Domestication of Lambda Phage Genes into a Putative Third Type of Replicative Helicase Matchmaker

Pierre Brézellec[1], Marie-Agnès Petit[2], Sophie Pasek[3], Isabelle Vallet-Gely[4], Christophe Possoz[4], and Jean-Luc Ferat[1,4,*]

[1]Universite de Versailles Saint-Quentin en Yvelines UFR des Sciences, France

[2]Micalis Institute, INRA, AgroParisTech, Université Paris-Saclay, Jouy-en-Josas, France

[3]Atelier de Bioinformatique, UMR 7205 ISYEB, CNRS-MNHN-UPMC-EPHE, Muséum d'Histoire Naturelle, Paris, France

[4]Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198, Gif-sur-Yvette cedex, France

*Corresponding author: E-mail: jean-luc.ferat@i2bc.paris-saclay.fr.

Accepted: June 26, 2017

Abstract

At the onset of the initiation of chromosome replication, bacterial replicative helicases are recruited and loaded on the DnaA-*oriC* nucleoprotein platform, assisted by proteins like DnaC/DnaI or DciA. Two orders of bacteria appear, however, to lack either of these factors, raising the question of the essentiality of these factors in bacteria. Through a phylogenomic approach, we identified a pair of genes that could have substituted for *dciA*. The two domesticated genes are specific of the *dnaC/dnaI*- and *dciA*-lacking organisms and apparently domesticated from lambdoid phage genes. They derive from *λO* and *λP* and were renamed *dopC* and *dopE*, respectively. DopE is expected to bring the replicative helicase to the bacterial origin of replication, while DopC might assist DopE in this function. The confirmation of the implication of DopCE in the handling of the replicative helicase at the onset of replication in these organisms would generalize to all bacteria and therefore to all living organisms the need for specific factors dedicated to this function.

Key words: replicative helicase, replication initiation, viral gene domestication, *dnaC*, *dciA*, lambda phage.

Replicative helicases are essential components of the replication machinery where they unwind double stranded DNA in front of the replisome, ensuring speed and processivity to the replication process. In bacteria, these enzymes are recruited at the onset of replication initiation by a nucleoprotein platform built through the ordered assembly of the initiator protein, DnaA, at the unique origin of replication of the chromosome, *oriC*, with the assistance of a factor that will tie the assembly of the replicative helicase hexamer around the parental DNA strand that serves as the lagging strand template with the formation of the replication machinery. This factor, whose function could be depicted as that of a Replicative Helicase Matchmaker (RHeMa), is specified by DnaC or DnaI (DnaC/I)—a family of proteins encoded by genes of phage origin—and DciA, the ancestral bacterial RHeMa (Forterre 1999; Bell and Kaguni 2013; Brezellec et al. 2016). The presence of either of these two RHeMa in bacterial genomes suggested that such systems were essential components of the initiation of replication and universally present in bacteria, as much as CDT1 and CDC6—their *alter ego* in eukaryotes and archaea (Bell and Kaguni 2013; Samson et al. 2016). Yet, we identified several species that apparently lack a *bona fide* RHeMa (Brezellec et al. 2016). Although the absence in most cases likely arose as a consequence of genome reduction during the process of symbiotization, the situation was more perplexing in a few gammaproteobacteria regrouping Cellvibrionales and Oceanospirillales, as *dciA* and *dnaC/I* appear to be missing in a complete clade of closely related and nonsymbiotic species. A phylogenetic analysis of these species shows indeed that *dciA* was lost in an ancestor of the Cellvibrionales and Oceanospirillales (fig. 1a, Materials and Methods), leading to the evolution of successful lineages of species and questioning the essentiality of RHeMa in bacteria.

Considering the ancestrality of *dciA* and the mutual exclusion reported earlier between *dciA* and *dnaC/I* (Brezellec et al. 2016), we hypothesized that if RHeMa is essential in bacteria, then, Cellvibrionales and *dciA*-deprived Oceanospirillales specify another type of RHeMa that emerged at the expense of

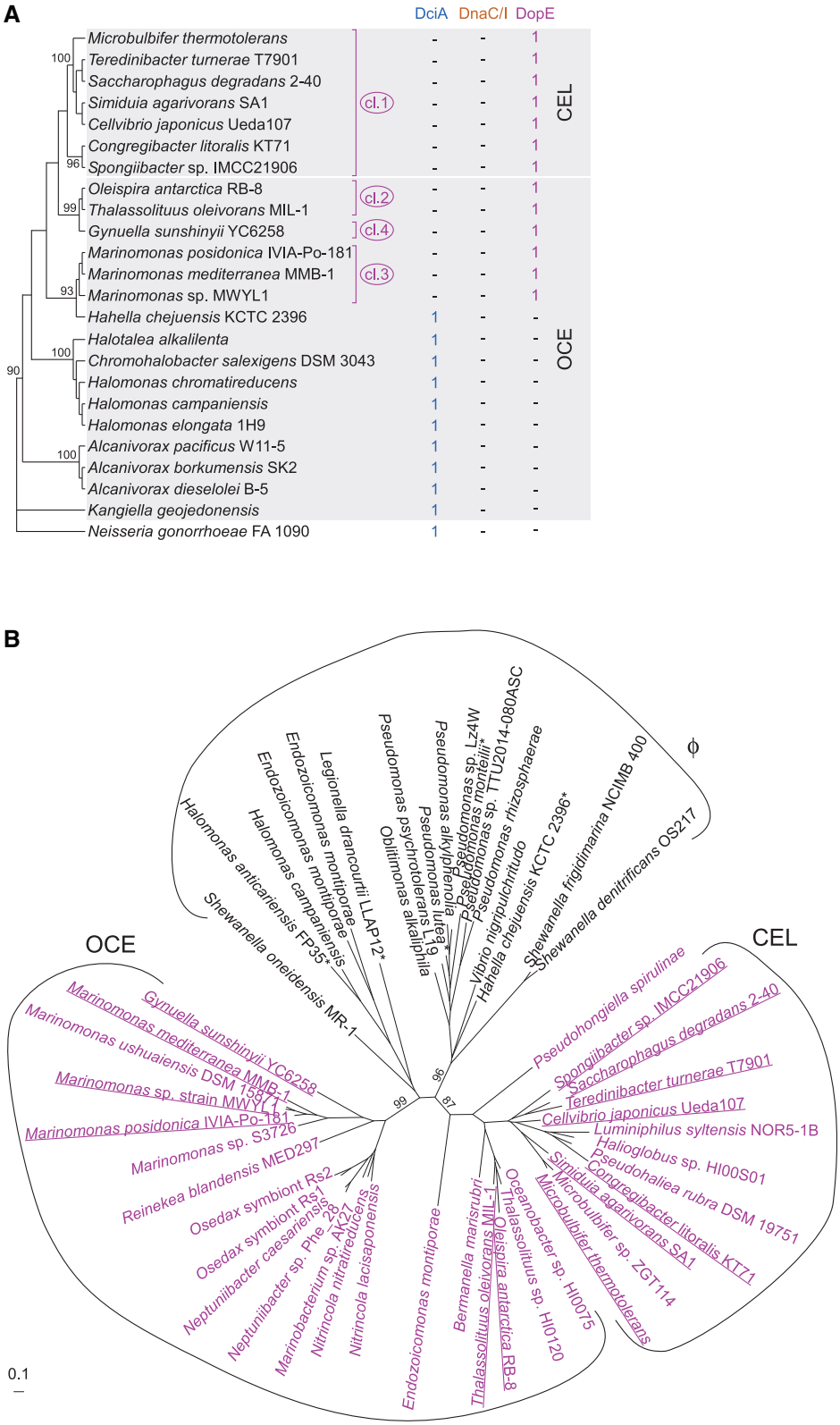

Fig. 1.—*dopE* is a domesticated phage replicative helicase operator gene. (*a*) Cladogram based on DnaA–DnaB–DnaX–DnaE–concatenated protein sequences of the Cellvibrionales and Oceanospirillales (Materials and Methods). The organisms in which a *λP*-like gene was domesticated (DopE) are indicated. Organisms specifying DciA and DnaC/I are also identified. The betaproteobacteria *Neisseria gonorrhoeae* was used as an out-group. DciA, DnaC/I,

dciA. In order to identify this potential RHeMa, we built an IN-group (regrouping genomes of Cellvibrionales and Oceanospirillales that lack *dciA* and *dnaC/I*) and an OUT-group (regrouping *dciA*-containing Oceanospirillales) and searched the Microscope database for genes systematically present in the IN-group and absent in the OUT-group (Materials and Methods). This first step returned 23 genes (sup_tab1), whose presence was assessed by BLAST (E-value cut-off = 10^{-4}) in the remaining genomes of the IN-group missing in the Microscope database. The list of 23 shrank to a single gene encoding a protein (named DopE—see below for description—in fig. 1*a* and supplementary table S1, Supplementary Material online), whose HMM profile is that of the P protein of the phage lambda (PF06992, Prob = 100%); P is the lambda phage RHeMa protein that brings the host replicative helicase DnaB to the phage replication initiation complex (Alfano and McMacken 1989).

The clear phage ancestry of this *λP* -like gene led us to consider a scenario of domestication of *λP* in an ancestor of these species to handle the bacterial replicative helicase in place of DciA, mirroring the DnaC/I situation (Brezellec et al. 2016). To investigate further this hypothesis, we asked whether the genes encoding the P-like proteins found in Cellvibrionales and Oceanospirillales that lack *dciA* and *dnaC/I* were resident, i.e., whether their phylogeny was congruent with that of the host and whether the protein sequences were under selection pressure. Proteins from identified organisms sharing homology with the P-like protein of *Saccharophagus degradans* (obtained by BLASTP against the reference proteome database, *E*-value cut-off = 10^{-4}) were collected to build a phylogenetic tree (fig. 1*b*). P-like proteins are distributed into two distinct groups. Twenty proteins display features typical of mobile elements. Their distribution is not congruent with the host phylogeny, branch length reflects a low selection pressure and the genes that encode these proteins are sparsely distributed in the orders represented. The analysis of their genomic context by two independent prophage detection methods (Lima-Mendez et al. 2008; Arndt et al. 2016) revealed that, except for a few cases where the contexts appear to be genomic, these *λP*-like genes are enclosed in prophages (supplementary table S2, Supplementary Material online) and embedded in genomes carrying a cellular RHeMa (supplementary table S3, Supplementary Material online). In contrast, the other P-like proteins (magenta) are all under stronger selection pressure (shorter branch length) and encoded by *dciA*- and *dnaC/I*-lacking organisms (supplementary tables S2 and S3, Supplementary Material online). Also, their distribution in Cellvibrionales and Oceanospirillales is congruent with the host phylogeny (compare the phylogenetic distribution of the species present in fig. 1*a* and *b*) and prophage detection methods revealed the absence of a prophage context in the DNA surrounding these genes (supplementary tables S2, Supplementary Material online). The inspection of the same regions, using the integrated microbial genome interface (Markowitz et al. 2014) led to the identification of four different classes of genomic contexts (fig. 2), whose distribution is congruent with our bacterial phylogeny (fig. 1*a*), all of them regrouping genes specifying genuine cellular functions and whose transcription direction alternates. Altogether, this indicates that the genes encoding the P-like proteins of the latter group are "resident" and suggests that a phage *λP* gene was domesticated in the *dciA*-containing common ancestor of these organisms at the expense of *dciA* (fig. 1*a*). We renamed the *λP*-like gene *dopE* for domesticated P element.

In the lambda genome, the *λO* gene is located immediately upstream of *λP*. *λO* encodes the initiator protein of lambda that builds the initiation complex and to which the P-hijacked host replicative helicase is brought. In addition, *λO* contains the phage origin of replication '*oriλ*' (Tsurimoto and Matsubara 1981). Interestingly, upstream of *dopE*, settles a gene that shares some features with the lambda *λO* gene (fig. 3). This *λO*-like gene is present in all *dopE* genomic contexts analyzed (fig. 2). The phylogenetic distribution of O-like proteins in bacteria (performed with protein sequences sharing homology with the O-like protein of *S. degradans*; same parameters as above), is congruent with that established for DopE (data not shown), suggesting that *dopE* and its neighboring gene co-evolved and were domesticated together. Accordingly, we renamed this gene *dopC*, for *dopE*-companion gene.

The homology between DopC and O is apparently restricted to the Nter extremity of the proteins. In DopC, the Cter domain "Φ_2220_C" of O is absent. Instead, the protein contains a domain found in the Cter of DnaT, the primosomal protein required for replication fork reactivation in *E. coli* (fig. 2) (Kornberg and Baker 1992). Moreover, we could not detect iteron-like sequences typical of phage replication origins (DnaA boxes, diagnostic of a bacterial origin of replication, were not found either) within the *dopC* locus (Zahn and Blattner 1985). Nonetheless, we investigated the possibility that this latter locus served as an origin of replication by scoring the cumulative GC skew over the entire genome of these bacteria (Materials and Methods) (Sernova and Gelfand 2008) (fig. 4). The lower inversion point of a V-like

Fig. 1. Continued

and DopE are indicated in cyan, brown, and magenta, respectively. Oceanospirillales (OCE) are split into *dciA*- or *dopE*-carying species, while Cellvibrionales (CEL) are all *dopE*-carrying species. (b) Phylogenetic tree of the resident (DopE) and nonresident P-like proteins. Strains in which a *λP*-like gene was domesticated (*dopE*) are in magenta. Underlined genomes are those used in (a). Scale bar represents 0.1 substitution per site. OCE: Oceanospirillales, CEL: Cellvibrionales, Φ: phage. Significant bootstrap values (> 70%) are indicated.*: phage environment could not be established around the *λP*-like gene within these species (supplementary table S2, Supplementary Material online).

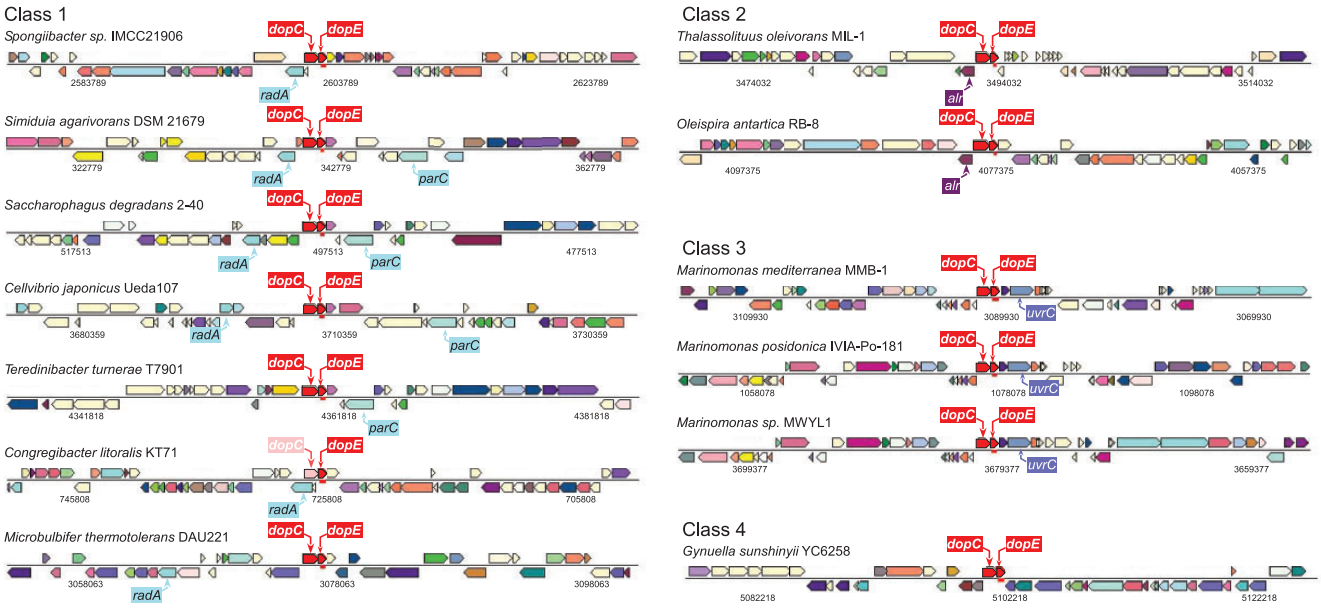

Fig. 2.—*dopCE* is distributed in 4 genomic context classes. The *dopE* gene of *Cellvibrio japanicus* was retrieved in the integrated microbial genome (IMG) interface at the JGI, with its locus tag (CJA_3124), and its homologs in the 12 complete genomes of Cellvibrionales and Oceanospirillales were then added in the gene list with the built-in Blast function. The "show neighborhood" tool was then used to display all 13 genetic contexts. The *dopE* gene and its neighboring *dopC* are shown in red. The position of the putative *dopC* gene in *Congregibacter litoralis* KT71 is shown in light pink as a careful inspection of the genome of this bacteria revealed the presence of a 253 residues long orf containing the characteristics of other O-like proteins, apparently missed during the annotation of this genome. In the display, genes receive a color when they belong to an identified Cluster of Orthologous Gene (COG). The result was reordered, to regroup similar contexts. Some genes typical of each class are highlighted.

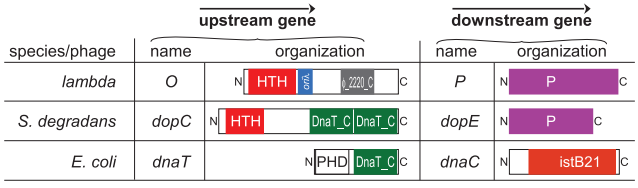

Fig. 3.—Organization, features, and relatedness of the putative bicomponent RHeMa DopCE. HHpred (Soeding et al. 2005) was fed with the protein sequences indicated to identify significant (prediction probability > 95%) protein domains (Pfam domain) or structural motifs (PDB accession) carried by the proteins. HTH: Helix-turn-helix_36 domain (PF13730), P: Replication protein P (PF06992), PHD = PhdYeFM_antitox (PF02604), Φ_2220_C: Conserved phage C-terminus (PF09524), DnaT_C: Primosomal protein DnaT Cter motif (PDB: 2ru8_A), istB21: IstB-like ATP binding protein (PF01695), *oriλ*: lambda origin of replication[11]. *Sd*: *Saccharophagus degradans* 2-40, *Ec*: *Escherichia coli* K12. Nter (N) and Cter (C) extremities of the proteins, as well as transcription directions (arrows) are given.

curve, which likely indicates the origin of replication, is located elsewhere than in the *dopC* locus in all 13 genomes analyzed (fig. 4). Instead, the lower inversion point is located within the *gidA-dnaA*, suggesting that replication initiates within this region. In support for this hypothesis, we noticed the presence of clusters of potential DnaA box sequences, whose organization is consistent with that expected for bacterial origins of replication, upstream of *gidA* (data not shown). Altogether, these data strongly suggest that *dopC* does not specify anymore a replication initiation protein. Yet, further molecular analyses will be required to identify precisely the bacterial origin of replication in these organisms.

Here, we identified a putative cellular RHeMa system, potentially enlightening another example of phage proteins diverted from their viral function to fulfill a cellular one. The experimental validation of DopE as a third RHeMa system in the last clade in which neither DnaC/I nor DciA was identified (Brezellec et al. 2016) would imply that RHeMa are needed components of the replication program in bacteria, like in archaea and eukaryotes.In turn, it would enlarge the variety of mechanisms operating bacterial replicative helicases during replication. The last conundrum regarding bacterial RHeMas is found in Thiotrichales. This order of gammaproteobacteria encompasses mostly *dciA*-containing species. Some species appear, however, to be devoid of *dciA*, *dnaC/I* or *dopE* gene. Intriguingly, RHeMa-lacking species are not regrouped in one phylogenetic clade, but scattered within the order (data not shown).

The close relatedness between P and DopE sequences suggests that the domesticated protein could perform like its viral ancestor during bacterial replication initiation. According to this hypothesis, DopE would bring free replicative helicases to the DnaA-built initiation complex formed at *oriC* rather than to the phage origin of replication. The fact that DopC does not serve as an origin of replication anymore gives weight to this hypothesis. Nonetheless, *dopC* is always maintained in *dopE*- containing genomes and genomically associated with

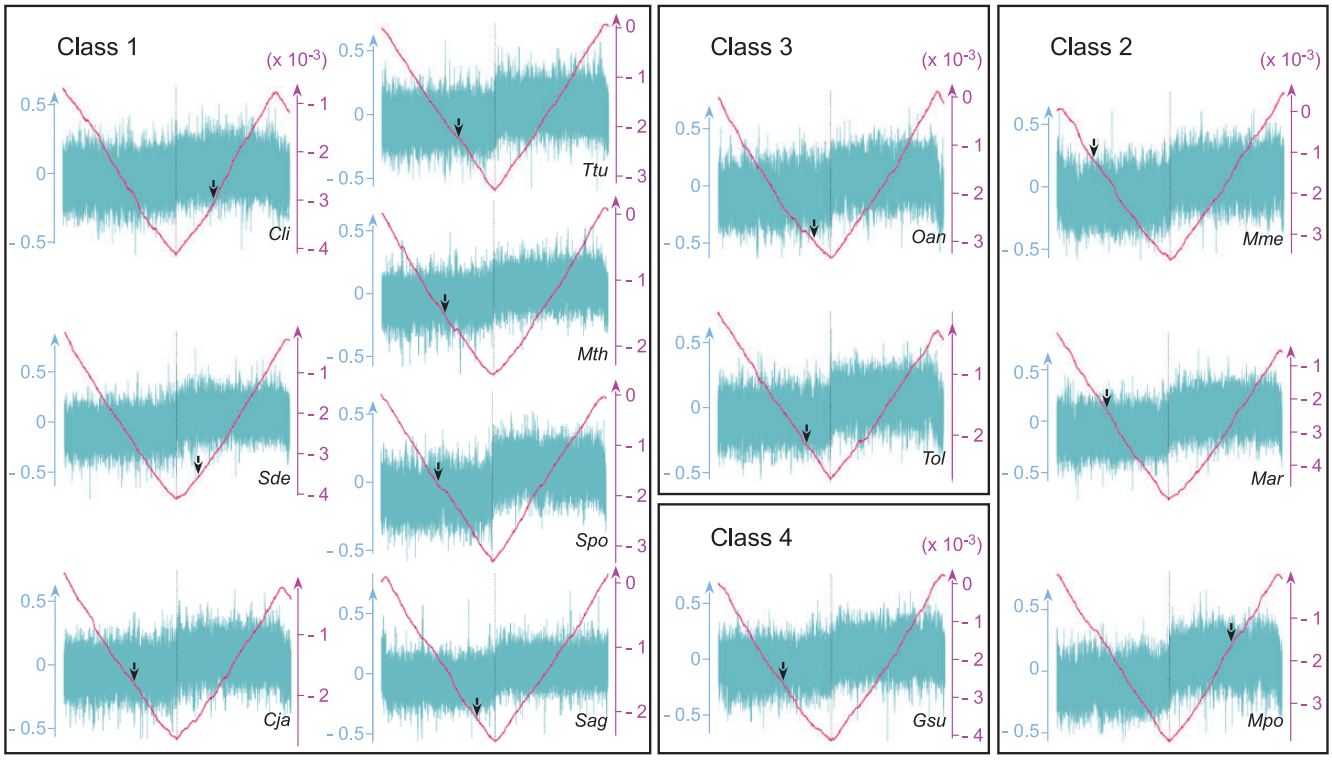

Fig. 4.—Replication is initiated within the *gidA-dnaA* region in *dopCE*-specifying organisms and not at *dopC*. GC (cyan) skew and cumulative GC skew (fushia) profiles of the Cellvibrionales and *dopCE*-containing Oceanospirillales genomes are represented (Materials and Methods). The analysis was performed on the four classes of genomes context. The position of *dopCE* in each genome is indicated with a black arrow, whereas that of *gidA-dnaA* (dashed line) is located at the minimum of the cumulative curve. *Cli*: *Congregibacter litoralis* KT71, *Sde*: *Saccharophagus degradans* 2-40, *Cja*: *Cellvibrio japonicus* Ueda107, *Ttu*: *Teredinibacter turnerae* T7901, *Mth*: *Microbulbifer thermotolerans*, *Spo*: *Spongiibacter* sp. IMCC21906, *Sag*: *Simiduia agarivorans* SA1, *Oan*: *Oleispira antarctica* RB-8, *Tol*: *Thalassolituus oleivorans* MIL-1, *Gsu*: *Gynuella sunshinyii* YC6258, *Mme*: *Marinomonas mediterranea* MMB-1, *Mar*: Marinomonas sp. MWYL1, *Mpo*: *Marinomonas posidonica* IVIA-Po-181, *Pae*: *Pseudomonas aeruginosa* PAO1, *Vch*: *Vibrio cholerae* N16961, *Eco*: *E. coli*.

dopE, suggesting that both genes specify synergistic functions. The substitution of *λO* phage features in DopC by a protein domain found in primosomal proteins, may link the function of the domesticated gene with replication forks reactivation. Alternatively, DopC might help recruit the DopE-bound replicative helicase during chromosomal replication initiation. Experimental investigation is now required to answer these questions.

The easiness with which a cellular machinery that manages the recruitment and the positioning of the replicative helicase on the replication initiation complex may be replaced by another one is remarkable. *dciA* was replaced in the bacterial domain multiple times by viral proteins, at least seven times by *dnaCI* (Brezellec et al. 2016) and potentially once or more by *dopE*. This suggests limited adjustments and likely requires the predomesticated gene to belong to a phage that employs the host replicative helicase for its own replication. As a matter of fact, we have not detected traces of domestication of the bacteriophage T4 RHeMa gene 59 in bacteria, likely because T4 specifies its own replicative helicase, unrelated to the bacterial ones. Yet, the wide distribution of domesticated RHeMas in the bacterial kingdom may reflect an additional selective advantage of the viral RHeMa over DciA. It is noteworthy, indeed, that RHeMa substitutions identified to date are unilateral and characterized by the replacement of the ancestral *dciA* by a domesticated phage gene. The mechanism through which P fulfills its function illustrates marvelously the selective advantage that can be associated with the domestication of viral RHeMas. In *Escherichia coli*, P titrates out the bacterial replicative helicase already complexed with DnaC for the benefit of the phage replication and to the expense of that of the genome (Mallory et al. 1990), inferring that the strength of the interaction between competing RHeMas and the replicative helicase governs over the choice of DNA that will be replicated. It is therefore conceivable that the domestication of a viral RHeMa gene brought a selective advantage to cells within which it was domesticated, by jeopardizing the replication and subsequently the proliferation of certain categories of phages.

Materials and Methods

Phylogenetic Analyses

An alignment of proteins (a crude alignment generated using the program ClustalW and refined by hand) was fueled into PhyLM (v. 3.0) and 100 bootstrap replicates were generated for each analysis (Guindon and Gascuel 2003). A consensus tree was eventually obtained by running the program CONSENSE and fed as an input tree into PhyML. Significant bootstrap scores (arbitrarily above 70%) are indicated.

Datasets

The following strains were used to identify *dopE* in RHeMa-lacking species through "Gene phyloprofile" at the Microscope platform. IN-group: *Cellvibrio japonicus* Ueda107, *Congregibacter litoralis* KT71, *Microbulbifer thermotolerans*, *Saccharophagus degradans* 2-40, *Simiduia agarivorans* SA1, *Spongiibacter* sp. IMCC21906, *Teredinibacter turnerae* T7901, *Gynuella sunshinyii* YC6258, *Thalassolituus oleivorans* MIL-1, *Oleispira antarctica* RB-8, *Marinomonas mediterranea* MMB-1, *Marinomonas posidonica* IVIA-Po-181, and *Marinomonas* sp. MWYL1. OUT-group: *Azotobacter vinelandii* DJ, *Chromohalobacter salexigens* DSM 3043, *Hahella chejuensis* KCTC 2396, *Halomonas elongata* DSM 2581, and *Pseudomonas aeruginosa* PAO1.

Gene Phyloprofile

"Gene phyloprofile" searches for genes systematically present in each species of the IN-group set, excluding genes sharing homology with a gene found in at least one genome of a species constituting the OUT-group set. "Gene phyloprofile" was run with the following parameters: Homology constraints within the IN-group set (minLrap $\geq$ 0.9, maxLrap $\geq$ 0, Identity $\geq$ 35%) and in the OUT-group set (minLrap $\geq$ 0.7; maxLrap $\geq$ 0; Identity $\geq$ 30%)

GC Skew

GC-skews have been computed as (G-C)/(G + C) on the whole genomic sequence. We used a window size of 200 nucleotides and a sliding step of 50 nucleotides. Both GC skew and cumulative GC skew have been computed and drawn using a script available on request (supplementary information, Supplementary Material online). Due to compositional strand bias between the lagging and the leading strand, the plot of the cumulative GC-skew typically produces a V-curve. The minimum value of this V-curve is commonly interpreted as the putative origin of replication (Sernova and Gelfand 2008).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We are thankful to B. Michel and lab members for critical discussions and helpful comments. This work was supported by the French National Research Council (CNRS) and the National Institute for Agricultural Research (INRA).

Literature Cited

Alfano C, McMacken R. 1989. Ordered assembly of nucleoprotein structures at the bacteriophage lambda replication origin during the initiation of DNA replication. J Biol Chem. 264:10699–10708.

Arndt D, et al. 2016. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 44:W16–W21.

Bell SP, Kaguni JM. 2013. Helicase loading at chromosomal origins of replication. Cold Spring Harb Perspect Biol 5:pii: a010124.

Brezellec P, Vallet-Gely I, Possoz C, Quevillon-Cheruel S, Ferat JL. 2016. DciA is an ancestral replicative helicase operator essential for bacterial replication initiation. Nat Commun. 7:13271.

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.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Kornberg A, Baker TA. 1992. DNA replication. New York: Freeman.

Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. 2008. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. Bioinformatics 24:863–865.

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.

Markowitz VM, Chen IM, Palaniappan K, et al. 2014. IMG 4 version of the integrated microbial genomes comparative analysis system. Nucleic Acids Res. 42:D560–D567.

Samson RY, Abeyrathne PD, Bell SD. 2016. Mechanism of archaeal MCM helicase recruitment to DNA replication origins. Mol Cell 61:287–296.

Sernova NV, Gelfand MS. 2008. Identification of replication origins in prokaryotic genomes. Brief Bioinform. 9:376–391.

Soeding J, Biegert A, Lupas AN. 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 33:W244–W248.

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.

Zahn K, Blattner FR. 1985. Binding and bending of the lambda replication origin by the phage O protein. EMBO J. 4:3605–3616.

Associate editor: Purificación López-García