Novel genomic island modifies DNA with 7-deazaguanine derivatives

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.
Novel genomic island modifies DNA with 7-deazaguanine derivatives

Jennifer J. Thiaville,a,b Stefanie M. Kellner,b,c 1 Yifeng Yuan,b,c 1 Geoffrey Hutinet,b 1 Patrick C. Thiaville,b Watthanachai Jampathong,b,c 1 Susovan Mohapatrab,c,d 1 Celine Brocher-Armanet, 1 André V. Letarov, 6 Roman Hillebrand,b,c, 1 Chanchal K. Malik, 1 Carmelo J. Rizzo, 5 Peter C. Dedon,b,c, 1 and Valérie de Crécy-Lagard,b,c, 4

aDepartment of Microbiology and Cell Science, University of Florida, Gainesville, FL 32611-0700; bDepartment of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; cCenter for Environmental Health Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139; dUniversité de Lyon, Université Lyon 1, CNRS, UMR5558, Laboratoire de Biométrie et Biologie Évolutive, F-69622 Villeurbanne, France; eWinogradsky Institute of Microbiology, Research Center of Biotechnology of Russian Academy of Sciences, Moscow 119071, Russia; fDepartment of Chemistry, Vanderbilt University, Nashville, TN 37235; gDepartment of Biochemistry, Center in Molecular Toxicology, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN 37232; hInfectious Disease Interdisciplinary Research Group, Campus for Research Excellence and Technological Enterprise, Singapore-MIT Alliance for Research and Technology, Singapore, Singapore 138602; and iUniversity of Florida Genetics Institute, Gainesville, FL 32610

Edited by Gregory J. Hannon, University of Cambridge, Cambridge, United Kingdom, and approved February 2, 2016 (received for review September 18, 2015)

The discovery of ~20-kb gene clusters containing a family of paralogs of tRNA guanosine transglycosylase genes, called tgtAS, alongside 7-cyano-7-deazaguanine (preQ0) synthesis and DNA metabolism genes, led to the hypothesis that 7-deazaguanine derivatives are inserted in DNA. This was established by detecting 2′-deoxy-preQ0 and 2′-deoxy-7-amido-7-deazaguanosine in enzymic hydrolysates of DNA extracted from the pathogen, Gram-negative bacteria Salmonella enterica serovar Montevideo. These modifications were absent in the closely related S. enterica serovar Typhimurium LT2 and from a mutant of S. Montevideo, each lacking the gene cluster. This led us to rename the genes of the S. Montevideo cluster as dpdA-K for 7-deazapurine in DNA. Similar gene clusters were analyzed in ~150 phylogenetically diverse bacteria, and the modifications were detected in DNA from other organisms containing these clusters, including Kineococcus radiotolerans, Comamonas testosteroni, and Sphingopyxis alaskensis. Comparative genomic analysis shows that, in Enterobacteriaceae, the cluster is a genomic island integrated at the leuX locus, and the phylogenetic analysis of the TgtAS family is consistent with widespread horizontal gene transfer. Comparison of transformation efficiencies of modified or unmodified plasmids into isogenic S. Montevideo strains containing or lacking the cluster strongly suggests a restriction–modification role for the cluster in Enterobacteriaceae. Another preQ0 derivative, 2′-deoxy-7-formamidino-7-deazaguanosine, was found in the Escherichia coli bacteriophage 9g, as predicted from the presence of homologs of genes of DNA extracted from the pathogen. The discovery of a novel modification system that inserts 7-deazaguanine derivatives in DNA, modifications thought until now to occur only in RNA, is an excellent illustration of the power of biological evolution to alter the ultimate function not only of the distinct proteins but also of entire metabolic pathways. The extensive lateral transfer of the gene cluster responsible for this modification highlights its significance as a previously unrecognized foreign DNA defense system that bacteria and phages use to protect their genomes. The characterization of these DNA modification pathways also opens the door to novel tools to manipulate nucleic acids.

DNA modification | restriction–modification | 7-deazaguanine | comparative genomics | queuosine

Hypemodifications of DNA requiring more than one synthetic enzyme are not as prevalent and chemically diverse as RNA hypermodifications, but around a dozen have been identified in DNA to date (1). The functions of most DNA hypermodifications are still not known, but some have roles in protection against restriction enzymes, whereas others affect thermal stability temperature, DNA packaging, or transcription regulation (2). For example, the hypermodified DNA base β-glucosyl-hydroxymethyluracil, or base β, is an epigenetic factor that regulates Pol II transcription–initiation in ketoplastids of trypanosomes (3). The recently discovered phosphorothioate (PT) modification of the DNA backbone in bacteria was found to perform different functions in different organisms (4–6). In Salmonella Cerro 87, PT occurs on each strand of a GAAC/GTTC motif as part of a restriction–modification (R–M) system, whereas in Vibrio cholerae FF75, which lacks PT restriction enzymes, PT occurs on one strand of CpGCA motifs, and the function remains unclear (6). In 2013, Iyer et al. described the computational prediction of 12 novel DNA hypermodification systems in phage and bacteria (7), demonstrating the potential diversity and complexity of modifications yet to be discovered. 7-Cyano-7-deazaguanine (preQ0) is a common precursor of the widespread RNA modifications queuosine (Q) and archaeosine (GQ) (8) and of pyrrolopyrimidines such as toyocamycin or tubercidin (9). In both Archaea and Bacteria, preQ0 is synthesized from GTP in a pathway that has been fully characterized in the last 10 y (Fig. L4). The first step catalyzed by GTP cyclodrolase I (GCHI; FolE) is shared with the tetrahydrofolate synthase pathway (10), and then three enzymes—6-carboxy-5,6,7,8-tetrahydropterin synthase (QueD), 7-carboxy-7-deazaguanine synthase (QueE), and 7-cyano-7-deazaguanine synthase (QueC)—lead to the formation of the preQ0 moiety (11, 12) (Fig. 1 L4).

The synthesis of GQ and Q diverge after the formation of preQ0. In Bacteria, the tRNA guanosine (34) transglycosylase (btGTG; EC 2.4.2.27), an enzyme that targets the G at position 34 of tRNAs with GUN anticodons (13) prefers the 7-deazapurine (preQ0) base in one step by NADPH-dependent enzyme preQ0 oxidoreductase (QueF) (14), but it can

Significance

The discovery of a novel modification system that inserts 7-deazaguanine derivatives in DNA, modifications thought until now to occur only in RNA, is an excellent illustration of the power of biological evolution to alter the ultimate function not only of the distinct proteins but also of entire metabolic pathways. The extensive lateral transfer of the gene cluster responsible for this modification highlights its significance as a previously unrecognized foreign DNA defense system that bacteria and phages use to protect their genomes. The characterization of these DNA modification pathways also opens the door to novel tools to manipulate nucleic acids.


The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

1J.J.T., S.M.K., and Y.Y. contributed equally to this work.
2Present address: Department of Chemistry, Chiang Mai University, Chiang Mai, Thailand 50200.
3Present address: WaVe Life Sciences, Boston, MA 02135.
4To whom correspondence may be addressed. Email: vzrcy@ufl.edu or pcdnodon@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1518570113/-/DCSupplemental.
also use preQ₀ when preQ₁ is absent (14). Thus, preQ₀ is reduced to preQ₂ and inserted in tRNAs by bTGT (Fig. L4). Two subsequent enzymatic steps, carried out by QueA and QueG, produce the final Q nucleoside (for recent review, see ref. 9). In Archaea, the tRNA guanosine (15) transglycosylase enzyme (aTGT; EC 2.4.2.48) is homologous to the bTGT enzyme and changes the G at position 15 with preQ₀ in nearly all tRNAs (15, 16). The preQ₀ is then modified to G⁺ by different types of amidotransferases [archaeosine synthase (ArcS), QueF-like, and glutamine amidotransferase class-II (GAT)-QueC] (17, 18) (Fig. 1L4). Although the bTGT and aTGT recognize a guanosine at different positions of the tRNA and use different substrates, key residues involved in base exchange and in zinc binding are conserved (19) (Fig. 24). In addition, signature residues involved in the differences in 7-deazaguanosine substrate recognition between the bTGT and aTGT enzymes have been identified (19) (Fig. 2B). The role of 7-deazaguanosine derivatives as precursors of modified bases in prokaryotic genomes showed that some organisms, such as pathogenic strains of E. coli (strain E22) or Salmonella enterica subsp. enterica serovar Montevideo, contained two homologs of the tgt gene, whereas most other sequenced E. coli or Salmonella species contained only one (SI Appendix, Table S1). Synteny analysis revealed that one of the two tgt genes clustered with the Q synthesis gene queA and encoded the experimentally characterized bTGT enzyme (24) (SI Appendix, Figs. S1 and S24), whereas the other, that we named tgtA5 (later changed to dpdA), was found in a different neighborhood context (Fig. 1B). Homologs of tgtA5 were found in nearly 264 complete prokaryotic genomes (SI Appendix, Table S1 and Fig. S10A and B), and the TgtA5 proteins possess divergent features from the bTGT that inserts preQ₂ at position 34 of tRNA (Fig. 2). Members of the TgtA5 family are larger proteins (average of ∼450 aa instead of ∼300 aa for bTGT), and only the core of TgtA5 shows significant similarity to the bTGT and aTGT enzymes (Fig. 2B). The key residues that catalyze the G exchange (Asp102 and Asp280 of bTGT and Asp95 and Asp249 of aTGT) (19), as well as the Zinc binding site (CXCXXCXH motif), are conserved in TgtA5 (Fig. 2B). Analysis of the substrate binding pocket suggests that TgtA5 binds preQ₀ like the aTGTs. The critical residues for preQ₀ binding by aTGT are GVVPL[L,M] at positions 196–201 of the P. horikoshii enzyme, differing from the bTGT preQ₁ binding pocket residues GLAVGE at position 230–235 of the Z. mobilis enzyme (19). Alignments of TgtA5 sequences with bTGT and aTGT showed the binding pocket residues resembled aTGT more than bTGT (G[ML]V[PL][KR] in

Results

A Bacterial TGT Variant, TgtA5, Must Be Involved in a PreQ₀-Dependent Pathway Different from Q Synthesis. Analysis of the distribution of all Q synthesis genes in bacteria was performed using the “dpd cluster” subsystem in the SEED database (23). Analysis of ∼12,000 bacterial genomes showed that some organisms, such as pathogenic strains of E. coli (strain E22) or Salmonella enterica subsp. enterica serovar Montevideo, contained two homologs of the tgt gene, whereas most other sequenced E. coli or Salmonella species contained only one (SI Appendix, Table S1). Synteny analysis revealed that one of the two tgt genes clustered with the Q synthesis gene queA and encoded the experimentally characterized bTGT enzyme (24) (SI Appendix, Figs. S1 and S24), whereas the other, that we named tgtA5 (later changed to dpdA), was found in a different neighborhood context (Fig. 1B). Homologs of tgtA5 were found in nearly 264 complete prokaryotic genomes (SI Appendix, Table S1 and Fig. S10A and B), and the TgtA5 proteins possess divergent features from the bTGT that inserts preQ₂ at position 34 of tRNA (Fig. 2). Members of the TgtA5 family are larger proteins (average of ∼450 aa instead of ∼300 aa for bTGT), and only the core of TgtA5 shows significant similarity to the bTGT and aTGT enzymes (Fig. 2B). The key residues that catalyze the G exchange (Asp102 and Asp280 of bTGT and Asp95 and Asp249 of Pyrococcus horikoshii aTGT) (19), as well as the Zinc binding site (CXCXXCXH motif), are conserved in TgtA5 (Fig. 2B). Analysis of the substrate binding pocket suggests that TgtA5 binds preQ₀ like the aTGTs. The critical residues for preQ₀ binding by aTGT are GVVPL[L,M] at positions 196–201 of the P. horikoshii enzyme, differing from the bTGT preQ₁ binding pocket residues GLAVGE at position 230–235 of the Z. mobilis enzyme (19). Alignments of TgtA5 sequences with bTGT and aTGT showed the binding pocket residues resembled aTGT more than bTGT (G[ML]V[PL][KR] in
Fig. 2. Comparison of TGT and TgtA5 proteins. (A) Schematic representation of the domain architecture and arrangement of TgtA5 proteins and bacterial and archaeal TGT proteins (bTGT and aTGT, respectively). The numbering of the upper and lower logos refers to the S. Montevideo TGT and TgtA5 sequences, respectively. Sequence logos in dashed boxes show the two conserved Asp residues of TgtA5 and the zinc binding sites of bTGT (Top) and aTGT (Bottom). C1 and C2 represent C-terminal domains unique to aTGT (17, 18). (B) Model and alignments of proposed substrate-binding pocket of TgtA5. The aligned cartoon representation (Top) of the pockets of S. Montevideo TgtA5 and P. horikoshii aTGT (PDB ID code 1IT8) was produced by PyMol (version 1.3). The catalytic residues of aTGT, ASP95, VAL197, VAL198, and ASP249 (red) (18) and their TgtA5 counterparts ASP95, MET208, VAL209, and ASP256 (cyan) are indicated in stick models. Dashed lines among stick models indicate the catalytic residues interacting with preQ. Sequence alignment (Bottom) of select aTGT, bTGT, and TgtA5 proteins was performed using MUSCLE (53). Dots indicate regions intentionally deleted for this figure. Dashes indicate gaps in the sequence alignment. Uniprot IDs for proteins included in multiple alignment are as follows: S. Montevideo TgtA5, E7VB44; P. balearica TgtA5, E15VY3; S. alaskensis TgtA5, Q1GPSS0; Comamonas testosteroni TgtA5, H1RKG1; K. radiotolerans TgtA5, A6WGA1; E. coli bTGT, P0A847; Z. mobilis bTGT, O67331; P. horikoshii aTGT, OSB848; Methanococcus aeolicus aTGT, ABUDR; Thermoplasma volcanium aTGT, Q977Z2; Picrophilus torridus aTGT, Q6L1W3; Ferroplasma acidarmanus aTGT, 50A2Q3.

TgtA5; Fig. 2B). Modeling of the S. Montevideo TgtA5 protein with the aTGT structure with preQ in the binding pocket demonstrated the similar placement of these binding pocket residues compared with the aTGT (Fig. 2B), supporting our hypothesis that TgtA5 binds preQ. Finally, tgtA5 clusters with the preQ synthesis genes in 94% of analyzed genomes, but it does not cluster with queF, suggesting that the substrate is preQ or a derivative, and not preQ. Sequence and genome context analyses predict that TgtA5 recognizes preQ as a substrate like the aTGT enzymes, but because TgtA5 proteins lack the tRNA binding PUA domain found in aTGTs (25), analyses strongly suggest that TgtA5 proteins do not target tRNAs (Fig. 24). Also, tgtA5 genes are found in organisms that lack the canonical Q synthesis gene tgt, such as Kineococcus radiotolerans (SI Appendix, Table S1), and we confirmed that K. radiotolerans lacked Q and preQ in tRNA (SI Appendix, Fig. S14). Finally, the tgt gene was deleted in the S. Montevideo strain that contained both a tgt and a tgtA5 gene, and RNA extracted from the Δtgt strain (YF630522) lacked Q (SI Appendix, Figs. S1B and S2), confirming that TgtA5 is not involved in incorporation of Q in tRNA.

Strong physical clustering was observed between tgtA5 with homologs of a DndB-like protein that is involved with the PT modification of DNA in bacteria (5) (Fig. 1B and SI Appendix, Table S1). These dndB-like genes were present in 123 of 134 (92%) tgtA5 gene clusters analyzed, and in 21 clusters (15%), there were two distinct copies of this gene flanking the tgtA5 gene (Fig. 1B and SI Appendix, Table S1 and Fig. S4). According to HHIPred analysis and previous work by Iyer et al. (7), DndB proteins contain a domain belonging to the superfamily of ParB nucleases involved in chromosome and plasmid partitioning (26, 27), suggesting a role in DNA recognition and binding. Studies have shown that DndB negatively regulates the PT modification (28–30) by binding to the promoter region and regulating expression of the DndBCDE operon (30). The S. Montevideo DndB-like protein (renamed DpdB) has 23% amino acid identity to the DndB protein of S. Cerro. Despite the low similarity, several residues are conserved among the PT-related DndB and TgtA5-clustered DndB-like proteins, including a QR doublet near the N terminus and a FXXXYX motif near the middle of the sequence. Most striking, however, is the strictly conserved DGQR motif in nearly all of the TgtA5-clustered DndB-like proteins, which differs by one residue from the DGQR motif conserved among the DndB proteins involved in PT modification (SI Appendix, Figs. S3 A and B). Although the PT-related DndB and the TgtA5-clustered DndB-like (DpdB) proteins share conserved motifs, they separate on a phylogenetic tree (SI Appendix, Fig. S3B), indicating that they comprise two subfamilies of the DndB-like family.

Strong physical clustering was also observed with tgtA5 and several other genes (Fig. 1 and SI Appendix, Figs. S4 and S5). Two genes of unknown function, which we call dpdC and dpdD, were present in 88% and 90% of the clusters analyzed, respectively. DpdC was predicted to encode a DUF328 domain-containing protein, and DpdD has little similarity to any known protein, although a small portion of the C terminus matched a DUF328 domain (SI Appendix, Fig. S5 and Table S2). In 98% of the bacteria analyzed, the tgtA5 genes also clustered with several other putative DNA-binding enzymes, including a member of the DEAD/DEAH box helicase family, a SNF2-type helicase, and a RecQ-like Superfamily II DNA helicase (Fig. 1 and SI Appendix, Figs. S4 and S5 and Tables S1 and S2). This grouping allowed us to speculate that TgtA5 is involved in introducing preQ-like modifications in DNA, a hypothesis previously proposed by Aravind and colleagues (7).

Nine of the analyzed genomes contained much larger TgtA5 proteins (~700 aa). The TgtA5 domain of these proteins is
similar to the rest of the TgtA5 family. The N-terminal half of the protein contains a DUF328 domain, like the one present in the DpdC of the S. Montevideo cluster. The genomic context of the longer tgtA5 genes includes preQ0 synthesis genes, similar to the other tgtA genes; however, one noticeable difference is the absence of some or all of the three genes conserved in the other clusters, the dndB-like dpdB, and dpdC and dpdD (see the Microbacterium chilensis DSM 9857 cluster in SI Appendix, Fig. S4 and others in SI Appendix, Fig. S6). Some of the long tgtA5 clusters encode a similar SNF2-type helicase, RecQ-like helicase, phospholipase-like domain-containing protein, and a DpdD, whereas others have other putative DNA-binding proteins and helicases, suggesting that the long TgtA5 protein is also involved in introducing a modification into DNA in these organisms.

**The tgtA5 Cluster Is Responsible for the Insertion of PreQ0 and of 7-Amido-7-Deazaguanine in DNA.** To test the hypothesis that 7-deazaguanine derivatives were inserted in the DNA of organisms that encode the tgtA5 cluster, a mass spectrometry-based approach was used to analyze DNA from two closely related Gram-negative bacteria possessing and lacking this gene cluster (S. Montevideo and S. Tyrphimurium LT2, respectively) and from the Gram-positive bacteria K. radiotolerans that also encodes the cluster (Fig. 1B). The strategy for discovering the 2-deoxynucleosides was based on an initial presumption of the presence of 2'-deoxynucleosides containing any of the six 7-deazaguanine nucleobase structures formed in the tRNA queuosine biosynthetic pathway shown in Fig. 1A: 2'-deoxyCpH2, 2'-deoxyCDG, 2'-deoxyQ, 2'-deoxyPreQ0, 2'-deoxyPreQ1, and 2'-deoxyG+. A search for each candidate 2'-deoxynucleoside was conducted by neutral loss analysis mass spectrometry, in which product ions resulting from loss of a 2-deoxyribose during collision-induced dissociation could be traced back to the original 2'-deoxynucleoside eluting from the HPLC column at a specific retention time. In both S. Montevideo and K. radiotolerans, small amounts of putative 2'-deoxyPreQ0 (dPreQ0) and a stronger signal for putative 2'-deoxyCDG (dCDG) were detected (SI Appendix, Fig. S7). Subsequent structural analysis revealed that the prediction of dCDG was incorrect and that the signal at m/z 311 was actually the M - H+ ion of 2-deoxypreQ0 (SI Appendix, Fig. S8). The identities of dPreQ0 and dADG were established by fragmentation analysis using high mass resolving power orbitrap mass spectrometry (QT time-of-flight QTOF) mass spectrometry (SI Appendix, Fig. S8) and by comparison with synthetic standards. Standards were also used to rule out detectable levels of dPreQ1 and 2'-deoxyArch (dG+) (SI Appendix, Fig. S8A). Using these standards, the optimal mass transitions (Fig. 3B and SI Appendix, Fig. S8 B-D) and retention times (SI Appendix, Fig. S9) of the modified 2'-deoxynucleosides were determined. Quantitative analysis by external calibration revealed ~1,600 dADG modifications per 106 nt in S. Montevideo and ~1,300 per 106 nt in K. radiotolerans (Fig. 3C). dPreQ0 levels were found to be significantly lower at 10 and 30 dPreQ0 per 106 nt, respectively (Fig. 3C). These results suggest that dADG is the main product of tgtA5 cluster, with dPreQ0 appearing as a side product. To confirm the role of the tgtA5 cluster in the insertion of dPreQ0 and dADG in DNA, a S. Montevideo derivative (YFY3022) with a 21-kb deletion eliminating nearly the entire cluster was constructed (Fig. 1B and SI Appendix, Fig. S2). Both the dPreQ0 and dADG modifications were absent in genomic DNA extracted from YFY3022 (Fig. 3C).

This discovery led us to rename the genes of the S. Montevideo cluster as dpdA-K (Fig. 1B), with dpd standing for 7-deazapurine in DNA.

**The dpd Cluster Is Horizontally Transferred and Found in Genomic Islands.** Analysis of the taxonomic distribution of the dpdA (tgtA5) gene (Fig. 4) showed that this gene is evenly distributed along the bacterial tree, suggesting either an ancestral origin in Bacteria accompanied by massive independent losses along the diversification of Bacteria or a more recent origin with propagation through horizontal gene transfers (HGTs). The phylogenetic analysis of the Dpda/TgtA5 homologs and the discrepancies observed between the topology of the resulting Bayesian and maximum likelihood dpdA/TgtA5 trees with the currently recognized systematics (compare Fig. 4 with SI Appendix, Fig. S10 A and B), such as the monophyly of Grammoproteobacteria or Archaear or the robust grouping of Herbaspirillum massiliense (Betaproteobacteria) with Sirevosa spitsbergense (Bacteroidetes) and a Verrucomicrobia bacterium but not with other Betaproteobacteria (bootstrap value, 100%; posterior probability, 1.00; SI Appendix, Fig. S10 A and B), strongly favor the HGT hypothesis.

To confirm that the presence of dpa4 is diagnostic of the presence of preQ0 and ADG in DNA, we analyzed the genomic DNA from a diverse set of organisms harboring the cluster (SI Appendix, Fig. S4 and Fig. 4). As shown in Fig. 3C, M. chilensis, Comamonas testosteronei, Sphingopyxis alaskensis, and Ferrimonas balearica all harbor dADG in DNA but in different quantities. Unlike the other strains analyzed, the DNA from M. chilensis contained only ADG and no detectable preQ0 (Fig. 3C). M. chilensis was the only strain tested with a long version of the dpdA and no dpdB cluster (SI Appendix, Fig. S4).

In the S. Montevideo and E. coli strains that harbor the dpd cluster, it is inserted adjacent to the LexX locus (SI Appendix, Fig. S11), a region that had previously been identified as highly variable (31). A pair of 19-bp direct repeats flanking the dpd cluster was identified (SI Appendix, Fig. S11), again indicative of a genomic island (32, 33). More recently, the sequencing of a large number of Salmonella strains confirmed this region as a novel variable island (SG12) that contained different types of restriction systems, toxin/antitoxin modules, and mobile elements (34). Of the sequenced S. Montevideo strains analyzed, 92% contain the dpdA cluster at the SG12 position.

Unlike the dpd islands of S. Montevideo and E. coli, neither tRNA genes nor direct repeats were identified in the region surrounding the K. radiotolerans dpd cluster. However, the region is flanked by mobile element proteins (e.g., a transposase-like protein and resolvase protein), and the IslandViewer software (35) identified this cluster as a genomic island. Additionally, the GC content of the ORFs is lower in this region compared with the rest of the genome (66% GC vs. 74% GC).

**The dpd Cluster of S. Montevideo Encodes an R-M System.** The discovery of a horizontally transferred DNA modification cluster logically suggested a potential role as a novel R-M system. To test this hypothesis, we compared the transformation efficiencies of the isogenic strains S. Montevideo WT and YFY3022 lacking the dpd cluster. pUC19 was propagated in each strain, and following extraction, 10 ng of the plasmid DNA was used to transform the WT and YFY3022 strains by electroporation. As shown in Fig. 5, pUC19 DNA extracted from YFY3022 transformed the WT strain with 100-fold less efficiency than pUC19 DNA extracted from the WT strain, indicating the unmethylated plasmid is restricted in the WT host. The pUC19 extracted from YFY3022 transformed the YFY3022 strain with about 1,000-fold higher efficiency than the WT strain, suggesting that one or more of the genes in the dpd cluster is responsible for this restriction. No difference was seen with the plasmid extracted from the WT strain. Liquid chromatography (LC)-MS/MS analysis confirmed that dPreQ0 and dADG were present in the pUC19 extracted from S. Montevideo WT and not the mutant strain (Fig. 3C).

**PreQ0 Derivatives in Phage.** R-M systems and genomic islands are often transferred through phage transduction. Several examples of phage-encoded tgt-like genes and preQ0 genes have already been reported in the literature, including Mycobacteriophage Rosebush (36), Streptococcus phage Dp-1 (37), and the E. coli...
In the characterization of phage 9g, Kulikov et al. (38) speculated that the restriction endonuclease-resistant nature of the phage DNA suggested it was heavily modified, and they proposed that tgt and preQ₀ synthesis genes were involved in inserting Q into the DNA. To evaluate the prevalence of tgt paralogs and PreQ₀ synthesis genes in phages, a similarity-based search was performed on all available phage genomes in the National Center for Biotechnology Information database. This revealed 36 bacteriophages and two archaeal viruses that encode a Tgt-like protein (SI Appendix, Table S5). Multiple sequence alignments of the phage Tgt-like proteins allowed the identification of catalytic residues (two conserved Asp residues) and of the preQ₀-binding pocket (SI Appendix, Fig. S13) (19). The zinc-binding residues that are conserved in the aTGT, bTGT, and TgtA5 families were not found in the phage homologs; however, a His residue (H196 of phage 9g Tgt-like protein) is conserved specifically in the phage enzymes.

The preQ₀ biosynthesis pathway genes (folE, queD, and queC) were identified in 16 of Tgt-containing phages (SI Appendix, Table S5), three of which contained a homolog of Gat-QueC that is involved in the synthesis of archaeosine in some archaea (18). Two of the phages harbored a QueF homolog involved in preQ₁ synthesis. Finally, one phage, phi13:1, encoded the three first genes of the preQ₀ synthesis pathway (folE, queD, and queE) but no tgt homolog (SI Appendix, Table S5).

The genomic contexts of the phage preQ₀/tgt clusters are different from those found in bacteria, however many of them encode DNA processing enzymes and could therefore insert 7-deazaguanosine derivatives in DNA. Some of these phages encode a homolog of ParB, an enzyme important for DNA binding and segregation (26), as previously pointed out by Aravind’s group (7) (e.g., Mycobacteriophage Rosebush). Other phages encode potential helicases and nucleases near the preQ₀ cluster [e.g., Gp11 of 9g and Gp39 of JenKI contain SnfII-like domains (39), Gp11 of Dp-1 is a RecU-like protein (40), Gp40 of JenKI encodes a putative exonuclease (41), and Gp10 of Dp-1 is a Cas4-like protein (42); SI Appendix, Fig. S12]. The presence of these nucleases is indicative of possible defense systems. The nature of the exact modification might differ with the specific phage, as the preQ₀...
pathway is found in Rosebush or BCD7, the preQ\(_0\) pathway in Dp-1, and the archaeosine pathway in 9g (SI Appendix, Table S5).

To test the hypothesis that some of these phages contained preQ\(_0\) derivatives, DNA from phage 9g, a phage predicted to insert archaeosine because of the presence of the gat-queC gene, was isolated and subjected to LC–MS/MS analysis as described above. As expected, 2'-deoxy-archaeosine (dG\(^+\)) was found in the phage DNA (Fig. 6B). dG\(^+\) quantities were extremely high, allowing quantification by both MS/MS and UV analysis that revealed a conversion of dG to dG\(^+\) by 25% and 27%, respectively.

**Discussion**

The discovery of 7-deazag derivatives in DNA of diverse bacteria and phages is a compelling example of the power of coupling in silico predictive approaches with bioanalytical validation (43, 44). These modifications would not have been identified if we had not purposely looked for them in very specific organisms. It was recently shown that ADG is an intermediate in the QueC-mediated reaction from CDG to preQ\(_0\) (45) (Fig. 1). ADG is the result of amidation of CDG, a reaction that occurs more rapidly compared with the subsequent dehydration to preQ\(_0\). DpdA is most likely involved in exchanging guanine in DNA for ADG or preQ\(_0\) in a base exchange reaction similar to bTGT and aTGT, but biochemical studies are required to test this hypothesis and determine if it can do it alone or if it requires other proteins encoded within the cluster. Because the majority of the detected modifications were dADG, it is likely that ADG is the preferred substrate of DpdA, and dPreQ\(_0\) may be present as the result of nonspecific insertion by DpdA. The presence of only dADG, and not dPreQ\(_0\), in the DNA of M. chliarophilus suggests that its DpdA has stricter substrate specificity. The DpdA of M. chliarophilus is a larger protein, with an additional 300 amino acids at the N terminus, and the substrate binding pocket of the longer DpdA proteins has a slightly different sequence motif (GGLAR vs. GGMVP of other DpdAs). The LA residues resemble the bTGT preQ\(_0\) binding pocket rather than the aTGT preQ\(_0\) pocket and may confer specificity for ADG.

The dpdB gene encodes a member of the DndB-like family. The DndB proteins appear to have a role in regulation of the PT modification of DNA. Deletion of dndB homologs in Streptomyces lividans and Salmonella Cerro led to increased PT modification (28, 43).

**Fig. 5.** Transformation efficiency of modified and unmodified pUC19 DNA. *(A) S. Montevideo WT and YYF3022 (ΔdpdC:dpdD::kan) transformed with 10 ng pUC19 extracted from either WT (modified) or ΔdpdC:dpdD::kan (unmodified) on LB agar plates containing ampicillin. (B) Transformation efficiencies of modified versus unmodified pUC19 in WT and ΔdpdC:dpdD::kan. Transformation efficiency per 1 ng DNA was calculated per 10\(^6\) viable cfu. The average of three experiments is shown, with error bars representing SE (*P < 0.05, two-tailed Student’s t test).
Further studies are needed to elucidate the roles of the 7-deazaguanine derivatives in bacterial and phage DNA, with potential functions varying among R–M systems, antirestriction systems, epigenetic marks, and unforeseen protective roles, as these modifications were found in organisms like *K. radiotolerans* that can resist radiation stress (48). We foresee that the molecular characterization of the enzymes involved in the synthesis, recognition, and cleavage of 7-deazaguanine derivatives in DNA could open the door to both biotechnological and antibacterial applications.

Materials and Methods

Bioinformatic Analyses. Taxonomic distribution and physical clustering analysis of tgtAS and preQ0 synthesis genes was performed on the public SEED server (pubseed.theseed.org/SubsysEditor.cgi) (23, 49). Results of the analysis are available in the dpd cluster subsystem and summarized in *SI Appendix, Table S1*. The taxonomic distribution of tgtAS was then visualized using the Interactive Tree of Life (iTol, itol.embl.de) (50). Further details on all bioinformatic analyses can be found in *SI Appendix, SI Materials and Methods*.

Strains, Media, and Growth Conditions. Strains used in this study are listed in *SI Appendix, Table S3*. *S. Montevideo* strains were routinely grown in LB (Tryptone 10 g/L, yeast extract 5 g/L, NaCl 5 g/L) at 37 °C. All other strains were grown in media and conditions as detailed in *SI Appendix, SI Materials and Methods*. *S. Montevideo* deletion constructs were made using the linear recombination method described by Datensk and Wanner (51). Oligonucleotides used for deletion and confirmation of mutants are listed in *SI Appendix, Table S4*. Further details can be found in *SI Appendix, SI Materials and Methods*.

Plasmid Restriction Test. Restriction of the plasmid pUC19 was tested as described in ref. 52 and is detailed in *SI Appendix, SI Materials and Methods*.

DNA Preparation. Total DNA was extracted from bacteria and phage with phenol-chloroform followed by alcohol precipitation, as detailed in *SI Appendix, SI Materials and Methods*.

DNA Analysis. DNA was enzymatically hydrolyzed to 2′-deoxynucleosides as described in *SI Appendix, SI Materials and Methods*. Modified 2′-deoxynucleosides were initially detected by LC/MS analysis. Subsequently, dPreQ0 and dADG were identified using tandem mass spectrometry. dPreQ0, dADG, dCDG, and dGβ′ were achieved by LC–MS/MS using external calibration curves. Details can be found in *SI Appendix, SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank Rémi Zallot, Zdenek Paris, and Juan Alfonzo for help with the determination of the presence of queuosine in tRNA; Sophie Alvarez for LC–MS/MS analyses of tRNA; Anne Oudard for sharing her local database of complete and ongoing genome projects; Max Teplitski for the *S. Montevideo* strain; and Paul Guilig and Kelly Rice for use of laboratory facilities. Mass spectrometric analyses were performed in the Bioanalytical Facilities Core of the MIT Center for Environmental Health Sciences, which is supported by National Institute for Environmental Health Sciences Grant ES002109. This work was supported by the National Institutes of Health Grant GM70641 (to V.d.C.-L.), Deutsche Forschungsgemeinschaft (S.M.K.), French National Agency for Research Grant ANR-10-BIFIN-01-01 and the Institut Universitaire de France (to C.B.-A.), Russian Scientific Foundation Grant RSF 15-15-00134 (to A.V.L.; this grant supported the work on phage 9g and preparation of its DNA and A.V.L.’s work on data analysis and writing), and the National Research Foundation of Singapore through its Singapore-MIT Alliance for Research and Technology (P.C.D.).


