VILNIUS UNIVERSITY

Darius Kazlauskas

COMPUTATIONAL ANALYSIS OF DNA REPLICATION PROTEINS IN DOUBLE-STRANDED DNA VIRUSES

Summary of doctoral Dissertation

Physical Sciences, Biochemistry (04 P)

Vilnius, 2014

The dissertation work was carried out at the Department of Bioinformatics, Institute of Biotechnology, Vilnius University during 2010 – 2014.

Scientific supervisor:

dr. Česlovas Venclovas (Vilnius University; Biomedical sciences, biology – 01 B and Physical sciences, biochemistry – 04 P)

The dissertation is defended at the Council of Biochemistry science direction of Vilnius University:

Chairman:

prof. dr. Rimantas Daugelavičius (Vytautas Magnus University, Physical sciences, biochemistry – 04 P)

Members:

dr. Mart Krupovič (Institut Pasteur, Biomedical sciences, biology – 01 B)

prof. dr. Edita Sužiedėlienė (Vilnius University, Physical sciences, biochemistry – 04 P)

dr. Rolandas Meškys (Vilnius University, Physical sciences, biochemistry – 04 P)

dr. Giedrė Tamulaitienė (Vilnius University, Physical sciences, biochemistry – 04 P)

The thesis defence will take place at the Institute of Biotechnology, Vilnius University (Graičiūno 8, LT-02241 Vilnius, Lithuania) on 4th of December, 2014, at 11 a.m.

The summary of doctoral dissertation was sent on 3rd of November, 2014.

The thesis is available at the Library of Vilnius University and at website of Vilnius University: http://www.vu.lt/naujienos/ivykiu-kalendorius

VILNIAUS UNIVERSITETAS

Darius Kazlauskas

DVIGRANDĖS DNR VIRUSŲ DNR REPLIKACIJOS BALTYMŲ ANALIZĖ KOMPIUTERINIAIS METODAIS

Daktaro disertacija

Fiziniai mokslai, biochemija (04 P)

Vilnius, 2014

Disertacija rengta 2010 – 2014 metais Vilniaus universiteto Biotechnologijos instituto Bioinformatikos skyriuje.

Mokslinis vadovas:

dr. Česlovas Venclovas (Vilniaus universitetas; biomedicinos mokslai, biologija – 01 B ir fiziniai mokslai, biochemija – 04 P)

Disertacija ginama Vilniaus universiteto Biochemijos mokslo krypties taryboje:

Pirmininkas:

prof. dr. Rimantas Daugelavičius (Vytauto Didžiojo universitetas, fiziniai mokslai, biochemija – 04 P)

Nariai:

dr. Mart Krupovič (Pastero institutas, biomedicinos mokslai, biologija – 01 B)

prof. dr. Edita Sužiedėlienė (Vilniaus universitetas, fiziniai mokslai, biochemija – 04 P)

dr. Rolandas Meškys (Vilniaus universitetas, fiziniai mokslai, biochemija – 04 P)

dr. Giedrė Tamulaitienė (Vilniaus universitetas, fiziniai mokslai, biochemija – 04 P)

Disertacija bus ginama viešame Biochemijos mokslo krypties posėdyje 2014 gruodžio 4 d. 11 val. Vilniaus universiteto Biotechnologijos instituto aktų salėje.
Adresas: V.A. Graičiūno 8, LT-02241 Vilnius, Lietuva.

Disertacijos santrauka išsiuntinėta 2014 lapkričio 3 d.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje ir VU svetainėje adresu: http://www.vu.lt/naujienos/ivykiu-kalendorius

TABLE OF CONTENTS

INTRODUCTION

Ability to replicate and evolve are two distinct features of all living entities. Duplication of genetic information is carried out by replication proteins. Composition of DNA replication machinery is similar in all free-living cellular organisms. In contrast, replication in double-stranded (ds) DNA viruses is very diverse. It is well studied in T7 and T4 phages, herpes, polyoma and papilloma viruses, however, these groups make up only about 10% of known dsDNA viruses. How do lesser known viruses replicate? Do they use variations of already known replication systems? Or perhaps, they use novel replication strategies? DsDNA viruses are not only diverse, but they also vary in genome size. For example, genomes of smallest dsDNA viruses (polyoma) are 500 times smaller than that of the largest Pandora viruses (genome size – 2500 kbp). Genome size in free-living cellular organisms also varies. For example, genome size difference between human and the smallest eukaryote (*Ostreococcus tauri*) is ~260-fold. However, they have the same components of replication machinery. Is this true for dsDNA viruses? Or maybe, the diversity and genomic distribution of viral DNA replication proteins depends on virus genome size?

We attempted to answer questions mentioned above by performing a detailed computational analysis of DNA replication proteins in dsDNA viruses. Using current state-of-the-art computational methods we identified and characterized replication proteins (DNA polymerases, processivity factors, clamp loaders, primases, helicases, single-stranded DNA binding proteins, primer removal proteins, DNA ligases and topoisomerases) and analyzed their distribution patterns in genomes of dsDNA viruses.

This study was carried out in two stages. At first we analyzed DNA replicases (DNA polymerases, processivity factors, clamp loaders). The analysis revealed dependency between DNA replicase components and the viral genome size. We found that small viruses (<40 kbp) use protein-primed DNA replication or rely on replication proteins from the host. Large viruses (>140 kbp) have their own RNA-primed replication apparatus often supplemented with processivity factors and sometimes by clamp loaders to increase replication speed and efficiency. The only seeming exception from the latter general pattern was eliminated after finding B-family DNA polymerases in large phiKZ phages. Next, we asked whether the distribution of other viral DNA replication proteins depends on genome size. It turned out that as the genome size increases viruses tend to encode their own replication proteins more frequently. Latter insight led us to a search for "missing" replication components in large genomes. This has resulted in the discovery of single-stranded DNA binding (SSB) proteins in largest eukaryotic viruses. Surprisingly, these proteins turned out to be homologs of SSB proteins previously thought to be specific for T7-like phages. Another surprise came from the analysis of DNA helicases. We found out that replicative helicases are the most common replication proteins in dsDNA viruses. In addition, our analysis revealed that the component of herpesviral helicase-primase complex (UL8) is a highly diverged and inactivated B-family DNA polymerase.

The aim of this study was to analyze DNA replication proteins in double-stranded DNA viruses using computational methods.

Specific objectives:

1. To computationally identify and characterize DNA replication proteins by analyzing viral genomes and proteins.
2. To check for the presence and the nature of dependency between DNA replication proteins and viral genome size.

Scientific novelty

This work is the first large scale computational study of all replication proteins from dsDNA viruses. The analysis revealed a dependency between DNA replication proteins and viral genome size. We newly discovered and characterized SSB proteins in NCLD viruses and B-family DNA polymerases in large phiKZ phages. We detected a significant similarity between poxviral SSB (I3) and bacterial SmpB. We also revealed that the component of herpesviral helicase-primase complex (UL8) is a highly diverged and inactivated B-family DNA polymerase.

Practical value

Relationship between DNA replication proteins and viral genome size, discovered in this study, enables one to predict the completeness of DNA replication machinery in newly sequenced dsDNA virus genomes. In addition, dsDNA viruses are usually pathogenic (herpes, papilloma, polyoma, adeno viruses) or agricultural pests (African swine fever, nimaviruses). Thus, new knowledge about DNA replication of *Pseudomonas aeruginosa* phiKZ phages, poxviruses and herpesviruses, presented in this study, may help scientists to better understand and fight diseases.

1. METHODS

1.1. Databases

Non-redundant ("nr") protein sequence database was downloaded from NCBI: "ftp://ftp.ncbi.nlm.nih.gov/blast/db/". Viral protein and genome sequences were also obtained from NCBI using the address: "http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?taxid=10239". All the genomes of dsDNA viruses were subjected to the six-frame translation using Virtual Ribosome (1) or Bioperl (2). Family *Polydnaviridae* was excluded from the analysis because these viruses have a distinct genome organization (split in small segments), and their genome acts only as a vector for transmission of parasitic wasp genes (3).

1.2. Sequence similarity searches

Standard sequence searches were performed using PSI-BLAST (4) and jackhmmer (5). Searches were run iteratively against the nr70 sequence database (the non-redundant database with no more than 70% identity between any sequences) until convergence using E-value=1e-03 or a more stringent inclusion threshold. Programs HHsearch (6) and Condor (http://mindaugas.ibt.lt/condor/) were used for sensitive homology search. Sequence profiles of viral proteins were generated by running two or three iterations against nr70 database using the E-value=1e-03 inclusion threshold. HHsearch or Condor with default parameters were then used to search the PDB (http://www.pdb.org/), SCOP (7) and Pfam (8) databases. HHsearch and Condor results with probability >20% and E-value <10, respectively, were extracted and analyzed for the presence of DNA replication proteins.

1.3. Sequence clustering

DNA replication proteins were clustered according to their pairwise similarity using CLANS (9). The similarity in CLANS is represented with P-values derived from BLAST or PSI-BLAST E-values. For clustering divergent proteins, their pairwise similarity was quantified using PSI-BLAST. For each sequence, CLANS was configured to run two iterations of PSI-BLAST using the E = 1e-03 inclusion threshold against the reference database (nr70) to generate a sequence profile. The last PSI-BLAST iteration with the obtained profile was performed against the database of sequences to be clustered.

1.4. Identification of replication proteins

Replication proteins of dsDNA viruses were identified using criteria listed below (arranged in decreasing priority order):

1. Similarity to characterized DNA replication proteins.
2. Presence (absence) of active site and other conserved regions.
3. Protein contains domain of other DNA replication protein (for example, DNA helicase and primase).
4. Protein is encoded in the vicinity of DNA replication proteins.

1.5. Multiple sequence alignments

Multiple sequence alignments were constructed with MAFFT (10) optimized for accuracy (parameter L-INS-i) or MUSCLE (11) using default parameters. If sequences had homologs with known structures, PROMALS3D (12) with default parameters was used instead.

1.6. Prediction of protein secondary structure and disordered regions

Genesilico (13) server was used to predict protein secondary structure and disordered regions. If protein structure was known, secondary structure was calculated using DSSP (14).

1.7. Genome comparison and the analysis of gene context

Alignment of genomes and the inspection of gene neighborhoods were performed using ACT (Artemis Comparison Tool) (15).

1.8. Genome filtering

To obtain a more representative genome set, highly similar genomes were removed. All genomes were grouped according to their nucleotide and protein pairwise similarities using LAST (16) and CLANS, respectively. Genomes with local nucleotide sequence identity >70% or those which had more than 70% homologous proteins were filtered out.

1.9. Homology modeling and structure analysis

Initial sequence-structure alignments were constructed based on alignments produced by PSI-BLAST-ISS (17), COMA (18), GeneSilico, I-TASSER (19), RaptorX (20), pGenTHREADER (21), FFAS-3D (22), HHpred (23) servers and subsequently modified during an iterative modeling process (24). Modeller 9v10 (25) was used for 3D model construction. The quality of resulting models was assessed by the ProSA-web server (26) and then compared to the quality of the corresponding main structural templates whose missing loops where modeled-in with Modeller prior to evaluation. Conservation scores, derived from the multiple sequence alignments, were mapped onto the surfaces of models with ConSurf (27).

1.10. Analysis of electrostatic properties

Surface electrostatic maps were calculated using the APBS (28) plugin in PyMol (29). Calculation of theoretical isoelectric points (pIs) was performed using the "Isoelectric point" program from the EMBOSS software package (30). Non-conserved N- and C-termini were removed from the sequences before the pI calculation.

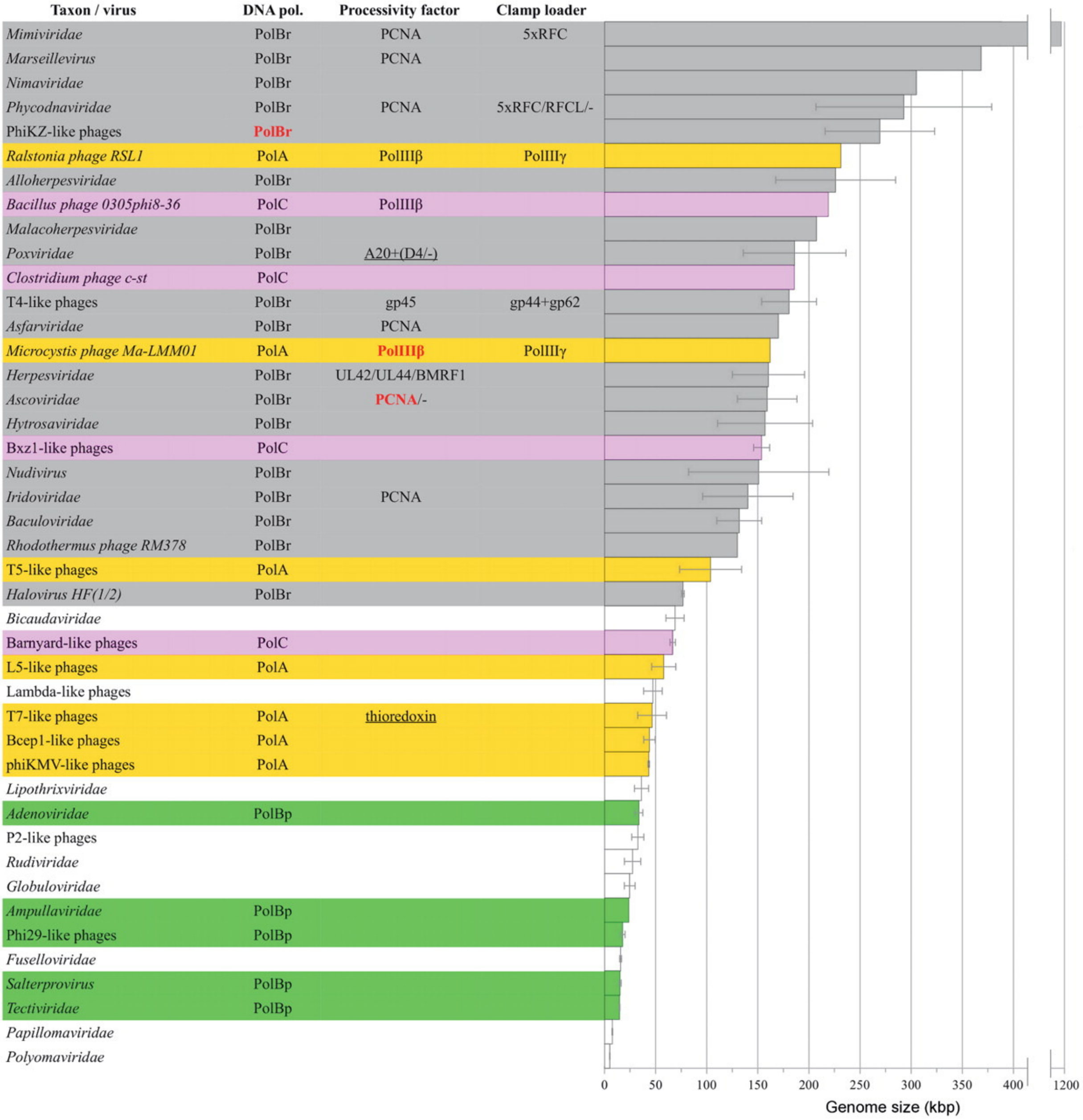
2. RESULTS AND DISCUSSION

This study was carried out in two stages. At first, DNA replicases were investigated. The analysis revealed that the presence and the nature of DNA replicases encoded in the genomes of dsDNA viruses is related to the genome size. This observation led us to a detailed study and discovery of DNA polymerases in large phiKZ phages. Next we asked whether other replication proteins also show the genome size dependency. The analysis revealed that as genome size increases viruses tend to encode their own replication proteins more frequently. Further studies of viral replication proteins led us to even more insights and discoveries. Surprisingly, it turned out, that DNA helicase is the most common viral replication protein. While analyzing helicases, we discovered that the component (UL8) of herpesviral helicase-primase is inactivated B-family polymerase. The examination of viral single-stranded DNA-binding (SSB) proteins revealed that the largest eukaryotic DNA viruses have at least two distinct SSB families. All the topics mentioned above are presented in detail in the following sections.

2.1. DNA replicase components and the genome size

The available fully sequenced genomes of dsDNA viruses were analyzed for the presence of DNA replicase components. In all, genomes of 808 viruses including 458 (57%) bacteriophages, 317 (39%) eukaryotic and 33 (4%) archaeal viruses were examined. Specifically, we looked for DNA polymerases, polymerase processivity factors (DNA sliding clamps) and clamp loader subunits. We detected DNA polymerases in about half of the analyzed viral genomes. In addition to either known or previously annotated enzymes, for the first time we identified highly divergent DNA polymerases in phiKZ-like bacteriophages. We found a significantly smaller fraction of genomes (<20%) coding for homologs of DNA sliding clamps that may serve as DNA polymerase processivity factors. We newly discovered remote homologs of cellular DNA sliding clamps in *Microcystis phage Ma-LMM01* and the *Ascoviridae* family. DNA sliding clamps that form rings (PCNA, polIIIβ, gp45) need a multimeric clamp loader for their loading onto DNA. In line with this prerequisite, we detected clamp loader subunits only in genomes carrying genes of DNA sliding clamp homologs. Yet, surprisingly, not all PCNA or polIIIβ homologs are accompanied by clamp loader subunits.

Overall, the results revealed a great variety of DNA replicase components and their combinations in dsDNA viruses. The variety is much larger than it is in all three domains of cellular life combined and seemingly without any discernible pattern. However, we reasoned that if the increase in viral genome size requires improved processivity properties of a DNA replicase we should be able to detect this dependency even in the face of this overwhelming variety. Indeed, the arrangement of viral taxonomic groups according to their average genome size revealed a clear trend (Fig. 2.1). Viruses having smallest genomes (<40 kb) either have a B-family protein-primed DNA polymerase or do not have a DNA polymerase at all. Viruses with larger genomes (40–140 kb) have their own DNA polymerases more often. These polymerases usually belong to A-, rarely to B- or C-families. Viruses having largest genomes (>140 kb) always encode DNA polymerases (most often B-family RNA-primed), frequently have processivity factors and sometimes clamp loader subunits.



Taxon / virus	DNA pol.	Processivity factor	Clamp loader
Mimiviridae	PolBr	PCNA	5xRFC
Marseillevirus	PolBr	PCNA	
Nimaviridae	PolBr		
Phycodnaviridae	PolBr	PCNA	5xRFC/RFCL/-
PhiKZ-like phages	**PolBr**		
Ralstonia phage RSL1	PolA	PolIIIβ	PolIIIγ
Alloherpesviridae	PolBr		
Bacillus phage 0305phi8-36	PolC	PolIIIβ	
Malacoherpesviridae	PolBr		
Poxviridae	PolBr	A20+(D4/-)	
Clostridium phage c-st	PolC		
T4-like phages	PolBr	gp45	gp44+gp62
Asfarviridae	PolBr	PCNA	
Microcystis phage Ma-LMM01	PolA	**PolIIIβ**	PolIIIγ
Herpesviridae	PolBr	UL42/UL44/BMRF1	
Ascoviridae	PolBr	**PCNA**/-	
Hytrosaviridae	PolBr		
Bxz1-like phages	PolC		
Nudivirus	PolBr		
Iridoviridae	PolBr	PCNA	
Baculoviridae	PolBr		
Rhodothermus phage RM378	PolBr		
T5-like phages	PolA		
Halovirus HF(1/2)	PolBr		
Bicaudaviridae			
Barnyard-like phages	PolC		
L5-like phages	PolA		
Lambda-like phages			
T7-like phages	PolA	thioredoxin	
Bcep1-like phages	PolA		
phiKMV-like phages	PolA		
Lipothrixviridae			
Adenoviridae	PolBp		
P2-like phages			
Rudiviridae			
Globuloviridae			
Ampullaviridae	PolBp		
Phi29-like phages	PolBp		
Fuselloviridae			
Salterprovirus	PolBp		
Tectiviridae	PolBp		
Papillomaviridae			
Polyomaviridae			

Fig. 2.1 DNA replicase components in dsDNA viral genomes. Viral taxonomic groups are arranged by their average genome size. DNA pol., DNA polymerase type; PolA, A-family; PolBr, B-family DNA polymerase that uses RNA as a primer; PolBp, B-family DNA polymerase that uses protein as a primer; PolC, C-family. Coloring scheme: white, no polymerases found; green, PolBp; yellow, PolA; gray, PolBr; pink, PolC. Newly identified replicase components are labeled in bold red font. Processivity factors, non-homologous to the cellular ones, are underlined. Minus sign indicates that the processivity factor is missing in some viruses within the taxonomic group. Error bars indicate standard deviation from the mean genome size.

However, the representation of various viral taxonomic groups differs significantly. In addition, some taxons show quite large variation of the genome size. Therefore, we next asked whether or not the observed pattern of distribution of replicase components depends on the taxonomic classification of viruses. To address this question, we arranged individual genomes according to their size without dividing into taxonomic groups and plotted the observed frequency of a particular DNA replicase component against the moving average of the genome size (Fig. 2.2). To reduce sample bias in this analysis, we performed pairwise genome comparisons and retained only 236 viral genomes that were <70% identical to each other. Again, the plot showed a clear relationship between DNA replicase components and the genome size, indicating that this is a general property and not the result of taxon-specific division.

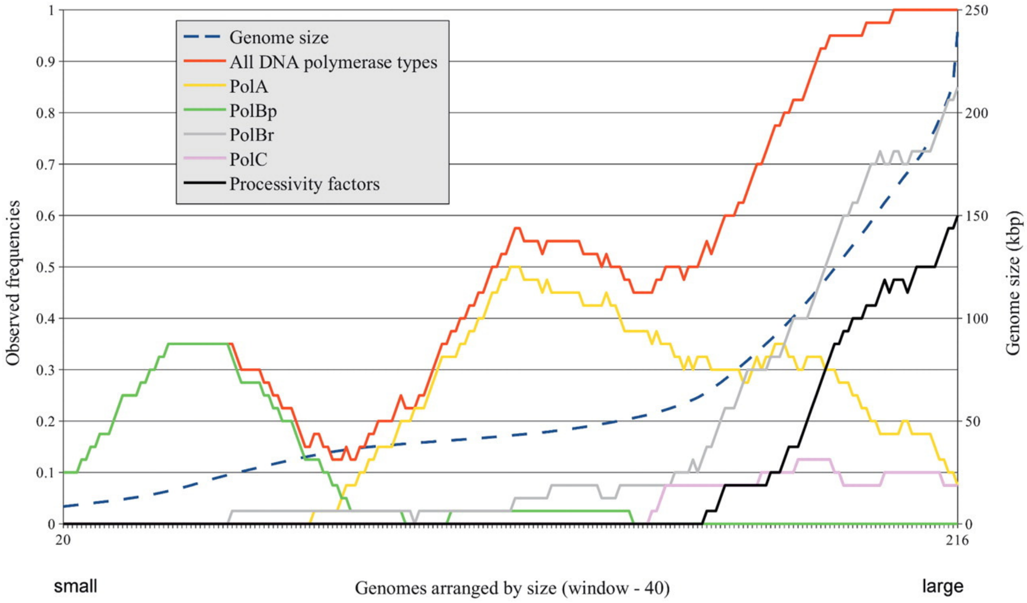


Fig. 2.2 Dependence between the observed frequencies of viral DNA replicase components and the genome size of dsDNA viruses. X-axis—genomes arranged by their size (from smallest to largest); major y-axis (left)—observed frequencies of various DNA replicase components in viral genomes; minor y-axis (right)—genome size (kbp). The genome size and the observed frequencies of DNA replicase components were averaged using the moving window of 40 genomes and a single-genome step. Broken blue line corresponds to the averaged genome size. Solid lines correspond to averaged observed frequencies of individual DNA replicase components.

2.2. Analysis of DNA replicases

Having established a general dependency of the presence and the type of viral DNA replicase components on the genome size (Figs. 2.1 and 2.2), we were nonetheless puzzled by the substantial number of seeming exceptions. While DNA polymerases are present in all taxonomic groups above the certain genome size, processivity factors and clamp loaders are not. If we assume that DNA replicase processivity properties become more important as the genome size increases, how to rationalize the absence of DNA sliding clamps and clamp loaders in some taxons with the large average genome size? To address this question, we performed a detailed analysis of sequence and structure properties of DNA polymerases, sliding clamp homologs and clamp loader subunits. Results of this analysis for each of the three components of DNA replicases are presented in separate sections below.

2.2.1. DNA polymerases

Major DNA polymerase groups. We identified DNA polymerases in 415 out of the 808 analyzed genomes of dsDNA viruses. The majority of DNA polymerases (255 genomes) belong to B-, less frequently (132) to A-, and very rarely (28) to the C-family. No polymerases of the archaeal D-family were detected. B-family polymerases are present in viruses that infect organisms from all three domains of life. In contrast, we found

A- and C-family polymerases only in bacteriophage genomes. The greatest diversity by far is among B-family members (Fig. 2.3, PolB), followed by the distantly related A-family (Fig. 2.3). Based on sequence similarity, PolB polymerases can be divided into three distinct clusters: one including protein-primed (PolBp), and two that include RNA-primed (PolBr) polymerases (Fig. 2.3) The small PolBr cluster consists of highly divergent PolBrPhiKZ polymerases identified in this study for the first time (Fig. 2.3).

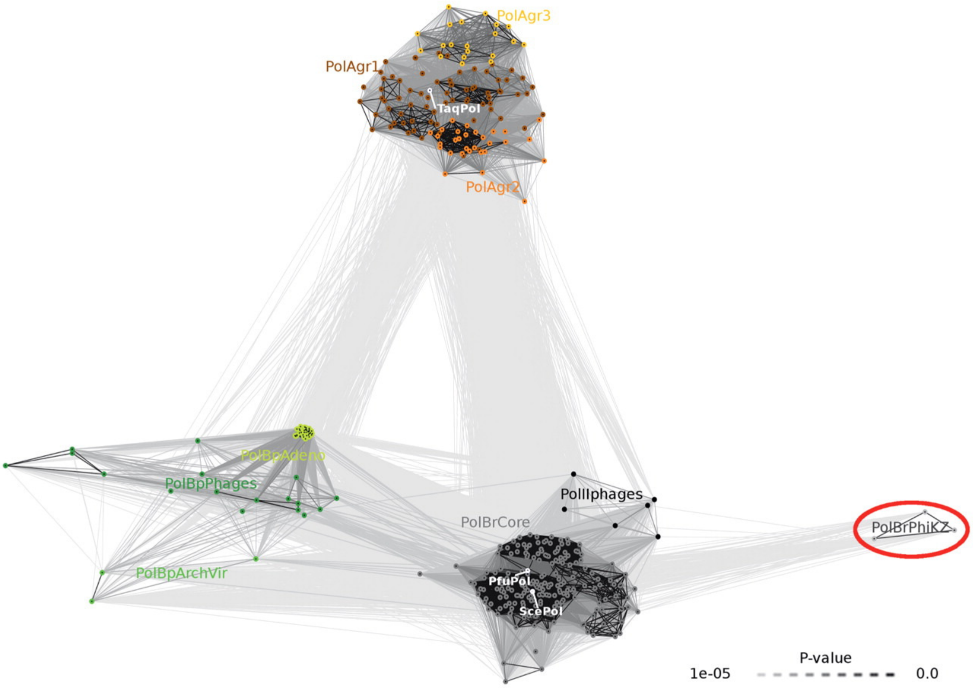


Fig. 2.3 DNA polymerases of A- and B-families clustered by the pairwise sequence similarity. Nodes represent individual sequences. Lines connect sequences with P≤1e-05. Line shading corresponds to P-values according to the scale in the bottom-right corner (light and long lines connect distantly related sequences). A-family DNA polymerases are represented using shades of orange, PolBp—shades of green, PolBr—shades of gray; well-known cellular DNA polymerases are shown in white. Newly identified DNA polymerases are marked with the red ellipse. ArchVir, archaeal viruses; Adeno, *Adenoviridae*; gr, group; PhiKZ, phiKZ-like phages; Pfu, *Pyrococcus furiosus*; Sce, *Saccharomyces cerevisiae*; Taq, *Thermus aquaticus*.

PhiKZ-like viruses have a genome that is almost twice as large as that of T4 phage (e.g. *Pseudomonas phage 201phi2* — 317 kb, T4 — 169 kb), yet no DNA polymerases were found in their genome sequences during previous analyses (31-33). Since our initial data suggested that the absence of a polymerase gene in viral genomes of this size is highly unlikely, we performed a particularly thorough analysis of the genomes of PhiKZ-like phages. Not surprisingly, standard homology detection methods (BLAST, RPS-BLAST and PSI-BLAST) failed to detect statistically significant similarity between predicted proteins of these phages and any known polymerases. Only when we applied very sensitive homology search methods based on profile-profile comparison, we were able to identify putative polymerases. Thus, HHsearch matched *Pseudomonas phage EL* hypothetical protein (gi: 82700954) and the RB69 (T4-like) phage DNA polymerase gp43 with high statistical significance (89% probability). COMA for the same phage EL protein also identified a B-family DNA polymerase (from *Thermococcus sp.*) as the best match (E=4e-07). The putative EL polymerase and its homologs in the other two phiKZ-like phages apparently include all the polymerase domains characteristic of gp43 except for the N-terminal region, which harbors the 3'–5' exonuclease domain. Interestingly, the 3'–5' exonuclease domain in these phages has been detected previously as a separate ORF (33). Thus, 3'–5' exonuclease and polymerase activities in these phages appear to reside in two separate polypeptide chains (Fig. 2.4). To further validate the polymerase assignment we analyzed the motifs, essential for the DNA polymerase function. Both sequence motifs harboring active site residues are conserved between RB69 gp43 and predicted polymerases in all three phiKZ-like phages (Fig. 2.4, B). In particular, as illustrated with a 3D model of the predicted EL polymerase active site, both aspartates (Fig. 2.4, C) involved in the coordination of metal ions are absolutely conserved.

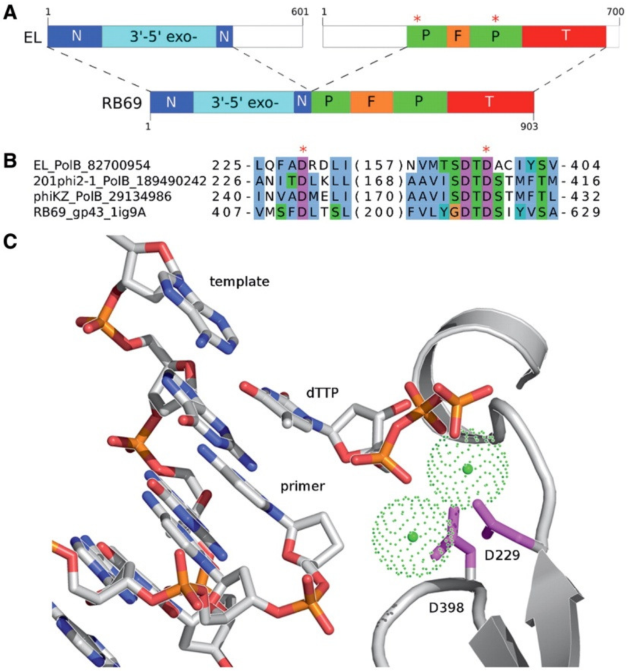

A
EL
1
601
1
700
N
3'-5' exo-
N
P
F
P
T
RB69
N
3'-5' exo-
N
P
F
P
T
1
903
B
EL_PolB_82700954 225-LQFADRDLI(157)NVMTSDTDACIYSV-404
201phi2-1_PolB_189490242 226-ANITDLKLL(168)AAVISDTDSTMFTM-416
phiKZ_PolB_29134986 240-INVADMELI(170)AAVISDTDSTMFTL-432
RB69_gp43_1ig9A 407-VMSFDLTSL(200)FVLYGDTDSIYVSA-629
C
template
dTTP
primer
D229
D398

Fig. 2.4 Comparison of DNA polymerases from phiKZ-like phages and the RB69 phage. (A) Correspondence of structural domains in *Pseudomonas phage EL* 3'–5' exonuclease (gi: 82700984) and DNA polymerase with those in the RB69 DNA polymerase. N, N-terminal; P, palm; F, fingers; T, thumb. Red stars indicate positions of the active site aspartates (D229 and D398). The correspondence was derived using COMA server. (B) Alignment of the DNA polymerase active site motifs. Sequence labels consist of the phage acronym, the protein name, and the gi number (PDB code in the case of RB69). (C) A 3D model of the *Pseudomonas phage EL* DNA polymerase active site complexed with the primed DNA and the incoming dTTP based on the ternary complex of the RB69 DNA polymerase and the DNA (PDB code: 1ig9). A fragment of the polymerase active site is shown in cartoon representation. Side chains of the active site aspartates coordinating two metal ions (green spheres) are shown as pink sticks.

A-family DNA polymerases could be subdivided into three groups. The most diverse group, PolAgr1, contains phages such as phiKMV, L5, N4, T5, SPO1, RSL1 and Ma-LMM01. Interestingly, the SPO1 DNA polymerase has the additional uracil-DNA glycosylase (UDG) domain at its N-terminus. It has been hypothesized that the UDG domain may serve as the intrinsic polymerase processivity factor (34). According to our analysis, the T5 DNA polymerase, which is highly processive (35), also has the UDG domain-like extension at the N-terminus. Taking into account that UDG (D4) in complex with A20 confers DNA polymerase processivity in eukaryotic vaccinia virus (36), the role of the UDG domain as the intrinsic polymerase processivity factor is quite likely. Groups 2 and 3 consist of T7-like and Bcep1-like viruses respectively.

Viral C-family DNA polymerases have domain organization similar to that of *E. coli* polIIIα (37). The conservation extends from the N-terminal PHP domain and includes the polymerase active site as well as the "fingers" domain. However, the C-terminal region following the "fingers" domain does not show significant similarity to the *E. coli* replicative polymerase suggesting that it may include different structural domains. Only the DNA polymerase from *Bacillus phage 0305phi8-36* (gi: 154622917) appears to extend sequence conservation past the "fingers" domain and into the OB-domain. In addition, this polymerase has a sequence motif (1131-EEDLL-1135) that aligns to the polIIIβ interaction motif in *E. coli* polIIIα (920-QADMF-924) suggesting that it may utilize a DNA sliding clamp to achieve the processivity. Incidentally, the *Bacillus phage 0305phi8-36* has the largest genome of those found to carry a C-family polymerase, and the only one among them in which we found a polIIIβ homolog (gi: 154622720).

Distinct subgroups of RNA-primed B-family DNA polymerases. The application of a more stringent clustering procedure (using CLANS coupled with BLAST instead of PSI-BLAST) revealed a number of subgroups within the large PolBrCore cluster. Since most PolBrCore polymerases are present in viruses with fairly large genomes, we analyzed polymerase sequences from poorly characterized subgroups to obtain hints as to the possible DNA replication processivity mechanisms. Polymerases of T4-like phages and herpesviruses that utilize DNA sliding clamps as processivity factors are known to possess characteristic clamp-binding motifs at their C-termini (38). Therefore, we looked for the presence of any clamp-binding motifs in all remaining subgroups. We readily identified a putative PCNA-interacting motif (the consensus sequence QxxIxxFF, where x is any amino acid) within the C-terminus of phycodnaviral DNA polymerases. In other subgroups we either did not find any clamp-binding motifs, the alignments of C-terminal regions were too variable or the number of sequences was too small to make a definite conclusion. In addition to clamp-binding motifs we looked for the presence of additional domains. It turned out that the members of three outlying subgroups (*Malacoherpesviridae*, *Alloherpesviridae* and *Nimaviridae* families) feature additional sequence regions compared with typical PolBrCore representatives. Although we were unable to confidently assign any known functional/structural domains to these additional polymerase regions, their very presence suggests that these three viral families may have evolved alternative processivity mechanisms for the efficient replication of their large genomes.

2.2.2. Processivity factors

Diversity and taxonomic distribution. Similarly as in the case of DNA polymerases, we asked whether each of the analyzed viral genomes encodes a polymerase processivity factor. In particular, we looked for homologs of either cellular (PCNA and polIIIβ) or viral (gp45, UL42, UL44 and BMRF1) DNA sliding clamps. As a result, in addition to already characterized or annotated sliding clamps, we discovered two new putative processivity factors: a PCNA homolog in the family *Ascoviridae* and a polIIIβ homolog in the Ma-LMM01 phage. All sliding clamp homologs identified in viral genomes were pooled together with representatives of cellular sliding clamps (PCNA and polIIIβ) and clustered. The results shown in Fig. 2.5 indicate that, just like DNA polymerases, viral DNA sliding clamp homologs are significantly more diverse than their cellular counterparts. Two major clusters correspond to PCNA and polIIIβ families. PolIIIβ homologs were found only in phages, while all PCNA homologs (except for

PCNA from the archaeal virus *Natrialba phage PhiCh1* and some baculoviruses) were found in eukaryote-infecting nucleo-cytoplasmic large DNA viruses (Fig. 2.1). PCNA homologs from iridoviruses infecting cold-blooded vertebrates form a distinct subgroup in the PCNA cluster (Fig. 2.5, CBvertIrido). In addition to two major clusters corresponding to PCNA and polIIIβ families, there are two compact outlying groups: gp45 and UL42. Gp45 includes DNA sliding clamps from T4-like phages, UL42 is found in *Herpesviridae*, both groups having structurally characterized representatives (39,40). Three additional divergent families of viral sliding clamps (UL44, BMRF1 and G8R) are not included in Fig. 2.5 as the clustering procedure was unable to link these families and any other clamps. However, it is known that herpesviral UL44 and BMRF1 are structurally similar to UL42 and other DNA sliding clamps (41,42). G8R is a remote PCNA homolog (43) found in vaccinia virus and other members of the *Chordopoxvirinae* subfamily, however, it does not act as a processivity factor in DNA replication (44).

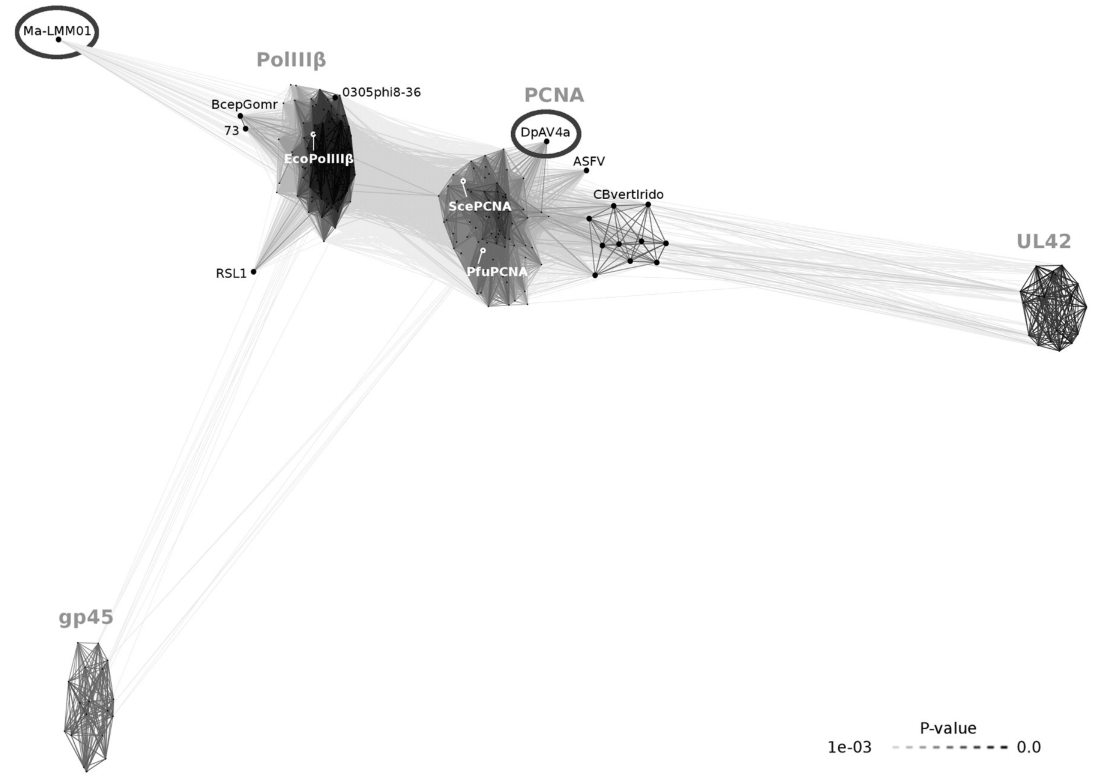


Fig. 2.5 DNA sliding clamps and their homologs grouped by the pairwise sequence similarity. Newly identified sliding clamp homologs are marked with ellipses. Ma-LMM01, *Microcystis phage Ma-LMM01*; RSL1, *Ralstonia phage RSL1*; 73, *Pseudomonas phage 73*; BcepGomr, *Burkholderia phage BcepGomr*; 0305phi8-36, *Bacillus phage 0305phi8-36*; Eco, *Escherichia coli*; ASFV, *African swine fever virus*; DpAV4a, *Diadromus pulchellus ascovirus 4a*; CBvertIrido, cold-blooded vertebrate animal iridoviruses.

We detected polIIIβ homologs in only twelve phages. Of the 12 polIIIβ homologs, seven have a typical length and five are shorter, covering only the second and third domains of polIIIβ. A full-length distant polIIIβ homolog in Ma-LMM01 phage was identified (the HHsearch probability of 96%) for the first time. The Ma-LMM01 polIIIβ is coded (locus tag: MaLMM01_gp176) near other DNA replication proteins (45), supporting its putative processivity factor function.

A number of the identified viral sliding clamp homologs may have been acquired through the horizontal gene transfer (patchy taxonomic distribution, high similarity to corresponding host proteins, the absence of a DNA polymerase in the viral genome). For example, only nine out of 53 baculoviruses have PCNA homologs, and seven of those show high similarity to PCNAs from mosquitoes and moths. For one of baculoviruses, *Autographa californica nucleopolyhedrovirus*, it has been shown that its own PCNA is not required for genome replication (46). As polIIIβ and PCNA homologs, likely acquired through horizontal gene transfer, are either known or can be assumed to be dispensable for DNA replication, we did not include them in the summary presented in Figs. 2.1 and 2.2.

Unexpectedly, we did not find homologs of any known processivity factors in some viral families with the large average genome size. These include eukaryotic *Nimaviridae*, *Alloherpesviridae*, and *Malacoherpesviridae* families as well as phiKZ-like phages and *Clostridium phage c-st* (Fig. 2.1). However, as discussed in the "Polymerases" section, DNA polymerases of the three eukaryotic viral families are atypical B-family members with additional uncharacterized domains. The *Clostridium phage c-st* DNA polymerase is one of the C-family polymerases having a divergent C-terminal region. These observations suggest that viruses from these families may use different mechanisms to ensure DNA replication processivity. In the case of PhiKZ-like phages, whether or not processivity factors are indeed absent from their genomes remains an open question.

Electrostatic properties. DNA sliding clamp distribution in viral genomes (Fig. 2.1) shows that *Bacillus phage 0305phi8-36* and several families of eukaryotic viruses carrying correspondingly polIIIβ and PCNA genes in their genomes totally lack clamp loader subunits. Since a clamp loader is needed to open and load ring-shaped polIIIβ or PCNA onto DNA, this finding raised a question as to how these sliding clamps may function. One possibility is that these viruses use a clamp loader of the host. Another possibility is that these clamps do not form a closed ring and, similarly to UL42 or UL44, bind DNA directly without the need for a clamp loader. While the first possibility cannot be explored using computational approaches, the second one can.

One of the observed differences between non-ring sliding clamps (e.g. UL42, UL44) and the ring-forming ones (PCNA, polIIIβ) is that the former have an increased positive charge located on the DNA-binding face (47,48). To explore the electrostatic properties of all the identified viral sliding clamp homologs, we calculated their theoretical pIs. In addition, we constructed 3D models for representatives of viral PCNA homologs and analyzed electrostatic properties of their surfaces. The obtained data was then compared to structurally and functionally characterized cellular and viral processivity factors (Fig. 2.6). It turned out that pIs of sliding clamp homologs show a striking correlation with the presence/absence of clamp loader subunits in corresponding viral families. Thus, *Phycodnaviridae* and Mimivirus PCNAs, predicted to be orthologous, have electrostatic properties similar to ring-shaped sliding clamps. In contrast, electrostatic properties of G8R and PCNAs of *Asfarviridae* (ASFV), Irido-Asco viruses and Marseillevirus are more similar to herpesviral non-ring processivity factors. *Phycodnaviridae* and Mimivirus have RFC homologs, while *Asfarviridae*, Irido-Asco viruses and Marseillevirus do not. A similar correlation is observed for sliding clamp homologs in bacteriophages. PolIIIβ homologs in phages Ma-LMM01 and RSL1 (Fig. 2.6, PolIIIβ vir1) show much lower pI values than polIIIβ in *Bacillus phage 0305phi8-36* (PolIIIβ vir2). Phages Ma-LMM01 and RSL1 do encode clamp loader subunits, while Bacillus phage 0305phi8-36 does not. Hence, based on the electrostatic properties, DNA sliding clamp homologs from *Phycodnaviridae* and *Mimiviridae* are expected to form rings, while PCNA homologs in the remaining families and polIIIβ from the *Bacillus phage 0305phi8-36* are likely to bind the DNA directly, in a manner that does not require clamp loaders. According to pI values, PolIIIβ homologs of Ma-LMM01 and RSL1 phages are at the intermediate position between the characterized ring-forming and non-ring sliding clamps. However, the presence of clamp loader subunits (polIIIγ) in the corresponding genomes suggests that the closed-ring polIIIβ structure is more likely.

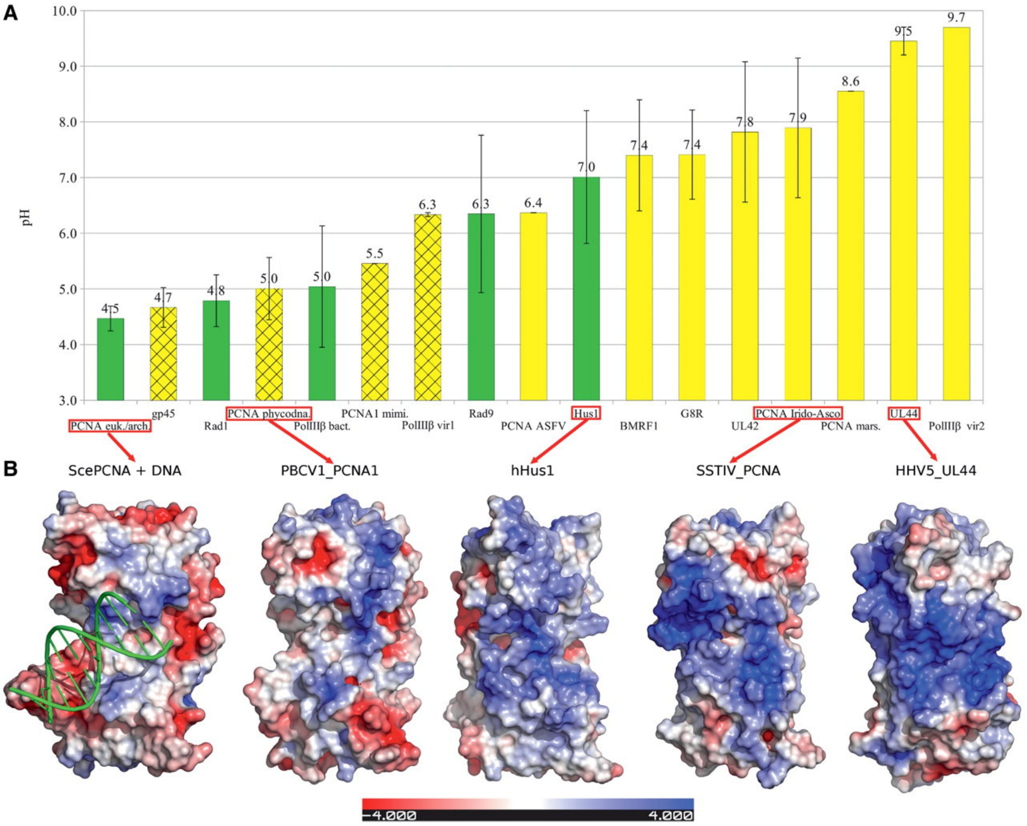


Fig. 2.6 Electrostatic properties of processivity factors and their homologs. (A) Average theoretical pIs of DNA sliding clamp subunits from cellular organisms (green bars) and viruses (yellow bars). Bars with the grid pattern correspond to viral sliding clamp homologs that are accompanied by clamp loader subunits in the genome. (B) Electrostatic potential maps of solvent accessible surface of five representatives (red color indicates negative, blue—positive potential; scale units—K_bT/e_c). All structures are shown in the same orientation as the ScePCNA complexed with DNA (PDB code: 3k4x). arch., *Archaea*; asco., — *Ascoviridae*; ASFV, *African swine fever virus*; euk., *Eukarya*; hHus1, *Homo sapiens* Hus1 (PDB code: 3g65), HHV5_UL44, *Human Herpesvirus 5* UL44 (PDB code: 1t6l); irido., *Iridoviridae*; PCNA mars., Marseillevirus PCNA (gi: 284504238); PCNA mim., Mimivirus PCNA (gi: 55664866); PBCV1_PCNA1, *Paramecium bursaria Chlorella Virus-1* PCNA1 (gi: 9631761); Sce, *S. cerevisiae*; SSTIV_PCNA, *Soft-shelled turtle iridovirus* PCNA (gi: 228861299); PolIIIβvir1, polIIIβ from *Microcystis phage Ma-LMM01* and *Ralstonia phage RSL1* (gi respectively: 117530347, 189233246); PolIIIβvir2, polIIIβ from *Bacillus phage 0305phi8-36* (gi: 154622720).

2.2.3. Clamp loaders

Compared to DNA polymerases and sliding clamps, homologs of clamp loader subunits are present in the fewest number of viral genomes. However, their genomic distribution appears to be highly non-random. We detected clamp loader subunits only in viruses with the largest genomes and only in those that also code for homologs of DNA sliding clamps. Moreover, as indicated above, the presence of clamp loader subunits correlates with electrostatic properties of DNA sliding clamps in the corresponding viral families. Hence, we found homologs of RFC subunits only in Mimivirus and *Phycodnaviridae*, the only two families that have PCNAs with electrostatic properties similar to those of ring-forming cellular PCNAs (Figs. 2.1 and 2.6). Mimivirus and its relative CroV code all five RFC subunits. Members of *Phycodnaviridae* family have only the largest RFC subunit homolog, similar to the archaeal large RFC subunit (RFCL). The exceptions include EsV-1, which encodes all five RFC subunits, and two other viruses (*Ostreococcus virus OsV5* and *Ostreococcus tauri virus 1*) that do not have any RFC subunit. Interestingly, the genomes of the latter two viruses are among the smallest in the family. Homologs of bacterial clamp loader subunits were identified in only two phages, RSL1 and Ma-LM001. In each case we found only a homolog of a single clamp loader subunit, polIIIγ. Both polIIIγ homologs have conserved P-loop, DEXX and SRC motifs (Fig. 2.7) suggesting that they are active ATPases. Again, polIIIβ homologs in these two phages have significantly lower pIs than polIIIβ in *Bacillus phage 0305phi8-36*, lacking any clamp loader subunit (Fig. 2.6). T4-like clamp loaders consisting of gp44 and gp62 subunits were identified only in T4-like phages.

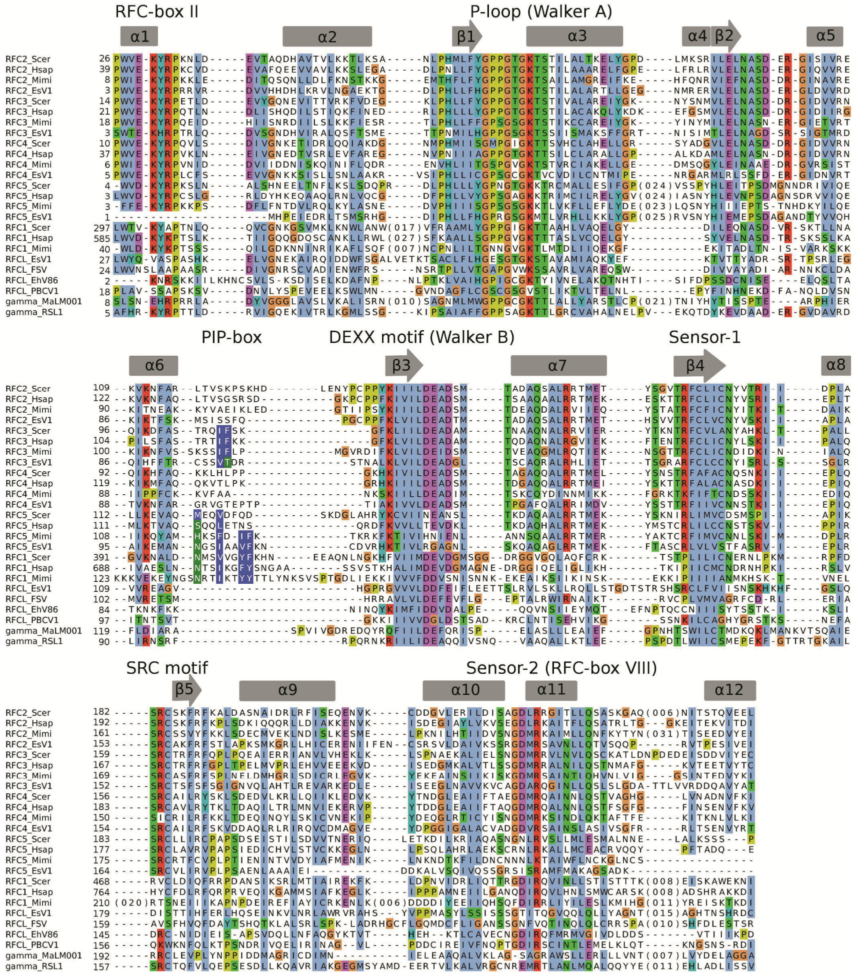


Fig. 2.7 Alignment of eukaryotic and viral clamp loader subunits. Sequence alignment is based on multiple structure superposition of experimental X-ray structures and homology models obtained using MUSTANG (49). Secondary structure of the yeast RFC3 subunit (PDB code: 1sxj) is shown above the alignment. EsV1, *Ectocarpus siliculosus virus 1*; EhV86, *Emiliania huxleyi virus 86*; FSV, *Feldmania species virus*; Mimi, Mimivirus.

2.3. Viral DNA replicases and genome size: relationship and its implications

Our results show that the presence and the nature of DNA replicases encoded in the genomes of dsDNA viruses is related to the genome size. This relationship can be defined as the tendency to encode polymerase processivity components in addition to the DNA polymerase more often as the genome size increases.

Viruses having genomes smaller than ~40 kb most often do not have their own DNA polymerases. However, if they do, it is usually a PolBp type DNA polymerase. Interestingly, this is seen in viruses infecting organisms from all domains of life. Coupled with the observation that PolBp polymerases disappear completely from larger viral genomes (Figs. 2.1 and 2.2), this suggests that properties of protein-primed B-family DNA polymerases might be optimal for this genome size range.

As the genome size increases (~40–140 kb) A-family polymerases take over. However, it is not clear whether the dominance of A-family polymerases in this genome size range is significant. The reason is that we detected A-family polymerases only in bacteriophages, and this particular size range is overrepresented with bacteriophage genomes. Nonetheless, even if we ignore the polymerase type, the typical feature of genomes in this size range is the lack of DNA sliding clamp homologs. It has been shown that *E. coli* polymerase I (A-family) is stimulated by the polIIIβ clamp (50). Therefore, the absence of sliding clamp homologs cannot be explained by the inability of polA to utilize sliding clamp as a processivity factor. Moreover, in two phages (Ma-LMM01 and RSL1) with large genomes (>150 kb) we detected an A-family polymerase, a polIIIβ homolog and a clamp loader subunit suggesting that the polIIIβ homolog may function as a processivity factor together with polA. On the other hand, some bacteriophages have evolved the increased processivity of A-family polymerases without using DNA sliding clamps. One such solution is the recruitment of thioredoxin from the host as observed in T7-like phages (51). The UDG-like domain in DNA polymerases of SPO1-like and T5-like phages may well be another solution, which is yet to be addressed experimentally.

The genome size range of 140 kb and larger is represented by eukaryotic viruses and bacteriophages. They all have their own DNA polymerases, typically of B-family. Our discovery of evolutionary distant DNA polymerases in phiKZ-like phages has eliminated the only seeming exception to this rule. DNA replicases in this size range often include DNA polymerase processivity factors and sometimes clamp loaders. Initially, there does not seem to be any discernible pattern as to the presence or absence of sliding clamp homologs and clamp loaders (Fig. 2.1). However, if we consider properties of DNA polymerases, homologs of sliding clamps and the presence of clamp loader subunits we get a fairly coherent picture.

Thus, we did not find any sliding clamp homologs in several groups of large dsDNA viruses. However, their DNA polymerases either have additional uncharacterized domains or non-homologous regions. It may be that these polymerases either possess an increased intrinsic processivity due to these additional/altered regions or use alternative processivity factors. On the other hand, the fact that we did not find any sliding clamp homolog in phiKZ-like phages is somewhat puzzling. Their polymerases, although evolutionary distinct, seem to possess a typical B-family architecture. In addition, two of the three polymerases at their C-termini feature a putative signature of a clamp-binding motif. It is quite possible that processivity factors are encoded in genomes of phiKZ-phages, but are too strongly diverged to be detected with current methods.

As it comes to the viral families that do have homologs of DNA sliding clamps, the intriguing finding was that a number of these families completely lack clamp loader subunits. However, the subsequent analysis of electrostatic properties of sliding clamp homologs was quite revealing. It showed that PCNA homologs from Irido-Asco, Asfarviruses and Marseillevirus as well as a polIIIβ homolog from *Bacillus phage 0305phi8-36*, all have elevated pIs (Fig. 2.6, A). Models of several representatives showed that most of the increased positive charge is localized to the DNA-interacting face (Fig. 2.6, B). This property is typical for well-characterized herpesviral processivity factors. They do not form rings; instead they bind DNA directly as monomers (UL42) or dimers (UL44). This suggests a similar direct DNA-binding mode for the sliding clamp homologs with the elevated pI and without clamp loaders in corresponding genomes.

Findings concerning viral clamp loaders are perhaps most puzzling compared to other replicase components. Only three eukaryotic viruses have a complete set of five RFC subunits corresponding to the eukaryotic clamp loader, RFC. As expected for functional RFC, all three viruses have characteristic P-loop and DEXX motifs in RFC1-4 subunits and also feature PCNA-interacting (PIP-box) motifs in RFC1, RFC3 and RFC5. Several members of *Phycodnaviridae* family have only a single homolog of the RFC large subunit. From studies with human and yeast RFC it is known that the RFC large subunit determines the specificity for the clamp (52). For example, RFC1 determines specificity for PCNA, while Rad17—for the 9-1-1 complex. Thus, it may be that the viral homolog of the large RFC subunit recruits four small RFC subunits of the host to form a pentameric complex specific for binding and loading viral PCNA. However, these RFC large subunits seem to completely lack PCNA-binding motifs and some have non-canonical ATPase motifs. It has been shown that the mutation in the ATP-binding motif of the large RFC subunit in yeast does not affect PCNA loading (53). Therefore, the ATPase activity may also be dispensable in viral RFC large subunits. It is not clear, though, how to reconcile the absence of a PCNA-binding motif with the expected specificity for the viral PCNA. Two large phages, Ma-LMM01 and RSL1, that have a bacterial clamp loading subunit homolog, polIIIγ, additionally have an A-family DNA polymerase and a homolog of polIIIβ sliding clamp. In these two cases it is also not clear what is the composition of the functional replicase. Does the viral polIIIγ recruit host clamp loader subunit(s) to produce a functional clamp loader specific for the viral polIIIβ? Or perhaps the composition of these clamp loaders is analogous to the T4 clamp loader, which is made of four copies of gp44 (polIIIγ homolog) and a single taxon-specific subunit gp62 (no detectable homologs outside the T4-like group)? To address these questions, computational methods can hardly substitute laboratory experiments.

Overall, our observed connection between the virus genome size and DNA replicase components might help in predicting the expected type and completeness of replicase components for newly sequenced viral genomes. In addition, our observations for DNA replicases in dsDNA viruses perhaps may have a more general significance. For example, symbiotic bacteria belonging to genus *Hodgkinia* and *Carsonella* have some of the smallest known cellular genomes of 144 and 160 kb size, respectively (54). It turns out that neither has a DNA sliding clamp or a clamp loader. However, somewhat larger genomes of symbionts *Sulcia cicada* (277 kb), *Buchnera Cc* (416 kb) and *Nanoarchaeum equitans* (491 kb) already have the complete set of DNA replicase components. With more large viral and small cellular genomes available, it will be interesting to see how universal the observed relationship is.

2.4. DNA replication proteins and virus genome size

In the first part of the study we have shown that the composition of DNA replicases in dsDNA viruses is genome size-dependent. Next, we asked whether other replication proteins also show genome size dependency. To answer this question, in addition to DNA replicase proteins, we identified viral replicative helicases, primases, single-stranded DNA-binding (SSB) proteins, ligases, primer removal proteins and topoisomerases. Our results regarding viral DNA replicases revealed that dsDNA viruses can be divided into three genome size groups: (<40 kbp, 40–140 kbp and >140 kbp) based on the type of DNA replicase. Thus, for each genome size group we counted frequencies of occurrence of replication proteins (Fig. 2.8). It turned out that as genome size increases dsDNA viruses tend to code their own replication proteins more often. The strongest genome size dependency is seen in DNA topoisomerases, DNA polymerases, sliding clamps and their loaders, primer removal proteins and ligases (Fig. 2.8). Our data show that SSB proteins have the weakest genome size dependency (Fig. 2.8). It is surprising because SSB proteins are found in all well characterized DNA replication machineries (55). Therefore we hypothesize that at least some SSB proteins may have escaped detection due to their high divergence rates and non-orthologous displacement of canonical OB-fold SSBs. This hypothesis is supported by our discovery of SSB proteins in nucleo-cytoplasmic large DNA viruses and characterization of non-canonical SSBs in poxviruses (see the section "Single-stranded DNA-binding proteins").

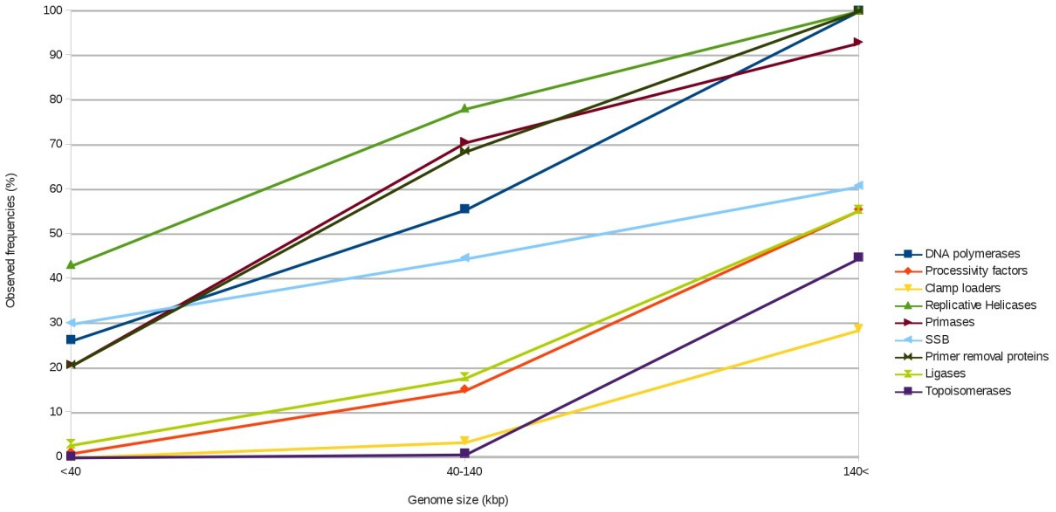


Fig. 2.8 Genome size groups of DsDNA viruses and observed frequencies of their DNA replication proteins.

2.5. DNA replication proteins and the taxonomy of virus host

It is known that bacterial DNA replication proteins are evolutionary distinct from those found in archaea and eukaryotes (56). Studies show that replication genes can be transferred horizontally between viruses and cells (57). Do viruses, infecting three domains of life, also have domain-specific DNA replication genes? Or perhaps, there are some "universal" viral replication gene sets?

To check for taxonomic dependency we divided viruses into three groups: bacterial (71%), archaeal (9%) and eukaryotic (20%). Bacteriophages make the largest portion of all viruses and are the only ones who have bacterial replication genes: DNA polymerases from A and C families, PolIIIβ and its loader, DnaB helicases, DnaG primases and SSBs (Fig. 2.9). Replication genes specific to archaea and eukaryotes (PCNA and RFC) are only observed in their viruses. Other replication proteins are found in viruses infecting organisms from at least two domains of life. For example, SF3 helicases, archaeo-eukaryotic primases (AEP) and B-family DNA polymerases are present in bacterial, archaeal and eukaryotic viruses (Fig. 2.9). Cellular organisms do not use SF3 family helicases for replication. Thus, it is likely that SF3 replicative helicases were not acquired from the host, but were already present in viruses infecting last universal common ancestor (LUCA). To sum up, viruses often have DNA replication proteins which are also found in their host, however, their type and presence/absence pattern is more dependent on the virus genome size than on the taxonomy of the host.

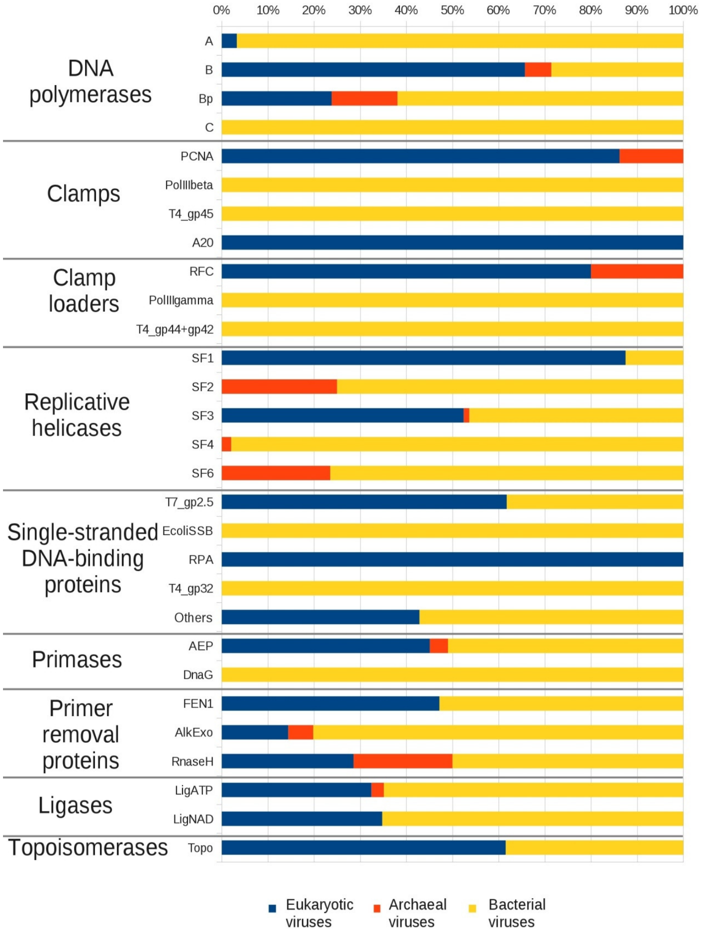


Fig. 2.9 DNA replication proteins and the taxonomy of virus host.

2.6. Replicative DNA helicases

Analysis of viral replicative helicases revealed that they are the most common replication proteins in dsDNA viruses. The study has shown that bacteriophages usually have homologs of bacterial replicative helicase (DnaB). Helicases from superfamilies SF3 and SF6 are found less often, however, they are more diverse. Herpesviruses possess SF1 helicases and have the best studied DNA replication machinery among eukaryotic viruses. However, their helicase-primase complex was still enigmatic. Here, we shed some light on this complex by finding out that one of its components (UL8) is inactivated B-family polymerase.

2.6.1. Replicative DNA helicase – most common replication protein in dsDNA viruses

Intuitively, it was expected to find DNA polymerases as the most common replication protein. However, after identification of replication proteins, it turned out that 70% of viruses have a replicative helicase and thus, it appeared to be the most frequent viral replication protein. What is the possible explanation for this phenomenon? Larger viruses (≥40 kbp) may have replicative helicases more often due to already described tendency to code their own replication proteins more frequently as the genome size increases. However, high observed frequencies of replicative helicases in small viruses (<40 kbp) is puzzling. Maybe, small viruses use replicative helicase for replication initiation and taking control of host replication proteins? For example, the latter strategy is used by small papilloma and polyoma viruses (58). We decided to analyze proteomes and literature data of viruses in which replicative helicases were not found. Part of those viruses use protein-primed DNA replication, which does not need a DNA helicase (59). The largest fraction of remaining viruses, without their own replicative helicase, have helicase loaders which are probably used for loading of host replicative helicase. After subtraction of viruses which use protein-primed DNA replication and addition of those that had helicase loaders, percentage of viruses, which have their own or can use the replicative helicase of the host, increases from 70% to 90%. Thus, the data show that the strategy to initiate genome replication using DNA helicase is very common and not limited to small eukaryotic viruses.

2.6.2. Herpesviral helicase-primase subunit UL8 is inactivated B-family polymerase

Herpesviral helicase-primase complex consists of subunits UL5, UL52 and UL8. UL5 is a superfamily I DNA helicase (60) related to yeast Pif1 and bacterial RecD (61). UL52 is a primase belonging to the superfamily of archaeo-eukaryotic primases (62). UL8 does not have any known catalytic activity (63). It appears to be important for the nuclear import of UL5 and UL52 and is known to interact not only with the UL5–UL52 subcomplex, but also with UL30, ICP8 and UL9 (64). Surprisingly, despite the essential role of UL8 in DNA replication, so far nothing is known about its structure and domain organization. Moreover, there are no known homologs of UL8 outside the herpesviruses (64).

Our systematic iterative sequence searches using Jackhmmer and UL8 proteins as queries provided an initial hint that the UL8 family, instead of being novel, might be related to B-family polymerases. For example, the search with *bovine herpes virus 2* UL8 (gi: 14161473) after four iterations produced a statistically significant match (*E*-value = 0.001) with the B-family polymerase (gi: 150401083) from *Methanococcus aeolicus*. The results of more sensitive profile–profile searches have further substantiated the initial finding. Thus, HHpred readily identified (95% probability) the relationship between HHV-1 UL8 and the B-family DNA polymerase from the archaeon *Thermococcus gorgonarius* (SCOP: d1tgoa2). This newly discovered relationship was also consistently supported by other sensitive homology detection and modeling servers (see methods). The identified similarity between UL8 and B-family polymerases is limited to the C-terminal half of UL8 sequences (~393-727 a.a. of HHV-1)(Fig. 2.10). The relationship is remote as the aligned region of UL8 proteins and B-family representatives shares only 9–12% sequence identity. Our attempts to identify any characterized homolog of the

UL8 N-terminal region were unsuccessful. This does not necessarily mean that the region harbors novel structural domain(s). B-family polymerases at their N-terminus typically have the exonuclease domain (65). Thus, it is conceivable that the UL8 N-terminal region represents an exonuclease-like domain diverged beyond recognition.

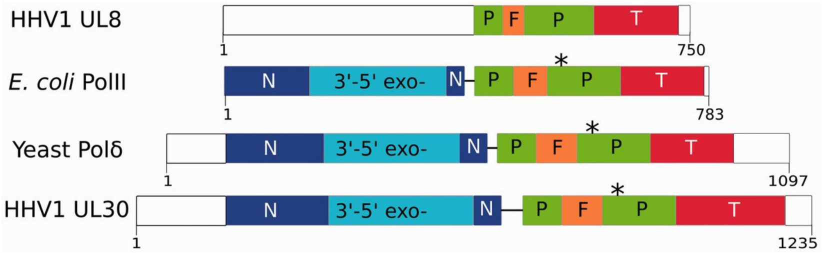


Fig. 2.10 Domain organization of HHV-1 UL8 and B-family DNA polymerases. N, N-terminal; P, palm; F, fingers; T, thumb. Asterisk indicates the polymerase active site ("DTD" motif).

To further corroborate the identified homology and to obtain insight into the structure of UL8 C-terminal region (UL8c), we constructed homology models for UL8c of HHV-1 (available at http://www.ibt.lt/bioinformatics/models/hhv1_ul8/) and its relatives from the Simplexvirus genus. Models were constructed using the structure of *Escherichia coli* PolII (PDB: 3k57) (66) as a template from the optimized sequence-structure alignments. To have reference points, we constructed computational models of several B-family polymerases using the same *E.coli* PolII structure as the modeling template. We chose those B-family polymerases, for which experimentally determined structures were available, and therefore, we could obtain the "ideal" (structure-based) alignment between them and PolII. We only selected B-family polymerases that were <30% identical to *E.coli* PolII, so as to make the situation more similar to that between UL8 and PolII. Thus, we chose PolB of phage RB69, yeast Pol δ, *Pyrococcus furiosus* PolB and HHV-1 DNA polymerase UL30 and generated models for them using structure-based alignments. Evaluation of models was performed with Prosa. Not surprisingly, Prosa *Z*-score values for UL8c models were worse than for the crystal structure of PolII used as a template. Nonetheless, some of UL8c models scored relatively high. Moreover, UL8c models of HHV-1 and some other herpesviruses scored better than all of the reference models, except for *P.furiosus* PolB. These results strongly suggest that UL8c and B-family polymerases are indeed structurally similar and imply that UL8c models are unlikely to contain serious flaws.

To understand the differences between UL8c and B-family polymerases, we performed a detailed analysis of sequence and structure motifs. A prominent feature of UL8c is the lack of the intact active site motif "DTD" (Fig. 2.10). Only the second aspartate from this motif is conserved in a number of α-herpesviruses. The fingers subdomain, which is important for recognition and binding of the incoming nucleotide (65), is reduced in UL8c. Furthermore, sequence region preceding fingers subdomain has a deletion in UL8c. In addition, UL8c lacks the "KKRY" motif, known to play an important role in stabilizing B-form of the DNA (67). Taken together, these features indicate that UL8 is not an active DNA polymerase, consistent with the failure to detect any enzymatic activity in UL8 using experimental approaches (63). Consistent with modifications of UL8c DNA-binding motifs, UL8 does not exhibit ssDNA or dsDNA binding on its own (68). However, UL8 appears to modulate DNA binding by UL5-UL52, the other two subunits of helicase–primase complex (69). Thus, sequence and structure analysis of

UL8 indicates that it lacks motifs necessary for polymerase activity and has undergone a reduction of DNA-binding motifs.

Protein–protein binding sites in UL8. UL8c structural models enabled us to look at known and putative binding sites mediating interactions with other proteins. Recently, systematic mutagenesis of charged residues into alanines has been carried out for HHV-1 UL8 and four replication-defective mutants were identified (70). Three of the mutants can be mapped onto the model of HHV-1 (Fig. 2.11). Mutants A49 (R640A, D642A and E644A) and A53 (R677A and R678A) displayed a defective interaction with UL52. The interaction between mutant A52 (D671A and E673A) and UL52 was only slightly impaired. In the HHV-1 UL8c model, these positions are close to positions substituted in mutant A53. Our model is in good agreement with these experimental findings, as all the positions affecting interaction with UL52 are on the same side of the model surface.

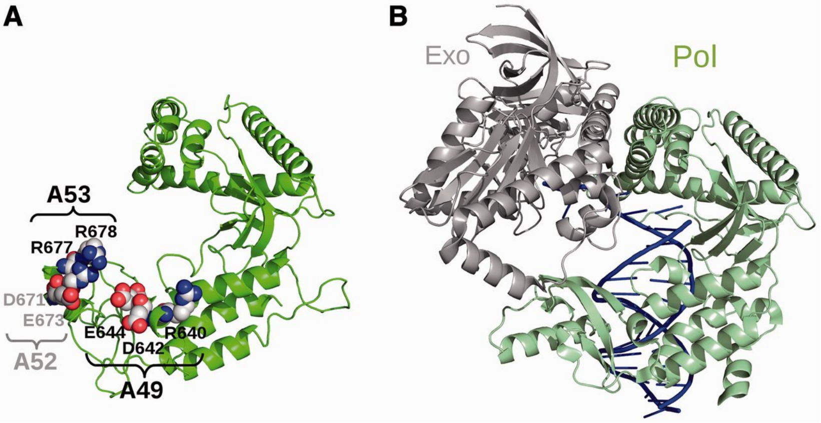


Fig. 2.11 A model of HHV-1 UL8c compared with *E.coli* PolII. (A) UL8c and its mutants that affect binding to UL52. Mutated residues are shown in space-filling representation (carbon atoms, gray/white; nitrogen, blue; oxygen, red) and are labeled. (B) Crystal structure of *E. coli* PolII complexed with DNA. Polymerase (Pol) and exonuclease (Exo) domains are colored in green and gray, respectively.

In addition, a putative protein-binding site is located within the very C-terminus of UL8, extending beyond the modeled structure. The C-terminal region features a short conserved hydrophobic motif (HHV-1 UL8 747-FLF-749) and is apparently disordered, as predicted by PrDOS (71) at 5% false positive rate and MetaDisorder3D (72). A conserved hydrophobic motif within the disordered region is often a signature of protein-binding site. It has been established that HHV-1 UL8 interacts with DNA polymerase, and indirectly the interaction site was mapped within the C-terminal region of UL8 (73). These results implicated UL8 residues, important for binding, just upstream of the conserved C-terminal motif. This suggests that the conserved hydrophobic motif is, perhaps, a secondary polymerase-binding site. Alternatively, it may mediate interaction with one of the other multiple binding partners of UL8.

UL52 and UL5 are related to eukaryotic PrimPol family and Pif1 helicases, respectively. Since UL8 functions as part of the helicase–primase complex, we decided also to look into homologous relationships of the other two components of the complex, UL52 and UL5. HHV-1 UL52 (gi: 9629434) is a 1058-residue long protein with the C-terminal half related to archaeo-eukaryotic primases (62). UL52 homologs, identified by sequence searches, fall into several clusters (Fig. 2.12). These include HHV-1 UL52 and other *Herpesviridae* sequences, UL52 homologs from mollusc and cold-blooded animal herpesviruses (*Malacoherpesviridae* and *Alloherpesviridae* families, respectively), A468R-like proteins from *Phycodnaviridae* family and a cluster of viral and eukaryotic homologs. The latter are exemplified by the recently characterized human PrimPol (CCDC111), which has both primase and polymerase activities and plays an important role in replication fork progression through sites of DNA damage (74,75). Human PrimPol is the first example of an eukaryotic protein harboring both primase and polymerase activities. Interestingly, in kinetoplastids this family of proteins has undergone lineage-specific expansion. For example, *Trypanosoma brucei* has four copies of PrimPol homologs. Two of the more distantly related ones to human PrimPol (TbPRI1 and TbPRI2) were shown to be active primases and to have roles essential for cell growth and kinetoplast DNA replication (76). The other two PrimPol-like proteins act as DNA polymerases with TbPPL1 also having a DNA primase activity (77). The PrimPol group includes viral proteins from *Marseilleviridae*, *Mimiviridae*, *Asfarviridae* and *Ostreid herpesvirus 1*. Surprisingly, the latter virus representing *Malacoherpesviridae* has three copies of primases. Two of them (OsHV1_ORF7 and OsHV1_ORF49) clustered with the PrimPol group, whereas the third one (OsHV1_ORF66) was more similar to herpesviral UL52 sequences.

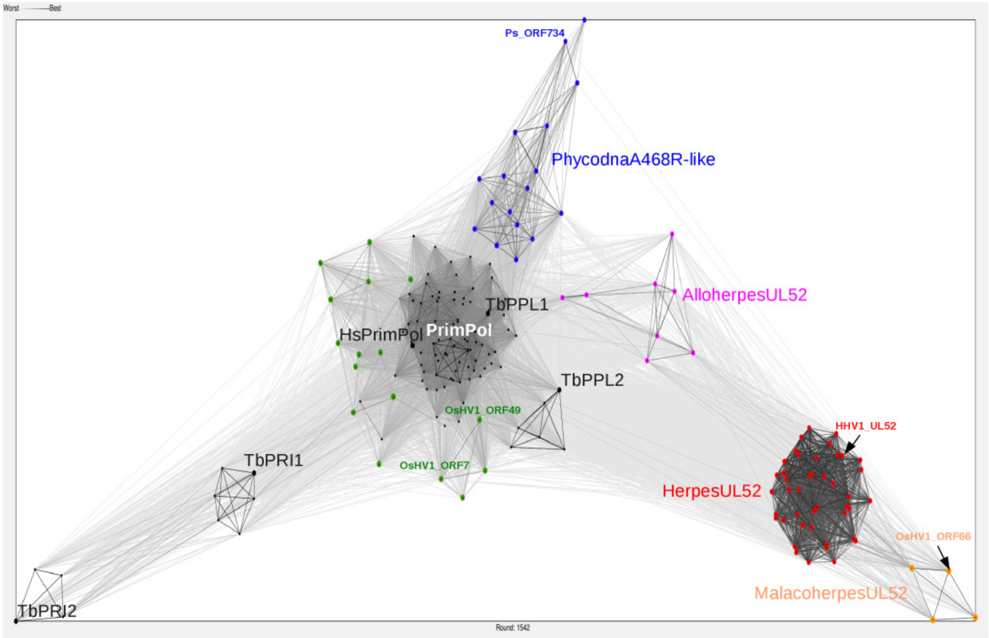


Fig. 2.12 UL52 proteins and their homologs clustered by pairwise sequence similarity. Lines represent sequence relationships (P-value≤1e-03). Protein sequences shown in green belong to the *Malacoherpesviridae* family and to the nucleo-cytoplasmic large DNA viruses. Abbreviations: Ps, *Pandoravirus salinus*; Tb, *Trypanosoma brucei*; Hs, *Homo sapiens*; HHV1, *Human herpesvirus 1*; OsHV1, Ostreid *herpesvirus 1*; Alloherpes, *Alloherpesviridae*; Herpes, *Herpesviridae*; Malacoherpes, *Malacoherpesviridae.*

UL5 is a member of the SF1 helicase superfamily having 5′-3′directionality. Using iterative sequence searches with HHV-1 UL5 (gi: 9629385), we readily detected similarity of UL5 to eukaryotic and viral Pif1 helicases as well as to more distantly related homologs of bacterial RecD and T4 Dda helicases. We also identified previously unannotated UL5 homologs in *Malacoherpesviridae* (Fig. 2.13). Clustering of sequence search results revealed that herpesviral UL5 are split into three groups mirroring the UL52 results. Eukaryotic Pif1 homologs belong to a group displaying the closest similarity to UL5. Pif1 is found in nearly all eukaryotes. Most eukaryotes including humans have a single Pif1 family helicase, but *Saccharomyces cerevisiae* has two (ScPif1 and ScRrm3). Interestingly, *T. brucei* has as many as eight Pif1 paralogs, mirroring the expansion of its PrimPol-like proteins. Yeast proteins represent some of the best characterized members of the Pif1 family. ScPif1 affects telomeric, ribosomal and mitochondrial DNA replication, as well as Okazaki fragment maturation (78). Recently, ScPif1 was also found participating together with Polδ in recombination-coupled DNA synthesis (79). ScRrm3 moves with the replication fork during the DNA replication and helps to pass difficult-to-replicate sites (80). ScRrm3 was found to interact with Polε, suggesting that it is a stable component of the replisome (80).

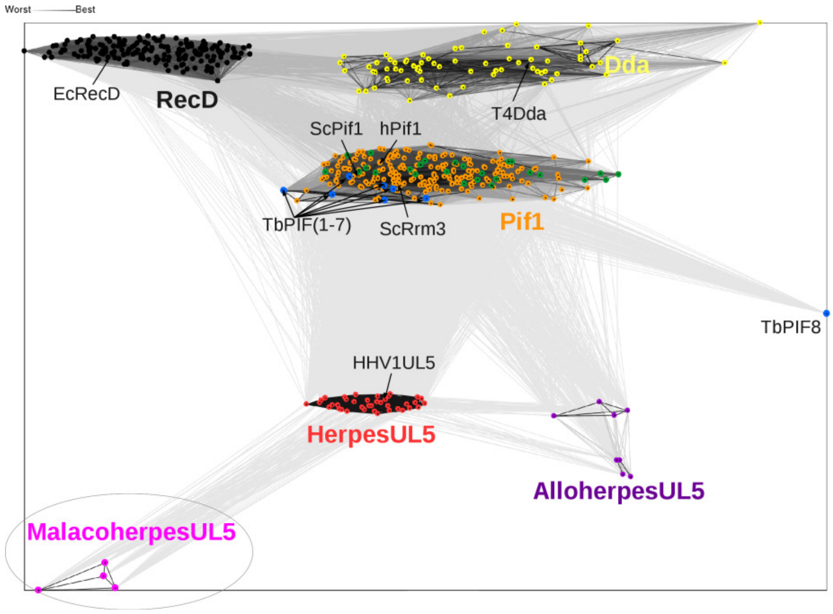


Fig. 2.13 Clusters of UL5 proteins and their homologs (P-value≤1e-07). Newly discovered UL5 homologs in *Malacoherpesviridae* family are marked with an ellipse. Pif1 homologs from the nucleo-cytoplasmic large DNA viruses, baculoviruses and phages are shown in green. Abbreviations: Sc, *Saccharomyces cerevisiae*; Ec, *Escherichia coli*; h, human; T4, T4 phage; Tb, *Trypanosoma brucei*.

Herpesviral helicase-primase: relationship with eukaryotic DNA replication. Our finding that UL8 is a homolog of B-family polymerases and that it has lost the active site explains why no catalytic activity of UL8 could be found. What could be the evolutionary origin of UL8? One possibility is that UL8 originated from the duplication of herpesviral DNA polymerase, UL30. Alternatively, UL8 could be derived from some ancestral form of a B-family polymerase. However, the sequence similarity with B-family polymerases is low, precluding a straightforward answer to this question. The analysis of homologs of UL52 and UL5, the other two subunits of helicase–primase complex, showed that their links to eukaryotic enzymes are even clearer. Thus, all three components of helicase–primase complex have mechanistic and perhaps functional similarities with the corresponding eukaryotic proteins. For example, UL8, an inactive polymerase, may be compared with the C-terminal region of Polε, which corresponds to an inactivated exonuclease-polymerase module serving as a protein-binding platform (81). UL5 has recently been found to interact with the polymerase UL30 (60), whereas yeast Pif1 interacts with Polδ (79). UL52–UL5 forms subassembly as part of helicase–primase complex. An interesting question is whether their eukaryotic homologs interact or at least cooperate in certain conditions. Although the direct evidence is lacking, there are some hints that they might. For example, PrimPol and Pif1 both help to bypass difficult-to-replicate sites (74,80). The observation of the correlated expansion of the PrimPol-like and Pif1 protein families in trypanosomes is another hint that they might be linked functionally if not physically. At least two pairs of trypanosomal PrimPol and Pif1 homologs participate in the same processes. Thus, TbPRI1 and TbPIF2 function in replication of DNA maxicircles while TbPRI2 and TbPIF1 are involved in replication and segregation of minicircles (76). Obviously, differences between the herpesviral and the eukaryotic counterparts are to be expected. However, the consideration of similarities may provide help in advancing the knowledge in both systems.

2.7. Single-stranded DNA-binding proteins

The SSB function in nature is strongly associated with the specific structural solution, oligonucleotide-binding (OB) fold, featuring a five-stranded β-barrel capped with an α-helix (82). To our knowledge, there are only two documented exceptions, where the structure of an SSB protein is unrelated to OB-fold. The two exceptions include adenovirus (83) and *Thermoproteales*, a clade of hyperthermophilic *Crenarchaea* (84).

Preliminary analysis of replication proteins in dsDNA viruses revealed that SSB proteins are found less frequently compared to DNA helicases, primases, DNA polymerases and primer removal proteins (Fig. 2.8.). This observation was surprising, given the essential role of SSB proteins in DNA replication. After examination of virus groups with "missing" SSB proteins we newly detected these proteins in large eukaryotic viruses. It turned out that nucleo-cytoplasmic large DNA viruses (NCLDV) have canonical OB-fold SSB proteins (homologs of T7 phage gp2.5) and that poxvirus SSB (I3) is related to bacterial Small protein B (SmpB).

2.7.1. Four NCLDV families have canonical OB-fold SSB proteins

We started by testing the hypothesis that NCLD viruses do possess classical OB-fold-containing SSB proteins, but with strongly diverged sequences. To this end, we performed iterative sequence searches with PSI-BLAST and jackhmmer. Searches were run against the nr70 sequence database using the *E*-value = 0.001 inclusion threshold and well-characterized SSB proteins from both cellular organisms and viruses as queries. To our surprise, when we used the T7 bacteriophage SSB protein (gp2.5) sequence (gi: 9627442) as a query, we detected statistically significant matches (*E*-value < 0.001) in four NCLDV families (*Phycodnaviridae*, *Mimiviridae*, *Iridoviridae* and Marseillevirus). We scrutinized this finding by testing whether these putative NCLDV SSB homologs are able to detect T7 SSB in a reverse search. For this test, we used HHsearch, to generate sequence profiles for several newly identified putative NCLDV SSB proteins and query them against profile databases. Searches against the sequence profile database derived from known structures (PDB) readily identified T7 gp2.5 with highly significant HHsearch probabilities (>95%). Taken together, these results provide convincing evidence that four NCLDV families encode T7-like SSB homologs. The detected relationship is not only reliable, but also specific because of distinct features of the T7 gp2.5 OB-fold domain. In T7 gp2.5, the capping α-helix is inserted between strands β2 and β3 in contrast to its typical position between β3 and β4.

The identified relationship between putative NCLDV SSB proteins and T7 gp2.5 suggests their similar role in DNA replication and recombination. This notion is strengthened by the genomic context. We noticed that frequently putative NCLDV SSB proteins are encoded in the vicinity of proteins involved in DNA replication and recombination, such as DNA polymerases, helicases, topoisomerases, clamp and clamp loader subunits. Moreover, transcriptomics data indicate that putative SSB genes in *Frog virus 3* (*Iridoviridae*), PBCV-1 (*Phycodnaviridae*) and Mimivirus are actively expressed early in the infection, together with DNA replication components (85-87).

Structure of NCLDV SSB proteins is consistent with the single-stranded DNA binding function. To get a further insight regarding structure and function of putative NCLDV SSB proteins, we constructed a homology model for a representative (gi: 340025795) from the phycodnavirus PBCV-1. The model was built using iterative optimization of both the sequence–structure alignment and the set of protein structures used as modeling templates (24). The initial PBCV-1 SSB model was constructed using the crystal structure of T7 gp2.5 (PDB: 1je5) as the single template. However, our analysis suggested that the capping α-helix (α1 in T7 gp2.5; Fig. 2.14) and the adjacent region of PBCV-1 SSB might be more similar to the corresponding OB-domain substructure of the herpesviral SSB protein (ICP8, PDB: 1urj). Indeed, the combination of T7 and herpesviral SSB structures led to a substantial improvement. According to the energy estimation with ProSA-web, the final PBCV-1 SSB model (available at: http://www.ibt.lt/bioinformatics/models/pbcv1_ssb/) fared better than the T7 gp2.5 structure, for which missing loops were modeled-in before the evaluation. Thus, evaluation results suggested that the model is likely to be sufficiently accurate for the exploration of functional properties.

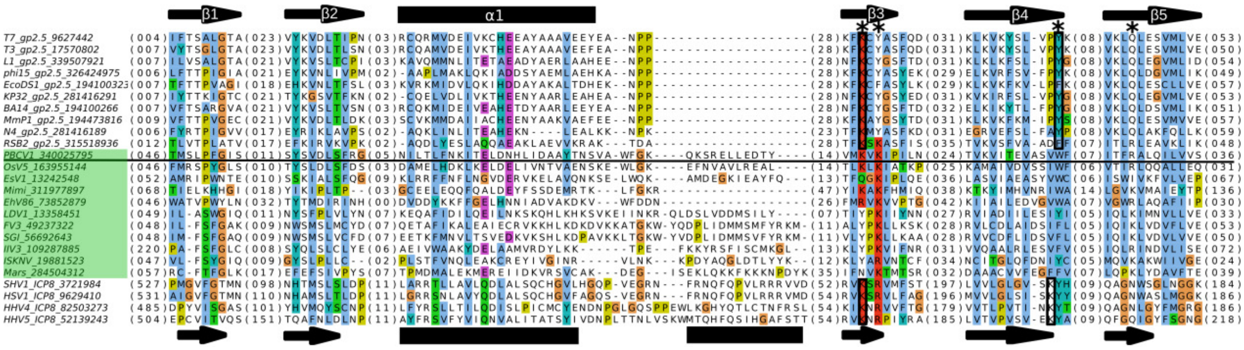


Fig. 2.14 Alignment of the OB-fold region of SSB proteins from T7-like phages, NCLDVs and herpesviruses. Secondary structures of T7 gp2.5 and herpes simplex virus 1 ICP8 are shown, respectively, above and below the alignment. NCLDV SSB residues predicted to interact with ssDNA are indicated with asterisks above the alignment. Residues, whose importance in the interaction with ssDNA was established experimentally in T7 and Suid herpes virus 1 SSBs, are indicated with the black frame. For T7 gp2.5, these are K109 and Y158 and for SHV1 ICP8 K756 and K970. NCLDV SSB protein labels have green background. PBCV-1 SSB is underlined.

To see whether the modeled PBCV-1 SSB structure indeed is suggestive of the single-stranded DNA (ssDNA) binding, we examined its surface properties (Fig. 2.15). Consistent with the predicted ssDNA-binding function, we observed an increased positive electrostatic potential in the region, which is thought to bind ssDNA in T7 gp2.5 (88). Moreover, we established that the same PBCV-1 SSB surface region features some of the highest evolutionary conservation (Fig. 2.15, B). Based on the conservation and/or electrostatic properties, we identified four residues that are most likely to participate in the ssDNA binding (Figs. 2.14 and 2.15).

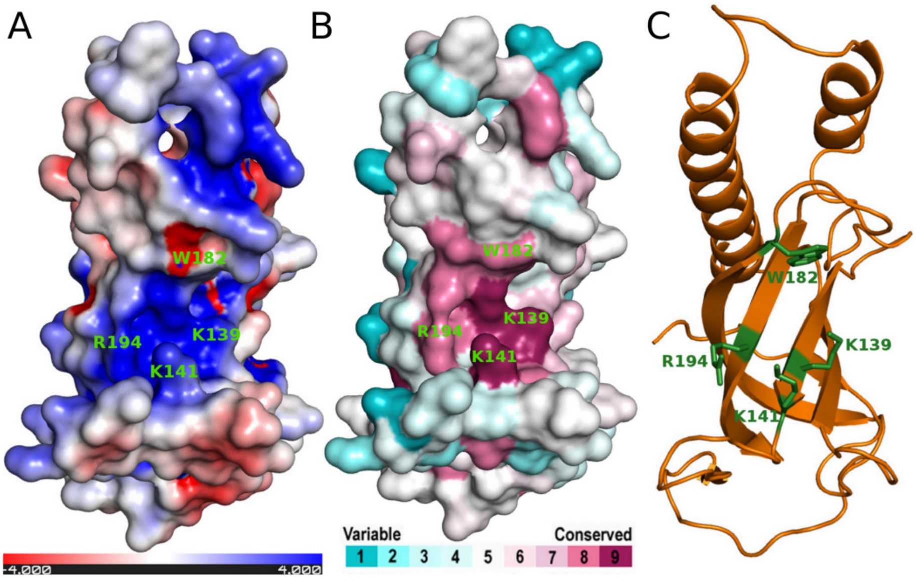


Fig. 2.15 PBCV-1 SSB structural model and its putative ssDNA-binding site. Residues predicted to be important for ssDNA binding are labeled. (A) The surface electrostatic potential map. (B) The position-specific conservation of NCLDV SSB proteins mapped onto the surface of the PBCV-1 SSB model. (C) Cartoon representation of the model with labeled residues shown as sticks.

The involvement of at least two of these residues in ssDNA binding is supported by experimental studies of T7 and herpesviral SSB proteins. Suid herpes virus SSB K756 corresponding to PBCV-1 SSB K139 is critical for ssDNA binding (89), while the point mutation of respective T7 2.5 residue (K109I) alters the ssDNA-binding mode (88). The Y158C point mutation renders T7 SSB defective in ssDNA binding (88), presumably by disrupting the stacking interaction with a nucleotide base. Consistent with such a role, the corresponding position in NCLDV SSB proteins (W182 in PBCV-1 SSB) is occupied exclusively by aromatic residues. The SSB protein of Suid herpes virus also has an aromatic residue (Y971) in this position. Although the importance of Y971 has not been addressed directly, the mutation of adjacent lysine (K970) significantly affected ssDNA binding (89). The importance of the remaining two positions in NCLDV SSB proteins (K141 and R194 in PBCV-1) has no experimental support coming from either T7 or herpesviral SSBs. However, their location within the putative ssDNA-binding cleft and a fairly strong conservation suggest a likely role in ssDNA binding.

2.7.2. Poxviral SSB proteins (the I3 family) are evolutionary distinct

In poxviruses, we did not detect either T7 gp2.5-like or any other OB-fold-containing SSB homolog. On the other hand, vaccinia virus, the prototypic poxvirus, has long been known to encode the ssDNA-binding protein, named I3 (90). Several reports have provided convincing evidence that I3 is the replicative SSB protein and a major player in viral DNA recombination (91,92). However, despite extensive efforts, no relationship between I3 and any other protein family could be established (92).

To search for homologous proteins related to the I3 family, we first performed standard PSI-BLAST and jackhmmer runs (up to five iterations, inclusion *E*-value = 0.001) against the nr70 sequence database using representative I3 sequences as queries. However, these searches revealed no statistically significant matches outside of the I3 family. To increase sensitivity, we performed the same searches with the more permissive *E*-value inclusion threshold (0.01). This time, jackhmmer (but not PSI-BLAST) readily detected small protein B (SmpB) sequences as significant matches. To test whether the I3 alignment with SmpB is indicative of true homology or is just a spurious match, we performed additional analyses. We noticed that only the central region of I3 sequences was aligned to SmpB. It is commonly known that sequence segments of low complexity or unrelated domains may hinder homology detection. Thus, we reasoned that if I3 and SmpB are homologous, the removal of unaligned regions of I3 should improve the detection of SmpB. This indeed turned out to be the case. One of the N- and C-terminally truncated I3 sequences (*Crocodilepox virus* I3; gi: 115531731) in the jackhmmer search was now able to detect SmpB using a more stringent inclusion threshold (*E*-value = 0.001).

Next, we performed homology searches using HHsearch. Initially, we generated profiles based on representative full-length I3 sequences and queried them against the PDB-derived profile database. HHsearch detected SmpB from *Aquifex aeolicus* (PDB: 1p6v) as the best match. The most significant results (HHsearch probability = 89%) were obtained using the *Yoka poxvirus* I3 (gi: 345107239) profile. This result alone would not be considered entirely reliable. However, HHsearch and jackhmmer returned essentially identical alignments for the same central region of I3 (Fig. 2.16). Moreover, when we ran queries with profiles based on truncated I3 sequences, HHsearch results in detecting SmpB have improved further, reaching the 95% probability (e.g. I3; gi: 41057468). Collectively, all these homology search results strongly suggest common ancestry for I3 and SmpB families.

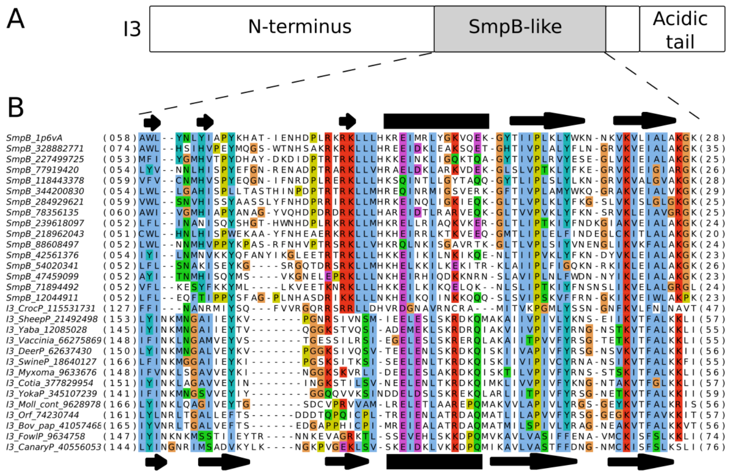


Fig. 2.16 Similarity between poxviral SSB (I3) and SmpB families. (A) Distinct sequence regions of I3. (B) Multiple alignment of SmpB and I3 representatives. The structure-derived secondary structure of *A.aeolicus* SmpB (PDB: 1p6v) and the predicted secondary structure of Vaccinia I3 are shown above and below the alignment, respectively.

SmpB is a bacterial protein that binds tmRNA, a hybrid RNA molecule having functions of both transfer RNA and messenger RNA. SmpB adopts a distinct structural fold with no significant similarity to any other structures (93). Despite that, SmpB was proposed to contain an "embedded" OB-fold (93), a proposition that appears to be misleading. Our analysis revealed that the SmpB structure is a *bona fide* duplication of the βαββ structural motif (Fig. 2.17). Thus, the pseudosymmetric SmpB structure, made of two consecutive repeats, is in stark contrast to the repeat-free asymmetric OB-fold.

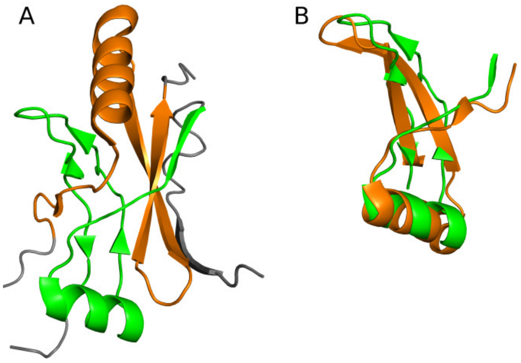


Fig. 2.17 SmpB consists of two structurally similar motifs. (A) Structure of *A. aeolicus* SmpB (B) Structurally similar motifs of *A. aeolicus* SmpB superimposed using Dali.

Interestingly, only the second SmpB repeat is aligned with I3 in its entirety (Fig. 2.16), suggesting its stronger evolutionary conservation. Consistent with this notion, the aligned regions of both SmpB and I3 harbor a number of functionally important residues (92,94). The central SmpB-like region of I3 is followed by the C-terminus, enriched in aspartates and glutamates and predicted to feature a significant intrinsic disorder. In this regard, the I3 C-terminus is reminiscent of unstructured acidic tails involved in protein–protein interactions in canonical bacterial and phage SSB proteins.

2.7.3. Taxonomic distribution of NCLDV SSBs and its evolutionary implications

We mapped the resulting distribution of NCLDV SSB proteins onto the consensus phylogenetic tree of NCLD viruses (Fig. 2.18). As NCLDVs are thought to be monophyletic (43), the distribution is suggestive of T7-like SSB being already present in the ancestral virus. This is consistent with the idea that bacteriophages played an important role in the early NCLDV evolution (43,95,96). Indeed, many of NCLDV genes essential for genome replication (DNA primase-helicase, NAD-dependent ligase and Holliday junction resolvase) show bacteriophage origins (97). Our newly detected T7 gp2.5-like protein family is yet another addition to this list. On the other hand, the unrelated I3 family, confined to poxviruses, is likely the result of a more recent non-orthologous replacement event.

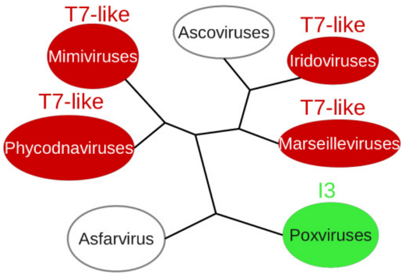


Fig. 2.18 Distribution of T7-like and I3 SSB families mapped onto the consensus NCLDV evolutionary tree (adopted from (97)).

CONCLUSIONS

1. The presence and the nature of DNA replicases encoded in the genomes of dsDNA viruses is related to the genome size. Small viruses (<40 kbp) use protein-primed DNA replication or rely on replication proteins from the host. Larger viruses tend to encode their own DNA polymerases more often. Largest viruses (>140 kbp) have their own RNA-primed replication apparatus often supplemented with processivity factors and sometimes by clamp loaders.

2. PhiKZ phages have highly divergent B-family DNA polymerases.

3. As genome size increases viruses tend to encode their own replication proteins more frequently. Replicative helicase is the most common replication protein in dsDNA viruses.

4. Herpesviral helicase-primase subunit UL8 is inactivated B-family polymerase.

5. Largest eukaryotic viruses (NCLDV) have at least two families of SSB proteins, which are evolutionary distinct.

LIST OF PUBLICATIONS

1. **Kazlauskas D.**, Venclovas Č. (2014). Herpesviral helicase-primase subunit UL8 is inactivated B-family polymerase. *Bioinformatics*, 30:2093-2097.
2. **Kazlauskas D.**, Venclovas Č. (2012) Two distinct SSB protein families in nucleo-cytoplasmic large DNA viruses. *Bioinformatics*, 28:3186-3190.
3. **Kazlauskas D.**, Venclovas Č. (2011) Computational analysis of DNA replicases in double-stranded DNA viruses: relationship with the genome size. *Nucleic Acids Res*, 39:8291-305.

CONFERENCE PRESENTATIONS

1. **Kazlauskas D.**, Venclovas Č. Viral DNA replication: new insights and discoveries from large scale computational analysis. *European Conference on Computational Biology*, Strasbūras, Prancūzija, 2014 09 06-10 (oral and poster presentations).
2. **Kazlauskas D.**, Venclovas Č. Viral DNA replication: new insights and discoveries from large scale computational analysis. *XIIIth International Conference of Lithuanian Biochemical Society*, Birštonas, Lietuva, 2014 06 18-20 (poster presentation).
3. **Kazlauskas D.**, Venclovas Č. Computational identification and analysis of Single-Stranded DNA Binding Proteins from Nucleo-Cytoplasmic Large DNA Viruses. *„Viruses of Microbes“ Meeting of International Society for Viruses of Microorganisms.* Briuselis, Belgija, 2012 07 16-20 (poster presentation).
4. **Kazlauskas D.**, Venclovas Č. Computational analysis of DNA replicases in double-stranded DNA viruses: Relationship with the genome size. *International Society for Computational Biology / European Conference on Computational Biology*, Viena, Austrija, 2011 07 14-20 (poster presentation).

FINANCIAL SUPPORT

This work was funded by Howard Hughes Medical Institute (55005627); Ministry of Education and Science of Lithuania, European Community's Seventh Framework Programme (FP7-REGPOT-2009-1 project 245721 "MoBiLi") and European Social Fund under the Global Grant measure (VP1-3.1-ŠMM-07-K-03-004).

CURRICULUM VITAE

Name	Darius Kazlauskas
Date of birth	1985 12 20
Work adress	Department of Bioinformatics, Institute of Biotechnology, Vilnius University V. A. Graičiūno 8 Vilnius LT-02241 Lithuania
Phone	+37060191333
E-mail	dariausk@gmail.com
Education and professional background	
2008	B.Sc. Molecular Biology Vilnius University
2010	M.Sc Biochemistry Vilnius University
2010-2014	PhD student of Biochemistry at Institute of Biotechnology, Vilnius University
2006-2014	Bioengineer at Department of Bioinformatics, Institute of Biotechnology, Vilnius University
Since 2013	Junior Scientist at Department of Bioinformatics, Institute of Biotechnology, Vilnius University

ACKNOWLEDGEMENTS

I am grateful to my scientific supervisor Česlovas Venclovas for his guidance in my research. I would like to thank Mindaugas Margelevičius, Albertas Timinskas, Ieva Mitašiūnaitė-Besson and Rytis Dičiūnas for technical support. I am also thankful Justas Dapkūnas, Kęstutis Timinskas, Kliment Olechnovič, Mart Krupovič and Elena Bakienė for advices and discussions. I would like to thank my student Algirdas Grybauskas for his help while analyzing viral DNA helicases. I am grateful to Giedrius Sasnauskas, Edita Sužiedėlienė and Ana Venclovienė who reviewed my manuscripts. I would like to thank Peter Kerpedjiev for providing articles.

Lastly, I sincerely thank my wife Natalija and my parents for their love, support and understanding.

REZIUMĖ

Sugebėjimas daugintis yra būtina visų gyvų esybių savybė. Genetinės informacijos kopijavimą atlieka replikacijos baltymai. Laisvai gyvenantys ląsteliniai organizmai turi tuos pačius replikacijos baltymus nepriklausomai nuo genomo dydžio. O kaip yra virusuose, kurių genetinę informaciją kaip ir ląstelinių organizmų atveju taip pat koduoja dvigrandė (dg) DNR? DgDNR virusai pasižymi ne tik replikacijos baltymų įvairove, bet ir plačiu genomų dydžio spektru. Jų genomo dydis gali skirtis net 500 kartų (nuo 5 iki 2500 kbp). DNR replikacija gerai ištirta T7, T4 faguose ir herpes virusuose, tačiau kitų dg DNR virusų dauginimasis vis dar lieka paslaptimi. Ar mažiau žinomi virusai naudoja jau charakterizuotų replikacijos baltymų variacijas? O gal jiems būdingos dar nežinomos DNR replikacijos strategijos?

Norint atsakyti į aukščiau pateiktus klausimus buvo atlikta pirma tokios plačios apimties analizė, kurioje tirti replikacijos baltymai visuose žinomuose bakterijų, archėjų ir eukariotų dgDNR virusuose. Naudojant pažangiausius kompiuterinius metodus buvo identifikuoti ir charakterizuoti virusų replikacijos baltymai bei nustatyti jų pasiskirstymo genomuose dėsningumai. Tyrimas vyko dviem etapais. Iš pradžių tirti replikazės baltymai (DNR polimerazės, procesyvumo veiksniai, žiedo užkėlimo kompleksai). Vėliau tyrimas papildytas kitais replikacijos baltymais (DNR praimazės, helikazės, viengrandę DNR surišantys baltymai, pradmens pašalinime dalyvaujančios nukleazės, DNR ligazės ir topoizomerazės).

Šio darbo metu atskleista, kad replikazės komponentų (ne)buvimas genome ir jų tipai priklauso nuo viruso genomo dydžio: maži virusai (<40 kbp) dažnai neturi DNR polimerazių arba turi B-šeimos polimerazę, kaip pradmenį naudojančią baltymą. Didėjant genomui virusai vis dažniau koduoja nuosavą DNR polimerazę, o patys didžiausi atstovai (>140 kbp) dažnai papildomai turi procesyvumo veiksnį ir kartais jo užkėlėją. Nustatyta, kad didėjant genomui virusai dažniau koduoja ne tik nuosavus replikazės, bet ir kitus replikacijos baltymus. Šios įžvalgos paskatino atidžiau išanalizuoti didelių genomų phiKZ fagus ir eukariotų virusus bei juose naujai rasti atitinkamai DNR polimerazes ir viengrandę DNR surišančius baltymus. Netikėtai, pastarieji buvo giminingi ne eukariotų, bet T7 fagų baltymams. Nuostabą kėlė ir tai, kad dažniausias virusų replikacijos baltymas yra ne DNR polimerazė, o helikazė. Ją turėjo net 70% visų dgDNR virusų. Beje, detali herpes virusų helikazės-praimazės analizė atskleidė dar vieną netikėtumą. Nors herpes virusų replikacijos aparatas yra vienas geriausiai ištirtų, tačiau iki šiol nebuvo žinoma, kad jų helikazės-praimazės komplekso baltymas UL8 yra inaktyvuota B-šeimos DNR polimerazė.

REFERENCES

1. Wernersson, R. (2006) Virtual Ribosome--a comprehensive DNA translation tool with support for integration of sequence feature annotation. *Nucleic Acids Res*, **34**, W385-388.
2. Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, **12**, 1611-1618.
3. Federici, B.A. and Bigot, Y. (2010) In Pontarotti, P. (ed.), *Evolutionary Biology - Concepts, Molecular and Morphological Evolution*. Springer Berlin Heidelberg, pp. 229-248.
4. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-3402.
5. Eddy, S.R. (2011) Accelerated Profile HMM Searches. *PLoS computational biology*, **7**, e1002195.
6. Söding, J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951-960.
7. Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, **247**, 536-540.
8. Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res*, **40**, D290-301.
9. Frickey, T. and Lupas, A. (2004) CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, **20**, 3702-3704.
10. Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, **30**, 772-780.
11. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**, 1792-1797.
12. Pei, J., Kim, B.H. and Grishin, N.V. (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res*, **36**, 2295-2300.
13. Kurowski, M.A. and Bujnicki, J.M. (2003) GeneSilico protein structure prediction meta-server. *Nucleic Acids Res*, **31**, 3305-3307.
14. Joosten, R.P., te Beek, T.A., Krieger, E., Hekkelman, M.L., Hooft, R.W., Schneider, R., Sander, C. and Vriend, G. (2011) A series of PDB related databases for everyday needs. *Nucleic Acids Res*, **39**, D411-419.
15. Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.A., Barrell, B.G. and Parkhill, J. (2005) ACT: the Artemis Comparison Tool. *Bioinformatics*, **21**, 3422-3423.
16. Frith, M.C., Wan, R. and Horton, P. (2010) Incorporating sequence quality data into alignment improves DNA read mapping. *Nucleic Acids Res*, **38**, e100.
17. Margelevičius, M. and Venclovas, Č. (2005) PSI-BLAST-ISS: an intermediate sequence search tool for estimation of the position-specific alignment reliability. *BMC Bioinformatics*, **6**, 185.
18. Margelevičius, M., Laganeckas, M. and Venclovas, Č. (2010) COMA server for protein distant homology search. *Bioinformatics*, **26**, 1905-1906.

19. Roy, A., Kucukural, A. and Zhang, Y. (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols*, **5**, 725-738.

20. Kallberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H. and Xu, J. (2012) Template-based protein structure modeling using the RaptorX web server. *Nature protocols*, **7**, 1511-1522.

21. Lobley, A., Sadowski, M.I. and Jones, D.T. (2009) pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics*, **25**, 1761-1767.

22. Xu, D., Jaroszewski, L., Li, Z. and Godzik, A. (2013) FFAS-3D: improving fold recognition by including optimized structural features and template re-ranking. *Bioinformatics*.

23. Söding, J., Biegert, A. and Lupas, A.N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res*, **33**, W244-248.

24. Venclovas, Č. and Margelevičius, M. (2009) The use of automatic tools and human expertise in template-based modeling of CASP8 target proteins. *Proteins*, **77**, 81-88.

25. Šali, A. and Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, **234**, 779-815.

26. Wiederstein, M. and Sippl, M.J. (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res*, **35**, W407-410.

27. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. and Ben-Tal, N. (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res*, **38**, W529-533.

28. Baker, N.A., Sept, D., Joseph, S., Holst, M.J. and McCammon, J.A. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A*, **98**, 10037-10041.

29. Schrodinger, LLC. (2010).

30. Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*, **16**, 276-277.

31. Hertveldt, K., Lavigne, R., Pleteneva, E., Sernova, N., Kurochkina, L., Korchevskii, R., Robben, J., Mesyanzhinov, V., Krylov, V.N. and Volckaert, G. (2005) Genome comparison of Pseudomonas aeruginosa large phages. *J Mol Biol*, **354**, 536-545.

32. Mesyanzhinov, V.V., Robben, J., Grymonprez, B., Kostyuchenko, V.A., Bourkaltseva, M.V., Sykilinda, N.N., Krylov, V.N. and Volckaert, G. (2002) The genome of bacteriophage phiKZ of Pseudomonas aeruginosa. *J Mol Biol*, **317**, 1-19.

33. Thomas, J.A., Rolando, M.R., Carroll, C.A., Shen, P.S., Belnap, D.M., Weintraub, S.T., Serwer, P. and Hardies, S.C. (2008) Characterization of Pseudomonas chlororaphis myovirus 201varphi2-1 via genomic sequencing, mass spectrometry, and electron microscopy. *Virology*, **376**, 330-338.

34. Weigel, C. and Seitz, H. (2006) Bacteriophage replication modules. *FEMS Microbiol Rev*, **30**, 321-381.

35. Andraos, N., Tabor, S. and Richardson, C.C. (2004) The highly processive DNA polymerase of bacteriophage T5. Role of the unique N and C termini. *J Biol Chem*, **279**, 50609-50618.

36. Druck Shudofsky, A.M., Silverman, J.E., Chattopadhyay, D. and Ricciardi, R.P. (2010) Vaccinia virus D4 mutants defective in processive DNA synthesis retain binding to A20 and DNA. *J Virol*, **84**, 12325-12335.

37. Lamers, M.H., Georgescu, R.E., Lee, S.G., O'Donnell, M. and Kuriyan, J. (2006) Crystal structure of the catalytic alpha subunit of E. coli replicative DNA polymerase III. *Cell*, **126**, 881-892.

38. Dalrymple, B.P., Kongsuwan, K., Wijffels, G., Dixon, N.E. and Jennings, P.A. (2001) A universal protein-protein interaction motif in the eubacterial DNA replication and repair systems. *Proc Natl Acad Sci U S A*, **98**, 11627-11632.

39. Moarefi, I., Jeruzalmi, D., Turner, J., O'Donnell, M. and Kuriyan, J. (2000) Crystal structure of the DNA polymerase processivity factor of T4 bacteriophage. *J Mol Biol*, **296**, 1215-1223.

40. Zuccola, H.J., Filman, D.J., Coen, D.M. and Hogle, J.M. (2000) The crystal structure of an unusual processivity factor, herpes simplex virus UL42, bound to the C terminus of its cognate polymerase. *Mol Cell*, **5**, 267-278.

41. Appleton, B.A., Loregian, A., Filman, D.J., Coen, D.M. and Hogle, J.M. (2004) The cytomegalovirus DNA polymerase subunit UL44 forms a C clamp-shaped dimer. *Mol Cell*, **15**, 233-244.

42. Murayama, K., Nakayama, S., Kato-Murayama, M., Akasaka, R., Ohbayashi, N., Kamewari-Hayami, Y., Terada, T., Shirouzu, M., Tsurumi, T. and Yokoyama, S. (2009) Crystal structure of epstein-barr virus DNA polymerase processivity factor BMRF1. *J Biol Chem*, **284**, 35896-35905.

43. Iyer, L.M., Aravind, L. and Koonin, E.V. (2001) Common origin of four diverse families of large eukaryotic DNA viruses. *J Virol*, **75**, 11720-11734.

44. Boyle, K. and Traktman, P. (2009) In Raney, K. D., Gotte, M. and Cameron, C. E. (eds.), *Viral Genome Replication*. Springer US, pp. 225-247.

45. Yoshida, T., Nagasaki, K., Takashima, Y., Shirai, Y., Tomaru, Y., Takao, Y., Sakamoto, S., Hiroishi, S. and Ogata, H. (2008) Ma-LMM01 infecting toxic Microcystis aeruginosa illuminates diverse cyanophage genome strategies. *J Bacteriol*, **190**, 1762-1772.

46. Kool, M., Ahrens, C.H., Goldbach, R.W., Rohrmann, G.F. and Vlak, J.M. (1994) Identification of genes involved in DNA replication of the Autographa californica baculovirus. *Proc Natl Acad Sci U S A*, **91**, 11212-11216.

47. Komazin-Meredith, G., Santos, W.L., Filman, D.J., Hogle, J.M., Verdine, G.L. and Coen, D.M. (2008) The Positively Charged Surface of Herpes Simplex Virus UL42 Mediates DNA Binding. *J Biol Chem*, **283**, 6154-6161.

48. Loregian, A., Sinigalia, E., Mercorelli, B., Palu, G. and Coen, D.M. (2007) Binding parameters and thermodynamics of the interaction of the human cytomegalovirus DNA polymerase accessory protein, UL44, with DNA: implications for the processivity mechanism. *Nucleic Acids Res*, **35**, 4779-4791.

49. Konagurthu, A.S., Whisstock, J.C., Stuckey, P.J. and Lesk, A.M. (2006) MUSTANG: a multiple structural alignment algorithm. *Proteins*, **64**, 559-574.

50. Lopez de Saro, F.J. and O'Donnell, M. (2001) Interaction of the beta sliding clamp with MutS, ligase, and DNA polymerase I. *Proc Natl Acad Sci U S A*, **98**, 8376-8380.
51. Bedford, E., Tabor, S. and Richardson, C.C. (1997) The thioredoxin binding domain of bacteriophage T7 DNA polymerase confers processivity on Escherichia coli DNA polymerase I. *Proc Natl Acad Sci U S A*, **94**, 479-484.
52. Indiani, C. and O'Donnell, M. (2006) The replication clamp-loading machine at work in the three domains of life. *Nat Rev Mol Cell Biol*, **7**, 751-761.
53. Schmidt, S.L., Gomes, X.V. and Burgers, P.M. (2001) ATP utilization by yeast replication factor C. III. The ATP-binding domains of Rfc2, Rfc3, and Rfc4 are essential for DNA recognition and clamp loading. *J Biol Chem*, **276**, 34784-34791.
54. McCutcheon, J.P. (2010) The bacterial essence of tiny symbiont genomes. *Curr Opin Microbiol*, **13**, 73-78.
55. Shereda, R.D., Kozlov, A.G., Lohman, T.M., Cox, M.M. and Keck, J.L. (2008) SSB as an organizer/mobilizer of genome maintenance complexes. *Critical reviews in biochemistry and molecular biology*, **43**, 289-318.
56. Leipe, D.D., Aravind, L. and Koonin, E.V. (1999) Did DNA replication evolve twice independently? *Nucleic Acids Res*, **27**, 3389-3401.
57. Filee, J., Forterre, P. and Laurent, J. (2003) The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies. *Research in microbiology*, **154**, 237-243.
58. Mansky, K.C., Batiza, A. and Lambert, P.F. (1997) Bovine papillomavirus type 1 E1 and simian virus 40 large T antigen share regions of sequence similarity required for multiple functions. *J Virol*, **71**, 7600-7608.
59. Salas, M. (1991) Protein-priming of DNA replication. *Annu Rev Biochem*, **60**, 39-71.
60. Weller, S.K. and Coen, D.M. (2012) Herpes simplex viruses: mechanisms of DNA replication. *Cold Spring Harbor perspectives in biology*, **4**, a013011.
61. Hodgman, T.C. (1988) A new superfamily of replicative proteins. *Nature*, **333**, 22-23.
62. Iyer, L.M., Koonin, E.V., Leipe, D.D. and Aravind, L. (2005) Origin and evolution of the archaeo-eukaryotic primase superfamily and related palm-domain proteins: structural insights and new members. *Nucleic Acids Res*, **33**, 3875-3896.
63. Dodson, M.S. and Lehman, I.R. (1991) Association of DNA helicase and primase activities with a subassembly of the herpes simplex virus 1 helicase-primase composed of the UL5 and UL52 gene products. *Proc Natl Acad Sci U S A*, **88**, 1105-1109.
64. Muylaert, I., Tang, K.W. and Elias, P. (2011) Replication and recombination of herpes simplex virus DNA. *J Biol Chem*, **286**, 15619-15624.
65. Rothwell, P.J. and Waksman, G. (2005) Structure and mechanism of DNA polymerases. *Adv Protein Chem*, **71**, 401-440.
66. Wang, F. and Yang, W. (2009) Structural insight into translesion synthesis by DNA Pol II. *Cell*, **139**, 1279-1289.
67. Franklin, M.C., Wang, J. and Steitz, T.A. (2001) Structure of the replicating complex of a pol alpha family DNA polymerase. *Cell*, **105**, 657-667.

68. Parry, M.E., Stow, N.D. and Marsden, H.S. (1993) Purification and properties of the herpes simplex virus type 1 UL8 protein. *J Gen Virol*, **74 (Pt 4)**, 607-612.

69. Chen, Y., Bai, P., Mackay, S., Korza, G., Carson, J.H., Kuchta, R.D. and Weller, S.K. (2011) Herpes simplex virus type 1 helicase-primase: DNA binding and consequent protein oligomerization and primase activation. *J Virol*, **85**, 968-978.

70. Muylaert, I., Zhao, Z., Andersson, T. and Elias, P. (2012) Identification of conserved amino acids in the herpes simplex virus type 1 UL8 protein required for DNA synthesis and UL52 primase interaction in the virus replisome. *J Biol Chem*, **287**, 33142-33152.

71. Ishida, T. and Kinoshita, K. (2007) PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res*, **35**, W460-464.

72. Kozlowski, L.P. and Bujnicki, J.M. (2012) MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinformatics*, **13**, 111.

73. Marsden, H.S., McLean, G.W., Barnard, E.C., Francis, G.J., MacEachran, K., Murphy, M., McVey, G., Cross, A., Abbotts, A.P. and Stow, N.D. (1997) The catalytic subunit of the DNA polymerase of herpes simplex virus type 1 interacts specifically with the C terminus of the UL8 component of the viral helicase-primase complex. *J Virol*, **71**, 6390-6397.

74. Garcia-Gomez, S., Reyes, A., Martinez-Jimenez, M.I., Chocron, E.S., Mouron, S., Terrados, G., Powell, C., Salido, E., Mendez, J., Holt, I.J. *et al.* (2013) PrimPol, an Archaic Primase/Polymerase Operating in Human Cells. *Mol Cell*, **52**, 541-553.

75. Mouron, S., Rodriguez-Acebes, S., Martinez-Jimenez, M.I., Garcia-Gomez, S., Chocron, S., Blanco, L. and Mendez, J. (2013) Repriming of DNA synthesis at stalled replication forks by human PrimPol. *Nat Struct Mol Biol.*

76. Bezalel-Buch, R., Yaffe, N. and Shlomai, J. (2013), *Trypanosomatid Diseases*. Wiley-VCH Verlag GmbH & Co. KGaA, pp. 243-260.

77. Rudd, S.G., Glover, L., Jozwiakowski, S.K., Horn, D. and Doherty, A.J. (2013) PPL2 translesion polymerase is essential for the completion of chromosomal DNA replication in the African trypanosome. *Mol Cell*, **52**, 554-565.

78. Bochman, M.L., Sabouri, N. and Zakian, V.A. (2010) Unwinding the functions of the Pif1 family helicases. *DNA Repair (Amst)*, **9**, 237-249.

79. Wilson, M.A., Kwon, Y., Xu, Y., Chung, W.H., Chi, P., Niu, H., Mayle, R., Chen, X., Malkova, A., Sung, P. *et al.* (2013) Pif1 helicase and Poldelta promote recombination-coupled DNA synthesis via bubble migration. *Nature*, **502**, 393-396.

80. Azvolinsky, A., Dunaway, S., Torres, J.Z., Bessler, J.B. and Zakian, V.A. (2006) The S. cerevisiae Rrm3p DNA helicase moves with the replication fork and affects replication of all yeast chromosomes. *Genes & development*, **20**, 3104-3116.

81. Tahirov, T.H., Makarova, K.S., Rogozin, I.B., Pavlov, Y.I. and Koonin, E.V. (2009) Evolution of DNA polymerases: an inactivated polymerase-exonuclease module in Pol epsilon and a chimeric origin of eukaryotic polymerases from two classes of archaeal ancestors. *Biol Direct*, **4**, 11.

82. Murzin, A.G. (1993) OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. *Embo J*, **12**, 861-867.

83. Tucker, P.A., Tsernoglou, D., Tucker, A.D., Coenjaerts, F.E., Leenders, H. and van der Vliet, P.C. (1994) Crystal structure of the adenovirus DNA binding protein reveals a hook-on model for cooperative DNA binding. *EMBO J*, **13**, 2994-3002.

84. Paytubi, S., McMahon, S.A., Graham, S., Liu, H., Botting, C.H., Makarova, K.S., Koonin, E.V., Naismith, J.H. and White, M.F. (2012) Displacement of the canonical single-stranded DNA-binding protein in the Thermoproteales. *Proc Natl Acad Sci U S A*, **109**, E398-405.

85. Legendre, M., Audic, S., Poirot, O., Hingamp, P., Seltzer, V., Byrne, D., Lartigue, A., Lescot, M., Bernadac, A., Poulain, J. *et al.* (2010) mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in Mimivirus. *Genome Res*, **20**, 664-674.

86. Majji, S., Thodima, V., Sample, R., Whitley, D., Deng, Y., Mao, J. and Chinchar, V.G. (2009) Transcriptome analysis of Frog virus 3, the type species of the genus Ranavirus, family Iridoviridae. *Virology*, **391**, 293-303.

87. Yanai-Balser, G.M., Duncan, G.A., Eudy, J.D., Wang, D., Li, X., Agarkova, I.V., Dunigan, D.D. and Van Etten, J.L. (2010) Microarray analysis of Paramecium bursaria chlorella virus 1 transcription. *J Virol*, **84**, 532-542.

88. Hyland, E.M., Rezende, L.F. and Richardson, C.C. (2003) The DNA binding domain of the gene 2.5 single-stranded DNA-binding protein of bacteriophage T7. *J Biol Chem*, **278**, 7247-7256.

89. Wu, S.L., Li, C.C., Ho, T.Y. and Hsiang, C.Y. (2009) Mutagenesis identifies the critical regions and amino acid residues of suid herpesvirus 1 DNA-binding protein required for DNA binding and strand invasion. *Virus Res*, **140**, 147-154.

90. Rochester, S.C. and Traktman, P. (1998) Characterization of the single-stranded DNA binding protein encoded by the vaccinia virus I3 gene. *J Virol*, **72**, 2917-2926.

91. Gammon, D.B. and Evans, D.H. (2009) The 3'-to-5' exonuclease activity of vaccinia virus DNA polymerase is essential and plays a role in promoting virus genetic recombination. *J Virol*, **83**, 4236-4250.

92. Greseth, M.D., Boyle, K.A., Bluma, M.S., Unger, B., Wiebe, M.S., Soares-Martins, J.A., Wickramasekera, N.T., Wahlberg, J. and Traktman, P. (2012) Molecular genetic and biochemical characterization of the vaccinia virus i3 protein, the replicative single-stranded DNA binding protein. *J Virol*, **86**, 6197-6209.

93. Dong, G., Nowakowski, J. and Hoffman, D.W. (2002) Structure of small protein B: the protein component of the tmRNA-SmpB system for ribosome rescue. *EMBO J*, **21**, 1845-1854.

94. Gutmann, S., Haebel, P.W., Metzinger, L., Sutter, M., Felden, B. and Ban, N. (2003) Crystal structure of the transfer-RNA domain of transfer-messenger RNA in complex with SmpB. *Nature*, **424**, 699-703.

95. Iyer, L.M., Balaji, S., Koonin, E.V. and Aravind, L. (2006) Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res*, **117**, 156-184.

96. Yutin, N., Wolf, Y.I., Raoult, D. and Koonin, E.V. (2009) Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. *Virology journal*, **6**, 223.

97. Koonin, E.V. and Yutin, N. (2010) Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. *Intervirology*, **53**, 284-292.