico* analysis of viral genomes and corresponding proteomes. Based on both experimental characterization and computational analyzes of the repertoire of putative DNA replication genes several major types of dsDNA genome replication systems can be distinguished. (1) A common type is the one that mirrors major molecular functions associated with RNA-primed DNA replication in cellular organisms. However, this type of DNA replication machineries may differ greatly in completeness in various prokaryotic viruses. Some, such as phage T4, have essentially autonomous DNA replication machinery, which does not require assistance from the host. Others encode only few components of their own and rely on the host to provide the remaining ones (e.g., phage lambda). (2) Protein-primed DNA replication system specific to viruses and other mobile genetic elements (MGEs). (3) Rolling circle DNA replication (RCR) system, in which the key component is a viral replication initiation protein

Table 1 Classification of prokaryotic viruses

Genome type	*Archaeal viruses*	*Bacterial viruses*
dsDNA	Order: **Caudovirales** Families: *Myoviridae* (HGTV-1) *Siphoviridae* (HCTV-1) Order: ***Ligamenvirales*** Families: *Lipothrixviridae* (AFV1) *Rudiviridae* (SIRV2) Other families: *Ampullaviridae* (ABV) *Bicaudaviridae* (ATV) *Clavaviridae* (APBV1) *Fuselloviridae* (SSV1) *Globuloviridae* (PSV) *Guttaviridae* (APOV1) *Ovaliviridae* (SEV1) *Pleolipoviridae* (His2) *Portogloboviridae* (SPV1) *Sphaerolipoviridae* (SNJ1) *Tristromaviridae* (PFV1) *Turriviridae* (STIV)	Order: **Caudovirales** Families: *Ackermannviridae* (Limestone) *Herelleviridae* (SPO1) *Myoviridae* (T4, P2) *Podoviridae* (T7, phi29) *Siphoviridae* (λ) Other families: *Corticoviridae* (PM2) *Plasmaviridae* (L2) *Sphaerolipoviridae* (IN93) *Tectiviridae* (PRD1)
ssDNA	*Pleolipoviridae* (HRPV-1) *Spiraviridae* (ACV)	*Inoviridae* (M13) *Microviridae* (phiX174)
dsRNA	N/A	*Cystoviridae* (phi6)
ssRNA (+)	N/A	*Leviviridae* (MS2)
ssRNA (−)	N/A	N/A

Note: Viral families according to the ICTV Master Species List 34, 2018b with representatives indicated in parentheses. Families common to both archaeal and bacterial viruses are underlined. N/A, viruses with the indicated genome structure have not been identified (not available).

(Rep) whereas all or nearly all other proteins required for DNA replication are recruited from the host. (4) DNA replication systems that use other strategies to propagate viral genome (see below).

RNA-Primed DNA Replication

DNA replication systems in many prokaryotic viruses, particularly in bacteriophages, follow the same organizational principles as the corresponding systems of their cellular hosts. Archaea and bacteria use RNA-primed DNA replication systems that are made up of components which may not always be evolutionarily related, but perform the same function. At the most general level, cellular DNA replication machineries include the following key functional components: replicative helicases, primases, replicative DNA polymerases, DNA sliding clamps and clamp loaders, single-stranded DNA binding proteins, primer removal nucleases, DNA ligases and topoisomerases (**Fig. 2**). Typically, viral RNA-primed DNA replication machineries represent hybrid systems that are composed of proteins encoded by the viral genome and those provided by the host. According to the ratio of their own/host-provided DNA replication proteins viruses may differ considerably. Some viruses encode DNA replication systems that are essentially autonomous or nearly autonomous, whereas others encode only one or two DNA replication proteins and are largely dependent on the host to provide the rest. There is a continuum between these extreme cases. A global survey of viral-encoded DNA replication proteins has shown that both archaeal viruses and bacteriophages most frequently encode their own replicative helicases and not, as might be expected, DNA polymerases. Primases are also more frequent than DNA polymerases in both archaeal viruses and bacteriophages.

In addition, it was observed that certain functional categories of viral DNA replication proteins are tightly coupled. One such tightly coupled group comprises replicative helicase, primase and DNA polymerase. Viruses encoding either a DNA polymerase or primase nearly always encode a replicative helicase. The inverse co-occurrence is not nearly as stringent. Thus, in many cases, viruses, which encode replicative helicases, lack the genes for primases and DNA polymerases and likely rely on the corresponding proteins of the host. This asymmetry suggests that viral primases and DNA polymerases are constrained to work together with viral helicases. Such a coupling reflects intimate physical and functional interactions between these three proteins. Primase and helicase are often fused into a single polypeptide chain as exemplified by the primase-polymerase encoded by gene 4 of phage T7. There are

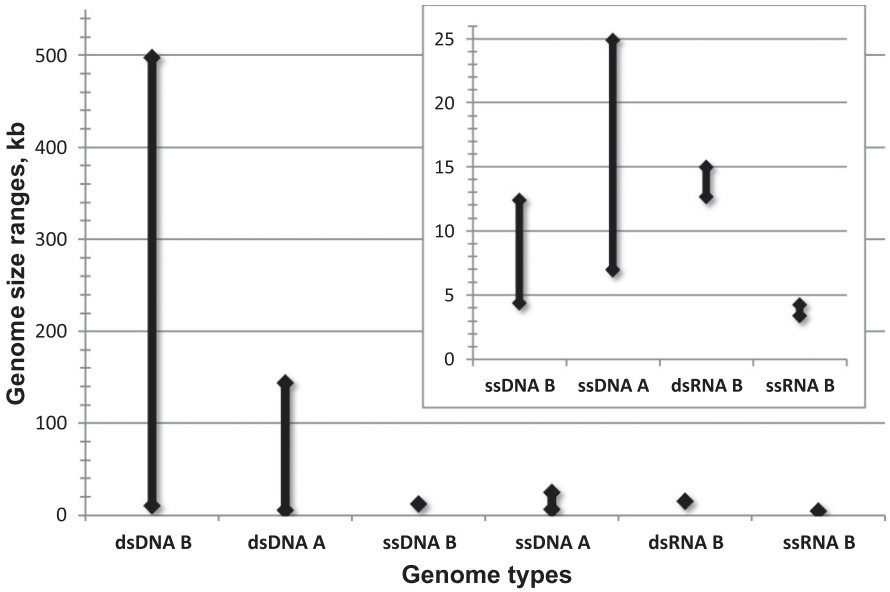

Fig. 1 Genome size ranges of prokaryotic viruses with different genome structure. B, bacteriophages, A, archaeal viruses. The inset shows a zoom-in on viruses with ssDNA and RNA genomes.

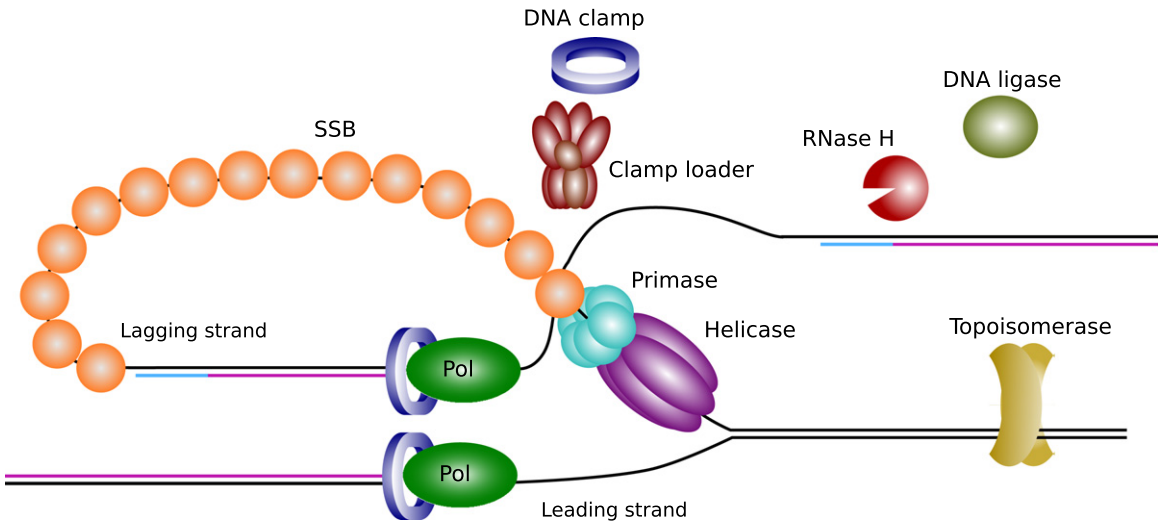

Fig. 2 Schematic representation of DNA replication fork based on bacteriophage T4 RNA-primed DNA replication system. The leading strand is synthesized continuously whereas the lagging strand is synthesized as Okazaki fragments. Newly synthesized strands are shown in magenta, RNA primers are represented as cyan-colored regions.

also cases where all three proteins (helicase, primase and polymerase) are fused together. One such example is the gp55 protein (Acc: YP_001828703) of *Lactococcus* phage 1706, and another one is the Pas55 protein (Acc: YP_024841) of *Actinoplanes* phage phiAsp2. Unfortunately, no experimental data are available on the molecular mechanism of these multifunctional proteins.

Another strong, unidirectional functional link was observed between viral DNA polymerases and their accessory factors, DNA sliding clamps and clamp loaders. It was found that all viruses which encode clamp loaders also encode DNA sliding clamps and DNA polymerases. This co-occurrence pattern is a direct reflection of steps involved in the assembly of a processive DNA replicase. The clamp loader loads the clamp, which in turn tethers DNA polymerase to the DNA enabling processive DNA synthesis. Again, the opposite link between the presence of a viral polymerase and accessory proteins is weak. The reason for this is that some viral DNA polymerases have high intrinsic processivity whereas others utilize host proteins to increase the processivity of DNA synthesis. A well-known example of host-supplied processivity factor is the case of *Escherichia coli* phage T7. By itself the T7 DNA polymerase is a low-processivity enzyme, however, upon binding the *E. coli* thioredoxin the processivity of the DNA polymerase increases by about 1000-fold.

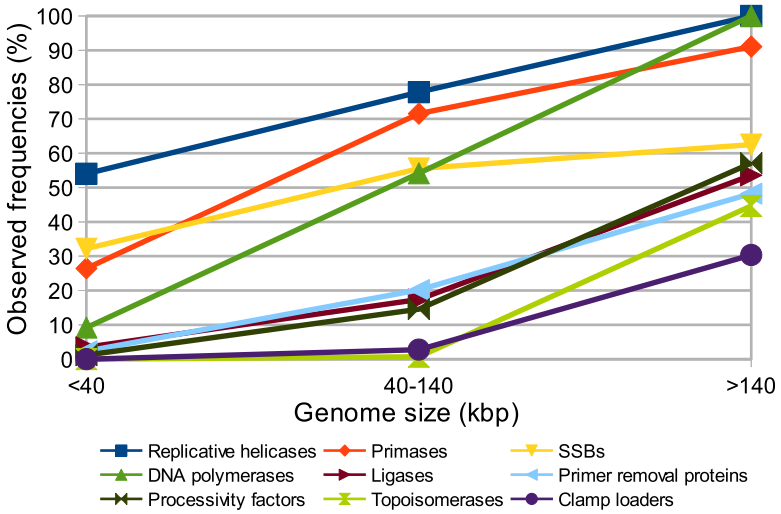

Fig. 3 Observed frequencies of DNA replication proteins encoded in genomes of prokaryotic dsDNA viruses. The figure is adapted from Kazlauskas, D., Krupovic, M., Venclovas, Č., 2016. The logic of DNA replication in double-stranded DNA viruses: Insights from global analysis of viral genomes. Nucleic Acids Research 44, 4551–4564.

Table 2 Prokaryotic dsDNA viruses having the most complete DNA replication machineries

Taxon/Virus	Genome size (kb)	DNA polymerase	DNA clamp	Clamp loader	Helicase	Primase	SSB	Primer removal protein	Ligase	Topoisomerase
Bacillus phage G	498	■	□	□	■	■	■	■	■	■
T4-like	135-359	■	■	■	■	■	■	■	▨	▨
Ralstonia phage RSL1	231	■	■	■	■	■	■	■	■	■
Sphingomonas phage PAU	219	■	■	■	■	■	■	■	■	■
Clostridium phage c-st	186	■	□	□	■	■	■	□	■	■
Microcystis aeruginosa phage Ma-LMM01	162	■	■	■	■	■	□	□	□	□
Cyanophage S-TIM5	161	■	□	□	■	■	□	■	□	■
Halovirus HVTV-1	102	■	■	■	■	■	□	■	□	□
Vibrio phage vB_VpaS_MAR10	79	■	■	■	■	■	■	■	■	□
Clostridium phage phiCTP1	59	■	■	■	■	■	■	□	■	□
T7-like	37-42	■	□	□	■	■	■	■	■	□

Note: DNA replication proteins are marked by their presence in the genome (present, black square; absent, white square; present in some members, black-and-white square). Viruses are arranged by the genome size and colored by their host (yellow, bacteriophages; pink, archaeal viruses). Adapted from Kazlauskas, D., Krupovic, M., Venclovas, Č., 2016. The logic of DNA replication in double-stranded DNA viruses: Insights from global analysis of viral genomes. Nucleic Acids Research 44, 4551–4564.

What determines the completeness of the viral genome-encoded DNA replication machinery? It appears that one of the key factors is the size of the viral genome. In a sense, viral genome may be compared to a computer disk – the larger the disk, the more data it can hold (**Fig. 3**). For example, both archaeal and bacterial viruses with dsDNA genomes exceeding 140 kb all encode their own replicative helicases and DNA polymerases. However, the correlation between the genome size and the completeness of the viral-encoded DNA replication machinery is quite noisy and far from perfect. The increased coding capacity of viral genome is not always a decisive factor. This can be clearly seen in the list of prokaryotic viruses with the most complete DNA replication machineries (**Table 2**). For example, phage T4 with the 169 kb genome has all the components needed for the genome replication, whereas a phage with the 498 kb genome (*Bacillus* phage G) lacks the recognizable accessory subunits (DNA sliding clamp and clamp loader) of a DNA polymerase.

Not only the completeness of viral DNA replication machineries is highly variable, but also components making up these machineries are highly diverse. In particular, DNA replication proteins used by bacteriophages often differ from those used by their bacterial hosts. For chromosomal DNA replication bacteria use the DnaB-type superfamily 4 (SF4) helicases, TOPRIM domain-containing primase (i.e., *E. coli* DnaG), C-family DNA polymerase, processivity β-clamp and bacterial clamp loader, OB-fold-containing SSB protein. RNA primers in bacteria are excised by RNase HI and Pol I 5′-3′ exonuclease domain homologous to FEN1. Nicks are then sealed by a NAD+ – dependent DNA ligase. DNA replication machineries encoded by bacteriophages represent a mosaic of proteins typical of bacterial hosts, proteins of archaeo-eukaryotic type and those specific to viruses and other MGEs. Thus, bacteriophages are not simply mirroring DNA replication systems employed by their bacterial hosts. There are a number of components notably different from the bacterial ones. Among these are Superfamily 3 (SF3) helicases that are specific to MGEs and that are not involved in chromosomal replication of cellular organisms. Bacteriophages also frequently encode archaeo-eukaryotic primases (AEP) that have unrelated structural fold to bacterial TOPRIM primases. Most interestingly, C-family DNA polymerases that represent replicative polymerases in bacteria, are rarely found in bacteriophages. Instead, most abundant DNA polymerases in bacteriophages are represented by A- and B-family enzymes, both featuring the same structural fold of their catalytic domain. A-family polymerases in bacteria are ubiquitous, but their role is limited to participation in gap-filling during replication and in DNA repair. B-family polymerases in eukaryotes and archaea represent replicative enzymes whereas in bacteria their presence is sporadic. Thus, B-family polymerases provide another link between bacteriophage and archaeal/eukaryotic replication machineries. Some bacteriophages have DNA polymerase accessory proteins that also appear to be of archaeal/eukaryotic origin. The case in point is DNA sliding clamps and clamp loaders of T4-like bacteriophages. Both archaeal/eukaryotic (PCNA) and T4 (gp45) clamps are homotrimers, unlike the dimeric bacterial β-clamps. In addition, the crystal structure of T4 clamp loader (gp44/62) revealed that it represents a minimal version of the archaeal/eukaryotic RFC clamp loader. Archaeal/eukaryotic proteins participating in the lagging strand synthesis, most notably ATP-dependent DNA ligase, are also common in bacteriophage genomes.

Components of DNA replication systems encoded by archaeal viruses generally are homologous with the archaeal counterparts. However, there are fewer complete genomes of archaeal viruses, they tend to be smaller than in the case of bacteriophages and they are extremely diverse. Therefore, it is difficult just yet to make broader generalizations regarding components of RNA-primed replication systems in archaeal dsDNA viruses. Notably, some previous analyzes of complete archaeal genomes failed to find typical SSB proteins featuring OB-fold. However, a more recent metagenomics study identified archaeal members of the *Caudovirales*, dubbed Magroviruses, that possess dsDNA genomes of 65–100 kb in size and encode a nearly complete replication apparatus of apparent archaeal origin including SSB proteins. In addition to typical archaeal/eukaryotic proteins, just like bacteriophages, archaeal viruses were found to encode SF3 helicases, specific to MGEs.

Thus, there are some DNA replication proteins of prokaryotic viruses that are spread across both bacterial and archaeal domains. The 'universal' proteins include SF3 helicases, AEP primases, B-family DNA polymerases, RNase H and ATP-dependent DNA ligases. Only RNase H is commonly found in both bacteria and archaea, the remaining proteins, except SF3 helicases, are of archaeal/eukaryotic origin. If the similarity is considered at the structure level, the set of common archaeal/eukaryotic proteins could be extended even further into DNA sliding clamps and clamp loaders (PCNA/T4 gp45 and RFC/(T4 gp44/gp62)).

Protein-Primed DNA Replication

Some bacterial viruses with relatively small linear dsDNA genomes utilize the so-called protein-primed DNA replication. Cellular replicases use the 3′-hydroxyl (3′-OH) group provided by a nucleic acid primer (RNA or DNA) to initiate DNA synthesis. In contrast, protein-primed DNA polymerases utilize the OH group presented by a specific serine, threonine or tyrosine residue of a terminal protein (TP). Protein-primed DNA replication systems are distinctly different from those utilizing nucleic acid primers. Unlike the latter, protein-primed DNA replication systems require only few components to be fully functional. Such systems typically include a distinct B-family DNA polymerase, TP and an atypical single-stranded DNA binding protein (SSB). The DNA replication machinery of a protein-primed DNA replication system may be exemplified by one of the best studied such systems from *Bacillus subtilis* phage phi29 (**Fig. 4**). Phage phi29 genome is a linear approximately 19 kb-long dsDNA with covalently attached TP to each 5′-end. DNA replication is initiated by binding of the heterodimer of phi29 DNA polymerase and the free TP to the genomic DNA ends. Phi29 DNA polymerase uses the OH group of a specific serine residue as a primer bypassing the need for a primase. During the extension stage the phi29 DNA polymerase continues to synthesize DNA in a standard DNA-primed fashion and the displaced single-stranded regions are covered by SSB. Once the replication of a single DNA duplex is completed, there are two newly synthesized linear DNA duplexes with TP covalently attached to the 5′-ends. The structure of the phi29 DNA polymerase, so far the only experimentally determined 3D structure of a protein-primed DNA polymerase, has been instrumental in understanding structural and functional features of protein-primed DNA polymerases. Phi29 and other protein-primed DNA polymerases have two additional subdomains, Terminal Protein Region 1 (TPR1) and 2 (TPR2), that distinguish them from B-family members involved in conventional RNA-primed DNA replication. The TPR1 subdomain is involved in interaction with the TP. The palm, thumb and TPR2 subdomains form an internal sliding clamp-like structure that encircles the upstream duplex DNA, providing the enzyme with its inherently high processivity without the assistance of processivity factors. In addition, the TPR2 subdomain of the phi29 DNA polymerase couples processive DNA synthesis with the strand displacement in downstream dsDNA, making the function of a replicative helicase unnecessary. Based on identified genes for protein-primed DNA polymerase,

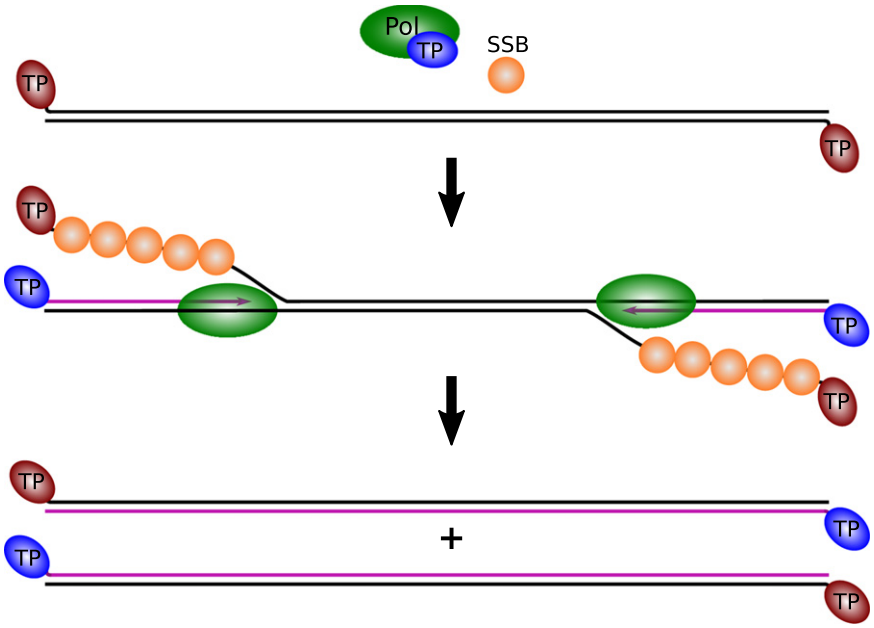

Fig. 4 Schematic representation of protein-primed replication of bacteriophage phi29. Pol, polymerase; TP, terminal protein; SSB, single-stranded DNA binding protein. Newly synthesized strands are shown in magenta.

five groups of archaeal viruses are also believed to replicate their genomes in this manner. These groups include bottle-shaped ampullaviruses, spindle-shaped salterprovirus His1, pleolipovirus His2, *Sulfolobus ellipsoid virus 1* and *Methanosarcina* Spherical Virus (MetSV). However, in these cases protein-primed DNA replication systems have not been studied experimentally. Apparently, the protein-primed DNA replication system is optimal for small-sized viral genomes, because these systems have only been identified in viruses with dsDNA genomes of no more than 50 kb.

Rolling Circle DNA Replication

Some archaeal and bacterial dsDNA viruses lacking typical components of either RNA-primed or protein-primed DNA replication systems utilize the so-called rolling circle DNA replication (RCR) mode. The signature of RCR is the presence of the multifunctional replication initiation protein (RCR-Rep). In many cases, RCR-Rep is the only viral protein needed for viral DNA replication. Therefore, RCR replication systems are prevalent in small viral genomes. The RCR replication mode is not specific to dsDNA prokaryotic viruses. In fact, RCR was first discovered and characterized in *E. coli* bacteriophage phiX174 and other ssDNA bacteriophages. Subsequently, RCR was identified as the replication mode of many other genetic elements such as bacterial and archaeal plasmids as well as a mechanism used by plasmids for conjugative DNA transfer. DsDNA bacteriophages that utilize the RCR replication mode include bacteriophages such as P2, 186, HP1 and PM2. The general mechanism of dsDNA bacteriophage genome replication using the RCR mode can be illustrated with that of bacteriophage P2 (Fig. 5). Following the entry of linear P2 DNA into the host cell, DNA circularizes due to the cohesive DNA ends and is sealed by the cellular DNA ligase. The RCR-Rep protein (protein A) of bacteriophage P2 nicks a circular DNA at the replication origin generating the covalent protein-DNA intermediate and a free 3′-OH. The latter primes leading strand DNA synthesis by the bacterial replicase, which uses the non-cleaved strand as a template. After the full circle is replicated, protein A cuts the newly generated origin and also acts as a ligase, producing a covalently closed circular ssDNA molecule. In this way, RCR-Rep is able to perform multiple rounds of cleavage and ligation at the origin by what has been termed a "flip-flop" mechanism. The circular ssDNA is converted to dsDNA by the replication machinery of the host.

Some archaeal dsDNA viruses are also known to encode RCR-Rep homologs. Among those archaeal viruses are rudiviruses that all have a candidate RCR-Rep protein. The structure of the RCR-Rep protein (gp119) of *Sulfolobus islandicus* rod-shaped virus 1 (SIRV1) has been solved and it revealed similarity to the HUH endonuclease superfamily. DNA replication of another rudivirus, SIRV2, has been studied experimentally. These studies revealed that although SIRV2 RCR-Rep protein can both nick and seal DNA like canonical RCR-Rep proteins, the RCR replication mode cannot fully explain the experimental observations. Based on the observed multimeric and highly branched DNA replication intermediates, it was suggested that SIRV2 could employ a combination of strand-displacement, rolling-circle and strand-coupled genome replication mechanisms. Whether all of these mechanisms occur during each cycle of viral DNA replication remains unclear. In another archaeal virus, sphaerolipovirus SNJ1 (*Sphaerolipoviridae*), the RCR-Rep protein of the HUH endonuclease superfamily has been experimentally shown to be essential for

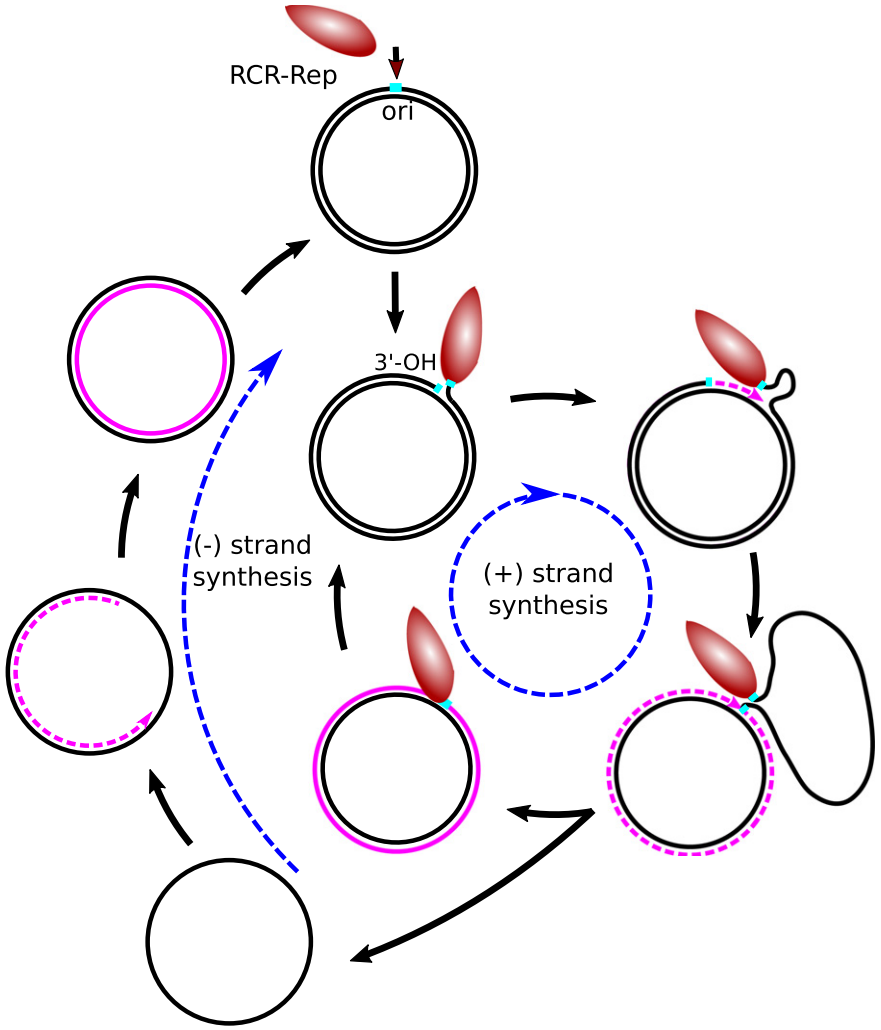

Fig. 5 Schematic illustration of the rolling-circle DNA replication based on the bacteriophage P2 replication mechanism. Phage RCR-Rep protein cuts the circular dsDNA at the replication origin (cyan) generating the covalent protein-DNA intermediate and a free 3′-OH group. Host's replicase utilizes the 3′-OH group as a primer and synthesizes the new (+) strand (magenta) while displacing the original strand. After the full circle is replicated, RCR-Rep nicks the newly generated origin at the same time sealing the displaced strand to produce a closed ssDNA circle. Host's replication machinery converts circular ssDNA to dsDNA by synthesizing complementary (−) strand (magenta). In the case of ssDNA viruses replicating through RCR, the latter process is the initial step in DNA replication.

viral genome replication. Furthermore, point mutations confirmed that the inferred catalytic residues of RCR-Rep are essential for its functionality. As pointed out above, the RCR mode of replication can be used by either dsDNA or ssDNA genomes. This is nicely illustrated in the case of haloarcheal pleomorphic viruses. For example, *Haloarcula hispanica* pleomorphic virus 1 (HHPV-1) with circular dsDNA genome is closely related to *Halorubrum* pleomorphic virus 1 (HRPV-1) with circular ssDNA genome and both encode a homolog of RCR-Rep protein.

Since RCR is not the main replication mode in dsDNA prokaryotic viruses, but is prevalent in ssDNA viruses, the detailed discussion on the RCR mechanism and RCR-Rep proteins is provided in the section devoted to the replication of ssDNA viruses (see below).

Other DNA Replication Mechanisms

There are prokaryotic dsDNA viruses lacking any of the components indicative of a specific DNA replication system such as a replicative helicase (suggestive of RNA-primed DNA replication), a protein-primed DNA polymerase (protein-primed DNA replication) or an RCR-Rep homolog (suggestive of the RCR replication mode). In such cases dsDNA bacteriophages frequently encode either integrases, DNA recombination proteins or various replication initiation proteins, including homologs of bacterial replisome organizer DnaA and helicase loader DnaC. These observations suggest that in these cases viral genome is either

integrated into the host genome and propagated by the replication of the host genome or other proteins in the replication initiation stage are used to recruit cellular replication proteins. There are also cases when no typical DNA replication proteins were detected in the viral genomes. In particular this is true in the case of archaeal viruses. These viruses might either have highly diverged DNA replication proteins that cannot be detected by current computational approaches or use new replication strategies. The existence of the latter possibility is exemplified by *Acidianus* filamentous virus 1 (AFV1), which appears to use a novel strand displacement mechanism.

Genome Replication of Prokaryotic ssDNA Viruses

Viruses with ssDNA genomes are among the smallest viruses and encode as few as two proteins – one for capsid formation and the other one for genome replication. SsDNA bacteriophages have circular genomes and are classified into two families, *Microviridae* and *Inoviridae*. The *Microviridae* family comprises ssDNA phages with small icosahedral capsids including archetypal phage phiX174. Members of this family have circular ssDNA genomes in the 4.4–6.1 kb range. The *Inoviridae* family comprises filamentous ssDNA phages such as the M13 phage. Inoviruses also have circular ssDNA genomes and their size is in the 4.5–12.4 kb range. Officially recognized ssDNA bacteriophages represent only a small fraction in comparison with dsDNA bacteriophages. However, recent metagenomic studies found that ssDNA bacteriophages are present in various environments and that their abundance and diversity have been underestimated.

Archaeal viruses that have ssDNA genomes belong to two officially recognized families, *Pleolipoviridae* and *Spiraviridae*. Members of *Pleolipoviridae* similarly to the ssDNA bacteriophages have rather small ssDNA genomes ranging from 7 to 10.6 kb. Strikingly, the family *Pleolipoviridae* comprises viruses not only with ssDNA genomes, but also with either linear or circular dsDNA genomes. This observation complicates the traditional virus classification based on genome type. Not all archaeal ssDNA viruses are small. The sole member of the *Spiraviridae* family, *Aeropyrum* coil-shaped virus (ACV) infecting an extreme aerobic hyperthermophile (*A. pernix*) was found to have an unusually large (24.9 kb) ssDNA genome. So far, the ssDNA genome of ACV is the largest among all known ssDNA viruses.

Rolling Circle Replication

Most bacteriophages possessing ssDNA genome replicate using the RCR mechanism. In general, genome replication of ssDNA bacteriophages is similar to the RCR-using dsDNA viruses (see **Fig. 5**). The key difference is that in the case of ssDNA phages there is an additional stage, the conversion of ssDNA into dsDNA. The whole process can be generally divided into three stages as exemplified by the phage phiX174 replication. During the first stage, the infecting (+)ssDNA is converted by host replication proteins into a covalently closed dsDNA called replicative form DNA (RF). During the second stage, RF DNA is amplified. RCR-Rep (protein gpA) nicks the DNA at the origin of replication forming the covalent 5′-phosphotyrosine intermediate and a free 3′-OH. The 3′-OH serves as a primer for host DNA replicase which uses (−)strand as a template to synthesize DNA by "peeling off" the (+)strand. After one round of rolling circle synthesis, RCR-Rep cuts the newly generated origin and acts as a ligase, producing a covalently closed circular (+)ssDNA molecule. Newly generated (+)ssDNA genomes are again converted into dsDNA circular molecules. During the last stage of DNA synthesis, the ssDNA genome is concurrently synthesized and packaged into the viral procapsid.

The key role in ssDNA replication belongs to a multifunctional RCR-Rep protein. Early on it was noticed that RCR-Rep proteins are not monophyletic and that they often cluster with corresponding proteins from rolling-circle plasmids found in bacteria and archaea. Bacteriophages with ssDNA that replicate through the RCR mechanism use RCR-Rep proteins belonging to at least two evolutionarily and structurally distinct superfamilies, the HUH endonucleases and the TATA-box binding protein-like enzymes.

The superfamily of HUH endonucleases is named after the conserved metal binding HUH motif, consisting of two His residues separated by a bulky hydrophobic residue (U). In addition, HUH endonucleases have either one or two catalytic Tyr residues that form the covalent 5′-phosphotyrosine intermediate during nicking of ssDNA. The phage phiX174 RCR-Rep (protein gpA) represents the founding member of the RCR-Rep HUH endonucleases. It uses two catalytic Tyr residues to initiate and terminate rolling-circle replication and to spin off multiple circles of phiX174 ssDNA. Remarkably, despite decades of studies the structure of phiX174 gpA, the founding member of HUH endonuclease superfamily, remains unsolved. Peculiarly, the closest homolog with the solved three-dimensional structure appears to be a protein (PDB id: 2X3G) from dsDNA archaeal virus SIRV1. HUH endonucleases have the ferredoxin-like fold which is also known as the RNA recognition motif (RRM). HUH endonucleases also share similarity with origin binding proteins of small dsDNA viruses suggesting that there are intricate and ancient evolutionary relationships between these proteins.

RCR-Rep proteins of the second class, typified by GP2 of phage M13, belong to the Pfam family *Phage_CRI* (PF05144) and the related *Rep_trans* (PF02486) family of plasmid replication initiation proteins. Only recently the first three-dimensional structures of *Rep_trans* family members have been solved. These structures revealed that RCR-Rep proteins of *Rep_trans* family feature TATA-box binding protein-like (TBP-like) structural fold, which is entirely different from the RRM fold of HUH endonucleases. The metal binding site is also organized differently. *Rep_trans* representatives do not have the HUH motif. Instead, the metal binding is mediated by three acidic residues. However, despite structural differences and differently organized metal binding sites of the two types of RCR-Rep proteins, both types use an active site tyrosine residue as the nucleophile during the nicking/religation reactions. Moreover, the side chains of the corresponding catalytic tyrosine residues adopt a similar orientation with respect to the metal ions.

Therefore, these two different types of RCR-Rep proteins may represent a case of convergent evolution for catalyzing the same nicking and religation reactions. Interestingly, even within the same phage family RCR-Rep proteins may be of different types as exemplified by *Inoviridae*. Thus, the RCR-Rep protein from *Xanthomonas* phage Lf is a member of the HUH endonuclease superfamily, whereas the one from phage M13 belongs to the TBP-like superfamily.

Most ssDNA archaeal viruses, similarly to ssDNA bacteriophages, likely replicate their genomes via RCR or related rolling hairpin replication mechanisms as implied by the identified RCR-Rep homologs. A candidate replication initiator (Rep) family gene was identified in HRPV-1, a representative of the *Pleolipoviridae* family and the first characterized archaeal virus having ssDNA genome. The HRPV-1 Rep gene is homologous to RCR-Rep HUH endonucleases. On the other hand, no potential candidate showing significant sequence homology or sharing conserved motifs with known RCR-Rep proteins could be identified in ACV, a virus with the largest ssDNA genome, suggesting that it might use a unique mechanism of genome replication.

Genome Replication of Prokaryotic RNA Viruses

RNA viruses are divided into three distinct classes depending on the nature of their genome: positive-sense (+) single-stranded RNA (ssRNA) viruses, double-stranded (ds) RNA viruses, and negative-sense (−)ssRNA viruses. Although RNA viruses infecting eukaryotes are very common, the known representation of prokaryotic RNA viruses is quite narrow. Currently, there are only two recognized families of RNA bacteriophages, the *Leviviridae* family consisting of (+)ssRNA bacteriophages and the *Cystoviridae* family comprising segmented dsRNA bacteriophages. So far, no (−)ssRNA bacteriophages have been isolated. At present, there are no known RNA viruses that infect archaea. Although the putative presence of archaeal (+)ssRNA viruses has been reported, this finding remains unsubstantiated. Recent metagenomics surveys in diverse environments have substantially expanded known host range and spectrum of genome architectures of RNA bacteriophages. Nonetheless, it appears that the prokaryotic RNA virome is both significantly smaller and less diverse than the DNA virome.

The *Leviviridae* family comprises bacteriophages with (+)ssRNA genomes. The family includes four recognized species (Enterobacteria phage Qβ, Enterobacteria phage F1, Enterobacteria phage MS2, and Enterobacteria phage BZ13). Bacteriophages of the *Leviviridae* family have a monopartite (+)ssRNA genome of 3.3–4.3 kb in size and are among the simplest and smallest of known viruses. Notably, the genome of phage MS2 was the first ever genome to be fully sequenced.

The *Cystoviridae* family for over two decades was represented by only one recognized virus species, *Pseudomonas* phage phi6. Recently, a number of additional dsRNA phages have been isolated from various environmental samples and six of them (Pseudomonas phages phi8, phi12, phi13, phi2954, phiNN and phiYY) have been fully sequenced. All seven ICTV-recognized species of the *Cystoviridae* family have a dsRNA genome, which is divided into three separate segments: large (L), medium (M) and small (S). Individual segments range in size from 2.9 to 6.4 kb. The total genome size varies from 12.7 kb (phi2954) to 15.0 kb (phi8).

Replication Using an RNA-Dependent RNA Polymerase

Bacterial hosts lack the capability to synthesize complementary strands using the RNA strand as a template. Thus, all characterized RNA phages encode their own RNA-dependent RNA polymerase (RdRp). Notably, RdRp is the only universal protein not only in RNA phages, but in all known RNA viruses. In (+)ssRNA phages, the genome serves simultaneously as a genome template and as messenger RNA (mRNA). Since replication and translation run in the opposite directions, there is a competition between the two processes. Viral RdRp assembles with host proteins (ribosomal protein S1, translation elongation factors EF-Tu and EF-Ts) to form the active RNA polymerase holoenzyme (replicase). The replicase then initiates synthesis of the negative RNA strand by replicating through the (+) ssRNA genome. In turn, the newly synthesized (−)ssRNA strand is used to produce new (+)ssRNA genomes for the viral progeny.

In the case of dsRNA phages, RdRp (P2) is encoded in the largest genome fragment. The cystoviral RdRp first initiates synthesis of the plus-strands by unwinding each of the dsRNA segments and using the minus-strands as template. Newly synthesized (+)ssRNA segments (+S, +M, +L) are utilized as mRNAs by the host translational machinery. Once phage proteins are produced, the empty procapsids are self-assembled and the three (+)ssRNA segments are packaged into each of them. Following packaging, P2 initiates a single round of minus-strand RNA synthesis to recapitulate the dsRNA phage genome.

RdRps of (+)ssRNA and dsRNA phages as well as those of eukaryotic RNA viruses are evolutionary related. Viral RdRps belong to the class of right-hand polymerases that share the same RRM structural fold of their "palm" domain. This class also includes single-subunit DNA-dependent RNA polymerases, reverse transcriptases and DNA polymerases of A, B and Y families. Based on both sequence and structure analysis it was suggested that RdRps may represent the most ancient group of right-hand polymerases in agreement with the RNA World hypothesis, because RdRps could serve as replicases of RNA genomes. Phylogenetic analyzes have further suggested that RdRps of dsRNA viruses evolved from (+)ssRNA viruses pointing to their ancient origin.

Concluding Remarks

Genome replication strategies employed by prokaryotic viruses are extremely diverse. This diversity in part can be explained by differences in type (DNA or RNA) and topology (linear or circular, double- or single-stranded, contiguous or segmented) of nucleic

acids encoding viral genomic information. However, even those prokaryotic viruses that, like cellular organisms, have dsDNA genomes, show remarkable diversity both in replication strategies and in composition of their DNA replication machineries. Interestingly, bacterial viruses often encode DNA replication proteins typical not of bacteria, but of archaea and eukaryotes, pointing to an ancient and complex evolutionary history of dsDNA replication systems of bacteriophages. Other viral DNA replication strategies such as protein-primed DNA replication and rolling circle replication evolved within a broader context of mobile genetic elements and apparently also have deep roots. Viral RNA-dependent RNA polymerases are believed to represent the most ancient group of right-handed polymerases in accord with the RNA world theory. Therefore, studies of viral proteins involved in genome replication might hold important clues for the emergence and the evolution not only of viral but also of cellular DNA replication machineries. A very fast pace at which genomic and metagenomic data for viruses are currently accumulating provides a solid basis for such studies.

Acknowledgments

The author thanks Darius Kazlauskas and Mart Krupovic for comments and suggestions. This work was in part supported by the Research Council of Lithuania [09.3.3-LMT-K-712–01–0080].

Further Reading

Callanan, J., Stockdale, S.R., Shkoporov, A., *et al.*, 2018. RNA phage biology in a metagenomic era. Viruses 10.
Carr, S.B., Phillips, S.E., Thomas, C.D., 2016. Structures of replication initiation proteins from staphylococcal antibiotic resistance plasmids reveal protein asymmetry and flexibility are necessary for replication. Nucleic Acids Research 44, 2417–2428.
Černý, J., Černá Bolfíková, B., De, A.Z.P.M., Grubhoffer, L., Růžek, D., 2015. A deep phylogeny of viral and cellular right-hand polymerases. Infection Genetics and Evolution 36, 275–286.
Chandler, M., De La Cruz, F., Dyda, F., *et al.*, 2013. Breaking and joining single-stranded DNA: The HUH endonuclease superfamily. Nature Reviews Microbiology 11, 525–538.
Dellas, N., Snyder, J.C., Bolduc, B., Young, M.J., 2014. Archaeal viruses: Diversity, replication, and structure. Annual Review of Virology 1, 399–426.
Depamphilis, M., Bell, S., 2010. Genome Duplication. Taylor & Francis Group.
Devoto, A.E., Santini, J.M., Olm, M.R., *et al.*, 2019. Megaphages infect Prevotella and variants are widespread in gut microbiomes. Nature Microbiology 4, 693–700.
Kazlauskas, D., Krupovic, M., Venclovas, Č., 2016. The logic of DNA replication in double-stranded DNA viruses: Insights from global analysis of viral genomes. Nucleic Acids Research 44, 4551–4564.
Koonin, E.V., Dolja, V.V., 2013. A virocentric perspective on the evolution of life. Current Opinion in Virology 3, 546–557.
Kornberg, A., Baker, T.A., 2005. DNA Replication. University Science.
Krupovic, M., 2013. Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. Current Opinion in Virology 3, 578–586.
Krupovic, M., Cvirkaite-Krupovic, V., Iranzo, J., Prangishvili, D., Koonin, E.V., 2018. Viruses of archaea: Structural, functional, environmental and evolutionary genomics. Virus Research 244, 181–193.
Krupovic, M., Forterre, P., 2015. Single-stranded DNA viruses employ a variety of mechanisms for integration into host genomes. Annals of the New York Academy of Sciences 1341, 41–53.
Malathi, V.G., Renuka Devi, P., 2019. ssDNA viruses: Key players in global virome. Virusdisease 30, 3–12.
Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., *et al.*, 2016. Uncovering Earth's virome. Nature 536, 425–430.
Prangishvili, D., Bamford, D.H., Forterre, P., *et al.*, 2017. The enigmatic archaeal virosphere. Nature Reviews Microbiology 15, 724–739.
Prangishvili, D., Koonin, E.V., Krupovic, M., 2013. Genomics and biology of Rudiviruses, a model for the study of virus-host interactions in Archaea. Biochemical Society Transactions 41, 443–450.
Rumnieks, J., Tars, K., 2018. Protein-RNA interactions in the single-stranded RNA bacteriophages. Subcellular Biochemistry 88, 281–303.
Salas, M., De Vega, M., 2016. Protein-primed replication of bacteriophage Phi29 DNA. Enzymes 39, 137–167.
Székely, A.J., Breitbart, M., 2016. Single-stranded DNA phages: From early molecular biology tools to recent revolutions in environmental microbiology. FEMS Microbiology Letters 363.
Wawrzyniak, P., Płucienniczak, G., Bartosik, D., 2017. The different faces of rolling-circle replication and its multifunctional initiator proteins. Frontiers in Microbiology 8, 2353.
Weigel, C., Seitz, H., 2006. Bacteriophage replication modules. FEMS Microbiology Reviews 30, 321–381.
Wolf, Y.I., Kazlauskas, D., Iranzo, J., *et al.*, 2018. Origins and evolution of the global RNA virome. mBio 9.

Relevant Websites

http://ictv.global
International Committee on Taxonomy of Viruses (ICTV).
https://www.ncbi.nlm.nih.gov/genome/viruses/
NCBI Viral Genomes.
http://dmk-brain.ecn.uiowa.edu/VOG/
The database of Prokaryotic Virus Orthologous Groups, or pVOGs.
https://viralzone.expasy.org/
Viral Zone.
http://www.virology.ws/
Virology blog.