Four additional natural 7-deazaguanine derivatives in phages and how to make them

Cui, Liang; Balamkundu, Seetharamsing; Liu, Chuan Fa; Ye, Hong; Hourihan, Jacob; Rausch, Astrid; Hauß, Christopher; Nilsson, Emelie; Hoetzinger, Matthias; Holmfeldt, Karin; Zhang, Weijia; Martinez-Alvarez, Laura; Peng, Xu; Tremblay, Denise; Moinau, Sylvain; Solonenko, Natalie; Sullivan, Matthew B.; Lee, Yan Jiun; Mulholland, Andrew; Weigele, Peter R.; de Crécy-Lagard, Valérie; Dedon, Peter C.; Hutinet, Geoffrey

Published in:
Nucleic Acids Research

DOI:
10.1093/nar/gkad657

Publication date:
2023

Document version
Publisher's PDF; also known as Version of record

Document license:
CC BY-NC

Citation for published version (APA):
Four additional natural 7-deazaguanine derivatives in phages and how to make them

Liang Cui¹,‡, Seetharamsing Balamkundu¹,², Chuan-Fa Liu², Hong Ye², Jacob Hourihan³, Astrid Rausch³, Christopher Haupt³, Emelie Nilsson⁴, Matthias Hoetzinger⁴, Karin Holmfeldt⁴, Weijia Zhang⁵, Laura Martinez-Alvarez⁶,⁷, Xu Peng⁸, Denise Tremblay⁶,⁷, Sylvain Moinau⁰,⁶,⁷, Natalie Solonenko⁹, Matthew B. Sullivan⁹,¹⁰, Yan-Jiun Lee¹¹, Andrew Mulholland¹¹, Peter R. Weigele¹¹,¹², Valérie de Crécy-Lagardœ⁵,¹², Peter C. Dedon¹,¹³ and Geoffrey Hutinetœ³,⁴

¹Singapore-MIT Alliance for Research and Technology, Antimicrobial Resistance Interdisciplinary Research Group, Campus for Research Excellence and Technological Enterprise, Singapore 138602, Singapore, ²School of Biological Sciences, Nanyang Technological University, Singapore, ³Department of Microbiology and Cell Science, University of Florida, Gainesville, FL 32611, USA, ⁴Centre for Ecology and Evolution in Microbial Model Systems (EEMiS), Linnaeus University, 391 82 Kalmar, Sweden, ⁵Department of Biology, University of Copenhagen, Copenhagen N, Denmark., ⁶Département de biochimie, de microbiologie et de bio-informatique, Faculté des sciences et de génie, Université Laval, Québec City, Québec, Canada, ⁷Groupe de recherche en écologie buccale, Faculté de médecine dentaire, Université Laval, Québec City, Québec, Canada, ⁸Félix d’Hérelle Reference Center for Bacterial Viruses, Université Laval, Québec City, Québec, Canada, ⁹Department of Microbiology, Ohio State University, Columbus, OH, USA, ¹⁰Department of Civil, Environmental, and Geodetic Engineering, and Center of Microbiome Science, Ohio State University, Columbus, OH, USA, ¹¹Research Department, New England Biolabs, Ipswich, MA 01938, USA, ¹²University of Florida, Genetics Institute, Gainesville, FL 32610, USA and ¹³Department of Biological Engineering and Center for Environmental Health Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received April 14, 2023; Revised July 20, 2023; Editorial Decision July 21, 2023; Accepted August 06, 2023

ABSTRACT

Bacteriophages and bacteria are engaged in a constant arms race, continually evolving new molecular tools to survive one another. To protect their genomic DNA from restriction enzymes, the most common bacterial defence systems, double-stranded DNA phages have evolved complex modifications that affect all four bases. This study focuses on modifications at position 7 of guanines. Eight derivatives of 7-deazaguanines were identified, including four previously unknown ones: 2′-deoxy-7-(methylamino)methyl-7-deazaguanine (mdPreQ₁), 2′-deoxy-7-(formylaminomethyl)-7-deazaguanine (fdPreQ₁), 2′-deoxy-7-deazaguanine (dDG) and 2′-deoxy-7-carboxy-7-deazaguanine (dCDG). These modifications are inserted in DNA by a guanine transglycosylase named DpdA. Three subfamilies of DpdA had been previously characterized: bDpdA, DpdA₁, and DpdA₂. Two additional subfamilies were identified in this work: DpdA₃, which allows for complete replacement of the guanines, and DpdA₄, which is specific to archaeal viruses. Transglycosylases have now been identified in all phages and viruses carrying 7-deazaguanine modifications, indicating that the insertion of these modifications is a post-replication event. Three enzymes were predicted to be involved in the biosynthesis of these newly identified DNA modifications: 7-carboxy-7-deazaguanine decarboxylase (DpdL), dPreQ₁ formyltransferase (DpdN) and dPreQ₁ methyltransferase (DpdM), which was experimentally validated and harbors a unique fold not previously observed for nucleic acid methylases.

⁷To whom correspondence should be addressed. Tel: +1 352 846 0959; Email: ghutinet@ufl.edu
‡The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Present addresses:
Astrid Rausch, Hirsch-Apotheke Wöllstein, Alzeyer Str. 5, 55597 Wöllstein, Germany.
Christopher Haupt, Institute of Pharmaceutical Technology and Buchmann Institute for Molecular Life Sciences, Max-von-Laue-Str. 9, 60438 Frankfurt am Main, Germany.
Matthias Hoetzinger, Faculty of Biosciences, Norwegian University of Life Sciences, 1433 Ås, Norway.

© The Author(s) 2023. Published by Oxford University Press on behalf of Nucleic Acids Research. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
binding site and the catalytic residues on the same domain. Members of the bimodular subgroup contain two repeating domains: the N-terminal domain with the NADPH binding site, and the C-terminal domain with the catalytic residues (19). PreQ₁ is inserted in tRNA by the bacterial tRNA transglycosylase bTGT (20) and further modified in two steps to produce Q (3). PreQ₀ is directly inserted in tRNA in archaea by arcTGT, where it is further modified into G′. The distant TGT paralog, ArcS (21), as well as Gat-QueC, a fusion protein of QueC and a glutamine amidotransferase (22), and QueF-L, a paralog of the unimodular QueF that lacks the NADPH-dependent reduction activity (22,23), have been found as interchangeable proteins for this reaction.

G′ was the first 7-deazaguanine derivative found in phages, replacing 25% of the Gs in the dsDNA genome of Enterobacteria phage 9g. The enzymes FolE, QueD, QueE and Gat-QueC are all encoded by this phage, as is DpdA1 (4), which inserts preQ₀ into DNA (7). A related phage, Escherichia phage CAjan, that encodes QueC rather than Gat-QueC, was found to replace 32% of its Gs with preQ₀ (7). Furthermore, two other 7-deazaguanine derivatives were discovered in phage genomes: 7-amido-7-deazaguanine (ADG), that modifies Campylobacter phage CP220 DNA at 100% (6), and preQ₁ that modifies 30% of the guanines in Halovirus HVTV1, a virus that encodes a QueF (7). No DpdA was previously detected in these two last viruses. In bacteria, bDpdA, in complex with DpbB, inserts preQ₀ into DNA, that is further modified into ADG by DpcD (5).

In our effort to expand the set of phages that contain 7-deazaguanine derivatives in their genome, we identified four unique 7-deazaguanine derivatives not previously observed in DNA: 7-deazaguanine (DG), 7-(methylamino)-methyl-7-deazaguanine (mpreQ₁), 7-(formylamino)methyl-7-deazaguanine (preQ₁) and 7-carboxy-7-deazaguanine (CDG). We predicted and validated a preQ₁ methyltransferase enzyme and predicted the involvement of five additional proteins in the synthesis of these modifications, including two additional subfamilies of DpdA, DpdA3 and DpdA4, and three enzymes with unprecedented chemistry.

**INTRODUCTION**

Because of their intrinsic properties, such as resistance to nucleases (1), or fluorescence quenching (2), 7-deazaguanine derivatives have long been employed in synthetic biology. Two of these derivatives are tRNA modifications, queuine (Q) and archaeosine (G′). They are respectively involved in the avoidance of translational errors and in tRNA stabilization (3). Recently, 7-deazaguanine derivatives have been found in DNA as components of restriction/modification systems in bacteria (4,5), and anti-restriction systems in phages (4,6,7). Epigenetic modifications are common among phages (8–11) to resist to various bacterial defense systems (11–16).

Members of a transglycosylase superfamily are responsible for the incorporation of 7-deazaguanine derivatives into both tRNA and DNA. Proteins of the Tgt subgroup modify tRNA, while DpdA subgroup proteins modify DNA (3), both by replacing the target guanine with a specific 7-deazaguanine derivative. Tgt enzyme 7-deazaguanine substrates differ between organisms. One of these substrates is queuine (q), which is inserted at position 34 of the GUN anticodon in tRNAs in eukaryotes and in certain bacteria (3). 7-aminoethyl-7-deazaguanine (preQ₁) is inserted at the same position in most bacteria. 7-cyano-7-deazaguanine (preQ₀) is inserted at position 15 or 16 of many tRNAs in archaea (3). Similarly, all DpdA enzymes tested thus far insert preQ₀ in DNA (4,5,7), and sequence specificity has been identified for one of them (17). DpaA homologs are divided in three groups: bacterial DpaA (bDpaA), and two phage DpaA (DpaA1 and DpaA2) (7). Of note, DpaA homologs have not been identified in some of the phages that contain modified 7-deazaguanine derivatives (6,7).

PreQ₀, the key intermediate in all experimentally validated pathways is synthesized from guanosine triphosphate (GTP) by a pathway involving four proteins (FolE, QueD, QueE and QueC, see Figure 1A) found in archaea, bacteria, and some phages (3,7). The pathways then diverge, producing various modifications. PreQ₀ is reduced by QueF into preQ₁ in bacteria through a NADPH dependent reaction (18). QueF proteins can be categorized into two subgroups. Members of the unimodular subgroup harbor the NADPH binding site and the catalytic residues on the same domain. Members of the bimodular subgroup contain two repeating domains: the N-terminal domain with the NADPH binding site, and the C-terminal domain with the catalytic residues (19). PreQ₁ is inserted in tRNA by the bacterial tRNA transglycosylase bTGT (20) and further modified in two steps to produce Q (3). PreQ₀ is directly inserted in tRNA in archaea by arcTGT, where it is further modified into G′. The distant TGT paralog, ArcS (21), as well as Gat-QueC, a fusion protein of QueC and a glutamine amidotransferase (22), and QueF-L, a paralog of the unimodular QueF that lacks the NADPH-dependent reduction activity (22,23), have been found as interchangeable proteins for this reaction.

G′ was the first 7-deazaguanine derivative found in phages, replacing 25% of the Gs in the dsDNA genome of Enterobacteria phage 9g. The enzymes FolE, QueD, QueE and Gat-QueC are all encoded by this phage, as is DpdA1 (4), which inserts preQ₀ into DNA (7). A related phage, Escherichia phage CAjan, that encodes QueC rather than Gat-QueC, was found to replace 32% of its Gs with preQ₀ (7). Furthermore, two other 7-deazaguanine derivatives were discovered in phage genomes: 7-amido-7-deazaguanine (ADG), that modifies Campylobacter phage CP220 DNA at 100% (6), and preQ₁ that modifies 30% of the guanines in Halovirus HVTV1, a virus that encodes a QueF (7). No DpdA was previously detected in these two last viruses. In bacteria, bDpdA, in complex with DpbB, inserts preQ₀ into DNA, that is further modified into ADG by DpcD (5).

In our effort to expand the set of phages that contain 7-deazaguanine derivatives in their genome, we identified four unique 7-deazaguanine derivatives not previously observed in DNA: 7-deazaguanine (DG), 7-(methylamino)methyl-7-deazaguanine (mpreQ₁), 7-(formylamino)methyl-7-deazaguanine (preQ₁) and 7-carboxy-7-deazaguanine (CDG). We predicted and validated a preQ₁ methyltransferase enzyme and predicted the involvement of five additional proteins in the synthesis of these modifications, including two additional subfamilies of DpdA, DpdA3 and DpdA4, and three enzymes with unprecedented chemistry.

**MATERIALS AND METHODS**

**Strains and plasmids**

All strains and plasmids used in this study are referenced in Supplementary Tables S1 and S2, respectively.

**Mass spectrometry analysis**

DNA analysis followed our previous publication (7). Purified DNA (10 μg) was hydrolyzed in 10 mM Tris–HCl (pH 7.9) with 1 mM MgCl₂ with Benzonase (20U); DNase I (4 U), calf intestine phosphatase (17 U) and phosphodiesterase (0.2 U) for 16 h at ambient temperature. Following passage through a 10 kDa filter to remove proteins, the filtrate was analyzed by liquid chromatography–coupled triple quadrupole mass spectrometry (LC–MS/MS).

Quantification of the modified 2′-deoxynucleosides (dADG, dQ, dPreQ₀, dPreQ₁, mdPreQ₁, dG′, dCDG and mdA) and the four canonical deoxyribonucleosides (dA,
**Figure 1.** Proteins involved in the 7-deazaguanine DNA modifications pathway. (A) Proposed pathway for the biosynthesis of all eight 7-deazaguanine derivatives in DNA and (B) detection of the gene encoding for these proteins in phage genomes (full table in sup. table). Numbering in the table correspond to the gene numbering in each phage. Coloring match with the proteins. In shades of green are the DpdA1 to 4 (7-deazaguanine derivative DNA transglycosidase). In shade of blues, the enzyme leading to preQ0: FolE (GTP cyclohydrolase I, EC 3.5.4.16), QueD (CPH1 synthase, EC 4.1.2.50), QueE (CDG synthase, EC 4.3.99.3), QueC (preQ0 synthase, EC 6.3.4.20), QueC* (proposed ADG synthase, homologue of QueC). In shades of orange, protein modifying preQ0 further: QueF (NADPH-dependent 7-cyano-7-deazaguanine reductase, EC 1.7.1.13), ArcS2 (core domain of the archaeosine synthase, PF5591), Gat-QueC (glutamine amidotransferase class-II domain fused to QueC). In shades of purple, protein discovered and described in this study: DpdL (proposed CDG decarboxylase), DpdM (proposed preQ, or dPreQ methylase), DpdN (proposed preQ, or dPreQ formylase). Question marks are proposed reactions that have not been proven. Dashed arrows are previously unpublished reactions. Molecule abbreviations: guanosine tri-phosphate (GTP), dihydronopterin triphosphate (H2NTP), 6-carboxy-5,6,7,8-tetrahydropterin (CPH4), 7-carboxy-7-deazaguanine (CDG), 7-amido-7-deazaguanine (ADG), 7-deazaguanine (DG), 7-cyano-7-deazaguanine (preQ0), 7-aminomethyl-7-deazaguanine (dCDG), 2′-deoxy-7-amido-7-deazaguanine (dADG), 2′-deoxy-7-deazaguanine (dDG), 2′-deoxy-7-cyano-7-deazaguanine (dPreQ0), 2′-deoxy-7-aminoethyl-7-deazaguanine (dPreQ), 2′-deoxy-7-(aminomethyl)methyl-7-deazaguanine (mdPreQ1), 2′-deoxy-7-(aminoformyl)methyl-7-deazaguanine (fdPreQ1). Molecules 1 through 4 (circled numbers) were discovered in this study.
dT, dG and dC) was achieved by liquid chromatography-coupled triple quadrupole mass spectrometry (LC-MS/MS) and liquid chromatography-coupled diode array detector (LC-DAD), respectively. Aliquots of hydrolysed DNA were injected onto a Phenomenex Luna Omega Polar C18 column (2.1 × 100 mm, 1.6 µm particle size) equilibrated with 98% solvent A (0.1% v/v formic acid in water) and 2% solvent B (0.1% v/v formic acid in acetonitrile) at a flow rate of 0.25 ml/min and eluted with the following solvent gradient: 2–12% B in 10 min; 12–25% B in 1 min; hold at 25% B for 5 min. The HPLC column was coupled to an Agilent 1290 Infinity DAD and an Agilent 6490 triple quadrupole mass spectrometer (Agilent, Santa Clara, CA). The column was kept at 40°C and the auto-sampler was cooled at 4°C. The UV wavelength of the DAD was set at 260 nm and the electrospray ionization of the mass spectrometer was performed in positive ion mode with the following source parameters: drying gas temperature 200°C with a flow of 14 l/min, nebulizer gas pressure 30 psi, sheath gas temperature 400°C with a flow of 11 l/min, capillary voltage 3000 V and nozzle voltage 500 V. Compounds were quantified in multiple reaction monitoring (MRM) mode with the following transitions: m/z 310.1→194.1, 310.1→177.1, 310.1→293.1 for dADG; m/z 394.1→163.1, 394.1→146.1, 394.1→121.1 for dG; m/z 292.1→176.1, 176.1→159.1, 176.1→52.1 for dPreQ; m/z 296.1→163.1, 296.1→121.1, 296.1→279.1 for dPreQ₂; m/z 310.1→163.1, 310.1→121.1 for mdPreQ; m/z 309.1→193.1, 309.1→176.1, 309.1→159.1 for dG⁺; m/z 311.1→177.1, 311.1→78.9 for dCGD and 266.1→150.1, 266.1→108.1, 266.1→55.1 for m²dA. External calibration curves were used for the quantification of the modified 2'-deoxynucleosides and the four canonical deoxyribonucleosides. The linear range for each compound was evaluated using a serial dilution of standard solution, and the solution of each concentration was measured two times. The calibration curves were constructed based on peak area of the chromatogram versus the corresponding concentration of each standard. A linear regression with r² > 0.99 was obtained in the entire range of concentrations (Supplementary Figure S1). The quantification of modified 2'-deoxynucleosides in phage DNA samples was based on the relationship between peak area and concentration deduced from the calibration curve. The concentration of each modification was divided by summed concentration of the four canonical deoxyribonucleosides and then multiplied by 10⁶ to get modification per 10⁶ nucleotides. The limit of detection (LOD), defined by a signal-to-noise ratio (S/N) of 3, ranged from 0.1 to 1 fmol for the modified 2'-deoxynucleosides. Data acquisition and processing were performed using MassHunter software (Agilent, Santa Clara, CA).

Unknown DNA modification analysis was performed using Agilent 1290 ultrahigh pressure liquid chromatography system equipped with DAD and 6550 QTOF mass detector managed by a MassHunter workstation. The column used for the separation was a Waters ACQUITY HSS T3 column (2.1’×150 mm, 1.8 µm). The oven temperature was set at 45°C. The gradient elution involved a mobile phase consisting of (A) 0.1% formic acid in water and (B) 0.1% formic acid in acetonitrile. The initial condition was set at 2% B. A 25 min linear gradient to 7% B was applied, followed by a 15 min gradient to 100% B which was held for 5 min, then returned to starting conditions over 0.1 min. Flow rate was set at 0.3 ml/min, and 2 µl of samples was injected. The electrospray ionization mass spectra were acquired in positive ion mode. Mass data were collected between m/z 100 and 1000 Da at a rate of two scans per second. The electrospray ionization of the mass spectrometer was performed in positive ion mode with the following source parameters: drying gas temperature 250°C with a flow of 14 l/min, nebulizer gas pressure 40 psi, sheath gas temperature 350°C with a flow of 11 l/min, capillary voltage 3500 V and nozzle voltage 500 V. Two reference masses were continuously infused to the system to allow constant mass correction during the run: m/z 121.0509 (C₂H₃N₄) and m/z 922.0098 (C₁₈H₁₆O₈N₅P₂F₂₄). Raw spectrometric data were analyzed by MassHunter Qualitative Analysis software (Agilent Technologies, US).

Protein sequence detection in phages

HHpred online tool (https://toolkit.tuebingen.mpg.de/tools/hhpred) (24,25) was used with default setting against the pfam database (Pfam-A.x35) (26) to investigate the deduced proteins encoded by genes flanking the 7-deazaadenosine modification genes in *Cellulophaga* phage phiSM, *Cellulophaga* phage phiST, and Halovirus HVT-1. DpdL, DpdM, DpdA3 and DpdA4 were predicted this way. DpdN was discovered by looking at the annotations of the genes in the vicinity of the 7-deazaadenosine modification genes in *Flavobacterium* phage vb_FspM_immuto-2-6A. These proteins were then used as queries to retrieve homologs in the proteome of viruses publicly available in NCBI GenBank database (July 2022) using psiBLAST version 2.13.0 (27), with at most three iterations. Other previously discovered proteins involved in the 7-deazaadenosine derivative DNA modifications (Supplementary Data S1) (7) were used to identify homologs in viral genomes encoding for at least one of DpdL, DpdM, DpdN, DpdA3 or DpdA4 using BLASTp version 2.13.0 (28). HHpred and expert annotation were used to sort these proteins and curate false positives. All protein matches are summarized in Supplementary Data S1.

Alignments, trees and structures

Protein sequences were collected from the NCBI database, using the protein id collected from the detection. Multiple sequence alignments were generated using MAFFT (29) online server (version 7, https://mafft.cbrc.jp/alignment/server/), with default settings and then visualized using Jalview version 2.11.2.4. Clustering trees were generated using Graph Splitting (30) online server (version 2.0, http://gs.bbs.u-tokyo.ac.jp/), with default settings. Protein structures were predicted using the multimeric collab notebook of AlphaFold2 (version 2.2.4 (31), https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb). Protein structures were visualized using ChimeraX version: 1.5rc202210241843 (32), and already published protein structure were imported from PDB (https://www.rcsb.org,
Autodock Vina was used to predict the docking of chemicals in enzymes.

RESULTS

7-(Methylamino)-methyl-7-deazaguanine in cellulophaga phage phiSM DNA

Cellulophaga phage phiSM resisted to a wide variety of restriction enzymes containing at least a guanine in their specific recognition site (Supplementary Figure S2A). In addition, phiSM encodes a complete set of dPreQ1 synthesis genes, including DpdA, FolE, QueD, QueE, QueC and QueF (Figure 1, Supplementary Data S1) and thus should harbor preQ1 in its genome, as previously observed for Halovirus HVTV-1 (7). To test this hypothesis, we used liquid chromatography coupled to diode array UV detection and a tandem mass spectrometer (LC–UV–MS/MS) to analyze the nucleosides obtained from enzymatic digestion of phiSM genomic DNA, as we previously described (4,7). A 2′-deoxynucleoside form of preQ1 (dPreQ1) was indeed detected (3790 modifications per 10^6 nucleotides, ~1.1% of the Gs, Table 1). In addition to the UV peaks for the four canonical nucleosides, dA, dC, dT and dG, an unknown UV peak with a mass of 310 Da was observed at a retention time of 6.5 min (Figure 2A, 212 modifications per 10^6 nucleotides, ~0.1% of the Gs, Table 1). The collision-induced dissociation (CID) MS/MS spectra of the unknown peak revealed the protonated 2′-deoxyribose ion ([M + H]^+ at m/z 117) and its further dehydration ions (m/z 99 and 81), confirming that the unknown peak corresponded to a non-canonical nucleoside (Figure 2B). The CID MS/MS spectra of dPreQ1 and the unknown modification showed very similar patterns. Both compounds showed fragment [M + H]^+ ions at m/z 163 and m/z 279, indicating that the unknown modification could be a derivative of dPreQ1 (Figure 2B). The mass of the unknown modification is 14 Da larger than that of dPreQ1, implying that it is a methylated product of dPreQ1. The high-resolution mass spectrometry (HRMS) of the protonated [M + H]^+ ion (m/z 310.1511) of the unknown modification matches the theoretical mass of protonated methylated-dPreQ1 very well (m/z 310.1515, mass error = 1.29 ppm). The MS/MS spectra of the unknown modification revealed a fragment ion with a loss of 31 Da (m/z 310 → m/z 279), corresponding to a methylamino group (Figure 2B). The loss of the methylamino group was observed in the MS/MS spectra at low CID energy, indicating that the methyl group is likely linked to the 7-amino group, which is less stable than the linkage to the 2-amino group in CID MS/MS experiment.

We chemically synthesized 7-(methylamino)methyl-2′deoxy-7-deazaguanine (mdPreQ1, Supplementary Scheme S1), which was purified by HPLC and characterized by NMR and HRMS, to test whether methylation at the 7-amino position of dPreQ1 corresponds to the unknown DNA modification. The standard was then analyzed using LC–UV–MS/MS. The retention time and MS/MS spectra of mdPreQ1 standard were identical to those of the unknown non-canonical nucleoside (Supplementary Figure S3), confirming that the unknown modification is mdPreQ1. The same modification was identified in Cellulophaga phages phi38:2 and phi47:1 (Supplementary Figure S4), which are related to phage phiSM (34).

Phage and bacterial QueF have similar function

Given the presence of dPreQ1 in DNA (7) and its subsequent modification to mdPreQ1, one must ask which precursor molecule (preQ0 or preQ1) is directly inserted in the genome. Indeed, to date, all characterized DpdAs insert preQ1 into DNA (5,7). Hence, the phage QueF should behave like the Archaeal QueF-L (22) and generate preQ1 from preQ0 inserted into DNA. However, if phage QueF proteins are similar to the bacterial QueF (35) and form preQ1 base from preQ0 base, then the DpdA of these phages should have changed in substrate specificity to insert preQ1 in DNA. QueF family sequences were collected from phages (Supplementary Data S1 and (7)) and compared to the sequences of three experimentally validated QueF proteins: the bimodular QueF of Escherichia coli (NP_417724), the unimodular QueF of Bacillus subtilis (NP_389258) and the QueF-L of Pyrococcus caldofontis (WP_011848915). Surprisingly, no phage QueF sequences aligned with QueF-L (Alignment S1). The bimodular sequence aligned with half of the phage QueF sequences (Supplementary Figure S5A and B), while the unimodular one aligned with the other half (Supplementary Figure S5B). The NADPH binding motif, E(S/L)K(S/A)hK(l/Y)(Y/F/W), and most of the amino acids characteristic of the QueF family sequences (Supplementary Figure S5B, starred conserved residues (19,36)) were conserved in all phage sequences with the exception of a tyrosine (Y221 in E. coli, Y87 in B. subtilis, Y52 in P. caldofontis) in the unimodular phage sequences. This degree of conservation strongly suggests that the phage QueF proteins are NADPH-dependent preQ0 reductases.

To validate this prediction, an E. coli ΔqueF mutant was transformed with plasmids expressing queF genes from three phages/viruses, namely Cellulophaga phage phiSM, Vibrio phage VH7D, and Halovirus HVTV-1. We observed that expression of the phiSM and VH7D queF genes, but not of the HVTV-1 one, complemented the ΔqueF strain’s Q-deficiency phenotype (Supplementary Figure S5C).
Figure 2. HPLC–UV–MS analysis of digested genomic DNA samples from bacteriophage phiSM, phiST and vB_FspM_immuto_2–6A. (A) The HPLC-UV chromatogram on top was obtained from calf thymus DNA to show the retention of the canonical nucleosides. PhiSM shows a fifth peak with a mass of 310 Da. The dG peak disappeared, and a new peak was detected in phiST and vB_FspM_immuto_2–6A with a mass of 267 Da and 324 Da, respectively. (B) The MS/MS spectra and proposed CID fragmentation of dDG ([M + H]^+ = 267.1), dpreQ1 ([M + H]^+ = 296.1), mpreQ1 ([M + H]^+ = 310.1), fdpreQ1 ([M + H]^+ = 324.1). Molecule abbreviations: dC, 2'-deoxycytidine; dA, 2'-deoxyadenosine; dG, 2'-deoxyguanosine; dT, 2'-deoxymydridine.
Because HVT-1 is a virus infecting a hyper-saline archaeon, *Haloarcula valismortis*, expressing its *queF* gene in *E. coli* in a low salt environment may have been challenging. Nonetheless, these experiments confirmed that phage QueF, like its bacterial counterpart, catalyzed the reduction of preQ0 to preQ1.

To confirm that phage DpdA encoded in QueF-like reductase switched specificity to preQ1, we cloned phiSM *dpdA1* and VH7D *dpdA2* in pBAD24 vector and expressed them in several mutants of *E. coli*. In our experiments, phiSM DpdA1 was found to be inactive, while VH7D DpdA2 inserted preQ1 in DNA (2765 modifications per 10^6 nucleotides), proving that this DpdA substrate specificity indeed adapted to preQ1. Interestingly, VH7D DpdA2 also inserted preQ0 at a lower efficiency (712 modifications per 10^6 nucleotides) in a strain that does not produce preQ1 and accumulates preQ0 (*ΔqueF*, see pathway in Figure 1A), as well as CDG at a very low efficiency (67 modifications per 10^6 nucleotides) in a strain that accumulates CDG (*ΔqueC*, see Figure 1A).

### Prediction and validation of a preQ1 methyltransferase

Phages that harbor the *mdPreQ1* modification should encode a methyltransferase that appends a methyl group onto the nitrogen of the methylamino group of preQ1 in genomic DNA. There are four genes coding for proteins of unknown function in the cluster of genes encoding the preQ1 pathway in phage phiSM, namely CEPG_00048, CEPG_00054, CEPG_00056 and CEPG_00057. The proteins CEPG_00048 and CEPG_00057 were ruled out as candidates because they encode short proteins (~60 amino acids) and are not found by psiBLAST in other phages encoding a deazaguanine DNA modification pathway (Supplementary Data S2 and S3). CEPG_00056 homologs were observed in closely related *Cellulophaga* phages and in eukaryotic herpes viruses (Supplementary Data S4). This candidate was eliminated because no deazaguanine DNA modification was ever found in eukaryotic viruses (7) and eukaryotes do not produce any preQ1 (3). Finally, CEPG_00054 homologs were found in seven other phages, including *Vibrio* phages phi-Grn1, phi-ST2, and VH7D, which were predicted to encode preQ1 modification pathways (Figure 1B, Supplementary Data S1 and S5). This protein belongs to the DUF3109 family (Supplementary Data S6) and has an *E. coli* homolog, YkgJ, which is annotated as a zinc or iron binding protein, making CEPG_00054 the leading candidate for the missing preQ1 methyltransferase, and tentatively renaming it DpdM.

We found that the genome of *Vibrio* phages phi-Grn1 and phi-ST2 DNA, encoding DpdM homologs (Figure 1B, Supplementary Data S1), were also modified with mdPreQ1 (Supplementary Figure S6, Table 1) at a rate of 0.01% of the Gs for both phages (35 and 44 modifications per 10^6 nucleotides, respectively, Table 1). Finally, expressing the predicted VH7D *dpdM* gene in an *E. coli* strain already expressing the *dpdA2* of Vibrio phage VH7D resulted in the formation of low levels of mdPreQ1 in plasmid DNA (Table 2). Taken altogether, these data linked mdPreQ1 with the presence of DpdM (Figure 1B, Supplementary Data S1).

As shown above with VH7D *dpdA2* expression alone, when both the VH7D *dpdA2* and *dpdM* genes were expressed in a *ΔqueF* background, which does not produce preQ1 but accumulates preQ0, preQ0 was inserted into bacterial DNA at a ~5-time lower efficiency. Similarly, when a *ΔqueC* background that accumulates CDG was used, CDG was found in DNA with a ~40-fold decrease in efficiency (Table 2).

### DpdM proteins likely bind two metals

Although the initial amino acid sequence analysis of DpdM from *Cellulophaga* phage phiSM revealed a CxxxCxxCC metal binding motif (Supplementary Data S6), this motif was missing in the Vibrio phage phi-ST2 homolog. We found that the *orf* encoding this protein was miscaled and discovered that by selecting a start codon 171 nucleotides prior to the originally predicted one now resulted in a polypeptide containing the CxxxCxxCC motif (phi-ST2 corrected in Alignment S2).

The tertiary and quaternary structures of DpdM from both *Cellulophaga* phage phiSM and *Vibrio* phage VH7D were predicted using AlphaFold2. Both proteins were predicted to be monomeric, with only a few amino acids interacting between monomers (data not shown). The phiSM DpdM prediction (Supplementary Figure S7A) had
a higher confidence score than the VH7D prediction (Supplementary Figure S7B). We found small domains around the VH7D predicted structure with unknown function as shown in the alignment. However, the core parts of the protein were well aligned (Supplementary Figure S7C).

The phiSM DpdM structure contains a tunnel that has an electro-positively charged groove on one side (Figure 3A), which could be a candidate site for DNA binding, and a second groove on the opposite side (Figure 3B), which could be a site for a methyl donor binding. Surprisingly, majority of the conserved residues are clustered around this tunnel (Figures 3C and D). The CxxxCxxCC motif appears to be divided into two metal binding sites rather than one. The CxxxCxxC motif (orange in Figure 3E; representing C33, C38 and C41), is a known motif for a Fe₄S₄ cluster and SAM binding (37), but the presence of a fourth cysteine, C150 (red in Figure 3E), in the pocket would disrupt the Fe₄S₄ binding and may bind another metal instead, as well as another different methyl donor. It appears that the fourth cysteine in the CxxxCxxCC motif is involved in another metal binding pocket containing three other cysteine residues (yellow in Figure 3F; representing C42, C92, C102 and C112). Both these metal binding pockets are found in the DpdM tunnel implying that they both participate in the transfer of the methyl group from the methyl donor to preQ₁ in DNA.

The tunnel observed in phiSM DpdM, the positively charged groove (Figure 3A), and the binding site on the opposite side of the protein (Figure 3B), are absent in the VH7D DpdM structure (Supplementary Figure S7D and E). If the conserved residues in both VH7D and phiSM DpdM proteins are mostly clustered at the same place (Supplementary Figure S7F and G), the two metal binding sites are located in one side of the enzyme in VH7D DpdM (Supplementary Figure S7H and I; in orange C45, C50, C52; in red C222; in yellow C53, C120, C129, C146).

**Discovery of 7-deazaguanine in *cellulophaga* phage phiST DNA**

Cellulophaga phage phiST resisted to a wide variety of restriction enzymes containing at least a guanine in their specific recognition site (Supplementary Figure S2B). In addition, phiST encodes FoIE, QueE and QueE but not QueC or DpdA (Figure 1; Supplementary Data S1) (7). The product from the reaction catalyzed by QueE, 7-carboxyl-7-deazaguanine (CDG; see Figure 1A) (3) was not detected in this phage DNA using LC-UV-MS/MS analysis. Meanwhile, peaks corresponding to three of the canonical nucleosides, dA, dC, and dT were observed, along with an unknown peak with a mass of 267 Da and retention time of 8.2 min (Figure 2A). The CID MS/MS spectra of the unknown peak revealed the protonated nucleobase [B + H]⁺ ion at m/z 151, corresponding to the glycosidic bond cleavage with the loss of a neutral 2'-deoxyribose (m/z 116). The presence of a protonated 2'-deoxyribose ion (m/z 117) and its dehydration ions (m/z 99 and 81) confirmed that the unknown peak corresponds to a non-canonical nucleoside (Figure 2B). A signal for dG was not detected, suggesting that it had been completely replaced by the unknown non-canonical nucleoside. Although CDG was not detected, the unknown modification could have been a CDG derivative.

The mass of the unknown modification is 1 Da less than dG and the HRMS of its protonated [M + H]⁺ ion (m/z 267.1094) matches well with the theoretical mass of the protonated [M + H]⁺ ion of the decarboxylated derivative of dCDG, 2'-deoxy-7-deazaguanine (dDG, m/z 267.1093, mass error = 0.37 ppm). We analyzed a synthetic dDG standard by LC-UV-MS/MS and found that its retention time and CID MS/MS spectra matched those of the unknown non-canonical nucleoside (Supplementary Figure S8), confirming that the unknown modification was dDG. The same modification was found in Cellulophaga phages phi19:2 and phi13:1 (Supplementary Figure S9), which are related to phage phiST (34).

**Prediction of a decarboxylation leading to 7-deazaguanine**

The discovery of dDG suggested that phiST encodes a CDG decarboxylase that could remove the carboxyl moiety of CDG to form DG. Between the CDG pathway and the polymerase genes of phiST lie five genes coding for protein of unknown function: CGPG_00064, CGPG_00065, CGPG_00066, CGPG_00067 and CGPG_00068. Other phages containing 7-deazaguanine modifications pathway do not encode CGPG_00064 and CGPG_00066 homologs (Supplementary Data S7 and S8). CGPG_00068 encodes a dUTPase (99.15% probability matching to PF08761.14 by HHpred, Supplementary Data S9) or MazG (98.4% probability matching to PF12643.10 by HHpred, Supplementary Data S9), which has been shown to hydrolyze dNTP in phages (38). CGPG_00065 is a distant homolog of a TGT/DpdA (98.83% probability matching to PF01702.21 by HHpred, Supplementary Data S10) and its function is discussed in the sections below. CGPG_00067 is highly similar to QueD (99.89% probability matching to PF01242.22 by HHpred, Supplementary Data S11). T-fold enzymes like QueD bind pterins or purines (39), and three of them are involved in preQ₀ synthesis (3). This gene was also found in other phages encoding for a DpdA, FoIE, QueE, and QueD but not QueC (Figure 1B; Supplementary Data S1 and S12). Because of these findings, CGPG_00067 was chosen as the best candidate for the missing decarboxylation and renamed DpdL.

To investigate structural differences between QueD and DpdL, we aligned the sequences of QueD from *E. coli* (NP_417245.1) and *B. subtilis* (NP_389256.1) with all the proposed decarboxylase phage protein sequences (Alignment S3). Both proteins share three histidines and two glutamic acids, but the position of the fourth histidine differs in the multiple alignment. The signature motif of QueD CxxxCxxHGH (40) is also changed to LxxxHRHxF in DpdL. Both histidines of the motif coordinate the zinc ion in the active site, and the cysteine is required for the catalysis of the reaction. Because the glycine residue is not involved in ligand binding or catalysis, changing it to arginine would not change any essential properties of the active site. The conversion of cysteine to leucine does, as QueD is inactive without this cysteine (41). The predicted structure of DpdL indicated that it would catalyze the reaction on the base (Supplementary Text, Supplementary Figure S10) via an alkaline decarboxylation mechanism involving zinc or other bound metal. This would imply that the specificity
of the co-encoded DpdA would be changed from preQ₀ to DG.

We expressed dpdL genes from phage phiST and Acidovorax phage ACP17 in E. coli alongside their respective dpdA genes, but we were unable to detect any dDG in this heterologous system (data not shown). Proteins may be inactive in E. coli due to temperature, salt, or codon optimization differences with their host organisms, or other unknown enzymes may be required to complete the reaction.

A DpdA is encoded in all phages that harbor 7-deazaguanine derivatives

As previously stated, CGPG_00065 is a distant homolog of a TGT/DpdA and is also found in Campylobacter phages (Figure 1B; Supplementary Data S1 and S13), which have been previously shown to be modified by ADG (6). This DpdA3 family had not previously been identified (6) and is the most logical candidate for the enzyme inserting a 7-deazaguanine derivative in the DNA of both phiST and Campylobacter phages (6).

It is difficult to predict the substrate specificity of the DpdA3 family (Figure 1A). DpdA3 is unlikely to insert preQ₀ as the full pathways are absent in phiST and the Campylobacter phages stop the synthesis at CDG (6). As a result, DpdA3 may insert CDG, a common precursor of dADG and dDG. Because the nucleoside form of ADG was detected in the cytoplasm of Campylobacter jejuni infected with phage CP220 (6), the DpdA3 might have shifted their substrate specificity to insert DG or ADG.

With the discovery of the DpdA3 subfamily, only a few of the phages/viruses identified in our previous study remained with no encoded DpdA (7). We reanalyzed the genome of Haloviruses HVT1-1, which is modified with preQ₁. HVT1_69 gene product had a 100% probability of matching with PF20314.1, a domain of unknown function (DUF6610), by HHpred, but also 92.5% with PF01702.21, a tRNA-guanine transglycosylase (Supplementary Data S14). Furthermore, homologs of this protein were found to be encoded in other archaeal viruses that also contain preQ₁ synthesis genes, as well as a singleton modification gene in a few other viral genomes, including Halorubrum phage HF2 (Figure 1B; Supplementary Data S1 and S15). With the discovery of this final DpdA subgroup, renamed DpdA4, all phages known to harbor a 7-deazaguanine in their DNA encode a DpdA family protein, which now could be considered a signature protein family for the presence of such DNA modifications.

7-(Formylamino)-methyl-7-deazaguanine in flavobacterium phage vB_FspM_immuto_2-6A DNA

Flavobacterium phage vB_FspM_immuto_2-6A DNA resisted to a wide variety of restriction enzymes containing at least a guanine in their specific recognition site (Supplementary Figure S2C). In addition, vB_FspM_immuto_2-6A encodes DpdA3, FoE, QueD, QueE, QueC and QueF (Figure 1, Supplementary Data S1) and should thus have complete guanosine replacement to preQ₁. However, dPreQ₁ was not detected in this phage genome using LC-UV-MS/MS analysis. Meanwhile, peaks corresponding to three of the canonical nucleosides, dA, dC and dT, as well as an unknown peak at a retention time of 9 min with a mass of 324 Da were observed in the LC-UV-MS/MS analysis of this phage DNA (Figure 2A). The CID MS/MS spectra of the unknown peak revealed fragment ions at m/z 208,
117, 99 and 81, which could be attributed similarly to the loss of 2′-deoxyribose to form [B + H]^+ ion and protonated 2′-deoxyribose ion and its further dehydration ions, respectively, confirming the unknown peak is a noncanonical nucleoside (Figure 2B). The dG peak was not detected, indicating that it has been completely replaced by the unknown non-canonical nucleoside. The CID MS/MS spectra of preQ1, mdPreQ1, and the unknown modification showed very similar pattern, with fragment [M + H]^+ ions observed at m/z 163 and m/z 279 for all three compounds, indicating that the unknown modification could also be a dPreQ1 derivative (Figure 2B). The [B + H]^+ ion of fdPreQ1 is stronger than the [B + H]^+ ion of mdPreQ1 and dPreQ1, but the predominant ion of all three modifications is at m/z 163, which corresponding to the loss of 2′-deoxyribose (m/z 116) together with the amino group for dPreQ1 and the methylamino group (m/z 31) for mdPreQ1. The unknown modification had a mass of 28 Da greater than dPreQ1, corresponding to one additional carbon and one oxygen (formyl group) or two additional carbons and four hydrogens (ethyl or dimethyl group). The unknown modification mass (m/z 324.1313) matched well with the theoretical mass of protonated [M + H]^+ ion of formyl-dPreQ1 (m/z 324.1308, mass error = 1.56 ppm, Supplementary Figure S9) but not with the theoretical mass of protonated ethyl or dimethyl-dPreQ1 (m/z 324.1672, mass error = 112.32 ppm). The MS/MS spectra of the unknown modification at low CID energy revealed a fragment ion with a loss of 45 Da (m/z 324 → m/z 279), corresponding to a formylamino group. This suggested that the formyl group was most likely linked to the 7-amino group, which is less stable than the 2-amino group in CID MS/MS experiment. To test our hypothesis, we chemically synthesized fdPreQ1, which was then purified using HPLC and characterized using NMR and HRMS. (Supplementary Scheme S2, Supplementary Figure S11). The standard was then analyzed using LC-UV-MS/MS. The standard retention time and MS/MS spectra were identical to those of the unknown noncanonical nucleoside, confirming that the unknown modification is fdPreQ1 (Supplementary Figure S12). This finding suggested that the vB_FspM_immuto_2-6A genome may encode a formyltransferase that adds a formyl group to dPreQ1.

**Prediction of a preQ1 formyltransferase**

A protein annotated as PF00551 formyltransferase gene is encoded close to the 7-deazaguanine insertion site of phage vB_FspM_immuto_2-6A (locus tag KNV73_0067, Figure 1B, Supplementary Data S1). This protein was used to identify similar proteins in other viral genomes (Figure 1B; Supplementary Data S1, S16). We found six phages that encode a similar protein and shared the entire pathway from FolE to QueF, including a DpdA, and 15 other phages that encode a similar protein but lacked any 7-deazaguanine modification genes. These sequences could be divided into three groups, according to a multiple sequence alignment (Supplementary S4) and a clustering cladogram (Supplementary Figure S13). One of them include four proteins that are co-encoded with the modification pathway (DpdA and FolE to QueF, Supplementary Data S1): YP_010114479.1, of phage vB_FspM_immuto_2-6A, as well as CAB4142580.1 and CAB5221950.1, all three encoded by uncultured Caudovirales phages and renamed DpdN. The other two groups appear to be unrelated to the fdPreQ1 modification because group 2 is encoded by phage that do not encode the proteins involved in the modification pathway and group 3 contains members that are longer forms of the formylase, which are likely to be involved in other reactions (Supplementary Data S1). As DpdN is a member of the same superfamily as the enzyme PurN, which catalyzed the formylation of 5-phospho-riboisyl-glycinamide in the purine synthesis pathway (42), it likely uses the same formyl donor 5-methyl-5,6,7,8-tetrahydrofolate (Supplementary Text and Supplementary Figure S14).

**7-Carboxy-7-deazaguanine in sulfolobus virus STSV-2**

Because Sulfolobus virus STSV-2 encodes DpdA and ArcS (Figure 1, Supplementary Data S1), it should harbor dGt in its genome. As we previously described (4,7), we used LC-UV-MS/MS to analyze the nucleosides obtained from enzymatic digestion of STSV-2 genomic DNA. dCDG, but not dGt, was detected at a rate of 0.04% of the Gs (149 modifications per 106 nucleotides, Supplementary Figure S15, Table 1).

There was no other neighboring gene that was clearly shared with other phages or viruses (data not shown). Surprisingly, the host archaea, _Sulfolobus tengchongensis_, does not encode any proteins involved in the G or Gt biosynthesis pathway (data not shown). We believed that its ArcS evolved to revert preQ0 into CDG. To investigate this, we aligned STSV-2 ArcS sequence with canonical ArcS proteins (21) and with homologs previously identified in other viruses (7) (Alignment S5). The phage/virus ArcS corresponds to only the core catalytic domain of the canonical ArcS (PF17884.4 annotated as DUF5591, 99.9% similar for STSV2-16 encoded by Sulfolobus virus STSV2, Supplementary Data S17, and 99.9% for VPGF_00169 encoded by Vibrio phage nt-1, Supplementary Data S18). It has previously been demonstrated that the ArcS have a high degree of diversity (21). Initially, four domains were identified in ArcS (Nt, C1, C2 and PUA). The PUA domain is specific to RNA binding, the Nt domain is similar to the TGT catalytic domain and the C1 domain is specific to ArcS and contains the catalytic core of the functions. These four domains are found in others, but in some organisms, the Nt domain is separated from the other three domains. In some archaea, the C1 domain is encoded independently, as in the phages. The C1 domain’s specific motif, PC-X3-KPY-X2- S-X2-H (21), was conserved in STSV2 ArcS but slightly degenerated in Vibrio phage nt-1 ArcS (Supplementary Figure S16, Alignment S5).

We decided to test the ArcS of phage nt-1 because the ArcS from a hyperthermophile organism might be inactive in our _E. coli_ double plasmid test system, as hypothesized previously (7). Both nt-1 dpa2 and arcS were cloned in pBAD24 and pBAD33 vectors, expressed in _E. coli_, and the plasmids were extracted. We found that DpreQ0 is inserted into DNA when nt-1 DpdA2 is expressed alone, and dGt is present when nt-1 ArcS is co-expressed (Table 3). This suggested that STSV2 ArcS, which is less degenerate than nt-1 ArcS, may have the same function, generating dGt.
Therefore, additional STSV2 proteins yet to be identified are required to catalyze the insertion of CDG in DNA in this virus.

**DISCUSSION**

In this current era of active discoveries of new bacterial defence systems against phages driven by genomic data mining (12–14), the identification of phage counter defences (11), including DNA modifications, is also rising. This study focused on a group of guanine modifications, known as 7-deazaguanines, where the nitrogen in position 7 of guanine is replaced by a carbon allowing an easier addition of various side chains at this position. In a previous study, we presented four side chains, namely dPreQ₁ and dG⁺ (7) in the genome of some viruses, and dADG and dPreQ₀ in both phage and bacterial DNA (4,7). Here, we have doubled the number of 7-deazaguanine derivatives identified in DNA, with the description of four new epigenetic marks (a) two modifications that represent further modification of dPreQ₁, namely mdPreQ₁ and fdPreQ₁; (b) one precursor of preQ₀, namely dCDG; (c) and one unprecedented natural 7-deazaguanine, dDG. In addition, we identified five previously undescribed families of viral enzymes involved in the synthesis of these modified bases (Figure 1).

When entering a cell, viral DNA may face various types of host nucleases. As such, 7-deazaguanines present chemical properties that enhance the persistence of viral DNA upon infection. For instance, deazanucleosides lower the general binding affinity of protein to DNA (43). We previously showed that 7-deazaguanines protect DNA from restriction enzymes at various levels depending on the type of modification (7). It has been demonstrated that the replacement of the guanine by a 7-deazaguanine in the recognition site of EcoRV protected against cleavage by theendonuclease (44). Viruses have likely evolved different DNA modification strategies, including the addition of deazapurines in their genome, to counteract nucleic acid-based defence systems (11–16). It is also tempting to speculate that bacteria likely have evolved anti-phage systems targeting 7-deazaguanine. Consequently, in this ‘arms race’ with their hosts, viruses may have been driven to diversify their 7-deazaguanine into various derivatives in their genome, thereby explaining the presence of various deazapurines in viral DNA.

Additionally, the study of these 7-deazaguanine modification pathways in both RNA and DNA offers examples of paralogous enzymes shifting substrate specificity (4), or the catalyzed reaction (40), non-orthologous displacements (22,45) and the involvement of yet to be discovered enzymes. In this study, we discovered and annotated two proteins that hypermodified dPreQ₁: a dPreQ₁ methyltransferase, now named DpdM, found in Cellulophaga phage phiSM, and a potential preQ₁ or dPreQ₁ formyltransferase, DpdN, leading to fdPreQ₁, in the genome of Flavobacterium phage vB_FspM_immuto_2–6A. We showed that the Vibrio phage VH7D DpdM homolog methylated dPreQ₁ into mdPreQ₁ in vivo (Table 2). Based on the analysis of the protein structures, we propose that DpdM methylates preQ₁ already inserted in DNA using two metal groups (Figure 3 and Supplementary Figure S5). We also identified a paralog of QueD in Cellulophaga phage phiST, DpdL, that we propose to promote alkaline decarboxylation of CDG to lead to the dDG modification.

The DpdA family is a good example of substrate specificity shift. Indeed, we showed that the substrate specificity of some DpdA has changed toward different 7-deazaguanines. For example, the change in substrate specificity between dPreQ₀ and dPreQ₁ seems to have occurred several times in evolution, as phages acquired both monomodal and bimodal QueF (Figure 3), and almost all DpdA sub-families have a member that may insert preQ₁ into DNA (Supplementary Data S1). We predict that dPreQ₀ was the first 7-deazaguanine DNA modification, as it is the modification that requires the fewest enzymes. Interestingly, Vibrio phage VH7D DpdA2 inserted various 7-deazaguanine derivatives in its DNA with different efficiencies (Table 2). Vibrio phage nt-1 DpdA2 did not insert preQ₁ in DNA in our assay but could insert preQ₀ and possibly G⁺ (Table 3). It was previously reported that this phage harbored three 7-deazaguanine DNA modifications, at various levels (7). DpdA2 family exhibits promiscuity for substrates specificity.

The DpdA family is also a good example of proteins changing slightly their affinity to fit the needs of organisms, and a good case study for protein classification. Thus far, the DNA transglycosylases that insert preQ₀ into DNA have been classified into three subfamilies: bDpdA (4), DpdA1, and DpdA2 (7). We identified two other subfamilies in viruses, DpdA3 and DpdA4. All viruses encoding 7-deazaguanine synthesis genes now encode a member a DpdA subgroup. Thus, we propose that all 7-deazaguanine DNA modifications reported to date are post-replication modifications. Interestingly, the efficiency of insertion by DpdA members varied between subgroups. bDpdA appears to have a low insertion rate, less than 0.1% of the Gs (4), similar to what was observed for DpdA2 (7) (Table 1). However, modification levels vary from 0.1 to 30% of the Gs for DpdA1 (7). The genome of the only DpdA4 encoding virus tested (Halovirus HVT1) was modified at 30% (7). DpdA3 is the most efficient, completely modifying the genomes of Campylobacter phage CP220 (6), Cellulophaga phage phiST, and Flavobacterium phage vB_FspM_immuto_2–6A (Table 1). Our attempts to
test members of the DpdA3 and DpdA4 families in our *E. coli* model were unsuccessful.

Finally, these pathways still leave open questions, such as the presence of dCDG in Sulfolobus virus SVST-2. This virus only encodes DpdA and ArcS homologs. Furthermore, its model does not modify its tRNA with 7-deazaguanines. The proteins encoded in the vicinity of *dpa* and *arcS* were not found in any other phage or virus encoding a 7-deazaguanine modification pathway. We do not know what the source of 7-deazaguanine for this virus is, nor how it ends up on dCDG. Indeed, we expected dG instead because Vibrio phage nt-1 ArcS produces dG (Table 3). Hence, we renamed this enzyme ArcS2 to differentiate from its tRNA-acting homolog (21). Future research will be needed to answer these questions.

**DATA AVAILABILITY**

Protein model predicted by AlphaFold2 were deposited in ModelArchive:

- Cellulophaga phage phiSM DpdM DOI: https://www.modelarchive.org doi/10.5452/ma-tqgw7
- Vibrio phage VH7D DpdM DOI: https://www.modelarchive.org doi/10.5452/ma-117yh
- Cellulophaga phage phiST DpdL DOI: https://www.modelarchive.org doi/10.5452/ma-bxwuk
- Flavobacterium phage vB_FspM_immuto_2–6A DpdN DOI: https://www.modelarchive.org doi/10.5452/ma-t6vzw

**SUPPLEMENTARY DATA**

Supplementary Data are available at NAR Online.

**FUNDING**

Human Frontier Science Program [HFSP-RGP0024/2018 to V.d.C.L., S.M.]; National Institutes of Health [GM70641 to V.d.C.L., ES031576 to P.C.D.]; National Research Foundation of Singapore under the Singapore-MIT Alliance for Research and Technology Antimicrobial Resistance IRG and from Agilent ACT-UR program [4658, 4773]; S.M. holds the Canada Research Chair in Bacteriophages. Funding for open access charge: National Institutes of Health [GM70641].

Conflict of interest statement. None declared.

**REFERENCES**
