Bacteriophage replication modules

supplement: compendium of origins and proteins

Christoph Weigel & Harald Seitz

Max-Planck-Institut für molekulare Genetik, Ihnestrasse 73, Berlin, Germany

Correspondence: Christoph Weigel, Technical University Berlin, Faculty III, Inst. for Biotechnology, Sekr. MA 5-11, Straße des 17. Juni 136, D-10623 Berlin, Germany.
Tel. +49 30 84131614,
e-mail: weigel@molgen.mpg.de

Received 22 March 2005, revised 6 September 2005, accepted 8 November 2005

doi: 10.1111/j.1574-6976.2006.00015.x

Editor: Ramón Díaz Orejas

Abbreviations: bp base pair(s); cpϕ cryptic prophage, in most cases defect, i.e. non-inducible; cSDR constitutive stable DNA replication; D-loop DNA-based displacement loop; DSB double-strand break; dsDNA double-stranded DNA; EMSA electrophoretic mobility shift assay; F4 family 4 (of helicases), e.g. $DnaB_{Eco}$; HGT horizontal gene transfer; hmUra 5-(hydroxymethyl)-2'-deoxyuridine; HTH helix-turn-helix DNA-binding motif; iSDR inducible stable DNA replication; NMR nuclear magnetic resonance; NTP nucleoside 5'-triphosphate; orf open reading frame; PCNA proliferating cell nuclear antigen (sliding clamp); ϕ (phi) prefix to mark bacteriophage (replicon); p prefix to mark plasmid (replicon); pϕ prophage; ppDR protein-primed DNA replication; RCR 'rolling circle' DNA replication; RDR recombination-dependent DNA replication; RF replicative form (of ssDNA phages); R-loop displacement loop, based on a RNA·DNA hybrid; RFC replication factor C (clamp loader); SDR stable DNA replication;σDR sigma(σ)-type DNA replication; SF2 superfamily 2 (of helicases), e.g. ϕT4 UvsW; SSB single-strand (DNA) binding protein; SAP single-strand (DNA) annealing protein; ssDNA single-stranded DNA; TBD thioredoxin binding domain; tDR transcript-driven initiation of DNA replication; (ts) temperature-sensitive mutation;θDR theta(θ)-type DNA replication; TP terminal protein; UDG uracil-DNA glycosylase.

Contents

Contents (continued)

C1. Introduction

This supplement of the review 'bacteriophage replication modules' provides a comprehensive basis of knowledge *and* data – for the discussion of replication mechanisms as well as for the discussion of phage replication modules. It is structured in a way that allows rapid access to particular replication components.

Within the individual chapters of each section, the role of the respective component for bacteriophage replication is discussed first in the context of analogous components from plasmids and chromosomes. We then review the genetics and biochemistry of well-known examples for phage replication origins (Section C2.) and replication proteins (Section C3.). Each chapter concludes with a discussion of homologous or similar origins/proteins that were – in their majority – detected by (DNA/protein) BLAST searches and have not yet been studied experimentally. This latter part within each chapter will appear less 'readable' due to numerous Figures and Tables designed to present the available data, but it should be easily 'searchable', and thus provide a compendium.

Note: all sections of the main body of the review 'bacteriophage replication modules' referred to in the following are marked with 'BRM' (see also last page). All numbers of Sections, Tables, and Figures of this compendium are 'tagged' with the prefix 'C' in order to facilitate unambiguous navigation between both parts of the review. The table of contents and all tables, figures, citations and *internal* cross-references are hyperlinked in the PDF version (not visible).

C2. Bacteriophage replication origins

A replication origin is defined as a specific segment of a replicon where the synthesis of daughter DNA molecules commences. Prokaryotic replicons contain one single replication origin. There are two exceptions from this rule: i. phage or plasmid replicons that propagate by the rolling circle mechanism of DNA replication (RCR) contain two replication origins, one for (+)- strand synthesis and one for (–)-strand synthesis (BRM Section 2.1.), and ii. joint replicons that use different replication origins for different stages of their life-cycles. E.g. phage P1 replication is initiated in the θ-mode at *oriL* (within the *repL* gene) during the lytic cycle, and also in the θ-mode at *oriR* during propagation as (plasmid) prophage of its host. Conjugative plasmids use dif-

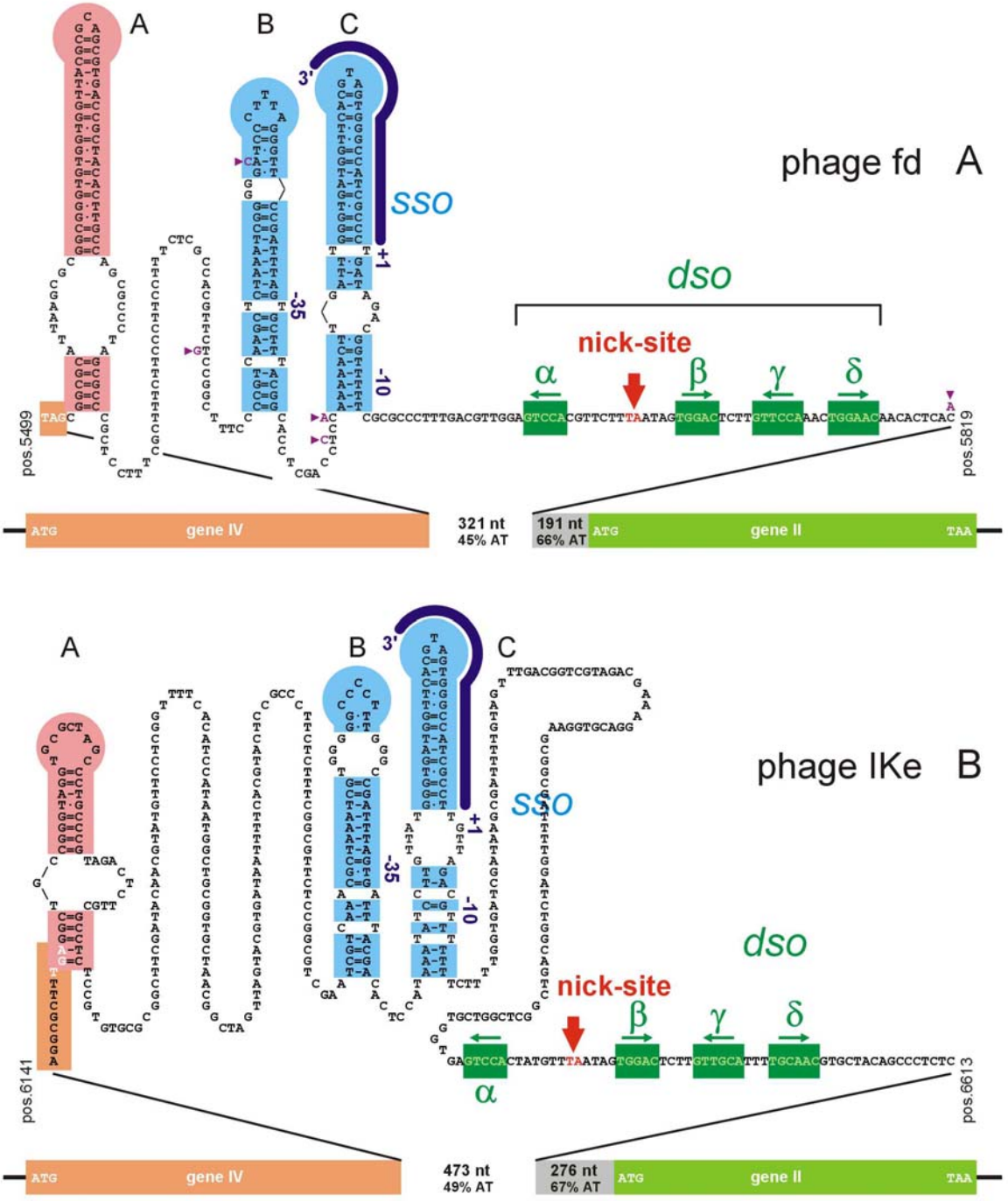

Fig. C1 Structure of the ϕfd and ϕIKe replication origins.
A. Sequence and structural elements of the ϕfd [J02451] origin region. The sequence of the viral (+)-strand is shown for pos. 5499 to 5819. Stemloops (A = pink, B+C = light blue), the promoter structural motifs (-35 region, -10 region, and +1 transcript start) with the priming transcript (dark blue line) of the *sso*, and the structural motifs of the *dso* (gpII binding-sites α, β, γ, δ = dark green; nick-site = red + red arrowhead) were assigned according to Horiuchi [181]. Genes IV, II, and the intergenic region are shown in a different scale. Sequence and structural elements of the origin region are identical in ϕf1 and ϕM13 [NC_003287] except for the residues indicated and marked by a dark violet triangle. **B**. Sequence and structural elements of the ϕIKe [NC_002014] origin region. The sequence of the viral (+)-strand is shown for pos. 6141 to 6613. Numbering and colour code for structural elements as in **A**.

ferent origins for DNA transfer during conjugation and during propagation in their hosts. Formally, also most bacterial chromosomes are 'joint replicons' because they contain varying numbers of prophages. But since initiation from prophage origins is normally tightly repressed, the replication origin *oriC* is the only 'active' replication origin of bacterial chromosomes under most growth conditions (replication restart sites are not considered 'replication origins'). Within this theoretical framework, different experimental approaches result in slightly differing origin definitions – besides their specific resolution.

Electron microscopy was frequently used in early studies for the analysis of bacteriophage replication intermediates. Origins can be mapped by this technique to the smallest detectable 'replication bubbles', or by extrapolating a common 'origin' from a series of bubbles of different size. By this technique, the positions of replication origins can be pinpointed to regions of approximately 100 bp in length (mostly expressed as a position corresponding to '%' of the genome). Despite its somewhat limited resolution, electron microscopy can contribute valuable details, e.g. indications for unidirectional and/or bidirectional replication, and also indications for coupled leading- and lagging-strand DNA synthesis and/or displacement synthesis. Replication intermediates of RCR and θDR can be distinguished by electron microscopy but a discrimination between replication intermediates of θDR and tDR is not possible.

Genetics defines the replication origin of a prokaryotic replicon as the only element that is strictly required *in cis*. The experimental strategy for the isolation of the *E. coli* chromosomal replication origin, *oriC*, based on this definition [437,286]. Also, this definition has been widely exploi-ted to detect the replication origins of various phages – the history of the λ*dv* plasmids may serve as example (BRM Section 3.). In combination with mutational analysis and DNA sequencing, the genetic approach can define a replication origin more exactly than any other method – at least with respect to size. However, the raw DNA sequence does not allow for a prediction of the replication mechanism operating on this origin unless it is reasonably similar to a known replication origin. Even if the replication mechanism is known, the set of *trans*-acting factors has to be determined in subsequent experiments – mostly, but not always a trivial task.

The major contribution of protein biochemistry to the definition of the replication origin(s) of a given replicon has always been the elucidation of the complete set of factors – cognate and host-encoded – that are required to obtain *de novo* DNA synthesis *in vitro*. But in addition, for most experimental systems the first indications that the conformation of the origin-containing DNA substrate – single/double-stranded, linear/circular, relaxed/negatively supercoiled – determines its proficiency to direct DNA synthesis came from biochemical studies. Biochemistry can determine the first nucleotides synthesised for the leading- and the lagging-strand, but this may lead to a biased definition of the origin if the primase acts distributively rather than at fixed positions (compare e.g. [406] and [169]).

State-of-the-art analyses of replication origins include i. the genetic definition, ii. the DNA sequence, iii. the biochemically defined 'catalogue' of factors required for initiation, and iv. footprinting analyses which determine the nucleic acid-protein interaction(s) and the DNA distortions that characterise the 'DNA unwinding' – either for 'nicking' or for 'melting'. We wish to emphasise that the interaction of a replication origin with its cognate initiation protein(s) is a highly dynamic process and determined by the properties of both, the nucleic acid and the protein(s). Horiuchi and co-workers found a point mutation in the ϕf1 gpII initiator that renders the protein proficient for unwinding at reduced levels of origin superhelicity [171]. On the other side, Leonard and co-workers describe point mutations in *E. coli oriC* that allow initiation by ADP-DnaA, which is inactive under 'normal conditions' [250] (Section C3.1.2.). The dynamic nature of the origin-initiator interaction has until now frustrated attempts to describe the origin-unwinding reaction in thermodynamically sound terms. We can expect considerable progress by forthcoming analyses employing the novel 'single molecule analysis' technique (reviewed in [433,434]).

The replication mechanism driving the propagation of a replicon determines the structural elements that constitute its replication origin. We will therefore discuss origins for 'rolling circle' DNA replication (RCR) and theta (θ)-type DNA replication (θDR) in separate chapters. The ϕ29 'replication origin' for protein-primed DNA replication (ppDR) is discussed in BRM Section 2.3. . Also, we will not discuss in this section the replication origins for transcription-driven DNA replication (tDR) of the phages T7 and T4. The promoters that serve as replication origins for these phages are discussed in BRM Sections 2.4. & 2.5., respectively. We will discuss instead the replication origins of ϕc2 and related phages, which probably also replicate by tDR.

C2.1. Origins for 'rolling circle' DNA replication (RCR)

We characterise in BRM Section 2.1. 'rolling circle' DNA replication (RCR) as a mechanism that avoids topological problems during replication of circular dsDNA molecules by directing the synthesis of both DNA daughter-strands in consecutive steps. Both synthesis steps employ different initiation mechanisms – and initiation proteins – and different replication origin

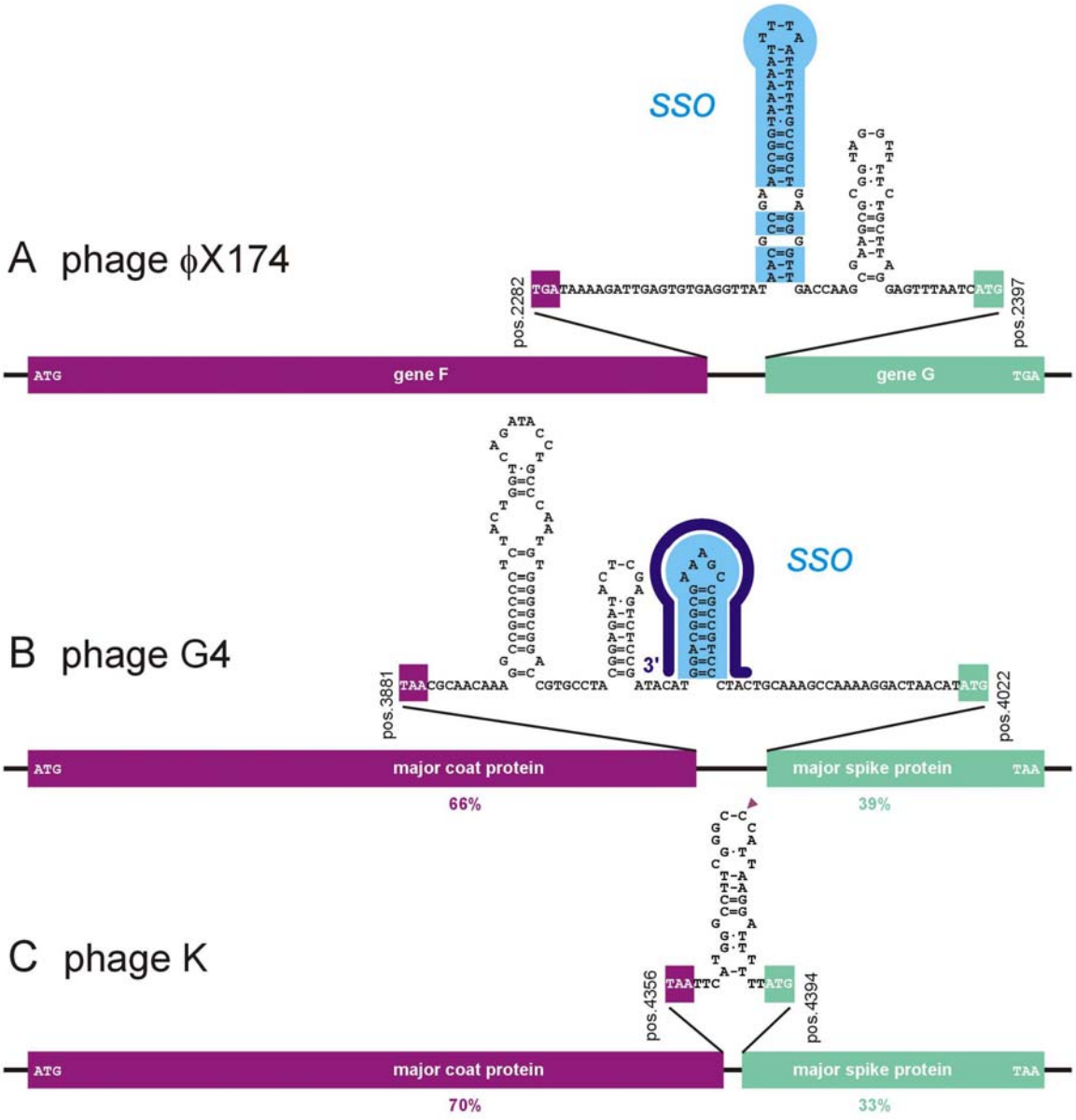

Fig. C2 Structure of the ϕX174, ϕG4, and ϕK single-strand origins (*ssos*).
A. Sequence and structural elements of the ϕX174 [NC_001422] *sso*. The sequence of the viral (+)-strand is shown for pos. 2282 to 2397. The light blue stemloop corresponds to the PAS site detected by Shlomai and Kornberg [413]. Genes F, G, and the intergenic region are shown in a different scale. Sequence and structural elements of the *sso* region are identical in ϕS13 [NC_001424]. **B**. Sequence and structural elements of the ϕG4 [NC_001420] *sso*. The sequence of the viral (+)-strand is shown for pos. 3881 to 4022. The light blue stemloop corresponds to the structure, and the dark blue line to the DnaG-dependent transcript detected by Bouché *et al.* [41]. The genes for the major coat protein, the major spike protein, and the intergenic region are shown in a different scale; %-values indicate similarity to the corresponding proteins of ϕX174 (ident. res.). **C**. Sequence and structural elements of the putative ϕK [NC_001730] *sso*. The sequence of the viral (+)-strand is shown for pos. 4356 to 4394. Sequence and structural elements of the *sso* region are identical in ϕα3 [NC_001330] except for the missing C residue in the steemloop marked with a lilac triangle. The genes for the major coat protein, the major spike protein, and the intergenic region are shown in a different scale; %-values indicate similarity to the corresponding proteins of ϕX174 (ident. res.).

structures. Therefore, phage replicons that propagate by RCR contain a 'double-strand origin' (*dso*) for the synthesis of single-stranded DNA circles – the viral (+)-strand – and a 'single-strand origin' (*sso*) for the synthesis of the complementary strand – the viral (–)-strand. The *dso* contains the binding site(s) for the cognate initiator protein responsible for the 'nicking reaction' that precedes RCR. Replication initiation driven by a *sso* on single-stranded DNA involves host RNA polymerase (ϕfd), primase (ϕG4), or a restart primosome (ϕX174). A *sso* is characterised by DNA secondary structure elements that are specifically recognised by the initiation protein(s).

The replication origin region of the filamentous phage fd of *E. coli*. The structural and functional elements of the almost identical replication origin regions of the filamentous phages fd, f1 and M13 have been extensively studied by Horiuchi and co-workers [181].

The *dso* of ϕfd spans a "core region" of ~40 bp, which is absolutely required for viral (+)-strand synthesis. The *dso* contains the four binding-sites α–δ for the phage-encoded initiator gpII (Section C3.1.1.) arranged as inverted repeats, and the nick-site between α and β (Fig. C1; A). While the specific sequence of the gpII binding-site α is not absolutely required for initiation, the sequence around and including α is necessary for the termination reaction. The AT-rich region downstream of the *dso* is important for origin function but its exact role unknown. A deletion of the AT-rich region can be compensated for by mutations elsewhere in the genome, and also insertion in this region – as in cloning vectors derived from ϕf1 or ϕM13 – do not abolish origin function [181].

In the 'left' part of the origin region, stemloop A is thought to serve as 'morphogenetic signal' for packaging of viral (+)-strands into phage heads and therefore not part of the replication origin in the strict sense [181], stemloops B and C constitute the *sso*. Both stemloops B and C together create a partially single-stranded promoter structure with the −35 region residing within stemloop B, and the −10 region and the transcription start within stemloop C. Except for the direct neighbourhood of the −35 and −10 regions, the base- pairing in the stems is more important than a specific nucleotide sequence, and the entire structure is recognised by *E. coli* RNA polymerase holoenzyme ($\alpha\beta\beta'\sigma^{70}$). The impressive promoter strength – approximately 15-fold stronger than the *lac*UV5 promoter – is probably due to the partial single-strandedness of the otherwise non-canonical promoter [174,173,172].

In Figure C1, we show the *dso* regions of ϕfd and ϕIKe without the possible additional stemloop structures D and E – encompassing four gpII-binding sites α through δ – because they are irrelevant for initiation by gpII on the doublestranded RF (compare to Fig. 1 in [181]). Initiation by nicking occurs after the sequential binding of gpII to all 4 binding-sites, which would be impossible if α + β form part of stemloop D, and γ + δ part of stemloop E, respectively, on the viral (+)- strand DNA [151,170]. The (putative) stemloops D + E might nevertheless have a biological role, e.g. preventing untimely binding of gpII and nicking of the viral ssDNA. We note that also in a number of *dso*s of RCR plasmids the nick-sites are embedded in (putative) stemloop structures [208].

According to Konings and co-workers, the filamentous phage IKe – infecting host cells carrying IncI2 plasmids – is a distant relative of the F-specific phages fd, f1 and M13 [343]. The genome organisation of ϕIKe is strikingly similar to that of ϕfd in the gene IV/gene II region. In addition, two short stretches of (almost perfect) sequence homology between ϕIKe and ϕfd allowed the straightforward identification of stemloops B + C (*sso*) and the nick-site region (*dso*) (Fig. C1; B). Noteworthy, the individual motifs are separated by intervening sequences of different lengths and varying sequences. This observation emphasises that *sso* and *dso* have to be considered independent structural motifs. The co-localisation of these motifs in ϕfd (ϕf1, ϕM13) seems to be rather the exception than the rule: *sso* and *dso* do not co-localise in ϕX174 and related phages (see below), and also not in various RCR-plasmids [208].

Together with gpII of ϕfd and ϕIKe, the (putative) initiator protein gpII of phage I2-2 belongs to group 3 of RCR-initiators (Section C3.1.1.; Table C9). Except for a comparable size and the conserved 'active tyrosine'-motif 3 however, ϕI2-2 gpII is unrelated to ϕfd gpII (<20% ident. res.) but shows instead significant homology to the orf1(Rep) protein of *Xanthomonas campestris* RCR-plasmid pXV64 (36% ident. res.) [476], and to the RepC protein of *Moraxella sp.* RCR-plasmid pTA144up (39% ident. res.). Konings and co-workers assumed therefore that ϕI2-2 inherited its replication module from a RCR-plasmid [427]. However, the genome organisation of ϕI2-2 is comparable to that of ϕfd and ϕIKe: the genes IV and II are separated by an intergenic region of 835 nt. The part proximal to gene II can be classified as 'AT-rich' (65% A+T) in contrast to the distal part (51% A+T), and a pu-tative stemloop structure analogous to stemloop A could be readily identified directly downstream of gene IV (motifs as in Fig. C1). The few short and patchy sequence homologies to either ϕfd or ϕIKe did not allow pinpointing the putative *sso* and *dso* structures (not shown). This example demonstrates that RCR-origin structures cannot always be simply deduced by comparing nucleotide sequences of (more distantly) related phages.

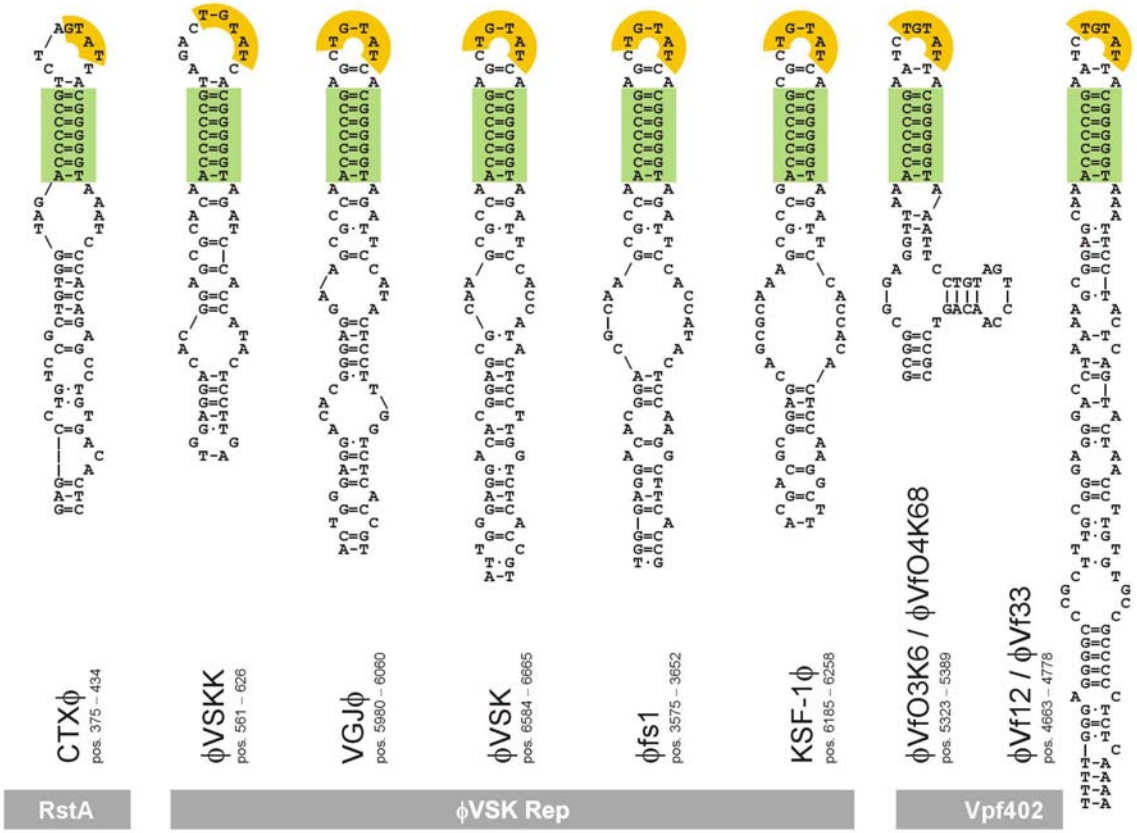

Fig. C3 (Putative) replication origin (*dso*) of CTXϕ and related vibriophages.
Secondary structure of the region containing the conserved *dso*-motif (light green) of CTXϕ taken from Moyer *et al*. [306] with slight modifications; the putative nick-site is shown in orange. The positions of the regions shown were taken from the entries for CTXϕ [(partial sequence) VCU83796], ϕ VSKK [NC_003311], VGJϕ [NC_004736], ϕ VSK [NC_003327], ϕ fs1 [NC_004306], KSF-1ϕ [NC_006294], ϕ Vf12 [NC_005949], ϕ Vf33 [NC_005948], ϕ VfO4K68 [NC_002363], and ϕ VfO3K6 [NC_002362]. Protein names given in the grey bar in the lower part refer to (putative) initiator subgroups as discussed in Section C3.1.1. .

The *dso* and *sso* replication origins of the isometric phage ϕX174 of E. coli. The *dso* of ϕX174 is located at a distance of ~1900 nt from the *sso*. The *dso* of ϕX174 was mapped to a discrete 30 bp stretch within the A gene (see Table C1; corresponding to amino acid res. 107–116) and the position of the nick-site has been determined as A_{4307} [14,187,115]. The *dso*s of phages ϕX174, ϕS13, ϕG4, ϕU3, ϕG14, ϕK, ϕα3, and ϕSt-1 are virtually identical, and all are recognised by ϕX174 A protein (Table C1) [166]. Although the A proteins of the isometric phages are close homologues (similarity ranging from 43% to 96%; see group 2A in Table C8), the nucleotide sequence similarity drops significantly to either side of the 30 bp *dso* core sequence. The ϕX174 *dso* lacks detectable secondary structures or sequence repeats, and the binding-site(s) of A protein is not known exactly.

Despite their uniform *dso*s, the isometric phages posses different *sso*s, and also the mechanism for complementary strand synthesis differs from that of the filamentous phages (see above and BRM Section 2.1.). Shlomai and Kornberg could show that efficient *in vitro* complementary strand synthesis of ϕX174 depends on the host PriA protein and assembly of the restart primosome on a primosome-assembly site (PAS) located between the genes F and G (Fig. C2; A) [413]. The sequence containing the PAS site is identical in ϕX174 and in ϕS13, and ϕS13 complementary strand synthesis most likely follows the ϕX174 pathway. In contrast, *in vitro* complementary strand synthesis of ϕG4 depends on the synthesis of a specific ~26–28 nt long primer by *E. coli* DnaG [41]. Bouché *et al.* could locate the template of the RNA primer to a stemloop structure in the intergenic region between the genes coding for the major coat protein and major spike protein (Fig. C2; B). Despite considerable protein sequence similarity between ϕX174 and ϕG4 for F protein/ major coat protein and G protein/ major spike protein, respectively, the nucleotide sequen-

Table C1 Double-strand origin (*dso*) of ϕX174 and related phages.

ϕ	gene	acc.	5'-flanking	pos.	*dso*	pos.	3'-flanking
ϕ X174	A	NC_001422	AGTGCTCCCC	4299	CAACTTG↓ATATTAATAACACTATAGACCAC	4328	CGCCCCGAAG
ϕ S13	A	NC_001424	....	382	.......↓.....	411	.CG.....
ϕ G4	A	NC_001420	C......GGA	500	.......↓.....	529	AAA...CT..
ϕ U3	-	M10630	C......GGA	125	.......↓.....	154	AC...AA...
ϕ G14	-	M10632	.......GGA	55	.......↓.....	84	AC...TA...
ϕ K	A	NC_001730	GTGTGCAG..	1001	.......↓...A..G.....	1030	...A..CC.T
ϕ α3	A	NC_001330	GTGTGCTG..	1002	.......↓...A..G.....	1031	..TA..CC.T
ϕ St-1	-	[1])	GTGTGCTG..	-	.......↓...A..G.....	-	..TA

Identity with the nucleotide sequence of the ϕX174 gene A sequence is indicated by '.' . The nick-site is indicated by an arrow. [1]) partial sequence taken from Heidekamp *et al.* [166].

ce and (putative) secondary structure elements in the intergenic regions are widely different. The *sso*s of the closely related phages ϕK and ϕα3 are not known, but – by analogy to ϕX174 and ϕG4 – a (putative) secondary structure in the intergenic region between the genes coding for the major coat protein and major spike protein could serve this purpose (Fig. C2; C).

The ϕX174 A protein is distantly related to the (putative) replication protein encoded by orf4 of ϕchp1 (Table C8; groups 2A + 2B). The replication of the small chlamydial phages has yet to be studied, and their replication origins remain to be determined.

The *dso* of *V. cholerae* phage CTXϕ. The small ssDNA phages were instrumental in developing the present models for prokaryotic replication and have been widely used as smart cloning vehicles. However, their biological importance was generally underestimated until it became known that they can be responsible for the spread of pathogenicity factors: e.g. the cholera toxin genes are disseminated among *Vibrio cholerae* strains by a filamentous phage, CTXϕ [95]. Waldor and co-workers analysed the replication of CTXϕ, which may reside in (mostly multiple) tandem copies as prophage in the genomes of *V. cholerae* strains. The mechanism of CTXϕ prophage formation in *V. cholerae* is not known, but lysogenic host cells release mature phages [379]. Using an elegant genetic technique developed by Horiuchi [179], Waldor and co-workers could show that a (presumed) *dso* of CTXϕ present twice on a test plasmid is reduced to a single chimeric origin after replication [306]. This result offered a (mechanistically) simple explanation for the puzzling observation that unit-length closed-circular viral (+)-strands were produced from tandemly arranged chromosomal copies: initiation of RCR at the first *dso* is followed by ring closure at a second *dso* in the 'prophage array' (BRM Section 2.1.). Supporting these results, the phage-encoded RstA protein shares the 'active tyrosine' motif 3 with numerous initiators for RCR of phages and plasmids (Section C3.1.1.). We note, however, that neither the function of RstA as initiator for RCR has yet been proven biochemically, nor the position and functionality of the presumed nick-site confirmed.

A number of filamentous vibriophages encode (putative) initiators for RCR with significant similarity to the CTXϕ RstA protein (see Section C3.1.1. and Table C8; subgroup 1A). We were interested whether also the *dso* of CTXϕ is conserved in these phages. Nucleotide sequence comparison with the 'stemloop 2' defined by Moyer *et al.* [306] as query led to the detection of corresponding structures in all phage genomes (Fig. C3). Although we show the conserved structures as stemloops in Figure C3, we should cautiously speak of 'stretches with dyad symmetry' rather, until their stemloop structures are experimentally proven. The conserved 'stemloop structures' are embedded in an only partially conserved context but are invariably located in intergenic regions in the respective genomes. Although not a stringent proof, the presence of a highly conserved 'stemloop 2'-like structure in all compared phage genomes encoding similar initiator proteins supports the hypothesis that they constitute the *dso*s of these phages. A *sso* has yet to be identified for CTXϕ and the other vibriophages.

The *dso* of *E. coli* phage P2. Besides the group of phages with small single-stranded DNA genomes (<10 kb for the RF) that replicate via RCR, a second group of phages with mid-sized dsDNA genomes (~30 kb) is known to replicate by this mechanism. The best known phage of the latter group is *E. coli* phage P2, for which the nick-site of the dso could be located within the 3'-end of gene A encoding the initiator protein

Table C2 Double-strand origin (*dso*) of ϕP2 and related phages.

ϕ / pϕ	gene	5'-flanking	pos.	*dso*	pos.	3'-flanking	ident.	translation [1])
ϕ P2	A	CGGCATCGCC	29885	GCGC**CT**C**G**↓**GA**GTCC**TG**TC**A**A**TAACTGT**	29911	GGAAAGCTCA	x	AAPRSPVNNC
ϕ L-413C	gpA	. . . . .	28691	. . . . . ↓	28717	. . . . .	97%	. . . . .
ϕ Wϕ	gpA	. . . . .	28077	. . . . . ↓	28103	. . . . . CTG	96%	. . . . .
Y. pseudo. pϕ	yptb1766	. . CATCT.G.	2119424	. . . . . ↓	2119398	. . GTTAGGTG	56%	. . . . .
ϕ PSP3	gp36	. . CTCCT. . G	28618	C.CT. . . . ↓ . . . . T.	28644	ACGGGAAG.G	39%	.PS. . S. . . .
ϕ 186	A	. . CTTCT. . G	28568	C.CT. . . . ↓ . . . . T.	28594	ACGGGAAG.G	37%	APS. . S. . . .
ϕ Fels-2	stm2729	G. . . GCTTG.	2870666	. . C. . . T. ↓ . . C. . G. . G.	2870640	CCCCTTGCTG	34%	C. . WTRG. . .
pϕ CP-933T	z2978	. . . . CCT. . .	2676644	. . C. . . T. ↓ . . C. . G. . G.	2676670	CCCCGTG.AC	31%	A. . WTRG. . .
ϕ HP2 / ϕ HP1	Rep	. . CACG.AGT	7444	. . . . . T. ↓ . . C.TG. . . . . G.	7470	AACCGCTCA.	31%	VERS.TA. . S
ϕ PM2	p12	. . A. . G. . A.	3504	. . C. . . T. ↓ . . . CA. . . AA.	3530	AACC.C.C.G	33%	D. . W.TE. . .
ϕ fs2	orf716	ACA. . ATTGG	2357	.T. . . . T. ↓ . . C. . G. . . . . T.	2383	C.CC.TA.TT	33%	IGCLGL.SLT
ϕ V86	-	C. . . TCT. . .	5922	. . CT. . . . ↓ . . . . T.	5948	ACG [2])	36%	. . S. . S. . . .
ϕ K139	Rep	C. . . TCT. . .	10212	. . CT. . . . ↓ . . . . T.	10238	ACGG.T.C.T	31%	. . S. . S. . . .

The nick-site in the ϕP2 *dso* (within the coding region for A protein) determined by Liu and Haggård-Liungquist is indicated by an arrow [257]. Identity with the nucleotide sequence of the ϕP2 gene A sequence is indicated by '.' ; bold letters (in the ϕP2 A lane) indicate positions conserved in all 13 sequences . ident. = BLAST similarity (% ident. res.) as in Table C9; group 4 . 1) reading frame in the *dso* region; identity with ϕP2 A protein is indicated by '.' . 2) end of sequenced region. Sequences were taken from: ϕ P2 [NC_001895], ϕ L-413C [NC_004745], Wϕ [NC_005056], *Y. pseudotuberculosis* pϕ [NC_006155], ϕ PSP3 [NC_005340], ϕ 186 [NC_001317], ϕ K139 [NC_003313], ϕ V86 (partial sequence) [AF008938], Fels-2 pϕ [NC_003197], ϕ fs2 [NC_001956], pϕ CP-933T of *E. coli* O157:H7 EDL933 [NC_002655], ϕ PM2 [NC_000867], and ϕ HP2 [NC_003315] (the DNA sequence of ϕ HP1 [NC_001697] is identical in this region). For each (putative) *dso*, the relative position within the gene is shown as projection onto the protein sequence in Fig. C6 .

[256,257] (BRM Section 2.1.).In Section C3.1.1., we discuss the similarity of 12 known phage-encoded homologues of ϕP2 A protein. Here we can show that a stretch of 27 bp around the ϕP2 nick-site is well conserved in all 13 initiator genes, suggesting an identical nick-site in all sequences (Table C2). This 27 bp 'core' is contained within the region of 35 bp suggested by Liu and Haggård-Ljungquist as *dso* by comparison of the ϕP2 and ϕ186 sequences with other phage or plasmid *dso*s [257]. Except for the ϕL413-C and Wϕ genes that are (almost) identical with the entire ϕP2 A sequence, the similarities drop immediately on either side of the 27 bp 'core' sequence. This is reminiscent of the *dso*s of ϕX174 and related phages, where the conservation of the nucleotide sequence was also confined to the 30 bp 'core' (see above). The 'ϕP2- type' *dso* shares more features with the ϕX174-type *dso*: i. the 3'-end of the nick contains the conserved CTXG motif also present in a number of plasmid *dso*s, ii. all sequences lack detectable direct or inverted repeats within the *dso* and the adjacent regions, and iii. they also lack a detectable AT-rich region. In contrast to the ϕX174 A gene however, the (putative) *dso*s are located close to the 3'-end of the genes. Except for ϕf2, ϕV86, and ϕK139, the putative *dso*s are located at exactly the same position in the genes as the ϕP2 *dso* (the position of the putative *dso*s within the initiator genes is shown in Fig. C6).

Given the reasonable protein sequence similarity of the ϕP2 A homologues and the conserved position of the *dso*s, the observed nucleotide sequence conservation of the putative *dso*s may simply reflect the requirement for conservation of structurally/functionally important amino acid residues. The degree of conservation of the amino acid residues in the *dso* region reflects the overall similarity of the proteins, and does not allow for a clear decision (Table C2; rightmost column). However, alternative 'translation' of the ϕf2 *dso* in the −1 frame results in a sequence WVPWTRVI-NC, which is closer to the ϕP2 A sequence. This indicates that the conservation of the nucleotide sequence is important, supporting the tentative assignment of the double-strand origin to this part of initiator genes of the 13 phages.

Vibriophage ϕfs2 is a typical filamentous phage – only distantly related to ϕfs1, though (see above) – with a ssDNA genome size of 8.7 kb (RF). It is the first example for a small phage replicon with a ϕP2-type replication module for RCR. Also *Alteromonas* phage MP2 has a relatively small genome (10.1 kb dsDNA). ϕMP2 is unique among the known bacteriophages because its capsid contains lipids [272]. However, its initiator protein p12 (Section C3.1.1.) and its (putative) *dso* are clearly related to the 'ϕP2-type'. Both examples demonstrate that the ϕP2-type replication module for RCR is not confined to the closer relatives of *E. coli* phage P2. However, plasmid replicons with a ϕP2-type module are not (yet) known, which suggests that this module is not promiscuously distributed among different genetic elements. We note that neither for ϕP2 nor for any other of

the mentioned phages the *sso*s have yet been determined.

Conclusions. Despite accomplishing identical functions and containing the nick-site for initiation of RCR, the double-strand origins discussed above differ considerably with respect to their specific structural elements. The ϕfd *dso* contains repeats as binding sites for the initiator, flanking the nick-site. The *dso* of CTXϕ contains a region of dyad symmetry that may form a stemloop, with the nick-site in the loop. Neither of the above-mentioned structural elements could be detected in the ~30 bp long *dso*s of ϕX174 and ϕP2. There are thus three clearly distinct types of phage *dso*s for RCR, but the more subtle differences in the initiation and termination reactions that can be expected thereof remain to be studied.

The *dso*s of the phages also differ with respect to their localisation within the replicon. The *dso*s of the filamentous *E. coli* phages are located directly upstream of gene II encoding the initiator in an intergenic region. Also the *dso*s of the vibriophages are located in intergenic regions, but unlinked to the initiator gene. In contrast, the *dso*s of the isometric phages and of the ϕP2-group are located within the coding region of the initiator: the ϕX174-type *dso*s in the 5'-region, and the ϕP2-type *dso*s in the 3'-region, respectively. A localisation of the *dso* within the cognate initiator gene is a feature also found in plasmids replicating by RCR, e.g. in pT181 and related plasmids (Table C8; group 1B) [210]. Origin-containing initiator genes are known, in addition, for phage replicons propagating in the θ-mode, e.g. the lambdoid phages (see below). Although this 'double layer' of genetic information imposes strong selective pressure on the initiator gene – simultaneous selection for a functional origin and for a functional protein – we have to realise that this intricate organisation is apparently evolutionary successful.

C2.2. Origins for DNA replication in the 'theta-mode' (θDR)

All known prokaryotic origins for DNA replication in the theta(θ)-mode have a strikingly similar architecture irrespective of whether they drive the replication of plasmids, phages, or chromosomes (Fig. C4). In the generalised 'origin design', a stretch of 100-200 bp containing a number of short repeated sequence motifs (iterons) is located on one side of an AT-rich region of ~50 bp in length. Initiation occurs after binding of the cognate initiator protein to the iterons on a negatively supercoiled template: concomitant with oligomerisation of the iteron-bound initiators, a long-range distortion of the DNA occurs that is sufficiently strong to promote the 'melting' of a short stretch of neighbouring bases within the AT-rich region. The partially single-stranded (unwound) region serves as entry site for primosomal and replisomal proteins, culminating in the establishment of a replication fork(s).

Two types of iterons can be discriminated in origins for θDR: inverted repeats (*ori*λ) or direct repeats (pPS10 *oriV*, *E. coli oriC*) (Fig. C4). Direct repeats can be perfect (*oriV*) or degenerate (*oriC*). They can be arranged head-to-tail (*oriV*) or with irregularly alternating orientation (*oriC*). Regularly alternating orientations of direct repeats would appear as array of larger (imperfect) palindromes. Iterons of the inverted repeat type can be either perfect or imperfect palindromes, as in *ori*λ (Fig. C4; A). We note that although any of these motifs are usually detectable in putative origin sequences at first glance their correct classification requires knowledge of the binding-mode of the cognate initiator protein. As a rule of thumb, initiators bind as monomers to direct repeats (RepA→*oriV*, DnaA→*oriC*), while dimeric initiators bind to inverted repeats (O→*ori*λ). Iterons of either type can be regularly spaced (*oriV*, *ori*λ) or irregularly spaced (*oriC*). A regular iteron spacing can coincide with helical phasing (*oriV*: 4 iterons with a length of 22 bp, directly adjacent to each other = 8 helical turns), but also may deviate from it. It could be shown for phage λ *ori*λ and *E. coli oriC* that disruption of the phasing of the iterons had a more pronounced (negative) effect on origin function than had their mutation [299,291,484,239]. This latter observation emphasises that the spacing of the iterons is dictated by the architecture of the nucleoprotein complex that is responsible for origin unwinding. Even in the best-studied model systems however, the architecture of this nucleoprotein complex is not known to the necessary detail – present efforts concentrate on the elucidation of the initiator-initiator interactions and on describing the path of the origin DNA within the complex (see e.g. [250]).

In addition to the iterons that serve as binding-sites for the initiator protein, many origins contain additional repeats for which a function is not yet known (*ori*λ, *oriV*). They are mostly direct repeats and their consensus sequence differs from that of the iterons. These 'secondary repeats' are often located more closely to (or within) the AT-rich region than the initiator-binding sites. The observation that *E. coli* DnaA binds to such repeats in the AT-rich region of *oriC* when they become single-stranded during unwinding has so far not been repeated in other systems and may be a unique property of the DnaA-type of initiators (Section C3.1.2.) [423].

The regulation of replication initiation at origins for θDR is as strikingly different as is their overall architecture similar. In the pPS10 system, two types of binding-sites for the RepA initiator exist that allow for the regulation of initiation in response to the oligomerisation

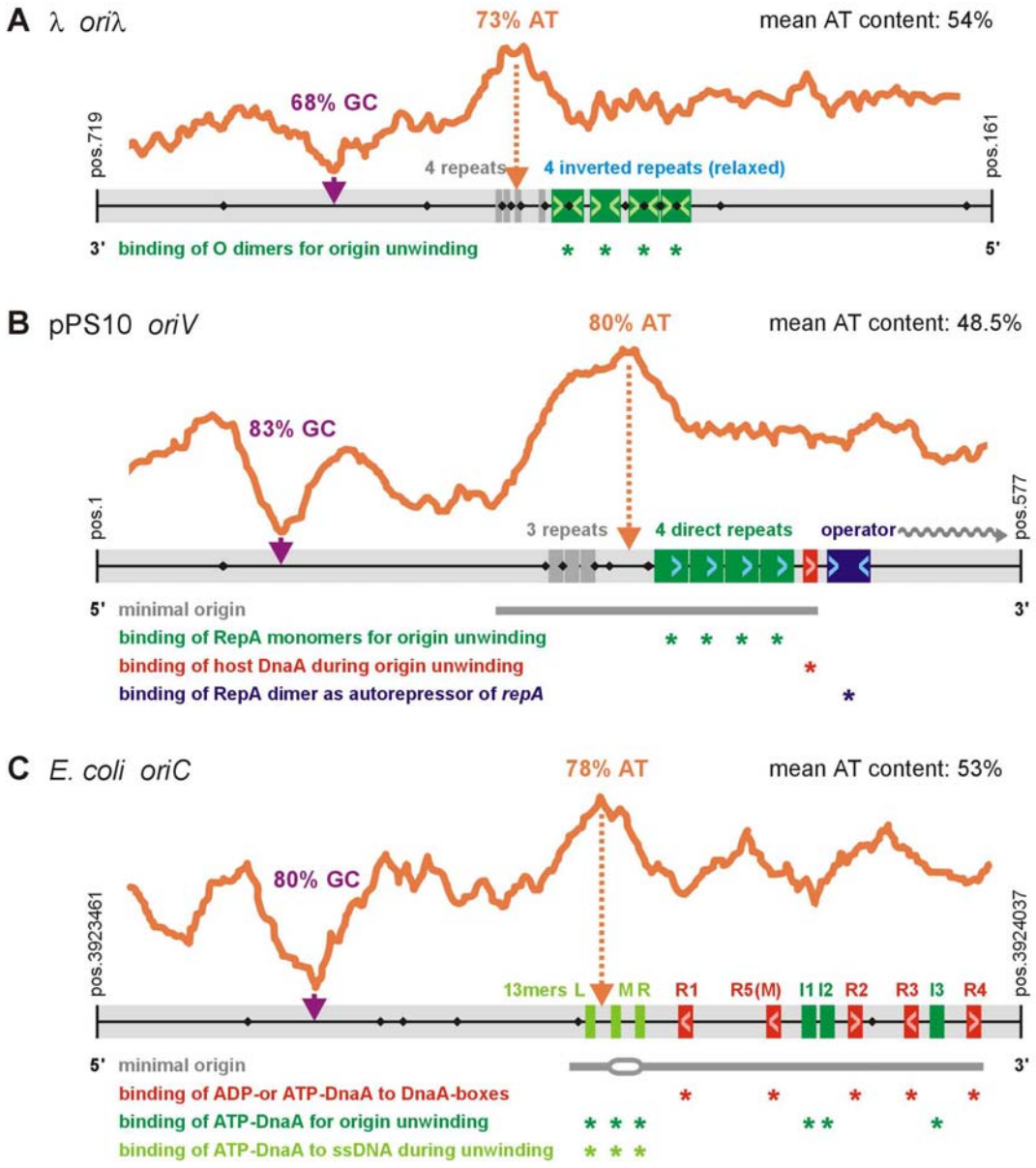

Fig. C4 Structural elements of three origins for replication in the θ-mode.
Origin regions are shown as grey bar, structural elements as coloured boxes; all sequence motifs are shown exactly to scale. Red boxes = 'canonical' DnaA binding-sites (DnaA boxes [389, 239]); dark green boxes = binding sites of the cognate initiator protein that are crucial for origin opening (I-boxes in the case of *oriC* [250]); light green boxes = 'ATP-DnaA boxes' bound as ssDNA by ATP-DnaA during origin unwinding [423]; blue boxes = initiator binding sites not required for origin unwinding; grey boxes: repeats of unknown function. '<' and '>' indicate directionality/dyad symmetry within the protein-binding sites. The function of a DNA-binding protein at a particular binding-site is indicated by '*' and additional explanatory text. '♦' indicate stretches of ≥4 A (or T bases, respectively). Note that in **A** the conventional 5'→3' orientation of the sequence was reversed to allow for a better comparison with **B** and **C**. AT-plots (orange) were computed by the spin v1.1 software from the Staden Package [425], using a window of 41 nt (therefore, the plots are shown for -20 bp from either end only). The AT-plots in **A**, **B**, and **C** are shown in the same scale with the upper edge of the sequence bar as baseline. The positions of the (partial) sequences analysed for AT-content are shown. The genetically defined 'minimal origin' is indicated by a dark grey bar in **B** and **C**. **A** : partial sequence of the phage λ *o* gene [J02459] (see text for details). **B** : partial sequence of plasmid pPS10 [X58896] (for details see [143]), the genetically defined 'minimal origin' is indicated by a dark grey bar. **C** : partial sequence of the *E. coli* K12 chromosome [U00096] (for details see [290, 250]), The genetically defined 'minimal origin' is indicated by a dark grey bar and the unwound region by an open bubble [232].

state of RepA (reviewed in [143]). Both binding-sites share the motif GGACAG, which is present once in each *oriV* iteron, and – as inverted repeat – twice in the operator of the *repA* promoter (Fig. C4; B).

Monomeric RepA binds to the *oriV* iterons. Dimeric RepA acts as autorepressor: *repA* transcription is repressed by binding of a RepA dimer to the operator site. In contrast to the *oriV*/RepA paradigm, binding of DnaA to *oriC* is not regulated by the type of the iterons. The DNA-binding *and* the oligomerisation properties of *E. coli* DnaA depend crucially on the bound nucleotide: ADP-DnaA and ATP-DnaA bind as monomers – and with comparable affinity – to DnaA boxes that match the consensus TTATNCACA (Fig. C4; C: R1, R2, R4) [389]. Binding to 'low affinity' DnaA boxes that deviate from the consensus (Fig. C4; C: R3, R5) occurs only when a close-by 'high affinity' DnaA box promotes cooperative binding, and requires ATP-DnaA. Also the non-canonical binding sites (Fig. C4; C: I-boxes, ATP-DnaA boxes in the L, M, R 13mer region) are only bound by ATP-DnaA and require neighbouring 'high affinity sites' for cooperative binding [424,423,250]. Different from the above systems, λ O protein binds as dimer to single iterons, and further oligomerisation occurs among iteron-bound dimers (see below). A specific role for the O monomer and iteron half-sites as targets for binding of O monomers are not known. Unlike when bound to DNA, O is highly unstable in solution. Therefore the initiation of λ replication may be simply regulated by the availability of O, and may not involve more intricate regulatory loops as known for the initiators of plasmids and chromosomes.

The AT-rich region is a prominent feature in the three origins shown in Figure C4 and also readily detected in many replication origins for θDR. In some origins that have been studied experimentally (*Saccharomyces cerevisiae* ARS, Simian Virus SV40), spontaneous unwinding of negatively supercoiled DNA occurs close to AT_{max} in the absence of divalent cations – not necessarily at *exactly* the same position where initiator-induced unwinding occurs [460,252]. This region has been termed 'DNA unwinding element' (DUE) therefore, and it could be shown for *E. coli oriC* that it is the helical 'instability' that is required, and not the exact sequence [231]. In all three origin sequences shown in Figure C4, the AT-rich region is flanked on one side by the iteron-region, and on the opposite side by a GC-rich region. Although this 'GC-peak' adds to the overall similarity of the three origin regions, it does not play a (known) role for origin functioning and is not part of the genetically defined 'minimal' origins of pPS10 and *E. coli*.

A feature of origin sequences that is important for the DNA distortion required for unwinding is sequence-induced (intrinsic) curvature. Intrinsic curvature is detectable by an abnormal electrophoretic behaviour of DNA in gels, and was first described for the minicircle DNA of *Leishmania tarentolae* kinetoplasts. *ori*λ DNA was the first prokaryotic DNA for which curvature could be demonstrated [499,500]. Intrinsic curvature of DNA can in many instances be predicted for a sequence with periodically spaced A runs of ~5-6 bp in length. This is the case for *ori*λ, but not for *oriV* and *oriC* (see Fig. C4; '♦' in the sequence bar). Minimal *oriC* DNA has a close-to normal electrophoretic behaviour that varies however slightly depending on the degree of *dam* methylation (H.Seitz and C.Weigel, unpublished). A functional role has not been demonstrated for the region with significant intrinsic curvature immediately downstream of minimal *oriC* (not shown in Fig. C4; C) [215]. Intrinsic curvature is thus not a sufficient criterion for the detection/definition of replication origins. Initiator-induced unwinding requires negative superhelicity of the origin DNA. We refer the reader to BRM Section 2. for a detailed discussion of this topic.

The three examples given in Figure C4 show origins for θDR which differ with respect to localisation in their cognate replicon. The phage λ replication origin, *ori*λ, is located within the *O* initiator gene, *oriV* of pPS10 upstream of the *repA* initiator gene, and *E. coli oriC* at a distance of ~40 kb from the *dnaA* gene – in contrast to *B. subtilis* where *oriC* and *dnaA* co-locate. As already pointed out in the previous chapter, there is no preference for origin localisation in prokaryotic replicons.

The replication origin *ori*λ of phage λ. The λ initiator protein O binds to the relaxed palindromic iteron $\underline{\text{ATCCC}}T_{/C}AAAA_{/T}C_{/T}G_{/A}G_{0\text{-}1}R\underline{\text{GGGAT}}_{/AC}$ present four times in *ori*λ as part of the *O* gene (Fig. C4; Table C3). The iterons are arranged as an array with exact helical phasing, and 5'- upstream of the AT-rich region at a distance of approximately two helical turns (23 bp). A considerable variation exists among the number of iterons and their lengths even among phage initiators with highly conserved protein sequence: there are e.g. three iterons in gene 15 of ϕ80, but 6 in the *O* gene of ϕH19-B. In both cases, the 'backbone' of the relaxed palindromic iteron is highly similar to the *ori*λ iterons and iteron spacing corresponds exactly to helical phasing, i.e. one helical turn (Table C3). It could be shown for *ori*λ that only three of the four iterons are absolutely required for origin function [299]. The phenomenon that initiator-binding sites defined by footprinting may be disabled without grossly impairing origin function is also known for *E. coli oriC* [23,239,472]. As a consequence, the number of iterons detected in *ori*λ and other phage origins for θDR (see below) is probably higher in most cases than the minimal number required for origin functioning. The interaction of λ O with a single iteron and with the entire *ori*λ is discussed in more detail in Section C3.1.2. .

Table C3 (Putative) replication origins of Gram(−)-specific phages; part A.

(pro)phage	gene	nt	AT	consensus sequence of iteron(s)	nt	#	groups	type	spacing	pos.	d
ϕ D3	orf73	990	424	TGAGACCAAAWCWKCTCACTC	21	2	1+2+1	rIR [1])	-	5'	13
				TGAGC-$N_{8/13}$-TCTC	17/22	2		rIR	-		
ϕ P22	gp18	816	451	GGTAAAA-N_{1-6}TWACC	13-18	3	3	rIR	-	5'	25
ϕ 80	15	909	466	TTCCCGAAAWC-N_{0-1}GGWA	15-16	3	3	rIR	5 nt / 2 t	5'	24
λ / ϕ 21	*O*	900	459	ATCCCT$_{/C}$AAAA$_{/T}$C$_{/T}$G$_{/A}$G$_{0-1}$RGGGAT$_{/AC}$	19-20	4	4	rIR	1 t	5'	23
ϕ 4795	O	936	496	ATCCCT$_{/C}$AAAA$_{/T}$C$_{/T}$RGGGA$_{/G}$W	19	4	4	rIR	1 t	5'	22
ϕ H-19B	O	939	493	CCCTC$_{0-1}$AAAA$_{/T}$YG$_{0-1}$AGGG	14-15	6	6	rIR	1 t	5'	22
ϕ 933W	O	939	490	CCCT$_{/C}$AAAA$_{/T}$C$_{/T}$(A$_{/G}$)$_{0-1}$A$_{/G}$GGG	14-15	6	6	rIR	1 t	5'	19
cpϕ CP-933V	z3356	939	496	CCCT$_{/C}$ (C$_{/T}$)$_{0-1}$A$_{/C}$AAAC$_{/T}$G$_{/TA}$A$_{/G}$GGG	14-15	6	6	rIR	1 t	5'	24
cpϕ CP-933X	z1868	744	376	TTACCC$_{/T}$G$_{/A}$AAT$_{/A}$C$_{/G}$AGGTAA	17	5	5	rIR	1 t	5'	17
pϕ LambdaSo	O	1017	385	–							
			478	ATTGTCRRAACCGAYAG	17	2	2	rep	2 t	5'	21
P. putida	pp3894	831	409	GTGT$_{/G}$A$_{/T}$GAAA$_{/T}$A$_{/CT}$B$_{0-1}$W$_{0-1}$A$_{0-1}$CAC$_{/G}$	13-16	3	3	rIR	–	5'	5
ϕ Nil2	o	822	442	GCT$_{/C}$N$_{7-11}$AGC	13-17	4	1+3	rIR	2 t	5'	29
ϕ HK97	gp54	822	446	GCT$_{/C}$N$_{7-11}$AGC	13-17	4	1+3	rIR	2 t	5'	33
cpϕ Gifsy-2	stm1014	984	628	ATCCCYGAAAAC$_{/G}$RGGGAW	18	6	4+3	rIR	2 t	5'	22
				GTACCTGAAACCAGGGGAT	18	1					
cpϕ CP-933P	z6070	978	433	TC$_{/T}$AAMDG$_{/T}$TTGA	11	4	8	rIR	6 nt	5'	14
				C$_{/G}$A$_{/G}$AA$_{/T}$T$_{/A}$AT$_{/A}$TG$_{/A}$	9	4					
			502	GMASAKTKG	9	2	3	rIR	6 nt	5'+3'	3/2
P. putida	pp1551	855	415	AAACMG	6 [2])	4	2+2	rep	1 t/4 t/1 t	5'	19
ϕ Aaϕ23	O	783	409	AC$_{/T}$AC$_{/AT}$T$_{/A}$VG$_{/C}$NT$_{/AG}$T$_{/AG}$A$_{/G}$	11	7	-	rep	-	5'(+3')	1/0
				AC$_{/T}$AT$_{/A}$VG$_{/C}$NT$_{/AG}$T$_{/AG}$A$_{/G}$	10	1					
ϕ P27	L17	927	493	T$_{/GC}$CA$_{/T}$G$_{/A}$AT$_{/A}$T$_{/C}$C$_{/AT}$T$_{/A}$	9	8	1+7+2	rep	1.5 t	5'	18
				KCAK$_2$WSKW	9	2					
ϕ SfV	orf39	819	400	A$_{/C}$TTCAN$_{7-10}$GTG$_{0-1}$CA	17-20	5	6	rep	~2 t	5'	16
				TTGCACCAGGGGGTAGTGC	19	1		rep			
ϕ ST64B	sb42	819	403	A$_{/C}$TTCAN$_{7-10}$GTG$_{0-1}$CA	17-20	5	6	rep	~2 t	5'	19
				TTGCACCAGGGGGTAGTGA	19	1		rep			

The replication origin *ori*L of phage ϕSPP1. The approximate localisation of the ϕSPP1 replication origin could be inferred from analysis of phage replication mutants with an 'immediate stop' phenotype and mapped to a region containing the genes 38, 38.1, 39, and 40 [284,55,341,475]. Alonso and co-workers were the first to show that the gene order in the replication origin region of ϕSPP1 closely resembles that of other lambdoid phages [341]. They could demonstrate specific binding of G*38*P protein to a 393 bp DNA segment from within the coding sequence of gene 38. A prominent AT-rich region approximately in the middle of gene 38 is flanked 5'-upstream by an array of five direct repeats (AB boxes) with the consensus sequence $A_{/G}G_{/A}AAC_{/A}GA_{/G}C_{/A}A_{/T}$ (Table C5). Together with the single intervening base pair, the 10 bp long iterons are aligned in a phasing of exactly one helical turn per iteron. Footprint analysis confirmed that the iterons are the binding-sites for the G*38*P initiator protein (Section C3.1.2.) [284,297]. The importance for origin function of the two more degenerate 'ab boxes' located at a distance of 47 bp upstream of the major iteron cluster is presently not known (Table C5).

A second origin-like structure, *ori*R, is present in the ϕSPP1 genome at a distance of ~32 kb from *ori*L. Compared to *ori*L, *ori*R lacks one AB box in the major iteron cluster, and although G*38*P binds efficiently to *ori*R, unwinding does not occur [341,297]. Alonso and co-workers could show that *ori*R is instrumental for the switch from θDR to σDR during replication of ϕSPP1. In their model, DNA-bound G*38*P 'initiates' this switch by blocking replication fork progression through *ori*R (BRM Section 2.2.) [12].

The replication origin *oriJ* of the *E. coli* Rac prophage. Díaz and Pritchard identified the replication ori-

Table C4 (Putative) replication origins of Gram(−)-specific phages; part B.

(pro)phage	gene	nt	AT	consensus sequence of iteron(s)	nt	#	groups	type	spacing	pos.	d
cpϕ CP-933M	z1337	1203	487	AAG$_{3-4}$CTT$_{/G}$	8-9	3	3	rIR	1 t	5'	161
			643	AAG$_4$CTT	9	3	3	rIR	1 t	5'	60
				AAGAACAGGAACAGGA	16	2	2	rep	2 t	5'	14
ϕ Bcep22	gp26	774	391	C$_{/G}$CC$_{/A}$C$_{/T}$T$_{/C}$T$_{/GC}$C$_{/GT}$GG$_{/A}$A$_{/G}$A$_{1-2}$M-G$_{/CA}$HCG	16-17	5	1+4	rep	3 t + 3x 2 t	5'	24
ϕ E125	gp60	993	435	AAAG$_{/C}$GGTGCA	10	4	9	rep	-	5'	11
				GGGTTCA	7	3					
				TGCACCC (= inverted GGGTGCA)	7	2		rep[5])			
P. luminescens	plu3473	753	250	YTGA$_{/C}$AC$_{/T}$C$_{/A}$	7	6	-	rep	-	5'+3'	10/5
B. bronchis.	bb1683	957	415	CACCCCCGCA	10	7	7	rep	-	5'	27
cpϕ CP-933O	z2048	1041	346	GGCAGAACAAGCA|AATTGTGGTGTACC, AGTGTTGCCGAG|CCTGAATACACC, GGTGCTGTTAAT|AAAAACAGTACC, GGTTATGTTTCAGCACAAAACAGACC	24-28	4	4	rIR	-	5'+3'	-
S. enterica	sty2063	789	421	TTGTAACGCTCGCGGCGTTACAA, GTTACAAACG|AGCGAGTTAC, GTAACGCTGCCAGCGTTAC, AGCGTTAC	19-20	3,5	3+$^1/_2$	IR	-	5'	11
cpϕ CP-933R	z2397	789	421	TGTAACGCTCGCGGCGTTACA, GTAACGCTGCCAGCGTTAC, AGCGTTAC, GTTACAAACG|AGAGAGTAAC	19-20	3,5	3+$^1/_2$	IR	-	5'	11
cpϕ Rac	*ydaU*	858	490	C$_{/GA}$A$_{/T}$G$_{/T}$VTGTT$_{/C}$	8	6	1+1+4	rep	1 t	5'	24
S. flexneri	sf0873	633	292	not detected	-	-	-	-	-	-	-
ϕ ε15	-	678	397	GTCCGCN$_5$GTCCGCN$_7$GCGGACN$_5$-GCGGAY	41	2	1+1	IR	-	5'	22
B. bronchis.	bb2207	825	517	CGTNACG|CGTNACG	14	5.5	0.5+1+1+3	IR[4])	-1 / -2 nt	5'	29
X. fastidiosa	xf1560	906	400	AAACCCA	7	5	2+7	rep	[1 t]	5'+3'	-
				A$_{/TG}$AA$_{/G}$C$_{/G}$CC$_{/T}$A$_{/G}$ (1 mismatch)		6	+1+1				
H. somnus	hsom1669	714	334	TCGGAWA$_{/CT}$T|AA$_{/G}$ATCCGA	16	4	4	IR	2 t	5'	20
cpϕ CP-933N	z1773	1038	475	GTCACG$_{/TA}$A$_{/GTC}$	7	11	6+5	rep	1 t	5'	17
cpϕ CP-933U	z3124	1038	466	G$_{/A}$TCACG$_{/T}$A$_{/GTC}$	7	11	6+5	rep	1 t	5'	8
ϕ ST64T	gp18	900	511	TGTCCAA$_{/T}$C$_{/A}$G$_{/A}$	9	6	3+3	rep	2 nt / 1 t	5'	17
ϕ HK022	O	900	502	TACCAGCA	8	3	1+6+3	rep	[3 nt /1 t]	5'	11
				T$_{/A}$A$_{/CG}$C$_{/T}$CAG$_{/CA}$HN	8[3])	7					
ϕ VT2-Sa	O	900	502	TACCAGCA	8[3])	4	1+6+3	rep	[3 nt /1 t]	5'	11
				T$_{/A}$A$_{/G}$C$_{/T}$CAG$_{/CA}$HN	8	6					
ϕ HK620	O	900	505	TACCAGCA	8[3])	4	1+9	rep	[3 nt /1 t]	5'	14
				T$_{/A}$A$_{/G}$C$_{/T}$C$_{/A}$A$_{/T}$G$_{/CA}$C$_{/TA}$D	8	6					

Legend to Tables C3 + C4 . nt = length of gene; AT = AT$_{max}$; nt = length of iteron; number of iterons; groups = clustering of iterons; type = iteron type (rep: repeat, IR: inverted repeat, rIR: relaxed inverted repeat); spacing = spacing between iterons, indicated as distance (nt) between outside bases or helical turns (t) calculated from the iteron middle; pos. = position of iterons/major iteron cluster relative to AT$_{max}$ ('AT-peak'); d = distance (nt) between the proximal iteron and AT$_{max}$. Inverted repeats in the iterons are indicated by boxes. Abbreviations used in the 'consensus' column: R = A or G; Y = C or T; M = A or C; K = G or T; S = C or G; W = A or T; H = A or T or C; D = A or G or T; B = G or C or T; V = A or C or G; N = any. 1) relaxed inverted repeat. 2) repeats embedded in 2 rIR: GA(GT)(T/C)CG**AAACAG**ACCA**AAAC(A/C)G**CCGACTC. 3) 2 perfect repeats: TG**TACCAG-CA**GAT**TACCAGCA**AAT**TACCAC**. 4) nested palindromes: cgtcacg**CGTA**A**Acg**TCACG**C**GTA**Acg**tacg**CGTGACG**. 5) 2 iterons arranged with the inverted iterons in two 19 and 21 bp palindromes, respectively. DNA sequences for Table C3 were taken from: ϕD3 [NC_002484], ϕP22 [NC_002371], ϕ80 [X13065], λ / ϕ21 [J02459], ϕ4795 [NC_004813], ϕH-19B [AF034975], ϕ933W [NC_000924], cpϕ CP-933V [NC_002655], cpϕ CP-933X [NC_002655], pϕ LambdaSo [NC_004347], *P. putida* [NP_746024], ϕNil2 [ECO413274], ϕHK97 [NC_002167], cpϕ Gifsy-2 [NC_003197], cpϕ CP-933P [NC_002655], *P. putida* [NP_743708], ϕAaϕ23 [NC_004827], ϕP27 [NP_543069], ϕSfV [NC_0034441], and ϕST64B [NC_004313]. DNA sequences for Table C4 were taken from: cpϕ CP-933M [NC_002655], ϕBcep22 [AY349011], ϕE125 [NC_003309], *P. luminescens* [BX571870], *B. bronchiseptica* [NC_002927], cpϕ CP-933O [NP_287511], *S. enterica* [NP_456424], cpϕ CP-933R [NC_002655], cpϕ Rac [NC_000913], *S. flexneri* [NC_004337], ϕε15 [NC_004775], *B. bronchiseptica* [NC_002927], *X. fastidiosa* [NC_002488], *H. somnus* [NZ_AABO02000023], cpϕ CP-933N [NC_002655], cpϕ CP-933U [NC_002655], ϕST64T [NC_004348], ϕHK022 [NC_002166], ϕVT2-Sa [NC_000902], and ϕHK620 [NC_002730].

Table C5 (Putative) replication origins of Gram(+)-specific phages; part A.

(pro)phage	gene	nt	AT	consensus sequence of iteron(s)	nt	#	groups	type	spacing	pos.	d
ϕ SPP1	gp38	771	439	A/GG/AAAAC/AGA/GC/AA/T (AB boxes) A/TG A/TA/TAAA/CGAC/TA (ab boxes)	10 10	5 2	2+5	rep	1 t	5'	25
ϕ mv4	orf292	876	565 670	CAGAAGTTGGAACAA -	14	2	2	rIR	2 t	5'	118
ϕ r1t	orf11	735	382	TGGAACC	7	4	4	rep	1 t	5'	27
ϕ 3626	gp34	753	268	-							
			451	CCTTGTAGCAAAAT	14	3	3	IR	2 t	5'	32
ϕ 13	phi13_ 15	861	493	TGGAAAAACCGTTAACGGTAAAACCA TGGAAAAACCGTTAATGGTAAATCGC	26	2	2	rIR	4 nt	5'	29
ϕ SLT	orf256	771	400	TGGAAAAACCGTTAACGGTAAAACCA TGGAAAAACCGTTAATGGTAAATCGC	26	2	2	rIR	4 nt	5'	35
ϕ ETA	orf22	771	409	TGGAAAAACCGTTAACGGTAAAACCA TGGAAAAACCGTTAATGGTAAATCGC	26	2	2	rIR	4 nt	5'	35
ϕ LL-H	orf299	900	469	G/AG/AAA/TVBC/A	8	6	6	rIR	1.5 t	5'	44
ϕ 315.1	spyM3_0691	762	448	ATGGAACTGTAAAAAATACAGTATCGGAACTGGAAATTTTACAG	44	2	2	rIR	5 t	5'	27
S. pyogenes	sps1159	836	520	ATGGAACTGTAAAAAATACAGTATCGGAACTGGAAATTTTACAG	44	2	2	rIR	5 t	5'	23
S. pyogenes	spyM18_1803	939	487	AGAGTTAGAATAAAGAGA	17	4	4	rIR	1 nt	5'+3'	–
L. gasseri	lgas 0588	924	459	AAGATAAGTTAAG AWGATAA	13 7	2 3	1+4	rIR	2 nt		
pϕ LambdaBa04	ba3819	759	448 520	CAGAGAYTA -	9	3	3	rIR	1 t	5'	8
ϕ TP901-1	REP	819	382	TTTTCCAGATGTGG YCWGAT	14 6	2 2	4	rep	1 nt 9 nt	5'	45
ϕ PV83	orf20	804	469	2x GTCACGTGAC; 1x GTCACGCAAC	10	3	3	rIR	1 t	5'	40
pϕ LambdaBa02	ba4121	717	382	CCACAAAATGACCCAGTGG CCACCAGTGGAAAACCACCGG	19/22	2	2	rIR	8 nt	5'	44
ϕ BC6A52	bc2563	747	400	RTGGARRWWYCAB ATGRARWWWCMC	13 12	4 2	6	rIR	2 nt	5'	33
ϕ PVL	orf46	894	442	TCAA/GHMC/AA/GAC	10	8	1+7	rep	1 t	5'	18
ϕ N315	sa1791	894	451	TCAA/GHMC/AA/GAC	10	8	1+7	rep	1 t	5'	27
ϕ A118	gp49	933	448	AC/TAA/TCG/ADHHD	10	10	1+2+6+1	rep	1 t	5'+3' [1])	36

gin *oriJ* of the Rac prophage of *E. coli* K12 by its property to drive replication of a selectable plasmid [105]. The location of *oriJ* within in the ~13 kb insert of the original plasmid pLG-2 could be narrowed down to ~2.5 kb, encompassing the *ydaSTUV* genes (H.Seitz and C.Weigel, unpublished). In analogy to the arrangement of the λ replication genes and origin, Wegrzyn and co-workers proposed for *oriJ* four inverted repeats (consensus sequence GCATCTN$_{11\text{-}13}$AGATG-C) upstream of an AT-rich region within the *ydaU* gene as putative binding sites of the YdaU initiator [354]. These 4 and 2 additional more degenerate inverted repeats are rather widely spaced, and located preferentially in regions enriched for GC within the *ydaU* 5'-half. We could detect, in addition, six non-palindromic iterons (Table C4). When compared with the arrangement of the iterons in other phage replication origins (see below), the size of these latter iterons, their number, their regular spacing, and their position closely upstream of the 'AT-peak' in the *ydaU* gene correspond better to a generalised design of a phage replication origin for θDR than the 4 inverted repeats proposed by Wegrzyn and co-workers.

Replication origins of lambdoid (pro)phages. Experimental data on the replication origin structure of lambdoid phages other than those discussed above are scarce. By sequence comparison and genetic analysis of phage hybrids, the replication origin of ϕP22 was map-

ped within gene 18 encoding the initiator (analogous to λ *O*) [346,15].

The conserved arrangement of structural elements in the replication origins of λ and ϕSPP1 led us to assume that the replication origins should be easily detectable also for other lambdoid (pro)phages: i. by the presence of an AT-rich region located approximately in the middle of the (putative) initiator gene, and ii. by the presence of iterons 5'-adjacent to the AT-rich region. For the analysis, we chose the 40 (putative) initiator genes of Gram(−)-specific phages and the 40 (putative) initiator genes of Gram(+)-specific phages discussed in Section C3.1.2. .

A+T plots revealed a prominent 'AT-peak' in ~90% of the analysed genes. The 'AT-peak' located approximately to the middle of most genes (Fig. C5; see also Tables C3 – C6). The AT-rich region 3'-adjacent to the iterons in the λ *O* gene and ϕSPP1 gene 38 appeared as 'AT-peaks' in the A+T plots, suggesting that this peak is indicative of an origin-specific AT-rich region also in the other phage initiator genes. We found one single prominent 'AT-peak' in 73% of the (putative) initiator genes, and two centrally located 'AT-peaks' in 16%. In the remaining sequences (11%), either multiple peaks were found, or no significant AT-clustering. Interestingly, no 'AT-peak' could be detected in the λ *cI* gene encoding the λ repressor; λ CI protein is comparable in size and architecture to λ O (N-terminal DNA- binding domain, C-terminal oligomerization domain) [58].

All but 3 of the 80 initiator gene sequences contain easily detectable repeats, in >90% of the cases clustered 5'-upstream of an 'AT-peak'. Although only assigned theoretically, the presence of such short repeated sequences in virtually all (putative) initiator genes suggests that these repeats represent iterons, i.e. binding sites for the phage replication initiator, and are thus part of the replication origins of these phages (Tables C3 – C6). However, with respect to type (simple or inverted repeat), number, spacing, distance to the 'AT-peak', and orientation to each other these (putative) iterons present a colourful bouquet. Only in those cases where the initiator proteins show a high degree of similarity in their N-termini (>60% ident. res.; Section C3.1.2.) the iterons found in the respective genes are also very similar or identical. With one exception discussed below, we did not detect any features of the iterons that could discriminate replication origins of phages of Gram(+) hosts from those of Gram(−) hosts.

We found iterons of the 'simple repeat'-type (rep) as in *ori*L of ϕSPP1 in roughly one half of the initiator genes. Iterons of the 'relaxed inverted repeat'-type (rIR) as in oriλ or perfect palindromic iterons (IR) were found in the remaining genes. In some cases a decision whether the iterons were of the 'rep'- or the 'rIR'-type was impossible because simple repeats were found embedded in larger (relaxed) inverted repeats. E.g. in the prophage gene pp1551 of *P. putida* KT2440 two AAACMG repeats are embedded twice in a larger (relaxed) inverted repeat gagtcgAAACAGaccaAAACCGccgactc, which is found twice 5'-upstream of a prominent 'AT-peak'. There is thus no clear preference for a particular iteron type in the (putative) phage replication origins. The length of 'rep'-type iterons varies between 6 and 34 nt, with a mean of ~10 nt. The length of 'rIR/IR'-type iterons varies between 8 and 41 nt, with a mean of ~20 nt. In most cases, 'rIR'-type iterons with a length of ~20 nt are characterised by short inverted repeats flanking a less conserved central part. These motifs are reminiscent of the DNA-binding sites of transcription factors of the λ CI-type, which bind to DNA as dimers with each monomer binding to one half-side of the binding-site, prior to forming higher-order nucleoprotein complexes by protein-protein oligomerisation [58].

The number of iterons in the individual initiator genes varies between 2 and 11, with a mean of ~5. The *Streptococcus agalactiae* prophage gene sag0559 seems unique in that it contains one single 31 nt long iteron of the 'rIR'-type, which is located 36 nt upstream of a prominent 'AT-peak'. As a rule of thumb, we found higher numbers of iterons for short iteron sequences and lower numbers for longer ones. Invariably, however, we found a clustering of iterons at a distance of ~20 nt from the calculated maximum (ATmax) of the 'AT-peak' (measured from the 3'-end of the proximal iteron) for initiator genes of phages and prophages of Gram(−) hosts. For the genes of phages from Gram(+) hosts, the mean distance is more close to ~30 nt; the significance of this difference with respect to phage host range is unclear. Note that the values for ATmax in AT-plots is a calculated value and does not indicate the position where DNA unwinding could possibly occur. Therefore the distance between ATmax of the 'AT-peak' and the proximal iteron given in Tables C3 – C6 cannot be more but a crude estimate of the spacing between the iteron region and the AT-rich region of a putative phage replication origin. Correspondingly, however, the distances between the 3'-end of the iteron proximal to the AT-rich region and the position where unwinding commences are generally in the order of 10-50 nt in chromosomal and plasmid replication origins [98,290].

There also exists a great variation with respect to iteron spacing. In close to one half of the initiator genes we found a spacing of exactly one helical turn between individual iterons (centre-to-centre) when they are part of the iteron cluster 5'-adjacent to the 'AT-peak'. Such a regular spacing of 1 helical turn is particularly observed for short iteron sequences. In the other half of the (putative) initiator genes we found an iteron spacing of exactly 2, 3, and 4 helical turns, less frequently a regular spacing deviating from complete helical turns, and only in a

Table C6 (Putative) replication origins of Gram(+)-specific phages; part B.

(pro)phage	gene	nt	AT	consensus sequence of iteron(s)	nt	#	groups	type	spacing	pos.	d
L. innocua	lin2412	933	445	AC$_{/T}$AA$_{/T}$CG$_{/A}$DHHD	10	11,5	1+1+2 +6.5+1	rep	1 t	5'+3' [1])	33
L. innocua	lin0086	912	445	THABAA$_{/C}$A$_{/C}$BDD	10	8	2+6	rep	1 t	5' [1])	31
			634	THABAA$_{/C}$A$_{/C}$BDD	10	2	1+1	rep	–	3'	9
L. monocyt.	lmo2317	975	493	A$_{/T}$G$_{/A}$TDA$_{/G}$HG	7	7	7	rep	≤1 t	5'	23
ϕ SM1	gp19	798	475	GTAAC|GTTAC GTAAC	10 5	2 2	2+0,5+0,5	IR	1 t	5'	39
B. cereus	bc0955	864	343 514	not detected	–	–	–	–	–	–	–
B. anthracis	ba0933	861	343 523	not detected	–	–	–	–	–	–	–
B. anthracis	ba3331	741	339 520	GAAGAATGGAAG|ATTCCATTGTTG GCATTTTTGG|TA|TGCCTAAAGAA	24/22	2	2	rIR	–	3' 5'	–
pϕ Lp1	orf20	933	459	ATTAACGAG|CCAGCGTAT ATTAATGCT|GACATTAAT GTTAATGTC|GACATTAAT	18	3	1+2	rIR	3 t	5'	35
pϕ Lp2	orf20	978	337 487	AGTAGGYCARCTGGTT -	16	2	2	rep	32 nt	3' 5'	30 65
L. innocua	lin1244	918	411	CATTAATGTTGTAAATGATGACATTAACG AGTTAATGCC|AGCATTAAGT	16/25	2	2	IR	16 nt	3'	16
S. agalactiae	sag0559	861	442	TGGTGTACAAATGTTGTACAAATGTACACCA	31	1	1	rIR	–	5'	36
C. acetobu.	cac1934	840	-	GTAACGCTAA, GTAAC|GTTAC, GTAACGGAAT, GTAAC|GCAAC	10	4	4	rIR	≤1 t	–	–
ϕ P335	orf11	813	556	ATTGGACAATGTCCAAT GTCCAA	15 6	1 2	1+ 2x 0.5	IR rep	1 t	5'	137
ϕ bIL309	orf14	777	520	ATTGGACAATGTCCAAT GTCCAA	15 6	1 2	1+ 2x0.5	IR rep	1 t	5'	101
ϕ Tuc2009	orf16	783	259 535	CAACCA	6	4	4	rep	1 t	3' 5'	82 154
ϕ bIL285	orf16	783	274 538	CAACCA	6	4	4	rep	1 t	3' 5'	67 157
ϕ bIL286	bIL286p16	840	505	ATTAAN$_{5-8}$TTAAT	15/18	2	1+1	IR	–	5'	33
ϕ 11	phi11_15	807	478 538	GTAAC|GTTAC, GTTAC, GTAAC -	10/5	2/2	0.5+2.5	IR	1 t	5'	53
ϕ 7201	orf4	807	430	ACGGAACTAAAGAAAGTTCAGTTCAAGAACTAAA	34	2	2	rep [2])	4 t	5'	34
ϕ BK5-T	orf49	810	385	TA$_{T}$TACAV$_2$|V$_2$TGTAAA	16	4	4	rIR	6 nt / 2 t	5'	24

Legend to Tables C5 + C6 . nt = length of gene; AT = AT$_{max}$; nt = length of iteron; number of iterons; groups = clustering of iterons; type = iteron type (rep: repeat, IR: inverted repeat, rIR: relaxed inverted repeat); spacing = spacing between iterons, indicated as distance (nt) between outside bases or helical turns (t) calculated from the iteron middle; pos. = position of iterons/major iteron cluster relative to AT$_{max}$ ('AT-peak'); d = distance (nt) between the proximal iteron and AT$_{max}$. Inverted repeats in the iterons are indicated by boxes. Abbreviations used in the 'consensus' column: R = A or G; Y = C or T; M = A or C; K = G or T; S = C or G; W = A or T; H = A or T or C; D = A or G or T; B = G or C or T; V = A or C or G; N = any. 1) major cluster of iterons 5' to AT$_{max}$. 2) 2 repeats, each containing 2 16 nt long palindromes with the order > < < >. DNA sequences for Table C5 were taken from: ϕSPP1 [NP_690731], ϕmv4 [AAG31328], ϕr1t [NP_695039], ϕ3626 [NP_612863], ϕ13 [NP_803370], ϕSLT [NP_075484], ϕETA [BAA97608], ϕLL-H [AAL77546], ϕ315.1 [NP_664495], *S. pyogenes* [NP_802421], *S. pyogenes* [NP_607824], *L. gasseri* [ZP_0004-6421], pϕ LambdaBa04 [AAP27557], ϕTP901-1 [NP_112676], ϕPV83 [NP_061610], pϕ LambdaBa02 NP_846360], ϕBC6A52 [NP_832321], ϕPVL NP_058485], ϕN315 [NP_835538], and ϕA118 [NP_463514]. DNA sequences for Table C6 were taken from: *L. innocua* [NP_471742], *L. innocua* [NP_469432], *L. monocytogenes* [NP_465841], ϕSM1 [NP_862858], *B. cereus* [NP_830741], *B. anthracis* [NP_843439], *B. anthracis* [NP_845619], pϕ Lp1 [NP_784408], pϕ Lp2 [NP_785894], *L. innocua* [NP_470581], *S. agalactiae* [NP_687588], *C. acetobutylicum* [NP_348556], ϕP335 [NP_839902], ϕbIL309 [NP_076709], ϕTuc2009 [NP_108693], ϕbIL285 [NP_076588], ϕbIL286 [NP_076650], ϕ11 [NP_803268] ϕ7201 [NP_038304], and ϕBK5-T [NP_116541].

few cases a completely irregular spacing. Despite many attempts to resolve the stoichiometry and the three-dimensional structure of 'open complexes' [228] formed by initiator proteins with their cognate replication origins, their exact architecture(s) still remain enigmatic. It is even not known whether there exists a generalised architecture for 'open complexes' but it is assumed that a regular spacing of iterons within origins reflects a highly ordered structure of the 'open complex'. Based on electron microscopic studies, the Kornberg lab proposed a model for the 'open complex' formed by *E. coli* DnaA and *oriC* in which ~200 bp *oriC* DNA are wrapped around an ellipsoid DnaA core [86]. The great variation in iteron length and spacing found here in the set of 80 initiator genes leads us to speculate that it is more likely that 'open complexes' formed by phage initiators at their cognate origins have a three-dimensional structure resembling the RecA-filament [112], as was recently proposed also for the *E. coli* DnaA·*oriC* complex by Messer [290].

Alonso and co-workers detected a threefold repeated sequence AACAATGA within the AT-rich region of the ϕSPP1 *ori*L for which no function is known [12]. During our search for iterons in the set of 80 initiator genes we found that they contain – in addition to iterons – gene-specific repeats at a positions corresponding to those in ϕSPP1 *ori*L. In 69 of the 80 genes we found several small repeats with the sequence AAGAA, AAGA, or AGAA in close neighbourhood of the 'AT-peak', in many cases in addition to the gene-specific repeats mentioned above. These small AAGAA, AAGA, and AGAA repeats were also found on the complementary strand, albeit less frequently. Because in 22 of the 69 genes containing such small repeats these are found as part of larger, gene-specific repeats their presence might be meaningful in all 69 genes despite their short length. Invariably, the gene-specific repeats and the small repeats – we tentatively like to coin both 'secondary repeats' – are lo-cated closer to AT_{max} of the 'AT-peak' than the iterons, and mostly on both sides of AT_{max}. A straightforward (trivial) explanation for the existence of these AT-rich 'secondary repeats' in all but one of the 80 phage and prophage initiator genes is that they have evolved by micro-duplications and conserved simply because they contribute positively to origin unwinding. However, the 13mer repeats present within the AT-rich region of the chromosomal replication origin *oriC* of *E. coli* and related species overlap with recently detected ssDNA-binding sites for the initiator protein DnaA [423], which was long before only known to bind to the dsDNA *oriC* iterons, the DnaA boxes [290]. Therefore, the possibility cannot be excluded that the 'secondary repeats' within the phage replication origins represent preferential entry sites – on dsDNA, or ssDNA in the unwound origin – for one ore more of the proteins involved in the initiation of replication, i.e. initiator, SSB, helicase loader(s), helicase, or primase. Clearly, the preferential though not exclusive arrangement of these 'secondary repeats' on one strand adds asymmetry to the putative replication origins, which might be instrumental for the choice between unidirectional and/or bidirectional replication (a detailed listing of all 'secondary repeats' is available from the authors upon request).

Our search for structural elements of a replication origin in the set of 80 (putative) phage initiator genes revealed that the origin structures determined experimentally for the phages λ and ϕSPP1 are well conserved in the majority of the genes analysed. Therefore, λ and ϕSPP1 can be safely considered valid model systems for phage replicons that replicate by θDR.

Although the localisation of the replication origin within the initiator gene seems to be a common feature among phage replicons that replicate by θDR, also some plasmid replicons with the same characteristic have been described. Cohen and co-workers identified the replication origin of the 17 kb linear plasmid pSLA2 of *Streptomyces rochei* within its *rep1* initiator gene [71]. With 864 bp, the *rep1* gene has the typical length of a phage initiator gene, it contains a centrally located 'AT-peak', three (+ one degenerate) 21 bp long iterons of the 'rep'-type with regular spacing (2 helical turns) 5'-upstream of the 'AT-peak', and secondary repeats on either side of the 'AT-peak' (data not shown). Despite weak sequence similarity, the replication origin of the 12 kb *Streptomyces clavuligerus* plasmid pSCL shares the arrangement of structural elements with pSLA2. Interestingly, both replicons can drive bidirectional θDR of linear and circular forms of their plasmids [411,70], and encode in addition a $DnaB_{Eco}$-type helicase [NP_0442-86]. Although the *rep1* initiator protein of pSLA2 lacks similarity with any known or putative phage initiator protein, the weak similarity (~30% ident. res.) of the (putative) initiator of pSCL with the N-termini of L17 of ϕP27 and orf39 of ϕV (32% ident. res.) suggests that these plasmid replicons and phage replicons have a common evolutionary origin. The identification of the *rep2* gene located downstream of *rep1* in pSLA2 as $DnaB_{Eco}$-type helicase supports this hypothesis (BRM Section 3.) [71].

From our analysis one can derive a 'set of rules' which could – in the case of initiator genes – help to assign prophage gene functions in bacterial genomes with higher reliability than can be achieved by mere protein comparison. A candidate gene of ~900 bp / 300 res. in length (Fig. C5) with N- or C-terminal similarity to at least one of the 80 (putative) initiator protein sequences (Section C3.1.2.) should be considered as phage replication initiator containing the replication origin of this phage if: i. a prominent 'AT-peak' is located in its middle, ii. it contains several repeated sequences 5'-adjacent

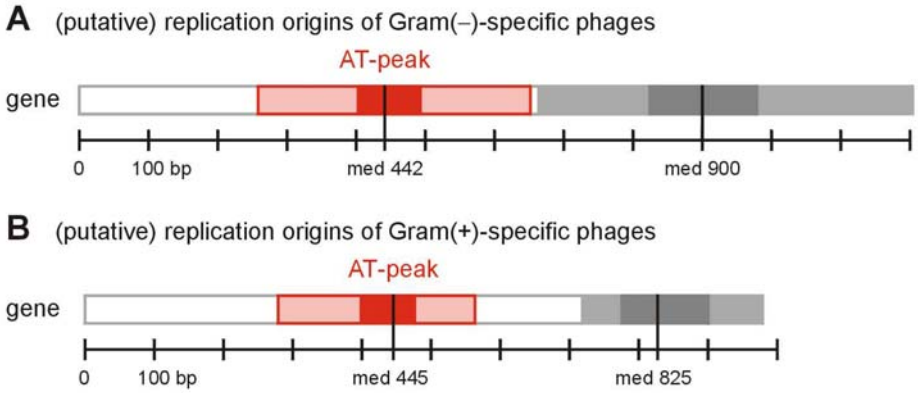

Fig. C5 (Putative) replication origins of lambdoid phages.
A Gram(–)-specific phages: summary of the 40 sequences from Tables C3 + C4 . **B** Gram(+)-specific phages: summary of the 40 sequences from Tables C5 + C6 . The length variation of the genes is indicated: light grey box: upper and lower quartil, respectively; dark grey box: 50% region; the median (med) is given as discrete value. AT-peaks were determined using the spin v1.1 software from the Staden Package [425], using a window of 41 nt. The variation of the position of the AT-peak in the genes is indicated: light red box: upper and lower quartil, respectively; red box: 50% region; the median (med) is given as discrete value.

to the 'AT-peak', and iii. it contains 'secondary repeats' to both sides of AT_{max}. These rules might even be translatable into robust algorithms for an automated search for prophage replication genes and origins in genomic sequences. But we expect from the great variability already apparent in our limited set of sequences that a closer inspection of candidate sequences by eye would still be necessary. Computer-aided searches for (pro)phage replication origins could profit from GC-skew analyses, in addition [152,375].

As examples, we discuss the *Shigella flexneri* prophage gene sf0873 and the *Bordetella bronchiseptica* prophage gene bb1683. The sf0873 protein sequence is highly similar to the *E. coli* cpϕ CP-933R putative initiator protein encoded by gene z2397 (93% ident. res.) and to the YdaU initiator protein of the *E. coli* Rac prophage (65% ident. res. in the C-terminus). With only 633 bp, the sf0873 gene is considerably shorter than most other origin-containing phage initiator genes. It contains 2 prominent 'AT-peaks', one located approximately in its middle. However, no iterons or 'secondary repeats' could be detected in the sequence. In addition, the gene is not found in a context of other replication genes (BRM Section 3.1.). Therefore, the indications for a replication origin within this gene are weak. In contrast, the 957 bp long gene bb1683 of *B. bronchiseptica* contains a very prominent 'AT-peak', preceded by a cluster of 7 iterons, and 'secondary repeats' to both sides of AT_{max}. Although the translated sequence has only moderate similarity to the ϕP27 initiator L17 (37% ident. res. in the N-terminal 75 res.), this gene should be considered a phage replication initiator gene containing the replication origin.

Other bacteriophage origins for θDR. Very recently, the ϕPY54 replication origin for θDR could be located in a segment of 212 bp from the 3'-part of the *repA* gene encoding the multifunctional helicase-primase protein [507]. A sub-segment of 54 bp contains four *dam* methylation sites – 3 as parts of short 6 bp repeats – and an AT-rich region with additional (different) repeats. These structural elements are conserved with respect to sequence and location in the related phages ϕN15 and ϕKO2 [365]. The coding capacity of the origin region is not important for RepA function, and the origin is functional if RepA is provided *in trans* [507].

The replication origin *oriI* of *E. coli* phage P4 resembles plasmid origins more than any known phage origin with respect to complexity. ϕP4 *oriI* has been defined as a bi-partite structure, consisting of *ori1* approximately 4.5 kb upstream of the α gene and the *crr* region downstream of α ([123]; reviewed in [47]). Replication is initiated at *ori1*, but *ori1* and *crr* are both required for origin (*oriI*) function. The relative orientation of *ori1* to *crr* is important but the distance between both sites is not. *Ori1* contains 6 'type 1' iterons GGTGAACAGA/T to which α protein binds for initiation; the 'type 1' iterons are arranged in irregularly alternating orientations but with helical phasing to both sides of an AT-rich region in a DNA segment of 123 bp in length. *Crr* con-

Table C7 (Putative) replication origins for θDR of other (non-lambdoid) phages.

phage	gene	location		AT	consensus sequence of iteron(s)	nt	#	groups	type	pos.	d
ϕ adh	orf771	3'	ig	+	AGTGTAG$_{/T}$G$_{/A}$GTT	11	4	2+2	rep	5'	45
ϕ A2	orf35	3'	ig	+	CGGGAG$_{/T}$G$_{/A}$AT$_{/A}$	9	4	3+1	rep	5'	18
ϕ PSA / ϕ 2389	pri	3'	ig	+	HRRATAGTT$_{/AG}$	9	4	1+1+2	rep	5'	0
ϕ DT1	orf36	3'	ig	+	GTTA$_{/GT}$C$_{/TG}$	5	10	1+3+3+3	rep	5'+3'	22
ϕ O1205	orf13	3'	ig	+	RGTTA$_{/TG}$C	6	16	4+4+4+4	rep	5'+3'	23
ϕ Sfi11	orf504	3'	ig	+	RGTT$_{/A}$A$_{/TG}$C$_{/TG}$	6	11	4+4+3	rep	5'+3'	24
ϕ 31	primase	3'	ig	+	T$_{/C}$GTTCCA	7	7	3+2+2	rep	5'+3'	9
ϕ bIL310	orf24	3'	ig	+	GT$_{/G}$T$_{/A}$A$_{/G}$C$_{/TA}$	5	9	3+4+2	rep	5'	12
ϕ 105	orf11	3'	ig	+	GTT$_{/C}$A$_{/CG}$C$_{/GTA}$	5	9	8+1	rep	5'+3'	30
ϕ P1	*repL*	int.	-	+	GTCAR$_3$TGAC$_{/T}$	11	2.5	2+0.5	rIR	5'	3

gene = gene upstream of (putative) origin; location = '3' indicates location 3'-downstream of 'gene', 'int.' indicates location within coding region of 'gene'; 'ig' indicates location in intergenic region. In the 'groups' column, underlining indicates the iteron group closest to AT_{max}. Other abbreviations and criteria as in legend to Table C6 . Sequences were taken from: ϕadh [NC_000896], ϕA2 [NC_004112], ϕPSA / ϕ2389 [NC_003291], ϕDT1 [NC_002072], ϕO1205 [NC_004303], ϕSfi11 [NC_002214], ϕ31 [LBA292531], ϕbIL310 [NC_002669], ϕ105 [NC_004167], and ϕP1 [NC_005856].

tains two direct repeats each 120 bp long. A number of (somewhat more degenerate) 'type 1' iterons are found in the *crr* repeats, less regularly spaced as in *ori1*, but also with alternating orientations. Origin function is diminished but not completely suspended upon deletion of one of the *crr* repeats. ϕP4 contains a second, cryptic replication origin, *ori2*, located in the 5'-half of the α gene [451]. The 36 bp long *ori2* sequence together with *crr* has been shown to confer replication proficiency to a test plasmid. *Ori2* lacks type 1 iterons and α protein did not bind to it *in vitro* [451].

In Section C3.1.2., we will discuss the 'initiator domain' of *Lactobacillus gasseri* phage adh and other phage-encoded helicase-primase enzymes that are related to the ϕP4 α protein. The replication origin of ϕadh was mapped to the intergenic region downstream of orf771 (encoding the ϕP4α-type primase-helicase) [4]. Applying the criteria developed for the lambdoid phages (see above), we could readily detect (putative) origin structures downstream of the ϕP4α-type helicase gene in all phage genomes of this group (Table C7) [124,426, 303,268] The detected iterons are invariably of the 'direct repeat' type and in several cases found on both sides of the AT-rich region, in this respect resembling the ϕP4 *ori1* configuration.

The phage P1 origin for theta(θ)-mode replication in the lytic cycle is located within the *repL* initiator gene as in the lambdoid phages (Table C7) [82].

C2.3. Other bacteriophage replication origins

The genome of the *Lactococcus lactis* phage c2 (22.2 kb) is organised in two divergently transcribed blocks of – early and late – genes separated by an intergenic region. This intergenic region, designated *ori*, was shown to support the replication of a selectable plasmid in *L. lactis* in the absence of phage-encoded proteins [469]. By analysing ϕc2 replication intermediates by two-dimensional gel electrophoresis, it could be shown that replication is initiated in the *ori* region *in vivo* and occurs in the θ-mode [60]. Deletion of the early promoter P_E1 in the *ori* region abolished origin function. Introduction of mutations into P_E1 or replacement of P_E1 with an unrelated but functional promoter did not abolish replication. Replacement of the P_E1 transcript template sequence by an unrelated sequence with a similar G+C content abolished replication, showing that the sequence encoding the transcript is essential for origin function. [394]. Phage c2 may thus represent the first example for a phage replicon that – like the ColE1-type plasmids – entirely depends on host functions for replication. Although it is presently not known to which extent possible secondary structures of the P_E1 transcript are important – and whether they are processed in the lactococcal host – we may assume that ϕc2 replicates by the tDR mechanism. Unlike ColE1 however, replication of ϕc2 occurs in a *polA* host and therefore does not require host PolA for transcript elongation [359]. We note that although the ϕc2 genome lacks replication genes, several (putative) recombination genes

could be readily identified (Section C3.6.2.). Also lactococcal phages of the ϕbIL67 and ϕ923-type replicate by tDR like ϕc2. All three phage groups contain a non-homologous P_E1 transcript template embedded in highly homologous PE1 and PE2 regions [359]. Also the origin regions of ϕbIL67 and ϕ923 can sustain plasmid replication in the absence of phage proteins but chimeric origins were inactive. The latter observation suggests that the particular secondary structure of each P_E1 transcript is important.

Hillier and co-workers have identified the replication origin of *L. lactis* phage sk1 by its property to drive replication of a selectable plasmid; the origin is located in the intergenic region between orfs 46 and 47 and includes the first 179 residues of orf 47 [68]. Although the intergenic region contains an AT-rich stretch and also short direct repeats, a closer analysis of this origin region is warranted: at present, neither the replication mechanism nor a putative initiator protein are known. ϕsk1 orf43 encodes a putative replication protein, but it may rather be a variant of the plasmid-encoded RepA-type or ϕPG1 gp59 F4-type helicases than a polymerase subunit as proposed by Schouler *et al.* [398].

A similar unsatisfactory situation exists for the replication origin of the *Leptospira biflexa* (Spirochaetes) phage LE1. Girons and co-workers could show that the ϕLE1 prophage exists as plasmid in its host, and the replication origin could be located on a ~2 kb fragment encompassing orfs 3–5. An AT-rich region and a number of direct repeats could be identified, but the replication mechanism and a putative initiator are not known [144].

C3. Bacteriophage replication proteins

Having set the stage with a discussion of the different molecular mechanisms driving phage replicon propagation (BRM Section 2.) and origin structures in the preceding section (Section C2.) we will proceed here with the discussion of the bewildering variety of phage-encoded replication proteins. A discussion of the primosomal proteins – initiators, helicase loaders, helicases, and primases – will be followed by a discussion of replisomal proteins – DNA polymerases and their accessory proteins. We will complete this section with a survey of phage-encoded single-strand DNA binding proteins (SSBs) and recombination proteins. Although not usually considered in the context of replication, there is compelling evidence that recombination is closely interlocked with the replication of different types of phages replicons (see BRM Section 2.).

As outlined in the introduction, we will first discuss within each chapter of this section the genetic and biochemical characterisation of well-known proteins – with special emphasis on experimentally determined interactions among them, and the domains responsible for such interaction(s). This part will be accompanied by a discussion of homologous proteins or proteins with significant similarity that were detected by BLAST searches. Unless indicated otherwise, we performed BLAST searches with default settings: blastp; expect = 10.0; word size = 3; BLOSUM62 matrix; existence = 11; extension = 1; low complexity filter = off [5]. Bl2seq default settings: blastn; reward = 1; penalty = –2; gap x_dropoff = 50; expect = 10.0; word size = 11; [445]. All searches were performed using the NCBI BLAST- server(s) (http://www.ncbi.nlm.nih.gov/BLAST/). To increase the stringency in the discussion of 'protein similarity', we will exclusively refer to '% identical residues (ident. res.)'; for full-length comparisons, we define 'orthologous' or 'identical' for proteins which share >80% ident. res., 'significant similarity' for proteins with 40–80% ident. res., and 'similar' for proteins with >20% ident. res. . In a few cases, it was necessary – and mentioned explicitly in each case – to confine the 'region of similarity' to parts of the proteins, but usually not less than ~ 50. res. . More refined 'data mining' procedures can certainly reveal relationships among proteins that cannot be pinpointed by the 'crude' BLAST approach used here. However, such procedures require thorough discussions and are therefore not appropriate as tools for a review article. In several cases, BLAST results required further *in silico* analyses to gain reliability. In these cases, secondary structure predictions were performed using either the Jpred method or the PHD method, which gave comparable results; both methods claim to reach performances better than 70% [374,88]. The determination of protein tertiary structures is crucial for an understanding of the mechanisms driving replication on the molecular level. However, discussing replication *in toto* and replication modules in particular does not allow us to discuss each single step, and each single protein in minute details. Therefore, we will mention tertiary structures – where known – and refer the reader to the appropriate literature.

C3.1. Initiator proteins

Initiation of DNA replication in the 'rolling circle'-mode (RCR) and initiation in the theta-mode (θDR) depend on the positive action of a protein that is usually called 'initiator' – following the nomenclature proposed by Jacob, Brenner and Cuzin in the 'replicon model' [195]. The molecular mechanisms for both initiation reactions are fundamentally different (BRM Sections 2.1. + 2.2.). Reflecting these differences, initiators for RCR and θDR are functionally and structurally unrelated: initiators for RCR can be classified as site-specific topoisomerases, initiators for θDR are related to transcrip-

tion factors. Both types of initiators have one highly specific property in common: they direct primosomal proteins – helicases and primases – to the replication origin, i.e. they function as 'replisome organiser'.

C3.1.1. Initiator proteins for 'rolling circle' DNA replication (RCR)

Proteins that initiate replication in the 'rolling circle'-mode perform four functions: i. they bind to the replication origin of their cognate replicon and promote duplex unwinding in a small region (nick-site) adjacent to the initiator-binding site(s), ii. they catalyse the breakage of the phosphodiester backbone of one strand in the single-stranded region at a specific site and transfer the 5'-phosphate end to a tyrosine residue within their polypeptide chain – creating a covalent protein-DNA bond – and leave the free 3'-OH end exposed as primer for DNA synthesis, iii. they direct a helicase to the unwound region, thus functioning as 'replisome organiser', and iv. they participate in the termination reaction by catalysing the transfer of the covalently bound 5'-phosphate end of the displaced strand to its 3'-OH end, liberating a covalently closed circular ssDNA molecule. In the known replication systems, these initiators are not involved in the conversion of single-stranded phage genomes into the double- stranded replicative form (RF), or do not participate in the synthesis of the lagging-strand.

Genes encoding RCR-initiators are found in prokaryotic plasmids and phages, but also in archaeal plasmids [276] and plant gemini viruses [429,226]. The RCR-initiators share sequence motifs containing 'active tyrosine' residue(s) involved in the (transient) covalent linkage to DNA with topoisomerases and relaxases (Tra, Mob), proteins that initiate the DNA transfer of conjugative plasmids [334,470]. (Super)families and motifs were defined for the different types of Rep and Mob proteins on the basis of a in-depth sequence comparison [188,227]. From these motifs, we will only consider 'motif 3' containing the catalytically active tyrosine residue(s) in the following; the functional importance of the other motifs remains to be established. Specific features of plasmid-encoded RCR-initiators have been discussed in recent reviews [119,208,98,209]. The molecular mechanisms driving the replication of ssDNA phages have been concisely summarised by Kornberg and Baker [228], and later by Horiuchi [181].

Initiator proteins of phages with ssDNA genomes. Upon entry into the bacterial cell, the closed-circular single-stranded phage genome (+-strand or viral strand) is converted into double-stranded DNA, the replicative form (RF). Subsequently, the RF serves as template for the synthesis of progeny viral genomes by the 'rolling circle' mechanism [141]. The first stage is entirely dependent on host proteins. The second stage involves a phage-encoded function – the initiator for RCR – and a different subset of host proteins. The development of *in vitro* replication systems in the labs of Jansz, Hurwitz, Denhardt, and Kornberg was instrumental in dissecting the timely order of the individual enzyme-catalysed steps and in identifying the proteins involved [477,478,390,118,117,401,116,110]. These studies culminated in the definition of 'primosomal' and 'replisomal' proteins in the terms of enzymology – making replication of the *E. coli* phage ϕX174 the best-understood system in the 1970s. Also during these years, the determination of the nucleotide sequence of the ϕX174 genome by Sanger and co-workers in 1977 marks the onset of the era of genome sequencing [386].

Gene II protein (gpII) of phage fd. Meyer and Geider characterised ϕfd gpII biochemically as protein with endonuclease and topoisomerase activity. The endonuclease activity was found to be specific for ϕfd RF DNA (supercoiled dsDNA) because other supercoiled substrates were not cleaved. Also, neither the viral DNA (circular ssDNA) nor relaxed (doubly-closed circular) RF DNA were appropriate substrates. With a maximum of activity at pH 8.5 and 80 mM KCl, gpII introduced a single nick in supercoiled ϕfd RF at low Mg^{2+} concentrations at a specific site (Section C2.1.) [296]. At higher concentrations of Mg^{2+}, the topoisomerase activity of gpII became apparent: roughly one half of the supercoiled RF substrate was converted to the relaxed (doubly-closed circular) form [295].

Geider, Meyer and co-workers established an *in vitro* replication system that depended on gpII, and the *E. coli* proteins Rep, SSB, and DNA Pol III holoenzyme for the conversion of ϕfd RF DNA into viral (+)-strand) DNA. [138]. In another, partially heterologous *in vitro* system, viral DNA could be generated from RF DNA by the action of gpII, ϕT7 gene 4 helicase, and ϕT7 gene 5 DNA polymerase. The generation of closed circular ssDNA by both systems added support to the hypothesis that the cleaving/joining activities of gpII are required at successive steps during ϕfd replication. Although the *in vitro* system using ϕT7 components also produced closed circular ϕfd ssDNA, the yield was quenched by the strand-switching activity of the ϕT7 replisome – detected as double-stranded tails of the 'rolling circle' structures [164]. This observation may lead to an answer to the puzzling question why – in all known cases – RCR depends on the *E. coli* Rep helicase *in vivo* and *in vitro*, rather then on the replicative helicase DnaB. DnaB is involved in the coupling of leading- and lagging-strand synthesis as part of the *E. coli* replisome, and the same

Table C8 RCR initiators, part A.

φ / pφ / plasmid	gene	res.	ident.	motif 3
1A. CTXφ-group			cons.	**S**RI Y **WRI** Y **NK**AL**E**Q A KAQL
CTXφ	RstA	276	x	SRI Y WRI Y NKAAQL
φ VGJφ	orf359	359	43%	SAI Y WRI Y NKKLEQ
φ KSF-1φ	orf1	367	38%	SAI Y WRI Y NKKLEQ
φ VSK	Rep	368	37%	SRI Y WRI Y NKAAQL
φ VSKK	VSKKp3	367	38%	SAI Y WRI Y NKKLEQ
φ Vf12	Vpf402	402	36%	SRI Y WRI Y NKALEQ
φ Vf33	Vpf402	402	36%	SRI Y WRI Y NKALEQ
φ VfO4K68	Vpf402	402	35%	SRI Y WRI Y NKALEQ
φ VfO3K6	Vpf402	402	35%	SRI Y WRI Y NKALEQ
φ fs1 [1])	fs1p12	166	33%	
	fs1p13	208	46%	SAI Y WRI Y NKKLEQ
2A φX174-group			cons.	VxF Y **VAK** Y Vx**KK**SD N W T QFV
φ X174	A	513	x	VGF Y VAK Y VNKKSD
φ S13	A	513	96%	VGF Y VAK Y VNKKSD
φ G4	A	554	61%	VGF Y VAK Y VNKKSD
φ K (*E. coli*)	A	494	44%	VAW Y VTK Y VAKQSD
φ α3	A	494	43%	VAW Y VTK Y VAKQSD
φ chp1	orf4	399	25%	NIF Y VAR Y VQKKFV

φ / pφ / plasmid	gene	res.	ident.	motif 3
1B pT181-group			cons.	**S**xR F **IRI** Y **NK**Kx**E**R K Q V L R Q
pT181	RepC	314	x	SNR F IRI Y NKKQER
pCW7	RepN	314	86%	SDR F IRI Y NKKQER
pTZ12	Rep	314	83%	SDR F IRI Y NKKQER
pC223	Rep	314	77%	SNR F IRI Y NKKKER
S. agalactiae	sag0224	332	29%	SEK Q VRL Y NKKKEQ
pRS2	Rep	319	27%	SEK Q IRL Y NKRIER
2B φchp1-group			cons.	xhx Y **VA**R Y xx**KK**xx T **K**
φ chp1	orf4	399	x	NIF Y VAR Y VQKKFV
φ MH2K	Vp4	315	30%	SAS Y VAR Y AAKKLA
φ 3	orf4	315	30%	SAG Y VAR Y SLKKVN
φ 2	orf4	336	29%	SAG Y VAR Y SLKKVN
φ CPAR39	p7	327	27%	SAG Y VAR Y SLKKVN
φ CPG1	p9	263	28%	SAG Y VAR Y SLKKVN
φ 4	gene 2	320	28%	SAN Y TAR Y TTKKLG
φ X174	A	513	25%	VGF Y VAK Y VNKKSD

holds for the gene 4 helicase in the φT7 replisome. It would be highly interesting to learn whether the *E. coli* Rep helicase – or PcrA, its orthologue in *B. subtilis* – is inefficient for the coupling of leading- and lagging-strand DNA synthesis. A reason for such a hypothetical inefficiency could be the inability of Rep helicase to recruit the DnaG primase.

The binding of gpII to the φfd replication origin was studied to great detail in the lab of Horiuchi. Initially, it was found by filter-binding assays that gpII binds to φfd DNA in the superhelical or linear form. The gpII binding-site(s) could be mapped close to the site of initiation of DNA synthesis, overlapping the nick-site. The nick-site itself was found to be dispensable for gpII binding [180]. Using more advanced techniques – e.g. EMSA, DNase I footprinting and methylation interference – two distinct complexes between φfd origin DNA and gpII could be defined: complex I forms at low gpII concentrations at two inverted repeats (~25 bp) to the right of the nick-site (binding-sites β + γ; Fig. C1), complex II requires approximately twice the amount of gpII, and the bound region comprises of ~ 40 bp, including the nick-site (binding-sites α + δ). Binding of gpII to the origin in complex II occurs to the inverted repeat (β + γ) as in complex I, to a third repeat at the right of the inverted repeat (δ) and to the nick-site to the left without sequence specificity (α). This latter observation indicates that protein-protein interactions among gpII protomers stabilise the entire complex, allowing one protomer to cleave the nick-site [151]. Also it was found that complex I induces bending of the origin DNA, an effect that was even more pronounced for complex II [170]. Analysis of gpII complexes formed on supercoiled φfd RF DNA by the *in vitro* $KMnO_4$-footprinting technique revealed that complex II induces a local unwinding in the nick-site region as a prerequisite for efficient nicking [170].

In 1999, Horiuchi and co-workers could isolate the covalent complex between gpII and an artificial substrate mimicking nicked-origin DNA. By peptide sequencing, the covalent link could be traced down to Y_{197} . Substituting Y_{197} by phenylalanine resulted in a protein with wildtype DNA- binding properties *in vitro*. Also, the mutant protein induced origin bending like the wildtype protein but did not show any origin-nicking or topoisomerase activity [10]. The 'active' tyrosine of φfd

Table C9 RCR initiators, part B.

ϕ / pϕ / plasmid	gene	res.	ident.	motif 3
3 pTLC-group			cons.	Rxx x hxh Y xKxxEx Q S G S
pTLC	Cri	522	x	RSF E LCI Y TKHDEI
ϕ IKe	gpII	421	34%	RHR T LVA Y LKHVEV
ϕ M13	gpII	410	32%	RHR T LVA Y LKHVEL
ϕ f1	gpII	410	32%	RHR T LVA Y LKHVEL
ϕ fd	gpII	410	32%	RHR T LVA Y LKHVEL
pTA144up	RepC	336	24%	QFK G STL Y WGKGST
ϕ I2-2	gpII	344	24%	RRW S IKA Y SKYDEV
plasmid RCR initiator-groups lacking phage homologues (5+6 taken from [98])				
5 pLS1 / pMV158	RepB	210	x	NVE N MYL Y LTHESK
6 pC194	Rep	232	x	ELY E MAK Y SGKDSD
7 pJV1	Rep	528	x	DDV A LIE Y LTKNQD

ϕ / pϕ / plasmid	gene	res.	ident.	motif 3
4 ϕP2-group			cons.	Axx Y IAK Y IxKNID P V T L LN
ϕ P2	A	761	x	AAG Y IAK Y ISKNID
ϕ L-413C	gpA	761	97%	AAG Y IAK Y ISKNID
ϕ Wϕ	gpA	761	96%	AAG Y IAK Y ISKNID
Y. pseudo. pϕ	yptb1766	760	56%	AAG Y IAK Y IAKNID
ϕ PSP3	gp36	708	39%	ATG Y VAK Y ISKNID
ϕ 186	A	694	37%	ATG Y VAK Y ISKNID
ϕ Fels-2	stm2729	809	34%	PTS Y IAK Y ISKNID
pϕ CP-933T	z2978	940	31%	PTX Y IAT Y IGKNLD
ϕ HP2 / ϕ HP1	Rep	775	31%	ATA Y IAK Y IAKNID
ϕ PM2	p12	634	33%	AVG Y IAK Y ISKNIN
ϕ fs2	orf716	716	33%	AVG Y VAK Y LSKNID
ϕ V86 [2])	-	[532]	36%	ATG Y IAK Y ISKNID
ϕ K139	Rep	800	31%	ATG Y IAK Y ISKNID

The position of the signature motif 3 [188] containing the conserved active tyrosine are: in pT181 res. 184-198, ϕf1 res. 190-204, in pLS1 res. 92-106, in pJV1 res. 268-282, in pC194 res. 207-221, in ϕfs2 res. 437-453. The active tyrosine residues are boxed; bold letters indicate conservation of residiues in motif 3 among subgroups. ident.: % identical residues. cons.: consensus sequence. Sequences were taken from: CTXϕ [AAN06952], VGJ ϕ [NP_835472]; KSF-1 ϕ [YP_087070]; ϕVSK [NP_542355]; ϕVSKK [NP_536619]; ϕVf12 [YP_031686]; ϕVf33 [YP_031679], ϕVfO4K68 [NP_059541], ϕVfO3K6 [NP_059531], ϕfs1 [NP_695201, NP_695202], ϕX174 [NP_040703], ϕS13 [AAG29971], ϕG4 [AAL51008], ϕK (*E. coli*) [NP_043942], ϕα3 [NP_039590], ϕchp1 [NP_044320, NP_044321], ϕMH2K [NP_073537], ϕ3 [YP_022485], ϕ2 [NP_054653], ϕCPAR39 [NP_063900], ϕCPG1 [NP_510879], ϕ4 [NP_598335], ϕIke [NP_040570], ϕM13 [NP_510885], ϕf1 [CAA23876], ϕfd [AAA32303], ϕI2-2 [NP_039615], ϕP2 [NP_046795], ϕL-413C [NP_839887], Wϕ [NP_878237*]; Y. pseudotuberculosis* pϕ [YP_070292], ϕPSP3 [NP_958093], ϕ186 [NP_052289], ϕPM2 [NP_049896], ϕ Fels-2 [NP_461656], *E. coli* pϕ CP-933T [NP_288356], ϕHP2 [NP_536816], ϕHP1 [NP_043478], ϕK139 [NP_536641], ϕV86 [AAB66811], ϕfs2 [NP_047364], pLS1 [NP_040421], pJV1 [NP_044352], pC194 [CAA24585], pTA144up [NP_052248], pTLC [NP_862693], pT181 [NP_040469], pCW7 [AAA26669], pTZ12 [AAA72571], pC223 [CAA30291], *S. agalactiae* [NP_687259], pRS2 [NP_443751]. Table C8: 1) Overlapping orfs encode the proteins fs1p12 and fs1p23. Both predicted proteins show significant similarity to CTXϕ RstA in the N- and C-terminus, respectively. The splitting may be a mere sequencing artefact (reading-frame shift), therefore. Table C9: 2) the length of the unnamed ϕV86 protein is given in brackets because the sequence may be only partly correct (see Legend to Fig. C6 for details).

gpII is embedded in the motif 3-variant RHRTLV-A$\underline{Y}_{197}$LKHVEL (res. 190-204) which is conserved in other initiators (see below). Neither the domain responsible for DNA- binding, nor the interaction domain with the host Rep helicase have yet been determined for gpII, and also a crystal structure is not available.

Protein A of phage ϕX174. Synthesis of ϕX174 viral-strand circles *in vitro* requires gene A protein and in addition the *E. coli* proteins Rep, SSB, and DNA Pol III holoenzyme [118]. During replication, A protein nicks ϕX174 RF DNA and binds covalently to it; a complex with 1:1 stoichiometry could be isolated [116,110,50]. It was shown that the DNA bound covalently to A corresponds to (+)-strand DNA di-rectly adjacent to the replication origin, and that the bond involved tyrosine residues of A protein [377]. By peptide sequencing it could be shown that either of the two tyrosine residues in the motif YVALKYVNK can bind covalently to DNA [462]. The 'active' tyrosines of ϕX174 A protein are embedded in the motif 3-variant VGFY$_{343}$VAKY$_{347}$VN-KKSD (res. 340-353) which is conserved in other initiators (see below). Neither the domain responsible for DNA- binding, nor the interaction domain with the host Rep helicase have yet been determined for A protein, and also a crystal structure is not available.

RCR initiators of ssDNA phages. The results of BLAST searches suggests the definition of three distinct groups of RCR-initiators of ssDNA phages:

i. the ϕX174/ϕchp1-group. The orf14 protein of *Chlamydia* sp. phage ϕchp1 is only weakly similar to ϕX174 A protein, but all proteins detected by BLAST searches with either 'prototype' as query contain a well conserved motif 3 of the 'two- tyrosine'-type (Table C8; groups 2A + 2B). Historically, studies of ϕS13 replication preceded the extended studies of ϕX174 [448].
ii. the CTXϕ/pT181-group. The proteins of the CTXϕ- and pT181-subgroups are of comparable size but there is no detectable sequence similarity among them (Table C8; groups 1A + 1B). The proteins of the CTXϕ-subgroup contain a conserved motif 3 of

the 'two-tyrosine'-type, while the proteins of the pT181-subgroup contain a single tyrosine in their motif 3. Despite this difference which certainly reflects subtle functional differences of the proteins, the motif 3-variants present in both subgroups are more closely related to each other than to any other known motif 3-variant. The CTXϕ- subgroup can be further divided in three 'branches': i. the CTXϕ RstA-type, ii. the ϕVSK Rep-type (ϕVGJ, KSF-1ϕ, ϕVSKK, ϕfs1), and iii. the ϕVf12 Vpf402-type (ϕVf33, ϕVfO3K6, ϕVfO4K68). Within each branch, the proteins are virtually identical, but the similarities to proteins of the other two branches does not exceed 45% (ident. res.) and excludes the N-termini.

iii. the pTLC-group. The Cri protein of plasmid pTLC is similar to the gene II proteins of phages M13, f1, and fd, and all proteins of this group contain a motif 3 variant of the 'one tyrosine'-type.

We have listed in Tables C8 + C9 groups of plasmid- and phage-encoded RCR-initiators by sequence similarity which in all cases corresponded with a conserved variant of motif 3. Groups 1A + 1B and 3 contain proteins encoded by phages and plasmids. Plasmid-encoded initiators belonging to groups 2 or 4 (ϕP2-group, discussed below) could not been found. Also *vice versa*, phage-encoded initiators with similarity to pLS1/ pMV158 (group 5), pC194 (group 6), and pJV1 (group 7) were not detected. This shows that there is no promiscuous exchange of genes for these functionally analogous proteins among the different replicon types. However, a report of Waldor and co-workers suggests that an intricate molecular interplay exists between the cholera-toxin encoding CTXϕ phage (Table C8; group 1A) of *V. cholerae* and the plasmid pTLC (Table C9; group 3) [379].

Recently, Chopin and co-workers have characterised ϕB5, a filamentous phage with ssDNA genome of a Gram-positive host [77]. However, the putative replication protein orf9 does not bear any resemblance with any of the known initiators discussed above, and the replication modus ϕB5 remains unknown therefore.

Initiator proteins of phages with dsDNA genomes: P2 A protein. Haggård-Ljungquist and co-workers could show that the 76.7 to 91.6% region of bacteriophage ϕP2 (genome size 33.6 kb, dsDNA) is necessary and sufficient to drive plasmid replication [258]. This region contains – among several orfs with unknown function – the replication genes A and B, and the replication origin located within gene A (pos. 89±1%). The helicase loader encoded by gene B is discussed in the following chapter, the ϕP2 replication origin in Section C2.1. . These authors could show that the isolated A gene is sufficient to drive ϕP2 mini-chromosome replication. Egan and co-workers could show that gene A protein (CP87) of ϕ186 is also sufficient to drive plasmid replication [417].

ϕP2 A protein contains an $AAGY_{450}IAKY_{454}ISKNID$ motif that corresponds to motif 3 defined by Koonin and Ilyina [227]. Both tyrosine residues could be shown to be essential for RCR of ϕP2 [325]. Y_{454} is apparently more proficient in the initial nicking reaction while the nicking activity of Y_{450} can be observed when Y_{454} is covalently linked to a specific oligonucleotide. These observations led Haggård-Ljungquist and co-workers to suggest a coupling of termination and re- initiation of ϕP2 replication by A protein, mediated by the two tyrosine residues. It is known that ~100 res. of the C-terminus are dispensable for ϕP2 A functioning during replication *in vivo* and *in vitro* [325], but domains involved in DNA-binding and the interaction with B protein or the host helicases – DnaB and Rep [59,43] – have not yet been identified.

(Putative) initiators for RCR with similarity to ϕP2 A were detected in the genomes of several phages of γ-proteobacterial hosts, and also encoded by prophage genes in the genomes of many γ-proteobacteria (Table C9; group 4). The similarities to ϕP2 A and the sizes of these proteins vary considerably, and a sequence alignment could only reveal several patches of homology in a 'core region' of ~300 res. in length (Fig. C6) but no true consensus sequence. Interestingly, the alignment suggests the existence of several distinct sub-families, which share a (predicted) comparable secondary structure in the 'core region'. The motif containing the two catalytically important tyrosine residues is well conserved (Table C9; group 4), and located at a corresponding position in the C-terminal part of the 'core region' in all sequences (Fig. C6; light-blue vertical bar). The A proteins of ϕP2 and ϕ186, respectively, have both been shown to function as initiators although their similarity is limited. From the above we conclude that all proteins of the ϕP2-group are *bona fide* initiators (Table C9; group 4).

A distantly related member of the ϕP2-group of RCR-initiators is the Arp protein encoded by the 1282 bp long phasyl (phage-plasmid hybrid) [140]. With only 285 res. Arp is considerably shorter than ϕP2 A and a reasonable alignment of Arp with the ϕP2-group could not be obtained (not shown). However, similarity (27% ident. res.) could be detected for a stretch of ~150 residues within the 'core region' between Arp and the pϕFels-2 stm2729 protein, and Arp contains the motif 3-variant IGRYVGKYISKGIE at a corresponding position.

Plasmid-encoded initiator proteins for RCR with significant similarity to ϕP2 A could not be detected by BLAST searches, but there is a puzzling similarity of ϕP2 A protein to the orf2 protein [SWISS YR72_ECO-

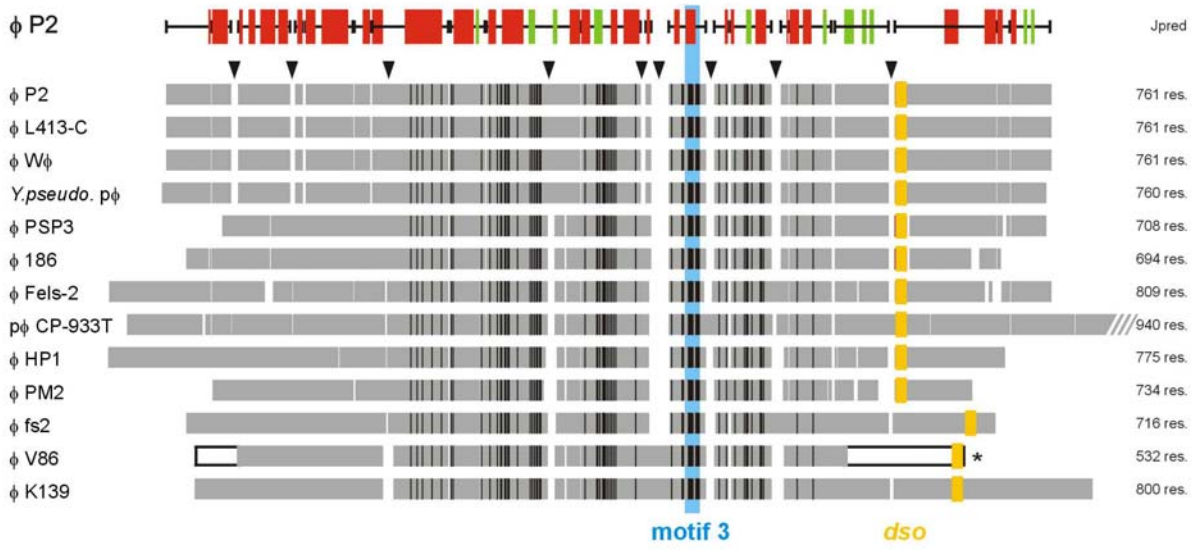

Fig. C6 ϕP2 A-type (pro)phage-encoded initiators for RCR.
First line: secondary structure prediction for ϕP2 A by the Jpred method [88]. Colour code: red = predicted α-helical region; green = predicted β-stranded region; black line = unstructured/no prediction; the sequence was split into segments to allow comparison with the sequence alignment below. Following lines: sequence alignment of proteins with similarity to ϕP2 A by 'MultAlin' [84]. For gene names and accession numbers see Table C9 . Aligned sequences are shown as grey blocks. Gaps are shown in white. Identical residues in all sequences are indicated by black bars. The vertical light-blue bar indicates the position of motif 3; orange squares mark the position of the (putative) replication origin (*dso*) in the respective gene (see Table C2; Section C2.1. for discussion). The sequence of pϕ CP-933T z2978 protein was shortened at the C-terminus to fit to the figure. The sequences of the homologous ϕK139 Rep and (unnamed) ϕV86 proteins are identical on the nucleotide level (99.5%). We therefore tentatively corrected the ϕV86 sequence at two positions (indicated by open boxes) i. choosing the alternate GUG start codon 40 triplets upstream of the AUG assigned as start in entry AF008938 allows the perfect alignment with the ϕK139 Rep protein in the N-terminus (the Rep gene also starts with GUG). ii. 2 additional bases inserted shortly after codon 527 (original sequence) result in a reading-frame shift and premature stop, *in silico* correction restores the ϕK139 Rep reading frame. The sequenced part of the ϕV86 gene ends after codon 678 (counted from the corrected N-terminus), i.e. one triplet downstream of the putative *dso* (indicated by '∗').

LI] encoded by the *E. coli* retro-element Ec67 [256]. However, the similarity does not exceed 27% ident. res. for the N-terminal ~350 res & does not include motif 3. Therefore, a role for this protein for retron propagation remains speculative.

C3.1.2. Initiator proteins for DNA replication in the 'theta-mode' (θDR)

Proteins that initiate replication in the θ-mode perform two major functions: i. they bind to the replication origin and promote duplex unwinding, and ii. they direct other replication proteins – the helicase loader(s) and/or the replicative helicase – to the exposed single-stranded regions, thus functioning as a 'replisome organiser'.

Initiators bind to the iterons present in the replication origins for θDR adjacent to an AT-rich region (Section C2.2.), and binding occurs – depending on the iteron type – in the monomeric form to non-palindromic iterons and in the dimeric form to palindromic iterons. In addition to binding of initiator protomers to the iterons, unwinding requires protein·protein interactions between two ore more protomers. In most cases, the properties for DNA-binding and for oligomerisation are located in distinct regions of the proteins, which retain their partial function upon separation. Therefore, prokaryotic initiators are composed of – at least – two distinct domains: a DNA-binding domain, and an oligomerisation domain. A third domain contiguous with – or overlapping – the oligomerisation domain is required for the physical interaction with the replicative helicase, or the helicase loader(s).

(Almost) all sequenced bacterial genomes contain *dnaA* genes, and DnaA may be the 'universal' initiator for the replication of chromosomal replicons in bacteria. *E. coli* DnaA binds to dsDNA with its C-terminal domain 4 [376]. The N- terminal domain 1 of DnaA and – in addition – a recently identified structural motif in the central NTP-binding domain 3 are responsible for oligomerisation [473,121]. The physical interaction between DnaA and the replicative helicase DnaB could be shown to involve two domains of either protein: res. 24-86 (domain 1) and 130-148 (domain 3) of DnaA and res. 154-210 (βγ-fragment) and 1-156 (α-fragment) of DnaB, respectively [278,404]. It is not known whether *E. coli* DnaA interacts specifically also with the helicase loader, DnaC, but experimental results supporting this possibility are not available (see following chapter). In contrast to the known initiators of bacterial plasmids

and phages, DnaA is a NTP- binding protein, and the proficiency of DnaA for origin unwinding is strictly dependent on the ATP-bound form [423]. The properties and functions of DnaA have been reviewed by Skarstad and Boye [418], and more recently by Messer [290].

The molecular architecture of various initiators of plasmids replicating by the θ-mode – including e.g. pPS10 RepA, R6K π, R1 RepA – is comparable to that of DnaA: a N-terminal dimerisation/oligomerisation domain, and a C- terminal DNA binding domain. However, there is no apparent sequence similarity between the plasmid initiators and DnaA, and the structural motifs performing analogous functions are clearly different. A physical interaction between the plasmid initiator and the host replicative helicase was shown for R6K π and pSC101 RepA, and involves similar regions in DnaB as found for DnaA [362,91]. The structures and functions of plasmid initiators have been reviewed by Giraldo [142] and del Solar and colleagues [98].

Besides being essential for replication initiation, DnaA serves as transcription factor in E. coli [293]. Most importantly, DnaA acts as repressor of its own gene in a complex regulatory feedback loop involving other factors in addition. Autoregulation is also known for plasmid initiators and serves as part of the copy-number control mechanisms operating in plasmid replicons [98, 97]. A role as transcription factor has not been found for the initiator proteins of bacteriophages – λ O and ϕSPP1 G38P – and a role of these proteins as autorepressors can be excluded.

We discuss in detail bacteriophage Lambda O protein, and Gene 38 protein (G*38*P) of *B. subtilis* phage SPP1. Few other phage-encoded replication initiators of the 'λ O'-type have attained attention. We could detect by BLAST searches 140 putative 'λ O'-type initiator sequences from phages or prophages in the sequenced bacterial genomes, and we present (theoretical) evidence that most if not all proteins are *bona fide* initiators for the replication of their cognate (pro)phages. In addition, the RepL initiator for bacteriophage P1 replication in the lytic cycle and the C-terminal 'initiator domain' of phage P4 alpha (α) protein and related multifunctional helicases are discussed.

Bacteriophage Lambda O protein. ***Gene***tic experiments with hybrid phages obtained by *in vivo* crosses between λ and ϕ80 suggested a dual role for λ O protein as initiator: recogni-tion of the λ replication origin by the N-terminus, and interaction with the λ P helicase loader by the C-terminus [440,133, 134]. The crossover points in these hybrids were mapped to the middle of the *O* and 15 genes, respectively, and it was demonstrated by complementation assays that only hybrid O proteins containing the λ O N-terminus showed specificity for *ori*λ. Similar complementation assays showed that a hybrid O protein carrying the C-terminus of ϕ80 gene 15 protein could only function in replication with the gene 14 product of ϕ80. It was concluded that the C-terminus of O protein interacts specifically with P. The physical interaction of O and P was later confirmed biochemically by Zylicz and co-workers [509] (see next chapter).

Matsubara and Tsurimoto analysed purified λ O for its binding specificity to the four *ori*λ iterons *in vitro* by exonuclease III protection, DNase I footprinting, and filter-binding assays [458,459]. At low protein concentration, only the inner two iterons were bound, while binding to all four iterons occurred at higher protein concentrations, covering up to ~95 bp of linear oriλ DNA fragments. The authors could not detect DNA-binding for a mutant O protein lacking 20 res. in the N-terminus, corroborating the results from genetic studies that the O N-terminus is responsible for specific binding to *ori*λ (see above). In addition, their results suggested that dimers of O bind to single iterons, i.e. eight protomers per oriλ. Wickner and Zahn confirmed the above results and could show, in addition, that the isolated λ O N-terminus is necessary and sufficient for specific binding to *ori*λ. The O N-terminal fragment cannot drive λ (λ*dv*) replication *in vitro* but inhibits this reaction by competing with wildtype O. They could also show that *ori*λ-bound O protein – but not the isolated O N-terminus – forms a stable complex with P [481].

Detailed structural studies of the λ O·oriλ complex and of the binding of O to isolated iterons were performed by Blattner and co-workers, employing a battery of *in vitro* techniques [499,501,396,395]. The emergent picture for the interaction of O with its cognate DNA binding-site(s) is: binding of O augments the intrinsic curvature of *ori*λ DNA (Section C2.2.), and the binding of O (dimers) already induces bending of the isolated iteron DNA. The 19 bp (imperfect) inverted repeat of the iteron is the minimal sequence required for efficient binding by O, and iteron binding + bending is exclusively dependent on the N-terminal domain of O suggesting that this domain has the property to dimerise. Asymmetrically located protected regions found by footprinting experiments for the O-bound iteron suggest, in addition, its partial distortion. The DNA distortions induced by O-binding in supercoiled *ori*λ DNA can be traced well into the AT-rich region. Up to ~84 bp of *ori*λ DNA are wrapped around – or partially buried within – a core of oligomerised O protomers with a diameter of ~50 Å.

The four 19 bp iterons of the λ replication origin, *ori*λ, posses dyad symmetry, are regularly spaced to each other (repeat: 1 helical turn), and located adjacent to the AT-rich region (Section C2.2.). Three iterons are sufficient for efficient replication, as suggested by the normal growth and burst size of λ*r93hot5* [299]. For *in*

vitro replication of *ori*λ- based plasmids, however, only the two iterons proximal to the AT-rich region seem to be required [480]. Similar observations were made with the DnaA-binding sites (DnaA boxes) in the *E. coli* replication origin, *oriC*: alterations of the helical phasing of the DnaA boxes had more severe negative effects than inactivation, and these effects were most pronounced for DnaA box R1, directly adjacent to the AT-rich region [239,290]. It should be kept in mind that the initiators bound to the iterons proximal to the AT-rich region are probably not only important for the unwinding reaction but also for the correct positioning of the replicative helicase [474].

When λ O protein remains bound to DNA after departure of the replication forks from *ori*λ it becomes resistant to degradation by ClpP/ClpX protease in the so-called 'inherited replication complex' [149,485,471, 446]. Initiation of new replication rounds from the inherited replication complex *in vivo* does not require *de novo* synthesis of O but still depends on host DnaA-dependent transcriptional activation of p_R, which triggers chaperone action for the liberation of the DnaB helicase from the λO·λP·DnaB complex (see below). It is not entirely clear at present whether and how the switch from θDR to σDR during λ replication is triggered by the inherited replication complex.

Gene 38 protein (G*38*P) of *B. subtilis* phage SPP1. Studies of the replication of phage SPP1 date back to the 1970s, and could reveal over the years the essential nature of the products of genes 38, 39 and 40. ϕSPP1 replication is independent of the host DnaA, DnaC helicase, the DnaB, DnaD and DnaI helicase loaders, but requires the host DnaG primase, host gyrase, and DNA polymerase III holoenzyme [55,341]. G*40*P was identified as the cognate helicase for ϕSPP1 replication (Section C3.3.), and G*39*P as the cognate helicase loader (Section C3.2.). G*38*P binds specifically two discrete regions in the SPP1 genome: *ori*L, the replication origin located within gene 38, and *ori*R, an origin-like structure at a distance of ~32 kb from *ori*L (Section C2.2.) [341]. Alonso and co-workers could measure specific, cooperative, and tight binding of G*38*P to *ori*L (K_{app} = 1 nm) and to *ori*R (K_{app} = 4 nm) *in vitro*, and protection of the AB iterons present in *ori*L and *ori*R – but not of the AT-rich region – by DNase I footprinting experiments [297]. As discussed in the following chapter, G*38*P does by itself not interact with its cognate helicase: when bound to the replication origin *ori*L, G*38*P interacts specifically with G*39*P in the G*39*P·G*40*P·ATP complex, which results in G*40*P binding to the origin [13]. The interaction domains of G*38*P and G*39*P have not yet been determined.

The generally accepted model for the complex between the initiator and its cognate origin DNA is that of a nucleosome-like structure, i.e. the DNA wrapped around a protein core – the λ O·*ori*λ complex ('O-some') may serve as example (see above) [446]. The core of the complexes would be held together by protein·protein interactions between iteron-bound initiator protomers, and the wrapping would create sufficient stress on the DNA helix to promote its unwinding. However, this model is not supported by electron microscopic analysis of the G38P·*ori*L complex [297]. This analysis suggested that the binding of G*38*P to the *ori* iterons results in DNA bending – as was observed for λ O (see above) –, stabilised by protein·protein interactions between G38P protomers. However, a 'shortening' of the DNA fragment – indicative of DNA wrapping – was not observed, which makes a nucleosome-like structure of the G38P·*ori*L complex unlikely. Until more experimental studies of initiator·origin complexes are presented the question of their higher-order structure(s) remains an intriguing issue.

Few other putative phage replication initiators have attained attention. Purified orf16(Rep) protein of ϕTuc-2009 was found to bind specifically to an internal fragment of the orf16 gene containing the putative replication origin, which qualifies the protein as initiator for the replication of this *L. lactis* phage [282]. The orf11 initiator of ϕr1t was identified by its similarity to ϕSPP1 G*38*P; (partially) purified orf11 protein binds to its own coding region, and the nucleoprotein complexes analysed *in vitro* support the notion of extensive oligomerisation of orf11 bound to the (putative) replication origin [508]. A preliminary genetic analysis of ϕTP901-1 replication identified REP(orf13) as replication protein containing the replication origin [333].

Initiator proteins of the Lambda O / SPP1 G*38*P-type. BLAST similarity searches were performed with the λ O and ϕSPP1 G*38*P protein sequences as query. A final set of 140 putative initiator sequences was accumulated in an iterative manner, i.e. the searches were repeated with the least similar (putative) phage initiator from the results lists as new query sequence until no further new matching sequences were found. In most cases, the results of BLAST searches for a given query sequence showed four types of matching sequences:

i. orthologues with similarity values close to 100% (ident. res.) and full-length homologues with values >60%;
ii. proteins with full-length similarity in the range from 40% to 60%;
iii. distantly related sequences with values in the range from 20% to 40% over the entire length, for the N-terminus, or for the C-terminus;
iii. distantly related sequences with values in the range from 20% to 40% over the entire length, for the N-terminus, or for the C-terminus;

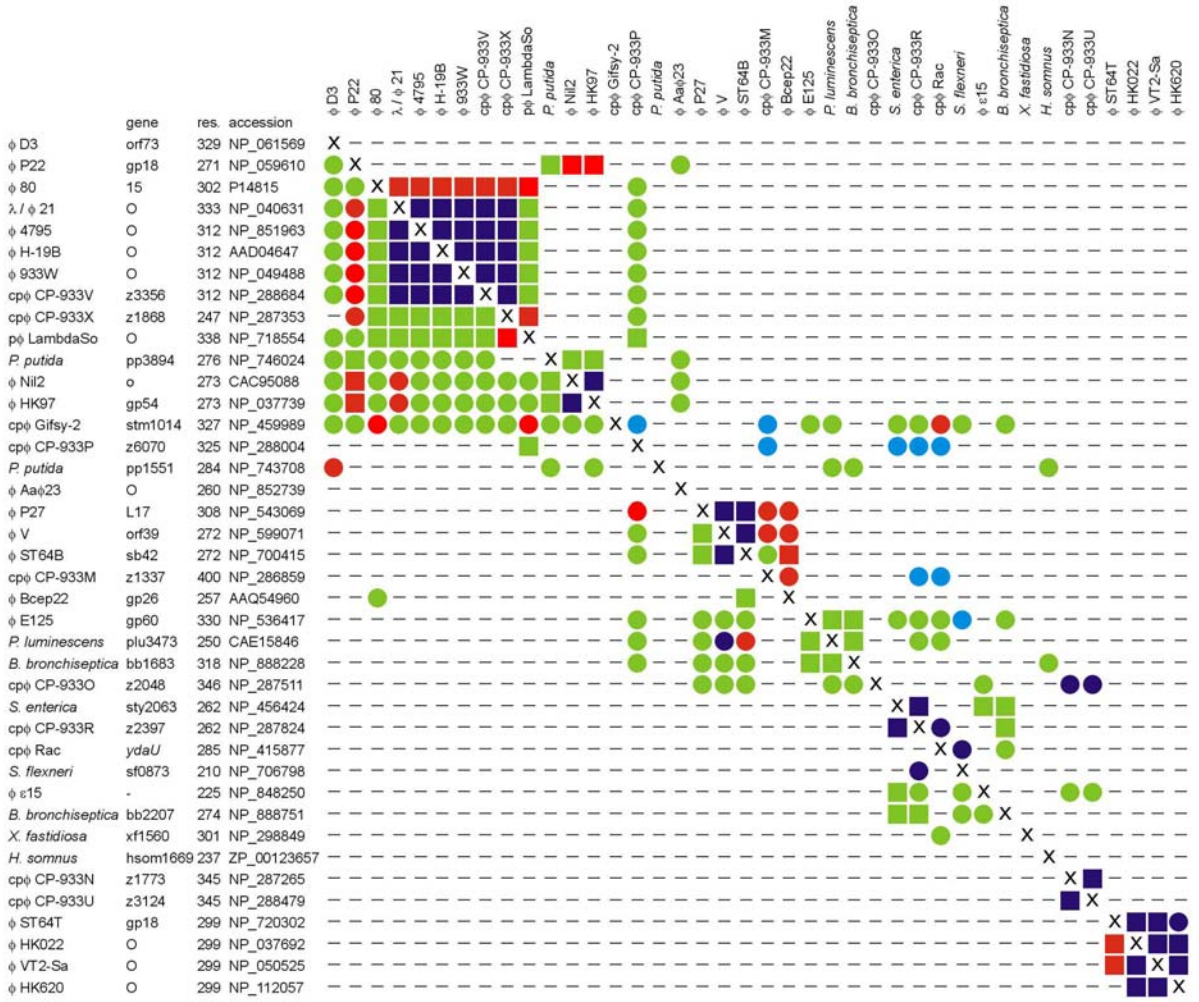

Fig. C8 Protein sequence similarity among initiator proteins of Gram(−)-specific phages.
BLAST searches were performed separately for the N-termini (res. N-~150) and C-termini (res. ~150-C) of the query-proteins. Similarities found for the N-termini are given below the diagonal; similarities found for the C-termini are given above the diagonal. Only regions of similarity with a length ~40 res. or more were considered significant except for a small region (~ 20 res.) with particularly high similarity (~70% ident .res.) at the extreme C-terminus in a subset of the proteins compared here (light blue filled circles). Similarities are shown in green (≥20-<40% ident.res.). red (≥40-<60% ident. res.). and blue (≥60% ident. res.). '–' no detectable similarity. Filled squares indicate similarity of two proteins in the N-terminus and in the C-terminus; filled circles indicate similarity in the N-terminus, or the C-terminus, respectively.

iv. proteins with highly significant similarity – even close to identity – exclusively to the N-terminus or to the C-terminus of the query sequence.

The majority of the matches were phage proteins or 'hypothetical phage proteins' from prophages in the sequenced bacterial genomes. Not surprisingly, a number of matches to putative transcription factors were obtained for the N-termini (putative DNA-binding domains). *Gene*tic drift can easily explain the similarity types i.–iii.. For type iii., detectable low similarity in either the N- or the C-terminus could have coincided with a degree of similarity in the second part of the proteins that was below the threshold level set by the BLAST parameters (~20% ident. res.). However, we observed for a number of proteins with full-length similarity in the ~40% to 60% range that the similarity for one domain was considerably higher than for the other, pointing to a two-domain architecture of all these proteins (Figs. C7 + C8). Similarity type iv. cannot be explained by genetic drift, and we have to assume instead an exchange of initiator domains among phages

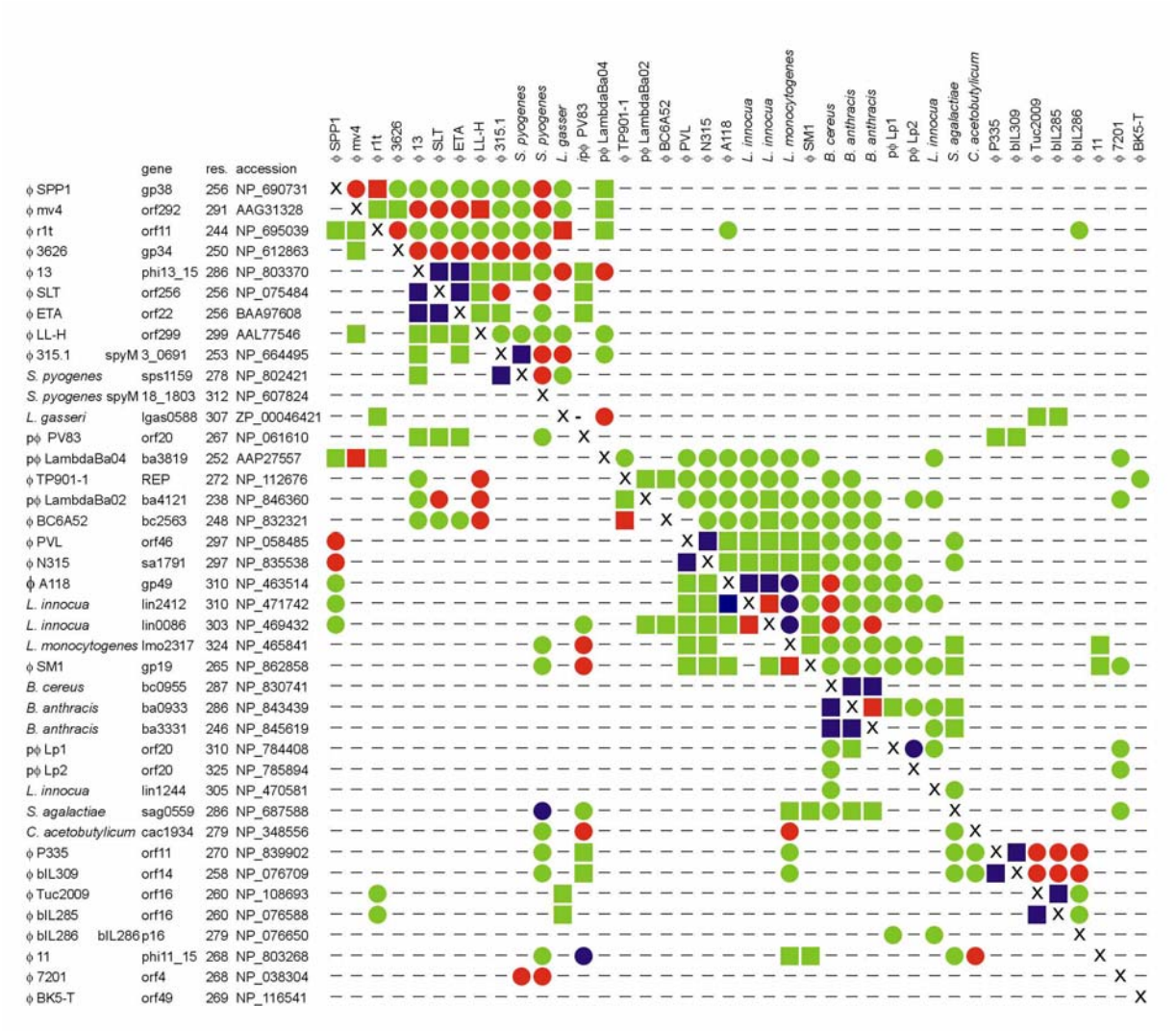

Fig. C7 Protein sequence similarity among initiator proteins of Gram(+)-specific phages.
BLAST searches were performed separately for the N-termini (res. N-~150) and C-termini (res. ~150-C) of the query-proteins. Similarities found for the N-termini are given below the diagonal; similarities found for the C-termini are given above the diagonal. Only regions of similarity with a length ~40 res. or more were considered significant. Similarities are shown in green (≥20-<40% ident.res.), red (≥40-<60% ident. res.), and blue (≥60% ident. res.). '–' no detectable similarity. Filled squares indicate similarity of two proteins in the N-terminus and in the C-terminus; filled circles indicate similarity in the N-terminus, or the C-terminus, respectively.

by recombination, resulting in the extensive mosaicism found for these genes encoding functional analogous proteins.

An instructive example is shown in Table C10 . The genes for the (putative) initiators in the genomes of the *Staphylococcus aureus* phages ϕPV83 and ϕ77, of a prophage from *S. aureus* Mu50, and of the *Burkholderia cepacia* (β-proteobacteria) phage ϕBcep22 are located directly upstream of (putative) helicase loader genes of the $DnaC_{Eco}$-type (see next chapter). This type of arrangement of (putative) replication genes is discussed as 'initiator-helicase loader' (IL-type) replication module in BRM Section 3.1. . All four $DnaC_{Eco}$-type proteins show full-length similarity of varying degrees to each other; the high similarity between the genes of ϕ77 orf013 and ϕPV83 orf21 corresponds to the significant similarity of the DNA sequences of both genes (Fig. C9; C). In contrast, the similarities among the (putative) initiators are 'patchy', and confined to either the N-terminus or the C-terminus. The *S. aureus* pϕ protein sav0868 shows almost complete identity to the ϕ77 orf016 N-terminus and the ϕPV83 orf20 C-terminus, but no detectable similarity to the respective other domains. A comparison of the DNA sequences suggests that entire blocks were

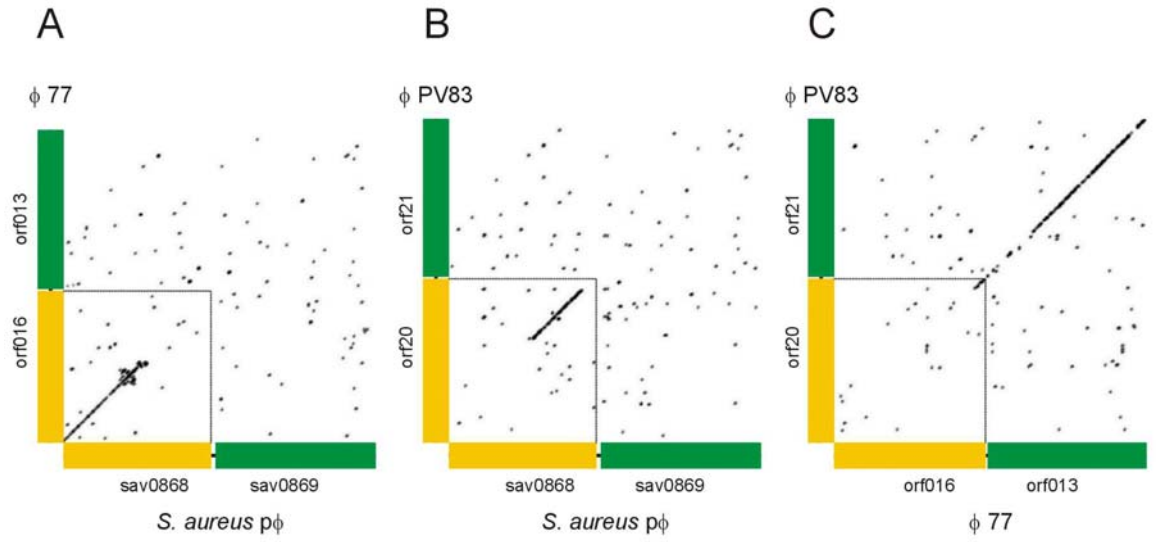

Fig. C9 DNA sequence similarity among (putative) replication genes: ϕ77orf016/orf013, ϕPV83 orf20/orf21, and *S. aureus* prophage sav0868/sav0869.

Dot-plot matrix analysis of the DNA sequences of the (putative) initiator and helicase loader genes of phages ϕ77 (pos. 32867-34399, 1532 bp [NC_005356]), ϕPV83 (pos. 9476-11062, 1586 bp [NC_002486]), and an *S. aureus* pϕ (pos. 927419-928948, 1529 bp [NC_002758]). Dot-plots were obtained using 'method 2' (K_tuple value = 8) of the dot matrix subprogram of the DNAMAN® software (version 4.0; Lynnon Inc.). **A** : *S.aureus* pϕ X ϕ77; **B** : *S.aureus* pϕ X ϕPV83; **C** : ϕ77 X ϕPV83. Yellow blocks: putative initiator; green blocks: (putative) $DnaC_{Eco}$-type helicase loader; black line: intergenic region. Gene sizes are shown exactly to scale. Dotted lines indicate the 3'-ends of the initiator genes.

Table C10 Sequence similarity among (putative) replication proteins: ϕ77orf016/orf013, ϕPV83 orf20/orf21, and *S. aureus* prophage sav0868/sav0869.

ϕ / pϕ	gene	res.	accession	*S. aureus* pϕ N	*S. aureus* pϕ C	ϕ PV83 N	ϕ PV83 C	ϕ 77 N	ϕ 77 C
initiator									
S. aureus pϕ	sav0868	243	NP_371392	x	x				
ϕ PV83	orf20	267	NP_061610	–	95	x	x		
ϕ 77	orf016	247	NP_958644	98	–	–	32	x	x
ϕ Bcep22	gp26	257	NP_944256	28	–	–	–	26	–

gene	res.	accession	*S. aureus* pϕ	ϕ PV83	ϕ 77
helicase loader					
sav0869	261	NP_371393	x		
orf21	257	NP_061611	43	x	
orf013	259	NP_958645	44	87	x
gp27	265	NP_944257	26	30	30

BLAST (bl2seq; [445]) similarity searches were performed for the N- and C-termini of the initiator proteins separately, and for the (putative) helicase loaders with the complete sequences. Values are % ident. res.; '–' no detectable similarity.

exchanged by recombination among ancestors of these (pro)phages (Fig. C9; A+B). As was also observed for most of the (putative) initiator genes with high N-terminal similarity, the similarity on the DNA level extends from the 5'-end to a region downstream of the iteron region defining the putative replication origin (Fig. C9; A: note small strings parallel to the diagonal indicative of repeats close to the middle of the (putative) initiator genes sav0868 and orf016). Contrarily, ϕ77 orf016 and ϕPV83 orf20 show only moderate similarity to each other, confined to the C-termini, but with complete identity for the C-terminal 20 res. (Fig. C9; C: see DNA sequence identity in the extreme 3'- ends of the initiator genes). It is obvious that conventional methods used to determine evolutionary distances in a set of analogous proteins on the basis of mutation rates in the full-length

sequences would produce highly biased results for mosaic proteins like the phage initiators.

From the 140 (putative) initiator sequences found by 'saturation' BLAST searches 40 were chosen as examples from Gram(+)-specific (pro)phages (Fig. C7) and another 40 from Gram(−)-specific (pro)phages (Fig. C8) for a matrix-type comparison. The number of 40 was chosen so that the matrix would still fit to a page-size figure; the selected sequences represent closely related as well as seemingly unrelated sequences. Similarities between (putative) initiators of Gram(+)-specific and Gram(−)-specific (pro)phages were rarely found, and mostly confined to the N-terminal DNA-binding domains with similarity values in the range of ~30% or below. Although we cannot exclude interbreeding between phages of both groups – including exchange of parts of their (putative) initiator genes by recombination – such rare events are hardly detectable by crude methods like BLAST comparisons.

The matrices shown in Figs. C7 + C8 give an simultaneous synopsis of similarities calculated separately for the N- and C- termini because it was clear from the above that only this procedure would allow for a meaningful comparison. The sequences were arranged manually in a way that nearly identical sequences were placed as closely as possible to each other. Around these 'foci', the remaining sequences were placed according to their decreasing similarity. It is immediately apparent from the matrices that several subgroups exist within both sets. Tentatively, we would call them the λ, the ϕP27/ pϕ Rac, and the ϕVT2-Sa subgroups for the Gram(−)-specific phages. For the Gram(+)-specific phages, we would tentatively define the ϕETA, the ϕA118, and the ϕP335 subgroups. However, the 'borders' of these subgroups are fuzzy, and there are examples in both matrices for sequences that fit into different subgroups with their N- and C- termini, respectively, e.g. the initiators of the Gram(−)- specific prophages Gifsy-2 and CP-933P, and the initiators of the Gram(+)-specific phages ϕSPP1, ϕ11 and ϕPV83. These protein sequences belonging to different subgroups support the notion of recombinatorial crosstalk between the subgroups, and create a 'link' – on the 'BLAST level' – of unrelated sequences, e.g. ϕSPP1 G38P and ϕBK5-T orf49. For both sets, however, the sample size is not sufficient for a prediction whether further sequences would consolidate the existing subgroups, create novel subgroups, or add more examples for sequences belonging to different subgroups with each of their two domains.

For most of the initiator proteins from both sets with significant similarity to each other, a correlation exists with the type of replication module driving the replication of these phages (BRM Section 3.1.). However, there are exceptions: the initiators of the Gram(−)-specific phages ϕBcep22 gp26, ϕV orf39, and ϕP27 L17 have significantly similar C-termini, but ϕBcep22 has an IL-type replication module, ϕV an 'Initiator-solo' (I-solo type) replication module, and ϕP27 an 'Initiator-helicase loader-helicase' (ILH-type) replication module. Also, the initiators of the Gram(+)-specific phages ϕSPP1 G*38*P and ϕr1t orf11 are similar, but ϕSPP1 has an ILH-type replication module, and ϕr1t an IL-type module. Therefore, reliable predictions about the replication mechanism of a given (pro)phage can only be made after an analysis of the complete set of replication genes in its genome, particularly with respect to the point of entry of the host replication machinery (see below).

In conclusion, there are four lines of (theoretical) evidence that most if not all proteins analysed in the 2 matrices are *bona fide* initiators for the replication of their cognate (pro)phages: i. although a direct one-to-one comparison of two proteins fails to reveal similarity for most of them, it is possible to find 'connecting' protein(s) with similarity to both, ii. all proteins are of similar size and show the two-domain architecture of λ O with N-terminal DNA-binding and C-terminal oligomerisation/interaction domains, iii. for most of the initiator genes a co-localisation with other replication genes – i.e. helicase loader and/or helicase genes – can be detected in a defined part of the respective phage genomes (BRM Section 3.1.), and iv. in >90% of the 80 initiator genes, (putative) replication origin structures could be detected located approximately in the middle of the genes (Section C2.2.).

Several (putative) initiator proteins of Gram(+)-specific (pro)phages contain in their C-terminal domains regions of significant similarity to the DnaB and/or DnaD helicase loaders of their hosts. This topic is discussed in more detail in BRM Section 3.1.1. . In a few instances, we found similarities of the N-termini of phage initiators to plasmid-encoded proteins. The *Streptomyces* sp. plasmid pSCL contains a replication module strongly resembling the IH-type phage replication module (BRM Section 3.1.3.). The N-terminus of the putative initiator of pSCL shows similarity to the N-terminus of *E. coli* phage P27 initiator L17 (27% ident. res.). ϕP27 L17 also shows similarity (31.% ident. res.) in its N-terminus to the N-terminus of the RepC protein of pTAV320 (RSF1010-type; host: *Paracoccus pantotrophus*, α-proteobacteria). Taken together, there are only few indications for a closer relationship of plasmid and phage initiator proteins. An intriguing finding is the N-terminal similarity (25% ident. res.) of the *Bacteroides thetaiotaomicron* (Chlorobium/Bacteroides group) gene bt1515 to ϕSPP1 G*38*P and ϕPVL orf46 because this gene is followed downstream by a (putative) $DnaB_{Eco}$-type helicase, reminiscent of the IH-type replication module (BRM Section 3.1.3.). No phages specific for species of this phylon have yet been described at

the sequence or molecular level. These two genes may indicate that prophages similar to Gram(+)-specific phages are present in the genomes of species from this phylon.

RepL, the initiator for bacteriophage P1 replication in the lytic cycle. *E. coli* cells lysogenic for phage P1 carry the prophage either integrated into their chromosomes or as low-copy number plasmid [76,259]. Plasmids containing the ϕP1 R (plasmid) replicon have been among the favourite models for studies of plasmid replication and plasmid copy number control [98,74]. The L (lytic) replicon – responsible for ϕP1 propagation in the lytic cycle – has attracted considerably less attention.

Early genetic analyses suggested that different replicons might be responsible for the propagation of ϕP1 as plasmid or as phage. Experimentally, the phage burst size and specific DNA synthesis were measured in a set of *E. coli* strains carrying temperature-sensitive mutations in replication genes. Phage propagation could be shown to depend on host DnaC, DnaG, and Pol III holoenzyme, but did not require DnaA and DnaB [165]. In contrast, ϕP1 plasmid replication crucially depended on host DnaA and DnaB, and – in addition – on DnaC, DnaG, and Pol III holoenzyme [163]. Analysis of ϕP1 replication intermediates by electron microscopy revealed theta-shaped molecules early after infection of *E. coli* and predominantly sigma-shaped molecules later on. Such σ-shaped molecules were rarely found in *recA* host cells. Cohen concluded from these observations that initiation of ϕP1 replication occurs in the θ-mode and switches to the σ-mode later depending on host recombination functions [81]. A switch from θDR to σDR is also known for λ and ϕSPP1 replication but is dependent only in the case of ϕSPP1 on the cognate recombination functions (Section C3.6.2.). In independent studies, Hansen [162] and Cohen + Sternberg [82] could show that the ϕP1 L replicon is contained within a DNA segment that carries the *c4*, *ant*, *kilA* – all three non-essential for replication – and *repL* genes. Transcription of the *kilA* and *repL* genes is driven by an upstream promoter, P53, which is negatively regulated by the ϕP1 cI repressor. Insertional inactivation of the *repL* gene abolishes L replicon activity. A putative replication origin structure is located within the *repL* gene, comparable to the localisation of the replication origin in 'λ *O*-type' initiators (Section C2.2.). P53-driven transcription is required for L replicon activity but the promoter can be replaced by e.g. the *lac* promoter. Schuster and co-workers could show that *repL* mRNA transcription is also subject to negative control by an 180 nt long antisense transcript [167]. The ϕP1 RepL protein has not yet been studied biochemically.

The ϕP1 *ban* (DnaB-analogue) gene encodes the cognate replicative helicase, an orthologue of *E. coli* DnaB (Section C3.3.) Because ϕP1 replication from *oriL* requires host DnaC but not DnaB, we have to assume that RepL directs a Ban_6DnaC_6 double-hexamer(s) to the unwound region, but experimental evidence is lacking. Phage proteins with detectable similarity to ϕP1 RepL are not known.

The 'initiator domain' of phage P4 alpha (α) protein and related helicases. Like P1 but for different reasons, *E. coli* phage P4 belongs to the group of replicons that withstand a straightforward classification: it is a natural occurring phage-plasmid hybrid. All protein components of the P4 virus capsid stem from *E. coli* phage ϕP2, and progeny can only be obtained from *E. coli* hosts lysogenic for ϕP2 – or from mixed infections. Therefore, ϕP4 is usually described as satellite phage of ϕP2, although the 11.6 kb-long dsDNA genome of ϕP4 is completely unrelated to the ϕP2 genome with respect to replication functions. Following infection of a host that is lysogenic for ϕP2, the circular ϕP4 genome can either enter the lytic cycle or integrate into the host chromosome as prophage, mediated by its cognate integration system. In a $\phi P2^-$ host, the ϕP4 prophage can either integrate or replicate as a free plasmid. Integration is preferred, however, and maintaining ϕP4 as plasmid under laboratory conditions requires continuous counter-selection. The extraordinary life-style of ϕP4 has been reviewed in detail by Dehò, Ghisotti, and colleagues [47].

Replication of ϕP4 in the lytic cycle or as free plasmid depends on the alpha (α) protein, host SSB, and DNA Pol III holoenzyme, but is independent of DnaA, DnaB, DnaC, and DnaG [253,22,43,233,103]. ϕP4 α is a multifunctional polypeptide that exhibits primase activity (Section C3.4.), helicase activity (Section C3.3.), and binds specifically to the ϕP4 replication origin, *oriI* (Section C2.2.). ϕP4 α binds to the decameric iterons GGTGAACAGA/T in *ori1* and *crr*, which both constitute the P4 *oriI* replication origin [123, 506]. ϕP4 replication proceeds bidirectionally from *ori1* in the θ-mode [234]. Because it is responsible for origin unwinding, ϕP4 α has the properties of an initiator and is discussed in this context.

The ϕP4 α C-terminus has DNA binding activity, and retains this activity upon isolation [505]. In addition, mutations were found in the α C-terminus that render the protein insensitive to negative regulation by the Cnr protein (see below). This suggested to Lanka and co-workers that the C- terminal DNA-binding and Cnr-interaction domains overlap [503]. Dehò, Ghisotti and co-workers could show by a yeast two-hybrid approach that α protein can oligomerise, and this interaction domain was also located in the α C-terminus [452]. The

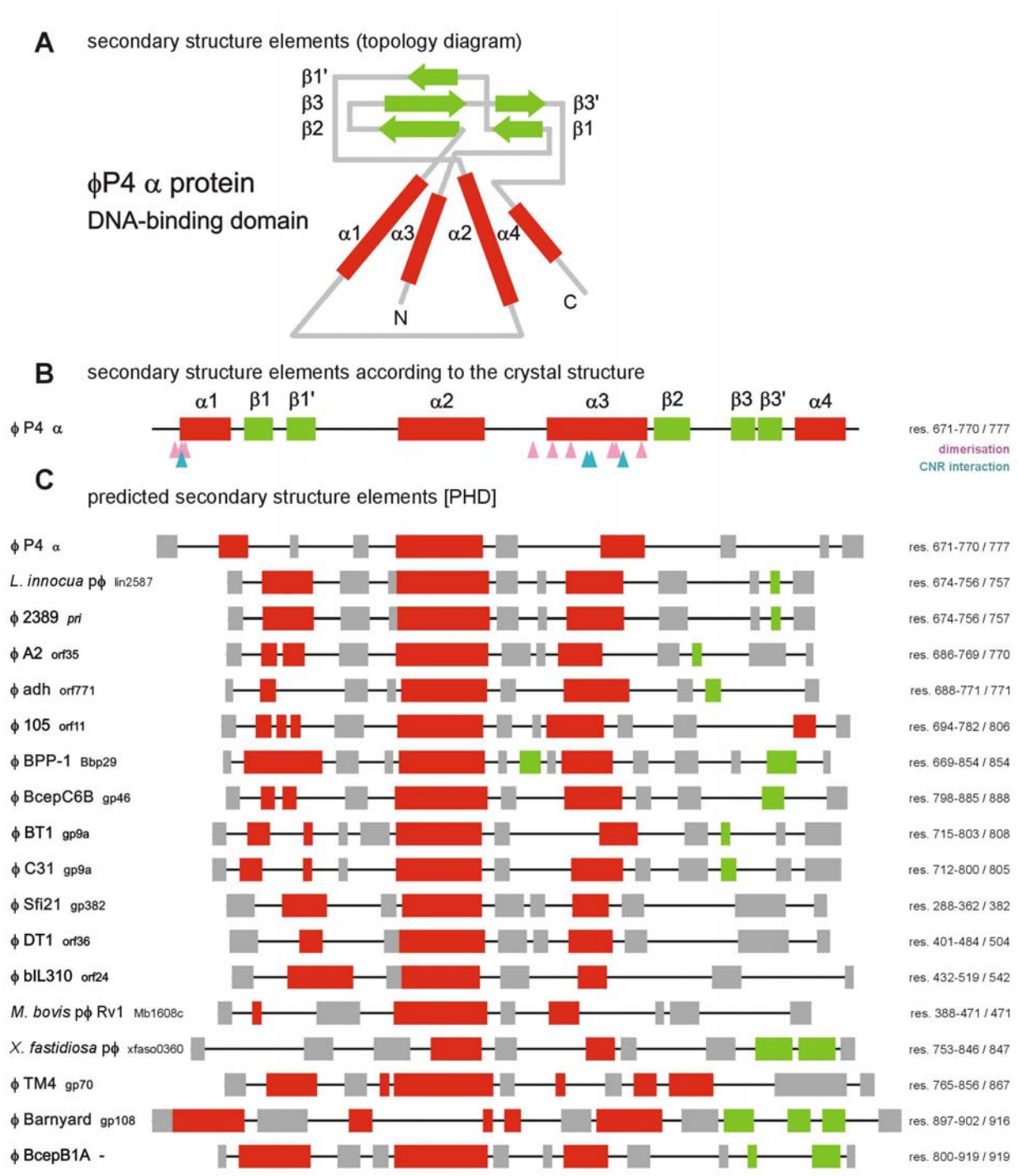

Fig. C10 (Putative) ϕP4 α-type initiator domains.
A topology diagram of the ϕP4 α DNA-binding domain adapted from Fig. 3B from the paper by Yeo *et al.* [492];. **B** linear projection of the secondary structure elements according to the crystal structure of the ϕP4 α DNA-binding domain [PDB 1KA8]; pink arrowheads point to residues involved in dimerisation, light blue arrowheads point to residues involved in the interaction with ϕP4 Cnr according to Yeo *et al.* [492]. **C** secondary structure prediction by the PHD method for the listed proteins. For accession numbers see Table C16 . Colour code: red = predicted α-helical region; green = predicted β-stranded region; grey = predicted loop; black line = unstructured/no prediction. The sequences were aligned along α-helix 2 of the ϕP4 α DNA-binding domain. Right column gives the regions of the proteins analysed by PHD and the length of the protein sequences.

three-dimensional structure of the ϕP4 α DNA- binding domain (or OBD, for origin-binding domain) is known at a resolution of 2.95 Å [PDB 1KA8]. Waksman and co-workers describe the α DNA-binding domain as double winged-helix with pseudo-twofold symmetry (see Fig. C10; A); they could pinpoint the residues responsible for dimerisation and for interaction with ϕP4 Cnr (see Fig. C10; B) [492].

In the lytic cycle of ϕP4, ~100 progeny copies accumulate in the host cell, in contrast to a copy number of ~40 for the plasmid state [2]. Therefore, mechanism(s) must operate that control ϕP4 plasmid copy number. One mechanism involves the phage encoded Cnr (copy number regulation) protein which down-regulates α activity through interaction (see above) [447,503]. It is presently not known exactly at which step inhibition of α by Cnr occurs: although it does apparently not inhibit DNA-binding of α, it may inhibit dimerisation of α bound to iterons in ori1 or crr. Alternatively, it may prevent the additional oligomerisation of α protomers bound to these two sites, which is necessary for origin unwinding. Modulation of the oligomerisation state of the initiator proteins is one out of several mechanisms for copy-number control operating in plasmid replicons [98]. The other known mechanism regulating ϕP4 copy number is a transcriptional network involving a highly complex and timely ordered interplay of activities of cognate transcription factors acting on several ϕP4 promoters (for details see Briani et al. [47]). The cnr gene is located immediately upstream of the α gene in the same transcription unit and the expression of both genes is co-regulated [96,350]. The available data suggest that ϕP4 replication strictly requires a balanced expression of Cnr and α [447].

BLAST searches with ϕP4 α as query led to the detection of numerous (pro)phage-encoded proteins with varying degrees of similarity, mostly confined to the helicase domain but in several cases also including the N-terminal primase domain (Sections C3.3. + C3.4.). In contrast, BLAST searches with the ϕP4 α DNA-binding domain (res. 671-770) led only to the detection of few orthologous prophage sequences (>80% ident. res.). In all proteins with similarity to the ϕP4 α helicase domain, the homologous region is located either at a position corresponding to that of the ϕP4 α helicase domain, or at the N-terminus, suggesting that – in the latter case – these proteins lack a primase domain. However, all proteins have a C-terminal domain of approximately 100 res. in length, similar in length but lacking detectable BLAST- similarity to the ϕP4 α DNA-binding domain. We speculated that this domain may be the DNA-binding domain of these proteins and performed a 'MultAlin' [84] analysis of a subset of these proteins. We obtained a reasonable alignment with gaps only introduced between β1+β1' and to both sides of β3', but no significant consensus sequence (not shown; for numbering of structural elements see Fig. C10; B). This subset of proteins was subjected to secondary structure prediction analysis by the PHD method [374]. The results were projected onto the secondary structure of ϕP4 α, which is known from the crystal structure (Fig. C10; C). Except for α-helix 2, the PHD method failed to predict the known secondary structure elements of ϕP4 α precisely: neither the β-stranded regions nor α-helix 4 were predicted, nor the length and positions of α-helices 1 and 3. However, there was no false prediction, and the 'loop' predictions are approximately correct. In spite of these limitations, the secondary structure prediction for ϕP4 α seems reliable enough to discuss the predictions for the other proteins in a meaningful way. Except for ϕBarnyard gp108, an equivalent of α-helix 2 is predicted at a corresponding position for all sequences. Also for α-helices 1 + 3, predictions could be obtained but less stringent than for α-helix 2. Except for ϕ105 orf11, no predictions for α-helices were obtained at the position of α-helix 4. The most surprising result was the prediction of β- stranded regions at positions roughly corresponding to the β-strands 2, 3 + 3'. Given the low similarity on the protein sequence level, the highly similar secondary structure prediction for all proteins indicates that they may be structural homologues. Together with the position C-terminal to a helicase domain and a rather uniform length, the results of the secondary structure predictions support the hypothesis that the C-terminal domains of the (putative) helicases represent 'initiator domains' alike the ϕP4 α DNA-binding domain. However, experimental results supporting this hypothesis are not yet available – and homology modelling on the basis of the known crystal structure of ϕP4 α beyond the scope of this review.

C3.2. Helicase loaders

Helicase loaders perform one or more of the following functions during primosome formation: i. they interact with the replicative helicase and promote its binding to unwound or forked DNA, ii. they interact with the DNA-bound pre-primosome, iii. they compete with SSB for binding to ssDNA, and iv. they link the replication of their cognate replicon to the host replication machinery in the case of numerous phage replicons. Genes encoding helicase loaders have so far not been described for genuine plasmid replicons.

A comparison of the primary sequence of several helicase loaders, or – more informative – of their secondary structure reveals that they are poorly related, if at all (Fig. C11). The similarity between the $DnaC_{Eco}$ and $DnaI_{Bsu}$ helicase loaders is entirely confined to the C-terminal nucleotide binding domain (Rossmann-fold

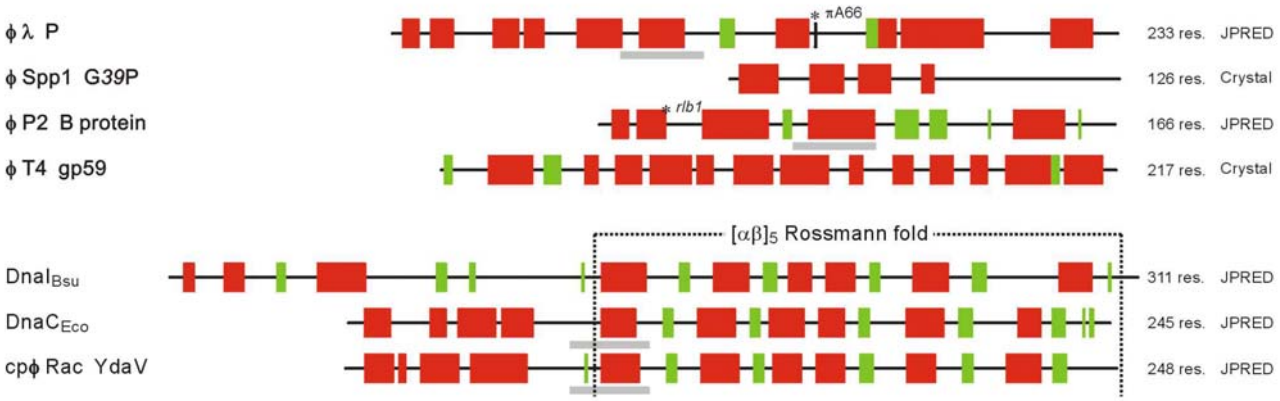

Fig. C11 (Predicted) secondary structure of helicase loaders.
For ϕT4 gp59 [PDB 1C1K] and ϕSPP1 G*39*P [PDB 1NO1] the secondary structure elements were taken from the crystal structures. Jpred secondary structure predictions were obtained for λ P [NP_040632], ϕSPP1 G*39*P [Q38151], ϕP2 B [P07696], ϕT4 gp59, $DnaI_{Bsu}$ [P06567], $DnaC_{Eco}$ [P07905], and cpϕ Rac YdaV [AAC74442] [88]. Colour code: red = predicted α-helical region; green = predicted β-stranded region; black line = unstructured. The length of the protein sequences and the sizes of the structural elements are shown to scale. The (approximate) positions of the *rIb1* mutation in ϕP2 B and the πA66 mutation in λ P are indicated by an asterisk. The dotted line indicates structural elements of the Rossmann fold [373].The grey bar underlining an α-helical region in λ P, $DnaC_{Eco}$, and cpϕ Rac YdaV show the positions of the hypothetical interaction sites with $DnaB_{Eco}$ (see text and Fig. C12 for details).

type) [373], as has already been noted previously [421]. In contrast, with 50% identity over their entire length $DnaC_{Eco}$ and the protein encoded by the *ydaV* (b1360) gene of the *E. coli* K12 Rac prophage are certainly orthologues [489], which is reflected in the similarity of their secondary structure predictions (see below).

The discussion focusses on the three phage-encoded helicase loaders that have been studied extensively: gp59 protein of phage T4 [495], P protein of phage λ , and G*39*P of *B. subtilis* phage SPP1 [341]. In these three model systems the participation of the cognate helicase loader in primosome formation is obligatory. B protein, the helicase loader of coliphage P2, seems to be required for lagging-strand synthesis [326]. $DnaC_{Eco}$ and $DnaI_{Bsu}$ are the helicase loaders for chromosome replication in *E. coli* and *B. subtilis*, respectively, and are discussed together with their phage-encoded homologues in detail at the end of the chapter.

Gene 59 protein (gp59) of phage T4. Gp59 is required for both types of ϕT4 replication initiation: origin-dependent replication (R-loops) early after infection and RDR (D-loops) later on [198]. Also, gp59 efficiently co-ordinates leading and lagging-strand synthesis by blocking replication until the gp41 helicase is loaded [198]. The crystal structure of gp59 protein has been determined, and the protein appears to be composed of two, largely α-helical domains [PDB 1C1K] [307]. Mueser and co-workers proposed a model that predicts N-terminal DNA-binding sites in gp59 as well as putative domains for the interactions with gp32 SSB and gp41 helicase, respectively [307]. It had been shown earlier that purified gp59 is a monomer in solution, and binds to both gp41 helicase and gp32 SSB [495]. Gp59 was also found to bind to single and double-stranded DNA, most tightly however to forked DNA substrates or DNA/RNA hybrids mimicking R-loops [495,198]. Gp59 binding to artificial substrates *in vitro* occurred without apparent sequence specificity, and binding sites for each fork arm required more than 6 but less than 12 single-stranded nucleotides for efficient helicase loading to the lagging-strand [495,199,198]. The N-terminus of gp59 interacts with the C-terminus of gp32 SSB [191]. Because gp32 SSB binds to ssDNA through its C-terminal DNA-binding domain it was proposed that the interaction with gp59 might weaken the association with ssDNA. Oligomerisation of gp59 requires its contact with gp32, and it was concluded from kinetic studies that gp59 oligomerises as a hexamer which forms stable complexes with the hexameric gp41 helicase in the presence of nucleotide [191]. In the gp59·gp41 double-hexamer, the six gp59 monomers are arranged in a head-to-head orientation, i.e. as trimers of dimers, according to Ishmael and co-workers [192]. The gp59 C-terminus interacts with the gp41·gp41 subunit interface, involving N- and C-terminus of gp41 [192]. Gp 41 helicase lacking 20 res. from the C-terminus does not form a complex with 59 protein on fork DNA [199]. Upon hexamerisation, a conformational switch of the gp41·gp59 complex is assumed to occur, which finally displaces gp32 from fork DNA [192].

Homologues of ϕT4 gene59 (gp59) are encoded by phages ϕT2 (99%; res. 26-217), ϕRB69 (85%; res. 1-217), ϕRB49 (58%; res. 1-216), ϕAeh1 (38%; res. 1-

126), and ϕKVP40 (33%; res. 13-216). BLAST searches could not identify other proteins in the data collections showing significant similarity with ϕT4 gp59. However, the identification of gp59 homologues in the genomes of phages ϕAeh1 and ϕKVP40 that infect γ-proteobacterial hosts other than *E. coli* shows that this type of helicase loader is not uniquely found in the T-even phages of *E. coli*.

Bacteriophage Lambda P protein. P protein is the prototype phage-encoded helicase loader that links the replication of its replicon to the replication machinery of the host: it interacts tightly with the initiator protein O [509] and also with the host replicative helicase $DnaB_{Eco}$, with which it forms a stable 3:6 complex [271]. Like its (functional) analogue $DnaC_{Eco}$, λ P exhibits ssDNA-binding activity in the complex with $DnaB_{Eco}$, which is further augmented by the interaction of the P·DnaB complex with oriλ-bound O protein, the so-called 'O-some' [106,244]. The interaction of P with $DnaB_{Eco}$ is significantly more stable *in vitro* than the DnaB·$DnaC_{Eco}$ interaction, which allows the re-direction of the host replicative helicase to the phage replicon [271]. The ATPase-, ssDNA-binding, and helicase properties of $DnaB_{Eco}$ remain suppressed in the P·DnaB complex and in the O·P·DnaB complex, and the action of chaperones is required for the release of the helicase from the complex as the last step of pre- primosome formation at oriλ [106,271].

Despite the wealth of data concerning the function of P during λ replication its molecular anatomy is poorly understood, and a crystal structure is not available. The ssDNA- binding domain within P has not yet been mapped. Wickner and Zahn could show that – in agreement with genetic observations – the O C-terminus interacts with P, but the O-interaction domain within P has not yet been determined (see above; [481]). There exist conflicting hypotheses concerning the interaction domain with $DnaB_{Eco}$. Ogawa and co-workers found that the N-termini of the homologous phage helicase loaders λ P and ϕ80 gene14 protein show a higher similarity in their N-termini, which suggested to them that this part of the proteins interacts with a host protein, most likely $DnaB_{Eco}$ [329]. On the other hand it was proposed that the P C-terminus interacts with $DnaB_{Eco}$ because the π(pi) mutations of the *P* gene map exclusively to the C-terminus of P and result in a lowered affinity of the mutant proteins for $DnaB_{Eco}$ [368, 224] (see Fig. C11).

A small stretch (~ 30 res.) of similarity exists between λ P, $DnaC_{Eco}$, and ϕP2 B, which we were unable to detect by BLAST analysis employing the complete protein sequences [326]. This region corresponds to α-helix 1 of the Rossmann fold of $DnaC_{Eco}$ with few neighbouring residues, and is also present in the ϕ80 gene14 protein, ϕ27 L18, and cpϕ Rac YdaV proteins (see Figs. C11 + C12). The identity between the λ P and ϕ80 gene 14 proteins raises significantly to 77% in this region, in contrast to 47% over-all identity, and 65% in the N-terminal 110 res. [329]. Roughly one third of the residues are conserved among several of the helicase loaders, but also in the *E. coli* DnaA protein. Interestingly, one of the two interaction domains of $DnaA_{Eco}$ with the $DnaB_{Eco}$ helicase has been mapped to this region [279,404]. Presently, experimental data in support of the hypothesis that this motif might be involved in an interaction of λ P-type or $DnaC_{Eco}$-type helicase loaders with the $DnaB_{Eco}$ helicase are not available.

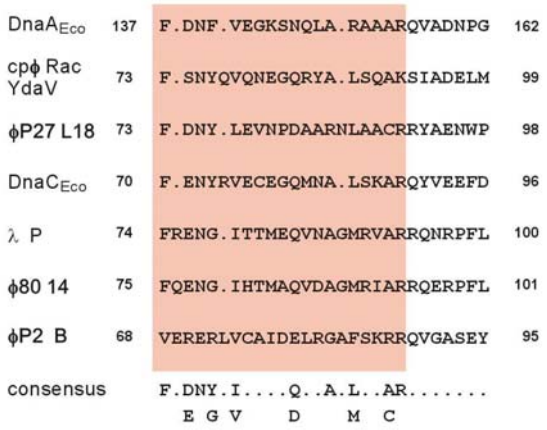

Fig. C12 Hypothetical interaction site of initiators and helicase loaders with $DnaB_{Eco}$.
Sequence alignement for $DnaA_{Eco}$ [P03004], pϕRac YdaV [AAC74442], ϕP27 L18 [NP_543070], $DnaC_{Eco}$ [P07905], λ P [NP_040632], ϕ80 14 [CAA31475], ϕP2 B [P07696]. The hypothetical site of interaction is shown as pink box, the regions in the individual sequences are indicated by residue numbering.

Full-length homologues of the λ P helicase loader are encoded in the genomes of phages ϕ21 (99% ident.), ϕH-19B (97% ident.), ϕ4795 (96% ident.), ϕ933W (96% ident.), and – as mentioned above – ϕ80 (47% identity with res. 1-201 of λ *P*). Various homologues of the λ *p* gene are also present in prophages within the genomes of several *E. coli*, *Salmonella* and *Shigella* strains, e.g. in *E. coli* O157:H7 EDL933 the genes z3355 of cpϕ CP-933V (96% ident. res.), and z1869 of cpϕ CP-933X (97% ident. res.). Two partial homologues probably represent truncated, non-functional genes: the hypothetical 77 residues long *ybcJ* gene product of *E. coli* K12 cpϕ DLP12 (96% identity with res.131-233 of λ P), and the hypothetical 103 residues long z0313 gene product of *E. coli* O157:H7 EDL933 cpϕ CP-933H. Three more prophage genes in *E. coli* O157:H7 EDL933 (z1774 of cpϕ CP-933N, z2094 of cpϕ CP-933O, and z3123 cpϕ CP-933U) show partial similarity with the λ P C-terminus but none with its N-terminus (~20% identity with res. 94-202 of λ P). This

observation points to the possibility that λ P-type helicase loaders are composed of two distinct domains, which might be found in various combinations also in yet unidentified (pro)phage-encoded helicase loaders.

Gene 39 protein (G*39*P) of *B. subtilis* phage SPP1. A crystal structure for the essential helicase loader G*39*P of ϕSPP1 phage has been obtained that shows a largely unstructured C-terminus of the protein [16] [PDB 1NO1]. G*39*P does not bind to DNA, but interacts with the initiator, G*38*P, and with the G*40*P helicase with the C-terminal 14 res. being essential for this interaction *in vitro* [13,17]. In the complex with G*40*P helicase and ATP, G*39*P inactivates the single-stranded DNA binding, ATPase and unwinding activities of G*40*P. G*38*P does by itself not interact with its cognate helicase in solution but with the G*39*P·G*40*P·ATP complex. When bound to the replication origin *ori*L, G*38*P interacts specifically with the G*39*P·G*40*P·ATP complex, which results in G*40*P binding to the origin. The dissociation of G*38*P·G*39*P heterodimers from the unstable intermediate complex is supposed to be required for subsequent helicase action [13]. No homologues of the ϕSPP1 G*39*P are presently found in the databases by BLAST searches.

B protein of phage P2. During RCR of ϕP2 *in vivo*, the displaced strand remains single-stranded in the absence of B protein [132]. It was therefore assumed that B protein is required to direct the host replicative helicase $DnaB_{Eco}$ to the displaced strand. Odegrip *et al.* could show that ϕP2 B protein interacts with $DnaB_{Eco}$ *in vitro*, [326]. They could show in addition that B protein competes with $DnaC_{Eco}$ for binding to $DnaB_{Eco}$, comparable to λ P. From the observation of partial suppression of an *E. coli dnaB*(ts) mutation by the *rlb1* mutation in the ϕP2 B gene it was speculated that the B N-terminus interacts with $DnaB_{Eco}$ (see Fig. C11). DNA-binding activity of ϕP2 B protein has not been shown, and its exact role during primosome formation for lagging-strand synthesis remains to be elucidated.

BLAST searches could identify full-length ϕP2 B homologues in phages L-413C (98%; res. 1-166) and Wϕ (98%; res. 1-166). Despite a comparable size, the limited similarity of ϕP2 B with gp77 of ϕRosebush (31%; res. 41-109) and with orf3 of ϕK139 (28%; res. 36-80) does not support a straight-forward qualification of these proteins as helicase loaders; this applies also to the *E. coli* K12 *ybgA* gene product (32%; res. 18-81).

***E. coli* DnaC and phage-encoded homologues.** DnaC was originally detected as host factor required for ϕX174 complementary strand synthesis [479,390]. Primosome formation on ϕX174 *in vitro* requires the *E. coli* primosomal proteins PriA, PriB, PriC, DnaT, the replicative helicase DnaB in complex with DnaC, and the DnaG primase, in addition to ϕX174 ssDNA and SSB [319]. The 'ϕX174-primo-some' was later renamed to 're-start primosome' when it became apparent that its formation is crucial for replication re-start and SDR (BRM Section 2.5.). Primosome formation at the *E. coli oriC* during initiation of chromosome replication *in vitro* requires DnaA, DnaB, DnaC, and DnaG (reviewed in [292]). DnaC binds and hydrolyses ATP and contains a nucleotide-binding domain of the Rossmann fold- type with the characteristic 'Walker A+B motifs' [373]. Together with DnaA proteins DnaC belongs to a superfamily of ATPases whose members are found throughout all phylogenetic kingdoms [225]. DnaC does not belong to the AAA+ family of chaperone-like ATPases [318] – as erroneously stated by Davey *et al.* [92] – because it lacks the AAA+-specific RFC-Box VII', RFC-Box VII", and Sensor II motifs [318].

Although probably evolutionary related, DnaC and DnaA are not functionally analogous proteins, and DnaC lacks the DNA-binding domain of DnaA [376]. However, both proteins use ATP-hydrolysis for a conformational switch, which dramatically alters their interaction properties with other proteins and DNA [424,92]. In the presence of ATP DnaC co-purifies with DnaB, and forms a stable double-hexameric complex [243]. The helicase action of DnaB requires the ATP-driven release of DnaC from the complex following its binding to DNA [131,467,468,120]. The double-hexameric $DnaB_6{\cdot}DnaC_6{\cdot}$ATP complex binds to ssDNA mediated by the cryptic DNA-binding property of DnaC [244]. Davey and co-workers could show that the ATP-form of DnaC in the $DnaB_6{\cdot}DnaC_6{\cdot}$ATP complex is required for its efficient binding to unwound *E. coli oriC* concomitant with an enlargement of the single stranded region, while ATP-hydrolysis is driven by the combined action of DnaB and ssDNA on DnaC [92]. Hydrolysis of bound ATP by DnaC is required to liberate the inactive helicase from the complex, in analogy to the action of chaperones required to separate DnaB from λ P during primosome formation at oriλ (see above).

By cryoelectron microscopy of $DnaB_6{\cdot}DnaC_6$ complexes and three-dimensional image reconstruction Bárcena, San Martin and co-workers obtained structures in which six DnaB monomers were assembled as three asymmetric dimers in a hexamer with threefold and sixfold symmetries on its both sides, respectively. The six DnaC protomers bound tightly to the sixfold face of the DnaB hexamer [382, 20]. Despite these efforts, the molecular anatomy of $DnaC_{Eco}$ is poorly understood, and a crystal structure is not available. The ssDNA-binding domain within DnaC has not yet been mapped. The *dnaC810* mutations was shown to confer to the mutant protein the property of helicase loading to SSB-covered ssDNA in the absence of the re-start primosomal pro-

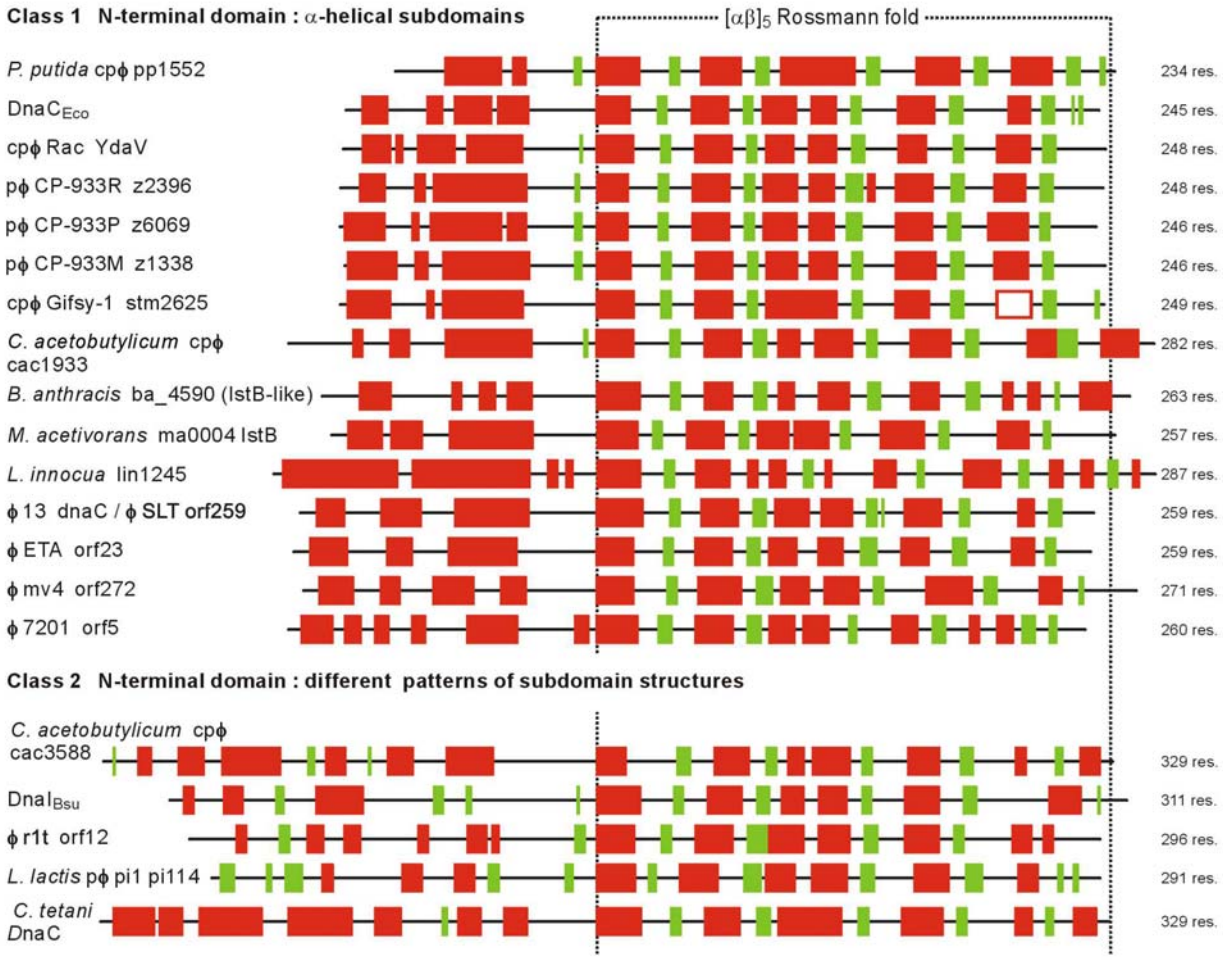

Fig. C13 (Putative) helicase loaders; part A.
Jpred secondary structure predictions [88] were obtained for *P. putida* cpϕ pp1552 [NP_743709], $DnaC_{Eco}$ [P07905], cpϕRac YdaV [AAC74442], pϕ CP-933R z2396 [NP_287823], pϕ CP-933P z6069 [NP_288003], pϕ CP-933M z1338 [NP_286860], cpϕ Gifsy-1 stm2625 [NP_461560], *C. acetobutylicum* cpϕ cac1933 [NP_348555], *B. anthracis* ba_4590 [NP_657950], *M. acetivorans* ma0004 [NP_614978], *L. innocua* pϕ lin1245 [NP_470582], ϕ13 dnaC [NP_803371] / ϕSLT orf259 [NP_075485],ϕETA orf23 [NP_510917], ϕmv4 orf272 [AAG31329], ϕ7201 orf5 [NP_038305], *C. acetobutylicum* cpϕ cac3588 [NP_350171], $DnaI_{Bsu}$ [P06567], ϕr1t orf12 [NP_695040], *L. lactis* pϕ pi1 pi114 [NP_266605], and *C. tetani* DnaC [NP_780838]. Colour code: red = predicted α-helical region; green = predicted β-stranded region; black line = unstructured. The length of the protein sequences and the sizes of the structural elements are shown to scale. The dotted line indicates structural elements of the Rossmann fold [373].

teins and may therefore have an increased affinity for DNA [385]. However, the E176G alteration in the *dnaC810* allele is located close to the Walker B-motif within the Rossmann fold and may therefore influence the reaction of the mutant protein with its bound nucleotide rather than affecting its DNA-binding property directly. Also $DnaA_{Eco}$ binds to ssDNA [423,474] but mediated by its C-terminal domain 4, which is not present in DnaC (F.Blaesing, and H.Seitz; unpublished). Unlike λ *P*, which interacts with the O initiator, $DnaC_{Eco}$ apparently does not interact with $DnaA_{Eco}$ in solution (F.Blaesing, and H.Seitz; unpublished); it is not known whether DnaC interacts with DnaA in the $DnaB_6$·$DnaC_6$·ATP double-hexameric complex. During re-start primosome formation, DnaC in the $DnaB_6$· $DnaC_6$·ATP complex interacts with the PriABC·DnaT pre-primosome, most likely with the DnaT moiety, but the DnaC subdomains involved in this interaction are not known [319]. Kaguni and co-workers presented evidence from mutational analyses that the N-terminus of $DnaC_{Eco}$ is responsible for the interaction with $DnaB_{Eco}$; All 6 mutations characterised in this study map to the three predicted N-terminal α-helices (see Fig. C13) [266].

Homologues of the *E. coli* K12 *dnaC* gene are present in the genomes of all sequenced *E. coli* strains (100% ident.), *Salmonella* (and sub-species) strains

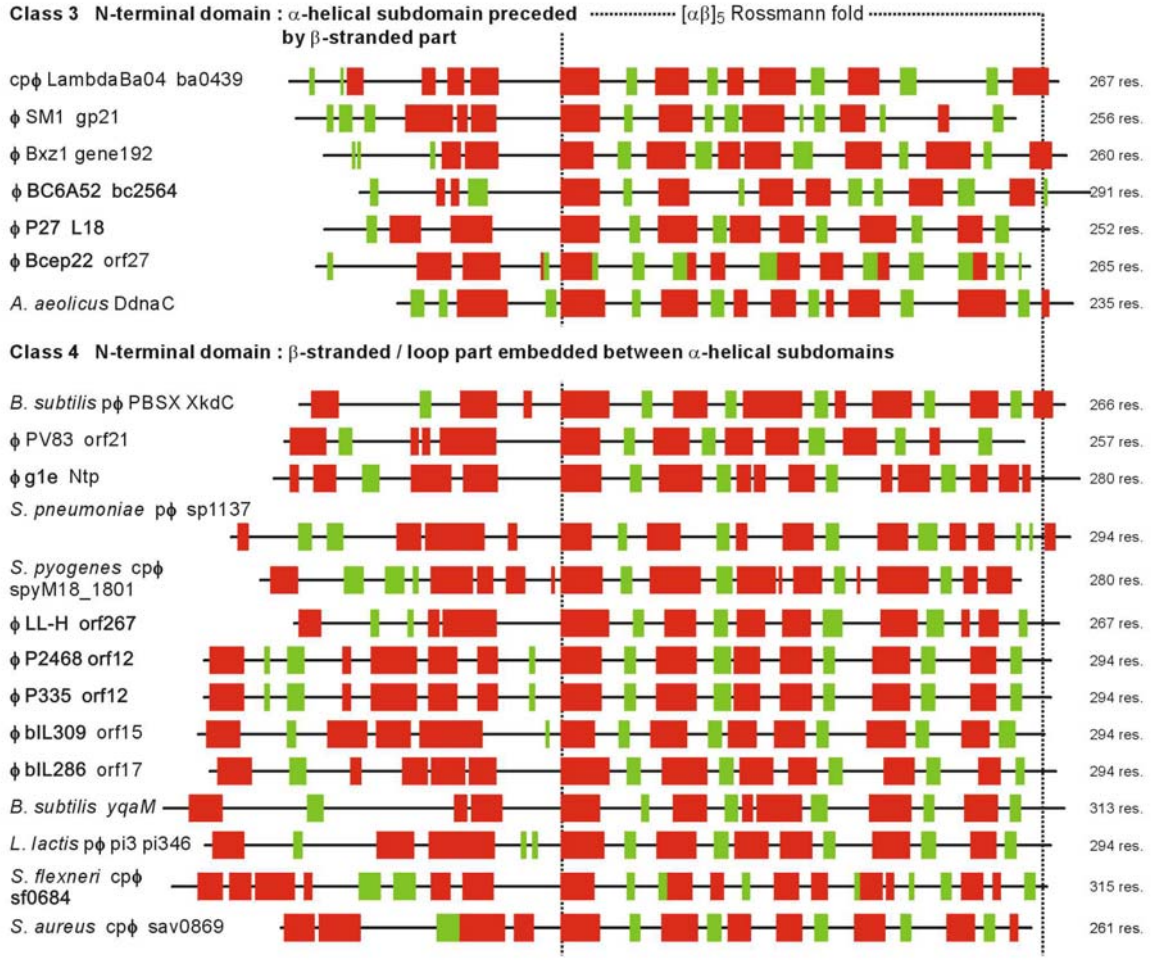

Fig. C14 (Putative) helicase loaders; part B.
Jpred secondary structure predictions [88] were obtained for *B. anthracis* cpϕ LambdaBa04 ba0439 [NP_842981], ϕSM1 gp21 [NP_862860], ϕBxz1 gene192 [NP_818243], ϕBC6A52 bc2564 [NP_852569], ϕ27 L18 [NP_543070], ϕBcep22 orf27 [NP_944257], *A. aeolicus* DnaC [NP_213618], *B. subtilis* pϕ PBSX XkdC [NP_389135], ϕPV83 orf21 [NP_061611], ϕg1e Ntp [NP_695183], *S. pneumoniae* cpϕ sp1137 [NP_345607], *S. pyogenes* cpϕ spyM18_1801 [NP_607823],ϕLL-H orf267 [AAL77547], ϕP335 orf12 [NP_839903], ϕbIL309 orf15 [NP_076710],ϕbIL286 orf17 [NP_076651], *B. subtilis* YqaM [P45910], *L. lactis* pϕ pi3 pi346 [NP_267575], *S. flexneri* cpϕ sf0684 [NP_706613], and *S. aureus* cpϕ sav0869 [NP_371393]. Colour code: red = predicted α-helical region; green = predicted β-stranded region; black line = unstructured. The length of the protein sequences and the sizes of the structural elements are shown to scale. The dotted line indicates structural elements of the Rossmann fold [373].

(~93% ident.), *Shigella flexneri* (100% ident.), *Klebsiella pneumoniae* (93% ident.), and *Buchnera* strains (65% ident.). *Y. pestis* is a close relative of *E. coli* with which it shares most relevant replication genes (*dnaA* 87% , *dnaB* 84%, *dnaG* 74%, *priA* 73%, *dnaE* 88% ident. res.) but a *dnaC* gene is lacking from the genomes of both sequenced *Y. pestis* strains [450]. It is likely, therefore, that the 'classical' helicase loader $DnaC_{Eco}$ is a very specific solution for helicase loading during primosome formation in few Enterobacteria. In fact, the distribution of *dnaC* genes among γ-proteobacteria seems to be even more restricted than the distribution of the *B. subtilis* *dnaI* gene among Bacillaceae (see also BRM Section 4.3.).

Annotations of 'DnaC-like' genes are found for various sequenced (pro)phage genomes in the databases. BLAST searches with $DnaC_{Eco}$ as query produce numerous hits (>250) due to the presence of the Rossmann fold-type nucleotide binding domain. We excluded all the hits produced by enterobacteriaceal DnaC proteins (discussed in the preceding paragraph) and by DnaA proteins from the further analysis. As mentioned above, *dnaA* and *dnaC* genes are structurally but not functionally related. Also we excluded – for simplicity – all

hits produced by proteins which are considerably smaller or larger than DnaC. For the majority of the remaining proteins, the similarity with $DnaC_{Eco}$ was confined to the C-terminal 2/3 of the protein (res. ~70-234), i.e. the Rossmann fold. We focussed our interest on those proteins that showed similarity with $DnaC_{Eco}$ also in the N-terminal portion of the protein. Among the proteins which showed identities ranging between >20% and ~40% were numerous (pro)phage-encoded 'putative replication' proteins, and – indistinguishable from the former – a significant number of *istB* genes encoding the small subunit of two-protein transposases. Because the BLAST-approach failed to detect 'DnaC-type' helicase loaders reliably we performed a JPred-based secondary structure prediction analysis for 40 of the proteins showing similarity with $DnaC_{Eco}$ [88,87] (Figs. C13 + C14). With few exceptions, the Jpred method could identify the $(\alpha\beta)_5$-structure of the Rossmann fold, which made up between 1/2 and 2/3 of the entire structure of the analysed proteins. Note that in almost all cases the third α-helix is predicted to be either longer than the others, or bi-partite – as was found in the known crystal structures of other AAA- or AAA+-type NTPases (see e.g.[154]). For the majority of the proteins analysed this C-terminal nucleotide-binding domain appears to be separated from the N-terminus by an unstructured stretch of ~25 res. in length. Tentatively, we consider the proteins as two-domain proteins with a flexible connecting linker. Although we find four formally distinct classes of N-terminal structures as depicted in Figs. C13 + C14 we cannot relate known or possible functions to any of these classes unequivocally. Within each class there are examples of structurally diverse but also structurally closely related proteins. In class 4, the best similarity is found between the proteins of ϕPV83 and ϕg1e. In class 3, the proteins of *Aquifex aeolicus* and ϕP27 are certainly structurally very similar. In class 1, the proteins of ϕ13, ϕSLT, ϕETA, and ϕmv4 are structurally more related among each other than with the small group of truly DnaC-like proteins of the *E. coli* Rac, Gifsy-1 (and Gifsy-2, not shown), CP-933M, CP-933P, and CP-933R prophages. Note that the predicted structures of two *istB*-like proteins closely resemble that of $DnaC_{Eco}$. We conclude therefore that neither BLAST searches nor secondary structure predictions allow for a clear discrimination between putative DnaC homologues and IstB proteins, for example. Thus, the identification of phage encoded helicase loader genes certainly requires additional criteria. We explore one such possible criterion – modular gene arrangement – in BRM Section 3. .

Neither the YdaV protein of the *E. coli* Rac prophage, nor the putative helicase loader of the *S. typhimurium* Gifsy-1 prophage were able to suppress temperature-sensitivity of an *E. coli dnaC28* mutant strain (B.Wróbel, G.Wegrzyn, H.Seitz, and C.Weigel, unpublished). On the other hand, autonomously replicating *oriJ*-plasmids [105,104] lacking the *ydaV* gene can be obtained but they are lost rapidly even with selection (R.Diáz O., H.Seitz, and C.Weigel; unpublished). This may indicate that the host DnaC protein can at least partially substitute for YdaV in the replication of the *oriJ* plasmid.

The helicase loading mechanism in *B. subtilis*. Genetic and biochemical analyses have shown that helicase loading during primosome formation follows different routes in *E. coli* and *B. subtilis* chromosome replication, for initiation at *oriC* as well as for replication re-start. Primosome formation in *B. subtilis* involves the $DnaB_{Bsu}$, $DnaD_{Bsu}$, and $DnaI_{Bsu}$ proteins, which have no obvious homologues in *E. coli*. $DnaB_{Bsu}$ and $DnaD_{Bsu}$ bind to DNA, and both proteins oligomerise [277, 53]. $DnaB_{Bsu}$ and $DnaD_{Bsu}$ have similar C-Termini (~100 res.), which is not detectable by BLAST analysis but becomes apparent by cross-wise comparison of DnaB and DnaI proteins from other Bacillaceae (BRM Section 3.1. 1.). The $DnaD_{Bsu}$·$DnaB_{Bsu}$ hetero-oligomer or $DnaD_{Bsu}$ alone interact with DNA-bound $PriA_{Bsu}$ *in vitro* [277], and also the interaction of $DnaD_{Bsu}$ with $DnaA_{Bsu}$ has been demonstrated by yeast two-hybrid analysis [190]. Marsin and co-workers propose a sequence of events for re-start primosome formation that might also be applicable for DnaA-primosome formation: DNA-bound $PriA_{Bsu}$ attracts $DnaD_{Bsu}$, which in turn interacts with $DnaB_{Bsu}$ [277]. This pre-primosome subsequently binds $DnaI_{Bsu}$ through a $DnaB_{Bsu}$-$DnaI_{Bsu}$ interaction [421,464]. $DnaB_{Bsu}$ and $DnaI_{Bsu}$ jointly recruit monomers of $DnaC_{Bsu}$ (the $DnaB_{Eco}$ orthologue) resulting in hexamerisation of the replicative helicase around single-stranded DNA [189,464, 52].

BLAST searches readily identify *dnaI*, *dnaD*, and *dnaB* genes among the members of the Bacillales and Lactobacillales sub-groups of the Firmicutes (Gram-positives). The detection of *bona fide* homologues of the *B. subtilis dnaI* gene in the genomes of other bacteria or in phage genomes is difficult by this approach due to conservation of the Rossmann fold in functionally analogous (*dnaC*), related (*istB*), or unrelated genes (*dnaA*). The orf12 gene product (296 rs.) of *L. lactis* ϕr1t may represent a true $DnaI_{Bsu}$ homologue because it shows 23% identity over a stretch of 270 res.. Other phage-encoded helicase loaders of the DnaI-type (similarity extending beyond the Rossmann fold towards the N-terminus) could not be detected by BLAST searches. Because ϕr1t does not encode genes similar to the *L. lactis dnaB* and *dnaD* genes the orf12 gene product might be used by this phage replicon to link its own replication to that of the host machinery (see Fig. C14). This hypothesis is supported by the observation that orf11 preceding

orf12 in the ϕr1t genome shows partial similarity with the gene38 initiator of *B. subtilis* ϕSPP1 (Section C3.1.2. and BRM Section 3.1.). In addition, the latter observation supports the notion that DnaC in *E. coli* and DnaI in *B. subtilis* perform analogous functions during helicase loading. *dnaD* genes and the *dnaB-dnaI* gene pairs are exclusively found in the genomes of the Bacillales and Lactobacillales sub-groups of the Firmicutes, and not present in the genomes of Clostridia and Mollicutes (analysis restricted to the fully sequenced genomes). This makes it very likely that the helicase loading mechanism operating in *B. subtilis* chromosome replication is specific for these two bacterial subfamilies. The partial similarity of several phage-encoded (putative) initiators with *dnaD* genes, and the partial similarity of $DnaB_{Bsu}$ with the ϕA118 gene49 (gp49) protein is discussed in BRM Section 3.1.1. .

C3.3. Helicases

Helicases separate the complementary strands of duplex nucleic acids in a processive manner, depending on NTP hydrolysis as energy supply. All known genomes of prokaryotes, eukaryotes, and also the genomes of many viruses and several plasmids encode helicases of several different types, underscoring the fundamental importance of helicases for all metabolic processes involving duplex nucleic acids (DNA, RNA, DNA-RNA hybrids): replication, recombination and repair, transcription, translation, and RNA splicing.

Helicases can be monofunctional proteins that interact with proteins involved in upstream and/or downstream pro??cesses, or alternatively domains of multifunctional proteins [466]. Helicases are usually classified with respect to the polarity of the strand to which they remain bound during duplex unwinding: 5'→3' or 3'→5' helicases. Helicases can function as monomers but many helicases – in particular replicative helicases – form oligomers, often homo-hexamers. The structure and function of hexameric helicases – including phage-encoded helicases – as well as present models for helicase movement have been extensively reviewed by Patel and Picha [339]. These authors also propose a very useful 'pathway' to guide studies of oligomeric helicases: i. the nature of the bound nucleotide dictates the oligomerisation state and the conformation required for DNA binding, ii. DNA binding stimulates the NTPase-activity, which in turn supplies the energy required for conformational changes that result in translocation along the bound DNA/RNA and duplex unwinding (see Fig. 1 in ref.[339]).

The replicative helicases ϕT7 gene 4A, ϕT4 gp41, and ϕP1 ban belong to the family 4 helicases with *E. coli* DnaB as the prototype, and will be discussed first together with their phage-encoded homologues. As we will show then, the importance of the helicase type represented by the ϕP4 α primase-helicase has been somewhat underestimated in the past. Next we will discuss the phage-encoded superfamily 2 helicases (SF2), which are encoded too frequently in phage genomes to be neglected. Their function as replicative helicases has yet to be demonstrated, but the ϕN15 RepA protein is the best candidate for such a role. Finally we shall briefly discuss phage-encoded superfamily 1 (SF1) helicases, represented by the Dda helicases of the ϕT4 group of phages. For a thorough discussion of the classification of the various types of helicases by their conserved structurally and functionally important motifs we refer to Hall and Matson [158].

Family 4 helicases (5'→3' helicases)

*Phage T7 gene 4 helicase.*The gene A4 primase-helicase protein is involved in the initiation of replication, in DNA synthesis, and in recombination of *E. coli* phage T7 [436,222, 220,221]. The ϕT7 replisome is formed by three phage-encoded proteins: gene 2.5 SSB, gene4A primase-helicase, and gene 5 DNA polymerase together with host thioredoxin [111,369].

Gene 4 of ϕT7 encodes two proteins: the full-length gene 4A or 63 kDa protein, and – starting from codon 64 – a shorter gene 4B or 56 kDa protein that lacks the N-terminal Zn- finger motif. The gene 4B protein is active as helicase but lacks primase activity [111,29]. The gene 4A protein is absolutely required for phage propagation but optimal rates of DNA synthesis require both translation products, point-ing to a physiological role of the smaller gene 4B protein [288]. Both gene 4 proteins form hexamers in the presence of nucleotide, with or without Mg2+ [338,323,347]. 50-fold higher protein concentrations were required for hexamer formation in the absence of nucleotide, while hexamers were already observed at a concentration of 0.2 μM with dTMP [338]. Furthermore, gene 4A hexamers were shown to be more compact with dTTP than with dTDP [496]. Analysis of ϕT7 gene4A protein by electron microscopy revealed a two-tiered ring with six-fold symmetry, and with one DNA strand passing through the central cavity [114,497]. At present, RepA helicase of plasmid RSF1010 and the ϕT7 gene 4 helicase domain are the only F4-type helicases for which a crystal structure is known [PDB 1Q57] [387,415]. Acidic residues in the C terminus of the ϕT7 helicase have been implicated in binding the ϕT7 DNA polymerase [315]. Deletion of the 17 C-terminal residues of the gene 4A protein abolishes this interaction, but the mutant protein retains primase and helicase activities [322].

Homologues of the ϕT7 gene4A helicase are found in all phages that also encode DNA polymerases with high

Table C11 Family 4 helicases: subfamily of ϕT7 gene 4A-type helicases.

F4 motifs:				**1**$_{F4}$	**1a**$_{F4}$	**2**$_{F4}$	**3***$_{F4}$	**4***$_{F4}$
consensus				xhxhhxARxxhGKS SG T	VLxhSLEM G MM E	hIhhDY L H I	LKAhAxoLxxPhxxh xQ RG T V C	DLR x SGxIxQxADxIh L S
ϕT7 gene 4 subfamily consensus				EhIhhxSGSGhGKS Q V A T T	VGxhxxEE D	xIVLDxh II	Lxxhxxxxxhxhhh I CH I V S	DhR G SGALxQhSxxhI K GI A L
replicon	accession	gene	res.					
ϕ T7	P03692	gene 4A	566	EVIMVT**SG**SGM**GKS**	**VG**LA**MLEE**	V**II**L**DHI**	**LKG**FA**K**S**T**GV**V**LVVI CH	**DLR** G **SG**AL**RQ**L**SD**T**II**
ϕ A1122	NP_848279	p17	566	EVIMVT**SG**SGM**GKS**	**VG**LA**MLEE**	V**II**L**DHI**	**LKG**FA**K**S**T**GV**V**LVVI CH	**DLR** G **SG**AL**RQ**L**SD**T**II**
ϕ T3	P20315	gene 4	566	EVVMVT**SG**SGM**GKS**	**VG**LA**MLEE**	V**I**VL**DHI**	**LKG**FA**K**S**T**GV**V**LVVI CH	**DLR** G **SG**AL**RQ**L**SD**T**II**
ϕ YeO3-12	NP_052087	gene 4A	566	EVIMVT**SG**SGM**GKS**	**VG**LA**MLEE**	V**I**VL**DHI**	**LKG**FA**K**S**T**GV**V**LVVI CH	**DLR** G **SG**AL**RQ**L**SD**T**II**
ϕ gh-1	NP_813761	p15	562	EVILVT**SG**SGS**GKS**	C**G**VA**MLEE**	V**I**VL**DHI**	I**KA**FA**K**TKNVAVFVI CH	**DLR** G **SG**G**LRQ**L**SD**T**II**
ϕ P60	NP_570328	P60_18	531	ELVTIT**AG**SGT**GKS**	**VG**YIA**LEE**	W**I**VL**DHL**	**LR**SFVEE**T**GIGMILI SH	QL**R** G **S**HS**I**A**Q**L**SD**LVI
ϕ SIO1	NP_064752	p15	522	ELITFV**AG**TGV**GKT**	**VG**TLF**LEE**	F**II**L**DHI**	**LK**TLTVE**L**DICLLMV SH	**DIR** G TAG**I**G**Q**L**S**NI**I**
ϕ PaP3	NP_775217	p40	569	EIVGVGG**G**TGI**GKT**	**VG**TFL**LEE**	TVIL**DNM**	LA**G**MA**D**EL**G**IRVFIF SH	QFT G **S**RAMQRWCQLMI
ϕ SP6	NP_853570	gp10	661	QLIIVG**AG**SGV**GKT**	**VG**II**S**T**ED**	N**III**D**NL**	IGTIKDRHPVTIFLV SH	**DFR** G **SG**AI**G**FW**A**SYAL
ϕ K1-5	AAL84819	gp9	662	QLIIVG**AG**SGV**GKT**	**VG**II**S**T**ED**	N**III**D**NL**	IGTIKDRHPVTIFLV SH	**DFR** G **SG**AI**G**FW**A**SYAL
ϕ Felix01	NP_944967	p188	663	EITTLA**A**PSSV**GKS**	I**G**VIPV**ED**	**II**IL**DPI**	**L**LRRCKRYQYAQVNV CH	**DI**K G **SG**AYF**QIS**MNNI

Motif numbering is according to Hall + Matson [158]. Shorter versions of the motifs 3 and 4 were used here, indicated by their re-naming as 3*$_{F4}$ and 4*$_{F4}$, respectively. Bold letters indicate matches with the conserved residues of the family 4 consensus sequences. Residues within the individual motifs that are specific for the subfamily are boxed.

similarity to ϕT7 gene 5 (Section C3.5.), and SSBs with high similarity to ϕT7 gene 2.5 (Section C3.6.1.). A common characteristic of the helicases of this subfamily is the replacement of the conserved glutamine residue in motif 3* by histidine (Table C11).

Gene 41 (gp41) helicase of phage T4. Bacteriophage T4 replicates its genome using a complex of seven phage-encoded proteins: DNApolymerase (gp43), three polymerase accessory proteins (gp44, gp62, and gp45), SSB (gp32), primase (gp61), and helicase (gp41). The 54 kDa gp41 helicase is essential for DNA replication and recombination [494,381], and loaded to DNA by the gp59 helicase loader (see previous chapter).

ϕT4 gp41 forms ring-shaped hexamers in the presence of nucleotide, preferably ATP, and bound nucleotide-triphosphate is required for hexamer formation [108]. The half-life for ϕT4 gp41 dissociation from DNA was measured to be around 11 min [399]. Deletion of ~20 res. from the gp41 C- terminus strongly reduced the *in vitro* interaction with the ϕT4 clamp loader and clamp proteins gp44, gp62, and gp45, respectively [371] (Section C3.5.2.). ϕT4 gp41 and gp61 are believed to form the primosome that catalyses coupled leading- and lagging-strand DNA synthesis. When assayed *in vitro*, the reciprocal stimulation of the helicase and primase activities by gp41 and gp61, respectively, could be shown, emphasising their cooperation [465,176,370].

E. coli DnaB and ϕT4 gp41 both belong to the F4-family of helicases, and posses highly similar biochemical properties. The sequence similarity between the two proteins is in the range of 25% (ident. res.) over the C-terminal nucleotide- and DNA-binding domains, but their N-terminal domains seem completely unrelated: there is no detectable primary sequence similarity, and also the secondary structure prediction for both sequences does not reveal any conservation of structural features (Jpred, not shown). Full-length homologues of the ϕT4 gp41 helicase sharing specific variations of the F4-helicase family signature motifs are found in all genomes of phages that also encode homologues of the ϕT4 gp59 helicase loader, of the ϕT4 gp43 DNA polymerase (Section C3.5.), of the DNA polymerase accessory proteins ϕT4 *rnh*, gp44,45,62 , and of the ϕT4gp32 SSB (Section C3.6.1.). The similarity values for the proteins of this F4- type-helicase range from 45% for *Aeromonas* sp. ϕ65 to 85% for ϕRB69 (Table C12).

Gp65 helicase of phage D29. Four mycobacterial phages encoded F4-type helicases show specific variants of the F4 family consensus motifs, therefore we chose to place them in a distinct subfamily (Table C13).A prominent feature of these helicases is their small size (~270 res.). The direct comparison with $DnaB_{Eco}$ reveals that they lack the entire N-terminal domain of $DnaB_{Eco}$, which was shown to be absolute re-

Table C12 Family 4 helicases: subfamily of ϕT4 gp41-type helicases.

Family 4 motifs:				1_{F4}	$1a_{F4}$	2_{F4}	3^*_{F4}	4^*_{F4}
consensus				x h x hhxARxxhGKS SG T	VLxhSLEM G MM E	hIhhDY L H I	LKAhAxoLxxPhxxhxQ RG T V C	DL Rx S Gx Ix Qx ADxIh L S
ϕT4 gp41 subfamily consensus				[T] L [N] hhhAGxNVGKS	VhYhSMEM	IVIVDYL VIM	hRGhAVxxxxVxWTAAQ	Dh [AE] S [AG] Lx [AT] ADFhL [D] [HG]
replicon	accession	gene	res.					
ϕ T4	NP_049654	gp41	475	[T] L [N] VLM**AG**VNV**GKS**	VLYISMEM	I**I**IV**DYL**	**LRA**LAVETETVLWTAA**Q**	**D**I [AE] **S** [AG] LP [AT] **AD**FML
ϕ RB69	NP_861732	gp41	458	[T] L [N] VLM**AG**VNV**GKS**	VLYISMEM	V**I**MV**DYL**	**LRA**LAVESETVLWTAA**Q**	**D**I [AE] **S** [AG] LP [AT] **AD**FML
ϕ RB49	NP_891593	gp41	470	[T] L [N] VLL**AG**VNV**GKS**	VLYISMEM	VVIV**DYL**	**LRG**FFVKWDVVGWTAA**Q**	**D**T [AE] **S** [AG] LP [AT] **AD**FML
ϕ KVP40	NP_899258	gene 41	466	[T] L [N] LLL**AG**SNV**GKS**	VLYISMEM	IVIV**DYL**	I**RG**FAVEHNVAVWSAA**Q**	**D**I [AE] **S** [AG] LA [HT] **AD**LIL
ϕ Aeh1	NP_943891	gp41	485	[T] L [N] LIL**AG**VNV**GKS**	VLYVSMEM	IVIV**DYL**	**LRG**FAVEHNVVVWSAA**Q**	**D**I [AE] **S** [AG] LA [AT] **AD**FIL
ϕ 44RR2.8t	NP_932387	gp41	469	[T] L [N] VIM**AG**VNV**GKS**	VVYFSMEM	IVIV**DYL**	**LRG**LAVQHQVVLWTGA**Q**	**D**V [AE] **S** [AG] LP [AT] **AD**FML
ϕ 65	AAR90925	gp41	520	[T] L [N] VIL**AG**TGV**GKS**	VLYVSFEM	I**I**MV**DYL**	V**RG**LAIEKKLVAWSGA**Q**	**D**I [AD] **S** [YA] LL [HG] C**D**FVL

Motif numbering is according to Hall + Matson [158]. Shorter versions of the motifs 3 and 4 were used here, indicated by their re-naming as 3^*_{F4} and 4^*_{F4} , respectively. Bold letters indicate matches with the conserved residues of the family 4 consensus sequences. Residues within the individual motifs that are specific for the subfamily are boxed.

Table C13 Family 4 helicases: subfamily of ϕD29 gp65-type (putative) helicases.

F4 motifs:				1_{F4}	$1a_{F4}$	2_{F4}	3^*_{F4}	4^*_{F4}
consensus				xhxhhxAR x xhGKS SG T	VLxh SL EM G MM E	hIhhD Y L H I	L KA hAxoLxx P hxxh xQ RG T V C	DLR xSGxIxQx A DxIh L S
ϕD29 gp65 subfamily consensus				QLhLhCAG [P] GTGKS A G	STRA [VR] EQ AA G	LIVhD [N] h	L [HE] xxRETGS [C] hhGL [HH]	[LSG] hKGALGRV [P] Ehhh
replicon	accession	gene	res.					
ϕ D29 [1])	NP_046882	gp65	[232]	QLVLVC**AG** [P] GT**GKS**	STRA [VR] **E**Q	L**I**VV**D** [N] **I**	**L** [HE] M**A**RE**T**GS [C] VIGL [HH]	[LGG] IK**GQI**GRV [P] EMVL
ϕ L5	NP_039729	L5p62	268	QLVLVC**AG** [P] GT**GKS**	ATRA [VR] **E**Q	L**I**VV**D** [N] **I**	**L** [HE] MGRE**T**GS [C] VVGL [HH]	[LSG] IK**GQI**GRV [P] EM**I**L
ϕ Bxz2	NP_817653	gp64	269	QLALVC**AG** [P] GT**GKS**	SARA [VR] **E**G	LVVI**D** [N] V	**L** [HD] M**A**RS**T**SA [C] VVGL [HH]	[LSG] VK**GQI**SRV [P] EM**I**L
ϕ Bxb1	NP_075324	gp57	269	QLALIA**A**A [P] GGA**KS**	SARA [VR] **E**G	L**I**VI**D** [N] **I**	**L** [HE] K**A**RE**T**GA [C] IIGL [HH]	[LSG] IK**GQI**GRV [P] ELVA
pRSF1010	NP_044309	RepA	279	TVGALV**S**P G GA**GKS**	**V**IYL PA **E**D	QAVA**D** G **L**	M E**A** I**A**ADT**G**C S IVFL HH	GVS KANYGAPF **A D**RWF

Motif numbering is according to Hall + Matson [158]. Shorter versions of the motifs 3 and 4 were used here, indicated by their re-naming as 3^*_{F4} and 4^*_{F4} , respectively. Bold letters indicate matches with the conserved residues of the family 4 consensus sequences. Residues within the individual motifs that are specific for the subfamily are boxed. 1) Tentative translation of the sequence upstream of the rare start codon UUG used to define the N-terminal Met residue of ϕD29 gp 65 of this NCBI entry results in a complete motif 1_{F4} (underlining) and full-length similarity in the N-terminus to e.g. the ϕL5 L5p62 protein. Hence the length of the protein is shown in brackets. The sequence of the RSF1010 RepA helicase is included to allow for an easy comparison.

quired for *in vitro* helicase actvity of the latter [37]. Small size is also a feature of the RepA helicase (F4-type) of plasmid RSF1010 with which the ϕD29 gp65-type helicases share no detectable sequence similarity, though (<20% ident. res.; see Table C13 for comparison). Purified $RepA_{RSF1010}$ binds nucleotide, shows GTP-dependent 5'→3' helicase activity in the presence of divalent cations, and forms a hexameric ring structure like $DnaB_{Eco}$ [392]. The crystal structure of $RepA_{RSF1010}$ protein is known [PDB 1G8Y] [320]. In contrast to $DnaB_{Eco}$, hexamer formation by RepA does not require bound nucleotide, and nucleotide binding is co-ordinated between two adjacent protomers in the hexamer, while $DnaB_{Eco}$ protomers bind nucleotide individually

Table C14 Family 4 helicases: subfamily of *E. coli* DnaB-type helicases.

F4 motifs:				**1**$_{F4}$	**1a**$_{F4}$	**2**$_{F4}$	**3***$_{F4}$	**4***$_{F4}$
consensus				xhxhhxAR x xhGKS SG T	VLxhSLEM G MM E	hIhhDYL HI	LKAhA x oLxxPhxxh x Q RG T V C	DLRxSGxIxQ x ADxIh L S
DnaB$_{Eco}$ consensus				xLhhhAAR P xMGKT GG V S	hhhhSLEM M	hIhhDYL H	LKxhA K ELxxPVhxL S Q R D V T	DLRESGxIEQ D ADhhh S DT C
replicon	accession	gene	res.					
ϕ P1	YP_006543	ban	454	DLIIVA**AR** P SM**GKT** **VL**VF**SLEM** M**I**M**IDYL** **LKA**LA K E**L**QV**P**VVAL S **Q** **DLRESG**A**IEQ** D **AD**L**I**M
ϕ D3	NP_061570	orf74	446	QMIVIAG**R** P AM**GKT** **VL**VI**SLEM** L**I**VI**DYL** M**K**LLA R EIGC**P**VLPL C **Q** **DLRESG**A**IEQ** D **AD**IVM
ϕ SPP1	NP_690733	gene 40	442	NFVLIA**AR** P SM**GKT** **V**NLH**SLEM** IVM**IDYL** **LK**KMA R E**L**DV**V**VIAL S **Q** **DLRESG**QLE**Q** D **ADII**E
ϕ HK97	NP_037740	gp55	458	DLVIIA**AR** P GM**GKT** **VL**IF**SMEM** L**I**MA**DYL** **LKA**MA K D**L**KT**P**VISL S **Q** **DLRDSG**S**IEQ** D **AD**S**I**I
ϕ P27	NP_543071	L19	463	DLVFIA**AR** P SM**GKT** **VL**LFT**MEM** LVVV**DYL** **LK**GLA K SGGF**P**LIAL S **Q** **DL**K**NSGE**I**EA** D **AD**I**I**L
ϕ Nil2	CAC95089	P	458	DLVIIA**AR** P GM**GKT** **VL**IF**SMEM** L**I**MA**DYL** **LKA**MA K D**L**RT**P**VFSL S **Q** **DLRDSG**S**IEQ** D **AD**S**I**I
ϕ P22 [1])	NP_059611	gene 12	458	DLVIIA**AR** P GM**GKT** **VL**IF**SMEM** L**I**MA**DYL** **LKA**MA K D**L**KT**P**VISL S **Q** **DLRDSG**S**IEQ** D **AD**S**I**I
ϕ ST104	YP_006384	gene 12	458	DLVIIA**AR** P GM**GKT** **VL**IF**SMEM** L**I**MA**DYL** **LKA**MA K D**L**KT**P**VISL S **Q** **DLRDSG**S**IEQ** D **AD**S**I**I
ϕ 3626	NP_612864	orf35	427	EFIVLG**AR** P SM**GKT** **VL**YIQ**L**D**M** V**II**V**DHI** **LK**SIA K E**L**NVAMVGL C **Q** **DLR**DT**GS**I**EE** D **AD**V**I**G
ϕ 11	NP_803269	16	413	QLIVIA**AR** P SV**GKT** TSFF**SLE**T V**I**FI**DYL** **LK**IIA N E**T**GAIIVLL S **Q** **D**MKE**SGG**I**EA** D **A**SLAM
ϕ CJW1	NP_817531	gp82	420	NLVVVGG**R** P GA**GKS** **SL**VF**SVEM** VVMV**DY**A **LK**VA**A** K **Q**L**HMV**VVLA S **Q** N**L**AGGDSLGR D **AD**VVL
ϕ Lc-Nu	AAG43560	orfA	416	RLLTIG**AR** P GV**GKS** **V**DMF**SLEM** LAIV**DYL** F**K**VLT N E**L**GI**P**IVLL S **Q** **DLRESG**S**IEQ** D **S**NAVG
ϕ MAV1	NP_047257	RepB	276	QLYIFAG**R** P AT**GKT** LVFF**SLEM** A**I**FI**DHL** **LK**KLA R K**L**NVNVFAI S **Q** **DLRESG**S**IEQ** D **AD**VVL
B. cereus pϕ	NP_830742	bc0956	431	DFVVLG**AR** P SM**GKT** **VG**LF**SLEM** L**I**IV**DYL** **LK**LLA R E**L**NVCVVAL S **Q** **DLR**ET**GQ**I**EQ** D **AD**V**I**M
ϕ 315.1	NP_664494	SpyM3_0690	446	QLIILA**AR** P AM**GKT** **V**AVF**SLEM** L**I**VI**DYL** **LK**ILA K E**L**KV**P**VIAL S **Q** **DLRESG**S**IEQ** D **AD**I**V**A
ϕ HK022 [2])	NP_037693	P	478	SLFVIG**AR** P KM**GKT** A**L**MF**SLEM** M**I**LV**DYL** **LK**NLA K E**L**DC**V**VVLL T **Q** **DSR**DT**GQ**I**EQ** D **C**D**Y**WV
ϕ Aaϕ23	NP_852741	p	485	SLVALG**AR** P KC**GKT** V**L**LF**SLEM** F**I**GI**DYL** **LK**NLA R EMDC**V**VLLL T **Q** **DSR**DT**GQ**I**EQ** E **C**D**Y**WL
E. coli	P03005	*dnaB*	471	DLIIVA**AR** P SM**GKT** **VL**IF**SLEM** L**I**MI**DYL** **LKA**LA K E**L**NV**P**VVAL S **Q** **DLRES**QS**IEQ** D **AD**L**I**M
B. subtilis	P37469	*dnaC*	454	DLIIVA**AR** P SV**GKT** **V**AIF**SLEM** M**I**LI**DYL** **LK**SIA R E**L**QV**P**VIAL S **Q** **DIRESG**S**IEQ** D **AD**IVA

Motif numbering is according to Hall + Matson [158]. Shorter versions of the motifs 3 and 4 were used here, indicated by their re-naming as 3*$_{F4}$ and 4*$_{F4}$, respectively. Bold letters indicate matches with the conserved residues of the family 4 consensus sequences. Residues within the individual motifs that are specific for the subfamily are boxed. 1) a virtually identical helicase (98% ident. res.) is encoded by phage ϕST104 (YP_006384). 2) virtually identical helicases (>98% ident. res.) are encoded by phages ϕSf6 (NP_958219), ϕVT2-Sa (NP_050526), ϕHK620 (NP_112058), and ϕST64T (NP_720303). The sequences of the *E. coli* DnaB and the homologous *B. subtilis* DnaC helicase are included to allow for an easy comparison.

[320]. It has been reported that RepA performs a C3–C6 conformational switch despite the lack of an equivalent of the DnaB$_{Eco}$ N- terminal domain (cited in [320]). Since none of the ϕD29 gp65-type helicases has been characterised structurally or biochemically so far, the results obtained with RepA$_{RSF1010}$ may provide a rough guideline for their understanding. Also in this subfamily the conserved glutamine residue in motif 3* is replaced by histidine (Table C11).

E. coli *DnaB and phage-encoded homologues*. The DnaB protein (52 kDa) is the replicative helicase of *E. coli*, and required for both initiation and elongation of DNA replication [246,228]. All completely sequenced bacterial genomes have highly similar *dnaB* genes, and in addition a number of plasmid and phage replicons. Homologues of DnaB are also encoded by several mitochindrial and chloroplast replicons. Although readily detected by sequence comparison few DnaB proteins have been characterised.

Biochemical methods including gel-filtration and cross-linking, and also studies by electron microscopy could establish that *E. coli* DnaB forms hexamers in solution [367,240, 241,8,54,383,498,382]. In the absence of Mg^{2+} ions and nucleotides DnaB formes mostly trimers, which can be converted to hexamers by the addition of Mg^{2+} or high concentration of salt [54]. When bound to nucleotide, DnaB oligomerises to hexamers with threefold symmetry (C3) and sixfold symmetry (C6) in addition to intermediates [498]. Also, the preferred conformation is greatly influenced by the pH of the sample buffer [107] .The significance of the

two symmetry forms and their interconversion for helicase function are not well understood. The C6 symmetry was found for the $DnaB_6C_6$ double-hexamer, and it may represent the inactive conformation because the ATPase activity of DnaB is inactivated in the $DnaB_6C_6$, complex [20].

The dimensions of the *E. coli* DnaB hexamer (Ø: 12.5–14 nm, cavity: 3–4 nm, heigth: 5.7 nm; taken from [339]) change in the presence of different nucleotides. Hexamers formed in the presence of ADP are somewhat more compact than those formed in the presence of AMP-PNP [498]. When bound to DNA, the DnaB hexamers appear even more compact [197].

As discussed in Section C3.1.2., the physical interaction between *E. coli* DnaA and DnaB was shown to involve two domains of either protein: res. 24-86 (domain 1) and 130-148 (domain 3) of DnaA and res. 154-210 (βγ-fragment) and 1-156 (α-fragment) of DnaB, respectively [278,404]. Kaguni and co-workers presented evidence from mutational analyses that the N-terminus of $DnaC_{Eco}$ is responsible for the interaction with $DnaB_{Eco}$, probably with a site within the γ-fragment [266] (Section C3.2.). From the observation that the similarity between the λ *P* helicase loader and its ϕ80 homologue, gene 14 protein, is considerably higher in the N-terminus than in the C-terminus, Ogawa and co-workers deduced that the C-termini of the helicase loaders could interact with their cognate initiators, and the conserved N-termini with a host protein, possibly DnaB [329]. These results support the hypothesis of Chattoraj and co-workers that the known helicase loaders contact DnaB via their N-termini [326]. In contrast however, λ π mutations with the phenotype of reduced affinity to DnaB map exclusively in the C-termini of mutant *P* proteins [224]. Certainly however, the interaction of DnaB with the different types of helicase loaders is even more complex: image reconstruction for electron micrographs of $DnaB_6C_6$ double-hexamers suggests multiple (putative) sites of mutual interactions, with one DnaC protomer contacting two neighbouring DnaB protomers of the hexamer [20].

E. coli DnaB can unwind duplex DNA *in vitro* without any accessory proteins but this activity is stimulated by DnaG primase [265]. The interaction between the *E. coli* DnaB and DnaG has always been difficult to analyse *in vitro* because of instability of the complex. Tougu and Marians could show that the DnaG C-terminus is responsible for interaction with DnaB [455]. By analysing mutant DnaB proteins in priming reactions, Chang and Marians found that residues in the DnaB N-terminus (α-fragment) are responsible for a productive DnaB-DnaG interaction [69]. This is at variance with the observation of Bastia and co-workers who found that a region partially overlapping with the interaction sites for plasmid initiators and DnaA in the N-terminus of the βγ-fragment is important [1]. A recent study by Soultanas and co-workers suggests that the flexible linker – connecting the DnaB α- and βγ-fragments – is involved in the interaction with DnaG in addition to several residues in the N-terminal α-fragment, and it seems possible that up to three DnaG monomers bind to one DnaB hexamer [449].

In the nomenclature of Hall and Matson [158], the signature motifs 1_{F4} and $1a_{F4}$ of Family 4-helicases are located in the nucleotide-binding β domain of $DnaB_{Eco}$ (res. 157-302), and represent the 'Walker A' and 'Walker B' motifs, respectively; motifs 2_{F4} to 4_{F4} are located in the $DnaB_{Eco}$ γ domain (res. 303-471), responsible for DNA binding [37] (See Tables C11 – C14, C16).

With 78% identical residues, the ϕP1 ban (DnaB analogue) protein shows extremely high overall similarity to *E. coli* DnaB, in the 3'-half of the gene also on the DNA level. This is in fact the highest similarity found so far between a (pro)phage-encoded protein and a host-encoded orthologue (Table C14). Not surprisingly therefore, ban greatly influenced the development of models for horizontal gene transfer (HGT) among prokaryotic replicons. Following its detection by genetic experiments [89,328], biochemical analysis established that ban and DnaB are orthologues with very similar properties, which form functional hetero-hexamers [242,400, 454]. Only recently it was shown that ban supports growth of an E. coli dnaB°(null) mutant strain albeit not at all physiological conditions [249]. The role of ban for ϕP1 replication is not clear: in the (plasmid) prophage state, ban is repressed, and suppresion of a host dnaB mutation was first observed in a ϕP1 mutant expressing ban constitutively [168]. Since the availability of E. coli DnaB is restricted to ~4–8 hexamers/cell under normal growth conditions it seems possible that de-repression of the gene encoding the cognate helicase allows for phage propagation under certain conditions – in analogy to the situation found for the ϕP1 SSB [26].

As discussed in Section C3.1.2 & C3.2., the initiation of replication of *B. subtilis* phage SPP1 requires the joint action of the phage-encoded G*38*P initiator, the G*39*P helicase loader, and the G*40*P helicase. G*40*P bound to the unwound *ori*L recruits the host DnaG primase, and both proteins together the host DNA polymerase III holoenzyme [341,297]. The DNA unwinding activity of G*40*P was shown *in vitro*, and – similar to ϕT4 gp41 and *E. coli* DnaB – the helicase-activity of G*40*P is stimulated by the DnaG primase, suggesting that the primase and helicase interact during DNA replication [11,13]. G*40*P was shown to form a ring-shaped hexamer with a two-tiered structure by electron microscopy [19]. Comparable to *E. coli* DnaB, two types of hexamers were observed, one with a C6- and the other with a C3-symmetry.

Table C15 Family 4 helicases: miscellaneous phage-encoded (putative) helicases.

F4 motifs:				1_{F4}	$1a_{F4}$	2_{F4}	3^*_{F4}	4^*_{F4}
consensus				xhxhhxARxxhGKS SG T	VLxhSLEM G MM E	hIhhDYL HI	LKAhAxoLxxPhxxhxQ RG T V C	DLRxSGxIxQxADxIh L S
replicon	accession	gene	res.					
ϕ Rosebush	NP_817815	gp54	920	GLSCIIGPPGV**GKS**	**V**LYMPG**E**G	EVGN**D**L**L**	FDQIRRH**T**NAGVLIVHH	TGPTGTVDPMQG**D**IVL
ϕ PG1	NP_943837	gp59	915	GLTSIIGPPGI**GKS**	**V**LYLPG**E**G	DLGN**D**LM	YDKLREL**T**GAGVCVVHH	TGPNGSVDPMQGEIVL
ϕ RM378	NP_835606	p019	416	ELGIVMLPSGW**GKS**	**V**IYFT**LE**L	VVLI**DY**A	**LR**LIAKVYNTAVWSAS**Q**	YIAD**S**FAKVVEI**D**FGM
pϕ SPBc2	NP_046682	*yorI*	504	KFYLRSSITGG**GKT**	STVI**S**T**EM**	YVYF**DYI**	ILLLMSDKLKGLCNKYD	Y**LR**G**S**KA**I**ADKT**D**AAM
ϕ T5	YP_006947	D6	502	ETLLLGGWRGT**GKS**	APYFSI**EM**	AKMS**D**FY	A**R**YGDKV**T**VALLDYIN**Q**	**DGR**TRMSKGILDSADM
ϕ VpV262	NP_640278	p17	408	DHVIIFGPVNA**GKS**	**V**LYVGN**E**D	ICIV**D**QA	**LR**MLYKRMKI**V**GVSVT**Q**	VFGS**R**REVAAQ**AD**VMI
ϕ Bxz1	NP_818244	gp193	418	ELAAIA**A**YTKV**GKS**	**V**M.FT**LE**M	FVLI**DQ**L	ETSRSSVGEL**P**TFLAI**Q**	NFAN**S**SM**IEQ**TV**D**IAL
ϕ K	YP_024501	orf71	480	EVGLII**A**PTGR**GKS**	**V**LYIA**LEE**	VVII**DY**P	I**R**RLSQQYGF**V**CWTLA**Q**	HVEG**S**RK**I**VNAVEVSL
ϕ KMV	NP_877454	orf15	397	ASVLVA**A**PPDA**GKT**	ILALDP**EE**	VVFW**D**MM	V**R**EMAVRHDFISFMTW**Q**	C**L**KD**S**KTAVQG**A**VDVQ
ϕ Xp10 [1])	NP_858988	xp10p41	261	DSIIWA**AR**PDQ**GKT**	HEQGV**LE**T	-	-	-
	NP_858987	xp10p40	139	-	-	ANAP**D**EV	V**R**DYASIEGYAAINTS**Q**	M**L**KD**S**KTAKAG**A**CEAI

Motif numbering is according to Hall + Matson [158]. Shorter versions of the motifs 3 and 4 were used here, indicated by their re-naming as 3^*_{F4} and 4^*_{F4} respectively. Bold letters indicate matches with the conserved residues of the family 4 consensus sequences. 1) The two orfs encoding the N- and C-terminus of the putative ϕXp10 helicase, respectively, are separated by an untranslated stretch of 82 bp; the comparison to the highly similar ϕKMV orf14 protein suggests that this gene split is not a sequencing artefact.

The function of the small RepB helicase (276 res.) for the replication of the mycobacterial phage ϕMAV1 (15.6 kb) is not known. However, RepBMAV1 adds another example to the group of smaller versions of DnaB-type helicases, but this one is more related to $DnaB_{Eco}$ than RepBRSF1010, or ϕD29 gp65 (see above, and Table C14).

Miscellaneous phage-encoded family 4 helicases. BLAST searches with various $DnaB_{Eco}$-type helicases as query yielded matches with several phage-encoded (putative) helicases that – considering the reasonable conservation of the signature motifs of the F4-type helicases – clearly belong to this family (Table C15). Because they lack the specific 'signature residues' of the DnaB-type subfamily within the motifs we placed them in a somewhat heterogeneous subfamily. The ~500 C-terminal residues of the ϕRosebush gp54 and ϕPG1 gp59 proteins contain the signature motifs and show similarity to the other helicases, the function of the extended N-termini of these proteins is not known (Section C3.4.).

Helicases with similarity to phage P4 alpha (α) primase-helicase. Replication of the circular ϕP4 prophage *in vitro* requires the phage-encoded α protein, host (*E. coli*) DNA PolIII holoenzyme, SSB, and gyrase; the host replication proteins DnaA, DnaB, DnaC, and DnaG are not required [103]. Thus, the initiator, helicase and primase functions are performed by the α protein (reviewed in [47]). ϕP4 Replication *in vitro* proceeds (mostly) bidirectionally from *ori*, and elongation occurs probably as cooupled leading- and lagging-strand synthesis, without a requirement for host DnaG primase and DnaB helicase [103]. Therefore, α is not only responsible for priming of leading-strand synthesis, but also for priming of lagging-strand synthesis. This suggests, in turn, an association of α with *E. coli* DNA PolIII holoenzyme. In short: α is the replicative helicase for ϕP4 replication. However, and unlike *E. coli* DnaB, ϕP4 α has been shown to unwind DNA in the 3'→5' direction *in vitro* (with respect to the polarity of the strand to which it binds), and its oligomeric status during helicase action is not known [506].

When expressed as isolated domain, the N-terminus of ϕP4 α could be shown to retain its primase function [505], discussed in the following chapter (Section C3.4.). Lanka and co-workers could show that the isolated C-terminus of α still binds specifically to DNA, but helicase activity could only be detected for constructs expressing the nucleotide-binding and the DNA-binding domains in a contiguous polypeptide [504]. This situation is reminiscent of RNA polymerases where the polymerising activity also cannot be uncoupled from the helicase activity [466]. The specific binding of α to the octamer repeats in the ϕP4 *ori* qualifies this protein as initiator, as discussed in Section C3.1.2. . We will refer

Table C16 (Putative) helicases of the ϕP4α-type.

ϕP4 α motifs:				**I** $_{P4}$	**II** $_{P4}$	**III** $_{P4}$	**IV** $_{P4}$	**V** $_{P4}$
consensus				hxxxN**G**hh**D** S N	EL x x x h**Q** x hh**GY**Sh W AA	hhhh**G**x**G**xN**GK**S D G	h**K**xhTGG**D**xh N VT **E**	Px.hRxxxxGhWR**R**hxhhP**F** K Y
replicon	accession	gene	res.					
ϕ P4 / ϕ R73	P10277	alpha	777	IGFRN**G**VL**D**	EKRDVILAALFMVL	LEVT**GP**G**G**S**GK**S	L**K**AITGG**D**AV	**P**MRFTDRSGGVSR**R**RVIIH**F**
ϕ BPP-1 [1])	NP_958698	Bbp29	854	LGVGN**G**VV**D**	DMIGFF**Q**RLI**GY**SL	AIPY**GSG**SN**GK**S	I**K**SMTGG**E**PL	**P**I.VKGDDHAIWR**R**LLPVP**F**
ϕ BcepC6B	AAT38405	gp46	888	LGVGN**G**AV**D**	EQVEFF**Q**RLV**G**YAL	IIPH**GTG**SN**GK**S	I**K**AMTGG**D**PI	**P**I.VKGDDHAIWR**R**LMLVP**F**
ϕ BT1	NP_813724	gp9a	808	LSFRN**G**VV**D**	ELAEYM**Q**RLT**GY**GI	AVLW**GKG**SN**GK**S	L**K**RVTGK**D**KV	**P**K.FKSQDEGLWR**R**VKLIP**F**
ϕ C31	NP_047955	gp9a	805	LSFAN**G**VV**D**	DLVGYMRRLV**GY**GI	AVLW**GKG**AN**GK**S	L**K**RVTGK**D**KV	**P**K.FKSQDEGLWR**R**VKLIP**F**
ϕ TM4	NP_569803	gp70	867	LNTPS**G**VV**N**	SMVEYV**Q**RLA**GY**AA	PFLF**GAG**SN**GK**S	V**K**LLTGG**D**VL	**P**D.VKAGGTSFFR**R**FRLIP**F**
ϕ BcepB1A	AAT37747	-	919	IGAPNCII**D**	ELIVFFKRLM**GY**AL	VIPY**GHG**AN**GK**S	V**K**SLTGD**D**TI	**P**V.IKGSDNGIWR**R**IMMIP**F**
X. fastidiosa pϕ	ZP_00040564	xfaso0360	847	LNCTN**G**TV**D**	PLSDFL**Q**RWF**GY**CA	AVMY**GMG**RN**GK**S	V**K**QATGG**D**SL	**P**V.IKGQDVGIWS**R**LMLIP**F**
ϕ Barnyard	NP_818646	gp108	916	LAVAN**G**VVE	SYRRDV**Q**IIL**G**HAL	IIFK**G**AANT**GK**S	M**K**TATGN**D**YI	**P**E.IEGHDKALRE**R**LRVIS**F**
ϕ A2	NP_680517	orf35	770	LNTIN**G**YV**D**	ELIDYL**Q**KAI**GY**SL	FILY**G**N**G**RN**GK**S	V**K**ELTGG**D**MV	**P**I.IRGTDDGIWR**R**LMLIP**F**
ϕ 2389 / ϕ PSA	NP_511033	*pri*	757	LNTQN**G**YI**N**	ELINYI**Q**KAV**GY**SL	FILF**G**N**G**RN**GK**S	V**K**QLTGG**D**KV	**P**I.IRGRDDGIWR**R**LHLVP**F**
ϕ adh	NP_050131	orf771	771	LNTPS**G**YI**D**	ELIHFV**Q**KII**GY**SL	FILY**G**N**G**RN**GK**S	V**K**QMTGG**D**TL	**P**K.IYGTDEGIWR**R**LVIIP**F**
L. innocua pϕ	NP_471917	lin2587	757	LNTQN**G**YI**N**	ELINYM**Q**KAV**GY**SL	FILF**G**N**G**RN**GK**S	V**K**QLTGG**D**KV	**P**I.IRGRDDGIWR**R**LHLVP**F**
ϕ 105	NP_690795	orf11	806	FNCEN**G**VI**D**	EIIEFL**Q**KAI**GY**SL	FFLF**G**N**G**RN**GK**S	V**K**QITGG**E**KM	**P**I.VKGSDEGIWR**R**IRLVP**F**
ϕ DT1 / ϕ κ3	NP_049424	orf36	504	L.VKN**G**IY**D**	ELVELL**W**QVI**AA**SL	IWLV**G**N**G**ND**GK**G	F**N**SVVTG**E**PV	**P**V.FKNKSNGTYR**R**IVIIP**F**
ϕ Sfi11 [2])	NP_056712	orf504	504	L.VKN**G**IY**D**	ELVELL**W**QVI**AA**SL	IWLV**G**N**G**ND**GK**G	F**N**SVVTG**E**PV	**P**V.FKNKSNGTYR**R**IVIIP**F**
ϕ Sfi19	NP_049955	orf508	506	L.VKN**G**IY**D**	ELVELL**W**QVI**AA**SL	IWLV**G**N**G**ND**GK**G	F**N**SVVTG**E**PV	**P**V.FKNKSNGTYR**R**IVIIP**F**
ϕ Sfi21	NP_050002	orf382	382	L.VKN**G**IY**D**	ELVELL**W**QVI**AA**SL	IWLV**G**N**G**ND**GK**G	F**N**SVVTG**E**PV	**P**V.FKNKSNGTYR**R**IVIIP**F**
ϕ bIL310	NP_076775	orf24	542	IPVAN**G**IF**N**	ELVSLL**W**QIIS**A**ST	VWLV**GKG**ND**GK**G	YFSVVTG**D**PV	**P**K.FRNKSNGTYR**R**LLIVP**F**
ϕ 31	CAC04163	primase	492	EVESTASR**D**	ELVKLL**W**QVIS**A**SL	IWFV**GEG**ND**GK**G	F**N**SVVTG**E**PV	**P**K.VRNKTNGTYR**R**FLIIP**F**
M. bovis pϕ Rv1	NP_855261	Mb1608c	471	LNVAN**G**TL**D**	GVRGFV**Q**RLA**G**VGL	AILI**G**V**G**AN**GK**S	I**K**RLTGG**D**TI	**P**R.VPGDDTAIWR**R**IRVVP**F**

Motifs I - V were detected by BLAST alignement of the sequences as the regions with the highest degree of conservation. Bold typing indicates particularly well conserved residues. 1) identical sequences are present in *Bordetella* sp. phages ϕBMP-1, and ϕBIP-1. 2) identical sequences are present in *S. thermophilus* phages ϕsfi18, and ϕO1205.

to the central nucleotide-binding domain as 'helicase domain' in the following, and to the C-terminus as 'DNA-binding domain'.

BLAST searches with ϕP4 α as query (entire sequence or res. 390-777) led to the detection of >20 phage-encoded proteins with significant similarity, and many more (putative) prophage sequences in the genomes of bacteria from different phyla. Within the helicase domain (ϕP4 α: res. 390-670) we could detect five particularly well conserved short stretches, which we propose as signature motifs I_{P4}–V_{P4} although it is presently not possible to correlate them – except for motif III_{P4} – with known structural or functional elements (Table C16). Mutations of G506 or K507 in motif III_{P4} were found to result in a reduction or complete loss of nucleotide hydrolysis proficiency and helicase activity of α protein [504]. Motif III_{P4} 'hhhhGxGxNGKS' corresponds to the 'A' (Walker A') motif 'hhhxGPxGTGKS' proposed earlier by Koonin and co-workers for a group of helicases that are now placed into the helicase superfamily 1 (SF1) [148,158]. Although a rather formal approach, it should be noted that among the variants of the 'Walker A' or 'P-loop' motifs found in nucleotide-binding proteins – e.g. the AAA+- family – the specific positioning of the glycine residues is highly conserved within the subfamilies. We tend to place the ϕP4 α-type helicases in a separate family rather then grouping them into SF1, therefore. The signature motifs 'B' and 'C' suggested by Koonin and co-workers were also detected but considered less useful due to their degeneracy (see Figs. C15 + C16) [148]. The choice of signature motifs I_{P4} – V_{P4} is supported by comparing the secondary structure predictions for three ϕP4 α-type helicases that showed <25% identity to each other: all motifs are positioned within a fairly similar (predicted) secondary structure context (Fig. C16).

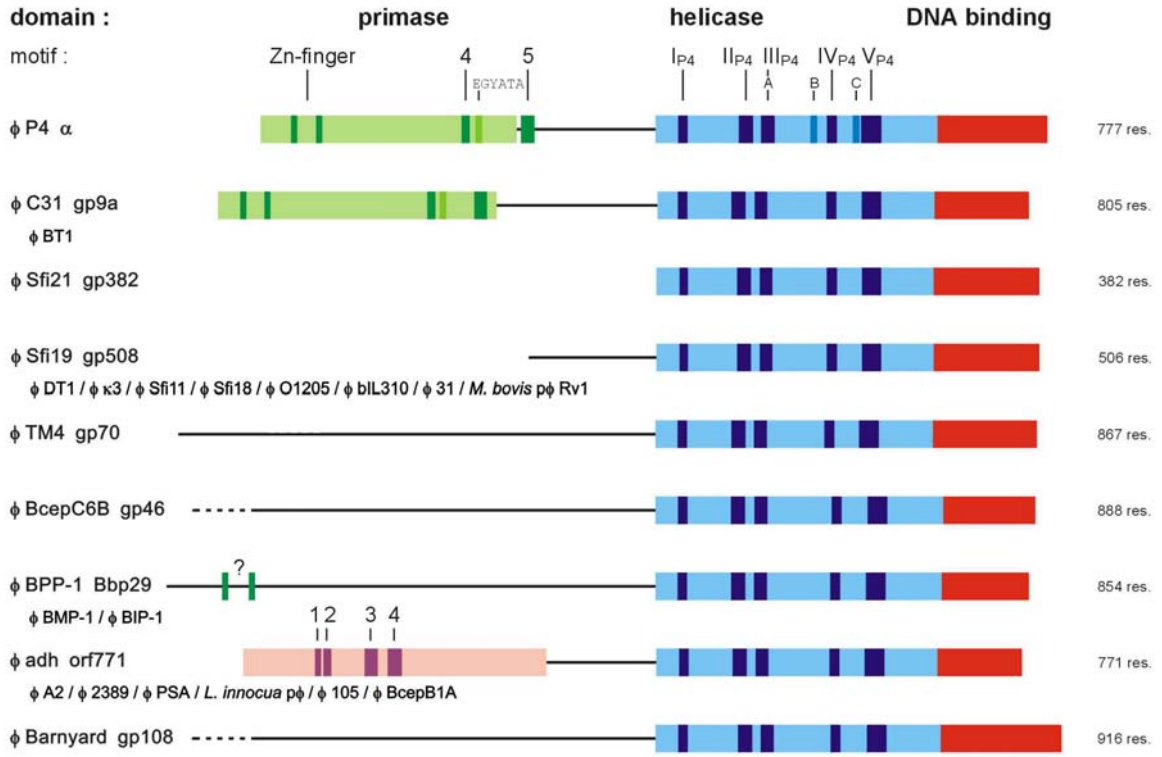

Fig. C15 Anatomy of ϕP4 α-type (putative) helicases.
All (putative) helicase sequences, and the size/position of the signature motifs are drawn to scale. For accession numbers and signature motif numbering see Table C16 for the helicase motifs, and Tables C23 + C25 for the primase motifs; the A, B + C motifs proposed by Gorbalenya *et al.* for ϕP4 α are indicated at their respective positions for this protein only [148]. In addition, the 'EGYATA' motif conserved in some primases is indicated (see text for details [431, 335]). Line length indicates the size of the protein (dotted line: protein sequence longer than shown), domains are shown as coloured blocks. Colour code: red = DNA-binding domain of ϕP4 α [PDB 1KA8] and homologous regions detected by BLAST comparison + secondary structure prediction Jpred (not shown; [88]); blue = helicase domain, the ϕP4 α-type helicase signature motifs are shown in dark blue (size of the domain tentatively adjusted to the smallest detected helicase domain of ϕSfi21 gp382); green = primase domain, the ϕP4 α-type primase signature motifs are shown in dark green (domain size in ϕP4 α is shown according to the minimal primase domain determined by Ziegelin *et al.* [505]); pink = conserved domain of unknown function, the signature motifs are shown in dark pink. Phages encoding (putative) helicases with identical or highly similar anatomy are indicated below the prototype sequence by smaller font size. '?' weak similarity to Zn-finger motif.

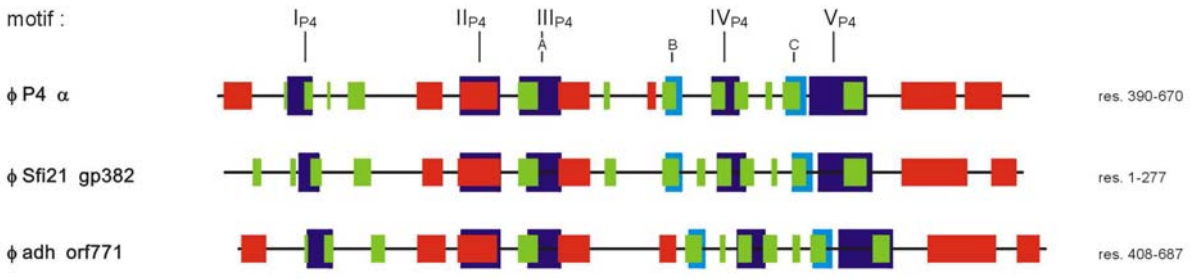

Fig. C16 Secondary structure prediction for the ϕP4 α helicase domain.
The (putative) helicase domain sequences, and the size/position of the signature motifs are drawn to scale. For signature motif numbering see Table C16 . The A, B + C motifs proposed by Gorbalenya *et al.* for ϕP4 α are indicated at their respective positions [148]. Secondary structure predictions were obtained from the Jpred server [88]. Colour code: red = predicted α-helical region; green = predicted β-stranded region. Light blue = A, B + C motifs according to Gorbalenya *et al.* ; dark blue = signature motifs proposed in this study (see Text for details).

Table C17 Superfamily II (SF2) helicases: ϕT4 UvsW-type helicases.

SF2 motifs:				**I** SF2	**Ia** SF2	**II** SF2	**III** SF2	**VI** SF2
consensus				hhxxxoGxGKT S	xhhhxPoo	hhhDExH	hxhSATxxx TGS	QxxGRxxR
replicon	accession	gene	res.					
ϕ T4	NP_049796	UvsW	587	ILNLPT[S]A**GKS**	ILIIV**P**TT	MMN**DEC**H	FGL**SGS**LRD	**Q**TI**GR**VL**R**
ϕ RB69	NP_861886	UvsW	504	ILNLPT[S]A**GKS**	ILIIV**P**TT	MMN**DEC**H	FGL**SGS**LRD	**Q**TV**GR**VL**R**
ϕ RB49	NP_891741	UvsW	500	MLVLPT[S]A**GKS**	VLIIV**P**TT	VIV**DEC**H	IGM**TGS**PRD	**Q**SI**GR**AL**R**
ϕ 44RR2.8t	NP_932529	UvsW	493	ILNLPT[S]A**GKS**	VLVIV**P**TT	VMN**DEC**H	YGL**SGS**LRD	**Q**TI**GR**VL**R**
ϕ KVP40	NP_899623	UvsW	507	LLNLPT[S]A**GKS**	VLIIV**P**TT	LLV**DEC**H	IGL**TGS**LRD	**Q**SI**GR**IL**R**
ϕ Aeh1	NP_944130	UvsW	503	IGNLPT[S]A**GKS**	VLVLV**P**TT	IIV**DEC**H	IGM**SGS**LRD	**Q**SI**GR**VL**R**

Motif numbering is according to Hall + Matson [158]. Motifs IV and V were omitted due to their high degree of degeneration. Bold letters indicate matches with the conserved residues of the SF2 consensus sequences. Residues within the individual motifs that are specific for the subfamily are boxed.

The proteins with similarity to ϕP4 α show a considerable variation in size indicating that they are probably also multi-domain proteins (see Table C16). A tentative anatomy of these proteins is shown in Figure C15, using ϕP4 α as 'template'. The length of the helicase domain is very similar in all proteins (~270 res.), and the signature motifs are evenly spaced. ϕSfi21 gp382 was the smallest protein detected in the databases with full-length similarity to ϕP4 α and might therefore represent the 'minimal version' of this type of helicases. All proteins shown in Figure C15 have a C-terminal domain with significant similarity in length and sequence to the ϕP4 α DNA-binding domain. ϕSfi21 gp382 and ϕSfi19 gp508 lack any N-terminal extension and are unlikely to function as 'primase-helicase' proteins. In analogy to the function of α for ϕP4 replication, we speculate that these proteins function as 'initiator-helicases' (Section C3.1.2.). ϕC31 gp9a is a full-length homologue to ϕP4 α, containing a clearly detectable N-terminal primase domain (see next chapter). The function of the long N-terminal extensions – responsible for the size variation – of the remaining proteins is not known, but they might represent yet unknown primase domains (see next chapter).

In the past, the α protein of satellite phage P4 has been studied mainly as a rare example of a multifunctional replication protein, comparable to the large T antigen of simian virus 40 (SV40). Here we can show that α-type helicases are found as widespread among phage replicons as the $DnaB_{Eco}$- type helicases. In contrast to DnaB however, homologues of ϕP4 α are not found in the sequenced bacterial genomes, except for the many α-type proteins encoded by (putative) prophages. It should be interesting to learn why chromosomal replicons rely on DnaB as their replicative helicase, and (with few exceptions) on DnaA as replication initiator, for which a phage-encoded homologue has not yet been found – contrary to the expectation of Campbell and Botstein [62].

Phage-encoded superfamily 2 helicases. Next to genes encoding F4-type helicases, genes encoding superfamily 2 helicases (SF2) are found most frequently in phage genomes. Clear evidence for the participation of SF2-type helicases in the initiation and elongation steps of replication has not yet been presented, but the SF2-type helicase UvsW of ϕT4 may participate in the later (RDR) stages of ϕT4 replication. SF2- type helicases homologous to ϕT4 UvsW are present in the other completely sequenced genomes of the ϕT4 group phages (Table C17), and could be detected in further >30 (pro)phage genomes (Table C18). The listing in Table C18 is complete for phage-encoded SF2 helicases, but numerous (putative) prophage sequences detected in the fully sequenced genomes of species from all phyla have been omitted. We note that for several phage genome entries in the NCBI database the SF2-type genes were annotated 'putative helicase', and the genes showing similarity to ϕP4-α annotated 'putative primase' although the region of similarity did not include the N-terminal primase domain of ϕP4α (e.g. ϕSfi21, ϕSfi11) [102]. This imprecision calls for more expert assistance during gene function annotation procedures.

In addition to the SF2-type helicases with a size of ~450-650 res., four large phage-encoded proteins (~1300 res.) with significant homology to the RepA protein of ϕN15 were detected, which show partial similarity to the ϕP4 α primase- helicase in their N-terminal (putative) primase domain, but have a central nucleotide binding domain that – as judged from the signature motifs – suggests their grouping in the superfamily 2 (Table C18). Very recent studies suggest that these pro-

Table C18 Phage-encoded superfamily II (SF2) helicases.

SF2 motifs:				I_{SF2}	Ia_{SF2}	II_{SF2}	III_{SF2}	VI_{SF2}
consensus				hhxxxoGxGKT S	xhhhxPoo	hhhDExH	hxhSATxxx TGS	QxxGRxxR
replicon	accession	gene	res.					
ϕ K	AAO47518	orf69	582	IAHLATNG**GKT**	VAFFTGST	MIV**DEAH**	IAL**TGS**IDK	**Q**RI**GR**AL**R**
ϕ Bcep1	NP_944368	gp60	614	IAAIVT**GGKS**	VMSLA**P**SM	VLV**DEAH**	VGM**TAT**DFI	**Q**IV**GR**GL**R**
ϕ Bcep781	NP_705681	gp58	614	IAAIVT**GGKS**	VMSLA**P**SM	VLV**DEAH**	VGM**TAT**DFI	**Q**IV**GR**GL**R**
ϕ Bcep43	NP_958163	gp57	614	IAAIVT**GGKS**	VMSLA**P**SM	VLV**DEAH**	VGM**TAT**DFI	**Q**IV**GR**GL**R**
ϕ T5	P11107	D10	450	IINGKP**GFGKT**	TLVICTNT	VIV**DEVH**	IGL**SGT**LKR	**Q**LA**GR**VQ**R**
ϕ PG1	NP_943831	gp53	559	GVVLPT**GTGKS**	VLMVAHRG	ILW**DEVH**	AGF**TAT**MRR	**Q**MI**GR**AL**R**
ϕ Rosebush	NP_817811	gp50	571	GVVLPT**GAGKS**	VVALAHRA	ILW**DEFH**	CGF**TAT**MYR	**Q**MV**GR**AL**R**
ϕ VP16T	AAQ96511	orf44T	550	LAVLPT**GAGKT**	HRVIA**P**AK	SVY**DEGH**	LFV**TAT**PER	**Q**MV**GR**AL**R**
ϕ VP16C	AAQ96576	orf45C	718	LAVLPT**GAGKT**	HRIIA**P**AK	GVY**DEGH**	LFV**TAT**PER	**Q**MV**GR**AL**R**
ϕ BcepB1A	AAT37757	-	614	MAVSAT**GSGKT**	ARAGI**P**HN	LIG**DEGH**	VLF**TAT**PER	**Q**QW**GR**VL**R**
ϕ Fels-1	NP_459877	STM0900	527	MVYAPT**GSGKT**	VMFCV**P**YQ	LII**DEAH**	IGL**SGT**PFS	**Q**AI**GR**GL**R**
ϕ Asp2	AAT36801	Pas53	567	LVVLPT**GTGKT**	IGMVA**P**EL	VWV**DEAH**	VGF**TAT**PER	**Q**MI**GR**GT**R**
ϕ 105	NP_690792	ORF8	311	GFIGPS**GSGKT**	MQEAY**P**DA	GVV**D**TE**H**	VID**SLS**HNW	nd
L. innocua pϕ	NP_471919	lin2589	418	CCSLWL**GSGKS**	NDVDMKFV	III**DESH**	VGF**TAT**PVR	**Q**QSM**R**GM**R**
ϕ adh	NP_050128	orf455	455	LIQSPA**GSGKS**	WGVNMNLC	ILV**DEAH**	LGF**TAT**PVR	**Q**QSM**R**CM**R**
ϕ 2389 / ϕ PSA	NP_511031	hel	418	RCSLWV**GAGKS**	NEVDMSYV	III**DESH**	VGF**TAT**PVR	**Q**QSM**R**GM**R**
ϕ 31	CAC04160	helicase	448	IVQSPP[R]S**GKT**	NQVNMEYV	ILV**DEAH**	LFF**TGT**PIR	**Q**FAM**R**PLN
ϕ O1205	NP_695088	orf10	443	MVQSPP[R]S**GKT**	NGVDMNLV	ILI**DEAH**	LMF**TGT**PAR	**Q**FAM**R**ALN
ϕ Sfi19	NP_049952	orf443	443	MVQSPP[R]S**GKT**	NGVDMNLV	ILI**DEAH**	LMF**TGT**PAR	**Q**FAM**R**ALN
ϕ Sfi21 [1])	NP_049999	orf443	443	MVQSPP[R]S**GKT**	NGVDMNLV	ILI**DEAH**	LMF**TGT**PAR	**Q**FAM**R**ALN
ϕ κ3	AAL74161	hel3.1	443	MVQSPP[R]S**GKT**	NGVDMNLV	ILI**DEAH**	LMF**TGT**PAR	**Q**FAM**R**ALN
ϕ DT1	NP_049421	orf33	443	MVQSPP[R]S**GKT**	NGVDMNLV	ILI**DEAH**	LMF**TGT**PAR	**Q**FAM**R**ALN
ϕ A2	NP_680515	orf33	455	LLVSPA**GSGKS**	DEIDLQAT	IIT**DESH**	LGF**SAS**PWR	**Q**QSM**R**GM**R**
ϕ LE1	CAE14775	LE1-0071	501	LIIMAT**GTGKT**	SGLERAEH	III**DECH**	VGL**TAT**PDG	**Q**MV**GR**GT**R**
ϕ T1	YP_003923	22	672	FVKASVSA**GKT**	SIFSASLG	IGI**DECH**	FGM**TGS**EFR	**Q**LL**GR**GM**R**
ϕ mi7-9	P24125	-	452	MIQSPP**GSGKS**	KVHGV**P**LN	IIT**DEGH**	LGF**TAT**PWR	**Q**QSM**R**SM**R**
ϕ RM378	NP_835691	p104	670	LIADEM**GTGKT**	VLVVC**P**AS	VIV**DECH**	LFL**TGT**PIV	**Q**AED**R**LH**R**
ϕ Barnyard	NP_818603	gp65	653	VIADQP**GLGKT**	ILVVA**P**KT	VIV**DE**SQ	IAI**SGT**PFR	**Q**VEN**R**AYG
pϕ BC6A51	NP_831634	bc1861	397	IGVAPA**GSGKT**	PGIGRIGF	VVV**DEAH**	IGL**TAT**PSR	nd
ϕ 12 [2])	NP_803332	p26	423	GLFLDM**G**L**GKT**	MLVIA**P**KQ	VVI**DEL**S	IGL**TGT**PSP	**Q**ANA**R**LY**R**
ϕ BPP-1	NP_958716	Bbp47	464	AVWAGM**GMGKT**	SLVLA**P**LR	VVA**DES**T	IEL**TGT**PSP	**Q**IIE**R**IGP
ϕ APSE-1	NP_051002	P41	460	NVWAGM**GMGKT**	TLVLA**P**LR	IIA**DES**T	VNL**TGT**PSP	**Q**IIE**R**IGP
ϕ BcepNazgul	NP_919000	gp47	513	ALFIDL**GMGKT**	CLVIA**P**LR	VFI**DES**S	HLL**TAT**PAA	**Q**LI**GR**LA**R**
X. fastidiosa pϕ	NP_299806	xf2528	472	NLFVPM**GLGKT**	MLVIA**P**LR	VVA**DEC**S	IGL**TGT**PAP	**Q**IIE**R**IGP
S. pyogenes pϕ	AAR83209	-	458	AVILDM**GMGKT**	VLVIA**P**LR	VVI**DEL**S	VGL**TGT**PSS	**Q**TNA**R**LW**R**
S. pyogenes pϕ	NP_801687	sps0425	442	GLLLDM**GLGKT**	ILIVA**P**KK	VVI**DEL**S	VGL**TGT**PAP	**Q**ANA**R**LD**R**
ϕ VP5 / ϕ VP2	AAR92075	-	489	ALFMEM**GTGKT**	VLVIV**P**TP	VVV**DES**S	MIL**TGT**PIT	**Q**SED**R**NH**R**
ϕ PY54	NP_892081	RepA	1330	IVRAGM**GSGKS**	AKDIA**P**YA	FGL**DEA**T	ESIAM**T**DDH	**Q**MLR**R**DRT
ϕ KO2	YP_006615	gp35	1328	IVRAGM**GSGKS**	YQEMA**P**YA	FGFDEA**T**	DALAR**T**EDH	**Q**MLR**R**DRT
ϕ VP882	AAS38505	RepA	1239	ILRAPM**GSGKT**	MAVEM**P**YT	LCI**DEA**S	AAMRS**S**RRV	**Q**MMR**R**DRT
ϕ N15	AAC48876	RepA(gp37)	1228	IVRAGM**GSGKS**	YQEMA**P**YA	FGF**DEA**T	DALAR**T**EEH	**Q**MLR**R**DRT

Motif numbering is according to Hall + Matson [158]. Motifs IV and V were omitted due to their high degree of degeneration. Bold letters indicate matches with the conserved residues of the SF2 consensus sequences. Residues within the individual motifs that are specific for the subfamily are boxed. nd: not detected. 1) Homologues of the ϕSfi21 orf443 protein are encoded by phages ϕSfi18 and ϕSfi11 (including identical signature motifs). 2) the M_{AUG} in motif I_{SF2} residue (boxed) was assigned as start codon for this NCBI entry; alternative translation starting from a rare start codon located 28 triplets upstream results in a complete motif I_{SF2} (indicated by underlining) and full-length similarity in the N-terminus to the Bbp47 protein of ϕBPP-1.

Table C19 Superfamily I (SF1) helicases: ϕT4 Dda-type helicases.

SF1 motifs:				I_{S1}	II_{S1}	III_{S1}	V_{S1}	VI_{S1}
consensus				hhxGxAGoGKS A P T	hhhDExo	hhhhhGDxoQ	xxThxxxQGhohooV S K	VAhTRxoo G S
replicon	accession	gene	res.					
ϕ T4	NP_049632	Dda	439	TIN**G**PA**GTGKT**	LIC**DE**VS	TIIGI**GD**NK**Q**	AS**T**FHKA**QG**MSVDRA	**VGVTR**GRY
ϕ RB69	NP_861705	Dda	437	TIN**G**PA**GTGKT**	LLF**DE**AS	TVIGI**GD**RC**Q**	VS**T**FHKA**QG**MSVDTA	**VGTTR**GRF
ϕ RB49	NP_891582	Dda	463	TVR**G**AA**GTGKT**	IIV**DE**AS	VIIAV**GD**LY**Q**	AC**T**IHKS**QG**ISVDR**V**	**VGVTR**ARY
ϕ KVP40	NP_899402	Dda	411	TIS**G**PA**GSGKT**	LVV**EE**AS	RILAV**GD**KY**Q**	AS**T**VHKS**QG**TTVKG**V**	**VGLTR**PTD
ϕ Aeh1	NP_944136	Dda	454	TIS**G**PA**GSGKS**	LIV**DE**AS	QIIAI**GD**KH**Q**	AS**T**FHKS**QG**TTVIG**V**	**VGCTR**AQK
ϕ 44RR2.8t	NP_932377	Dda	439	TIR**G**PA**GTGKT**	LIC**DE**VS	VIIGI**GD**KA**Q**	AM**T**FHKS**QG**STFKNA	**VGNTR**ARE
ϕ PaP2	YP_024782	p54	683	ALT**G**AA**GTGKT**	LIIE**E**AS	QIIYL**GD**IN**Q**	CL**T**VHKS**QG**SEWRK**V**	**TALTR**AKE

Motif numbering is according to Hall + Matson [158]. Motifs I_{SF1} + IV_{SF1} were omitted due to their degeneracy. Bold letters indicate matches with the conserved residues of the SF1 consensus sequences.

teins perform multiple functions for the replication of their cognate replicons comparable to ϕP4 α, including a role of RepA as replicative helicase [507,274].

Phage-encoded superfamily 1 helicases. The Dda helicase of ϕT4 has been characterised as a monomeric helicase with 5'→3' DNA duplex unwinding activity [360,317]. Together with the ϕT4 UvsX SAP, Dda can rescue stalled replication forks by re-activating DNA synthesis by ϕT4 gp43 DNA polymerase when the polymerase encounters a block in the template strand. This reaction may require two sequential template-switches, and Dda and UvsW may provide redundant functions [202]. Homologues of the ϕT4 Dda helicase are encoded by all (completely sequenced) phage genomes from the ϕT4-group (Table C19). Except for the ϕPaP2 gp54 protein, other phage-encoded proteins with significant similarity to the ϕT4 Dda helicase were not detected. The ϕPaP2 gp54 protein contains all signature motifs of SF1-type helicases, but has a N-terminal extension of ~150 res. of unknown function, and may be more closely related to the RecD-type of exodeoxyribonuclease V subunits.

The PcrA helicase of Gram(+) bacteria, and the Rep [PDB 1UAA] and UvrD helicases of Gram(–) bacteria are three closely related bacterial helicases belonging to the helicase superfamily 1. PcrA and Rep have been shown to be essential for the propagation of plasmids and phages that replicate by RCR (BRM Section 2.1.), and this has now also been shown for several *E. coli* plasmids in the case of UvrD [51]. None of these helicases plays a direct role during chromosome replication and their involvement in plasmid and phage replication is a long-standing enigma. Ehrlich and co-workers propose that these helicases may counter-act recombination processes of the RecFOR pathway that would upset chromosome replication [270,345]. Phage-encoded homologues of Rep or PcrA are not yet known.

C3.4. Primases

Primases are DNA-dependent RNA polymerases that synthesise short oligo-ribonucleotides on ssDNA as template, which are then elongated by DNA polymerases. In contrast to RNA polymerases, all known replicative DNA polymerases are incapable of *de novo* synthesis, and therefore require priming (Section C3.5.). Given the sterically correct positioning of a 3'-hydroxyl group, priming of leading-strand DNA synthesis can occur in principle by any one of five different ways:

i. by the 3'-end of DNA in a recombination intermediate (D-loop) or in a partially single-stranded region;
ii. by the 3'-end of an aborted RNA transcript which persists annealed to its template (R-loop);
iii. by the CCA-OH-3' end of a tRNA annealed to a unwound stretch of DNA (cauliflower mosaic virus (CAMV));
iv. by the hydroxyl group of an amino acid side chain (serine, threonine, tyrosine) of a DNA-bound protein (ϕ29);
v. by the 3'-end of a short RNA transcript synthesised in a template-dependent manner on unwound DNA by a primase.

In all studied cases, however, priming of lagging- strand synthesis during coupled leading- and lagging-strand synthesis is performed by a primase. Accordingly, all phage replicons that employ coupled leading- and lag-

Table C20 Phage-encoded primases of the ϕT7 gene 4A primase-type.

motifs:					Zn finger	4	5
consensus					xHxxCx 16 xCxxCx	xEGEhDxh	FDxDxxGxxAxx
replicon	accession	gene	res.	ident. res.			
ϕ T7	P03692	gene 4A	566	[1-245]	YHIPCD 16 FCYVCE	TEGEIDML	FDMDEAGRKAVE
ϕ A1122	NP_848279	p17	566	100%	YHIPCD 16 FCYVCE	TEGEIDML	FDMDEAGRKAVE
ϕ T3	P20315	gene 4	566	82%	FHAPCE 16 WCFVCE	TEGEIDAL	FDMDDAGRKAVE
ϕ YeO3-12	NP_052087	gene 4A	566	85%	YHIPCE 16 YCYACE	TEGEIDAL	FDMDDAGRKAVE
ϕ gh-1	NP_813761	p15	562	56%	KHVPCE 16 FCFACD	TEGEIDCL	FDMDDAGRAASQ
ϕ P60	NP_570328	P60_18	531	36%	RHEPCP 16 YCFSCG	TEGEFDAL	FDNDDAGIQAAE
ϕ SIO1	NP_064752	p15	522	33%	KHQPCQ 15 YCHVCK	TEGEFDAM	FDNDKAGQDAAM
ϕ Felix01	NP_944967	p188	663	27%	ACPRCG 17 TCFSCN	WEGEMECA	FDNDEAGAKATK
ϕ PaP3	NP_775217	p40	569	25%	ACPGCR 24 KCNRCG	SEDELSAM	HDADEEGRKSVE
ϕ SP6	NP_853570	gp10	661	28%	PCPACQ 19 YCNRGH	VGGELDAL	FDGDEIGQKLNQ
ϕ K1-5	AAL84819	gp9	662	30%	PCPACQ 19 YCNRGH	VGGELDAL	FDGDEVGQKQNQ

The presence of the N-terminal zinc finger motif of ϕT7 gene 4 primase is shown together with the spacing of the cysteine residues probably involved in zinc ion binding (indicated by gray shading). The signature motifs 4 + 5 correspond to the conserved functional important motifs of the TOPRIM domain [CDD 8119] defined by Aravind *et al*. [9]. Highly conserved residues are indicated by bold type.

Table C21 Phage-encoded primases of the ϕT4 gp61 primase-type.

motifs:					Zn finger	4	5*
consensus					xCxxCx 16 xCxxCx	hEGPhDSh	KDhNDhh
replicon	accession	gene	res.	ident. res.			
ϕ T4	P04520	gp61	342	[1-342]	RCPVCG 22 HCYNCN	LEGPIDSL	KDVNDMI
ϕ RB69	NP_861720	gp61	340	73%	RCPVCG 21 HCYNCQ	MEGPIDSL	KDVNDMV
ϕ RB49	NP_891587	gp61	342	51%	SCPICG 22 GCFNCN	LEGPIDSL	KDVNAMI
ϕ KVP40	NP_899268	gp61	352	40%	RCPICG 22 LHLRCS	VEGPIDSL	KDINDFI
ϕ Aeh1	NP_943889	gp61	344	46%	RCHICG 22 GCFNCG	FEGPIDSV	KDINDII
ϕ 44RR2.8t	NP_932385	gp61	334	53%	RCPICG 22 HCFNCG	MEGPLDSL	KDINDMI
ϕ 65	AAR90927	gp61	348	43%	KCPLCG 21 GCFNCG	LEGPIDSV	KDINDMV

The presence of the N-terminal zinc finger motif of ϕT4 gp61 primase is shown together with the spacing of the cysteine residues probably involved in zinc ion binding (indicated by gray shading). The signature motif 4 corresponds to the conserved functional important motifs of the TOPRIM domain [CDD 8119] defined by Aravind *et al*. [9], and was detected by alignment of the ϕT4 gp1 and ϕT5 *pri* sequences. Motif 5* was tentatively assigned to a conserved motif with the approximate position of motif 5 in a comparable secondary structure context. Highly conserved residues are indicated by bold type.

ging-strand DNA synthesis for their propagation either depend on the host primase or encode their cognate primase.

Because primer synthesis by primases depends on a single-stranded DNA template they are often closely associated with replicative helicases – phages provide examples for multifunctional proteins consisting of primase and helicase domains (see previous chapter). Although a physical interaction was also demonstrated for some primases with their cognate replicative DNA polymerase, a primase domains within a multi-domain DNA polymerase has not yet been detected. A closer association with DNA polymerases exists for eukaryotic primases, which can be purified as two distinct subunits of the four-protein Pol α complex (for references see [129]). The N-terminal primase domain of the ϕT7 gene 4A primase-helicase is among the best understood primases at the molecular level, has been studied extensively by various biochemical approaches, and is the only primase for which a complete crystal structure is available [204].

We discuss the different types of phage-encoded primases, including homologues of ϕT7 gene 4, of ϕT4 gene 61 (gp61), phage-encoded $DnaG_{Eco}$-type primases, and ϕP4 α-type primase domains. The structure and function of bacterial and phage-encoded primases has also been reviewed recently by Frick and Richardson [129].

The gene 4A primase domain of phage T7. As early as 1969, it became known that ϕT7 gene 4 is essential for replication [435], and that the protein shows primase activity [432]. It was shown that gene 4 mutants synthesise only small amounts of DNA hybridising exclusively to one parental strand – the leading-strand template as we know today. As primase, ϕT7 gene 4A protein has the preferred recognition sequence 5'-GTC on the single-stranded template, and the primers synthesised in vitro or in *vivo* and attached to Okazaki fragments were found to be mostly tetra-ribonucleotides with the sequence 5'AC(A/C)(AC) [432,412,391,405, 330].

Deletion analyses could show that the primase activity of gene 4A protein resides in the N-terminus, and the helicase activity in the C-terminus. These results were in accordance with earlier observations, that the two translation products of gene 4 ocurring *in vivo* differ with respect to these two activities depending on the start codon used. The truncated N- terminus (271 res.) retained the primase activity but lacked helicase activity [128]. The crystal structure was determined for a polypeptide containing the first 182 res. of the gene 4A primase domain, showing its organisation in two separate domains: the N-terminal Zn-binding domain, and the RNA polymerase domain [PDB 1NUI] [204]. Interestingly, primer extension by ϕT7 DNA gene 5 polymerase is already stimulated by the isolated Zn-binding domain [204].

A physical interaction between the ϕT7 gene 4A protein and the acidic C-terminus of the gene 2.5 SSB has been demonstrated *in vitro* [212] (Section C3.6.1.). The ϕT7 gene 2.5 SSB bound to ssDNA stimulates the rate of primer synthesis by gene 4A *in vitro*, and also stimulates primer extension by gene 5 DNA polymerase in a coupled system [316,289].

As mentioned in the preceding chapter, gene 4 of ϕT7 encodes two proteins: the full-length gene 4A or 63 kD protein, and – starting from codon 64 – a shorter gene 4B or 56 kDa protein that lacks the N-terminal Zn-finger motif. The gene 4B protein is active as helicase but lacks primase activity [111,29]. BLAST searches revealed the presence of genes encoding homologues of the ϕT7 gene 4 primase-helicase in all phage genomes that also encode DNA polymerases of the ϕT7-subfamily of the Pol I-type DNA polymerases (Table C11; see also BRM Section 3.4.). The ϕSP6 gp10 primase has been characterised biochemically and found to be largely comparable to ϕT7 gene 4A primase except that primer synthesis occurs preferentially at 5'-GCA trinucleotide template sequences [457]. A Met codon at pos. 64 as in ϕT7 gene 4 is present in the corresponding genes of phages ϕA1122, ϕT3, and ϕYeO3-12, and the translation pattern of these genes might be comparable to that of ϕT7 gene 4 (see above). Met codons close to the position of the secondary start codon in ϕT7 gene 4 were found in the ϕSIO1 p15 gene, and the ϕP60 gene 18, but not in the other primase-helicase genes of this subfamily (Table C20).

Gene 61 (gp 61) of phage T4. Genetic analyses of ϕT4 replication mutants and inspection of their progeny DNA molecules by electron microscopy led to the detection of gene 61, responsible for the phenotype of a reduced rate of DNA synthesis and accumulation of ssDNA [48,309,160]. The product of ϕT4 gene 61 (gp61) was then shown to be responsible for the requirement for NTPs in a T4 DNA *in vitro* replication system, qualifying gp61 as primase [302].

ϕT4 gp61 primase is a monofunctional protein that requires gp41 helicase for optimal activity [321,465, 176,370] reminiscent of the ϕT7 gene 4A protein containing both activities as separate domains in a contiguous polypeptide. The analysis of the 5'-ends of Okazaki fragments synthesised *in vivo* after infection of *E. coli* cells with ϕT4 or in an *in vitro* system revealed that ϕT4 gp61 primase synthesizes mostly pentanucleotide primers with the sequence 5'-ACNNN [236, 332,321,255]. Similar to the interaction of ϕT7 gene 4A primase with gene 2.5 SSB, also ϕT4 gp61 primase and gp32 SSB protein could be shown to interact *in vitro*: both proteins form a stable complex in the absence of DNA, and the negatively charged C-terminus of gp32 is responsible for this interaction (Section C3.6.1.) [56,57]. There is a complex interplay of gp61 with gp32 on one hand, and of both with the accessory proteins gp44, gp45, and gp62 (clamp-loader and clamp proteins; see Section C3.5.2) on the other hand. This interplay is thought to be instrumental in synchronising lagging-strand primer synthesis with subsequent primer elongation by gp43 DNA polymerase [66].

Except for *Aeromonas* sp. ϕ25 (sequence not yet complete), all phages of the ϕT4 group encode homologues of the gp 61 primase, and all proteins of this subfamily contain the N-terminal Zn-finger motif and the 'TOPRIM domain 'motifs (Table C21).

Phage-encoded *E. coli* DnaG-type primases. *E. coli* DnaG was the first bacterial protein shown to catalyse the synthesis of oligo-ribonucleotides required for *in vitro* replication [42]. Two standard assays for the analysis of primase activity *in vitro* were established by

Table C22 Phage-encoded (putative) primases of the *E. coli* DnaG-type.

motifs:				Zn finger		4	5
consensus				xCxxCx	16	xCxxCx	hEGPhDSh FDxDxxGxxAxx
replicon	accession	gene	res.				
E. coli	NP_417538	*dnaG*	581	CCPFHN	15	HCFGCG	VEGYMDVV YDGDRAGRDAAW
ϕ P4 / ϕ R73	P10277	alpha	777	RHQPCP	19	YCNQCG	LEGQNQAG ADRDLSGDGQKK
ϕ C31	NP_047955	gp9a	805	LCPAHS	18	TCRAGC	TEGPGDAL GDNDTAGVGFTL
ϕ BT1 [1])	NP_813724	gp9a	808	VCPAHA	18	TCRAGC	TEGPGDAL GDNDTAGTGFTT
B. subtilis pϕ SPBc2	NP_046683	*yorJ*	378	YRTVCH	18	HCYTEC	nd LDYKDSPADKGK
ϕ T5	YP_006949	*pri*	296	LNPDHD	15	HCLSCG	VEGIFDML LDNDASGNKAAQ
ϕ Bxz1	NP_818246	gp195	327	IVHSCL	26	VCWAFW	GEGRLDRV ADDDDAGRFMER
ϕ K	AAO47525	orf76	355	CCPFCG	19	HCKKCD	TEGVFDAL LDTDALDNNIDL
ϕ Xp10	NP_858989	xp10p42	280	RHAQCG	18	YCHRCG	TEDAISAY LDNDTGHSSGSN
ϕ KMV	NP_877453	orf14	274	HVLGCQ	20	YCYSCQ	TEDYLSAL LDGDPAGVRGSA
ϕ VpV262	NP_640276	p15	287	VCPCCR	17	LAYRCY	nd LDADATSKAASM
ϕ T1	YP_003921	24	306	PCPNCG	19	ICNSCG	TESFDVGI SDKDTLYMADDR
pϕ Fels-1	NP_459878	stm0901	322	ECPVCG	16	IC-VCG	LQEDNYLD ADRDENSATGLA
ϕ RM378	NP_835688	p101	310	LCPKCA	19	ICFRCG	FEGMFDML FDSDVSVEEISK
ϕ PY54	NP_892081	RepA	1330	nd		nd	IIGKLRGA LDNDQKSAAEGK
ϕ KO2	YP_006615	gp35	1328	nd		nd	VIGDLKGA LDNDRKSSAEGK
ϕ VP882	AAS38505	RepA	1239	nd		nd	IIGSLTDA ADNDAWKPHVGN
ϕ N15	AAC48876	RepA(gp37)	1228	nd		nd	LIGDLQGA LDNDRKSSAEGK

The presence of the N-terminal zinc finger motif is shown together with the spacing of the cysteine residues probably involved in zinc ion binding (indicated by gray shading). The signature motifs 4 + 5 correspond to the conserved functional important motifs of the TOPRIM domain [CDD 8119] defined by Aravind *et al.* [9]. Highly conserved residues are indicated by bold type. 'nd' not detected.

the Kornberg lab: in the first assay, primer synthesis by DnaG starts at a hairpin region in ϕG4 or ϕX174 DNA, which is left uncovered by SSB; the oligo- ribonucleotides synthesised in this assay are 26–29-nt in length [41,183,378]. The other assay – termed 'general priming reaction' – uses ssDNA, DnaB helicase and primase [6]; in this assay, the priming reaction depends on the ability of DnaB to load DnaG to the ssDNA. Primers synthesised by DnaG *in vitro* start with 5'GA but are considerably longer than those synthesised by the phage-encoded primases: 10–60 nt, with the majority around 11 nt [7,502]. The interaction of *E. coli* DnaG with the replicative helicase DnaB is weak *in vitro*, however, the orthologues from *Bacillus stearothermophilus* were shown to form a stable complex [34]. Proteolytic removal of <50 res. from the $DnaG_{Eco}$ C-terminus resulted in polypeptides that were still able to direct the synthesis of RNA primers although they were inactive in the helicase-dependent priming assay [265,455,438]. The biochemical analysis of the DnaG Q576A mutation – in the extreme C-terminus – was particularly rewarding: the interaction of the mutant protein with DnaB is significantly weakened [265], the initiation of bi-directional replication is less efficient than with the wildtype protein [169], and Okazaki fragments are significantly longer [455]. Intriguingly, the interaction of DnaB with DnaG is apparently not only necessary to direct the primase to its template but, in addition, modulates the DnaG conformation such that a reduction of sequence specificity allows more efficient lagging-strand priming [285,493,32].

As already addressed in the previous chapter, the initiator, helicase and primase functions required for ϕP4 replication are performed by the α protein: ϕP4 replication *in vitro* proceeds (mostly) bidirectionally from *ori*, and elongation occurs probably as coupled leading- and lagging-strand synthesis, without a requirement for host DnaG primase and DnaB helicase [103]. Primers synthesised by the ϕP4 α protein on ssDNA are 2–5-nucleotides long and begin with 5'-AG [506]. Host DnaG can only poorly substitute for the α E214Q primase(null) mutation in ϕP4 replication, and – *vice versa* –

Table C23 Split genes encoding (putative) primases of the *E. coli* DnaG-type.

motifs:						Zn finger		4	5
consensus				N	C	xCxxHx	15 xCxxCx	xEGYMDxh	xDGDxxGxxxAx
replicon	accession	gene	res.	ident.	ident.				
E. coli	NP_417538	*dnaG*	581	32%	28%	CCPFHN	15 HCFGCG	VEGYMDVV	YDGDRAGRDAAW
ϕ D29	NP_046874	gp58	129	[1-129]	-	LCPFHG	15 NCLACG	-	-
	NP_046873	gp57	152	-	[1-152]	-	-	TEGEIDAI	ADGDEPGMEFAK
ϕ L5	NP_039722	L5p55	130	84%	-	LCPFHG	15 NCMACG	-	-
	NP_039721	L5p54	152	-	80%	-	-	CEGELDTI	ADGDDAGMEFAR
ϕ Bxz2	NP_817647	gp58	135	69%	-	RCPFHG	15 RCFACP	-	-
	NP_817646	gp57	157	-	51%			TEGEIDAI	ADGDEPGLQFAN
ϕ Bxb1	NP_075318	gp51	171	62%	-	LCWHHE	15 NCLACS	-	-
	NP_075317	gp50	157	-	39%	-	-	AEGEIDAL	ADGDDAGMQFAE

The presence of the N-terminal zinc finger motif of *E. coli* DnaG is shown together with the spacing of the cysteine residues probably involved in zinc ion binding (indicated by gray shading). The signature motifs 4 + 5 correspond to the conserved functional important motifs of the TOPRIM domain [CDD 8119] defined by Aravind *et al.* [9]. Highly conserved residues are indicated by bold type.

α cannot suppress an *E. coli dnaG*(ts) strain [431]. The isolated N-terminus (253 res.) of α was shown to retain its primase function; note that in this truncated protein the TOPRIM motif 5 is located 5 residues downstream from the amber stop [505]. The primase(null) phenotype of the E214Q mutation suggests that the EGYATA motif downstream at short distance from motif 4 is functionally important (see Fig. C15). This motif is in most cases only poorly conserved in other phage-encoded DnaG-type primases, but several plasmid-encoded primases were detected that show conservation of this motif [431,335]. However, the functional role of this sequence motif is elusive.

BLAST searches revealed several phage-encoded proteins with significant similarity to DnaG (Table C22). In most of these proteins, the Zn-finger motif and the TOPRIM signature motifs 4 + 5 were readily detected, but for the sizes of the proteins there are clearly three different types: i. the proteins with a size of ~800 res. including ϕP4 α and two of its homologues, which also contain the ϕP4 α-type C-terminal helicase and DNA-binding domains (see previous chapter); ii. the proteins with a size of ~300 res. which probably represent the minimal version of a monofunctional primase; and iii. the large ϕN15 RepA-type proteins containing the TOPRIM in their N-termini, a variant of the ϕP4 α EGYATA motif, but lacking a detectable Zn-finger motif. The primase activity of the ϕN15 RepA protein has not yet been demonstrated, but the striking analogy of roles for the replication of their cognate replicons between the RepA and ϕP4 α suggests this function [507,364].

A unique organisation exists for the (putative) primase genes of four related Mycobacteriophages: a gene encoding a small polypeptide of ~130 res. containing a putative Zn-finger structural motif is located directly upstream of a gene encoding a ~150 res. polypeptide containing the primase signature motifs 4 + 5 (Table C23). These two open reading frames are separated by 1 nt in the genomes of ϕD29, ϕL5, and ϕBxz2. Also in the ϕBxb1 genome, gp51 and gp 50 are separated by the same 'reading-frame shift' but the stop codon present in the ϕD29 gp58 gene has been changed in the ϕBxb1 gp51 gene. This results in the 42 res. longer gp51 protein, and – as a consequence – the reading frames encoding gp51 and gp50 overlap. ϕBxb1 gp50 contains the signature motifs but is – as compared to its homologues in the other phages – considerably less well conserved, probably due to compensation for the overlap. Taken together, it is unlikely that the split of the genes is simply a sequencing artefact in all 4 sequences. Biochemical analysis would have to show that both polypeptides together exhibit primase activity but the genetic organisation is suggestive (see also Fig. C17). All four phages of this group encode highly similar DNA polymerases of the Pol I-type (Section C3.5.1.) and helicases (Section C3.3.). It is thus very likely that these phages inherited their unique primase genes together with the other replication genes from a common ancestor. A common origin of these mycobacterial phages was as also proposed by the Pittsburgh phage group from a whole-genome comparison [125,342].

The ϕT7 gene 4A and ϕP4 α primase domains, the ϕT4 gp61 primase and E. coli DnaG all contain the 'TO-

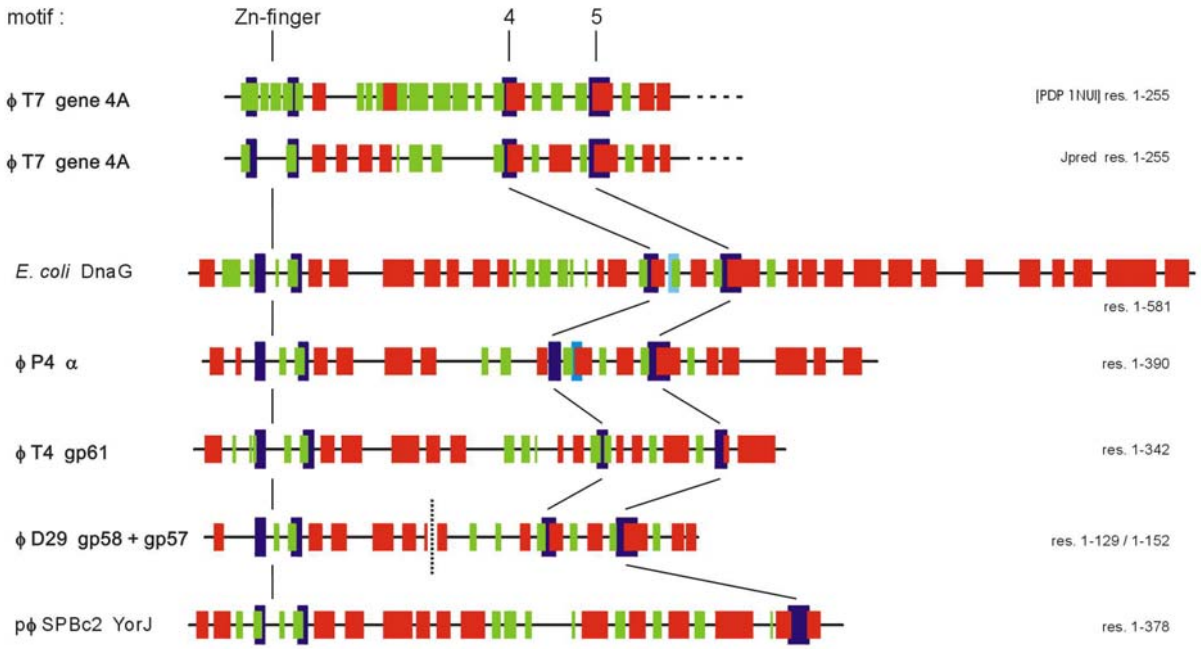

Fig. C17 Secondary structure prediction for DnaG-type primases/primase domains.
The (putative) primase/primase domain sequences, and the size/position of the signature motifs are drawn to scale. For signature motif numbering see Tables C23 + C24. Secondary structure predictions were obtained from the Jpred server [88]. Colour code: red = predicted α-helical region; green = predicted β-stranded region.Dark blue = signature motifs as listed in the Tables; light blue = EGYATA motif (see Text for details) [431]. The split of the two (putative) primase genes of ϕD29 is indicated by a dotted line.

Table C24 (Putative) primase domains of *Lactobacillus* sp. and *Burkholderia* sp. helicases.

motifs:					1	2	3	4
consensus					GhGFhh	hh.hGhDhD	xSGxGhHhhhRGx	hEhY..xxGRFFxxTG
replicon	accession	gene	res.	ident.				
ϕ A2	NP_680517	orf35	770	[1-300]	GLGFFF	GY.VGIDVD	LSGEGIHIIIRGK	VEMY..ESGRFFAMTG
ϕ 2389 / ϕ PSA	NP_511033	*pri*	757	50%	GLGFYF	PY.FGVDID	VSGTGIHIIAKGN	IEMY..PDGRFFVMTG
ϕ adh	NP_050131	orf771	771	50%	GLAFYF	GY.VGLDVD	QSGKGIHAVFKGK	YEMY..EAGRFFALTG
L. innocua pϕ	NP_471917	lin2587	757	49%	GLGFYF	PY.FGVDID	VSGTGIHIIAKGS	IEMY..PDGRFFVMTG
ϕ 105	NP_690795	ORF11	806	37%	GIGFMF	PF.IGIDID	PSGEGVHIIAKGK	LEVY..RHGRYFTFTG
ϕ BcepB1A	AAT37747	-	919	25%	GDGFTY	GFWLTADGG	SSGQGLHLFGRGV	LEFYTDKRGIAFGLSG

Motifs 1 - 4 were detected by BLAST alignment of the sequences as the regions with most significant conservation.

PRIM domain' (signature motifs 4 + 5; [9]) in addition to the N-terminal Zn-finger motif. However, none of the proteins shows any significant sequence similarity to the others, and their sizes differ considerably. The known crystal structures of the ϕT7 gene 4A primase domain and of a larger part of *E. coli* DnaG offered the possibility to compare these structures with each other, and to secondary predictions [205,204] (see Fig. C17). The Jpred algorithm gives a reliability value of ~75% for the predictions shown in Fig. C17, which seems reasonable. However, the algorithm is apparently biased towards the prediction of α-helices and less suited to predict β-stranded regions. The embedding of the Zn-finger motif in several β-strands is reflected by the prediction, as is the positioning of motifs 4 + 5 in α-helical regions. The region connecting motifs 4 + 5 differs in both crys-

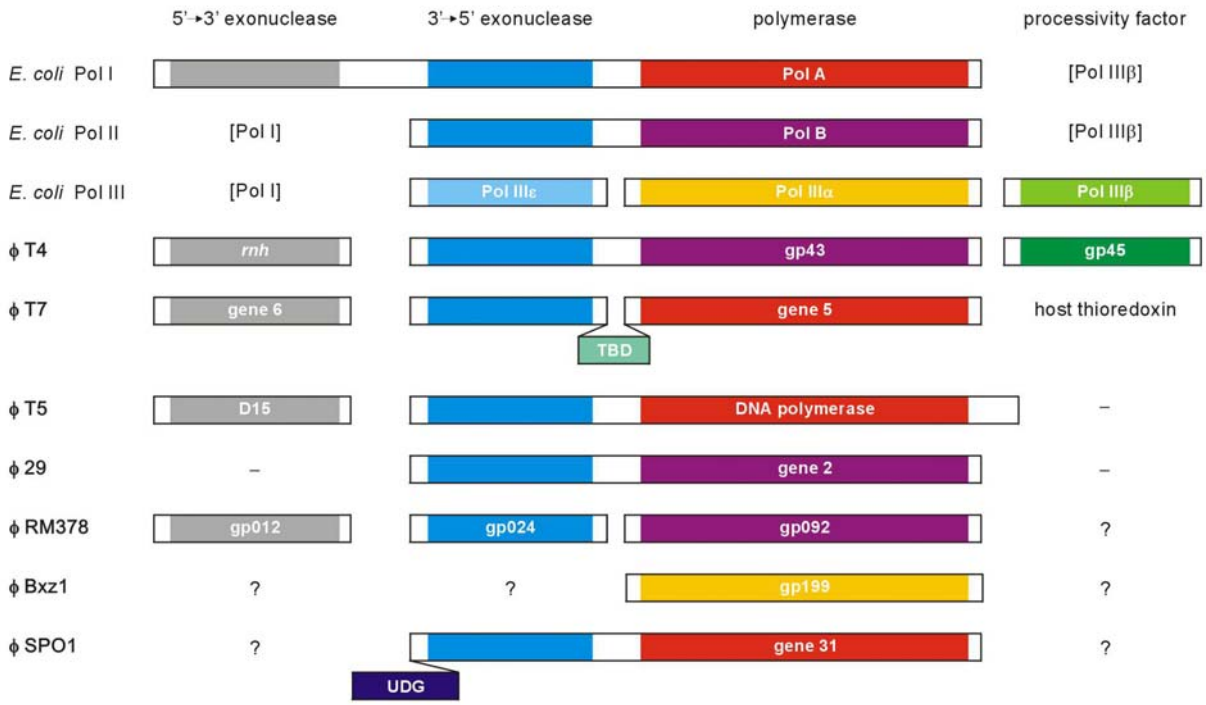

Fig. C18 Architecture of selected phage-encoded DNA polymerases.
Proteins are shown as open boxes (not to size; gene/protein name within box) with colours indicating the functional domain(s). Functional domains belonging to the same (sub)families are shown with identical colour (see text for details). 'UDG' uracil-d-glycosylase-like domain. 'TBD' thioredoxin-binding domain. '[Pol I]' indicates that the corresponding function is carried out by DNA Pol I (PolA). 'Pol IIIβ' indicates that the corresponding function can be carried out by the DNA Pol III β-subunit. '–' indicates that the corresponding function is not required for the respective replicon. A question mark indicates that a corresponding cognate gene of factor is not known yet for the respective replicon.

tal structures, and also the prediction cannot offer a plausible solution. In addition, the considerable length variation of the segment connecting the Zn-finger and motifs 4 + 5 makes it difficult to make any conclusive statement about the relationship of the proteins based on this type of analysis. Tentatively, we speculate that the ϕT7 gene 4A and ϕP4 α primase domain, ϕT4 gp61, and the split (putative) primase proteins gp58 and gp57 of ϕD29 are more related to each other than to any of the other two primases.

We have discussed the various types of primase-encoding genes found in phages – single, split, fused to helicase genes – that are all related to the prototype DnaG primase of *E. coli*, although rather distantly in most cases. We could not detect any phage-encoded primases with significant similarity to plasmid-encoded primases, e.g. to RepB of RSF1010 or to ColE2 Rep primase. Nevertheless, there may exist yet unknown primase genes: we were unable to detect – except for ϕT5 – primase-encoding genes in any of the (pro)phages that encode helicases (see previous chapter) and DNA polymerases of the ϕT5- and the ϕ12-subfamilies of Pol I-type DNA polymerases (BRM Section 3.4.).

Also, it is presently not possible to decide whether the N- terminal extensions of some very large F4-type and ϕP4 α-type helicases found in several phages can function as primases (see previous chapter; Fig. C15). E.g. several phages of Lactobacilli (Gram(+)) and *Burkholderia* sp. (Gram(–)) encode large proteins with significant similarity to other ϕP4 α- type helicases in their C-termini together with ~300 res. long N-termini of unknown function. These N-termini contain several well-conserved motifs in β-stranded regions (Jpred prediction; not shown) lacking any similarity to known primase motifs (Table C24). Whether these genes encode a novel type of primase-helicase enzyme needs to be clarified by biochemical experiments.

The observation that largely unrelated proteins could be classified by genetical and biochemical analysis as primases strongly suggests that such proteins evolved several times independently. This view is supported by the finding that for none of the known primases even a

distant relationship to DNA dependent RNA polymerases could be detected so far. For bacteriophage research, this situation is uncomfortable since it makes the prediction of gene functions for newly sequenced phage genomes difficult.

C3.5. DNA polymerases and accessory proteins

C3.5.1. DNA polymerases

DNA polymerases are the key enzymes for the propagation of DNA genomes. All known DNA polymerases synthesise polynucleotide in the 5'→3' direction, and complementary to a single-stranded DNA template. Replicative DNA polymerases are unable to initiate *de novo* DNA synthesis on a DNA template but require the existence of a primer containing a free hydroxyl group to start DNA elongation [228]. Since Kornberg's purification of *E. coli* Pol I in the 1950s of the last century DNA polymerases are among the most intensely studied molecules, genetically, biochemically, and structurally [30]. A highly instructive review by Baker and Bell [18] might help the reader to find a way through the vast amount of literature, but other reviews are also suggested as starting points [201,422,428,381,27,186,150].

DNA polymerases are composed of structurally and functionally defined domains: i. a domain responsible for template binding, primer binding, and 5'→3' polynucleotide synthesis, ii. a proofreading or 3'→5' exonuclease domain responsible for the removal of misincorporated nucleotides [388, 366], iii. a 'structure specific 5'→3' exonuclease' or RNase H domain for the removal of RNA primers, and iv. a processivity factor in replicative DNA polymerases. *E. coli* DNA Pol I combines three functions in a single polypeptide, but lacks processivity. In contrast, polymerase functions are contributed by separate subunits in the *E. coli* Pol III holoenzyme which lacks a cognate 5'→3' exonuclease. Pol IIIα (*dnaE*) is the polymerase subunit, Pol IIIε (*dnaQ*) the 3'→5' exonuclease subunit, and Pol IIIβ (*dnaN*) is the processivity factor, the sliding clamp (Section C3.5.2.). The Pol III γ/τ, δ, and δ' subunits are responsible for clamp-loading, and the τ, χ, θ, and ψ small subunits of DNA Pol III mediate communication among the subunits and with other components of the replisome, e.g. SSB, the DnaG primase and the replicative helicase DnaB. Phage-encoded DNA polymerases provide further examples of DNA polymerase functions organised in individual subunits or combined in integral proteins. ϕT4 DNA polymerase gp43 contains 3'→5' exonuclease and polymerase domains, and requires the ϕT4 gp45 sliding clamp for processivity. The DNA polymerase of ϕ29 contains 3'→5' exonuclease and polymerase domains, and possess high intrinsic processivity. ϕT7 gene 5 DNA polymerase contains 3'→5' exonuclease and polymerase domains, and recruits host thioredoxin as processivity factor. In several phage replicases separate proteins carry the 3'→5' exonuclease and polymerase functions, e.g. ϕBxz1, ϕRM378. In all phage replicons mentioned above, the 'structure specific 5'→3' exonuclease or RNase H is encoded by a separate gene (Fig. C18).

Most of the approximately 60 presently known phage-encoded (putative) DNA polymerases were detected by recent genome sequencing projects. All these putative DNA polymerases can be easily grouped into one of the known DNA polymerase families (see below). However, despite the wealth of structural data available for various DNA polymerases, four major topics exist where even a thorough comparison still cannot help to predict reliably the biochemical properties of a candidate DNA polymerase detected by genomics, emphasising the undiminished importance of biochemistry:

i. it is not predictable whether a candidate polymerase is a high-fidelity or an error-prone polymerase. E.g. *E. coli* Pol II (*polB*) and ϕT4 gp43 DNA polymerase are members of the same polymerase family but differ with respect to error rate during synthesis.
ii. the strand-displacement capacity is not predictable. E.g. the ϕ29 p2 and ϕT4 gp43 DNA polymerases are members of the same polymerase family, but ϕ29 p2 DNA polymerase is highly proficient in strand-displacement during processive DNA synthesis and can dispend with a helicase, ϕT4 gp43 polymerase in contrast requires its cognate processivity factor and helicase.
iii. the processivity of a DNA polymerase is not predictable. E.g. the ϕT7 gene 5 and the ϕT5 DNA polymerases belong both to the Pol I-family of DNA polymerases, ϕT5 polymerase is inherently capable of high processive DNA synthesis, while ϕT7 gene 5 requires host thioredoxin as processivity factor. As a second example, *E. coli* PolIIIα (*dnaE*) and ϕT4 gp43 polymerase belong to different polymerase families but both employ a sliding clamp as processivity factor [206].
iv. it is not possible to predict which molecular mechanism governs the coupling of leading- and lagging-strand synthesis. This co-ordinated synthesis requires a functional asymmetry in the replisome. It is also not possible to predict from the sequence of a DNA polymerase how this asymmetry is put into practice for a given replisome [283].

We confine ourselves to a discussion of the different types of DNA polymerases encoded by bacteriophages, and their functional organisation in domains or separate

proteins. For practical reasons we chose to sort the phage DNA polymerases by their similarity to one of the five known DNA polymerases of *E. coli*: Pol I (*polA*), Pol II (*polB*), Pol IIIα (*dnaE*), Pol IV (*dinB*), and Pol V (*umuC*) [363].

***E. coli* Pol I (*polA*)-type DNA polymerases.** The crystal structure of the Klenow fragment of *E. coli* Pol I is known since 1991 [25]. For a thorough discussion of Pol I structure and functions we refer the reader to the review of Joyce and Steitz [201]. By the BLAST approach, altogether 40 phage-encoded (putative) DNA polymerases were detected (Table C25). These 40 DNA polymerases include 3 arbitrarily chosen prophage-encoded genes, which were however too numerous – particularly among *Xylella* strains – to include them all. None of these (mostly putative) phage DNA polymerases contains the N-terminal 5'→3' exonuclease domain characteristic for orthologues of *E. coli* Pol I . All protein sequences contain the 3'→5' exonuclease signature motifs defined by Salas and co-workers [28], and the polymerase signature motifs originally defined by Braithwaite and Ito [44]. This with almost perfect conservation of the functionally important residues – the few exceptions are discussed further below. Because these signature motifs were defined by combining the results of biochemical analyses of proteins mutated at catalytically important positions with known crystal structures the relevance of their detection in a candidate protein is considered superior to a calculated similarity (BLAST) value [201].

We found a reasonable correlation between the similarity of the proteins and the presence of several conserved individual residues within the 4 signature motifs of the polymerase domain. The degree of conservation of these residues within the signature motifs reflects the phylogenetic relationship of these DNA polymerases, an issue beyond the scope of this review. Here, these conserved residues allowed the loose definition of 4 subfamilies of the phage-encoded Pol I-type DNA polymerases: the ϕT7, ϕ12, ϕD29, and ϕT5 subfamilies. The considerable length variation among the 40 DNA polymerases (between 545 and 993 res.) is discussed in the subfamily paragraphs.

Phage T7 gene 5-type subfamily. The crystal structure of the ϕT7 gene 5 DNA polymerase is known [PDB 1T7P] [109]. It is the prototype phage-encoded DNA polymerase of the Pol I family, and contains all conserved signature motifs of the 3'→5' exonuclease and polymerase domains. Extensive biochemical analyses revealed that co-ordination of the highly processive leading- and lagging-strand synthesis by this polymerase requires an intricate interplay of the ϕT7 gene 4 helicase-primase, the gene 2.5 SSB, and 2 polymerases at a specific DNA loop formed during synthesis of Okazaki fragments [248,247].

ϕT7 gene 5 DNA polymerase requires host thioredoxin for processivity [441]. A domain (~ 76 res.) responsible for thioredoxin binding (TBD) could be identified within the region connecting the 3'→5' exonuclease and polymerase domains [175]. As inferred from sequence similarity, TBDs are also present in the DNA polymerases of phages ϕT3, ϕA1122, ϕYeO3-12, and ϕgh-1. The domain character of the closely related TBD of ϕT3 DNA polymerase (97% ident. res.) was demonstrated by its functional transfer to a corresponding position in an *E. coli* Pol I-type DNA polymerase lacking this region [94]. The DNA polymerases of phages ϕSP6, ϕK1-5, and ϕFelix01 also contain an extended region connecting the 3'→5' exonuclease and polymerase domains, but there is no primary sequence similarity to the ϕT7 gene 5 TBD, and its function therefore unclear. The DNA polymerases of Cyanophage ϕP60 may also use thioredoxin as processivity factor: the locus P60_19 directly upstream of the polymerase gene encodes a putative thioredoxin. Roseophage ϕSIO1, and *Pseudomonas* sp.phage ϕPaP3 lack the extended interdomain region, and it is not known how these DNA polymerases maintain processivity. ϕPaP3 DNA polymerase seems to be encoded by two separate genes: while p39 is similar to the N-terminal ~190 res. of ϕT7 gene 5 DNA polymerase, the larger p32 protein is similar to the 'core' and contains the signature motifs.

The ϕT7 gene 5 DNA polymerase contains a small loop (EGDK motif) closely upstream of the signature motif 2a (res. 401-404). Deletion of this loop results in a mutant polymerase with a defect in primer utilisation and in its ability to contact the primase subunit of ϕT7 gene 4 helicase-primase [78]. The EGDK motif (EGDN in ϕgh-1 p18) is only conserved in those DNA polymerases of this subfamily that – like ϕT7 gene 5 – contain a TBD.

Phage T5 DNA polymerase-type subfamily. The DNA polymerase of ϕT5 has been characterised biochemically as a protein with high inherent processivity and a strong strand-displacement activity [245,73]. ϕT5 DNA polymerase possesses a longer than usual C-terminal extension (~ 50 res.) also present in the ϕSPO1 gene 31 DNA polymerase but lacking in all other Pol I-type DNA polymerases. This region of the protein is suspected to be responsible for the high intrinsic processivity of the protein but direct evidence has not yet been presented [245]. This group of seven phage DNA polymerases is rather heterogeneous, and individual sequences have only a very low similarity with ϕT5 DNA polymerase. They were grouped mostly because the second conserved aspartic acid residue in polymerase motif C is replaced by either serine or alanine; conserved residues

Table C25 Conserved motifs in the 3'→5' exonuclease and polymerase domains of *E. coli* Pol I-type polymerases.

domains				3'→ 5' exo domain			polymerase domain				
motifs:				ExoI	ExoII	ExoIII	2b	A 3	B 4	C 5	
consensus				DxE	Nx_{2-3}FD Y	YxxxD	TGRLSS	DhSxIELR	RxxAKxhNFGhhYG	Rhhh-VHDEhhh K	ident. res.
replicon	accession	gene	res.								
E. coli	P00582	*polA*	928	DTE	NL.KYD	YAAED	TGRLSS	DYSQIELR	RRSAKAINFGLIYG	RMIMQVHDELVF	
ϕ T7	P00581	gene 5	704	DIE	NGHKYD	YNVQD	TGRATH	DASGLELR	RDNAKTFIYGFLYG	AYMAWVHDEIQV	[331-704]
ϕ A1122	NP_848283	p21	704	DIE	NGHKYD	YNVQD	TGRATH	DASGLELR	RDNAKTFIYGFLYG	AYMAWVHDEIQV	98%
ϕ T3	NP_523320	gene 5	704	DIE	NGHKYD	YNVQD	TGRATH	DASGLELR	RDNAKTFIYGFLYG	AYMAWIHDEIQV	97%
ϕ YeO3-12	NP_052093	gene 5	704	DIE	NGHKYD	YNVQD	TGRATH	DASGLELR	RDNAKTFIYGFLYG	AQMGWIHDEVQI	89%
ϕ gh-1	NP_813764	p18	709	DIE	NGIKYD	YCEQD	TGRATH	DASGLELR	RGIAKTFIYAFLYG	CYMAWVHDELQI	56%
ϕ P60	NP_570330	P60_20	587	DIE	NIVNYD	YCARD	TGRQAH	DASGLELR	RKSGKGVTYCLIYG	YPLAFVHDEQQL	37%
ϕ SIO1	NP_064753	p16	580	DIE	LKDVAD	GFMLD	TGAVTH	DAAGIQLR	RPTAKTFIYAFLLG	KQCAWVHDEWQT	32%
ϕ PaP3	NP_775218	p39	175								30%
	NP_775225	p32	545	DLE	QWNIKD	TVQGD	TGRMRH	DGAGLELR	RDDAKTFIYAFIYG	WKVLDIHDEGQW	25%
ϕ SP6	NP_853574	gp14	849	DAE	NFLGYD	SKLTD	TFRMRH	DGAGLELR	RDMAKTFIYAFLYG	CGVANVHDEIQM	34%
ϕ K1-5	AAR90053	gp13	846	DAE	NFLGYD	SKLTD	TFRMRH	DGAGLELR	RDMAKTFIYAFLYG	CGIANVHDEIQM	33%
ϕ Felix01	NP_944958	p179	906	DTE	NGLGYD	GWKAD	TGRMTQ	DQNSAQLV	RKKAKNGIYALLFG	RWVCSYHDEVSL	23%
ϕ 12	NP_803318	p12	650	DIE	NA.NFE	YCIRD	TGRWAG	DFSAIEAR	RQKGKVSELALGYQ	KIVGHVHDEVIV	[251-650]
S. pyogenes pϕ	AAR83201	-	651	DIE	NA.SFE	YNRRD	TGRWAG	DFSAIEAR	RQKGKIAELACGYG	KIVGHVHDEVII	48%
S. pyogenes SSI-1	NP_801684	sps0422	640	DIE	NA.QFE	YCIQD	TGRWAG	DFSAIEAR	RQKGKISELALGYQ	GVVFHVHDEAII	55%
ϕ SPO2	P06225	gene L	648	DIE	NA.NFE	YCIQD	TGRWAG	DFSAIEAR	RQKGKVAELALGYQ	KTVMHVHDEAVL	56%
X. fastidiosa 9a5c	NP_299803	XF2525	726	DLE	NS.HFD	YAQRD	TGRWAG	DLSNIEGR	RQIGKVQELALGYG	SIVLTVHDEIIT	32%
ϕ BPP-1	NP_958711	Bbp42	688	DLE	NS.HFD	YAGLD	TGRWAG	DLSNIEGR	RQIGKVMELMLGYE	QIVLSVHDENIT	34%
ϕ APSE-1	NP_051006	P45	993	DLE	NGGMFD	YAGSD	TGRWAG	DLSNIEGR	RQIGKVMELGLGYG	EIVLTVHDEIIS	31%
ϕ BcepNazgul	NP_919002	gp49	693	DYE	NA.QFE	YNQQD	TQRWAG	DLSSIETV	RTMSKPAVLGAGYR	NIVLHVHDEIVT	31%
ϕ Vp16C	AAQ96570	orf39C	686	-	NI.TFE	YRYCD	TGRWSA	DFSAIEAV	KSLGKVAELASGYG	PVVMHTHDELIA	33%
ϕ Vp16T	AAQ96505	orf38T	714	DFE	NI.TFE	YAYCD	TGRWSA	DFSAIEAV	KSLGKVAELASGYG	GKCYA VHDITYA VHDDVLY	31%
ϕ Bcep1	NP_944374	gp66	655	DFE	NA.PFD	YALID	TARGAG	DLSGIEAR	RQVGKVADLALGFG	EIVHHVYDEALL	27%
ϕ Bcep781	NP_705685	gp57	655	DFE	NA.PFD	YALID	TARGAG	DVPKLEAF	RQVGKAADLALGFG	EIVHHVYDEALL	26%
ϕ Bcep43	NP_958166	gp61	656	DFE	NA.PFD	YALID	TARGAG	DVPKLEAF	RQVGKAADLALGFG	EIVHHIYDEALL	25%
ϕ D29	NP_046860	gp44	607	DTE	NA.SYD	YAGMD	TSRMSI	DYQAQELR	RKYAKTVNFGRVYG	YLRLPIHDEIVA	[213-607]
ϕ L5	NP_039708	L5p41	595	DTE	NA.SFD	YAGMD	TSRMSI	DYQAQELR	RKVGKTANFQKVYG	YLRLPIHDEIVA	71%
ϕ Bxz2	NP_817633	gp44	604	DTE	NA.SYD	YSGMD	TARMSI	DYQAQELR	RKYAKTVNFGRVYG	YLRLPIHDEILA	71%
ϕ Bxb1	NP_075308	gp41	608	DTE	NA.AFD	YAGMD	TARMST	DYQAQELR	RKVGKMVNFAYVFG	MIRLVIHDEVLA	52%
ϕ BT1	NP_813726	gp11	624	DTE	NA.MFD	YAGLD	TGRMSI	DFQAIEMR	RKVFKGAGFGKVYG	YMKLPIHDEIVF	42%
ϕ C31	NP_047956	gp11	617	DTE	NA.PFD	YAGLD	TGRMSI	DFQAIEMR	RKVFKGAGFGKVYG	YMKLPIHDEIVF	42%
ϕ Rosebush	NP_817817	gp56	609	DIE	NA.SFD	GAMAD	TGRMSY	DWAQIEPV	RPMAKVILLATMYG	HVQLAMHDELVV	32%
ϕ PG1	NP_943839	gp61	619	DIE	NA.PFD	YEGMD	TGRMSY	DWSQIEPV	RPTSKVVLLSSMYG	ELYLAMHDEVVV	28%
pϕ BC6A51	NP_831637	bc1864	802	DIE	NL.SFE	YSAED	TGRLNS	DFSGFELR	RTDAKAGNFGISYG	DMIAQIHDEIIF	27%
ϕ T5	P19822	-	829	DSE	NL.KFD	YAAKD	SGRLSS	DLTTAEVY	RQAAKAITFGILYG	KIVMLVHDSVVA	[301-829]
ϕ SPO1	AAA03732	gene 31	924	DLE	NG.KFD	YTLLD	TGRLSS	DYSQLELR	RTASKKIQFGIVYQ	RICITVHDSIVL	27%
ϕ VP5 / ϕ VP2	AAR92073	-	633	DTE	NA.SFD	YACPD	TGPALQ	DMASFEVR	QANAKQMNLGMIFN	SMILNTHDSYSM	23%
ϕ VpV262	NP_640280	p19	661	DLE	NA.KFD	LKLRD	THRLSS	DGAQIEFR	RTDAKPDTFKPLYG	FLVNTVHDSVIS	–
ϕ Xp10	NP_858986	p39	794	DLE	NI.KFD	GTGED	TGRLSS	DFTALEIY	RSKAKGFSFQRAYG	LIVNTVHDAAYI	31%
ϕ KMV	NP_877458	gp19	807	DLE	NLDGVD	GPCGD	TTRLSS	DYTALEVV	RKNIKPKAFSAQYG	FLINNVHDAVYT	26%
ϕ 1 (*V. harveyi*)	AAL85287	-	786	DLE	NF.KLD	YAGGD	TGRLAA	DFSQGELR	RNYGKPQNFGLLYG	KFMAMIHDALVS	23%

Initial BLAST searches with *E. coli* Pol I (*polA*) as query were refined in subsequent rounds: i. with parts of the Pol I sequence as query, ii. with *B. subtilis* Pol I sequence as query, iii. with known phage DNA polymerases as query (e.g. ϕT7, ϕT5), iv. with the least similar matches of preceding rounds as query, and v. by eye-scanning completely sequenced phage genomes for conserved motifs. The motifs ExoI, ExoII, and ExoIII were taken from Bernad *et al.* [28]. Polymerase domain motifs A - C for the Pol I family were taken from Joyce + Steitz [201] and [337], which correspond to motifs 3 - 5 of Delarue *et al.* [99]. The motifs 1 + 2 of Delarue *et al.* were omitted, and from the variation of motif 2 by Blanco *et al.* [38] only motif 2b included because motifs 1 + 2a could not be unambiguously identified in many phage DNA polymerase sequences. Bracketing indicates the proteins compared for similarity (% ident. res.) with ϕT7 gene 5, ϕ12 p12, ϕD29 gp44, and ϕT5 DNA polymerase, respectively. Conserved individual residues within the motifs are boxed. The micro-duplication in motif C of ϕVp16T orf38T is indicated by underlining. Bold letters indicate the functionally important residues (see text).

within the signature motifs were not detected (see Table C25). The spacing of the signature motifs in the DNA polymerases of phages ϕXp10 and ϕKMV suggests that these proteins contain an additional domain, in analogy to the TBD of ϕT7 gene 5 DNA polymerase. For these two proteins, however, the function of this interdomain-region is not known.

A rather unusual composition of functional domains is found in the ϕSPO1 gene 31 DNA polymerase: a stretch of ~200 residues N-terminal to the 3'→5' exonuclease and polymerase domains shows significant similarity with bacterial uracil-DNA glycosylases, e.g. 43% ident. res. over almost the entire length with *Thermus thermophilus* Hb8 UDG (205 res.). ϕSPO1 – together with Φe, ϕHI, ϕ2C, ϕSP8, and ϕSP82 – belongs to a group of large *B. subtilis* phages (genome size ~150 kb), which contain 5-(hydroxymethyl)-2'-deoxyuridine instead of thymine in their linear dsDNA genomes [177]. It is not known, whether the UDG-like domain of ϕSPO1 has UDG activity, and whether such a function has any relevance for the replication of hmUra containing DNA. Alternatively, an UDG-like polypeptide could be sim-ply employed as processivity factor for the polymerase by making use of its DNA-binding property. BLAST searches with ϕSPO1 gene 31 as query gave various matches in the NCBI database with proteins showing ~35% ident. res. confined to the N-terminal UDG-like domain (res. 1- ~200) . Matching proteins were <250 res. in size and their entries assigned 'DNA polymerase, bacteriophage-type'; such falsely assigned proteins include *Archaeoglobus fulgidus* DSM 4304 NP_071102, *Pyrococcus furiosus* DSM 3638 NP_579114, *Halobacterium* sp. NRC-1 NP_280750, *B. cereus* ATCC14579 NP_831308, and *Clostridium tetani* E88 NP_780781.

Phage D29 gp44-type subfamily. This group of nine (putative) DNA polymerases is characterised by a moderate to high similarity to Mycobacteriophage ϕD29 gp44 DNA polymerase. None of these proteins has yet been characterised biochemically. The proteins have a rather uniform size (~600 res.) and motif spacing, and several conserved residues within the signature motifs of the polymerase domain are specific for this subfamily. Only the DNA polymerase of ϕBC6A51 has a somewhat larger spacing between motifs ExoIII and 2a, but a BLAST search with this region only matched with the 3'→5' exonuclease–polymerase interdomain region of various bacterial Pol I homologues. It is therefore questionable whether this region has a special function.

Phage 12 p12-type subfamily. None of the 13 proteins of this group has been characterised biochemically. This group of (putative) DNA polymerases is very homogeneous with respect to size, sequence similarity (~40% ident. res.) and spacing of the conserved signature motifs (including conserved residues within the signature motifs). This is somewhat surprising because they are encoded by a great variety of phages of Gram-negative as well as Gram-positive hosts. None of the proteins has an N- or C-terminal extension, and also the interdomain region separating the 3'→5' exonuclease from the polymerase domain has the same length found in *E. coli* Pol I. There are therefore no hints how these (putative) DNA polymerases gain processivity. The extremely wide spacing of polymerase signature motifs B and C in the ϕAPSE-1 P45 DNA polymerase sequence (accession NP_051006) is artefactual, an intron sequence has not been removed during 'translation' of the ϕAPSE-1 genome sequence into accessible protein database entries.

The (putative) orf39C and orf38T DNA polymerases of *Vibrio* sp. phages ϕVp16C and ϕVp16T, respectively, deviate both from the canonical pattern. ϕVp16C orf39C DNA polymerase lacks the ExoI motif due to a truncated N-terminus: homology (91% ident. res.) between orf39C and orf38T begins with res. 91 of the latter. A simple sequencing and reading-frame assignment error can be excluded because the DNA sequence upstream of the ϕVp16C orf39C gene bears no similarity with the 5' region of the ϕVp16T orf38T gene [402]. The essential acidic residues in the polymerase motif C of ϕVp16T orf38T DNA polymerase cannot be unambiguously assigned because the region apparently contains a micro-duplication. Also the spacing of motifs B and C differs considerably from that observed in the orf39C and other proteins of this subfamily. The functionality of both DNA polymerases needs to be confirmed by biochemical analysis.

***E. coli* Pol II (*polB*)-type DNA polymerases.** DNA polymerase II of *E. coli* is involved in excision repair, translesion synthesis, and replication re-start, and encoded by the SOS-inducible *polB* gene (formerly *dinA*) [40,193,361]. The evolutionary origin of the *E. coli polB* gene is still subject to debate, but it does certainly not belong to the standard equipment of a bacterial genome, because orthologues are exclusively found in the genomes of several β- and γ-Proteobacteria. [217,122]. Structurally, Pol II is composed of an N-terminal 3'→5' exonuclease domain and a C-terminal polymerase domain. *E. coli* Pol II has a region of unknown function (~150 re.) preceding the 3'→5' exonuclease domain, and which is only found in the PolB orthologues. The C-terminal pentapeptide QLGLF was inferred to be involved in the interaction of Pol II with the β-clamp [90,262]. The three-dimensional structure of Pol II is not known, but all signature motifs characteristic for the Polα-family are readily detected (Table C26) [488,201]. Two groups of phage-encoded DNA polymerases share the

Table C26 Conserved motifs in the 3'→5' exonuclease and polymerase domains of *E. coli* Pol II-type polymerases.

domains				3'→ 5' exo domain			polymerase domain			
motifs:				ExoI	ExoII	ExoIII	A II	B III	C I	
consensus				**D**x**E**	Nx_{2-3}F**D**	**Y**xxx**D**	**D**hxSLYPS	**K**hxxNShYG	hYG**D**T**D**Shhh	ident. res.
replicon	accession	gene	res.							
E. coli	P21189	*polB*	783	**D**F**E**	NVVQF**D**	YNLK**D**	**D**YKSLYPS	**K**IIMNAFYG	IYG**D**T**D**STFV	
ϕ 29 / ϕ PZA	NP_040719	p2	572	**D**F**E**	NL.KF**D**	YIKN**D**	**D**VNSLYPA	**K**LMLNSLYG	IYC**D**T**D**SIHL	[1-572]
ϕ B103	Q37882	gp2	572	**D**F**E**	NL.KF**D**	YIKN**D**	**D**VNSLYPS	**K**LMFDSLYG	IYC**D**T**D**SIHL	81%
ϕ M2	P19894	G	572	**D**F**E**	NL.KF**D**	YIKN**D**	**D**VNSLYPS	**K**LMLNSLYG	IYC**D**T**D**SIHL	81%
ϕ GA-1	NP_073685	gene 2	578	**D**F**E**	NL.KF**D**	YLKH**D**	**D**VNSMYPA	**K**LMLNSLYG	IYA**D**T**D**SIHV	52%
ϕ Cp-1	Q37989	gene 5	568	**D**F**E**	NL.KF**D**	YIHV**D**	**D**INSMYPA	**K**IMLNSLYG	LYA**D**T**D**SLHL	33%
ϕ PRD1	P10479	gene I	553	**D**F**E**	NGGKF**D**	YLKG**D**	**D**VNSMYPH	**K**LILNSSYG	LYC**D**T**D**SIIC	24%
ϕ Bam35c	NP_943751	orf5	735	**D**T**E**	NL.DF**D**	YMEY**D**	**D**KNSLYPY	**K**LMQNALYG	AYC**D**T**D**SCAT	24%
ϕ 44AHJD / ϕ P68	NP_817305	orf10	755	**D**I**E**	NCNKY**D**	YIHN**D**	**D**INSSYPY	**K**VVLNGLYG	IYC**D**T**D**SLYM	21%
ϕ P1 [1])	NP_064636	P82	694	**D**F**E**	NSKNF**D**	YFR.**D**	**D**INSAYPY	**K**IKLNSLYG	IYG**D**T**D**SIMF	<20%
ϕ C1	NP_852013	orf7	784	**D**I**E**	NGAKY**D**	YEMV**D**	**D**LNSSYPT	**K**HYSKLFYG	WYA**D**T**D**SLYM	<20%
ϕ T4	P04415	gp43	898	**D**I**E**	NIEGF**D**	YNII**D**	**D**LTSLYPS	**K**ILINSLYG	AAG**D**T**D**SVYV	[1-898]
ϕ RB69	Q38087	gp43	903	**D**I**E**	NVESF**D**	YNII**D**	**D**LTSLYPS	**K**LLINSLYG	LYG**D**T**D**SIYV	61%
ϕ RB49	NP_891597	gp43	892	**D**I**E**	NTETF**D**	YNIG**D**	**D**LTSLYPS	**K**VLINSLYG	RYC**D**T**D**SVYV	55%
ϕ KVP40	NP_899330	gene 43	850	**D**I**E**	NIESF**D**	YQIQ**D**	**D**LTSLYPS	**K**ILINSLYG	IYG**D**T**D**SIYV	42%
ϕ Aeh1	NP_943895	gp43	919	**D**I**E**	NSEMF**D**	YLIR**D**	**D**LTSLYPH	**K**VLINSLYG	VYG**D**T**D**SIYL	38%
ϕ 25	AAP68686 AAP68688	gene 43A gene 43B	889	**D**I**E**	NSEGF**D**	YCVR**D**	**D**LTSLYPS	**K**ILINSLYG	CYI**D**T**D**SVYL	51%
ϕ 44RR2.8t	NP_932391 NP_932390	gp43A gp43B	889	**D**I**E**	NSEGF**D**	YCVR**D**	**D**LTSLYPS	**K**ILINSLYG	CYI**D**T**D**SVYL	50%
ϕ 65	AAR90916 AAR90917	gp43A gp43B	897	**D**I**E**	NSEGF**D**	YSIR**D**	**D**LTSLYPH	**K**VLINSLYG	IYC**D**T**D**SQYL	36%
ϕ RM378	NP_835611	p024	313	**D**I**E**	NV.IS**D**	YNAV**D**	-	-	-	–
ϕ RM378	NP_835679	p092	522	-	-	-	**D**FTSLYPS	**K**IMMNSMYG	IYSHT**D**SIFV	–

The motifs ExoI, ExoII, and ExoIII were taken from Bernad *et al.* [28]. Polymerase domain motifs A - C for the Polα family were taken from Joyce + Steitz [201], which correspond to motifs I - III of Wong *et al.* [488]. Bracketing indicates the proteins compared for similarity (% ident. res.) with ϕT4 gp43, and ϕ29 p2, respectively. Bold letters indicate the functionally important residues (see text). 1) *Mycoplasma* sp. ϕ P1 must not be mistaken as *E. coli* ϕ P1.

Table C27 Conserved motifs in the polymerase domain of *E. coli* Pol IIIα-type polymerases (DnaE); including the cognate 3'→5' exonuclease, DnaQ.

domain				3'→ 5' exo domain			polymerase		
motifs:				ExoI	ExoII	ExoIIIε	C	A	
consensus				**D**x**E**	Nx_{2-3}F**D**	Hx_4**D**	hP**D**h**D**hDh	hhKh**D**hLG	ident. res.
replicon	accession	gene	res.						
E. coli	P10443	*dnaE*	1160	-	-	-	MP**D**F**D**VDF	LVKF**D**FLG	[1-1160]
B. subtilis	O34623	*dnaE*	1115	-	-	-	MP**D**I**D**IDF	LLKM**D**FLG	37%
E. coli	P03007	*dnaQ*	243	**D**T**E**	NA.AF**D**	HGALL**D**	-	-	-
B. subtilis	P13267	*polC*	1437	**D**V**E**	NA.SF**D**	HRAIY**D**	VP**D**I**D**LNF	LLKL**D**ILG	23%
B. subtilis pϕ SPBc2	NP_046685	*yorL*	1305	-	-	-	LP**D**I**D**IDT	SVKF**D**LLT	22%
ϕ PIS136	AAL66178	DP1	1206	-	-	-	PP**D**I**D**IDF	LLKI**D**ILG	37%
ϕ Bxz1	NP_818250	gp199	1131	-	-	-	FP**D**I**D**VDF	GVKD**D**LLA	26%
ϕ Barnyard	NP_818618	gp80	1111	-	-	-	YP**D**I**D**LDF	MLKV**D**FLG	26%

The motifs ExoI and ExoII were taken from Bernad *et al.* [28]; motif ExoIIIε from Moser *et al.* [304] and deRose *et al.* [100]. Polymerase domain motifs C and A were taken from Pritchard and McHenry [355]. Bracketing indicates the proteins compared for similarity (% ident. res.) with *E. coli* Pol IIIα (*dnaE*). Bold letters indicate the functionally important residues (see text).

conserved signature motifs with *E.coli* Pol II despite the lack of significant sequence similarity: the group of ϕ29 p2-like DNA polymerases and the group ϕT4 gp43-like DNA polymerases (Table C26).

In marked contrast to *E. coli* Pol II and ϕT4 gp43 DNA polymerase (see below), the p2 DNA polymerase of *B. subtilis* phage ϕ29 is highly processive and exhibits a strong strand-displacement activity [372]. The latter property renders the replication of the 19.4 kb long dsDNA genome of ϕ29 independent of a helicase. ϕ29 DNA polymerase acts as monomer since the strand-displacement mechanism operating during ϕ29 replication does not require a coupling of leading- and lagging-strand DNA synthesis. For the details of protein-primed replication and ϕ29-type DNA polymerases we refer the reader to the recent review by Meijer, Salas and co-workers, and primary literature cited therein [287].

The DNA polymerases of several Gram(+)- specific phages and the Gram(−)-specific phage ϕPRD1 are similar in primary sequence, and of comparable size, as are the respective phage genomes [340]. It seems very likely, therefore, that also the biochemical properties of these rather compact DNA polymerases are very similar. This makes this group of enzymes excellent candidates for the elucidation of the molecular basis for inherent processivity and strand-displacement capability, not complicated by a coupling of leading- and lagging-strand DNA synthesis as in dimeric DNA polymerase holoenzymes.

The DNA polymerases of phages ϕBam35C, ϕ44AHJD/ ϕ P68, and ϕ C1 are considerably larger than ϕ29 p2 DNA polymerase. In all three proteins, the 3'→5' exonuclease domain is separated from the polymerase domain by a ~200 res. long region of unknown function. In analogy to the structure of ϕT7 gene 5 DNA polymerase this region might be required for binding of an (unknown) processivity factor. Alternatively it could be simply a flexible hinge region: BLAST searches with this region as query revealed limited similarity to putative hinge regions of several bi-modular transcription factors.

Pol II-type DNA polymerases involved in protein-primed replication are not confined to phage replicons: the linear mitochondrial plasmid pClK1 (6.7 kb) of the phytopathogenic fungus *Claviceps purpurea* codes for protein with C-terminal similarity (31% ident. res.; ~200 res.) to orf7 DNA polymerase of ϕC1. Interestingly, pC1K1 replicates via protein priming like ϕ29 [327].

Eight highly similar phage-encoded DNA polymerases make up the group of ϕT4 gp43-type DNA polymerases (Table C26). The crystal structure of the ϕRB69 gp43 protein has been solved [PDB 1CLQ], also in complex with the cognate clamp (see below) [407]. Thus this DNA polymerase is among the best understood structurally, and provides the prototype for the *E. coli* Pol II-type DNA polymerases [203]. Within the gp43 structure, the conserved signature motifs were found at positions indicative of their importance for protein function. Benkovic and co-workers could demonstrate the interaction of the ϕT4 gp43 C-terminus with the gp45 clamp [3]. The sequence of the extended C-terminal domain involved in the interaction with the cognate clamp is highly specific for this group of DNA polymerases. A coupling of leading- and lagging-strand synthesis by ϕT4 gp43 DNA polymerase is achieved by a dimeric 'holoenzyme', and a direct interaction of the two polymerase subunits could already be shown to involve a region overlapping the polymerase domain [381]. The details of the ϕT4 replisome were recently reviewed by Benkovic and co-workers [27].

To our knowledge, the (putative) p092 DNA polymerase of ϕRM378 is the only *E. coli* Pol II-type DNA polymerase lacking an N-terminal 3'→5' exonuclease domain – a key feature of Pol IIIα-type DNA polymerases (see below). However, a 3'→5' exonuclease is encoded separately in the ϕRM378 genome. The 3 polymerase signature motifs of p092 are located within the N-terminal half of the protein. It is not known whether the extended C-terminus (~200 res.) of p092 is necessary to achieve the topologically required orientation of the exonuclease and polymerase subunits within the DNA polymerase holoenzyme, or whether it is involved in contacting a yet unknown sliding clamp, as in ϕT4 gp43- type DNA polymerases.

***E. coli* Pol IIIα (*dnaE*)-type DNA polymerases.** BLAST searches with *E. coli* Pol IIIα (*dnaE*) as query led unexpectedly to the detection of four phage-encoded homologues (Table C27). The three-dimensional structure of the α-subunit of *E. coli* Pol III holoenzyme is not known but Pritchard and McHenry could identify functionally important residues by biochemical analysis of mutant proteins, and thus derive 2 signature motifs [355]. Despite their low overall similarity with Pol IIIα, the two signature motifs are well conserved and readily found at corresponding positions in all four phage-encoded proteins, which we consider *bona fide* DNA polymerases, therefore. Lacking an N-terminal 3'→5' exonuclease domain, all four DNA polymerases are structurally more related to DnaE than to PolC of *B. subtilis* (see Table C27). All four phages encoding Pol IIIα-like DNA polymerases infect Gram(+) hosts: *B. subtilis*, and three species of the Actinobacteria branch (high G+C Gram(+)-bacteria). ϕSPBc2 is also found as curable prophage in the genome sequence of *B. subtilis* 168 [491]. The genome sequences of the mycobacterial phages ϕBarnyard and ϕBxz1 were determined just recently [342]. Therefore we expect that more phage-encoded DNA polymerases of the *E. coli* Pol IIIα- type will be found in the future, adding this type of DNA polymera

Table C28 Conserved motifs in the phage-encoded 5'→3' exonucleases.

motifs:				A	B	C	D	E
consensus				hhhh**D**G	hhhh**D**x	hRxExxxxxxY**K**xxR	EA**D**DhhGxh	hhTx**D**KD
replicon	accession	gene	res.					
E. coli	P00582	*polA*	928	LILV**D**G	AVVF**D**A	FRDE..LFEHY**K**SHR	EA**D**DVIGTL	ISTG**D**KD
E. coli	P38506	*exo*	281	LLIV**D**A	VAVF**D**D	WRHQ..RLPDY**K**AGR	EA**D**DLAATL	IVST**D**KG
A. aeolicus	NP_214115	*pol*	289	LYIL**D**G	VVVF**D**A	TKRE.KIYADY**K**KQR	EA**D**DVIAYL	IYSP**D**KD
ϕ T4	NP_049859	*rnh*	305	ICLI**D**F	VLCI**D**N	WRRD..FAYYY**K**KNR	EA**D**DHIAVL	IISS**D**GD
ϕ RB69	NP_861940	*rnh*	290	IALA**D**F	VLCM**D**N	WRRD..FAYYY**K**KNR	EA**D**DHIGVL	IVAS**D**GD
ϕ RB49	NP_891817	*rnh*	315	VHLI**D**M	VLAF**D**S	RRDI...APYY**K**RNK	EA**D**DIIAVL	INSG**D**GD
ϕ KVP40	NP_899249	*rnh*	310	TKSF**D**G	LLTV**D**L	LRSN.VLKNKI**K**YPR	EA**D**DHIGVL	ITSS**D**GD
ϕ Aeh1	NP_944210	*rnh*	306	FCIF**D**F	IIAV**D**N	WRRQ..KYYFY**K**KHR	EA**D**DIIGVL	IVSS**D**GD
ϕ 44RR2.8t	NP_932573	*rnh*	307	VNII**D**A	IIAF**D**D	WRRD..LAWYY**K**KNR	EA**D**DIIAIL	ISSS**D**SD
ϕ T7	P00638	gene 6	300	ILVM**D**G	LAFT**D**S	WRKE.LVDPNY**K**TNR	EG**D**DVMGVI	IISC**D**KD
ϕ A1122	NP_848287	p25	300	ILVM**D**G	LAFT**D**S	WRKE.LVDPNY**K**ANR	EG**D**DVMGVI	IISC**D**KD
ϕ T3	P20321	gene 6	302	ILVM**D**G	LAFT**D**S	WRKE.LVDPNY**K**ANR	EG**D**DVMGVI	IISC**D**KD
ϕ YeO3-12	NP_052100	gene 6	303	VLVM**D**G	LAFT**D**D	WRKV.LVDETY**K**ENR	EG**D**DVMGII	LVSC**D**KD
ϕ gh-1	NP_813767	p21	314	ALDM**D**Y	ILSG**D**D	WRKE..VLETY**K**ANR	EG**D**DVCGIL	SVSC**D**KD
ϕ P60	NP_570332	P60_22	243	TLLI**D**A	LTFT**D**S	FRKD..VEPTY**K**GNR	EA**D**DVMGIL	LISP**D**KD
ϕ PaP3	NP_775229	p28	294	IAGI**D**G	LTGK**D**N	FRLELATIRKY**K**GTR	EA**D**DWLGVR	ACSR**D**KD
ϕ Felix01	NP_944975	p196	348	HVFI**D**S	GLTF**D**E	WERQ..TWKEA**K**SEK	EA**D**SIVIAK	LMSI**D**KD
ϕ Sp6	NP_853581	gp21	342	IALI**D**G	GAEC**D**A	FRLRLAFTKPY**K**GTR	EA**D**DLMSIA	IVSA**D**KD
ϕ K1-5	AAR90062	gp 20	342	IALI**D**A	AAEC**D**A	FRVRLAFTKPY**K**GQR	EA**D**DLMSIA	IVSL**D**KD
P. putida pϕ	NP_744425	PP2276	199	[QNLI**D**K]	LVFS**D**K	WRKT..VLPTY**K**HNR	EG**D**DVCGIL	VASI**D**KD
ϕ T5	P06229	D15	291	LMIV**D**G	IVLG**D**K	FRLE..HLPEY**K**GNR	EA**D**DMAAYI	LIST**D**GD
ϕ Xp10	NP_858984	p37	309	PLHV**D**G	AAGA**D**R	VHLS..ASNCT**K**GHR	EA**D**DGIAYC	VMSA**D**KD
ϕ VpV262	NP_640291	p30	323	RSLV**D**A	VADV**D**I	WRNEDATIQKY**K**GCR	EA**D**DSIAAT	VFSG**D**KD
ϕ KMV	NP_877461	orf22	313	TLVC**D**A	TRTL**D**T	TRVHLTAAGGA**K**AYR	EA**D**DGMMMD	IRSD**D**KD
ϕ RM378	NP_835599	p012	318	VNLI**D**L	GMII**D**D	LRKK..LLPQY**K**EHR	EA**D**DVIAHL	IVST**D**KD

Motifs A - E correspond to the proposal of Gutman and Minton [156] except that motif F was ommitted because it does not contain residues known to be involved in catalysis or DNA binding (see text). Bold letters indicate the functionally important residues (see text). Motif A of *P. putida* pϕ protein PP2276 is shown in parentheses because it is irregularly spaced to motif B.

ses to the well-known types of phage-encoded DNA polymerases of the Pol I- and Pol II-type.

None of the four phage-encoded Pol IIIα-like DNA polymerases has yet been analysed experimentally. It is therefore only possible to speculate about their points of conflict with or access to the replication machinery of their hosts. Some hints may be obtained from the present knowledge about the replicative DNA polymerase of *B. subtilis*, which also has not yet been thoroughly characterised. Since its early detection, the *B. subtilis* PolC protein was considered the functional equivalent of the α-subunit (*dnaE*) of *E. coli* Pol III holoenzyme [24]. In contrast to *E. coli* Pol IIIα, the $PolC_{Bsu}$ polymerase contains an N-terminal 3'→5' exonuclease domain that belongs to the 'dnaQ-subfamily' of exonucleases (see below) [21]. A homologue of the E. coli *dnaQ* gene is not found in the genome of *B. subtilis*. In addition to *polC*, a homologue of the *E. coli dnaE* gene is found in the *B. subtilis* genome, and Ehrlich and co-workers could show that both polymerases are essential for chromosome replication [101]. These authors also presented evidence suggesting that both polymerases form part of an asymmetric *B. subtilis* replisome, and that $DnaE_{Bsu}$ is responsible for lagging-strand synthesis. We hypothesise that the phage-encoded Pol IIIα- type DNA polymerases can replace the DnaE subunit in the asymmetric DNA polymerase holoenzyme of the host, thus linking the propagation of their cognate replicon to the host replication machinery.

Table C29 Conserved motifs in the phage-encoded 3'→5' exonucleases.

motifs:				ExoI	ExoII*	ExoIII	ExoIIIε
consensus				**D**x**E**	Nx_{2-4}F**D** Y	Yxxx**D**	Hxxxx**D**
replicon	accession	gene	res.				
E. coli	P00582	*polA*	928	**DTE**	NL..KY**D**	YAAE**D**	
E. coli	P03007	*dnaQ*	243	**DTE**	NA..AF**D**		HGALL**D**
E. coli	NP_416515	*sbcB*	475	**DYE**	NV..RF**D**		HDAMA**D**
ϕ T4	NP_049629	*dexA*	227	**DFE**	NI.APS**D**		HDCAK**D**
ϕ RB69	NP_861704	*dexA*	225	**DFE**	NLAPSE**D**		HDCAK**D**
ϕ RB49	NP_891579	*dexA*	222	**DFE**	[CEAVGV**D**]		HDAIH**D**
ϕ KVP40	NP_899271	*dexA*	230	**DYE**	LIPSAE**D**		HDCAK**D**
ϕ Aeh1	NP_943887	*dexA*	226	**DME**	LKPSMN**D**		HDCAK**D**
ϕ 44RR2.8t	NP_932376	*dexA*	221	**DWE**	NL.KPS**D**		HDCAKA
ϕ 65	AAR90930	*dexA*	219	**DFE**	LLLKPS**D**		HDCAR**D**
ϕ 12	NP_803310	p04	310	**DFE**	NA..LF**D**		HSAKF**D**
ϕ ST64B	NP_700403	sb30	189	**DLE**	NG.ANF**D**		HNALA**D**
ϕ PY54	NP_892096	p50	190	**DLE**	NG.ASF**D**		HNALA**D**
ϕ 186	NP_052286	orf81	194	**DTE**	NA..DY**D**		HRALA**D**
ϕ PaP3	NP_775218	p39	175	**DLE**	NI.LDY**D**	YCLQ**D**	
ϕ Xp10	NP_858982	p35	245	[**DD**I]	NG.KKF**D**	YNVV**D**	
ϕ KMV	NP_877463	orf24	348	**DIE**	NG.KRF**D**	YNID**D**	
ϕ RM378	NP_835611	p024	313	**DIE**	NV..IS**D**	YNAV**D**	
ϕ CJW1	NP_817562	gp115	261	**DIE**	NG.DRF**D**	YNIH**D**	
ϕ VWB	NP_958252	p10	263	**DLE**	NV..PY**D**	FHEL**D**	
ϕ Che8	NP_817375	gp37	273	**DVE**	NLLSQA**D**	YCAG**D**	

The ExoII motif Nx_{2-3}(F/Y)D of Bernad *et al.* [28] was changed to ExoII* motif Nx_{3-4}(F/Y)D to allow the alignment of the ϕRB69 *dexA* sequence. The motif ExoIIIε was taken from Moser *et al.* [304] and deRose *et al.* [100]. Bold letters indicate the functionally important residues (see text). The ExoI motif is shown in parentheses for the ϕXp10 p35 protein because it deviates from the consensus, it is however regularly spaced to motif ExoII.

***E. coli* Pol IV (*dinB*)-type and *E. coli* Pol V (*umuC*)-type DNA polymerases.** These two polymerases belong to the 'Y class' of errorprone DNA polymerases involved in translesion synthesis in *E. coli* and possibly also in stationary-phase induced mutagenesis [331]. They have homologues in *B. subtilis* (*yqjH*, *yqjW*) and many other bacterial genomes from different phyla [147,439].

BLAST searches identified the YolE protein of ϕSPBc2 as the only phage protein with significant similarity to *E. coli* DinB and UmuC. Because ϕSPB2c encodes the YorL DNA polymerase (see above), the involvement of YolE in replication of this phage under 'normal' conditions seems unlikely. Interestingly, in contrast to their virtually complete absence from phage replicons, genes encoding proteins with significant similarity to to *E. coli* DinB and UmuC polymerases are quite common among plasmid replicons from various incompatibility groups, e.g. IncL/M plasmid R471a, IncT plasmid R394, IncN plasmid R46, IncJ plasmid R391 (see also [344]).

5'→3' exonucleases. A 5'→3' exonucleolytic or RNase H activity is necessary for the removal of RNA primers after replication of dsDNA, prior to strand sealing by DNA ligase. Because this type of enzymatic activity requires a 5'-end for recognition but cuts DNA near the junction of the single- and double-stranded DNA it has been termed 'structure specific' 5'→3' exonuclease [201]. The role of this specific exonuclease clearly differs from that of the 5'→3' exonuclease required to create single-stranded 3'-OH overhangs required for recombination although the proteins may be structurally related (Section C3.6.2.).

The 5'→3' exonucleolytic activity of *E. coli* Pol I is located within the N-terminal domain (res. 1-323), and fully retained in a truncated Pol I protein (*resA1*;

Q298→amber), showing that the isolated domain is structurally and functionally intact [200]. This N-terminal domain is found in all 124 known bacterial Pol I (*polA*) orthologues, but in none of the 40 phage-encoded DNA polymerases of the *E. coli* Pol I-type (see above). However, several phages that encode DNA polymerases also encode a protein with limited similarity to the *E. coli* Pol I N-terminus, and *E. coli* exonuclease IX. [46].

Six particularly well conserved sequence motifs A–F were found by Gutman and Minton by alignment of ten 5'→3' exonuclease sequences, including 4 phage-encoded proteins [156]. Since then, the crystal structures of the ϕT5 D15 protein (PDB 1XO1) and ϕT4 RNase H (PDB 1TFR) were solved [65,308]. The results of mutation analysis in combination with both crystal structures led to the identification of a conserved lysine residue involved in DNA binding (motif C), and 4 aspartate residues essential for Mg^{2+}- binding, and thus catalysis (motifs A, B, D, E) [31,135]. The essential residues determined by Nossal and co-workers for ϕT4 RNase H are perfectly conserved in 21 phage- encoded genes, despite the lack of significant sequence similarity (>25% ident. res.) (Table C28) [31]. The region of similarity between the 21 proteins is in most cases confined to a stretch of ~50 residues encompassing motifs D and E. However, the spacing of the motifs is fairly uniform in all protein sequences. We infer from the latter and the perfect conservation of the important residues that the genes have been correctly assigned. Motif A is set in parentheses for the *P. putida* prophage locus PP2276 (Table C28) because the spacing to motif B is unusually short, and the protein might in fact be truncated at the N-terminus and non-functional. The phage-encoded exonucleases are found exclusively in three groups of phages: i. all phages encoding DNA polymerases of the ϕT4-subgroup of the *E. coli* Pol B-type encode a 5'→3' exonuclease/RNaseH (the complete DNA sequences of the *Aeromonas* sp. phages ϕ25 and ϕ 65 are not yet available); ii. within the ϕT7 subgroup of phages encoding *E. coli* Pol I-type DNA polymerases, a gene encoding a 5'→3' exonuclease/RNaseH with similarity to one of the proteins listed in Table C28 is only absent in the genome of phage ϕSIO1; iii. all phages of the heterogeneous ϕT5-subgroup encode a 5'→3' exonuclease (the ϕSPO1 genome has not yet been sequenced). Except for the functionally important residues, no other conserved group-specific residues could be detected.

13 bacterial genomes characterised by significant genome size reduction lack a *polA* gene (several *Mycoplasma* sp., *Ureaplasma* sp., *Wigglesworthia* sp., *Tropheryma* sp., *Buchnera* sp.). Nevertheless, all encode a protein similar to *E. coli* exoIX. Exo IX genes are, on the other hand, not pres??ent in all genomes encoding a *polA* orthologue. This observation emphasises the importance of the 5'→3' exonuclease/ RNaseH enzyme for the replication of bacterial genomes. A unique situation is found in the genome of *A. aeolicus*: the *pol* gene (locus aq_1628; falsly assigned 'DNA polymerase I 3'-5' exo domain') encodes a 5'→3' exonuclease/RNaseH enzyme (Table C28), the *polA* gene (locus aq_1967) encodes a ϕT7-subfamily DNA polymerase, i.e. lacking the N-terminal 5'→3' exonuclease/RNaseH domain of *E. coli* Pol I. However, the *A. aeolicus* PolA protein shares ~30% ident. res. with proteobacterial Pol I proteins and has no detectable similarity with any DNA polymerase of the ϕT7-subfamily. The origin of the *A. aeolicus polA* gene remains enigmatic, therefore.

3'→5' exonucleases. The 3'→5' exonucleolytic activity required for proofreading, i.e. the removal of misincorporated bases in a growing DNA strand, is contributed by the ε-subunit (*dnaQ*) in the *E. coli* DNA Pol III holoenzyme. Historically, the *dnaQ49* mutation was shown to result in mutator phenotype and led to the identification of the ε subunit of *E. coli* DNA Pol III [393,348]. In contrast, the proofreading activity resides within the N-terminal domain of all known phage-encoded DNA polymerases of the *E. coli* Pol I- or Pol II-type, with the exceptions of the ϕRM378 DNA polymerase. It is therefore questionable whether the 3'→5' exonucleases encoded by phage replicons are important for their replication, or involved in recombination rather (Section C3.6.2.).

Mutational analysis of the *B. subtilis* PolC protein and the NMR analysis of the *E. coli* Pol IIIε (*dnaQ*) protein suggest a significant structural similarity with the *E. coli* Pol I 3'→5' exonuclease domain [21,100]. Including the Pol IIIε structure, the alignment of 3'→5' exonucleases led to the definition of a novel conserved motif ExoIIIε, specific for the 'DnaQ subfamily' of exonucleases, which deviates from the canonical ExoIII motif of *E. coli* Pol I- and Pol II-type DNA polymerases [28]. The recombination protein SbcB of *E. coli* also contains a N-terminal domain belonging to this 'DnaQ subfamily' (see Table C29).

The DexA protein of phage ϕT4 belongs to the 'DnaQ subfamily' of exonucleases, but it is not essential for replication under laboratory conditions [137]. DexA might instead be involved in the degradation of the host chromosome and ϕT4-specific recombination [185]. All ϕT4 group phages encode orthologues of DexA (Table C29), and the conservation of the entire set of replication proteins in this group makes it very likely that DexA is also non-essential in the other phages (the complete *Aeromonas* sp. ϕ25 genome sequence is not yet known).

The ϕ12 p04 protein may be the cognate 3'→5' exonuclease of the ϕ12 p12 DNA polymerase. No other phage encoding DNA polymerases of the ϕ12 subfamily

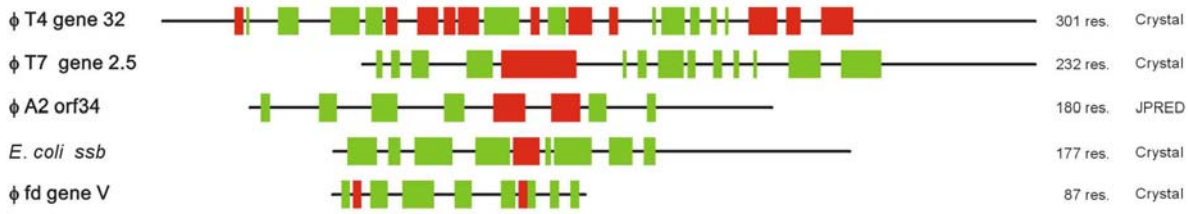

Fig. C19 Secondary structures of SSB proteins.
The secondary structure elements of ϕT4 gene 32 [PDB 1GPC], ϕT7 gene 2.5 [PDB 1JE5], ϕfdgene V [PDB 1YHA], and *E. coli ssb* [PDB 1KAW] were taken from the known crystal structures, and obtained by Jpred prediction for ϕA2 orf34 [88]. Colour code: red = (predicted) α-helical region; green = (predicted) β-stranded region; black line = unstructured/no prediction.

of Pol I-type polymerases encodes in addition a 3'→5' exonuclease, its occurrence in ϕ12 is unique, therefore. The ϕ12 p04 protein is special in a second aspect: it is rather large in size, the 3 signature motifs are found in the N-terminal half. BLAST searches revealed a significant similarity (~30% ident. res.) of its C-terminal half with C-termini of several NAD-dependent DNA ligases, while the N-terminal half gave matches exclusively with proteins of the 'DnaQ subfamily' of exonucleases. Any functional consequences of the apparent composite structure of this protein remain to be studied. Also the phages ϕST64B, ϕPY54, and ϕ186 encode exonucleases of the 'DnaQ subfamily' (Table C29). These phage replicons are devoid of a cognate DNA polymerase and the participation of the sb30, orf181, and p50 proteins, respectively, in their replication unlikely.

The remaining 7 phage-encoded 3'→5' exonucleases belong to the subfamily of Pol I-type 3'→5' exonucleases (Table C29). We infer from the observation that the ϕRM378 p092 polymerase lacks the canonical exonuclease domain a functional role of ϕRM378 p024 exonuclease during replication of this phage. The role of the remaining proteins is unknown at present.

C3.5.2. Sliding clamps, clamp loaders and DNA polymerase accessory proteins

Sliding clamps are processivity factors that tether the replicative DNA polymerase to the DNA template by encircling the DNA duplex with a ring-like structure. The sliding clamp of *E. coli* DNA polymerase III holoenzyme is formed by two DNA polymerase IIIβ-subunits. The β-clamp can attract other *E. coli* DNA polymerases in addition to DNA Pol III [263]. The archaeal and eukaryal PCNA clamps, and the ϕT4 clamp are homo-trimers. The proteins of the *E. coli*-type β-clamp, PCNA, and the ϕT4 gp45 clamp share no sequence similarity. Nevertheless, their three-dimensional ring structures derived from crystals can be readily superimposed, i.e. they are 'structural orthologues' (reviewed in [196]). Fueled by ATP hydrolysis, sliding clamps are assembled around the DNA duplex by clamp loader protein complexes: the $(\gamma/\tau)_3\delta\delta'$ complex in case of *E. coli* DNA polymerase III, the pentameric archaeal and eukaryal replication factor C complex (RFC) [324], and the pentameric ϕT4 gp44·gp62 (4:1) complex [453].

The crystal structures of the ϕT4 sliding clamp protein gp45 (PDB 1CZD) and of the highly related ϕRB69 gp45 protein (78% ident. res.; PDB 1B77, 1B8H) have been solved [407,298]. All phages encoding orthologues of the ϕT4 gp43 DNA polymerases also encode orthologues of the 228 res. long ϕT4 gp45 clamp: ϕRB69, *Aeromonas* sp. ϕ25 and ϕ65, ϕ44RR2.8t, ϕRB49, ϕAeh1, and ϕKVP40. The gp45 proteins vary slightly in length (221-330 res.), and their similarities with the ϕT4 protein range from 27% (ϕKVP40 gp45) to 78% (ϕRB69 gp45). Five monomers of the ϕT4 gp44 protein and one monomer of gp62 form the clamp loader complex [349]. Orthologues of the ϕT4 gp62 and gp44 proteins are encoded by all phages of this group. Since also the helicases and SSBs of the ϕT4-like phages are highly similar it can be safely assumed that replisome formation in these phages closely resembles the ϕT4 pathway, including the coupling of leading- and lagging-strand synthesis [381,3,456]. The ϕT4 clamp – loaded preferentially to primer-template junctions by the pentameric clamp loader – also acts as transcriptional activator during late transcription of ϕT4 together with the gp33 co-activator and the late σ-factor gp55 [384,487]. It is not precisely known, however, how a direction of the ϕT4 clamp to its cognate replicon is achieved, avoiding its titration by the host chromosome. The RFC small subunits of various sequenced archaeal and eukaryal genomes show significant similarity with the ϕT4 gp44 clamp loader subunit, which might in part simply be due to the presence of a nucleotide-binding domain of the AAA-type [324,93]. Homologues of the ϕT4 DNA polymerase, or the clamp, or the clamp loader proteins are not found in the genomes of other phages.

Phage-encoded proteins similar to either the the eukaryal PCNA clamp protein(s) or the *E. coli* β-clamp

(*dnaN*) were not be detected by BLAST searches. Also searches for phage-encoded proteins similar to the other subunits of *E. coli* DNA polymerase III, i.e. the γ(τ), δ, and δ' clamp loader subunits, or the χ, θ, and ψ subunits, were unsuccessful. It should be noted, however that genes encoding orthologues of the DNA Pol III χ, θ, and ψ subunits are only found in *E. coli* and few closely related species. *Streptococcus pyogenes* MGAS8232 encodes a protein (locus SpyM3_0690 as part of the pϕ N315.1 genome, also present in other *S. pyogenes* genomes) showing similarity (22% ident. res.) in the N-terminal 180 res. with *E. coli* Pol IIIδ'. The assignment 'putative DNA polymerase III delta prime subunit - phage associated' is nevertheless misleading because the protein is a fulllength homologue of the $DnaB_{Eco}$-type of helicases including all signature motifs (Section C3.3.; Table C28).

C3.6. Single-strand DNA binding and recombination proteins

C3.6.1. Single-strand DNA binding proteins (SSBs)

Single-strand DNA binding proteins lack enzymatic activity but play important roles as accessory proteins in DNA replication, recombination, and repair (reviewed in [294, 260]). Prokaryotic SSBs bind cooperatively to single-stranded DNA, and non-specifically, albeit with a marked preference for poly(dT) *in vitro* [261,235]. SSBs prevent degradation of ssDNA by nucleases, keep unwound DNA in an open conformation for subsequent steps during the initiation of replication, and stimulate the DNA strand exchange activity of RecA or RecT (see following chapter). A unique feature is the requirement of *E. coli* SSB as activator of ϕ N4 transcription [275]. SSBs of filamentous phages (e.g. ϕM13, ϕfd) or isometric phages (e.g. ϕX174, ϕG4) can inhibit replication, i.e. prevent the conversion of the ssDNA genomes into the double-stranded replicative form [45,130].

SSBs are essential for the replication of phages ϕT4 and ϕT7, the *E. coli* chromosomal replicon, and yeast, and it is assumed that SSBs are essential in all organisms (for references see [260,486]. Homologues of the *E. coli ssb* gene have been found in all presently known bacterial genomes (see below), in yeast mitochondria [461], and in the genomes of various phages and conjugative plasmids, e.g. F factor, R100, or Col1b. A specific role of the plasmid- and phage- encoded SSBs for their cognate replicons has been questioned because inactivation of the *ssb* gene did not prevent their propagation [146,184]. However, Lehnherr and co-workers could recently demonstrate that the ϕP1 SSB provides a selective advantage for wildtype over Δ*ssb-P1* phages in infection and propagation during stationary phase growth of the *E. coli* host [26]. It seems thus possible that plasmid- and phage-encoded SSBs perform important functions under conditions that are not tested in the laboratory routine.

The crystal structures of the SSB proteins encoded by ϕT4 gene 32 [PDB 1GPC], ϕT7 gene 2.5 [PDB 1JE5], ϕfd gene V [PDB 1YHA], and *E. coli ssb* [PDB 1KAW] have been solved. The proteins show no sequence homology and are structurally only distantly related. Despite some similarity in their secondary structures, BLAST searches did not reveal primary sequence similarities between the ϕA2 orf34 SSB and *E. coli* SSB (Fig. C19).

The SSB proteins have distinct biochemical properties: ϕT4 SSB binds to DNA as monomer, SSBs of filamentous phages as dimers, and *E. coli* SSB as homotetramers. *E. coli* SSB monomers bind non-specifically to ~15 nt long oligonucleotides. For the SSB homotetramer, binding sites of 35±3 nt and 65±5 nt were determined *in vitro*, depending on salt and protein concentrations, as well as temperature and pH, suggesting that 2 or all 4 subunits of a single tetramer, respectively, bind to ssDNA (for references see [260]). Lohman and Ferrari assume that *E. coli* SSB binding to ssDNA *in vivo* occurs by the binding mode in which ~65 nt ssDNA are wrapped around one tetramer. This binding mode of '*limited cooperativity*' results *in vitro* in the formation of beads consisting of double-tetramers separated by short protein- less DNA stretches that would allow other proteins to access the ssDNA [260]. The binding properties of SSB suggest that during the initiation of replication only very few tetramers actually bind to an unwound replication origin. A ~28 bp long stretch of the *B. subtilis oriC* was opened by the action of DnaA and HU *in vitro*, which was extended moderately by the addition of SSB [232].

The ϕT4, ϕT7, ϕA2, and *E. coli* SSBs possess highly acidic C-terminal tails of ~12 res. but not gene V SSB of ϕfd. It was shown that cooperative binding of *E. coli* SSB to ssDNA results in a conformational change of the protein exposing the C-terminus. SSB lacking up to 62 res. from the C-terminus is still able to form tetramers, and binds to ssDNA *in vitro* even tighter than wild-type SSB [482]. Comparable results were obtained also for ϕT4 gene 32 SSB [182,189], and Williams and coworkers proposed that these acidic C- termini are instrumental in signalling the SSB conformation to proteins involved in subsequent reactions [482]. For the acidic C-terminus of ϕT7 gene 2.5 SSB, an interaction with the gene 4A primase-helicase gene 5 polymerase could already be shown [213,214,212]. The *E. coli ssb-113* mutation results in a P→S exchange of the penultimate SSB residue, and – comparable to C-terminally truncated SSB – the SSB-113 mutant protein binds to ssDNA

Table C30 Phage encoded SSB proteins.

group 1: ϕ T4 gene 32-type SSBs	
genes	ϕT4 gp32, ϕT6 gp32, ϕT2 gp32, ϕRB69 gp32, ϕ44RR2.8t gp32, ϕRB49 p241, ϕAeh1 gp32, ϕSV14 gp32, ϕKVP40 gp32
lengths / similarity	295 - 322 res. / ≥48% ident. res.with ϕ T4 gp32, full length
conserved C-tail	acidic tail; 4 - 6 D/E res. within the 12 C-terminal res.
group 2: ϕ T7 gene 2.5-type SSBs	
genes	ϕT7 gene 2.5, ϕA1122 p12, ϕT3 gene 2.5, ϕYeO3-12 gene 2.5, ϕgh-1 p10
lengths / similarity	232 - 233 / ≥50% ident. res. with ϕ T7 gene 2.5, full length
conserved C-tail	12 res. acidic tail containing 4 - 8 E/D res.; C-terminal 3 res.: [G,E]DF
group 3: ϕ A2 orf34-type SSBs	
genes	ϕA2 orf34, ϕadh orf175, ϕPSA / ϕ2389 orf49, ϕDT1 orf34, ϕO1205 orf11, ϕsfi19 orf151, ϕ31 orf4, ϕsfi11 gp151, ϕsfi21 p37, ϕmi7-9 gp18c, ϕ105 orf10, ϕ ETA orf18
lengths / similarity	145 - 175 res. / ≥20% ident. res. with ϕ A2 orf34, full length
conserved C-tail	12 res. acidic tail containing 3 - 5 D/E res.; C-terminal 9 res.:el[S,e][D,e]xd[L,v]PF;
group 4: ϕ M13 gene V-type SSBs	
genes	ϕM13 gene V, ϕfd gene V, ϕf1 gene V, ϕIke gene V, ϕLf orf98, ϕCf1c p1
lengths	87-98 res. / ≥33% ident. res. with ϕ M13 gene V, full length
conserved C-tail	no pronounced acidic tail, no consensus
group 5: λ Ea10-type (putative) SSBs	
genes	λ *ssb*(*ea10*), ϕVT2-Sa *ea10*, ϕ933W *ea10*, ϕ 4795 *ea10*
lengths / similarity	122 res. / ≥99% ident. res. with λ ea10, full length
conserved C-tail	weakly acidic tail; C-terminal 12 res.: IFDSDDMTIKAA
group 6: other SSBs	
genes	ϕIf1 geneV, ϕPf3 orf78; from filamentous phages of Proteobacteria
	ϕN4 N4SSB, ϕPRD1 gene XII; from Gram(-) hosts
	ϕ29(ϕ PZA) gene 5A, ϕB103 gene 5, ϕNf gene 5, ϕGA-1 gene 5, ϕ P68 orf14, ϕ44AHJD orf14; from Bacillales phages with genome sizes of ~20 kb
	ϕ BK5-T orf48; from *Lactococcus* sp., acidic C-tail: APMNISDEDLPF
group 7: *E. coli*-type SSBs	
Gram(−) specific	
genes	ϕVP2 ssb, ϕVP5 ssb, ϕBcep22 gp16, ϕSf6 gp27, ϕP1 *ssb-P1*, ϕT1 27
lengths / similarity	157 - 169 res. / ≥48% ident. res. with *E. coli ssb*, full length
conserved C-tail	12 res. acidic tail containing 3 - 5 D/E res.; C-terminal 9 res.: m[D,e]fD[D,S][D,S]IPF;
Gram(+) specific	
genes	ϕTP901-1 p12, ϕLL-H orf299, ϕSPP1 G*36*P, ϕCJW1 gp102, ϕSLT orf147, ϕNIH1.1 gene 13, ϕul136 orf141, ϕTuc2009 orf15, ϕMM1 p13, ϕ7201 orf9, ϕbIL285 orf15, ϕPV83 orf18, ϕbIL286 orf15, ϕPVL orf45, ϕVO1 orf1, ϕ3626 p39, ϕA118 ORF60
lengths / similarity	139 -181 res. / ≥28% ident. res. with *E. coli ssb*, full length
conserved C-tail	12 res. acidic tail containing 3 - 5 D/E res.; C-terminal 8 res.: [d,e]ixd[D,q][L,v]P[F,v]

slightly tighter than wild-type SSB *in vitro* [72]. A possible interaction of *E. coli* SSB with the primosomal protein PriB was proposed based on biochemical experiments, but has not been substantiated [264]. More recently, a direct interaction of the acidic tail (15 res.) of *E. coli* SSB with the ψ subunit of DNA Pol III holoenzyme was shown by O'Donnell and coworkers to be necessary for efficient clamp loading [207]. They could also show that the SSB-113 mutant protein is defective in clamp loading. These findings support models for the initiation or the resumption of replication that consider SSB as an essential protein at the steps following helica-

se loading during replisome formation: SSB might be instrumental in co-ordinating the coupling of leading- and lagging-strand DNA synthesis [207,145]. Orthologues of the *E. coli holD* gene – encoding the ψ subunit of DNA Pol III holoenzyme – are only found in the genomes of the Enterobacteriales and a few other γ-subgroup Proteobacteria. However, the perfect conservation of the penultimate proline residue within a reasonably conserved acidic tail in all *E. coli*-type SSBs from Proteobacteria and Firmicutes, and in all ϕA2 orf34-type SSBs (see below) underscore the importance of the C-terminus of these SSBs. Database searches produced >50 matches for phage- encoded SSBs (prophages not included), which we sorted into 6 groups of similar proteins, and a seventh group containing sequences with no or few apparent homologues (Table C30). The largest group, group 7, containing the *E. coli*-type SSBs, is discussed last.

Group 1: phage T4 gene 32 (gp32) SSB. The 9 phages encoding homologues of the ϕT4 gene 32 single-strand DNA binding protein belong to the group of the T-even phages of Enterobacteriales, characterised by their intricate replication mode (BRM Section 2.5.). The high degree of similarity suggests that the biochemical properties of all proteins are very similar to those of ϕT4 gp32 SSB. All proteins contain a highly acidic C-terminal tail, which however lacks a detectable consensus sequence. In addition to the full-length homologues, the data bases list numerous homologues of ϕT4 gene 32 for which only fragments are known from phages ϕFS-α, ϕM1, ϕPST, ϕRB8, ϕRB10, ϕRB6, ϕRB15, ϕRB18, ϕSV76, ϕRB27, ϕRB32, ϕRB69, ϕRB3, ϕRB9, and ϕRB70.

Group 2: phage T7 gene 2.5 SSB. Like the SSBs of group 1, the 4 homologues of the ϕT7 gene 2.5 protein contain highly acidic C-terminal tails that lack a detectable consensus sequence. For the C-terminus this protein, specific interactions with gene 4 primase/helicase and gene 5 DNA polymerase was shown [223].

Group 3: phage A2 orf34 SSB. The 12 homologues of the ϕA2 orf34 SSB are exclusively found in the genomes of Bacilli phages (prophages not included). Although the similarity of the proteins among each other is rather low they all contain a well conserved acidic C-terminus. In addition to the overall structure (see above), the C-terminal tails of these SSBs are highly similar to those of *E. coli*-type SSBs of Gram(+)-specific phages (group 7). It is therefore reasonable to assume that the SSBs of both groups perform identical functions. The *ssb* genes of this group are in most cases closely located to (putative) replication genes of their cognate phage replicons, in the same transcription unit. The cognate SSBs are probably essential for the replication of the phages of this group because Kim and Batt could demonstrate inhibition of phage replication in an antisense approach [211]. Since in these experiments a premature termination of transcription of the (essential) ϕP4α-type helicase gene could have been the reason for the observed inhibition of phage replication, more experiments are warranted for a decisive statement. Notably, only 1 of the *ssb* genes of this group (ϕETA orf18) is found associated with one of the set of 40 replication initiator genes of Gram(+)-specific phages discussed in Section C3.1.2. . Also *vice versa*: with one exception (ϕBK5-T, see group 6), only *ssb* genes of group 7 are found associated with the replication initiator genes of the set of 40. Apparently, mosaicism, i.e. HGT among phages, did not mess up phage phylogeny entirely.

Group 4: phage M13 gene V SSB. The 6 orthologues of the ϕM13 gene V SSB are found in the genomes of filamentous phages with a genome size of ~7 kb. Despite their high degree of homology, these proteins lack a consensus C-terminus.

Group 5: phage λ Ea10 SSB. The precise function of the λ Ea10 SSB is not known, it is however not required for initiation of λ replication. The almost complete identity – including the weakly acidic tail – of the 4 Ea10 orthologues suggests that these proteins have a common origin in one of the four phages and were spread only recently.

Group 6: miscellaneous SSBs. This group contains 11 SSBs, which did not show any similarity to SSBs of one of the other groups. ϕBK5-T orf48 encodes a SSB that has no sequence similarity with either group 3 nor group 7 SSBs, except for the acidic C-terminus, which is highly similar to that of the group 7 SSBs of Gram(+)-specific phages. The 6 homologues of ϕ29 gene 5 SSB seem to be a unique type of SSB involved in the protein-primed replication of the linear genomes of related Bacillus phages [155,287]. Also *E. coli* phage ϕPRD1 replicates via the protein-primed mode, and its SSB might be involved in this process. The specific function of ϕN4 SSB is not known. This phage replicon requires *E. coli* SSB for its transcription (see above). The ϕf1 geneV and the ϕPf3 orf78 SSBs of filamentous phages have no apparent sequence similarity with group 4 SSBs, despite a comparable size.

The heterogeneity of the SSBs of this group suggests to us that in addition to the known SSB types more phage proteins with yet unknown function may be identified as SSBs. At present, it is clear that single-strand DNA binding proteins evolved several times, and that the *E. coli*-type SSB seems to be the evolutionary most successful, by prevalence.

Group 7: E. coli *SSB and phage-encoded homologues.* Homologues of *E. coli* SSB are encoded by 5 Gram(−)-specific phages and 17 Gram(+)-specific phages, and in addition by many prophages (see below). The 5 Gram(−)-specific phages possess different replication gene modules (BRM Section 3.1.). Most other phages of this host range having similar replication modules lack *ssb* genes. Therefore we presume that the *ssb* genes in the genomes of ϕVP2, ϕVP5, ϕBcep22, and ϕSf6 represent accessory genes required for specific growth conditions rather than for replication under standard laboratory conditions – similar to ϕP1 [26]. In contrast to the genomes of Proteobacteria-specific phages, *ssb* genes are more frequently found in the genomes of phages of Firmicutes hosts. As already observed for the SSBs of group 3, the *ssb* genes of the 17 Gram(+)-specific phages of group 7 are in many cases closely located to (putative) replication genes of their cognate phage replicons, probably in the same transcription unit. SSBs of this group might thus be essential for the replication of some of these phages even under normal conditions, but apparently not for SPP1 [475].

Two regions of the *E. coli*-type SSB proteins are conserved among all phage-encoded and also the chromosomally encoded SSBs: i. the N-terminal ~120 res. responsible for ssDNA-binding and tetramerisation (see above), and ii. the C-terminal ~15 res. . The connecting region is responsible for the observed variation in length of these SSBs, and poorly conserved. The short C-terminal consensus sequences determined for the SSBs encoded by the Gram(−)-specific and the Gram(+)-specific phages are slightly different but correspond well to the C-termini of the cognate SSBs of their hosts. They may be considered phylon-specific SSB- fingerprints. In addition to their conservation in the SSBs, C-terminal tails containing ~4 acidic residues and a terminal LPF motif are also found in the initiator proteins of the Gram(+)-specific phages ϕLL-H, ϕmv4, ϕr1t, and the *B. anthracis* pϕLambdaBa04. Although this feature is not found in other phage initiator proteins, and might be coincidental therefore. It could be rewarding to analyse the role of this putative signalling domain – by experiments.

81 of the 136 presently known bacterial genomes contain a single *ssb* gene encoding a protein with significant similarity to *E. coli* SSB. 24 genomes code for either SSBs with an atypical (e.g. *Mycoplasma* sp.) or truncated C-terminus, or contain a C-terminally truncated in addition to a full- length *ssb* gene. However, 32 genomes code for two or more full- length SSBs with significant similarity to SSB_{Eco}. Certainly, the 11 full-length and 4 truncated *ssb* genes in the genome of a *Phytoplasma* (sp. onion yellows) represents an extreme, but 3 - 4 *ssb* genes are frequently found in the genomes of Bacillales. On closer inspection, "surplus" *ssb* genes are in most cases readily identified as parts of prophages. But the significant similarity of all *E. coli*-type SSB proteins has already led to confusion in gene assignment for some bacterial genomes containing multiple *ssb* gene(s). E.g. the locus sa0353 of the NCBI genome entry for *S. aureus* N315 is marked as "single-strand DNA-binding protein of phage phi PVL"; the gene is located between *rpsF* and *rpsR*, a feature also conserved in the genomes of *Bacillus* sp & *Listeria* sp., respectively. Locus sa1792, the second SSB-encoding gene in *S. aureus* N315 and marked "single-strand DNA-binding protein" is next to sa1791 encoding a prophage initiator gene, which suggests that sa1792 is a prophage *ssb* gene. Locus sa0353 represents the chromosomal *ssb* gene, therefore.

The presence of multiple *E. coli*-type *ssb* genes as part of prophages in ~25% of the presently known genomes is explained by the observation that temperate phage replicons preferably contain *E. coli*-type *ssb* genes. However, also homologues of the temperate phage ϕA2 orf34 gene are found in the genomes of e.g. *L. innocua* (locus lin2588) and *S. pyogenes* (locus SpyM3_0960). There are, in addition, rare examples for the acquisition of *ssb* genes from virulent phages by chromosomal replicons. The *Xylella fastidiosa* locus XfasO1896 and the *X. campestris* locus XCC2059 encode proteins that are homologues of the ϕfd geneV SSB. Also the *rstB1* gene, part of the CTXϕ prophage of *V. cholerae*, is a ϕfd geneV homologue [95]. *P. aeruginosa* strain PAO1 contains a chromosomal copy of the ϕPf1 *ssb* gene (locus PA0720). The genome of *P. putida* KT2440 (locus PP2267) encodes a homologue of the ϕT7 gene2.5 SSB [397]. The phage-encoded *ssb* genes, which do not belong to the group 7, did not replace the chromosomal *E. coli*-type *ssb* gene as proposed by Forterre [126]. It seems more likely that phages acquired these *ssb* genes from their hosts by horizontal gene transfer (HGT) as suggested by Moreira [301] (see also BRM Section 4.). These observations emphasise the pivotal role of bacteriophages as vectors for HGT.

C3.6.2. Recombination proteins

Note in advance: temperate phage replicons can integrate into the chromosomes of their hosts by site-specific recombination. Subsequently, the 'prophage' is propagated as part of the host chromosome. The shut-off of prophage replication by a λ *cI*-type repressor can not be compared to copy- number control mechanisms operating in plasmid replicons and chromosomes. Also, the recombination process resulting in prophage formation is mechanistically unrelated to replication. We will therefore not discuss integration here [153]. We note however that the highly conserved phage integrase

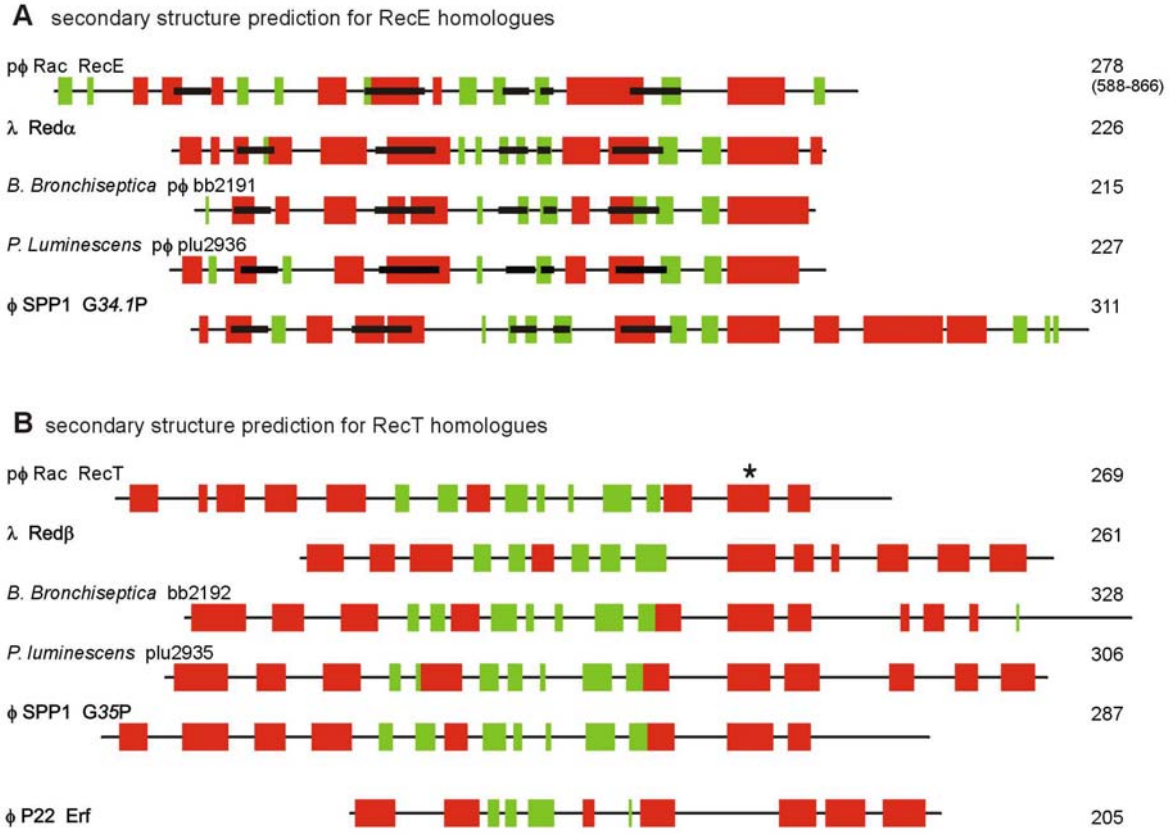

Fig. C20 Secondary structure prediction for RecE/RecT analogues.
The secondary structure of redα is shown according to the crystal data [PDB 1AVQ]. For the other proteins, secondary structures are shown according to Jpred prediction [88]. The length of the individual protein sequences is indicated by the black line; all protein sequences are shown at the same scale. Red = α-helices; green = β-strands. **A** Thicker black lines indicate the positions of the signature motifs defined by Vellani + Myers [463]. For the *E. coli* RecE sequence, only the functionally important res. 588-866 were included in the analysis [79]. **B** The helix containing the conserved MxxKT$x_{2\text{-}3}$(R/K) motif (weakly conserved in redβ) is marked with an asterisk.

genes have been instrumental for the detection of prophages – or their remnants – in bacterial genomes [63].

The stringent coupling of replication and recombination was for many years considered a hallmark of the extravagant 'life-style' of bacteriophage T4 – without paradigmatic consequences for the studies of both nucleic acid processing pathways in bacteria and other phages [305,139]. The coupled replication and recombination of ϕT4 was addressed in BRM Section 2.5., and we will only discuss ϕT4 endonuclease VII here in the context of phage-encoded Holliday junction resolvases.

Also the recombination-dependent replication of bacteriophage Mu was taken as an exception, and comparable rather to the propagation of prokaryotic transposons. Because ϕMu relies on the host replication/recombination machinery for propagation – except for its cognate transposase – we will not discuss it here, and refer the reader to several instructive reviews and papers [67,161,314,420]. During the last decade, however, and due to an increasing awareness of the pioneering work of Kogoma and his collaborators [218], the importance of recombination for chromosome replication in *E. coli* came into the focus of replication research [238,85] (BRM Sections 2.5. + 2.6.).

In this section we shall discuss only those phage-encoded recombination pathways and proteins for which a direct functional connection to phage replication has been shown, or can be assumed. Proteins include 5'→3' exonucleases, singlestrand annealing proteins (SAPs), and Holliday junction resolvases. This section concludes with a brief survey of phage-encoded (putative) homologues of other known *E. coli* recombination protein. Phage recombination pathways and proteins have been reviewed extensively by Kuzminov [238].

The λ Redα/Redβ (Exo/Bet) recombination pathway. The observation that λ recombination was not affected by a *recA* mutation of the *E. coli* host led to the detection of phage-encoded recombination functions [113]. Accordingly, λ mutants were isolated that grew normally in a wild-type host, but failed to propagate in *rec*$^-$ hosts. The corresponding λ genes were identified as the genes coding for λ exonuclease (Redα or Exo) [254,414,64], and for beta protein (Redβ or Bet) [216,310]. Redα and Redβ occur *in vivo* as an equimolar complex [310]. During stepwise 5'→3' exonucleolytic degradation of one strand, Redα loads Redβ to the other strand. Redβ was shown *in vitro* to promote reannealing of complementary single strands. The structure of Redα [PDB AVQ] but not that of Redβ is known. More recently, Stahl and co-workers could show by genetic analysis that the Rap(NinG) protein (see below) is the 'cognate' Holliday junction resolvase for the λ Redα/Redβ recombination pathway [443].

Younger readers may wonder why decades passed between the detection of the λ encoded recombination genes by genetic experiments [127,49], early biochemistry [64], and the elucidation of the three-dimensional structure of redα [229]. The initially vague concepts of homologous recombination had to be transformed into robust experimental models describing the molecular mechanisms and the actors driving it. So, (almost) all of what is known today about homologous recombination in bacteria can be traced back to the ingenious genetic analyses of various combinations of *E. coli rec* and λ *red* mutants [238].

The *E. coli* pϕ Rac RecE/RecT recombination pathway. ***Gene***tic analysis of *E. coli recBC* mutants (exonuclease V) revealed secondary suppressors termed *sbcA* that were able to (partly) restore recombination proficiency of the strain by activating the expression of exonuclease VIII [237,483]. The *sbcA* mutation(s) were mapped to the region of the cryptic Rac prophage of *E. coli* K12, and the gene encoding exonuclease VIII termed *recE*. Exonuclease VIII is a polypeptide of 866 res. in length but only the C-terminal 278 residues are required for exonuclease activity [79]. The function of the large N-terminal domain is not known. The *E. coli recBC sbcA* mutants not only allowed propagation of λ *red* phages – indicating that the activated function(s) are analogous to the λ *red* pathway – but also made the propagation of λ *red* phages independent of *recA*. The latter finding suggested that the *recE* region encoded a Redβ-like function in addition to exonuclease VIII [80]. Kolodner and co-workers determined the polypeptides expressed from the *recE* region: exonuclease VIII(RecE) and – from the same transcript – RecT with the characteristics of a SAP [159].

The ϕSPP1 G*34.1*P/G*35*P recombination pathway. As mentioned above, mutations in ϕSPP1 of either gene 34.1 or 35, respectively, result in a replication arrest phenotype, producing linear progeny molecules of subgenome length [475]. The specific length of these molecules led Alonso and co-workers to hypothesise that unidirectional replication proceeds normally in such mutants from *ori*L to *ori*R – a duplicated origin structure though inactive for initiation. At *ori*R, bound G*38*P creates a road block for the replication fork, which cannot be overcome due to the lacking recombination activity of either G*34.1*P or G*35*P [12]. These authors could show that purified G*35*P protein is a SAP with properties comparable to RecT, and able to anneal single-stranded *ori*L DNA to the homologous part of a supercoiled template [12]. *Gene* 34.1 was recently shown to encode for the cognate 5'→3' exonuclease, termed G*34.1*P [280] or Chu [463]. Together, these studies demonstrate that the exonuclease/SAP recombination pathway is similar in Gram(–)- and Gram(+)-specific phages with respect to the underlying molecular mechanism.

The ϕP22 Erf recombination protein. Early genetic analyses suggested that ϕP22 carries a RecA-like gene whose inactivation prevented the growth of the phage on an *E. coli recA* host [490,351,136]. The structure of ϕP22 Erf (essential recombination function) is not yet known. It was shown that the N-terminal ~140 res. of Erf are sufficient for ssDNA- binding and oligomerisation, the function of the C-terminus is not known [353,312]. The secondary structure prediction suggests that Erf is – at best – only distantly related with the RecT/Redβ SAPs although all known phage-encoded SAPs form fairly similar complexes with ssDNA (Fig. C20) [336]. In contrast to the phage recombination protein pairs discussed above, the ϕP22 Erf protein lacks a 'cognate' exonuclease. The arf gene downstream of *erf* in the P22 genome encodes a small protein (47 res.), which was shown by genetic analysis to be required for efficient recombination of ϕP22 [352]. Because of its small size it seems unlikely that arf possesses exonuclease activity but may instead recruit the yet unknown exonuclease. Since homologues of arf are not found in other phages this notion remains speculative.

Evolutionary relationship of the RecE/RecT, G*34.1*P/G*35*P, and Redα/Redβ protein pairs. The RecE/RecT recombination proteins encoded by the *E. coli* Rac prophage and the Redα/ Redβ proteins of λ perform analogous functions. Despite subtle differences in protein activity and the failure to obtain a functional complementation in reciprocal assays [313], and despite the lack of significant primary sequence similarity, these protein pairs may be evolutionary related [219,61, 194].

Table C31 Protein sequence similarity among RecE/RecT analogues.

		RecE	Redα	bb2191	plu2936	G*34.1*P
pϕ Rac	RecE	x				
λ	Redα	—	x			
B. bronchiseptica pϕ	bb2191	—	29%	x		
P. luminescens pϕ	plu2936	—	70%	28%	x	
ϕ SPP1	G*34.1*P	18% [1])	—	25%	27%	x

		RecT	Redβ	bb2192	plu2935	G*35*P
RecT	x					
Redβ	—	x				
bb2192	31%	—	x			
plu2935	28%	29%	41%	x		
G*35*P	40%	—	29%	28%	x	

Only the functionally important res. 588-866 of *E. coli* RecE were included in the analysis [79]. BLAST analysis was performed using the NCBI 'bl2seq' software with default settings except that the 'low complexity' filter was disabled [445]. 1) value taken from Ayora *et al.* [12].

Table C32 (Pro)phage-encoded ϕP22 Erf-like proteins.

A signature motifs:

	A	B
consensus	**Q**xx**G**ATSS**Y**hRK**Y**xhxxh**F**xh x$_{4-5}$ **D**x**D** **A** KR **L**	**KG**xx**S**DAh**KRA**AVxW**GIG**RY**LY** G CF F **V** E

B occurrence of motifs:

replicon	accession	gene	res. res.	BLAST ϕ P22 1-150	BLAST ϕ P22 151-205	BLAST ϕ D3 1-200	BLAST ϕ D3 201-275	BLAST ϕ ST64T 1-180	BLAST ϕ ST64T 181-235	motifs A	motifs B
C. acetobutylicum pϕ	NP_348558	cac1936	229	-	-	-	-	32	-		+
T. thermophilus	YP_005892	ttc1923	212	-	-	-	-	31	-		+
ϕ RM378	NP_835605	p018	164	-	-	-	-	39	-		+
ϕ bIL286	NP_076648	orf14	252	-	-	-	-	57	-		+
ϕ ul36	NP_663647	ORF252	252	-	-	-	-	57	-		+
ϕ ST64T	NP_720291	orf-235	235	-	76	-	-	x	x		+
ϕ Sf6	NP_958204	gene 28	235	-	63	-	-	94	81		+
ϕ ST104	YP_006369	ORF13	235	-	77	-	-	94	98		+
ϕ P22	NP_059596	*erf*	205	x	x	22	-	-	89	+	
ϕ H-19B	AAD04639	*erf*	201	88	74	23	-	-	32	+	
ϕ HK022	NP_037690	Erf	201	88	73	24	-	-	32	+	
ϕ HK97	NP_037726	gp40	201	89	74	24	-	-	32	+	
ϕ c2	NP_043534	e15	179	38	-	-	-	-	-	+	
ϕ MM1	NP_150142	ORF9	235	36	-	-	-	-	-	+	
ϕ 7201	NP_038308	ORF7	218	28	-	-	-	-	-	+	
S. agalactiae	NP_687594	sag0565	224	23	-	30	-	-	-	+	
ϕ NIH1.1	NP_438125	phiNIH1.1_12	224	23	-	30	-	-	-	+	
S. pyogenes MGAS8232	NP_607818	spyM18_1796	224	23	-	30	-	-	-	+	
ϕ PV83	NP_061607	orf17	212	24	-	26	-	-	-	+	
N. punctiforme	ZP_00106022	npun0391	257	23	-	23	-	-	-	+	
L. johnsonii pϕ Lj771	AAK27920	orf223	221	[42]	-	25	-	-	-	+	
ϕ LL-H	AAL77543	orf178b	178	-	-	24	-	-	-	+	
ϕ mv4	AAG31330	orf244	244	-	-	25	-	-	-	+	
ϕ T1	YP_003917	28	226	-	-	23	-	-	-	+	
ϕ SLT	NP_075480	orf216	216	-	-	-	-	-	-	+	
pcp9 *B.burgdorferi*	AAG36971	orf4	302	-	-	27	-	-	-	+	
L. gasseri pϕ	ZP_00046420	lgas0587	270	-	-	-	-	-	-	+	
L. johnsonii pϕ Lj928	NP_958522	ljo_1449	261	-	-	25	-	-	-	+	
D. vulgaris	YP_011247	dvu2032	237	-	-	27	-	-	-	+	
L. monocytogenes pϕ	NP_465842	lmo2318	232	-	-	26	-	-	-	+	
ϕ D3	NP_061548	orf52	275	-	-	x	x	-	-	+	
B. cepacia pϕ	ZP_00219218	Bucepa02006186	215	-	-	30	-	-	-	+	

Also The G*34.1*P/G*35*P protein pair encoded by *B. subtilis* phage ϕSPP1 belongs to this group of functional analogous exonuclease/SAP pairs, and Alonso and coworkers already pointed to the significant sequence similarity between RecT and G*35*P (40% ident. res.) [12]. The availability of many new genome sequences of (pro)phages and bacterial species made it attractive to re-evaluate the evolutionary relationship of these protein pairs. BLAST searches revealed protein sequences encoded by adjacent genes of (putative) prophages in the genomes of *B. bronchiseptica* (β- Proteobacteria) and *Photorhabdus luminescens* (γ-Proteobacteria) whose intermediate similarity values support the notion of a common origin of the phage- encoded exonuclease/SAP pairs (Table C31) [194]. This observation illustrates neatly the mutual benefits that phage research and genomic sequencing projects can achieve by combining their data.

The *P. luminescens* plu2935 protein shows significant similarity to RecT and Redβ. Likewise, the *P. luminescens* plu2936 protein is a homologue of Redα, but also similar to the *B. bronchiseptica* bb2191 protein, which in turn is similar tor G*34.1*P of ϕSPP1. Because the sequences differ considerably in size we performed a secondary structure prediction analysis in search for related structures (Fig. C20). Both groups of proteins show a fairly similar 'core' of alternating β-stranded and α-helical regions. We infer from this analysis that a common evolutionary origin of these phage-encoded exonuclease/SAP pairs is very likely. However, this analysis suggests that ϕP22 Erf is only very distantly related to the other 3 SAPs, if at all.

Other phage-encoded (putative) exonuclease/SAP gene pairs. As discussed above and also noted by others [463], phage-encoded exonucleases and SAPs form functional pairs and their genes are usually located close to each other in the genomes of their cognate replicons. A complex picture emerges when the BLAST similarities of the RecE/RecT, the G*34.1*P/G*35*P, the Redα/Redβ proteins, and the ϕP22 Erf protein to other phage-encoded (putative) exonuclease/SAP proteins are plotted in a matrix that includes the pairwise arrangement of the respective genes (Figs. C21 + C22). We analysed 91 phage and prophage sequences by this approach. A further known ~30 known sequences mostly from prophages were omitted because they did not form novel (sub)groups. Also we excluded the multiple exonuclease/SAP proteins encoded by *Phytoplasma* (sp. onion yellows) (see above for *Phytoplasma ssb* genes, Section C3.6.1). The matrix suggests the definition of six distinct groups:

Group I contains (putative) SAPs with similarity to ϕbIL286 orf 14 (Fig. C21; upper left part). The N-terminal ~150 res. of these proteins show weak similarity to the eukaryotic Rad52-type SAPs . The bipartite human Rad52 protein [PDP 1H2I] contains within its N-terminal DNA-binding domain (~200 res.) in a larger α-helix the motif **K**EAV**T**DGL**KRA**LRSF**G**NALGNC, which distantly resembles the motif B present in all proteins of this group at a corresponding position (Table C32) [416]. The Rad52 C-terminus might be responsible for the interaction with the Rad51 recombination protein [416]. Despite a reasonable conservation of the Ntermini, the SAPs of this group lack similarity in their C-termini. Moreover, The C-termini of the (putative) SAPs of ϕST64T, ϕSf6, and ϕST104 are highly similar to the C-terminus of ϕP22 Erf (Table C32). Possible protein-protein interaction partner(s) of ϕP22 Erf are not known, therefore we can only mention here the intriguing C-terminal similarity of two groups of SAPs with completely unrelated N-terminal DNA-binding domains. A 'cognate' 5'→3' exonuclease could not be detected in the phage genomes of this group.

Group II contains SAPs with similarity to the ϕP22 Erf protein (Fig. C21; upper left to middle part). The crude BLAST approach did not reveal any similarity to one of the other SAPs. The initial sorting by eye already suggested that several subfamilies exist within this group. A more detailed BLAST search identified motif A (pos. 97-124 in ϕP22 Erf) as the most highly conserved region in all sequences of this group. Motif A is contained within the presumed DNA-binding region of ϕP22 Erf but its exact functional role is not known [312]. We tentatively use this motif here to discriminate sequences belonging to group I from those belonging to group II (see above). Motif A could also be identified unequivocally at corresponding positions in sequences that failed to give BLAST matches with ϕP22 Erf or ϕD3 orf52 as query, e.g. ϕSLT orf216 (Table C32). The C-terminal ~50 res. of the Erf-like proteins are highly divergent; only for the closest relatives of ϕP22 Erf similarity was found to three sequences from group I (see above).

Although Erf was originally identified in the ϕP22 phage of enteric hosts, it is now clear that *erf* genes are present in many (pro)phage genomes covering a wide host range. BLAST searches with redα, G*34.1*P, or RecE as query did not lead to the detection of a 'cognate' 5'→3' exonuclease for this group. There are three notable exceptions: i. the ϕT1 gene 29 – directly adjacent to gene 28 encoding an Erf-like protein – shows significant similarity to the RecE-like proteins gp60 of ϕChe9c and orfB of a *Legionella pneumophila* prophage [356], and ii. the dvu2033 protein of a *Desulfovibrio vulgaris* prophage is similar to the G*34.1*P-like protein encoded by the *C. tetani* E88 prophage locus ctc02146 (not shown). The third exception is ϕD3: while orf 51 is a

Redα- like exonuclease, orf 52 is a Erf-like SAP (Fig. C21). These three examples emphasise that the pairwise occurrence of exonuclease/SAP genes in phage genomes is a recurring feature although other than the prototype protein pairs are possible.

Group III contains gene pairs with similarity to the λ Redα/Redβ proteins (Fig. C21; lower right part). An unequivocal sorting of the protein pairs in this group is only possible for the very close relatives of λ. There exist three minor subgroups: i. the ϕD3-type with a Redα-like exonuclease and a Erf-like SAP (see above), ii. the *B. bronchiseptica* prophage-type with a Redα-like exonulease (bb2191) and a RecT/G*35*P-like SAP (bb2192), and iii. the *P. rettgeri* prophage-type with a G*34.1*P-like exonuclease (orf59) and a Redβ-like SAP (orf68). A Redβ-like SAP gene but no 'cognate' exonuclease gene could be identified in the genomes of ϕLC3 and ϕ31.1 .

GroupIV contains proteins with similarity to the ϕSPP1 G*34.1*P/G*35*P protein pair (Fig. C22; upper left part). As can be deduced from the matrix, the G*35*P-like proteins of this group form a larger RecT-type family of phage encoded single-strand annealing proteins together with the SAPs of the ϕbIL309 and pϕRac groups. The RecT family has no detectable structural similarity to proteins from the Erf family, despite their analogous function. There is however a distant sequence similarity and (predicted) structural relatedness between members of the RecT family and Redβ proteins: the (putative) *P. luminescens* SAP plu2935 shows similarity to Redβ-like as well as RecT-like proteins (see also above).

Group V contains SAPs with similarity to the ϕbIL309 RecT protein (Fig. C22; lower right part). All members of this group are encoded by Gram(+)-specific phages. The enzymatic activity of the ϕbIL309 RecT protein – or of any other protein from this group – has not yet been demonstrated, and the function assigned only by the convincing similarity to RecT and G*35*P. BLAST searches with Redα, G*34.1*P and RecE as query – or with any of the exonucleases of these groups – did not lead to the detection of a 'cognate' 5'→3' exonuclease for this group. For some of the homologues of the ϕbIL309 orf13 protein we could detect some weak similarities to RecB and RuvA proteins. It seems thus possible that the orf13-like proteins are involved in recombination, although their inclusion in the matrix is purely speculative at present. However, the 'canonical' gene arrangement – the exonuclease gene preceding the SAP gene within the same transcription unit – is reversed in the ϕbIL309 RecT/orf13-like gene pairs.

Group VI contains proteins with similarity to the *E. coli* pϕRac RecE/RecT proteins (Fig. C22; lower right part). For several RecE-like proteins, a 'cognate' SAP could not be detected, e.g. for the *E. coli* prophage pϕCP-933P, or *Salmonella* sp. phage ϕST64B. Interestingly, the putative exonucleases encoded by the pϕCP-933U, P, and M show significant similarity to the DexA exonuclease of the ϕT4 group of phages – which however is a 3'→5' exonuclease.

We can draw the following conclusions:

i. within all groups, members with high, moderate and low similarity to each other were found. This suggests that the major possible variations within and among the groups defined here could be identified by analysing sequences from 91 (pro)phages.
ii. in roughly one half of the cases, exonuclease genes and SAP genes are present as pairs in phage genomes. For the other half, it is presently not possible to decide whether the genes are present as single genes or whether the 'cognate' partner has not yet been identified.
iii. the exonuclease and SAP genes occur in favoured combinations, but examples exist for any possible combination. This emphasises that the presence of both types of recombination proteins is more important for phage recombination than a particular exonuclease or SAP. We could however not detect any 'mixing' of the phage-encoded exonuclease/SAP gene pairs with the (functional analogous) chromosomal *recBCD*/*recA* genes.
iv. with regard to sequence similarity, there is no apparent distinction between exonuclease/SAP pairs from Gram(+)- and Gram(-)-specific phages, in contrast to the replication initiator proteins (Section C3.1.2.). This suggests that – consistent with experimental results (see above) – the phage-encoded recombination proteins perform their function independent from the host recombination machinery.
v. in most cases, the gene encoding the exonuclease precedes the SAP-encoding gene within the same transcription unit. This gene arrangement is independent of the type of exonuclease and SAP, and may have enabled successful recombination events combining different types of exonuclease and SAP genes. In some cases, small intervening orfs (<100 res.) were found to separate the exonuclease and SAP genes.
vi. the plasmid encoded Erf-like protein orf14 of the ~9 kb plasmid pcp9 of *B. burgdorferi* [430], and the (putative) 5'→3' exonuclease/SAP pairs of the 217 kb plasmid pCAR1 of *P. resinovorans* [269] and of the 199 kb plasmid pRts1 of *P. vulgaris* [311] are included in the matrix. No other plasmid-encoded proteins with similarity to either exonuleases or

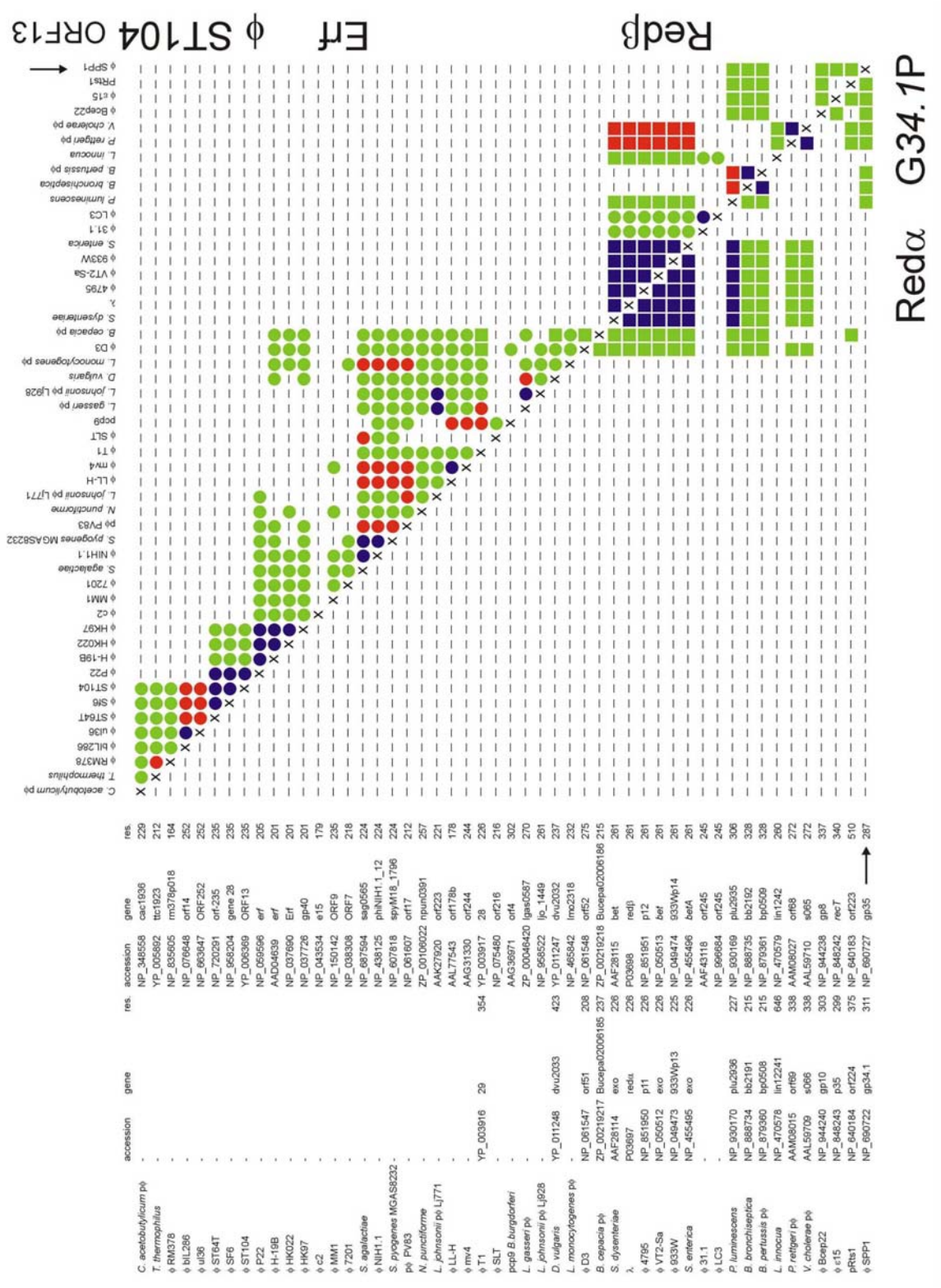

Fig. C21 Phage-encoded 5'→3' exonuclease / SAP protein pairs; part A.

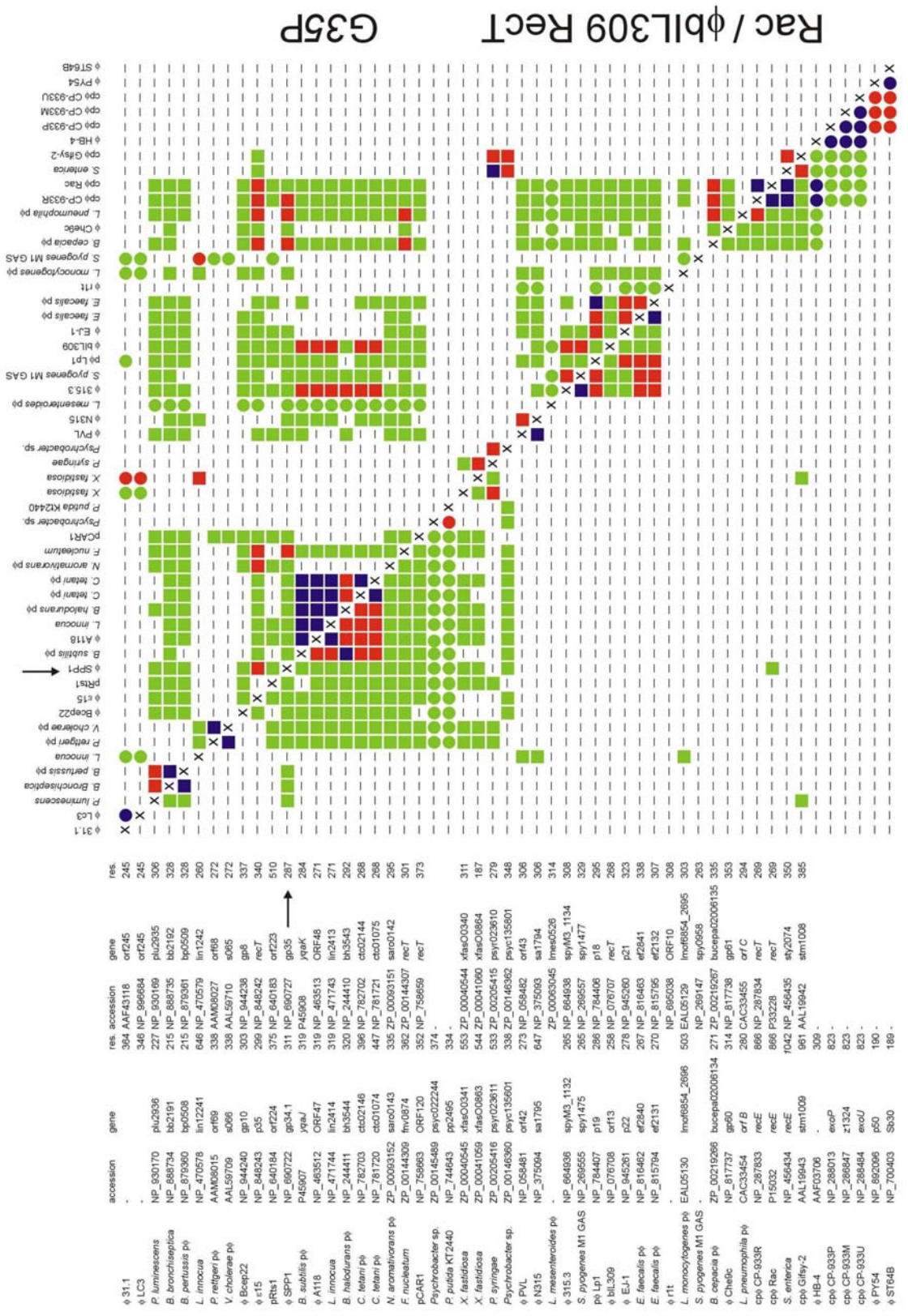

Fig. C22 Phage-encoded 5'→3' exonuclease / SAP protein pairs; part B.

Legend to Figs. C21 + C22 . The BLAST similarity values for the pairwise comparison of the exonuclease proteins are shown in the lower left part, i.e. below the diagonal. The values for the pairwise comparison of the SAP proteins are shown in the upper right part, i.e. above the diagonal. Colour code: blue = similarity 61-100% ident. res.; red = similarity 41-60% ident. res.; green = ~20%-40% ident. res.. Squares indicate that in the pairwise comparison of exonuclease and SAP sequences the respective proteins have been detected in both compared sets. Circles indicate that in the pairwise comparison one protein was not detected in at least one of the compared sets. For reasons of space the original matrix was split into two separate figures. The region of overlap between both figures is indicated by vertical + horizontal arrows at the positions of the ϕSPP1 genes.

SAPs were detected by BLAST searches. Therefore, the presence of recombination genes in plasmids is exceptional, and it is not known whether the mentioned recombination proteins play any role during propagation of their plasmid replicons.

Despite the pairwise occurrence of exonuclease/SAP genes in many phage genomes, other genes encoding (putative) recombination proteins are rarely found close by. Among the few exceptions are the RdgC-like protein encoded by the gene upstream of the RecE-homologue of a *Burkholderia cepacia* prophage (locus Bucepa02006134, not included in Fig. C22), and a RusA-like protein encoded upstream of the Erf-like protein encoded by a *S. pyogenes* prophage (locus spyM18_1796). SSB and replication genes, however, are found in many genomes of Gram(+)-specific phages in close vicinity of recombination genes, in some cases apparently within the same transcription unit, e.g. in ϕA118 (BRM Section 3.1.).

Holliday junction resolvases. DNA endonucleases that recognise and cleave the specific structure of four-way or crossover junctions have been termed Holliday junction resolvases. In the generally accepted model, the resolution of the Holliday junction(s) demarks the penultimate stage of homologous recombination, which is i. initiated by a 'loose end' and RecA-driven strand exchange/invasion, ii. leads to formation of a Holliday junction upon reciprocal strand exchange, iii. is continued by branch migration, and iv. is terminated by gap closure after resolution of Holliday junctions. In *E. coli*, branch migration is driven by the action of the RuvAB helicase, while Holliday junctions are resolved by the action of RuvC [61]. Recent experiments point to the possibility that Holliday junction resolution by RuvC may also trigger the initiation of RDR [403,281]. Replication fork arrest may lead to fork regression and the annealing of the newly synthesised strands, i.e. the formation of a Holliday junction as the 'ankle' of a chicken-foot structure, literally spoken. When the repair function of exonuclease V (RecBCD) is disabled, RuvC sets the strand scission(s) required to re-initiate replication via RDR. We speculate that a similar mechanism may also operate during the switch from θDR to σDR required to obtain phage genome concatemers (see above).

Phage-encoded Holliday junction resolvases include ϕT4 endonuclease VII (gp49), ϕT7 endonuclease I (gene 3), λ Rap, orf3 protein with limited similarity to *E. coli* RuvC from the lactococcal phage ϕbIL66, and RusA, encoded by the *E. coli* K12 cryptic prophage DLP12. Despite their highly divergent three-dimensional structures and unrelated protein sequences all these enzymes form dimers and require metal ions for enzymatic activity. The structure and func-tion of the phage-encoded Holliday junction resolvases have been comprehensively reviewed by Sharples who also defined the five groups of phage-encoded Holliday junction resolvases [408]. Therefore we confine ourselves here to a discussion of the many new sequences detected by BLAST searches for the five junction resolvase groups.

Gene 49 (gp49) endonuclease VII of phage T4. The crystal structure of the dimeric ϕT4 endonuclease VII is known [PDB 1EN7], and functionally important residues have been determined: D40, H41, N62, and E65 are closely located to the catalytic core, residues F72 and W87 in the C-terminal region are involved in dimerisation [35,36]. Endonuclease VII contains four cysteine residues in a specific arrangement that is also found in several type II restriction endonucleases. All phages of the ϕT4-group encode a homologous of endonuclease VII except for *Aeromonas* sp. phages ϕ65 and ϕ25 (sequences not yet completed) (Table C33). Motif E but not motif D is conserved in all endonuclease VII-type proteins detected by BLAST searches. The dimerisation of the proteins lacking a detectable motif C may follow a route different from that of ϕT4 gp49.

Notably, phages that encode a homologue of endonuclease VII but a DNA polymerase of the Pol I-type include ϕD29, ϕL5, ϕK1-5, and ϕSP6 (Section C3.5.1.). The *yajD* gene of the insect endosymbiont *Sodalis glossinidius* (γ-Proteobacteria) encodes a protein with similarity to ϕT4 gp49. By using *S. glossinidius* YajD as query, BLAST searches readily revealed homologues in the genomes of *E. coli*, *S. enterica*, *Yersinia pestis*, several other γ-Proteobacteria, and – notably – several *Streptomyces* species/strains. The function of these YajD proteins is not known but the conservation of the size and the structural motifs suggests that it may be closely related to the function of ϕT4 endonuclease VII. In the *E. coli* K12 genome, the *yajD* gene is located closely downstream of the *sbcC* and *sbcD* genes, which are related to the ϕT4 gp46/gp47 proteins. This suggests that these genes were originally acquired together from a ϕT4-like phage. The presence of a ϕT4 gp49 homologue in the genomes of ϕD29, ϕL5, ϕK1-5, ϕSP6, and in the genomes of the other phages listed in Table C33

Table C34 ϕT7 endonuclease I (gene 3) homologues.

motifs:					A	B
consensus					xYTP**D**FhhP	hhh**ET**KGx
replicon	accession	gene	res.	ident.		
ϕ T7	P00641	gene 3	149	x	TYTP**D**FLLP	IFV**ET**KGL
ϕ A1122	NP_848275	p13	151	87%	VYTP**D**FLLP	IFI**ET**KGL
ϕ T3	P20314	gene 3	152	81%	LYTP**D**FLLP	IFI**ET**KGL
ϕ YeO3-12	NP_052083	gene 3	153	86%	LYTP**D**FLLP	IFI**ET**KGL
ϕ gh-1	NP_813757	p11	147	60%	KYTP**D**FVLA	III**ET**KGI
P. putida KT2440 pϕ	NP_744417	pp2268	141	50%	KYTP**D**FALA	IIV**ET**KGR
ϕ P60	NP_570326	P60_orf16	66	47%	KYTP**D**FFLP	VIL**E**VKGF
	NP_570327	P60_orf17	54	43%		
ϕ SIO1	NP_064756	p19	134	47%	TYTP**D**FVLP	VII**ET**KGR

The motifs A + B represent highly conserved stretches identified by BLAST searches with ϕT7 gene 3 protein as query; conserved functional residues are shown in bold type [157].

suggests, in addition, that a *yajD* gene was acquired more than once by phages via HGT from a chromosomal copy – possibly even three times (!) by ϕOmega. This applies also to ϕYeO3-12, which encodes for two different (putative) Holliday junction resolvases: p47 as a protein similar to ϕT4 gp49, and a homologue of ϕT7 gene 3 protein. However, the issue of the spread of the ϕT4 gp49- like proteins among bacterial chromosomes and phages requires a deeper analysis before a firm statement can be made.

Gene 3 endonuclease I of phage T7. The crystal structure of the dimeric ϕT7 endonuclease I is known [PDB 1M0D], and functionally important residues have been determined: D55, E65, and T66 are closely located to the catalytic core and involved in manganese ion binding (Table C34) [157]. With two exceptions, the phages encoding a resolvase similar to ϕT7 gene 3 protein also encode a DNA polymerase with significant similarity to ϕT7 DNA polymerase (Section C3.5.1 and Table C25). A contiguous endonuclease protein is probably synthesised from the overlapping ϕP60 genes orf16 (N-terminus) and orf 17 (C-terminus), and the split into two genes may have resulted from a sequencing error (F. Chen; pers. communication) [75]. Phages ϕSP6 and ϕK1-5 encode ϕT7-type DNA polymerases but their cognate resolvases are similar to ϕT4 endonuclease VII (see above). *Gene*s encoding resolvases have not yet been identified in the genomes of ϕPaP3 and ϕFelix01, which also encode DNA polymerases of the ϕT7-type (see Table C25).

Rap protein of phage Lambda. The λ Rap(NinG) gene was identified by genetic analysis of recombination deficient λ mutants [178]. The function of Rap as a Holliday junction resolvase was confirmed by biochemical analysis [409]. The three-dimensional structure of λ Rap is not yet known. The particular arrangement of 8 cysteine residues in the λ Rap protein was proposed to mediate zinc ion binding, and is conserved throughout the entire group of Rap-like proteins (see Table C35) [408].

Two variants exist among the Rap-like proteins, which differ in the spacing of the two cysteine residues in motif A (Table C35). Because this difference in spacing is found also among otherwise highly conserved Rap homologues (> 90% ident. res.) it might not have functional consequences. In contrast to the RusA-like resolvases, the presence of a Rap encoding genes is entirely confined to Gram(–)-specific phages (see below). Interestingly – and also in contrast to the RusA proteins – only two distinct similarity groups exist among the Rap proteins: the proteins with almost perfect identity to λ Rap and the proteins with significant similarity, i.e with identities ranging from >30% to <50% (listed separately in Table C35). A group of proteins with similarities lower than 30% is missing, which was found for almost all phage recombination and replication proteins analysed in this study. This observation indicates that either the expected low-similarity genes are still missing from the data sources because the candidate phages have not yet been identified, or that the Rap gene is a very recent evolutionary development for which the precursor has yet to been identified.

Phage-encoded resolvases with similarity to E. coli *RuvC.* Chopin and co-workers could show that the orf2 and orf3 proteins of ϕbIL66 are responsible for a struc-

Table C33 ϕT4 endonuclease VII (gp49) homologues.

motifs:					A		B	C	D	E
consensus					xx**C**hh**C**xx		xh**DH**DHxxx	hh**C**xx**C**N xx**E**xxx	xx**h**xx	xhx**a**hx
replicon	accession	gene	res.	ident.						
ϕ T4	P13340	gp49	157	x	GK**C**LI**C**QR	27	HL**DH**DHELN	LL**C**NL**C**NAA**E**GQM	HK**F**NR	YLE**W**LE
ϕ RB69	NP_861776	gp49	157	89%	GK**C**PI**C**HR	27	HL**DH**DHELN	LL**C**NL**C**NAA**E**GQM	HK**F**NR	YLE**W**LE
ϕ RB49	NP_891632	p061	157	44%	GI**C**PL**C**QL	27	HL**DH**DHDLD	LLHSC**C**NRL**E**GSL	GD**F**KR	FVV**W**AT
ϕ KVP40	NP_899379	49	151	38%	DR**C**VL**C**GT	22	HT**DH**DHNTG	VL**C**RA**C**NTY**E**GVV	HK**F**TR	YVQ**W**LR
ϕ Aeh1	NP_943932	gp49	161	49%	GI**C**PL**C**KR	27	HL**DH**DHELN	LL**C**CF**C**NKF**E**GMV	HE**F**MR	YITQLK
ϕ 44RR2.8t	NP_932421	gp49	157	48%	GC**C**KL**C**KN	27	HL**DH**DHALE	LL**C**AR**C**NML**E**GMI	HK**F**IR	YIE**W**MR
ϕ D29	O64250	gp59	153	[34%]	GR**C**YI**C**QR	22	SV**DH**DHKTG	LL**C**TM**C**NKYILGW	-	CIEMLQ
ϕ L5	Q05272	L5p56	164	[41%]	GR**C**YI**C**QR	21	SV**DH**DHKTG	LL**C**TM**C**NKYTLGW	-	CIE**F**FK
ϕ K1-5	AAR90063	gp21	136	[36%]	WK**C**PL**C**GG	22	VL**DH**DHETG	VV**C**RG**C**NGA**E**GKI	GV**I**SG	QLQ**W**LE
ϕ SP6	NP_853582	gp22	136	[36%]	WK**C**PL**C**GG	22	VL**DH**DHETG	VV**C**RG**C**NGA**E**GKI	GV**I**SG	QLK**W**LE
ϕ S-PM2	AAP92668	g49	142	[40%]	GV**C**AI**C**KG	22	CV**DH**DHETG	LL**C**RN**C**NMMLGQV	-	-
ϕ Omega	NP_818437	gp138	160	[30%]	GR**C**AL**C**QR	21	AVE**H**DHKTG	IC**C**GP**C**NLGVLGH	-	DIA**F**FE
	NP_818388	gp89	166	[42%]	CANPG**C**RA	29	HL**DH**DHDCC	IL**C**PA**C**NLMLGSA	-	FRL**Y**GK
	NP_818499	gp199	229	[36%]	GK**C**ANPGC	33	PV**DH**DHACC	IL**C**PP**C**NTTLGQM	KR**L**RG	LRF**W**GV
ϕ KMV	NP_877462	gp23	146	[30%]	KL**C**PL**C**GK	23	VM**DH**DHETG	VLHRS**C**NTA**E**GKI	-	IIP**Y**LH
S. coelicolor A3(2)	AAC64277	orf8	170	[38%]	GL**C**VI**C**LK	17	HV**DH**CHKTG	VL**C**FN**C**NSAIGKL	-	GTS**W**KP
ϕ Bxz2	AAN01813	gp59	158	[43%]	GR**C**YG**C**RR	21	SV**DH**DHETG	LL**C**TA**C**NRNVLGH	-	FID**Y**LD
ϕ Bxb1	NP_075321	gp54	158	[35%]	GR**C**YI**C**RK	21	AV**DH**DHRTG	LLDTP**C**NRNVLGH	HL**G**DD	GID**Y**LE
ϕ Corndog	NP_817862	gp11	161	[38%]	GK**C**AI**C**QI	35	AV**DH**DHKMA	LL**C**GP**C**NQGIGRW	AL**V**RA	IQA**Y**TR
ϕ Xp10	NP_858983	35L	135	[26%]	GR**C**AI**C**QG	21	CL**DH**NHSTG	VLHRG**C**NSVLGLV	GV**F**AR	LAG**Y**LR
ϕ YeO3-12	NP_072072	p47	129	[31%]	GV**C**AI**C**GI	22	AV**DH**CHATG	LL**C**SA**C**NIALGKF	-	AIT**Y**LR
S. glossinidius	AAS66876	*yajD*	113	[30%]	WV**C**GR**C**SR	28	HI**DH**DHSNN	ML**C**LY**C**HDH**E**HEK	-	ASL**Y**GT
E. coli K12	P19678	*yajD*	115	-	WV**C**GR**C**SR	28	HI**DH**DHTNN	LL**C**LY**C**HDH**E**HSK	-	ADQ**Y**GT

The motifs A - C represent highly conserved stretches identified by BLAST searches with ϕT4 gp49 protein as query; conserved functional residues are shown in bold type [357]. the 4 conserved cysteine residues involved in metal ion binding in motifs A + C are indicated by gray shading, and spacing of the motifs shown in a separate clolum. Motifs C + D contain aromatic residues involved in protein dimerisation. In those cases where the region of similarity did not encompass the entire protein sequences, the identity values are shown in square brackets.

ture-specific endonuclease activity; while orf3 protein shows similarity to *E. coli* RuvC, the contribution of orf2 protein to this activity is still obscure [33]. BLAST searches revealed a group of orf3 homologues among lactococcal phages, and a second group of proteins with significant similarity (Table C36). All proteins share with *E. coli* RuvC the signature motifs A – D and F, but lack motif E*.

As was pointed out by Sharples and co-workers, *E. coli* RuvC [PDB 1HJR] and ϕbIL66 orf3 protein share several conserved catalytically important residues in a reasonably related structural environment but differ considerably outside this 'core', therefore simple BLAST searches with ϕbIL66 orf3 as query fail to detect similarity to *E. coli* RuvC (Table C36) [410]. These authors suppose that the structural differences between the two proteins reflect their different substrate specificities: while RuvC is highly specific for 'X' structures, i.e. four-way junctions, ϕbIL66 orf3 protein can cleave branched 'Y' structure, in addition. It should be interesting to include in future studies aimed to clarify the molecular basis of this functional divergence two proteins detected by our BLAST searches: *Fusobacterium nucleatum* RuvC and the protein encoded by locus psyc163101 of *Psychrobacter* sp. (Table C36; third group). Both proteins possess motif E* but show a reasonable similarity to ϕbIL66 orf3 protein (~25% ident. res.). However, the protein from *Psychrobacter* sp. is more closely related to *E. coli* RuvC (55% ident. res.) than to ϕbIL66 orf3 protein, but *F. nucleatum* RuvC is equally distantly related to both.

A fourth group of proteins with significant similarity to ϕbIL170 p63 is encoded exclusively by lactococcal phages. The signature motifs A – D are well conserved

Table C35 Phage- and prophage-encoded λ Rap-like proteins.

motifs:					A		B		C		D
consensus					RCx$_{2-4}$ Cx		hCxxxCG		PChSCG		QCxxCN
replicon	accession	gene	res.	ident.							
λ	P03770	Rap / NinG	204	x	RCKNDECR	11	WCSPECG	66	PCISCG	30	QCVVCN
ϕ 21	Q9XJQ4	NinG	204	98%	RCKNDECR	11	WCSPECG	66	PCISCG	30	QCVVCN
ϕ Sf6	NP_958228	gene 54	203	98%	RCKNDECR	11	WCSPECG	66	PCISCG	30	QCVVCN
ϕ ST104	YP_006391	NinG	203	96%	RCKNEECR	11	WCSPECG	66	PCVSCG	30	QCVVCN
ϕ P22	NP_059617	NinG	203	94%	RCKNEECR	11	WCSPECG	66	PCISCG	30	QCVVCN
pϕ CP-933K	NP_286501	z0953	207	89%	RCKNDECR	11	WCSPECG	69	PCISCG	30	QCVVCN
ϕ Nil2	CAC95102	NinG	201	91%	KCKI..CK	11	WCSPEHG	66	PCISCG	30	QCVVCN
ϕ LC159	AAN59917	orf2	190	91%	KCKI..CK	11	WCSPEHG	66	PCISCG	30	QCVVCN
ϕ H-19B	O48427	NinG	201	91%	KCKI..CK	11	WCSPEHG	66	PCISSG	30	QCVVCN
ϕ 933W / ϕ VT2-Sa	NP_049497	p37	201	92%	KCKI..CK	11	WCCPEHG	66	PCISCG	30	QCVVCN
ϕ 4795	NP_851974	ORF37	201	92%	KCKI..CK	11	WCCPEHG	66	PCISCG	30	QCVVCN
pϕ CP-933V	NP_288675	z3346	201	90%	KCKI..CK	11	WCCPEHG	66	PCISCG	30	QCVVCN
pϕ Gifsy-2	NP_459996	stm1021	203	47%	KCANKECR	11	VCSYQCA	68	GCISCG	30	QCDVYN
Y. pestis pϕ	NP_405642	ypo2090	201	41%	NCKV..CK	11	WCDEEHK	66	ACVSCG	30	QCAPCN
H. somnus pϕ	ZP_00122672	hsom0668	203	40%	KCKV..CG	11	WCSPECG	59	PCISCG	30	QCSVCN
B. cepacia pϕ?	ZP_00219525	bucepa02005864	190	37%	KCRE..CG	11	VCSPVCA	55	PCISCG	28	QCGPCN
P. syringae pϕ	ZP_00127017	psyr023592	194	38%	TCDNPACG	11	VCGWACG	55	PCVSCG	30	QCAPCN
ϕ D3	NP_061578	orf82	213	35%	KCQNTECG	11	VCSPACA	57	GCISCG	41	QCKACN
ϕ Aaϕ23	NP_852744	NinG	189	34%	KCKS..CG	11	VCSPKCA	60	PCIACG	30	GCIRCN

The motifs A - D represent the conserved stretches containing the 8 cysteine residues identified by BLAST searches with λ Rap as query; the cysteine residues potentially involved in metal ion binding indicated by gray shading, and spacing of the motifs shown in separate clolums.

in these proteins but they lack ~60–110 C-terminal residues as compared to ϕbIL170 p63 (Table C36). The function of these proteins remains to be studied but it seems doubtful whether they are functional Holliday junction resolvases. Interestingly, ϕvML3 contains genes for the longer as well as the shorter version of this protein; both proteins possess a truncated motif A at the extreme N-terminus.

Phage-encoded resolvases with similarity to E. coli *RusA.* In *E. coli ruvABC* mutants, homologous recombination can be rescued by activation of the *rusA* locus. The *rusA* gene is located in the region of the cryptic prophage DLP12 of *E. coli* K-12, and dimeric RusA has been shown by biochemical analysis to possess Holliday junction resolvase activity (reviewed in [408]). *E. coli* RusA has been crystallised [PDB 1Q8R], and the functionally important residues were determined: D70, D72 and D91 are likely to be involved in forming a magnesium ion-binding pocket involving both subunits of the RusA dimer [39,358]. In 2001, Sharples reported 11 homologues of RusA in the databases, but the number has increased in the meantime to >50 [408] (see Table C37).

We carried out the initial BLAST search for RusA-like proteins using *E. coli* RusA as query; in several subsequent steps, the most distantly related proteins were used as query until no further sequences with reasonable similarity were detected. With the exception of the RusA-like proteins of *A. aeolicus* (locus aq_1953) and *Bacillus* sp. plasmid pSL32 (see below), all proteins similar to *E. coli* RusA detected by BLAST searches are either phage-encoded or contained within (putative) prophage regions in bacterial chromosomes from a wide range of species. All RusA-like proteins have a comparable size, and reasonably well conserved motifs A + B. In one group however, the spacing within motif A differs by 3 residues; the functional consequence of this different spacing is not known.

The low copy-number plasmid pLS32 (~70 kb) of *Bacillus natto* encodes a protein of unknown function, which is highly similar to *E. coli* RusA (28% ident. res.) despite its longer C-terminus [442]. The N-termini of the putative replication initiator RepN of pLS32 and the ORF4 replication initiator of ϕ7201 (Section C3.1.2.) share significant sequence similarity (38% ident. res.). The replication origin of pLS32 was mapped within the

Table C36 Phage- and prophage-encoded RuvC-like proteins.

motifs:					**A**	**B**	**C**	**D**	**E***	**F**
consensus					LxhDhS	xTGYAHx	xERxx	hxDYx$_{1-4}$hxhExx	xQhKxxh	DDxxDAh
replicon	accession	gene	res.	ident.						
ϕ bIL170	NP_047175	p63	160	x	LAI**D**FS	NTGYAFR	LERAK	LFDYF..IYI**E**EP	ISNSMWC	D**D**MA**D**AF
ϕ sk1	NP_044999	ORF53	160	91%	LAI**D**FS	NTGYAFR	LERAK	LFDYF..IYI**E**EP	VPNSKWC	D**D**MA**D**AF
ϕ bIL66 / ϕ bIL67	AAA99046	orf3	160	92%	LAI**D**FS	NTGYAFR	LERAK	LFDYF..IYI**E**EP	IPNSKWC	D**D**QA**D**AF
ϕ 712	AAC63027	M3	161	95%	LAI**D**FS	NTGYAFR	LERAK	LFDYF..IYI**E**EP	ISNSMWC	D**D**QA**D**AF
ϕ c2	NP_043550	gene 12	161	45%	LAI**D**FS	GTGYAFR	WERTF	LKDYH..MAI**E**TP	IDNSKWC	DNMA**D**AY
ϕ vML3	P13004	-	152	44%	EFPS	GTGYAFR	WERTF	LKDYH..MAI**E**TP	IDNSKWC	DNMA**D**AY
ϕ 315.3	NP_795506	SpyM3_1128	170	24%	LSL**D**VS	GTGWAIF	YERAK	PFEDI...VI**E**QN	VNVSTWR	D**D**EA**D**AI
S. pyogenes	NP_607564	spym18_1491	170	23%	LSL**D**IS	GTGWALF	FERGR	LKKYY..CAV**E**KN	INVSTWR	D**D**EA**D**AI
ϕ EJ-1	NP_945263	p24	171	22%	LSL**D**IS	ATGWAVF	FERGR	HFESI...VV**E**KN	WTWRKYW	D**D**EA**D**AI
F. nucleatum	Q8RGS0	*ruvC*	190	24%	IGI**D**PG	IVGYGII	EERLE	YKPEF..MAI**E**DL	L**Q**V**K**IGI	D**D**AA**D**AL
Psychrobacter sp.	ZP_00145849	psyc163101	201	23%	IGI**D**PG	MTGYGIL	PERLK	YADEPIYTAI**E**QV	R**Q**I**K**QAV	Q**D**AA**D**GL
E. coli	NP_047175	*ruvC*	160	-	LGI**D**PG	VTGYGVI	PSRLK	QPDYF...AI**E**QV	R**Q**V**K**QTV	A**D**AA**D**AL
ϕ 923	AAP80756	orf L2	93	[41%]	LAI**D**FS	GTGYAFR	WERTF	LKDYH..MAI**E**TP	-	-
ϕ 5447 / ϕ 5469	AAN05716	L2	85	[41%]	LAI**D**FS	GTGYAFR	WERTF	LKDYH..MAI**E**TP	-	-
ϕ c6A	AAN05724	L2	84	[41%]	LAI**D**FS	GTGYAFR	WERTF	LKDYH..LAI**E**TP	-	-
ϕ 5440	AAN05712	L2	84	[40%]	LAI**D**FS	GTGYAFR	WERTF	LKDYH..MAI**E**TP	-	-
ϕ 943	AAN05707	L2	84	[40%]	LAI**D**FS	GTGYAFR	WERTF	LKDYH..MAI**E**TP	-	-
ϕ vML3	S05334	-	97	[38%]	EFPS	GTGYAFR	WERTF	LKDYH..MAI**E**TP	-	-

Motif B + C represent highly conserved stretches identified by BLAST searches with ϕbIL170 p63 protein as query. Motifs A, D, E* + F correspond to the RuvC signature motifs I - IV defined by Lilley + White [251]; the catalytically important residues are shown in bold type. The similarity values for protein sequences matching with the N-terminus of ϕbIL170 p63 only are shown in square brackets.

repN gene, and the plasmid may therefore represent a naturally occurring phage-derived plasmid replicon [442]. Alternatively – and in (weak) analogy to *E. coli* ϕP1 – pLS32 could be the prophage state of a yet unknown *Bacillus* sp. phage. It should be highly interesting to analyse the (putative) mechanism for copy number control in pLS32 because phage origins of the type apparently operating in this plasmid usually give rise to much higher copy numbers as have been determined for pSL32 (2-3 per chromosome)[442]. Other plasmid-encoded RusA-like proteins – or other Holliday junction resolvases – were not detected in extensive BLAST searches.

The discussion of phage-encoded Holliday junction resolvases shows that genes encoding such proteins are widespread among phages. Presently, five groups of resolvases are known, which share no structural features and are evolutionary only distantly related, if at all (but see [251]). This argues against the possibility that more and novel resolvases will be simply detectable by *in silico* analyses. Except for the Rap resolvase, which is only found in the genomes of phages and prophages with significant similarity to λ, the resolvases possibly spread among the various phage families by HGT, but there remain gaps. The known resolvase genes do not cover the entire spectrum of known phages: resolvase genes have not yet identified ϕSPP1 (but in ϕA118), ϕPaP3 and ϕFelix01 (but in ϕT7).

Other phage-encoded recombination proteins. We conclude the discussion of recombination proteins with a brief survey of phage-encoded homologues of the 'canonical' *E. coli* recombination proteins [230,61]. Several of them are probably not directly involved in chromosome replication, but we were curious to learn more about the possible mutual exchange of recombination protein-encoding genes between chromosomes and phage replicons.

RecA. Among the most highly conserved bacterial proteins, RecA has to date only two detectable homologues in phage replicons [267]: the gp117 protein of Mycobacteriophage ϕCJW1 (≥60% ident. res. with RecA proteins of several *Mycobacterium* sp.), and the gp201

Table C37 Phage- and prophage-encoded RusA-like proteins.

motifs:					A	B
consensus					xx**D**x**D**x$_{3/6}$	**K**xxx**D**	xhhx**DD**xxhx
replicon	accession	gene	res.	ident.		
pϕ DLP12 *E. coli* K12	P40116	*rusA*	120	x	RR**D**L**D**NLQ...**K**AAFD	GFWLD**D**AQVV
pϕ CP-933X *E. coli*	NP_287358	z1873	120	95%	RR**D**L**D**NLQ...**K**AAFD	GFWLD**D**AQVV
ϕ 82	Q37873	*rusA*	120	95%	RR**D**L**D**NLQ...**K**ASFD	GFWLD**D**AQVV
ϕ ST64T	NP_720315	Rus	131	53%	KR**D**L**D**NLP...**K**AVFD	GFWLD**D**GQID
ϕ HK620	NP_112065	*rus*	120	43%	RR**D**L**D**NIL...**K**GLLD	GFAED**D**EQFD
ϕ HK97	Q9MCN8	gp67	120	43%	RR**D**L**D**NIL...**K**GLLD	GFAED**D**EQFD
E. coli pϕ CP-933U	NP_288470	z3115	119	41%	RR**D**L**D**NIL...**K**APLD	GLLMD**D**EQFD
E. coli pϕ CP-933P	NP_287995	z6061	119	40%	RR**D**L**D**NIL...**K**APLD	GLLID**D**EQFD
E. coli pϕ CP-933O	NP_287519	z2057	119	39%	RR**D**L**D**NIL...**K**APLD	GVLMD**D**EQFD
E. coli pϕ CP-933M	NP_286866	z1344	124	41%	RR**D**L**D**NIL...**K**APLD	GLLMD**D**EQFD
E. coli pϕ CP-933N	NP_287277	z1785	123	40%	RR**D**L**D**NIL...**K**APLD	EVLID**D**EQFD
E. coli pϕ CP-933O	NP_287559	z2101	101	42%	RR**D**L**D**NIL...**K**APLD	GLLID**D**EQFD
ϕ SfV	NP_599075	orf43	129	38%	IR**D**L**D**NYN...**K**ALFD	GVWED**D**SQVK
ϕ ST64B	NP_700418	sb45	129	41%	IR**D**L**D**NYN...**K**ALFD	GVWED**D**SQVK
L. pneumophila pϕ	CAC33459	ORF G	120	31%	RR**D**I**D**NIL...**K**CLLD	GVYDD**D**NQID
ϕ Bcep1	NP_944331	gp23	131	41%	AW**D**VANRE...**K**LLSD	GFWRD**D**SQID
ϕ Bcep781	NP_705648	ORF20	136	33%	AW**D**VANRE...**K**LICD	GFWRD**D**SQID
ϕ Bcep43	NP_958127	gp23	128	33%	AW**D**VANRE...**K**LICD	GFWRD**D**SQID
pLS32	BAA24869	-	185	28%	RK**D**PNNFL...**K**VPFD	GVYLD**D**DVAL
A. aeolicus	NP_214335	aq_1953	152	26%	KR**D**I**D**NML...**K**SLWD	GVIKN**D**NLIF
S. pyogenes M1 GAS pϕ	NP_268913	spy0673	131	[64%]	KP**D**T**D**NLQ...**K**LLKD	GFWND**D**AQVA
ϕ TM4	NP_569805	gp72	213	[46%]	RP**D**L**D**KLA...RAILD	VVWID**D**SQVV
ϕ NIH1.1	NP_438128	15	117	[34%]	KP**D**L**D**NLE...**K**AVYD	VVWTD**D**NIIV
ϕ 315.4	NP_665052	SpyM3_1248	146	[34%]	KP**D**L**D**NLE...**K**AVYD	VVWTD**D**NIIV
ϕ r1t	NP_695042	orf14	131	[42%]	RP**D**L**D**NLM...**K**NLQD	RYYSD**D**SQIV
ϕ bIL285 / ϕ LC3	NP_076590	*rusA*	129	[42%]	RP**D**L**D**NLM...**K**NLQD	RYYSD**D**SQIV
ϕ Tuc2009 / ϕ ul36	NP_108696	orf19	129	[42%]	RP**D**L**D**NLM...**K**NLQD	RYYSD**D**SQIV
ϕ BK5-T	NP_116543	ORF51	133	[42%]	RP**D**L**D**NLM...**K**NLQD	RYYSD**D**SQIV
S. agalactiae pϕ	NP_687598	SAG0569	158	[33%]	KP**D**I**D**NLI...**K**AVFD	IVWSD**D**NIVC
ϕ 7201	NP_038311	orf10	153	[33%]	KP**D**I**D**NLI...**K**ALFD	IVWTD**D**SIVC
R. eutropha pϕ	ZP_00168225	Raeut330401	157	[40%]	KP**D**A**D**NVL...**K**AVKD	IVWVD**D**AQAV
ϕ PVL / *L. innocua* pϕ	NP_058487	orf 48	145	23%	RP**D**I**D**NYV...**K**AILD	IMFSD**D**GKIV
ϕ LL-H	AAL77552	orf139	139	25%	KP**D**T**D**NIE...**K**GIYD	LAYED**D**KQIV
C. tetani pϕ	NP_782962	ctc02433	222	27%	HP**D**TINIT...**K**SIFD	GVIIN**D**AQIR
ϕ Bxz1	NP_818256	gp205	151	-	KP**D**L**D**KLI...RAIGD	TVYTE**D**SRIV
N. aromativorans pϕ	ZP_00094972	saro1995	135	-	KP**D**G**D**NIL...**K**ALGD	IVWAD**D**SQVA
B. cepacia pϕ	ZP_00218864	Bucepa02006569	158	-	KP**D**A**D**NVV...**K**ALKD	VVYGD**D**GQVV
ϕ N315	NP_835540	sa1789	134	-	KP**D**I**D**NLI...**K**TVLD	HVWKD**D**NQIT
B. subtilis pϕ	P45911	*yqaN*	142	-	KP**D**V**D**NYV...**K**GVKD	LIYKD**D**SQVV
E. faecalis pϕ	NP_815791	ef2128	141	-	KP**D**L**D**NYF...**K**AVTD	LYKN**D**GQIAV
ϕ 31.1	AAF43122	orf139a	139	29%	RS**D**P**D**NLQPTL**K**AIMD	GLWSD**D**NHEV
ϕ ul36.1	AAF74065	orf109	109	[40%]	RS**D**P**D**NLQPTL**K**AIMD	GLWSD**D**NHEV
ϕ TP901-1	NP_112678	*rus*	139	29%	RS**D**P**D**NLQPTL**K**AIMD	GLWSD**D**NHEV
X. fastidiosa pϕ	ZP_00038274	xfasa0197	153	29%	LP**D**D**D**NMLARF**K**PYRD	ALGID**D**RRFV
ϕ A118	NP_463528	ORF63	134	-	KK**D**P**D**NIAFAK**K**FIFD	GFLEN**D**NLNY
ϕ MM1	NP_150151	MM1p20	133	[30%]	KL**D**PPNLYPTV**K**AIID	GIWTD**D**NHKV
ϕ Che9c	NP_817741	gp64	127	-	RR**D**NHNLWPTV**K**ALVD	GVVPD**D**DTEH
ϕ bIL309	NP_076712	orf17	136	-	KY**D**PPNYEPTS**K**ALID	GIWND**D**NYNV

For the prophages of the CP-933 series, only the genes of *E. coli* strain O157:H7 EDL933 but not for *S. flexneri* are listed.

protein of Mycobacteriophage ϕBxz1 . Gp117 can serve as example for recent horizontal gene transfer (HGT), i.e. the acquisition of a host gene by a phage replicon [301]. The origin of gp201 is more difficult to trace because it shows only ~30% similarity (ident. res.) to gp117 as well as to various bacterial RecA proteins of ϕBxz1 hosts. The significant divergence of the gp201 sequence from other RecA proteins suggests that it was acquired by ϕBxz1 much earlier than gp117 by ϕCJW1. Whether CJW1 gp117 and ϕBxz1 gp201 still function as SAPs has not yet been analysed.

*Exonuclease V (*recB, recC, recD*).* No phage-encoded proteins with significant similarity to the *E. coli* RecB and RecC proteins were detected by BLAST searches, and also the RecD protein shares only some limited similarity with the Dda proteins of the ϕT4-group of phages in the helicase domain (SF1-type) [419] (Section C3.3.).

RecJ. The YorK protein of *B. subtilis* ϕSPBc2 is the only phage-encoded protein with similarity (23% ident. res.; full-length) to *E. coli* RecJ . The *yorK* gene is located adjacent to the *yorL* gene encoding the DNA polymerase of ϕSPBc2; it has not been studied whether the putative 5'→3' exonuclease activity of the YorK protein is required for ϕSPBc2 replication, recombination, or both.

RecQ. Homologues of the *E. coli* RecQ helicase (superfamily II) are found in the genomes of bacteria from all phyla, and also in the genomes of archaea and eukarya. The only viral RecQ homologue (cac19131) is encoded by the fungal virus DpAV4, and was probably acquired from its host. The various phage-encoded SF2 helicases (Section C3.3.) may be functional analogues of RecQ but significant sequence similarity to *E. coli* RecQ could not be detected.

*RdgC. Gene*s encoding homologues of the nucleoid-associated protein RdgC are only found in proteobacterial genomes. This protein modulates RecA binding to ssDNA in concert with SSB and the RecFOR recombination proteins, and has been characterised only recently [380,300]. Phage-encoded homologues of RdgC are the HrdC protein of ϕP1 (61% ident. res.), ORF10 protein of *Vibrio harveyi* phage ϕVHML (41% ident. res.), and probably p45 protein of *Yersinia* sp. phage ϕPY54 (28% ident. res. in the N-terminal half).

SbcC, SbcD. The *E. coli* SbcB and SbcD proteins form a heterodimer with a complex substrate-specific exo- and endonuclease activity, and with a particular role in the removal of hairpin DNA [83]. SbcC and SbcD homologues are well conserved in bacterial genomes from most phyla, and found also in archaeal and eukaryal genomes. The only detectable phage-encoded proteins with significant similarity to SbcC/ SbcD are the gp46/gp47 proteins of phages of the ϕT4-group (see above).

RuvA, RuvB, and RuvC. No phage-encoded homologues of the RuvA and RuvB proteins were found by BLAST searches. The Holliday junction resolvase encoded by *L. lactis* phage ϕbIL66 shares the catalytic core residues with RuvC [33]; homologues of ϕbIL66 orf3 protein are found in several other lactococcal phages and prophages of *Streptococcus* sp. (see above).

BLAST searches did not reveal proteins with significant similarity to the *E. coli* RecF, RecO, RecR, RecG, RecN, exonuclease I (*sbcB*), helicase II (*uvrD/recL*), and helicase IV (helD) recombination proteins. Although the functional equivalence or similarity of several recombination pathways in bacteria and phages is well documented, this survey suggests that the genes/ proteins responsible for individual recombination pathways were kept largely separate during the long-lasting co-evolution of phages and their hosts. Clearly, the acquisition of bacterial genes encoding recombination proteins by phage replicons is the exception rather than the rule. Several prophage-encoded recombination proteins present in bacterial genomes were shown to constitute backup systems for mutated host recombination genes, but some require activation by mutation(s) in order to substitute for defective chromosomally encoded recombination proteins. There is thus far only one example where a phage recombination protein might have replaced a chromosomal recombination protein: the genome of *A. aeolicus* lacks detectable equivalents of the *E. coli* RuvABC proteins but encodes a homologue of the phage Holliday junction resolvase RusA [408].

C4. Concluding remarks

The still rapidly increasing number of completely sequenced genomes makes it impossible to predict a 'half-life' for the numerous figures and tables presented in this compendium. For quite a number of prophage genes present in bacterial chromosomes, however, we were able to deduce their function by (BLAST) sequence comparison combined with a search for signature motifs derived from homologues with known structure. In this respect, the compendium of phage replication origins and proteins may serve as a reliable starting-tool for function assignment for newly sequenced genomes at least for some time. On longer terms, it would be preferable to combine the knowledge-based approach presented here with automated approaches like the COG

(Clusters of Orthologous Groups of proteins) [444] or the CDD (Conserved Domains Database) [273] databases in an continuously updated online search tool.

In the introduction to the review (BRM Section 1.), we reason why a satisfactory definition of 'bacteriophages' is problematic. Also we mention that replication research has traditionally used phages as *experimental systems* rather than discussing their exact classification. Based on the results presented in this compendium, this somewhat sloppy approach seems largely justified: despite the enormous variability of phage-encoded replication functions, we rarely found plasmid- or chromosome-encoded homologues of phage replication proteins (except for prophages). There are only *very* few examples for gene exchange (HGT) between phage replicons and chromosomes. One example is the aquisition of the host *dnaB* gene encoding the replicative helicase by a number of lambdoid phages. Another example may be the (hypothetical) aquisition of *dnaC*-type helicase loader genes of lambdoid phages by several γ-proteobacteria and by the Bacillales. These issues are discussed in detail in BRM Sections 4.2. & 4.3., respectively. The activation of recombination functions encoded by resident prophages – e.g. *rusA*, *recET* – may serve as a third example, but in these cases one cannot speak of HGT in the strict sense.

There are also *very* few examples for an exchange of replication functions between phage and plasmid replicons. Clearly, plasmids and phages replicating by RCR share several features and their replication modules may have a common evolutionary origin. Phages however lack the intricate copy-number control mechanisms operating in RCR plasmids. Also, a closer inspection of the initiator genes encoded by RCR plasmids and phages reveals conserved structural motifs that are – with few exceptions – confined to either the phage-encoded or the plasmid-encoded proteins (Tables C8 + C9).

Coliphages P1 and P4 are among the best-studied naturally occurring joint replicons, but homology searches suggest that both are solitaires in the phage world. Plasmids pXO1, pTAV320, and pTA1060 encode (putative) replication proteins sharing reasonable similarity in their N-termini with several λ O-type initiator proteins. Since this N-terminal DNA-binding domain contains a HTH-motif, which is also present in a number of transcription factors with N-terminal DNA-binding domains, it would require a detailed analysis to determine whether these plasmid genes and the initiator genes of lambdoid phages have a common origin. The replication of two naturally occurring linear plasmids of *Streptomyces rochei* may be driven by phage-derived replication modules (Section C2.2.). Both encode λ O-type initiators with the replication origin contained within the initiator gene [71]. Both plasmids encode, in addition, a helicase of the $DnaB_{Eco}$-type in an arrangement typical for the 'IH-type' replication module (BRM Section 3.1.3.). Another example for a phage replication module driving plasmid replication may be the low copy-number plasmid pLS32 of *Bacillus natto* (Section C3.6.2.). Keeping these few exceptions in mind, we conclude that – with respect to replication – the term 'bacteriophage' can be used safely also in the future.

References

[1] Ajdic' D, McShan WM, McLaughlin RE, *et al*. (2002) Genome sequence of Streptococcus mutans UA159, a cariogenic dental pathogen. *Proc Natl Acad Sci USA* **99:** 14434-14439.

[2] Alano P, Dehò G, Sironi G & Zangrossi S (1986) Regulation of the plasmid state of the genetic element P4. *Mol Gen Genet* **203:** 445-450.

[3] Alley SC, Trakselis MA, Mayer MU, Ishmael FT, Jones AD & Benkovic SJ (2001) Building a replisome solution structure by elucidation of protein-protein interactions in the bacteriophage T4 DNA polymerase holoenzyme. *J Biol Chem* **276:** 39340-39349.

[4] Altermann E, Klein JR & Henrich B (1999) Primary structure and features of the genome of the *Lactobacillus gasseri* temperate bacteriophage (phi)adh. *Gene* **236:** 333-346.

[5] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

[6] Arai K & Kornberg A (1979) A general priming system employing only dnaB protein and primase for DNA replication. *Proc Natl Acad Sci USA* **76:** 4308-4312.

[7] Arai K & Kornberg A (1981) Mechanism of dnaB protein action. IV. *Gene*ral priming of DNA replication by dnaB protein and primase compared with RNA polymerase. *J Biol Chem* **256:** 5267-5272.

[8] Arai K, Yasuda S & Kornberg A (1981) Mechanism of dnaB protein action. I. Crystallization and properties of dnaB protein, an essential replication protein in Escherichia coli. *J Biol Chem* **256:** 5247-5252.

[9] Aravind L, Leipe DD & Koonin EV (1998) Toprim - a conserved ca-talytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. *Nucleic Acids Res* **26:** 4205-4213.

[10] Asano S, Higashitani A & Horiuchi K (1999) Filamentous phage replication initiator protein gpII forms a covalent complex with the 5' end of the nick it introduced. *Nucleic Acids Res* **27:** 1882-1889.

[11] Ayora S, Langer U & Alonso JC (1998) *Bacillus subtilis* DnaG primase stabilises the bacteriophage SPP1 G40P helicase-ssDNA complex. *Febs Lett* **439:** 59-62.

[12] Ayora S, Missich R, Mesa P, Lurz R, Yang S, Egelman EH & Alonso JC (2002) Homologous-pairing activity of the *Bacillus subtilis* bacteriophage SPP1 replication protein G35P. *J Biol Chem* **277:** 35969-35979.

[13] Ayora S, Stasiak A & Alonso JC (1999) The *Bacillus subtilis* bacteriophage SPP1 G39P delivers and activates the G40P DNA helicase upon interacting with the G38P-bound replication origin. *J Mol Biol* **288:** 71-85.
[14] Baas PD & Jansz HS (1976) Bacteriophage phiX174 DNA synthesis in a replication-deficient host: determination of the origin of phiX DNA replication. *J Mol Biol* **102:** 633-656.
[15] Backhaus H & Petri JB (1984) Sequence analysis of a region from the early right operon in phage P22 including the replication genes 18 and 12. *Gene* **32:** 289-303.
[16] Bailey S, Sedelnikova SE, Mesa P, Ayora S, Alonso JC & Rafferty JB (2003) Crystallization of the Bacillus subtilis SPP1 bacteriophage helicase loader protein G39P. *Acta Crystallogr D Biol Crystallogr* **59:** 1090-1092.
[17] Bailey S, Sedelnikova SE, Mesa P, *et al*. (2003) Structural analysis of Bacillus subtilis SPP1 phage helicase loader protein G39P. *J Biol Chem* **278:** 15304-15312.
[18] Baker TA & Bell SP (1998) Polymerases and the replisome: machines within machines. *Cell* **92:** 295-305.
[19] Bárcena M, Martin CS, Weise F, Ayora S, Alonso JC & Carazo JM (1998) Polymorphic quaternary organization of the *Bacillus subtilis* bacteriophage SPP1 replicative helicase (G*40* P). *J Mol Biol* **283:** 809-819.
[20] Bárcena M, Ruiz T, Donate LE, Brown SE, Dixon NE, Radermacher M & Carazo JM (2001) The DnaB.DnaC complex: a structure based on dimers assembled around an occluded channel. *EMBO J* **20:** 1462-1468.
[21] Barnes MH, Spacciapoli P, Li DH & Brown NC (1995) The 3'-5' exonuclease site of DNA polymerase III from gram-positive bacteria: definition of a novel motif structure. *Gene* **165:** 45-50.
[22] Barrett KJ, Gibbs W & Calendar R (1972) A transcribing activity induced by satellite phage P4. *Proc Natl Acad Sci USA* **69:** 2986-2990.
[23] Bates DB, Asai T, Cao Y, Chambers MW, Cadwell GW, Boye E & Kogoma T (1995) The DnaA box R4 in the minimal *oriC* is dispensable for initiation of *Escherichia coli* chromosome replication. *Nucleic Acids Res* **23:** 3119-3125.
[24] Bazill GW & Gross JD (1973) Mutagenic DNA polymerase in *B. subtilis*. *Nat New Biol* **243:** 241-243.
[25] Beese LS & Steitz TA (1991) Structural basis for the 3'-5' exonuclease activity of *Escherichia coli* DNA polymerase I: a two metal ion mechanism. *EMBO J* **10:** 25-33.
[26] Bendtsen JD, Nilsson AS & Lehnherr H (2002) Phylogenetic and functional analysis of the bacteriophage P1 single-stranded DNA-binding protein. *J Virol* **76:** 9695-9701.
[27] Benkovic SJ, Valentine AM & Salinas F (2001) Replisome-mediated DNA Replication. *Annu Rev Biochem* **70:** 181-208.
[28] Bernad A, Blanco L, Lazaro JM, Martin G & Salas M (1989) A conserved 3'-5' exonuclease active site in prokaryotic and eukaryotic DNA polymerases. *Cell* **59:** 219-228.
[29] Bernstein JA & Richardson CC (1988) A 7-kDa region of the bacteriophage T7 gene 4 protein is required for primase but not for helicase activity. *Proc Natl Acad Sci USA* **85:** 396-400.
[30] Bessman M, Kornberg A, Lehman I & Simms E (1956) Enzymic synthesis of deoxyribonucleic acid. *Biochim Biophys Acta* **21:** 197-198.
[31] Bhagwat M, Meara D & Nossal NG (1997) Identification of Residues of T4 RNase H Required for Catalysis and DNA Binding. *J Biol Chem* **272:** 28531-28538.
[32] Bhattacharyya S & Griep MA (2000) DnaB helicase affects the initiation specificity of *Escherichia coli* primase on single-stranded DNA templates. *Biochemistry* **39:** 745-752.
[33] Bidnenko E, Ehrlich SD & Chopin MC (1998) *Lactococcus lactis* phage operon coding for an endonuclease homologous to RuvC. *Mol Microbiol* **28:** 823-834.
[34] Bird LE, Pan H, Soultanas P & Wigley, DB (2000) Mapping protein-protein interactions within a stable complex of DNA primase and DnaB helicase from *Bacillus stearothermophilus*. *Biochemistry* **39:** 171-182.
[35] Birkenbihl RP & Kemper B (1998) Endonuclease VII has two DNA-binding sites each composed from one N- and one C-terminus provided by different subunits of the protein dimer. *EMBO J* **17:** 4527-4534.
[36] Birkenbihl RP & Kemper B (1998) Localization and characterization of the dimerization domain of holliday structure resolving endonuclease VII of phage T4. *J Mol Biol* **280:** 73-83.
[37] Biswas EE & Biswas SB (1999) Mechanism of DnaB helicase of *Escherichia coli*: Structural domains involved in ATP hydrolysis, DNA binding, and oligomerization. *Biochemistry* **38:** 10919-10928.
[38] Blanco L, Bernad A, Blasco MA & Salas M (1991) A general structure for DNA-dependent DNA polymerases. *Gene* **100:** 27-38.
[39] Bolt EL, Sharples GJ & Lloyd RG (1999) Identification of three aspartic acid residues essential for catalysis by the RusA holliday junction resolvase. *J Mol Biol* **286:** 403-415.
[40] Bonner CA, Randall SK, Rayssiguier C, Radman M, Eritja R, Kaplan BE, McEntee K & Goodman MF (1988) Purification and characterization of an inducible *Escherichia coli* DNA polymerase capable of insertion and bypass at abasic lesions in DNA. *J Biol Chem* **263:** 18946-18952.
[41] Bouché JP, Rowen L & Kornberg A (1978) The RNA primer synthesized by primase to initiate phage G4 DNA replication. *J Biol Chem* **253:** 765-769.
[42] Bouché JP, Zechel K & Kornberg A (1975) dnaG gene product, a rifampicin-resistant RNA polymerase, initiates the conversion of a single-stranded coliphage DNA to its duplex replicative form. *J Biol Chem* **250:** 5995-6001.
[43] Bowden DW, Twersky RS & Calendar R (1975) *Escherichia coli* deoxyribonucleic acid synthesis mutants: their effect upon bacteriophage P2 and satellite bacteriophage P4 deoxyribonucleic acid synthesis. *J Bacteriol* **124:** 167-175.
[44] Braithwaite DK & Ito J (1993) Compilation, alignment, and phylogenetic relationships of DNA polymerases. *Nucleic Acids Res* **21:** 787-802.
[45] Brayer GD & McPherson A (1985) A model for intracellular complexation between gene-5 protein and bacteriophage fd DNA. *Eur J Biochem* **150:** 287-296.
[46] Breyer WA & Matthews BW (2000) Structure of *Escherichia coli* exonuclease I suggests how processivity is achieved. *Nat Struct Biol* **7:** 1125-1128.
[47] Briani F, Dehò G, Forti F & Ghisotti D (2001) The plasmid status of satellite bacteriophage P4. *Plasmid* **45:** 1-17.
[48] Broker TR (1973) An electron microscopic analysis of pathways for bacteriophage T4 DNA recombination. *J Mol Biol* **81:** 1-16.
[49] Brooks K & Clark AJ (1967) Behavior of lambda bacteriophage in a recombination deficienct strain of *Escherichia coli*. *J Virol* **1:** 283-293.
[50] Brown DR, Roth MJ, Reinberg D & Hurwitz J (1984) Analysis of bacteriophage phi X174 gene A protein-mediated termination and reinitiation of phi X DNA synthesis. I. Characterization of the termination and reinitiation reactions. *J Biol Chem* **259:** 10545-10555.
[51] Bruand C & Ehrlich SD (2000) UvrD-dependent replication of rolling-circle plasmids in *Escherichia coli*. *Mol Microbiol* **35:** 204-210.
[52] Bruand C, Velten M, McGovern S, Marsin S, Serena C, Ehrlich SD & Polard P (2005) Functional interplay between the *Bacillus subtilis* DnaD and DnaB proteins essential for initiation and re-initiation of DNA replication. *Mol Microbiol* **55:** 1138-1150.

[53] Bruand C, Velten M, McGovern S, Marsin S, Serena C, Ehrlich SD & Polard P (2005) Functional interplay between the Bacillus subtilis DnaD and DnaB proteins essential for initiation and re-initiation of DNA replication. *Mol Microbiol* **55:** 1138-1150.
[54] Bujalowski W, Klonowska MM & Jezewska MJ (1994) Oligomeric structure of *Escherichia coli* primary replicative helicase DnaB protein. *J Biol Chem* **269:** 31350-31358.
[55] Burger KJ & Trautner TA. (1978) Specific labelling of replicating SPP1 DNA: analysis of viral DNA synthesis and identification of phage DNA-genes. *Mol Gen Genet* **166:** 277-285.
[56] Burke RL, Alberts BM & Hosoda J (1980) Proteolytic removal of the COOH terminus of the T4 gene 32 helix-destabilizing protein alters the T4 in vitro replication complex. *J Biol Chem* **255:** 11484-11493.
[57] Burke RL, Munn M, Barry J & Alberts BM (1985) Purification and properties of the bacteriophage T4 gene 61 RNA priming protein. *J Biol Chem* **260:** 1711-1722.
[58] Burz DS, Beckett D, Benson N & Ackers GK (1994) Self-assembly of bacteriophage lambda cI repressor: effects of single-site mutations on the monomer-dimer equilibrium. *Biochemistry* **33:** 8399-8405.
[59] Calendar R, Lindqvist B, Sironi G & Clark A.J (1970) Characterization of REP- mutants and their interaction with P2 phage. *Virology* **40:** 72-83.
[60] Callanan MJ, O'Toole PW, Lubbers MW & Polzin KM (2001) Examination of lactococcal bacteriophage c2 DNA replication using two-dimensional agarose gel electrophoresis. *Gene* **278:** 101-106.
[61] Camerini-Otero RD & Hsieh P (1995) Homologous recombination proteins in prokaryotes and eukaryotes. *Annu Rev Genet* **29:** 509-552.
[62] Campbell A & Botstein D (1983) Evolution of the Lambdoid Phages. In: Lambda II (Hendrix RW, Roberts JW, Stahl FW & Weisberg RA., Eds.), pp. 365-380. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA.
[63] Canchaya C, Proux C, Fournous G, Bruttin A & Bruessow H (2003) Prophage genomics. *Microbiol Mol Biol Rev* **67:** 238-276.
[64] Carter DM & Radding CM (1971) The role of exonuclease and beta protein of phage lambda in genetic recombination. II. Substrate specificity and the mode of action of lambda exonuclease. *J Biol Chem* **246:** 2502-2512.
[65] Ceska TA., Sayers JR, Stier G & Suck D (1996) A helical arch allowing single-stranded DNA to thread through T5 5'-exonuclease. *Nature* **382:** 90-93.
[66] Cha TA & Alberts BM (1990) Effects of the bacteriophage T4 gene 41 and gene 32 proteins on RNA primer synthesis: coupling of leading- and lag-ging-strand DNA synthesis at a replication fork. *Biochemistry* **29:** 1791-1798.
[67] Chaconas G, Harshey RM, Sarvetnick N & Bukhari A.I (1981) Mechanism of bacteriophage Mu DNA transposition. *Cold Spring Harbor Symp Quant Biol* **45(1):** 311-322.
[68] Chandry, PS, Moore SC, Boyce JD, Davidson BE & Hillier AJ (1997) Analysis of the DNA sequence, gene expression, origin of replication and modular structure of the *Lactococcus lactis* lytic bacteriophage sk1. *Mol Microbiol* **26:** 49-64.
[69] Chang P & Marians KJ (2000) Identification of a region of *Escherichia coli* DnaB required for functional interaction with DnaG at the replication fork. *J Biol Chem* **275:** 26187-26195.
[70] Chang PC & Cohen SN (1994) Bidirectional replication from an internal origin in a linear streptomyces plasmid. *Science* **265:** 952-954.
[71] Chang PC, Kim ES & Cohen SN (1996) Streptomyces linear plasmids that contain a phage-like, centrally located, replication origin. *Mol Microbiol* **22:** 789-800.
[72] Chase JW, L'Italien JJ, Murphy JB, Spicer EK & Williams KR (1984) Characterization of the *Escherichia coli* SSB-113 mutant single-stranded DNA-binding protein. Cloning of the gene, DNA and protein sequence analysis, high pressure liquid chromatography peptide mapping and DNA-binding studies. *J Biol Chem* **259:** 805-814.
[73] Chatterjee DK, Fujimura RK, Campbell JH & Gerard GF (1991) Cloning and overexpression of the gene encoding bacteriophage T5 DNA polymerase. *Gene* **97:** 13-19.
[74] Chattoraj DK (2000) Control of plasmid DNA replication by iterons: no longer paradoxical. *Mol Microbiol* **37:** 467-476.
[75] Chen F & Lu J (2002) Genomic Sequence and Evolution of Marine Cyanophage P60: a New Insight on Lytic and Lysogenic Phages. *Appl Environ Microbiol* **68:** 2589-2594.
[76] Chesney RH, Scott JR & Vapnek D (1979) Integration of the plasmid prophages P1 and P7 into the chromosome of *E coli*. *J Mol Biol* **130:** 161-173.
[77] Chopin MC, Rouault A, Ehrlich SD & Gautier M (2002) Filamentous phage active on the gram-positive bacterium *Propionibacterium freudenreichii*. *J Bacteriol* **184:** 2030-2033.
[78] Chowdhury K, Tabor S & Richardson CC (2000) A unique loop in the DNA-binding crevice of bacteriophage T7 DNA polymerase influences primer utilization. *Proc Natl Acad Sci USA* **97:** 12469-12474.
[79] Chu CC, Templin A & Clark AJ (1989) Suppression of a frameshift mutation in the *recE* gene of *Escherichia coli* K-12 occurs by gene fusion. *J Bacteriol* **171:** 2101-2109.
[80] Clark AJ, Sharma V, Brenowitz S, *et al*. (1993) Genetic and molecular analyses of the C-terminal region of the *recE* gene from the Rac prophage of *Escherichia coli* K-12 reveal the *recT* gene. *J Bacteriol* **175:** 7673-7682.
[81] Cohen G (1983) Electron microscopy study of early lytic replication forms of bacteriophage P1 DNA. *Virology* **131:** 159-170.
[82] Cohen G & Sternberg N (1989) Genetic analysis of the lytic replicon of bacteriophage P1.I. Isolation and partial characterization. *J Mol Biol* **207:** 99-109.
[83] Connelly JC, de Leau ES & Leach DR (1999) DNA cleavage and degradation by the SbcCD protein complex from *Escherichia coli*. *Nucleic Acids Res* **27:** 1039-1046.
[84] Corpet F (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res* **16:** 10881-10890.
[85] Cox MM, Goodman MF, Kreuzer KN, Sherratt DJ, Sandler SJ & Marians KJ (2000) The importance of repairing stalled replication forks. *Nature* **404:** 37-41.
[86] Crooke E, Thresher R, Hwang DS, Griffith J & Kornberg A (1993) Replicatively active complexes of DnaA protein and the *Escherichia coli* chromosomal origin observed in the electron microscope. *J Mol Biol* **233:** 16-24.
[87] Cuff JA & Barton GJ (2000) Application of Multiple Sequence Alignment Profiles to Improve Protein Secondary Structure Prediction. *Proteins* **40:** 502-511.
[88] Cuff JA, Clamp ME, Siddiqui AS, Finlay M & Barton GJ (1998) JPred: a consensus secondary structure prediction server. *Bioinformatics* **14:** 892-893.
[89] D'Ari R, Jaffé-Brachet A, Touati-Schwartz D & Yarmolinsky MB (1975) A dnaB analog specified by bacteriophage P1. *J Mol Biol* **94:** 341-366.
[90] Dalrymple BP, Kongsuwan K, Wijffels G, Dixon NE & Jennings PA (2001) A universal protein-protein interaction motif in the eubacterial DNA replication and repair systems. *Proc Natl Acad Sci USA* **98:** 11627-11632.
[91] Datta HJ, Khatri GS & Bastia D (1999) Mechanism of recruitment of DnaB helicase to the replication origin of the plasmid pSC101. *Proc Natl Acad Sci USA* **96:** 73-78.
[92] Davey MJ, Fang L, McInerney P, Georgescu RE & O'Donnell M (2002) The DnaC helicase loader is a dual ATP/ADP switch protein. *EMBO J* **21:** 3148-3159.

[93] Davey MJ, Jeruzalmi D, Kuriyan J & O'Donnell M (2002) Motors and switches: AAA+ machines within the replisome. *Nat Rev Mol Cell Biol* **3:** 826-835.
[94] Davidson JF, Fox R, Harris DD, Lyons-Abbott S & Loeb LA (2003) Insertion of the T3 DNA polymerase thioredoxin binding domain enhances the processivity and fidelity of Taq DNA polymerase. *Nucleic Acids Res* **31:** 4702-4709.
[95] Davis BM & Waldor MK (2003) Filamentous phages linked to virulence of *Vibrio cholerae*. *Curr Opin Microbiol* **6:** 35-42.
[96] Dehò G, Zangrossi S, Sabbattini P, Sironi G & Ghisotti D (1992) Bacteriophage P4 immunity controlled by small RNAs via transcription termination. *Mol Microbiol* **6:** 3415-3425.
[97] del Solar G & Espinosa M (2000) Plasmid copy number control: an ever-growing story. *Mol Microbiol* **37:** 492-500.
[98] del Solar G, Giraldo R, Ruiz-Echevarria MJ, Espinosa M & Díaz Orejas R (1998) Replication and control of circular bacterial plasmids. *Microbiol Mol Biol Rev* **62:** 434-464.
[99] Delarue M, Poch O, Tordo N, Moras D & Argos P (1990) An attempt to unify the structure of polymerases. *Protein Eng* **3:** 461-467.
[100] DeRose EF, Li D, Darden T, Harvey S, Perrino FW, Schaaper RM & London RE (2002) Model for the catalytic domain of the proofreading epsilon subunit of *Escherichia coli* DNA polymerase III based on NMR structural data. *Biochemistry* **41:** 94-110.
[101] Dervyn E, Suski C, Daniel R, Bruand C, Chapuis J, Errington J, Janniere L & Ehrlich SD (2001) Two Essential DNA Polymerases at the Bacterial Replication Fork. *Science* **294:** 1716-1719.
[102] Desiere F, Lucchini S, Bruttin A, Zwahlen MC & Bruessow H (1997) A Highly Conserved DNA Replication Module from *Streptococcus thermophilus* Phages Is Similar in Sequence and Topology to a Module from *Lactococcus lactis* Phages. *Virology* **234:** 372-382.
[103] Díaz Orejas R, Ziegelin G, Lurz R & Lanka E (1994) Phage P4 DNA replication in vitro. *Nucleic Acids Res* **22:** 2065-2070.
[104] Díaz R, Barnsley P & Pritchard RH (1979) Location and characterisation of a new replication origin in the E. coli K12 chromosome. *Mol Gen Genet* **175:** 151-157.
[105] Díaz R & Pritchard RH (1978) Cloning of replication origins from the E. coli K12 chromosome. *Nature* **275:** 561-564.
[106] Dodson M, Echols H, Wickner S, *et al*. (1986) Specialized nucleoprotein structures at the origin of replication of bacteriophage lambda: localized unwinding of duplex DNA by a six-protein reaction. *Proc Natl Acad Sci USA* **83:** 7638-7642.
[107] Donate LE, Llorca O, Bárcena M, Brown SE, Dixon NE & Carazo JM (2000) pH-controlled quaternary states of hexameric DnaB helicase. *J Mol Biol* **303:** 383-393.
[108] Dong F, Gogol EP & von Hippel PH (1995) The phage T4-coded DNA replication helicase (gp41) forms a hexamer upon activation by nucleoside triphosphate. *J Biol Chem* **270:** 7462-7473.
[109] Doublie S, Tabor S, Long AM, Richardson CC & Ellenberger T (1998) Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 A resolution. *Nature* **391:** 251-258.
[110] Dubeau L, Hours C & Denhardt DT (1981) The mechanism of replication of phiX174. XVII. Purification and partial characterization of the gene A and A proteins. *Can J Biochem* **59:** 106-115.
[111] Dunn JJ & Studier FW (1983) Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. *J Mol Biol* **166:** 477-535.
[112] Dunn K, Chrysogelos S & Griffith J (1982) Electron microscopic visualization of recA-DNA filaments: evidence for a cyclic extension of duplex DNA. *Cell* **28:** 757-765.
[113] Echols H & Gingery R (1968) Mutants of bacteriophage lambda defective in vegetative genetic recombination. *J Mol Biol* **34:** 239-249.
[114] Egelman EH, Yu X, Wild R, Hingorani MM & Patel SS (1995) Bacteriophage T7 helicase/primase proteins form rings around single-stranded DNA that suggest a general structure for hexameric helicases. *Proc Natl Acad Sci USA* **92:** 3869-3873.
[115] Eisenberg S, Griffith J & Kornberg A (1977) phiX174 cistron A protein is a multifunctional enzyme in DNA replication. *Proc Natl Acad Sci USA* **74:** 3198-3202.
[116] Eisenberg S & Kornberg A (1979) Purification and characterization of phiX174 gene A protein. A multifunctional enzyme of duplex DNA replication. *J Biol Chem* **254:** 5328-5332.
[117] Eisenberg S, Scott JF & Kornberg A (1976) An enzyme system for replication of duplex circular DNA: the replicative form of phage phi X174. *Proc Natl Acad Sci USA* **73:** 1594-1597.
[118] Eisenberg S, Scott JF & Kornberg A (1976) Enzymatic replication of viral and complementary strands of duplex DNA of phage phiX174 proceeds by seprate mechanisms. *Proc Natl Acad Sci USA* **73:** 3151-3155.
[119] Espinosa M, del Solar G, Rojo F & Alonso JC (1995) Plasmid rolling circle replication and its control. *FEMS Microbiol. Lett* **130:** 111-120.
[120] Fang LH, Davey MJ & O'Donnell M (1999) Replisome assembly at oriC, the replication origin of *E. coli* reveals an explanation for initiation sites outside an origin. *Mol Cell* **4:** 541-553.
[121] Felczak MM & Kaguni JM (2004) The Box VII motif of *Escherichia coli* DnaA protein is required for DnaA oligomerization at the *E. coli* replication origin. *J Biol Chem*
[122] Filee J, Forterre P, Sen-Lin T & Laurent J (2002) Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. *J Mol Evol* **54:** 763-773.
[123] Flensburg J & Calendar R (1987) Bacteriophage P4 DNA replication. Nucleotide sequence of the P4 replication gene and the cis replication region. *J Mol Biol* **195:** 439-445.
[124] Foley S, Lucchini S, Zwahlen MC & Bruessow H (1998) A short noncoding viral DNA element showing characteristics of a replication origin confers bacteriophage resistance to *Streptococcus thermophilus*. *Virology* **250:** 377-387.
[125] Ford ME, Sarkis GJ, Belanger AE, Hendrix RW & Hatfull GF (1998) Genome structure of mycobacteriophage D29: implications for phage evolution. *J Mol Biol* **279:** 143-164.
[126] Forterre P (1999) Displacement of cellular proteins by functional analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins. *Mol Microbiol* **33:** 457-465.
[127] Franklin NC (1967) Deletions and functions of the center of the phi80 -lambda phage genome. Evidence for a phage function promoting genetic recombination. *Genetics* **57:** 301-318.
[128] Frick DN, Baradaran K & Richardson CC (1998) An N-terminal fragment of the gene 4 helicase/primase of bacteriophage T7 retains primase activity in the absence of helicase activity. *Proc Natl Acad Sci USA* **95:** 7957-7962.
[129] Frick DN & Richardson CC (2001) DNA Primases. *Annu Rev Biochem* **70:** 39-80.
[130] Fulford W & Model P (1988) Bacteriophage f1 DNA replication genes. II. The roles of gene V protein and gene II protein in complementary strand synthesis. *J Mol Biol* **203:** 39-48.
[131] Funnell BE, Baker TA & Kornberg A (1987) *In vitro* assembly of a prepriming complex at the origin of the *Escherichia coli* chromosome. *J Biol Chem* **262:** 10327-10334.

[132] Funnell BE & Inman RB (1983) Bacteriophage P2 DNA replication. Characterization of the requirement of the gene B protein in vivo. *J Mol Biol* **167:** 311-334.
[133] Furth ME, McLeester C & Dove WF (1978) Specificity determinants for bacteriophage lambda DNA replication.I. A chain of interactions that controls the initiation of replication. *J Mol Biol* **126:** 195-225.
[134] Furth ME & Yates JL (1978) Specificity determinants for bacteriophage lambda DNA replication. II. Structure of O proteins of lambda-phi80 and lambda-82 hybrid phages and of a lambda mutant defective in the origin of replication. *J Mol Biol* **126:** 227-240.
[135] Garforth SJ, Ceska TA, Suck D & Sayers JR (1999) Mutagenesis of conserved lysine residues in bacteriophage T5 5'-3' exonuclease suggests separate mechanisms of endoand exonucleolytic cleavage. *Proc Natl Acad Sci USA* **96:** 38-43.
[136] Garzon A, Cano DA & Casadesus J (1998) The P22 Erf protein and host RecA provide alternative functions for transductional segregation of plasmid-borne duplications. *Mol Gen Genet* **259:** 39-45.
[137] Gauss P, Gayle M, Winter RB & Gold L (1987) The bacteriophage T4 dexA gene: sequence and analysis of a gene conditionally required for DNA replication. *Mol Gen Genet* **206:** 24-34.
[138] Geider K, Baumel I & Meyer TF (1982) Intermediate stages in enzymatic replication of bacteriophage fd duplex DNA. *J Biol Chem* **257:** 6488-6493.
[139] George JW, Stohr BA, Tomso DJ & Kreuzer KN (2001) The tight linkage between DNA replication and double-strand break repair in bacteriophage T4. *Proc Natl Acad Sci USA* **98:** 8290-8297.
[140] Gielow A, Diederich L & Messer W (1991) Characterization of a phage-plasmid hybrid (Phasyl) with two independent origins of replication isolated from *Escherichia coli*. *J Bacteriol* **173:** 73-79.
[141] Gilbert W & Dressler D (1968) DNA replication: the rolling circle model. *Cold Spring Harbor Symp Quant Biol* **33:** 473-484.
[142] Giraldo R (2003) Common domains in the initiators of DNA replication in Bacteria, Archaea and Eukarya: combined structural, functional and phylogenetic perspectives. *FEMS Microbiol Rev* **26:** 533-554.
[143] Giraldo R & Fernandez-Tresguerres ME (2004) Twenty years of the pPS10 replicon: insights on the molecular mechanism for the activation of DNA replication in iteron-containing bacterial plasmids. *Plasmid* **52:** 69-83.
[144] Girons IS, Bourhy P, Ottone C, Picardeau M, Yelton D, Hendrix RW, Glaser P & Charon N (2000) The LE1 bacteriophage replicates as a plasmid within *Leptospira biflexa*: construction of an *L. biflexa-Escherichia coli* shuttle vector. *J Bacteriol* **182:** 5700-5705.
[145] Glover BP & McHenry CS (1998) The chi psi subunits of DNA polymerase III holoenzyme bind to single-stranded DNA-binding protein (SSB) and facilitate replication of an SSB-coated template. *J Biol Chem* **273:** 23476-23484.
[146] Golub EI & Low KB (1986) Derepression of single-stranded DNA-binding protein genes on plasmids derepressed for conjugation, and complementation of an *E. coli ssb-* mutation by these genes. *Mol Gen Genet* **204:** 410-416.
[147] Goodman MF (2002) Error-prone Repair DNA Polymerases in Prokaryotes and Eukaryotes. *Annu Rev Biochem* **71:** 17-50.
[148] Gorbalenya AE, Koonin EV & Wolf YI (1990) A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses. *Febs Lett* **262:** 145-148.
[149] Gottesman S, Clark WP, Crecy-Lagard V & Maurizi MR (1993) ClpX, an alternative subunit for the ATP-dependent Clp protease of *Escherichia coli*. Sequence and *in vivo* activities. *J Biol Chem* **268:** 22618-22626.
[150] Grabowski B & Kelman Z (2003) Archeal DNA replication: eukaryal proteins in a bacterial context. *Annu Rev Microbiol* **57:** 487-516.
[151] Greenstein D & Horiuchi K (1987) Interaction between the replication origin and the initiator protein of the filamentous phage f1. Binding occurs in two steps. *J Mol Biol* **197:** 157-174.
[152] Grigoriev A (1999) Strand-specific compositional asymmetries in double-stranded DNA viruses. *Virus Res* **60:** 1-19.
[153] Groth AC & Calos MP (2004) Phage integrases: biology and applications. *J Mol Biol* **335:** 667-678.
[154] Guenther B, Onrust R, Sali A, O'Donnell M & Kuriyan J (1997) Crystal structure of the delta' subunit of the clamp-loader complex of *E. coli* DNA polymerase III. *Cell* **91:** 335-345.
[155] Gutiérrez C, Martin G, Sogo JM & Salas M (1991) Mechanism of stimulation of DNA replication by bacteriophage phi 29 single-stranded DNA-binding protein p5. *J Biol Chem* **266:** 2104-2111.
[156] Gutman PD & Minton KW (1993) Conserved sites in the 5'-3' exonuclease domain of *Escherichia coli* DNA polymerase. *Nucleic Acids Res* **21:** 4406-4407.
[157] Hadden JM, Declais AC, Phillips SEV & Lilley DMJ (2002) Metal ions bound at the active site of the junction-resolving enzyme T7 endonuclease I. *EMBO J* **21:** 3505-3515.
[158] Hall MC & Matson SW (1999) Helicase motifs: the engine that powers DNA unwinding. *Mol Microbiol* **34:** 867-877.
[159] Hall SD, Kane MF & Kolodner RD (1993) Identification and characterization of the *Escherichia coli* RecT protein, a protein encoded by the recE region that promotes renaturation of homologous single-stranded DNA. *J Bacteriol* **175:** 277-287.
[160] Hamlett NV & Berger H (1975) Mutations altering genetic recombination and repair of DNA in bacteriophage T4. *Virology* **63:** 539-567.
[161] Haniford DB & Chaconas G (1992) Mechanistic aspects of DNA transposition. *Curr Opin Genet Dev* **2:** 698-704.
[162] Hansen EB (1989) Structure and regulation of the lytic replicon of phage P1. *J Mol Biol* **207:** 135-149.
[163] Hansen EB & Yarmolinsky MB (1986) Host participation in plasmid maintenance: Dependence upon *dnaA* of replicons derived from P1 and F. *Proc Natl Acad Sci USA* **83:** 4423-4427.
[164] Harth G, Baumel I, Meyer TF & Geider K (1981) Bacteriophage fd gene-2 protein. Processing of phage fd viral strands replicated by phage T7 enzymes. *Eur J Biochem* **119:** 663-668.
[165] Hay N & Cohen G (1983) Requirement of *E. coli* DNA synthesis functions for the lytic replication of bacteriophage P1. *Virology* **131:** 193-206.
[166] Heidekamp F, Langeveld SA, Baas PD & Jansz HS (1980) Studies of the recognition sequence of phi X174 gene A protein. Cleavage site of phi X gene A protein in St-1 RFI DNA. *Nucleic Acids Res* **8:** 2009-2021.
[167] Heinrich J, Riedel HD, Rückert B, Lurz R & Schuster H (1995) The lytic replicon of bacteriophage P1 is controlled by an antisense RNA. *Nucleic Acids Res* **23:** 1468-1474.
[168] Heisig A, Severin I, Seefluth AK & Schuster H (1987) Regulation of the ban gene containing operon of prophage P1. *Mol Gen Genet* **206:** 368-376.
[169] Hiasa H & Marians KJ (1999) Initiation of bidirectional replication at the chromosomal origin is directed by the interaction between helicase and primase. *J Biol Chem* **274:** 27244-27248.
[170] Higashitani A, Greenstein D, Hirokawa H, Asano S & Horiuchi K (1994) Multiple DNA conformational changes induced by an initiator protein precede the nicking reaction in a rolling circle replication origin. *J Mol Biol* **237:** 388-400.

[171] Higashitani A, Greenstein D & Horiuchi K (1992) A single amino acid substitution reduces the superhelicity requirement of a replication initiator protein. *Nucleic Acids Res* **20:** 2685-2691.
[172] Higashitani A, Higashitani N & Horiuchi K (1997) Minus-strand origin of filamentous phage versus transcriptional promoters in recognition of RNA polymerase. *Proc Natl Acad Sci USA* **94:** 2909-2914.
[173] Higashitani N, Higashitani A, Guan ZW & Horiuchi K (1996) Recognition mechanisms of the minus-strand origin of phage f1 by Escherichia coli RNA polymerase. *Genes Cells* **1:** 829-841.
[174] Higashitani N, Higashitani A & Horiuchi K (1993) Nucleotide sequence of the primer RNA for DNA replication of filamentous bacteriophages. *J Virol* **67:** 2175-2181.
[175] Himawan JS & Richardson CC (1996) Amino acid residues critical for the interaction between bacteriophage T7 DNA polymerase and *Escherichia coli* thioredoxin. *J Biol Chem* **271:** 19999-20008.
[176] Hinton DM & Nossal NG (1987) Bacteriophage T4 DNA primase-helicase. Characterization of oligomer synthesis by T4 61 protein alone and in conjunction with T4 41 protein. *J Biol Chem* **262:** 10873-10878.
[177] Hoet PP, Coene MM & Cocito CG (1992) Replication Cycle of *Bacillus subtilis* Hydroxymethyluracil-Containing Phages. *Annu Rev Microbiol* **46:** 95-116.
[178] Hollifield WC, Kaplan EN & Huang HV (1987) Efficient RecABC-dependent, homologous recombination between coliphage lambda and plasmids requires a phage ninR region gene. *Mol Gen Genet* **210:** 248-255.
[179] Horiuchi K (1980) Origin of DNA replication of bacteriophage f1 as the signal for termination. *Proc Natl Acad Sci USA* **77:** 5226-5229.
[180] Horiuchi K (1986) Interaction between gene II protein and the DNA replication origin of bacteriophage f1. *J Mol Biol* **188:** 215-223.
[181] Horiuchi K (1997) Initiation mechanisms in replication of filamentous phage DNA. *Genes Cells* **2:** 425-432.
[182] Hosoda J & Moise H (1978) Purification and physicochemical properties of limited proteolysis products of T4 helix destabilizing protein (gene 32 protein). *J Biol Chem* **253:** 7547-7558.
[183] Hourcade D & Dressler D (1978) Site-specific initiation of a DNA fragment. *Proc Natl Acad Sci USA* **75:** 1652-1656.
[184] Howland CJ, Rees CE, Barth PT & Wilkins BM (1989) The *ssb* gene of plasmid ColIb-P9. *J Bacteriol* **171:** 2466-2473.
[185] Huang YJ, Parker MM & Belfort M (1999) Role of exonucleolytic degradation in group I intron homing in phage T4. *Genetics* **153:** 1501-1512.
[186] Huebscher U, Maga G & Spadari S (2002) Eukaryotic DNA Polymerases. *Annu Rev Biochem* **71:** 133-163.
[187] Ikeda JE, Yudelevich A & Hurwitz J (1976) Isolation and characterization of the protein coded by gene A of bacteriophage phiX174 DNA. *Proc Natl Acad Sci USA* **73:** 2669-2673.
[188] Ilyina TV & Koonin EV (1992) Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. *Nucleic Acids Res* **20:** 3279-3285.
[189] Imai Y, Ogasawara N, Ishigo-Oka D, Kadoya R, Daito T & Moriya S (2000) Subcellular localization of Dna-initiation proteins of *Bacillus subtilis*: evidence that chromosome replication begins at either edge of the nucleoids. *Mol Microbiol* **36:** 1037-1048.
[190] Ishigo-Oka D, Ogasawara N & Moriya S (2001) DnaD protein of *Bacillus subtilis* interacts with DnaA, the initiator protein of replication. *J Bacteriol* **183:** 2148-2150.
[191] Ishmael FT, Alley SC & Benkovic SJ (2001) Identification and Mapping of Protein-Protein Interactions between gp32 and gp59 by Cross-linking. *J Biol Chem* **276:** 25236-25242.
[192] Ishmael FT, Alley SC & Benkovic SJ (2002) Assembly of the Bacteriophage T4 Helicase. Architecture and Stoichiometry of the gp41-gp59 Complex. *J Biol Chem* **277:** 20555-20562.
[193] Iwasaki H, Nakata A, Walker GC & Shinagawa H (1990) The *Escherichia coli polB* gene, which encodes DNA polymerase II, is regulated by the SOS system. *J Bacteriol* **172:** 6268-6273.
[194] Iyer LM, Koonin EV & Aravind L (2002) Classification and evolutionary history of the single-strand annealing proteins, RecT, Redbeta, ERF and RAD52. BMC *Genomics* **3:** 8-8.
[195] Jacob F, Brenner S & Cuzin F (1963) On the regulation of DNA replication in bacteria. *Cold Spring Harbor Symp Quant Biol* **28:** 329-348.
[196] Jeruzalmi D, O'Donnell M & Kuriyan J (2002) Clamp loaders and sliding clamps. *Curr Opin Struct Biol* **12:** 217-224.
[197] Jezewska MJ & Bujalowski W (1996) Global conformational transitions in *Escherichia coli* primary replicative helicase DnaB protein induced by ATP, ADP, and single-stranded DNA binding. Multiple conformational states of the helicase hexamer. *J Biol Chem* **271:** 4261-4265.
[198] Jones CE, Mueser TC, Dudas KC, Kreuzer KN & Nossal NG (2001) Bacteriophage T4 gene 41 helicase and gene 59 helicase-loading protein: a versatile couple with roles in replication and recombination. *Proc Natl Acad Sci USA* **98:** 8312-8318.
[199] Jones CE, Mueser TC & Nossal NG (2000) Interaction of the Bacteriophage T4 Gene 59 Helicase Loading Protein and Gene 41 Helicase with Each Other and with Fork, Flap, and Cruciform DNA. *J Biol Chem* **275:** 27145-27154.
[200] Joyce CM, Fujii DM, Laks HS, Hughes CM & Grindley ND (1985) Genetic mapping and DNA sequence analysis of mutations in the *polA* gene of *Escherichia coli*. *J Mol Biol* **186:** 283-293.
[201] Joyce CM & Steitz TA (1994) Function and Structure Relationships in DNA Polymerases. *Annu Rev Biochem* **63:** 777-822.
[202] Kadyrov FA & Drake JW (2004) UvsX recombinase and Dda helicase rescue stalled bacteriophage T4 DNA replication forks in vitro. *J Biol Chem* **279(34):** 35735-35740
[203] Karam JD & Konigsberg WH (2000) DNA polymerase of the T4-related bacteriophages. *Prog Nucleic Acid Res Mol Biol* **64:** 65-96.
[204] Kato M, Ito T, Wagner G, Richardson CC & Ellenberger T (2003) Modular architecture of the bacteriophage T7 primase couples RNA primer synthesis to DNA synthesis. *Mol Cell* **11:** 1349-1360.
[205] Keck JL, Roche DD, Lynch AS & Berger JM (2000) Structure of the RNA polymerase domain of *E. coli* primase. *Science* **287:** 2482-2486.
[206] Kelman Z, Hurwitz J & O'Donnell M (1998) Processivity of DNA polymerases: two mechanisms, one goal. *Structure* **6:** 121-125.
[207] Kelman Z, Yuzhakov A, Andjelkovic J & O'Donnell M (1998) Devoted to the lagging strand-the subunit of DNA polymerase III holoenzyme contacts SSB to promote processive elongation and sliding clamp assembly. *EMBO J* **17:** 2436-2449.
[208] Khan SA (1997) Rolling-circle replication of bacterial plasmids. *Microbiol Mol Biol Rev* **61:** 442-455.
[209] Khan SA (2000) Plasmid rolling-circle replication: recent developments. *Mol Microbiol* **37:** 477-484.
[210] Khan SA & Novick RP (1983) Complete nucleotide sequence of pT181, a tetracycline-resistance plasmid from *Staphylococcus aureus*. *Plasmid* **10:** 251-259.

© 2006 Federation of European Microbiological Societies

[211] Kim SG, Bor YC & Batt CA (1992) Bacteriophage resistance in *Lactococcus lactis* ssp. *lactis* using antisense ribonucleic acid. *J Dairy Sci* **75:** 1761-1767.
[212] Kim YT & Richardson CC (1994) Acidic carboxyl-terminal domain of gene 2.5 protein of bacteriophage T7 is essential for protein-protein interactions. *J Biol Chem* **269:** 5270-5278.
[213] Kim YT, Tabor S, Bortner C, Griffith JD & Richardson CC (1992) Purification and characterization of the bacteriophage T7 gene 2.5 protein. A single-stranded DNA-binding protein. *J Biol Chem* **267:** 15022-15031.
[214] Kim YT, Tabor S, Churchich JE & Richardson CC (1992) Interactions of gene 2.5 protein and DNA polymerase of bacteriophage T7. *J Biol Chem* **267:** 15032-15040.
[215] Kimura T, Asai T, Imai M & Takanami M (1989) Methylation strongly enhances DNA bending in the replication origin region of the *Escherichia coli* chromosome. *Mol Gen Genet* **219:** 69-74.
[216] Kmiec E & Holloman WK (1981) Beta protein of bacteriophage lambda promotes renaturation of DNA. *J Biol Chem* **256:** 12636-12639.
[217] Knopf CW (2000) Molecular mechanisms of replication of herpes simplex virus 1. *Acta Virol* **44:** 289-307.
[218] Kogoma T (1997) Stable DNA replication: interplay between DNA replication, homologous recombination, and transcription. *Microbiol Mol Biol Rev* **61:** 212-238.
[219] Kolodner R, Hall SD & Luisi-DeLuca C (1994) Homologous pairing proteins encoded by the *Escherichia coli recE* and *recT* genes. *Mol Microbiol* **11:** 23-30.
[220] Kong D, Griffith JD & Richardson CC (1997) Gene 4 helicase of bacteriophage T7 mediates strand transfer through pyrimidine dimers, mismatches, and nonhomologous regions. *Proc Natl Acad Sci USA* **94:** 2987-2992.
[221] Kong D, Nossal NG & Richardson CC (1997) Role of the bacteriophage T7 and T4 single-stranded DNA-binding proteins in the formation of joint molecules and DNA helicase-catalyzed polar branch migration. *J Biol Chem* **272:** 8380-8387.
[222] Kong D & Richardson CC (1996) Single-stranded DNA binding protein and DNA helicase of bacteriophage T7 mediate homologous DNA strand exchange. *EMBO J* **15:** 2010-2019.
[223] Kong D & Richardson CC (1998) Role of the acidic carboxyl-terminal domain of the single-stranded DNA-binding protein of bacteriophage T7 in specific protein-protein interactions. *J Biol Chem* **273:** 6556-6564.
[224] Konieczny I & Marszalek J (1995) The requirement for molecular chaperones in lambda DNA replication is reduced by the mutation pi in lambda P gene, which weakens the interaction between lambda P protein and DnaB helicase. *J Biol Chem* **270:** 9792-9799.
[225] Koonin EV (1992) DnaC protein contains a modified ATP-binding motif and belongs to a novel family of ATPases including also DnaA. *Nucleic Acids Res* **20:** 1997-1997.
[226] Koonin EV & Ilyina TV (1992) Geminivirus replication proteins are related to prokaryotic plasmid rolling circle DNA replication initiator proteins. J Gen Virol **73(10):** 2763-2766.
[227] Koonin EV & Ilyina TV (1993) Computer-assisted dissection of rolling circle DNA replication. *Biosystems* **30:** 241-268.
[228] Kornberg A & Baker TA (1992) DNA Replication, WH Freeman and Company, New York, USA.
[229] Kovall R & Matthews BW (1997) Toroidal structure of lambda-exonuclease. *Science* **277:** 1824-1827.
[230] Kowalczykowski SC, Dixon DA, Eggleston AK, Lauder SD & Rehrauer WM (1994) Biochemistry of homologous recombination in *Escherichia coli*. *Microbiol Rev* **58:** 401-465.
[231] Kowalski D & Eddy MJ (1989) The DNA unwinding element: A novel, *cis*-acting component that facilitates opening of the *Escherichia coli* replication origin. *EMBO J* **8:** 4335-4344.
[232] Krause M & Messer W (1999) DnaA proteins of *Escherichia coli* and *Bacillus subtilis*: coordinate actions with single-stranded DNA-binding protein and interspecies inhibition during open complex formation at the replication origins. *Gene* **228:** 123-132.
[233] Krevolin MD & Calendar R (1985) The replication of bacteriophage P4 DNA in vitro. Partial purification of the P4 alpha gene product. *J Mol Biol* **182:** 509-517.
[234] Krevolin MD, Inman RB, Roof D, Kahn M & Calendar R (1985) Bacteriophage P4 DNA replication. Location of the P4 origin. *J Mol Biol* **182:** 519-527.
[235] Kuil ME, Holmlund K, Vlaanderen CA & van Grondelle R (1990) Study of the binding of single-stranded DNA-binding protein to DNA and poly(rA) using electric field induced birefringence and circular dichroism spectroscopy. *Biochemistry* 2**9:** 8184-8189.
[236] Kurosawa Y & Okazaki T (1979) Structure of the RNA portion of the RNA-linked DNA pieces in bacteriophage T4-infected *Escherichia coli* cells. *J Mol Biol* **135:** 841-861.
[237] Kushner SR, Nagaishi H & Clark AJ (1974) Isolation of exonuclease VIII: the enzyme associated with sbcA indirect suppressor. *Proc Natl Acad Sci USA* **71:** 3593-3597.
[238] Kuzminov A (1999) Recombinational repair of DNA damage in Escherichia coli and bacteriophage lambda. *Microbiol Mol Biol Rev* **63:** 751-813.
[239] Langer U, Richter S, Roth A, Weigel C & Messer W (1996) A comprehensive set of DnaA box mutations in the replication origin, *oriC*, of *Escherichia coli*. *Mol Microbiol* **21:** 301-311.
[240] Lanka E, Edelbluth C, Schlicht M & Schuster H (1978) *Escherichia coli* dnaB protein. Affinity chromatography on immobilized nucleotides. *J Biol Chem* **253:** 5847-5851.
[241] Lanka E, Geschke B & Schuster H (1978) *Escherichia coli dnaB* mutant defective in DNA initiation: Isolation and properties of the *dnaB* protein. *Proc Natl Acad Sci USA* **75:** 799-803.
[242] Lanka E, Mikolajczyk M, Schlicht M & Schuster H (1978) Association of the prophage P1ban protein with the dnaB protein of *Escherichia coli*. *J Biol Chem* **253:** 4746-4753.
[243] Lanka E & Schuster H (1983) The DnaC protein of *Escherichia coli*. Purification, physical properties and interaction with DnaB protein. *Nucleic Acids Res* **11:** 987-997.
[244] Learn BA, Um S-J, Huang L and McMacken R (1997) Cryptic single-stranded-DNA binding activities of the phage λ P and *Escherichia coli* DnaC replication initiation proteins facilitate the transfer of *E. coli* DnaB helicase onto DNA. *Proc Natl Acad Sci USA* **94:** 1154-1159.
[245] Leavitt MC & Ito J (1989) T5 DNA polymerase: structural--functional relationships to other DNA polymerases. *Proc Natl Acad Sci USA* **86:** 4465-4469.
[246] LeBowitz JH & McMacken R (1986) The *Escherichia coli dnaB* replication protein is a DNA helicase. *J Biol Chem* **261:** 4738-4748.
[247] Lee J, Chastain PD, Griffith JD & Richardson CC (2002) Lagging strand synthesis in coordinated DNA synthesis by bacteriophage t7 replication proteins. *J Mol Biol* **316:** 19-34.
[248] Lee J, Chastain PD, Kusakabe T, Griffith JD & Richardson CC (1998) Coordinated leading and lagging strand DNA synthesis on a minicircular template. *Mol Cell* **1:** 1001-1010.
[249] Lemonnier M, Ziegelin G, Reick T, Gomez AM, Diaz Orejas R & Lanka E (2003) Bacteriophage P1 Ban protein is a hexameric DNA helicase that interacts with and substitutes for *Escherichia coli* DnaB. *Nucleic Acids Res* **31:** 3918-3928.
[250] Leonard AC & Grimwade JE (2005) Building a bacterial orisome: emergence of new regulatory features for replication origin unwinding. *Mol Microbiol* **55:** 978-985.
[251] Lilley DM & White MF (2000) Resolving the relationships of resolving enzymes. *Proc Natl Acad Sci USA* **97:** 9351-9353.

[252] Lin S & Kowalski D (1994) DNA helical instability facilitates initiation at the SV40 replication origin. *J Mol Biol* **235:** 496-507.

[253] Lindqvist BH & Six EW (1971) Replication of bacteriophage P4 DNA in a nonlysogenic host. *Virology* **43:** 1-7.

[254] Little JW (1967) An exonuclease induced by bacteriophage lambda. II. Nature of the enzymatic reaction. *J Biol Chem* **242:** 679-686.

[255] Liu CC & Alberts BM (1980) Pentaribonucleotides of mixed sequence are synthesized and efficiently prime de novo DNA chain starts in the T4 bacteriophage DNA replication system. *Proc Natl Acad Sci USA* **77:** 5698-5702.

[256] Liu Y & Haggård-Liungquist E (1994) Studies of bacteriophage P2 DNA replication: localization of the cleavage site of the A protein. *Nucleic Acids Res* **22:** 5204-5210.

[257] Liu, Y & Haggård-Liungquist, E (1996) Functional characterization of the P2 A initiator protein and its DNA cleavage site. *Virology* **216:** 158-164.

[258] Liu, Y, Saha, S & Haggård-Liungquist, E (1993) Studies of bacteriophage P2 DNA replication. The DNA sequence of the cis-acting gene A and ori region and construction of a P2 mini-chromosome. *J Mol Biol* **231:** 361-374.

[259] Lobocka, MB, Rose, DJ, Plunkett, G, III, Rusin, M, Samojedny, A, Lehnherr, H, Yarmolinsky, MB & Blattner, FR (2004) Genome of bacteriophage P1. *J Bacteriol* **186:** 7032-7068.

[260] Lohman, TM & Ferrari, ME (1994) *Escherichia coli* single-stranded DNA-binding protein: multiple DNA-binding modes and cooperativities. *Annu Rev Biochem* **63:** 527-570.

[261] Lohman, TM & Overman, LB (1985) Two binding modes in *Escherichia coli* single strand binding protein-single stranded DNA complexes. Modulation by NaCl concentration. *J Biol Chem* **260:** 3594-3603.

[262] Lopez de Saro, FJ, Georgescu, RE, Goodman, MF & O'Donnell, M (2003) Competitive processivity-clamp usage by DNA polymerases during DNA replication and repair. *EMBO J* **22:** 6408-6418.

[263] Lopez de Saro FJ & O'Donnell M (2001) Interaction of the beta sliding clamp with MutS, ligase, and DNA polymerase I. *Proc Natl Acad Sci USA* **98:** 8376-8380.

[264] Low RL, Shlomai J & Kornberg A (1982) Protein n, a primosomal DNA replication protein of *Escherichia coli*. Purification and characterization. *J Biol Chem* **257:** 6242-6250.

[265] Lu YB, Ratnakar PVAL, Mohanty BK & Bastia D (1996) Direct physical interaction between DnaG primase and DnaB helicase of *Escherichia coli* is necessary for optimal synthesis of primer RNA. *Proc Natl Acad Sci USA* **93:** 12902-12907.

[266] Ludlam AV, McNatt MW, Carr KM & Kaguni JM (2001) Essential amino acids of *Escherichia coli* DnaC protein in an N-terminal domain interact with DnaB helicase. *J Biol Chem* **276:** 27345-27353.

[267] Lusetti SL & Cox MM (2002) The Bacterial RecA Protein and the Recombinational DNA Repair of Stalled Replication Forks. *Annu Rev Biochem* **71:** 71-100.

[268] Madsen SM, Mills D, Djordjevic G, Israelsen H & Klaenhammer TR (2001) Analysis of the genetic switch and replication region of a P335-type bacteriophage with an obligate lytic lifestyle on *Lactococcus lactis*. *Appl Environ Microbiol* **67:** 1128-1139.

[269] Maeda K, Nojiri H, Shintani M, Yoshida T, Habe H & Omori T (2003) Complete nucleotide sequence of carbazole/dioxin-degrading plasmid pCAR1 in *Pseudomonas resinovorans* strain CA10 indicates its mosaicity and the presence of large catabolic transposon Tn4676. *J Mol Biol* **326:** 21-33.

[270] Maisnier-Patin S, Nordström K & Dasgupta S (2001) Replication arrests during a single round of replication of the *Escherichia coli* chromosome in the absence of DnaC activity. *Mol Microbiol* **42:** 1371-1382.

[271] Mallory JB, Alfano C & McMacken R (1990) Host virus interactions in the initiation of bacteriophage lambda DNA replication. Recruitment of *Escherichia coli* DnaB helicase by lambda P replication protein. *J Biol Chem* **265:** 13297-13307.

[272] Mannisto RH, Kivela HM, Paulin L, Bamford DH & Bamford JK (1999) The complete genome sequence of PM2, the first lipid-containing bacterial virus To Be isolated. *Virology* **262:** 355-363.

[273] Marchler-Bauer A, Anderson JB, Cherukuri PF *et al*.. (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res* **33:** D192-D196 (Database Issue).

[274] Mardanov AV, Strakhova TS & Ravin NV (2004) RepA protein of the bacteriophage N15 exhibits activity of DNA helicase. *Dokl Biochem Biophys* **397:** 217-219.

[275] Markiewicz P, Malone C, Chase JW & Rothman-Denes LB (1992) *Escherichia coli* single-stranded DNA-binding protein is a supercoiled template-dependent transcriptional activator of N4 virion RNA polymerase. *Genes Dev* **6:** 2010-2019.

[276] Marsin S & Forterre P (1999) The active site of the rolling circle replication protein Rep75 is involved in site-specific nuclease, ligase and nucleotidyl transferase activities. *Mol Microbiol* **33:** 537-545.

[277] Marsin S, McGovern S, Ehrlich SD, Bruand C & Polard P (2001) Early steps of *Bacillus subtilis* primosome assembly. *J Biol Chem* **276:** 45818-45825.

[278] Marszalek J & Kaguni JM (1994) DnaA protein directs the binding of DnaB protein in initiation of DNA replication in *Escherichia coli*. *J Biol Chem* **269:** 4883-4890.

[279] Marszalek J, Zhang WG, Hupp TR, Margulies C, Carr KM, Cherry S & Kaguni JM (1996) Domains of DnaA protein involved in interaction with DnaB protein, and in unwinding the *Escherichia coli* chromosomal origin. *J Biol Chem* **271:** 18535-18542.

[280] Martinez-Jimenez MI, Alonso JC & Ayora S (2005) Bacillus subtilis Bacteriophage SPP1-encoded Gene 34.1 Product is a Recombination-dependent DNA Replication Protein. *J Mol Biol* **351:** 1007-1019.

[281] McGlynn P & Lloyd RG (2000) Modulation of RNA polymerase by (p)ppGpp reveals a RecG-dependent mechanism for replication fork progression. *Cell* **101:** 35-45.

[282] McGrath S, Seegers JF, Fitzgerald GF & van Sinderen D (1999) Molecular characterization of a phage-encoded resistance system in *Lactococcus lactis*. *Appl Environ Microbiol* **65:** 1891-1899.

[283] McHenry CS (2003) Chromosomal replicases as asymmetric dimers: studies of subunit arrangement and functional consequences. *Mol Microbiol* **49:** 1157-1165.

[284] McIntosh PK, Dunker R, Mulder C & Brown NC (1978) DNA of *Bacillus subtilis* bacteriophage SPP1: physical mapping and localization of the origin of replication. *J Virol* **28:** 865-876.

[285] McMacken R, Ueda K & Kornberg A (1977) Migration of *Escherichia coli* DnaB protein on the template DNA strand as a mechanism in initiating DNA replication. *Proc Natl Acad Sci USA* **74:** 4190-4194.

[286] Meijer M, Beck E, Hansen FG, Bergmans HE, Messer W, von Meyenburg K & Schaller H (1979) Nucleotide sequence of the origin of replication of the ***E. coli*** K-12 chromosome *Proc Natl Acad Sci USA* **76:** 580-584.

[287] Meijer WJ, Horcajadas JA & Salas M (2001) Phi29 family of phages. *Microbiol Mol Biol Rev* **65:** 261-287.

[288] Mendelman LV, Notarnicola SM & Richardson CC (1992) Roles of bacteriophage T7 gene 4 proteins in providing primase and helicase functions in vivo. *Proc Natl Acad Sci USA* **89:** 10638-10642.

[289] Mendelman LV & Richardson CC (1991) Requirements for primer synthesis by bacteriophage T7 63-kDa gene 4 protein.

Roles of template sequence and T7 56-kDa gene 4 protein. *J Biol Chem* **266:** 23240-23250.
[290] Messer W (2002) The bacterial replication initiator DnaA. DnaA and oriC, the bacterial mode to initiate DNA replication. *FEMS Microbiol Rev* **26:** 355-374.
[291] Messer W, Hartmann-Kühlein H, Langer U, Mahlow E, Roth A, Schaper S, Urmoneit B & Woelker B (1992) The complex for replication initiation of *Escherichia coli*. *Chromosoma* **102:** S1-S6.
[292] Messer W & Weigel C (1996) Initation of chromosome replication. In: *Escherichia coli* and *Salmonella,* Cellular and Molecular Biology (Neidhardt FC, Curtiss III R, Ingraham J, Lin ECC, Low KB, Magasanik B, Reznikoff WS, Riley M, Schaechter M & Umbarger HE, Eds.), pp. 1579-1601. ASM Press, Washington, D.C., USA.
[293] Messer W & Weigel C (1997) DnaA initiator - also a transcription factor. *Mol Microbiol* **24:** 1-6.
[294] Meyer RR & Laine PS (1990) The single-stranded DNA-binding protein of *Escherichia coli*. *Microbiol Rev* **54:** 342-380.
[295] Meyer TF & Geider K (1979) Bacteriophage fd gene II-protein. II. Specific cleavage and relaxation of supercoiled RF from filamentous phages. *J Biol Chem* **254:** 12642-12646.
[296] Meyer TF, Geider K, Kurz C & Schaller H (1979) Cleavage site of bacteriophage fd gene II-protein in the origin of viral strand replication. *Nature* **278:** 365-367.
[297] Missich R, Weise F, Chai S, Lurz R, Pedré X & Alonso JC (1997) The replisome organizer (G38P) of Bacillus subtilis bacteriophage SPP1 forms specialized nucleoprotein complexes with two discrete distant regions of the SPP1 genome. *J Mol Biol* **270:** 50-64.
[298] Moarefi I, Jeruzalmi D, Turner J, O'Donnell M & Kuriyan J (2000) Crystal structure of the DNA polymerase processivity factor of T4 bacteriophage. *J Mol Biol* **296:** 1215-1223.
[299] Moore DD, Denniston KJ & Blattner FR (1981) Sequence organization of the origins of DNA replication in lambdoid coliphages. *Gene* **14:** 91-101.
[300] Moore T, McGlynn P, Ngo HP, Sharples GJ & Lloyd RG (2003) The RdgC protein of *Escherichia coli* binds DNA and counters a toxic effect of RecFOR in strains lacking the replication restart protein PriA. *EMBO J* **22:** 735-745.
[301] Moreira D (2000) Multiple independent horizontal transfers of informational genes from bacteria to plasmids and phages: implications for the origin of bacterial replication machinery. *Mol Microbiol* **35:** 1-5.
[302] Morris CF, Sinha NK & Alberts BM (1975) Reconstruction of bacteriophage T4 DNA replication apparatus from purified components: rolling circle replication following de novo chain initiation on a single-stranded circular DNA template. *Proc Natl Acad Sci USA* **72:** 4800-4804.
[303] Moscoso M & Suarez JE (2000) Characterization of the DNA replication module of bacteriophage A2 and use of its origin of replication as a defense against infection during milk fermentation by Lactobacillus casei. *Virology* **273:** 101-111.
[304] Moser MJ, Holley WR, Chatterjee A & Mian IS (1997) The proofreading domain of *Escherichia coli* DNA polymerase I and other DNA and/or RNA exonuclease domains. *Nucleic Acids Res* **25:** 5110-5118.
[305] Mosig G (1998) Recombination and Recombination-dependent DNA Replication in Bacteriophage T4. *Annu Rev Genet* **32:** 379-413.
[306] Moyer KE, Kimsey HH & Waldor MK (2001) Evidence for a rolling-circle mechanism of phage DNA synthesis from both replicative and integrated forms of CTXphi. *Mol Microbiol* **41:** 311-323.
[307] Mueser TC, Jones CE, Nossal NG & Hyde CC (2000) Bacteriophage T4 gene 59 helicase assembly protein binds replication fork DNA. The 1.45 A resolution crystal structure reveals a novel alpha-helical two-domain fold. *J Mol Biol* **296:** 597-612.
[308] Mueser TC, Nossal NG & Hyde CC (1996) Structure of bacteriophage T4 RNase H, a 5' to 3' RNA-DNA and DNA-DNA exonuclease with sequence similarity to the RAD2 family of eukaryotic proteins. *Cell* **85:** 1101-1112.
[309] Mufti S & Bernstein H (1974) The DNA-delay mutants of bacteriophage T4. *J Virol* **14:** 860-871.
[310] Muniyappa K & Radding CM (1986) The homologous recombination system of phage lambda. Pairing activities of beta protein. *J Biol Chem* **261:** 7472-7478.
[311] Murata T, Ohnishi M, Ara T, *et al*. (2002) Complete nucleotide sequence of plasmid Rts1: implications for evolution of large plasmid genomes. *J Bacteriol* **184:** 3194-3202.
[312] Murphy KC, Casey L, Yannoutsos N, Poteete AR & Hendrix RW (1987) Localization of a DNA-binding determinant in the bacteriophage P22 Erf protein. *J Mol Biol* **194:** 105-117.
[313] Muyrers JP, Zhang Y, Buchholz F & Stewart AF (2000) RecE/RecT and Redalpha/Redbeta initiate double-stranded break repair by specifically interacting with their respective partners. *Genes Dev* **14:** 1971-1982.
[314] Nakai H, Doseeva V & Jones JM (2001) Handoff from recombinase to replisome: insights from transposition. *Proc Natl Acad Sci USA* **98:** 8247-8254.
[315] Nakai H & Richardson CC (1986) Interactions of the DNA polymerase and gene 4 protein of bacteriophage T7. Protein-protein and protein-DNA interactions involved in RNA-primed DNA synthesis. *J Biol Chem* **261:** 15208-15216.
[316] Nakai H & Richardson CC (1988) The effect of the T7 and *Escherichia coli* DNA-binding proteins at the replication fork of bacteriophage T7. *J Biol Chem* **263:** 9831-9839.
[317] Nanduri B, Byrd AK, Eoff RL, Tackett AJ & Raney KD (2002) Pre-steady-state DNA unwinding by bacteriophage T4 Dda helicase reveals a monomeric molecular motor. *Proc Natl Acad Sci USA* **99:** 14722-14727.
[318] Neuwald AF, Aravind L, Spouge JL & Koonin EV (1999) AAA(+): A class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes. *Genome Res* **9:** 27-43.
[319] Ng JY & Marians KJ (1996) The ordered assembly of the phiX174-type primosome .2. Preservation of primosome composition from assembly through replication. *J Biol Chem* **271:** 15649-15655.
[320] Niedenzu T, Roleke D, Bains G, Scherzinger E & Saenger W (2001) Crystal structure of the hexameric replicative helicase RepA of plasmid RSF1010. *J Mol Biol* **306:** 479-487.
[321] Nossal NG (1980) RNA priming of DNA replication by bacteriophage T4 proteins. *J Biol Chem* **255:** 2176-2182.
[322] Notarnicola SM, Mulcahy HL, Lee J & Richardson CC (1997) The acidic carboxyl terminus of the bacteriophage T7 gene 4 helicase/primase interacts with T7 DNA polymerase. *J Biol Chem* **272:** 18425-18433.
[323] Notarnicola SM, Park K, Griffith JD & Richardson CC (1995) A domain of the gene 4 helicase/primase of bacteriophage T7 required for the formation of an active hexamer. *J Biol Chem* **270:** 20215-20224.
[324] O'Donnell M, Jeruzalmi D & Kuriyan J (2001) Clamp loader structure predicts the architecture of DNA polymerase III holoenzyme and RFC. *Curr Biol* **11:** R935-R946.
[325] Odegrip R & Haggård-Liungquist E (2001) The two active-site tyrosine residues of the a protein play non-equivalent roles during initiation of rolling circle replication of bacteriophage p2. *J Mol Biol* **308:** 147-163.
[326] Odegrip R, Schoen S, Haggård-Liungquist E, Park K & Chattoraj DK (2000) The interaction of bacteriophage P2 B protein with *Escherichia coli* DnaB helicase. *J Virol* **74:** 4057-4063.

[327] Oeser B, Gessner-Ulrich K, Deing P & Tudzynski P (1993) pClK1 and pClT5 - Two Linear Mitochondrial Plasmids from Unrelated *Claviceps purpurea* Strains: A Comparison. *Plasmid* **30:** 274-280.
[328] Ogawa T (1975) Analysis of dnaB function of *Escherichia coli* K12 and the dnaB-like function of P1 prophage. *J Mol Biol* **94:** 327-340.
[329] Ogawa T, Ogawa H & Tomizawa J (1988) Organization of the early region of bacteriophage phi 80. *Gene*s and proteins. *J Mol Biol* **202:** 537-550.
[330] Ogawa T & Okazaki T (1979) RNA-linked nascent DNA pieces in phage T7-infected *Escherichia coli*. III. Detection of intact primer RNA. *Nucleic Acids Res* **7:** 1621-1633.
[331] Ohmori H, Friedberg EC, Fuchs RP, *et al*. (2001) The Y-family of DNA polymerases. *Mol Cell* **8:** 7-8.
[332] Okazaki T, Kurosawa Y, Ogawa T, *et al*. (1979) Structure and metabolism of the RNA primer in the discontinuous replication of prokaryotic DNA. *Cold Spring Harbor Symp Quant Biol* **43(1):** 203-219.
[333] Ostergaard S, Brondsted L & Vogensen FK (2001) Identification of a replication protein and repeats essential for DNA replication of the temperate lactococcal bacteriophage TP901-1. *Appl Environ Microbiol* **67:** 774-781.
[334] Pansegrau W & Lanka E (1991) Common sequence motifs in DNA relaxases and nick regions from a variety of DNA transfer systems. *Nucleic Acids Res* **19:** 3455
[335] Pansegrau W & Lanka E (1992) A common sequence motif among prokaryotic DNA primases. *Nucleic Acids Res* **20:** 4931
[336] Passy SI, Yu X, Li Z, Radding CM & Egelman EH (1999) Rings and filaments of beta protein from bacteriophage lambda suggest a superfamily of recombination proteins. *Proc Natl Acad Sci USA* **96:** 4279-4284.
[337] Patel PH, Suzuki M, Adman E, Shinkai A & Loeb LA (2001) Prokaryotic DNA polymerase I: evolution, structure, and "base flipping" mechanism for nucleotide selection. *J Mol Biol* **308:** 823-837.
[338] Patel SS & Hingorani MM (1993) Oligomeric structure of bacteriophage T7 DNA primase/helicase proteins. *J Biol Chem* **268:** 10668-10675.
[339] Patel SS & Picha KM (2000) Structure and Function of Hexameric Helicases. *Annu Rev Biochem* **69:** 651-697.
[340] Pecenkova T & Paces V (1999) Molecular phylogeny of phi29-like phages and their evolutionary relatedness to other protein-primed replicating phages and other phages hosted by gram-positive bacteria. *J Mol Evol* **48:** 197-208.
[341] Pedré X, Weise F, Chai S, Lüder G & Alonso JC (1994) Analysis of cis and trans acting elements required for the initiation of DNA replication in the Bacillus subtilis bacteriophage SPP1. *J Mol Biol* **236:** 1324-1340.
[342] Pedulla ML, Ford ME, Houtz JM, *et al*. (2003) Origins of highly mosaic mycobacteriophage genomes. *Cell* **113:** 171-182.
[343] Peeters BP, Peters RM, Schoenmakers JG & Konings RN (1985) Nucleotide sequence and genetic organization of the genome of the N-specific filamentous bacteriophage IKe. Comparison with the genome of the F-specific filamentous phages M13, fd and f1. *J Mol Biol* **181:** 27-39.
[344] Permina EA, Mironov AA & Gelfand MS (2002) Damage-repair error-prone polymerases of eubacteria: association with mobile genome elements. *Gene* **293:** 133-140.
[345] Petit MA & Ehrlich SD (2002) Essential bacterial helicases that counteract the toxicity of recombination proteins. *EMBO J* **21:** 3137-3147.
[346] Petri JB & Backhaus H (1984) Structural organization of the ori site of phage P22; comparison with other lambdoid ori sites. *Gene* **32:** 304-310.
[347] Picha KM & Patel SS (1998) Bacteriophage T7 DNA helicase binds dTTP, forms hexamers, and binds DNA in the absence of Mg2+. The presence of dTTP is sufficient for hexamer formation and DNA binding. *J Biol Chem* **273:** 27315-27319.
[348] Piechocki R, Kupper D, Quinones A & Langhammer R (1986) Mutational specificity of a proof-reading defective *Escherichia coli dnaQ49* mutator. *Mol Gen Genet* **202:** 162-168.
[349] Pietroni P, Young MC, Latham GJ & von Hippel PH (2001) Dissection of the ATP-driven reaction cycle of the bacteriophage T4 DNA replication processivity clamp loading system. *J Mol Biol* **309:** 869-891.
[350] Polo S, Sturniolo T, Dehò G & Ghisotti D (1996) Identification of a phage-coded DNA-binding protein that regulates transcription from late promoters in bacteriophage P4. *J Mol Biol* **257:** 745-755.
[351] Poteete AR & Fenton AC (1984) Lambda red-dependent growth and recombination of phage P22. *Virology* **134:** 161-167.
[352] Poteete AR, Fenton AC & Semerjian AV (1991) Bacteriophage P22 accessory recombination function. *Virology* **182:** 316-323.
[353] Poteete AR, Sauer RT & Hendrix RW (1983) Domain structure and quaternary organization of the bacteriophage P22 Erf protein. *J Mol Biol* **171:** 401-418.
[354] Potrykus K, Wrobel B, Wegrzyn A & Wegrzyn G (2000) Replication of oriJ-based plasmid DNA during the stringent and relaxed responses of Escherichia coli. *Plasmid* **44:** 111-126.
[355] Pritchard AE & McHenry CS (1999) Identification of the acidic residues in the active site of DNA polymerase III. *J Mol Biol* **285:** 1067-1080.
[356] Pugh JC & Ritchie DA (1984) Formation of phage T1 concatemers by the RecE recombination pathway of *Escherichia coli*. *Virology* **135:** 200-206.
[357] Raaijmakers H, Vix O, Toro I, Golz S, Kemper B & Suck D (1999) X-ray structure of T4 endonuclease VII: a DNA junction resolvase with a novel fold and unusual domain-swapped dimer architecture. *EMBO J* **18:** 1447-1458.
[358] Rafferty JB, Bolt EL, Muranova TA, Sedelnikova SE, Leonard P, Pasquo A, Baker PJ, Rice DW, Sharples GJ & Lloyd RG (2003) The structure of *Escherichia coli* RusA endonuclease reveals a new Holliday junction DNA binding fold. *Structure (Camb)* **11:** 1557-1567.
[359] Rakonjac J, Ward LJH, Schiemann AH, Gardner PP, Lubbers MW & O'Toole PW (2003) Sequence Diversity and Functional Conservation of the Origin of Replication in Lactococcal Prolate Phages. *Appl Environ Microbiol* **69:** 5104-5114.
[360] Raney KD & Benkovic SJ (1995) Bacteriophage T4 Dda helicase translocates in a unidirectional fashion on single-stranded DNA. *J Biol Chem* **270:** 22236-22242.
[361] Rangarajan S, Woodgate R & Goodman MF (2002) Replication restart in UV-irradiated *Escherichia coli* involving pols II, III, V, PriA, RecA and RecFOR proteins. *Mol Microbiol* **43:** 617-628.
[362] Ratnakar PVAL, Mohanty BK, Lobert M & Bastia D (1996) The replication initiator protein pi of the plasmid R6K specifically interacts with the host-encoded helicase DnaB. *Proc Natl Acad Sci USA* **93:** 5522-5526.
[363] Rattray AJ & Strathern JN (2003) Error-prone DNA polymerases: when making a mistake is the only way to get ahead. *Annu Rev Genet* **37:** 31-66.
[364] Ravin NV, Kuprianov VV, Gilcrease EB & Casjens SR (2003) Bidirectional replication from an internal ori site of the linear N15 plasmid prophage. *Nucleic Acids Res* **31:** 6552-6560.
[365] Ravin V, Ravin N, Casjens S, Ford ME, Hatfull GF & Hendrix RW (2000) Genomic sequence and analysis of the atypical temperate bacteriophage N15. *J Mol Biol* **299:** 53-73.

© 2006 Federation of European Microbiological Societies

[366] Reha-Krantz LJ (1998) Regulation of DNA polymerase exonucleolytic proofreading activity: studies of bacteriophage T4 "antimutator" DNA polymerases. *Genetics* **148:** 1551-1557.
[367] Reha-Krantz LJ & Hurwitz J (1978) The *dnaB* gene product of *Escherichia coli.* II. Single stranded DNA-dependent ribonuclease triphosphatase activity. *J Biol Chem* **253:** 4051-4057.
[368] Reiser W, Leibrecht I & Klein A (1983) Structure and function of mutants in the P gene of bacteriophage lambda leading to the pi phenotype. *Mol Gen Genet* **192:** 430-435.
[369] Richardson CC (1983) Bacteriophage T7: minimal requirements for the replication of a duplex DNA molecule. *Cell* **33:** 315-317.
[370] Richardson RW & Nossal NG (1989) Characterization of the bacteriophage T4 gene 41 DNA helicase. *J Biol Chem* **264:** 4725-4731.
[371] Richardson RW & Nossal NG (1989) Trypsin cleavage in the COOH terminus of the bacteriophage T4 gene 41 DNA helicase alters the primase-helicase activities of the T4 replication complex in vitro. *J Biol Chem* **264:** 4732-4739.
[372] Rodriguez I, Lazaro JM, Blanco L, *et al*. (2005) A specific subdomain in phi29 DNA polymerase confers both processivity and strand-displacement capacity. *Proc Natl Acad Sci USA* **102:** 6407-6412.
[373] Rossmann MG, Moras D & Olsen KW (1974) Chemical and biological evolution of nucleotide-binding protein. *Nature* **250:** 194-199.
[374] Rost B & Sander C (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **19:** 55-72.
[375] Roten CA, Gamba P, Barblan JL & Karamata D (2002) Comparative Genometrics (CG): a database dedicated to biometric comparisons of whole genomes. *Nucleic Acids Res* **30:** 142-144.
[376] Roth A & Messer W (1995) The DNA binding domain of the initiator protein DnaA. *EMBO J* **14:** 2106-2111.
[377] Roth MJ, Brown DR & Hurwitz J (1984) Analysis of bacteriophage phi X174 gene A protein-mediated termination and reinitiation of phi X DNA synthesis. II. Structural characterization of the covalent phi X A protein-DNA complex. *J Biol Chem* **259:** 10556-10568.
[378] Rowen L & Kornberg A (1978) Primase, the DnaG protein of *Escherichia coli*. An enzyme which starts DNA chains. *J Biol Chem* **253:** 758-764.
[379] Rubin EJ, Lin W, Mekalanos JJ & Waldor MK (1998) Replication and integration of a *Vibrio cholerae* cryptic plasmid linked to the CTX prophage. *Mol Microbiol* **28:** 1247-1254.
[380] Ryder L, Sharples GJ & Lloyd RG (1996) Recombination-dependent growth in exonuclease-depleted *recBC sbcBC* strains of *Escherichia coli* K-12. *Genetics* **143:** 1101-1114.
[381] Salinas F & Benkovic SJ (2000) Characterization of bacteriophage T4-coordinated leading- and lagging-strand synthesis on a minicircle substrate. *Proc Natl Acad Sci USA* **97:** 7196-7201.
[382] San Martin C, Radermacher M, Wolpensinger B, Engel A, Miles CS, Dixon NE & Carazo JM (1998) Three-dimensional reconstructions from cryoelectron microscopy images reveal an intimate complex between helicase DnaB and its loading partner DnaC. *Structure* **6:** 501-509.
[383] San Martin MC, Stamford NPJ, Dammerova N, Dixon NE & Carazo JM (1995) A structural model for the *Escherichia coli* DnaB helicase based on electron microscopy data. *J Struct Biol* **114:** 167-176.
[384] Sanders GM, Kassavetis GA & Geiduschek EP (1995) Rules governing the efficiency and polarity of loading a tracking clamp protein onto DNA: determinants of enhancement in bacteriophage T4 late transcription. *EMBO J* **14:** 3966-3976.
[385] Sandler SJ, Marians KJ, Zavitz KH, Coutu J, Parent MA & Clark AJ (1999) dnaC mutations suppress defects in DNA replication- and recombination-associated functions in priB and priC double mutants in *Escherichia coli* K-12. *Mol Microbiol* **34:** 91-101.
[386] Sanger F, Air GM, Barrell BG, *et al*. (1977) Nucliotide sequence of bacteriophage phi X174 DNA. *Nature* **265:** 687-695.
[387] Sawaya MR, Guo S, Tabor S, Richardson CC & Ellenberger T (1999) Crystal structure of the helicase domain from the replicative helicase-primase of bacteriophage T7. *Cell* 99: 167-177.
[388] Schaaper RM (1993) Base selection, proofreading, and mismatch repair during DNA replication in *Escherichia coli*. *J Biol Chem* **268:** 23762-23765.
[389] Schaper S & Messer W (1995) Interaction of the initiator protein DnaA of *Escherichia coli* with its DNA target. *J Biol Chem* **270:** 17622-17626.
[390] Schekman R, Weiner JH, Weiner A & Kornberg A (1975) Ten proteins required for conversion of phiX174 single-stranded DNA to duplex form in vitro. Resolution and reconstitution. *J Biol Chem* **250:** 5859-5865.
[391] Scherzinger E, Lanka E, Morelli G, Seiffert D & Yuki A (1977) Bacteriophage-T7-induced DNA-priming protein. A novel enzyme involved in DNA replication. *Eur J Biochem* **72:** 543-558.
[392] Scherzinger E, Ziegelin G, Bárcena M, Carazo JM, Lurz R & Lanka E (1997) The RepA protein of plasmid RSF1010 is a replicative DNA helicase. *J Biol Chem* **272:** 30228-30236.
[393] Scheuermann R, Tam S, Burgers PM, Lu C & Echols H (1983) Identification of the epsilon-subunit of *Escherichia coli* DNA polymerase III holoenzyme as the *dnaQ* gene product: a fidelity subunit for DNA replication. *Proc Natl Acad Sci USA* **80:** 7085-7089.
[394] Schiemann AH, Rakonjac J, Callanan M, Gordon J, Polzin K, Lubbers MW & O'Toole PW (2004) Essentiality of the early transcript in the replication origin of the lactococcal prolate phage c2. *J Bacteriol* **186:** 8010-8017.
[395] Schnos M, Zahn K, Blattner FR & Inman RB (1989) DNA looping induced by bacteriophage lambda O protein: implications for formation of higher order structures at the lambda origin of replication. *Virology* **168:** 370-377.
[396] Schnos M, Zahn K, Inman RB & Blattner FR (1988) Initiation protein induced helix destabilization at the lambda origin: A prepriming step in DNA replication. *Cell* **52:** 385-395.
[397] Scholl D, Kieleczawa J, Kemp P, Rush J, Richardson CC, Merril C, Adhya S & Molineux IJ (2004) Genomic analysis of bacteriophages SP6 and K1-5, an estranged subgroup of the T7 supergroup. *J Mol Biol* **335:** 1151-1171.
[398] Schouler C, Ehrlich SD & Chopin MC (1994) Sequence and organization of the lactococcal prolate-headed bIL67 phage genome. *Microbiology* **140(11)**: 3061-3069.
[399] Schrock RD & Alberts B (1996) Processivity of the gene 41 DNA helicase at the bacteriophage T4 DNA replication fork. *J Biol Chem* **271:** 16678-16682.
[400] Schuster H, Lanka E, Edelbluth C, Geschke B, Mikolajczyk M, Schlicht M & Touati-Schwartz D (1979) A dnaB-analog DNA-replication protein of phage P1. *Cold Spring Harbor Symp Quant Biol* **43(I):** 551-557.
[401] Scott JF, Eisenberg S, Bertsch LL & Kornberg A (1977) A mechanism of duplex DNA replication revealed by enzymatic studies of phage phi X174: catalytic strand separation in advance of replication. *Proc Natl Acad Sci USA* **74:** 193-197.
[402] Seguritan V, Feng IW, Rohwer F, Swift M & Segall AM (2003) Genome sequences of two closely related *Vibrio parahaemolyticus* phages, VP16T and VP16C. *J Bacteriol* **185:** 6434-6447.
[403] Seigneur M, Bidnenko V, Ehrlich SD & Michel B (1998) RuvAB acts at arrested replication forks. *Cell* **95:** 419-430.

[404] Seitz H, Weigel C & Messer W (2000) The interaction domains of the DnaA and DnaB replication proteins of *Escherichia coli*. *Mol Microbiol* **37:** 1270-1279.
[405] Seki T & Okazaki T (1979) RNA-linked nascent DNA pieces in phage T7-infected *Escherchia coli*. II. Primary structure of the RNA portion. *Nucleic Acids Res* **7:** 1603-1619.
[406] Seufert W & Messer W (1987) Start sites for bidirectional *in vitro* replication inside the replication origin, *oriC*, of *Escherichia coli*. *EMBO J* **6:** 2469-2472.
[407] Shamoo Y & Steitz TA (1999) Building a replisome from interacting pieces: sliding clamp complexed to a peptide from DNA polymerase and a polymerase editing complex. *Cell* **99:** 155-166.
[408] Sharples GJ (2001) The X philes: structure-specific endonucleases that resolve Holliday junctions. *Mol Microbiol* **39:** 823-834.
[409] Sharples GJ, Corbett LM & Graham IR (1998) lambda Rap protein is a structure-specific endonuclease involved in phage recombination. *Proc Natl Acad Sci USA* **95:** 13507-13512.
[410] Sharples GJ, Ingleston SM & Lloyd RG (1999) Holliday junction processing in bacteria: insights from the evolutionary conservation of RuvABC, RecG, and RusA. *J Bacteriol* **181:** 5543-5550.
[411] Shiffman D & Cohen SN (1992) Reconstruction of a *Streptomyces* linear replicon from separately cloned DNA fragments: existence of a cryptic origin of circular replication within the linear plasmid. *Proc Natl Acad Sci USA* **89:** 6129-6133.
[412] Shinozaki K & Okazaki T (1977) RNA-linked nascent DNA pieces in T7 phage-infected *Escherichia coli* cells. I. Role of gene 6 exonuclease in removal of the linked RNA. *Mol Gen Genet* **154:** 263-267.
[413] Shlomai J & Kornberg A (1980) An *E. coli* replication protein that recognizes a unique sequence within a hairpin region in ΦX174 DNA. *Proc Natl Acad Sci USA* **77:** 799-803.
[414] Shulman MJ, Hallick LM, Echols H & Signer ER (1970) Properties of recombination-deficient mutants of bacteriophage lambda. *J Mol Biol* **52:** 501-520.
[415] Singleton MR, Sawaya MR, Ellenberger T & Wigley DB (2000) Crystal structure of T7 gene 4 ring helicase indicates a mechanism for sequential hydrolysis of nucleotides. *Cell* **101:** 589-600.
[416] Singleton MR, Wentzell LM, Liu Y, West SC & Wigley DB (2002) Structure of the single-strand annealing domain of human RAD52 protein. *Proc Natl Acad Sci USA* **99:** 13492-13497.
[417] Sivaprasad AV, Jarvinen R, Puspurs A & Egan JB (1990) DNA replication studies with coliphage 186. III. A single phage gene is required for phage 186 replication. *J Mol Biol* **213:** 449-463.
[418] Skarstad K & Boye E (1994) The initiator protein DnaA: Evolution, properties and function. *Biochim Biophys Acta* **1217:** 111-130.
[419] Smith GR (2001) Homologous recombination near and far from DNA breaks: alternative roles and contrasting views. *Annu Rev Genet* **35:** 243-274.
[420] Sokolsky TD & Baker TA (2003) DNA gyrase requirements distinguish the alternate pathways of Mu transposition. *Mol Microbiol* **47:** 397-409.
[421] Soultanas P (2002) A functional interaction between the putative primosomal protein DnaI and the main replicative DNA helicase DnaB in *Bacillus*. *Nucleic Acids Res* **30:** 966-974.
[422] Sousa R (1996) Structural and mechanistic relationships between nucleic acid polymerases. *Trends Biochem Sci* **21:** 186-190.
[423] Speck C & Messer W (2001) Mechanism of origin unwinding: sequential binding of DnaA initiator protein to double-stranded and single-stranded DNA in the AT-rich region of the replication origin. *EMBO J* **20:** 1469-1476.
[424] Speck C, Weigel C & Messer W (1999) ATP- and ADP-DnaA protein, a molecular switch in gene regulation. *EMBO J* **18:** 6169-6176.
[425] Staden R, Beal KF & Bonfield JK (1998) The Staden Package. In: Bioinformatics Methods and Protocols, Vol.132: Computer Methods in Molecular Biology (Misener S & Krawetz SA, Eds.), pp. 115-130. The Humana Press Inc., Totowa NJ, USA.
[426] Stanley E, Walsh L, van der Zwet A, Fitzgerald GF & van Sinderen D (2000) Identification of four loci isolated from two *Streptococcus thermophilus* phage genomes responsible for mediating bacteriophage resistance. *FEMS Microbiol Lett* **182:** 271-277.
[427] Stassen AP, Schoenmakers EF, Yu M, Schoenmakers JG & Konings RN (1992) Nucleotide sequence of the genome of the filamentous bacteriophage I2-2: module evolution of the filamentous phage genome. *J Mol Evol* **34:** 141-152.
[428] Steitz TA (1999) DNA polymerases: structural diversity and common mechanisms. *J Biol Chem* **274:** 17395-17398.
[429] Stenger DC, Revington GN, Stevenson MC & Bisaro DM (1991) Replicational release of geminivirus genomes from tandemly repeated copies: evidence for rolling-circle replication of a plant viral DNA. *Proc Natl Acad Sci USA* **88:** 8029-8033.
[430] Stewart PE, Thalken R, Bono JL & Rosa P (2001) Isolation of a circular plasmid region sufficient for autonomous replication and transformation of infectious *Borrelia burgdorferi*. *Mol Microbiol* **39:** 714-721.
[431] Strack B, Lessl M, Calendar R & Lanka E (1992) A common sequence motif, -E-G-Y-A-T-A-, identified within the primase domains of plasmid-encoded I- and P-type DNA primases and the alpha protein of the *Escherichia coli* satellite phage P4. *J Biol Chem* **267:** 13062-13072.
[432] Stratling W & Knippers R (1973) Function and purification of gene 4 protein of phage T7. *Nature* **245:** 195-197.
[433] Strick T, Allemand J, Croquette V & Bensimon D (2000) Twisting and stretching single DNA molecules. *Prog Biophys Mol Biol* **74:** 115-140.
[434] Strick TR, Allemand JF, Bensimon D & Croquette V (2000) Stress-induced structural transitions in DNA and proteins. *Annu Rev Biophys Biomol Struct* **29:** 523-543.
[435] Studier FW (1969) The genetics and physiology of bacteriophage T7. *Virology* **39:** 562-574.
[436] Studier FW (1972) Bacteriophage T7. *Science* **176:** 367-376.
[437] Sugimoto K, Oka A, Sugisaki H, Takanami M, Nishimura A, Yasuda Y & Hirota Y (1979) Nucleotide sequence of *Escherichia coli* replication origin *Proc Natl Acad Sci USA* **76:** 575-579.
[438] Sun W, Tormo J, Steitz TA & Godson GN (1994) Domains of *Escherichia coli* primase: functional activity of a 47-kDa N-terminal proteolytic fragment *Proc Natl Acad Sci USA* **91:** 11462-11466.
[439] Sung HM, Yeamans G, Ross CA & Yasbin RE (2003) Roles of YqjH and YqjW, homologs of the *Escherichia coli* UmuC/DinB or Y superfamily of DNA polymerases, in stationary-phase mutagenesis and UV-induced mutagenesis of *Bacillus subtilis*. *J Bacteriol* **185:** 2153-2160.
[440] Szpirer J (1972) Control of development in temperate bacteriophages. IV. Specific action of N product at a transcription stop signal. *Mol Gen Genet* **114:** 297-304.
[441] Tabor S, Huber HE & Richardson CC (1987) *Escherichia coli* thioredoxin confers processivity on the DNA polymerase activity of the gene 5 protein of bacteriophage T7. *J Biol Chem* **262:** 16212-16223.
[442] Tanaka T & Ogura M (1998) A novel *Bacillus natto* plasmid pLS32 capable of replication in Bacillus subtilis. *Febs Lett* **422:** 243-246.

© 2006 Federation of European Microbiological Societies

[443] Tarkowski TA, Mooney D, Thomason LC & Stahl FW (2002) Gene products encoded in the *ninR* region of phage lambda participate in Red-mediated recombination. *Genes* Cells **7:** 351-363.
[444] Tatusov RL, Fedorova ND, Jackson JD, *et al*. (2003) The COG database: an updated version includes eukaryotes. BMC *Bioinformatics* **4:** 41-41.
[445] Tatusova TA & Madden TL (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett* **174:** 247-250.
[446] Taylor K & Wegrzyn G (1995) Replication of coliphage lambda DNA. *FEMS Microbiol Rev* **17:** 109-119.
[447] Terzano S, Christian R, Espinoza FH, Calendar R, Dehò G & Ghisotti D (1994) A new gene of bacteriophage P4 that controls DNA replication. *J Bacteriol* **176:** 6059-6065.
[448] Tessman ES (1966) Mutants of bacteriophage S13 blocked in infectious DNA synthesis. *J Mol Biol* **17:** 218-236.
[449] Thirlway J, Turner IJ, Gibson CT, Gardiner L, Brady K, Allen S, Roberts CJ & Soultanas P (2004) DnaG interacts with a linker region that joins the N- and C-domains of DnaB and induces the formation of 3-fold symmetric rings. *Nucleic Acids Res* **32:** 2977-2986.
[450] Thomson N, Sebaihia M, Cerdeno-Tarraga A & Parkhill J (2002) Sibling rivalry. *Trends Microbiol* **10:** 396-397.
[451] Tocchetti A, Galimberti G, Dehò G & Ghisotti D (1999) Characterization of the oriI and oriII origins of replication in phage-plasmid P4. *J Virol* **73:** 7308-7316.
[452] Tocchetti A, Serina S, Oliva I, Dehò G & Ghisotti D (2001) Cnr interferes with dimerization of the replication protein alpha in phage-plasmid P4. *Nucleic Acids Res* **29:** 536-544.
[453] Torgov MY, Janzen DM & Reddy MK (1998) Efficiency and frequency of translational coupling between the bacteriophage T4 clamp loader genes. *J Bacteriol* **180:** 4339-4343.
[454] Touati-Schwartz D (1979) A dnaB analog ban, specified by bacteriophage P1: genetic and physiological evidence for functional analogy and interactions between the two products. *Mol Gen Genet* **174:** 173-188.
[455] Tougu K & Marians KJ (1996) The extreme C terminus of primase is required for interaction with DnaB at the replication fork. *J Biol Chem* **271:** 21391-21397.
[456] Trakselis MA, Alley SC, Abel-Santos E & Benkovic SJ (2001) Creating a dynamic picture of the sliding clamp during T4 DNA polymerase holoenzyme assembly by using fluorescence resonance energy transfer *Proc Natl Acad Sci USA* **98:** 8368-8375.
[457] Tseng TY, Frick DN & Richardson CC (2000) Characterization of a novel DNA primase from the *Salmonella typhimurium* bacteriophage SP6. *Biochemistry* **39:** 1643-1654.
[458] Tsurimoto T & Matsubara K (1981) Purification of bacteriophage lambda-O protein that specifically binds to the origin of replication. *Mol Gen Genet* **181:** 325-331.
[459] Tsurimoto T & Matsubara K (1981) Purified bacteriophage lambda O protein binds to four repeating sequences at the lambda replication origin. *Nucleic Acids Res* **9:** 1789-1799.
[460] Umek RM & Kowalski D (1988) The ease of DNA unwinding as a determinant of initiation at yeast replication origins. *Cell* **52:** 559-567.
[461] Van Dyck E, Foury F, Stillman B & Brill SJ (1992) A single-stranded DNA binding protein required for mitochondrial DNA replication in *S. cerevisiae* is homologous to *E. coli* SSB. *EMBO J* **11:** 3421-3430.
[462] van Mansfeld AD, van Teeffelen HA, Baas PD & Jansz HS (1986) Two juxtaposed tyrosyl-OH groups participate in phi X174 gene A protein catalysed cleavage and ligation of DNA. *Nucleic Acids Res* **14:** 4229-4238.
[463] Vellani TS & Myers RS (2003) Bacteriophage SPP1 Chu is an alkaline exonuclease in the SynExo family of viral two-component recombinases. *J Bacteriol* **185:** 2465-2474.
[464] Velten M, McGovern S, Marsin S, Ehrlich SD, Noirot P & Polard P (2003) A two-protein strategy for the functional loading of a cellular replicative DNA helicase. *Mol Cell* **11:** 1009-1020.
[465] Venkatesan M, Silver LL & Nossal NG (1982) Bacteriophage T4 gene 41 protein, required for the synthesis of RNA primers, is also a DNA helicase. *J Biol Chem* **257:** 12426-12434.
[466] von Hippel PH & Delagoutte E (2003) Macromolecular complexes that unwind nucleic acids. *Bioessays* **25:** 1168-1177.
[467] Wahle E, Lasken RS & Kornberg A (1989) The dnaB-dnaC replication protein complex of *Escherichia coli*. I. Formation and properties. *J Biol Chem* **264:** 2463-2468.
[468] Wahle E, Lasken RS & Kornberg A (1989) The dnaB-dnaC replication protein complex of *Escherichia coli*. II. role of the complex in mobilizing dnaB functions. *J Biol Chem* **264:** 2469-2475.
[469] Waterfield NR, Lubbers MW, Polzin KM, Le Page RW & Jarvis AW (1996) An origin of DNA replication from *Lactococcus lactis* bacteriophage c2. *Appl Environ Microbiol* **62:** 1452-1453.
[470] Waters VL & Guiney DG (1993) Processes at the nick region link conjugation, T-DNA transfer and rolling circle replication. *Mol Microbiol* **9:** 1123-1130.
[471] Wegrzyn A, Wegrzyn G & Taylor K (1995) Plasmid and host functions required for lambda plasmid replication carried out by the inherited replication complex. *Mol Gen Genet* **247:** 501-508.
[472] Weigel C, Messer W, Preiss S, Welzeck M, Morigen and Boye E (2001) The sequence requirements for a functional *Escherichia coli* replication origin are different for the chromosome and a minichromosome. *Mol Microbiol* **40:** 498-507.
[473] Weigel C, Schmidt A, Seitz H, Tuengler D, Welzeck M & Messer W (1999) The N-terminus promotes oligomerisation of the *Escherichia coli* initiator protein DnaA. *Mol Microbiol* **34:** 53-66.
[474] Weigel C & Seitz H (2002) Strand-specific loading of DnaB helicase by DnaA to a substrate mimicking unwound oriC. *Mol Microbiol* **46:** 1149-1156.
[475] Weise F, Chai S, Lüder G & Alonso JC (1994) Nucleotide sequence and complementation studies of the gene 35 region of the *Bacillus subtilis* bacteriophage SPP1. *Virology* **202:** 1046-1049.
[476] Weng SF, Fan YF, Tseng YH & Lin JW (1997) Sequence analysis of the small cryptic *Xanthomonas campestris* pv. vesicatoria plasmid pXV64 encoding a Rep protein similar to gene II protein of phage 12-2. *Biochem Biophys Res Commun* **231:** 121-125.
[477] Wickner RB, Wright M, Wickner S & Hurwitz J (1972) Conversion of phiX174 and fd single-stranded DNA to replicative forms in extracts of *Escherichia coli*. *Proc Natl Acad Sci USA* **69:** 3233-3237.
[478] Wickner S & Hurwitz J (1974) Conversion of phiX174 viral DNA to double-stranded form by purified *Escherichia coli* proteins. *Proc Natl Acad Sci USA* **71:** 4120-4124.
[479] Wickner S & Hurwitz J (1975) Interaction of *Escherichia coli* dnaB and dnaC(D) gene products *in vitro*. *Proc Natl Acad Sci USA* **72:** 921-925.
[480] Wickner S & McKenney K (1987) Deletion analysis of the DNA sequence required for the in vitro initiation of replication of bacteriophage lambda. *J Biol Chem* **262:** 13163-13167.
[481] Wickner SH & Zahn K (1986) Characterization of the DNA binding domain of bacteriophage lambda O protein. *J Biol Chem* **261:** 7537-7543.

[482] Williams KR, Spicer EK, LoPresti MB, Guggenheimer RA & Chase JW (1983) Limited proteolysis studies on the *Escherichia coli* single-stranded DNA binding protein. Evidence for a functionally homologous domain in both the *Escherichia coli* and T4 DNA binding proteins. *J Biol Chem* **258:** 3346-3355.
[483] Willis DK, Satin LH & Clark AJ (1985) Mutation-dependent suppression of *recB21 recC22* by a region cloned from the Rac prophage of *Escherichia coli* K-12. *J Bacteriol* **162:** 1166-1172.
[484] Woelker B & Messer W (1993) The structure of the initiation complex at the replication origin, *oriC,* of *Escherichia coli*. *Nucleic Acids Res* **21:** 5025-5033.
[485] Wojtkowiak D, Georgopoulos C & Zylicz M (1993) Isolation and characterization of ClpX, a new ATP-dependent specificity component of the Clp protease of *Escherichia coli*. *J Biol Chem* **268:** 22609-22617.
[486] Wold MS (1997) Replication Protein A: A Heterotrimeric, Single-Stranded DNA-Binding Protein Required for Eukaryotic DNA Metabolism. *Annu Rev Biochem* **66:** 61-92.
[487] Wong K & Geiduschek EP (1998) Activator-sigma interaction: A hydrophobic segment mediates the interaction of a sigma family promoter recognition protein with a sliding clamp transcription activator. *J Mol Biol* **284:** 195-203.
[488] Wong SW, Wahl AF, Yuan PM, *et al.* (1988) Human DNA polymerase alpha gene expression is cell proliferation dependent and its primary structure is similar to both prokaryotic and eukaryotic replicative DNA polymerases. *EMBO J* **7:** 37-47.
[489] Wrobel B & Wegrzyn G (2002) Evolution of lambdoid replication modules. *Virus Genes* **24:** 163-171.
[490] Yamagami H & Yamamoto N (1970) Contribution of the bacterial recombination function to replication of bacteriophage P2. *J Mol Biol* **53:** 281-285.
[491] Yasbin RE, Fields PI & Andersen BJ (1980) Properties of *Bacillus subtilis* 168 derivatives freed of their natural prophages. *Gene* **12:** 155-159.
[492] Yeo HJ, Ziegelin G, Korolev S, Calendar R, Lanka E & Waksman G (2002) Phage P4 origin-binding domain structure reveals a mechanism for regulation of DNA-binding activity by homo- and heterodimerization of winged helix proteins. *Mol Microbiol* **43:** 855-867.
[493] Yoda K & Okazaki T (1991) Specificity of recognition sequence for *Escherichia coli* primase. *Mol Gen Genet* **227:** 1-8.
[494] Yonesaki T (1994) Involvement of a replicative DNA helicase of bacteriophage T4 in DNA recombination. *Genetics* **138:** 247-252.
[495] Yonesaki T (1994) The purification and characterization of gene 59 protein from bacteriophage T4. *J Biol Chem* **269:** 1284-1289.
[496] Yong Y & Romano LJ (1995) Nucleotide and DNA-induced conformational changes in the bacteriophage T7 gene 4 protein. *J Biol Chem* **270:** 24509-24517.
[497] Yu X, Hingorani MM, Patel SS & Egelman EH (1996) DNA is bound within the central hole to one or two of the six subunits of the T7 DNA helicase. *Nat Struct Biol* **3:** 740-743.
[498] Yu X, Jezewska MJ, Bujalowski W & Egelman EH (1996) The hexameric *E. coli* DnaB helicase can exist in different Quaternary states. *J Mol Biol* **259:** 7-14.
[499] Zahn K & Blattner FR (1985) Binding and bending of the lambda replication origin by the phage O protein. *EMBO J* **4:** 3605-3616.
[500] Zahn K & Blattner FR (1985) Sequence-induced DNA curvature at the bacteriophage lambda origin of replication. *Nature* **317:** 451-453.
[501] Zahn K & Blattner FR (1987) Direct evidence for DNA bending at the lambda replication origin. *Science* **236:** 416-422.
[502] Zechner EL, Wu CA & Marians KJ (1992) Coordinated leading- and lagging-strand synthesis at the *Escherichia coli* DNA replication fork. III. A polymerase-primase interaction governs primer size. *J Biol Chem* **267:** 4054-4063.
[503] Ziegelin G, Calendar R, Ghisotti D, Terzano S & Lanka E (1997) Cnr protein, the negative regulator of bacteriophage P4 replication, stimulates specific DNA binding of its initiator protein alpha. *J Bacteriol* **179:** 2817-2822.
[504] Ziegelin G, Calendar R, Lurz R & Lanka E (1997) The helicase domain of phage P4 alpha protein overlaps the specific DNA binding domain. *J Bacteriol* **179:** 4087-4095.
[505] Ziegelin G, Linderoth NA, Calendar R & Lanka E (1995) Domain structure of phage P4 alpha protein deduced by mutational analysis. *J Bacteriol* **177:** 4333-4341.
[506] Ziegelin G, Scherzinger E, Lurz R & Lanka E (1993) Phage P4 alpha protein is multifunctional with origin recognition, helicase and primase activities. *EMBO J* **12:** 3703-3708.
[507] Ziegelin G, Tegtmeyer N, Lurz R, Hertwig S, Hammerl J, Appel B & Lanka E (2005) The *repA* Gene of the Linear *Yersinia enterocolitica* Prophage PY54 Functions as a Circular Minimal Replicon in *Escherichia coli*. *J Bacteriol* **187:** 3445-3454.
[508] Zuniga M, Franke-Fayard B, Venema G, Kok J & Nauta A (2002) Characterization of the putative replisome organizer of the lactococcal bacteriophage r1t. *J Virol* **76:** 10234-10244.
[509] Zylicz M, Gorska I, Taylor K & Georgopoulos C (1984) Bacteriophage lambda replication proteins: formation of a mixed oligomer and binding to the origin of lambda DNA. *Mol Gen Genet* **196:** 401-406.

Note added in proof:

Due to the very recent change in the layout of FEMS Microbiology Reviews, the original numbering of Sections and Chapters is not preserved in the published main body of this review, but still referred to in this compendium. With the aim to minimise confusion resulting thereof, we present here a 'Table of Contents' of the main body with the original Section and Chapter numbering:

bacteriophage replication modules

Contents: main body