D274–D281 Nucleic Acids Research, 2021, Vol. 49, Database issue Published online 9 November 2020
doi: 10.1093/nar/gkaa1018

COG database update: focus on microbial diversity, model organisms, and widespread pathogens

Michael Y. Galperin*, Yuri I. Wolf, Kira S. Makarova, Roberto Vera Alvarez, David Landsman and Eugene V. Koonin*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

Received September 10, 2020; Revised October 13, 2020; Editorial Decision October 14, 2020; Accepted October 14, 2020

ABSTRACT

The Clusters of Orthologous Genes (COG) database, also referred to as the Clusters of Orthologous Groups of proteins, was created in 1997 and went through several rounds of updates, most recently, in 2014. The current update, available at https://www.ncbi.nlm.nih.gov/research/COG, substantially expands the scope of the database to include complete genomes of 1187 bacteria and 122 archaea, typically, with a single genome per genus. In addition, the current version of the COGs includes the following new features: (i) the recently deprecated NCBI's gene index (gi) numbers for the encoded proteins are replaced with stable RefSeq or GenBank\ENA\DDBJ coding sequence (CDS) accession numbers; (ii) COG annotations are updated for >200 newly characterized protein families with corresponding references and PDB links, where available; (iii) lists of COGs grouped by pathways and functional systems are added; (iv) 266 new COGs for proteins involved in CRISPR-Cas immunity, sporulation in Firmicutes and photosynthesis in cyanobacteria are included; and (v) the database is made available as a web page, in addition to FTP. The current release includes 4877 COGs. Future plans include further expansion of the COG collection by adding archaeal COGs (arCOGs), splitting the COGs containing multiple paralogs, and continued refinement of COG annotations.

INTRODUCTION

For the past 24 years, the Clusters of Orthologous Genes (COGs) database, also known as Clusters of Orthologous Groups of proteins, has been a popular tool for functional and comparative genomics of bacteria and archaea (1–4). Its relatively small collection of fewer than 5000 clusters of orthologous proteins (COGs) consists of the products of the most widespread bacterial and archaeal genes. These include, for example, ribosomal proteins, universal translation factors, aminoacyl-tRNA synthetases, subunits of the RNA polymerase, F-type and A/V-type ATP synthases, as well as many key metabolic and signaling enzymes.

In all these cases, identification of orthologs by cross-genome comparisons, using the COG-making algorithm to create COGs and the COGMaker method to add COG members from new genomes (5–7) allowed unequivocal functional assignments for proteins from diverse genomes, many of which have never been studied experimentally. These assignments reveal the advantages of functionally annotating a protein family as a whole rather than its individual members. The COG annotations including COG names were deliberately chosen to reflect the functional diversity among the members of the respective COG. Those COGs that included experimentally characterized members with substantially different functions were given composite names to reflect this functional diversity (2–4).

The 2014 release of the COGs provided a major update of the COG names and expanded the coverage to 711 genomes representing all bacterial and archaeal genera that had at least one member with a completely sequenced genome by the end of 2013 (3). Conversely, the three eukaryotic genomes (those of the baker's yeast *Saccharomyces cerevisiae*, fission yeast *Schizosaccharomyces pombe* and the microsporidian *Encephalitozoon cuniculi*) that had been included in the earlier versions of the COGs were discarded because any attempt to incorporate the vastly increased genomic diversity of eukaryotes was deemed to be outside the scope of the project.

In the new COG release described here, we substantially expand the genome collection covered by the COGs, update the annotation of many COGs, and start adding new COGs, at this time, only a limited number of clusters of orthologs that are linked to certain functionalities and have been described in detail in our recent publications. We also restored the lists of the key pathways and functional groups covered by the COGs, a functionality that was discarded in the

*To whom correspondence should be addressed. Tel:+1 301 435 5910; Email: galperin@ncbi.nlm.nih.gov
Correspondence may also be addressed to Eugene V. Koonin. Tel: +1 301 435 5913; Email: koonin@ncbi.nlm.nih.gov

Published by Oxford University Press on behalf of Nucleic Acids Research 2020.

previous version. Since the time of the latest COG release, the NCBI has shifted from gene index (gi) numbers for the encoded proteins to GenBank accessions and non-unique RefSeq identifiers, which resulted in a large number of broken links. Accordingly, an important technical task for this COG update was replacing the gi numbers with stable RefSeq or GenBank\ENA\DDBJ coding sequence (CDS) accession numbers. We expect that these amendments substantially increase the utility of the COGs for prokaryote genome analysis and annotation.

KEY CHANGES IN THE 2020 COG UPDATE

Expanded genome coverage

In the current update, the genome coverage of the COG database has been expanded to include representatives of all bacterial and archaeal genera that had complete genome sequences released by 1 April 2019, and a selection of genomes released after that date. Given the rapid progress in microbial genome sequencing, covering all sequenced genomes was clearly untenable: even excluding thousands of unfinished (draft) genomes left >16 000 (as of the start of 2019) complete genomes, as specified in the NCBI descriptions of the respective sequence assemblies. Therefore, selection of the organisms for inclusion in the COGs presented some challenges. To provide reasonable coverage of microbial diversity while keeping the number of organisms manageable, we adopted the following approach.

First, we kept most of the 711 organisms from the previous COG release, removing only some of the duplicate members of the same bacterial genus (*Bradyrhizobium*, *Frankia*, *Granulicella*, *Rickettsia*, *Spiribacter*) and keeping other bacterial (*Bacillus*, *Clostridium*, *Escherichia*, *Mycobacterium*, *Mycoplasma*, *Nostoc*, *Streptococcus*) and archaeal (*Pyrococcus*, *Thermoplasma*) genera with two or more members. In the case of the two former *Sulfolobus* species, one of these, *Sulfolobus solfataricus*, has been reassigned to the new genus *Saccharolobus* (8). Name changes have been also registered for 42 other organisms from the 2014 COG list (Supplementary Table S1), including some widely studied model organisms and widespread pathogens. These included reclassification of the Lyme disease-causing *Borrelia burgdorferi* and some other *Borrelia* spp. to the new genus *Borreliella* (9), and of the former *Clostridium* (*Peptoclostridium*) *difficile* to *Clostridioides difficile* (10; Table 1).

Second, we added representatives of newly sequenced bacterial and archaeal genera, including those with the *Candidatus* status. This amendment resulted in the addition of 507 bacterial and 36 archaeal genera. When there were several newly sequenced genomes from the same genus, the choice was typically made in favor of either (i) an organism included in the NCBI representative genome collection (https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/#representative_genomes), (ii) the better-studied organism (judging by the number of PubMed references) or (iii) the organism with the largest genome (and, hence, the largest number of encoded proteins). The increase in the number and diversity of complete prokaryote genomes occurred, primarily, via expansion of the already well-sampled phyla, such as Euryarchaeota (from 54 to 76 genera), Actinobacteria (from 73 to 149), Bacteroidetes (from 55 to 107), Firmicutes (from 73 to 166), and Proteobacteria (from 265 to 538 genera; see Supplementary Table S2). To further improve the coverage of the microbial diversity, we also added representatives of some incompletely characterized lineages, mostly, at the higher taxonomic levels, such as the DPANN superphylum of archaea, with examples listed in Table 2.

Table 1. Changes in the names of widely used model organisms from the 2014 COGs

Organism name in the 2014 release	Updated organism name
Bacillus halodurans C-125	*Alkalihalobacillus halodurans* C-125
Borrelia burgdorferi B31	*Borreliella burgdorferi* B31
Burkholderia xenovorans LB400	*Paraburkholderia xenovorans* LB400
Caulobacter crescentus CB15	*Caulobacter vibrioides* CB15
Chlamydophila pneumoniae CWL029	*Chlamydia pneumoniae* CWL029
Chlorobium tepidum TLS	*Chlorobaculum tepidum* TLS
Clostridium difficile 630	*Clostridioides difficile* 630
Klebsiella pneumoniae 342	*Klebsiella variicola* 342
Mesorhizobium loti MAFF 303099	*Mesorhizobium japonicum* MAFF 303099
Methanosaeta concilii GP6	*Methanothrix soehngenii* GP6
Planctomyces brasiliensis DSM 5305	*Rubinisphaera brasiliensis* DSM 5305
Sulfolobus solfataricus P2	*Saccharolobus solfataricus* P2

Finally, for some bacterial and archaeal genera that were already represented in the COG database, we added selected widely studied members, primarily, either popular model organisms or important pathogens (Supplementary Table S3). For example, although the previous releases of the COGs already included *Mycobacterium tuberculosis* and *Mycobacterium leprae*, in this version, we added *Mycobacterium avium*, *Mycobacterium marinum* and *Mycobacterium ulcerans*. In addition, the new version includes two former *Mycobacterium* spp., recently transferred into novel genera: *Mycolicibacterium smegmatis* and *Mycobacteroides abscessus* (11). The genomes from *Mycolicibacillus* and *Mycolicibacter*, two other genera that have been separated from the original *Mycobacterium* spp. (11), did not make the list. The *Streptomyces* genus was originally represented by *Streptomyces coelicolor*, which was later replaced by the much larger genome of *S. bingchenggensis*. In the current release, we restored *S. coelicolor* and added *S. griseus*, which was chosen over *S. avermitilis*, *S. cattleya* and *S. venezuelae* as the strain used for much more research. Likewise, we added three species of *Lactobacillus*, two species each of *Staphylococcus*, *Streptococcus*, *Clostridium* and *Klebsiella*, as well as important pathogens such as *Vibrio parahaemolyticus, Yersinia enterocolitica, Shigella flexneri* and *Leptospira interrogans* (Supplementary Table S3). The complete list of 1309 genomes included in the current release is presented in the Supplementary Table S6 and is also available on the NCBI FTP site at https://ftp.ncbi.nih.gov/pub/COG/COG2020/data/cog-20.org.csv.

Functional annotation of COGs

After the massive overhaul of the COG annotations in the 2014 release, the current version underwent smaller changes. Nevertheless, this release includes >250 COGs with substantial updates in functional annotation and

Table 2. Representatives of poorly characterized prokaryotic lineages in COGs

Organism name[a]	Taxonomy (phylum)	GenBank accession
ARCHAEA		
Ca. Mancarchaeum acidiphilum Mia14	DPANN group, Cand. Micrarchaeota	CP019964
Ca. Nanopusillus acidilobi	DPANN group, Nanoarchaeota	CP010514
archaeon GW2011_AR15	Unclassified archaea	CP010425
BACTERIA		
Armatimonadetes bacterium Uphvl-Ar1	Armatimonadetes	CP021423
Ca. Cyclonatronum proteinivorum	Balneolaeota	CP027806
Cand. division Kazan bacterium GW2011_GWA1_50_15	Cand. division Kazan-3B-28	CP011216
Cand. division SR1 bacterium Aalborg_AAW-1	Cand. division SR1	CP011268
Ca. Beckwithbacteria bacterium GW2011_GWC1_49_16	*Ca.* Beckwithbacteria	CP011210
Berkelbacteria bacterium GW2011_GWE1_39_12	*Ca.* Berkelbacteria	CP011213
Ca. Bipolaricaulis anaerobius	*Ca.* Bipolaricaulota	LS483254
Ca. Campbellbacteria bacterium GW2011_OD1_34_28	*Ca.* Campbellbacteria	CP011215
Ca. Babela massiliensis	*Ca.* Dependentiae	HG793133
Cand. division TM6 bacterium GW2011_GWF2_28_16	*Ca.* Dependentiae	CP011212
Ca. Dependentiae bacterium (ex *Spumella elongata* CCAP 955/1)	*Ca.* Dependentiae	CP025544
Ca. Gracilibacteria bacterium HOT-871	*Ca.* Gracilibacteria	CP017714
Ca. Peribacter riflensis	*Ca.* Peregrinibacteria	CP013062
Ca. Saccharibacteria bacterium YM_S32_TM7_50_20	*Ca.* Saccharibacteria	CP025011
Ca. Saccharibacteria oral taxon TM7x	*Ca.* Saccharibacteria	CP007496
Ca. Woesebacteria bacterium GW2011_GWF1_31_35	*Ca.* Woesebacteria	CP011214
Ca. Wolfebacteria bacterium GW2011_GWB1_47_1	*Ca.* Wolfebacteria	CP011209
bacterium AB1	unclassified Bacteria	CP017117
Vampirococcus sp. LiM	unclassified Bacteria	CP019384

[a]- *Ca.* stands for *Candidatus*, Cand. – for Candidate.

about the same number of COGs with minor name changes. In addition, annotations of approximately 60 COGs were updated to include protein domains from Pfam, InterPro or the NCBI's Conserved Domain Database (12–14). We also compared COG annotations to those in TIGRfam and SubtiWiki databases (15,16). The notable updates in the functional annotation include enzymes involved in the modification of 16S rRNA, 23S rRNA, and tRNA (17), molybdenum cofactor biosynthesis (18), cyclic nucleotide signaling (19) and biogenesis of the cell envelope (20). A list of some of the major changes in the COG annotation introduced in the current release is given in Supplementary Table S4. The annotation process also identified several obsolete entries in the UniProt and Pfam databases, which have been reported to the curators of these databases.

Updated COG web pages

The current release introduces several changes in the setup of the COG page compared to the one in the 2014 release. In addition to the COG name and the one-letter symbol(s) for the functional category(ies) corresponding to that COG, the top line now includes four more cells. The first of these shows the short—typically, gene-based—symbol for the COG, like those used previously to denote COGs in the NCBI's Conserved Domain Database (14). The second cell contains the name of the metabolic pathway and/or functional ensemble, if available, that the respective COG is assigned to. For the newly renamed COGs, two additional new cells contain links, respectively, to the reference(s) in PubMed that served as the basis for the new name and to a representative crystal structure in the PDB, where available.

As mentioned above, the increase in the number and diversity of the complete prokaryotic genomes occurred primarily via expansion of the already well-sampled taxa, which allowed us to retain the same arrangement of the key phyla. Based on the increased number of sequenced genomes, there are now separate sections for Deferribacteres (5 species), Negativicutes (10 species), Tissierellia (9 species), other gammaproteobacteria (6 species), and Verrucomicrobia (9 species).

In the previous releases of the COG database, gene entries in the COGs were denoted by either their GenInfo (gi) numbers or genomic locus tags in the NCBI protein database. The recent phasing out of the gi numbers (described in the NCBI Insights, see https://ncbiinsights.ncbi.nlm.nih.gov/2016/07/15/) and the shift of the entire database to GenBank protein accessions and non-unique RefSeq identifiers, coupled with the removal of numerous proteins from the database ((21,22), see https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/faq/#FAQ5), resulted in many broken links and forced us to adopt the same identifiers for COG proteins. As a result, proteins in COGs are now listed, primarily, using their RefSeq identifiers. Proteins from 10 popular model organisms ('Reference organisms' in RefSeq), such as *Escherichia coli* str. K-12 substr. MG1655, *B. subtilis* subsp. *subtilis* str. 168, *Pseudomonas aeruginosa* PAO1, and *Salmonella enterica* subsp. *enterica* serovar Typhimurium str. LT2, retain their original NP_xxxxxx or YP_xxxxxxxxx identifiers that refer to a respective (single) genome. Proteins from other organisms, however, are listed under WP_xxxxxxxxx identifiers, which usually refer to many (identical) proteins from multiple genomes, sometimes even coming from diverse taxa. Accordingly, the number of protein IDs in the database is now smaller than the number of genome loci that code for these proteins. WP_xxxxxxxxx identifiers are still linked to the entries in the NCBI protein database but finding specific proteins encoded in specific genomes requires examining the 'Identical Proteins' link, which takes the user to the Identical Protein

Groups database (https://www.ncbi.nlm.nih.gov/ipg/). The proteins that are missing in RefSeq are listed under their GenBank\ENA\DDBJ coding sequence (CDS) accession numbers.

The protein IDs are shown in full, even for multi-domain proteins where the COG covers only a part of the CDS. Accordingly, a single protein can show up in more than one COG. To determine whether the COG covers the entire protein or only a part of it, the user can simply click on the protein entry and/or check the domain organization of the respective protein via the CDD link. Mapping of the entire protein set to the COGs can be found at https://ftp.ncbi.nih.gov/pub/COG/COG2020/data/cog-20.org.csv.

Pathways and functional ensembles

The current release once again includes lists of COGs assigned to key metabolic pathways and functional assemblies. The metabolic pathways include glycolysis, pentose phosphate pathway, the TCA cycle, biosynthetic pathways for most amino acids and enzyme cofactors, biosynthesis and salvage pathways for purines and pyrimidines, and selected pathways of the lipid and murein biosynthesis. Some of these pathways are accompanied by charts prepared for the previous COG releases (23).

The functional ensembles include proteins of the large and small subunits of bacterial ribosomes, archaea-specific ribosomal proteins, enzymes catalyzing 16S rRNA, 23S rRNA, and tRNA modification and several enzymatic complexes involved in energy transformation, such as NADH dehydrogenase (Complex I), Na^+-translocating NADH:ubiquinone oxidoreductase (NQR), Na^+-translocating ferredoxin:NAD^+ oxidoreductase (RNF), F_oF_1-type ATP synthase and the A/V-type ATPase. This group also includes a set of the common CRISPR-associated (Cas) proteins that was expanded from 26 to 46 COGs (see below). We believe that these functional groupings of COGs can be useful for comparative studies and plan to further expand them.

Addition of new photosynthetic and sporulation COGs

In the past, the COG approach has been used to delineate clusters of orthologous genes for cyanobacteria (24,25), lactobacilli (26), spore-forming firmicutes (27,28) and archaea (7,29,30). Making use of those data, we expanded the COG database by adding 118 COGs for proteins involved in photosynthesis and 125 COGs for sporulation proteins. Detailed analysis and annotation of cyanobacterial and firmicute COGs showed that most of them overlapped with the existing Pfam and/or TIGRfam families. However, the inclusion of these protein clusters into the COGs allowed a new outlook at the phylogenetic distribution of the respective proteins, highlighting those organisms (and lineages) that lack the respective proteins. Some of the phyletic patterns were as expected, such as the absence of the entire photosystem II in the tiny genome of the nitrogen-fixing cyanobacterium *Candidatus* Atelocyanobacterium thalassa (formerly cyanobacterium UCYN-A; 31). This organism was also the only cyanobacterium that lacked several poorly characterized photosynthetic proteins, including orthologs of TPR-like protein Orf03 (At3g26580 in *Arabidopsis thaliana*) and Acclimation of photosynthesis to environment protein APE1 (At5g38660 in *A. thaliana*). This analysis also confirmed the previous report that *Gloeobacter violaceus* PCC 7421 does not encode photosystem I proteins PsaI, PsaJ, PsaK and PsaX and photosystem II proteins PsbQ, PsbY, PsbZ and Psb27 (32). In accordance with the previous predictions (24), *Candidatus* Melainabacteria bacterium MEL.A1, an early branching member of the Cyanobacteria/Melainabacteria lineage, was found not to encode a single gene of photosystem I or II components (33,34).

Other phyletic patterns revealed the absence of a single gene in otherwise complete photosynthetic machinery, perhaps, as a result of sequencing or annotation problems. Indeed, photosynthetic reaction complexes include some short proteins (PsaJ, 38 aa; PsaM, 31 aa; PsbL, 37–40 aa; PsbT = Ycf8, 32 aa; Psb30 = Ycf12, 39 aa) that could be easily missed by automated genome annotation. Identification of unexpected gaps in the phyletic profiles allowed us to identify several of such missed proteins (Ycf12 in *Prochlorococcus marinus*, PsaJ in *Anabaena cylindrica*, PsbL in *Fischerella* sp. NIES-3754 and *Planktothrix agardhii*) that were absent from the protein database but are actually encoded in the respective genomes. These missing proteins were identified by tBLASTn (35) search against the respective genomic sequences, translated, and reported to RefSeq. Once included in the NCBI protein database, these genes will be added to the respective COGs.

Likewise, the 26-aa forespore membrane curvature-sensing protein SpoVM has been often overlooked in firmicute genome annotation (27,36). However, examination of the phyletic pattern of SpoVM (COG5844) showed that it was actually missing in some spore-former genomes, in keeping with its apparent non-essentiality in Clostridia (37). In accordance with previous analyses (27,38), we identified a wide variety of non-spore-forming firmicutes (mostly, among Clostridia) that nevertheless encode the master regulator of sporulation Spo0A. Examination of the phyletic patterns of Spo0A-encoding asporogens showed that most of these organisms missed at least some core sporulation genes. We expect the patterns of presence and absence of photosynthetic and sporulation genes to be useful in further studies of these important processes.

The COG approach also allowed us to create separate COGs for certain widespread domain combinations that, although clearly visible in Pfam via its ‘Domain architectures’ tool, have not been so far analyzed in any detail. As an example, the CP12 protein (COG5767) regulates the Calvin cycle in cyanobacteria and chloroplasts by forming a complex with GAPDH and phosphoribulokinase (39,40). A cyanobacteria-specific fusion of CP12 with an N-terminal tandem of CBS domains forms a separate COG5792; the members of this COG appear to regulate phosphoribulokinase in a completely different manner (41). Likewise, cyanobacteria, similarly to plants, encode multiple copies of the GUN4 (= Ycf53) domain that was initially characterized in *A. thaliana GENOMES UNCOUPLED*4 mutant and has been shown to, first, regulate chlorophyll biosynthesis (by activating the Mg-chelatase), and second, to participate in plastid-to-nucleus signaling (42,43). Most proteins containing the GUN4 domain (Pfam domain PF05419) were included in COG5750. However, two widespread domain combinations of GUN4 were assigned to separate COGs. One, COG5751 that is represented in almost all cyanobacteria, consists of proteins in which the GUN4 domain is fused with an N-terminal α-helical domain (43,44). The other, COG5752, features a fusion of GUN4 with an N-terminal Ser/Thr/Tyr protein kinase domain that has been described previously (43) but never experimentally characterized. This distinct domain combination is found in up to four copies in certain cyanobacterial genomes and is, most likely, involved in an unknown signal transduction pathway.

The expanded set of Cas COGs

The recent exhaustive comparative analyses of the CRISPR-Cas systems and associated genes (45–47) allowed us to substantially expand the set of COGs that consist of Cas proteins, from 26 to 46. However, COG3512 was eliminated, and its members were moved into the other COG that includes Cas2 proteins, COG1343. The current set of Cas COGs includes all the major groups of Cas proteins, such as Cas1, Cas2, Cas3, Cas4, Cas5, Cas6, Cas7; several distinct families of Cas8; Cas10 and other proteins of the Cmr and Csm complexes of type III CRISPR-Cas systems; two major families of class 2 effectors, Cas9 and Cas12a, as well as two COGs that include CARF (CRISPR-associated Rossmann fold) domain-containing proteins, Csx1 and Csm6 (48).

USING THE COGs

Although there are several other databases of orthologous proteins in bacteria and archaea, such as eggNOG (49), KEGG Orthology (50,51), OMA (52,53), OrthoDB (54) and MBGD (55), which cover a wider selection of organisms than the COGs and provide comprehensive, automated functional annotation of their proteins, as well as specialized microbial genome databases with more narrow genome coverage (56,57), the COG database has several unique features. These include (i) classification of COGs into functional categories that allows an easy comparison of organisms based on their preference for certain types of metabolic, signal transduction, repair and other pathways; (ii) unification in a single COG of orthologous proteins from distant phylogenetic lineages, in many cases, with low (but significant) sequence similarity to each other, and (iii) phyletic patterns of presence–absence of (proteins from) the compared genomes in a given COG that facilitate functional annotation of new genomes and evolutionary inference. To our knowledge, the COG database remains the only tool that shows not only which proteins (protein families) are encoded in the given genome, but also which proteins (families) are missing in it because they are either not properly annotated (e.g. erroneously marked as pseudogenes) or not encoded in the genome at all (1,3,4,58). This makes the COGs a powerful tool for comparative and evolutionary analysis of prokaryote genomes.

As described previously (3,4,58), the most straightforward application of the COGs is the quality control of genome sequences. A set of core COGs, such as those for the key ribosomal proteins, is expected to be represented in every sequenced genome. The absence of such a COG member in a newly sequenced genome, unless it comes from an obligate intracellular parasite or symbiont with a highly reduced genome or has been lost in an entire lineage (59,60), could be a warning sign that a certain fraction of the genomic DNA is likely to be missing from the assembly. We detected several such missing ribosomal proteins, of which some simply were not properly translated, whereas others were not encoded in the available genome assemblies (see Supplementary Table S5 for examples).

Another use of the COGs involves examination of the consistency of the phyletic patterns, indicating the presence or absence of the respective protein families, among the COGs that belong to the same pathway or the same multi-subunit complex. Although not all the COGs in the same pathway or functional system are bound to have identical phyletic patterns, any deviations from the uniform pattern might indicate non-trivial evolutionary events and merit further investigation. An interesting example is the mitochondrial protein NDUFA12, a supernumerary (auxiliary) subunit A12 of the NADH:ubiquinone oxidoreductase (mitochondrial respiratory Complex I) that is required for the assembly of the complex and whose deficiency causes the Leigh syndrome (61,62). The core subunits of the NADH:ubiquinone oxidoreductase are widespread in bacteria and archaea (63), but NDUFA12 (COG3761) is represented exclusively in alphaproteobacteria, which is consistent with the alphaproteobacterial origin of the mitochondria. This gene is found in 148 of the 158 alphaproteobacterial genomes in the COGs, missing mostly in some (but not all) representatives of the family *Acetobacteraceae*. This subunit appears to have evolved within alphaproteobacteria and was inherited by the ancestral mitochondria, followed by the transfer of the gene into the nucleus. Meanwhile, this gene has been lost in several acetic acid bacteria that inhabit extremely nutrient-rich environments and might not require participation of NDUFA12 in the assembly of their Complex I.

Another notable example is the distribution of the c-di-AMP signaling systems. Recent analyses have identified c-di-AMP synthases and hydrolases in euryarchaea and in members of almost every bacterial phylum (19,64). However, certain members of the phylum Actinobacteria have been found to encode a typical DisA-type c-di-AMP synthase, but no (known) c-di-AMP-specific phosphodiesterase (19). If true, this would suggest unencumbered accumulation of c-di-AMP, which would be toxic for the cell (65). In short order, the missing c-di-AMP phosphodiesterase has been identified, characterized, assigned to a previously uncharacterized family (COG1524) within the alkaline phosphatase superfamily, and shown to be widespread among Actinobacteria (66,67). These examples show that analysis of identical, complementary, and mixed phylogenetic patterns can facilitate solving important biological problems.

The phyletic pattern of each specific COG can be viewed at and downloaded directly from the respective COG page, which shows the total number of organisms, genes, and proteins covered by that COG, as well as the representation of organisms from each phylogenetic group (Supplementary Figure 1). The list of proteins included in each particular COG can also be downloaded directly from the respective COG page in JSON format (in groups of 10) or extracted from the COG master file, https://ftp.ncbi.nih.gov/pub/COG/COG2020/data/cog-20.org.csv. This list of proteins can also be imported into Excel to highlight the organisms that are represented—and not represented—in the given COG. A comparison of the phyletic distribution of the COGs that belong to the same multisubunit enzyme, a metabolic pathway or a functional ensemble can be used for evaluating the phylogenetic distribution of the respective ensemble of COGs and for identifying deviations from the common pattern that could point either to problems with the annotation of certain genomes, non-orthologous gene displacement or unexpected loss of seemingly essential genes. Each of these situations merits further investigation. As a case in point, Supplementary File S2 contains an ordered list of the organisms in the current COG release (Supplementary Table S6) and a comparison (Supplementary Table S7) of the COGs representing six subunits of the Na^+-translocating NADH:quinone oxidoreductase (Na^+-NQR), a drug target for the recently developed new furanone antibiotic (68,69). This comparison can be used to delineate potential targets of this new antibiotic, which include such pathogens as *Chlamydia pneumoniae*, *Chlamydia trachomatis*, *Klebsiella aerogenes*, *Neisseria gonorrhoeae*, *Neisseria meningitidis*, *Pasteurella multocida*, *Porphyromonas gingivalis*, *Vibrio cholerae*, *Vibrio parahaemolyticus*, *Yersinia enterocolitica* and *Yersinia pestis* (Supplementary Table S7).

DATABASE CURATION AND FUTURE DIRECTIONS

The COG database is undergoing continued manual curation. COG annotations are regularly updated based on newly published functional data for the COG members. COG membership is also being updated, based, primarily, on the respective phyletic patterns. As an example, a recent examination of the phyletic patterns of the COGs for 50S and 30S ribosomal proteins has led to the identification, along with ∼500 ORFs genuinely missing in the respective genomes (59,60), of a variety of missed proteins, some with highly diverged sequences, others encoded by truncated or frameshifted ORFs, but also >70 full-size ORFs that had been overlooked in the course of genome annotation. Once these proteins are included in GenBank and/or RefSeq, they will be added to the respective COGs.

Further development of the COG database will be based, mostly, on the research projects carried out by the members of our team. Future plans include further expansion of the COG collection by adding archaeal COGs (arCOGs), splitting the COGs containing multiple paralogs, and continued refinement of COG annotations.

COG STATISTICS

The current version of COGs includes 1309 genomes from 1187 bacterial and 122 archaeal species that represent 1234 named genera, typically, with a single genome per genus. This is an almost two-fold increase from the 583 bacterial and 70 archaeal genomes that were represented in the 2014 COG release.

The database consists of 4877 COGs that include 3 236 575 unique genes mapped to 3 455 867 genomic loci, which represents 73.5% of the 4 401 819 proteins (4 110 746 bacterial and 291 073 archaeal proteins) encoded by the covered species.

DATA AVAILABILITY

The new version of the COGs is publicly available at https://www.ncbi.nlm.nih.gov/research/cog with service files available on the NCBI FTP site at https://ftp.ncbi.nih.gov/pub/COG/COG2020/. The previous versions of the COG database are available at https://ftp.ncbi.nih.gov/pub/COG/. All queries, comments, and suggestions for improvement regarding the COG database should be directed to the authors at cogs@ncbi.nlm.nih.gov.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Intramural Research Program of the U.S. National Library of Medicine at the National Institutes of Health. Funding for open access charge: Intramural Research Program of the U.S. National Library of Medicine at the National Institutes of Health.
Conflict of interest statement. None declared.

This paper is linked to: doi:10.1093/nar/gkaa1009.

REFERENCES

1. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
2. Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.
3. Galperin,M.Y., Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, **43**, D261–D269.
4. Galperin,M.Y., Kristensen,D.M., Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2019) Microbial genome analysis: the COG approach. *Brief. Bioinform.*, **20**, 1063–1070.
5. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
6. Kristensen,D.M., Kannan,L., Coleman,M.K., Wolf,Y.I., Sorokin,A., Koonin,E.V. and Mushegian,A. (2010) A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics*, **26**, 1481–1487.
7. Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2015) Archaeal Clusters of Orthologous Genes (arCOGs): An update and application for analysis of shared features between *Thermococcales*, *Methanococcales*, and *Methanobacteriales*. *Life (Basel)*, **5**, 818–840.
8. Sakai,H.D. and Kurosawa,N. (2018) *Saccharolobus caldissimus* gen. nov., sp. nov., a facultatively anaerobic iron-reducing hyperthermophilic archaeon isolated from an acidic terrestrial hot spring, and reclassification of *Sulfolobus solfataricus* as *Saccharolobus solfataricus* comb. nov. and *Sulfolobus shibatae* as *Saccharolobus shibatae* comb. nov. *Int. J. Syst. Evol. Microbiol.*, **68**, 1271–1278.
9. Adeolu,M. and Gupta,R.S. (2014) A phylogenomic and molecular marker based proposal for the division of the genus *Borrelia* into two genera: the emended genus *Borrelia* containing only the members of

the relapsing fever *Borrelia*, and the genus *Borreliella* gen. nov. containing the members of the Lyme disease *Borrelia* (*Borrelia burgdorferi* sensu lato complex). *Antonie Van Leeuwenhoek*, **105**, 1049–1072.

10. Lawson,P.A., Citron,D.M., Tyrrell,K.L. and Finegold,S.M. (2016) Reclassification of *Clostridium difficile* as *Clostridioides difficile* (Hall and O'Toole 1935) Prevot 1938. *Anaerobe*, **40**, 95–99.
11. Gupta,R.S., Lo,B. and Son,J. (2018) Phylogenomics and comparative genomic studies robustly support division of the genus *Mycobacterium* into an emended genus *Mycobacterium* and four novel genera. *Front. Microbiol.*, **9**, 67.
12. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
13. Mitchell,A.L., Attwood,T.K., Babbitt,P.C., Blum,M., Bork,P., Bridge,A., Brown,S.D., Chang,H.Y., El-Gebali,S., Fraser,M.I. *et al.* (2019) InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.*, **47**, D351–D360.
14. Lu,S., Wang,J., Chitsaz,F., Derbyshire,M.K., Geer,R.C., Gonzales,N.R., Gwadz,M., Hurwitz,D.I., Marchler,G.H., Song,J.S. *et al.* (2020) CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.*, **48**, D265–D268.
15. Haft,D.H., Selengut,J.D., Richter,R.A., Harkins,D., Basu,M.K. and Beck,E. (2013) TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res.*, **41**, D387–D395.
16. Zhu,B. and Stülke,J. (2018) SubtiWiki in 2018: from genes and proteins to functional network annotation of the model organism *Bacillus subtilis*. *Nucleic Acids Res.*, **46**, D743–D748.
17. de Crécy-Lagard,V., Ross,R.L., Jaroch,M., Marchand,V., Eisenhart,C., Brégeon,D., Motorin,Y. and Limbach,P.A. (2020) Survey and validation of tRNA modifications and their corresponding genes in *Bacillus subtilis* sp. *subtilis* strain 168. *Biomolecules*, **10**, E977.
18. Leimkühler,S. (2020) The biosynthesis of the molybdenum cofactors in *Escherichia coli*. *Environ. Microbiol.*, **22**, 2007–2026.
19. He,J., Yin,W., Galperin,M.Y. and Chou,S.H. (2020) Cyclic di-AMP, a second messenger of primary importance: tertiary structures and binding mechanisms. *Nucleic Acids Res.*, **48**, 2807–2829.
20. Ekiert,D.C., Bhabha,G., Isom,G.L., Greenan,G., Ovchinnikov,S., Henderson,I.R., Cox,J.S. and Vale,R.D. (2017) Architectures of lipid transport systems for the bacterial outer membrane. *Cell*, **169**, 273–285.
21. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
22. Haft,D.H., DiCuccio,M., Badretdin,A., Brover,V., Chetvernin,V., O'Neill,K., Li,W., Chitsaz,F., Derbyshire,M.K., Gonzales,N.R. *et al.* (2018) RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.*, **46**, D851–D860.
23. Koonin,E.V. and Galperin,M.Y. (2003) *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics*. Kluwer Academic, Boston.
24. Mulkidjanian,A.Y., Koonin,E.V., Makarova,K.S., Mekhedov,S.L., Sorokin,A., Wolf,Y.I., Dufresne,A., Partensky,F., Burd,H., Kaznadzey,D. *et al.* (2006) The cyanobacterial genome core and the origin of photosynthesis. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 13126–13131.
25. Mulkidjanian,A.Y. and Galperin,M.Y. (2013) A time to scatter genes and a time to gather them. Evolution of photosynthesis genes in bacteria. *Adv. Bot. Res.*, **65**, 1–35.
26. Makarova,K.S. and Koonin,E.V. (2007) Evolutionary genomics of lactic acid bacteria. *J. Bacteriol.*, **189**, 1199–1208.
27. Galperin,M.Y., Mekhedov,S.L., Puigbo,P., Smirnov,S., Wolf,Y.I. and Rigden,D.J. (2012) Genomic determinants of sporulation in *Bacilli* and *Clostridia*: towards the minimal set of sporulation-specific genes. *Environ. Microbiol.*, **14**, 2870–2890.
28. Yutin,N. and Galperin,M.Y. (2013) A genomic update on clostridial phylogeny: Gram-negative spore formers and other misplaced clostridia. *Environ. Microbiol.*, **15**, 2631–2641.
29. Makarova,K.S., Sorokin,A.V., Novichkov,P.S., Wolf,Y.I. and Koonin,E.V. (2007) Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol. Direct.*, **2**, 33.
30. Wolf,Y.I., Makarova,K.S., Yutin,N. and Koonin,E.V. (2012) Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. *Biol. Direct*, **7**, 46.
31. Zehr,J.P., Bench,S.R., Carter,B.J., Hewson,I., Niazi,F., Shi,T., Tripp,H.J. and Affourtit,J.P. (2008) Globally distributed uncultivated oceanic N_2-fixing cyanobacteria lack oxygenic photosystem II. *Science*, **322**, 1110–1112.
32. Inoue,H., Tsuchiya,T., Satoh,S., Miyashita,H., Kaneko,T., Tabata,S., Tanaka,A. and Mimuro,M. (2004) Unique constitution of photosystem I with a novel subunit in the cyanobacterium *Gloeobacter violaceus* PCC 7421. *FEBS Lett.*, **578**, 275–279.
33. Di Rienzi,S.C., Sharon,I., Wrighton,K.C., Koren,O., Hug,L.A., Thomas,B.C., Goodrich,J.K., Bell,J.T., Spector,T.D., Banfield,J.F. *et al.* (2013) The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. *eLife*, **2**, e01102.
34. Soo,R.M., Skennerton,C.T., Sekiguchi,Y., Imelfort,M., Paech,S.J., Dennis,P.G., Steen,J.A., Parks,D.H., Tyson,G.W. and Hugenholtz,P. (2014) An expanded genomic representation of the phylum Cyanobacteria. *Genome Biol. Evol.*, **6**, 1031–1045.
35. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
36. Kim,E.Y., Tyndall,E.R., Huang,K.C., Tian,F. and Ramamurthi,K.S. (2017) Dash-and-Recruit mechanism drives membrane curvature recognition by the small bacterial protein SpoVM. *Cell Syst.*, **5**, 518–526.
37. Ribis,J.W., Ravichandran,P., Putnam,E.E., Pishdadian,K. and Shen,A. (2017) The conserved spore coat protein SpoVM is largely dispensable in *Clostridium difficile* spore formation. *mSphere*, **2**, e00315-17.
38. Abecasis,A.B., Serrano,M., Alves,R., Quintais,L., Pereira-Leal,J.B. and Henriques,A.O. (2013) A genomic signature and the identification of new sporulation genes. *J. Bacteriol.*, **195**, 2101–2115.
39. Marri,L., Trost,P., Pupillo,P. and Sparla,F. (2005) Reconstitution and properties of the recombinant glyceraldehyde-3-phosphate dehydrogenase/CP12/phosphoribulokinase supramolecular complex of *Arabidopsis*. *Plant Physiol.*, **139**, 1433–1443.
40. McFarlane,C.R., Shah,N.R., Kabasakal,B.V., Echeverria,B., Cotton,C.A.R., Bubeck,D. and Murray,J.W. (2019) Structural basis of light-induced redox regulation in the Calvin-Benson cycle in cyanobacteria. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 20984–20990.
41. Hackenberg,C., Hakanpaa,J., Cai,F., Antonyuk,S., Eigner,C., Meissner,S., Laitaoja,M., Janis,J., Kerfeld,C.A., Dittmann,E. *et al.* (2018) Structural and functional insights into the unique CBS-CP12 fusion protein family in cyanobacteria. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 7141–7146.
42. Larkin,R.M., Alonso,J.M., Ecker,J.R. and Chory,J. (2003) GUN4, a regulator of chlorophyll synthesis and intracellular signaling. *Science*, **299**, 902–906.
43. Davison,P.A., Schubert,H.L., Reid,J.D., Iorg,C.D., Heroux,A., Hill,C.P. and Hunter,C.N. (2005) Structural and biochemical characterization of Gun4 suggests a mechanism for its role in chlorophyll biosynthesis. *Biochemistry*, **44**, 7603–7612.
44. Verdecia,M.A., Larkin,R.M., Ferrer,J.L., Riek,R., Chory,J. and Noel,J.P. (2005) Structure of the Mg-chelatase cofactor GUN4 reveals a novel hand-shaped fold for porphyrin binding. *PLoS Biol.*, **3**, e151.
45. Makarova,K.S., Wolf,Y.I., Alkhnbashi,O.S., Costa,F., Shah,S.A., Saunders,S.J., Barrangou,R., Brouns,S.J., Charpentier,E., Haft,D.H. *et al.* (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.*, **13**, 722–736.
46. Shmakov,S.A., Makarova,K.S., Wolf,Y.I., Severinov,K.V. and Koonin,E.V. (2018) Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, E5307–E5316.
47. Makarova,K.S., Wolf,Y.I., Iranzo,J., Shmakov,S.A., Alkhnbashi,O.S., Brouns,S.J.J., Charpentier,E., Cheng,D., Haft,D.H., Horvath,P. *et al.* (2020) Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.*, **18**, 67–83.

48. Makarova,K.S., Timinskas,A., Wolf,Y.I., Gussow,A.B., Siksnys,V., Venclovas,C. and Koonin,E.V. (2020) Evolutionary and functional classification of the CARF domain superfamily, key sensors in prokaryotic antivirus defense. *Nucleic Acids Res.*, **48**, 8828–8847.
49. Huerta-Cepas,J., Szklarczyk,D., Heller,D., Hernandez-Plaza,A., Forslund,S.K., Cook,H., Mende,D.R., Letunic,I., Rattei,T., Jensen,L.J. *et al.* (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*, **47**, D309–D314.
50. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
51. Aramaki,T., Blanc-Mathieu,R., Endo,H., Ohkubo,K., Kanehisa,M., Goto,S. and Ogata,H. (2020) KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, **36**, 2251–2252.
52. Altenhoff,A.M., Glover,N.M., Train,C.M., Kaleb,K., Warwick Vesztrocy,A., Dylus,D., de Farias,T.M., Zile,K., Stevenson,C., Long,J. *et al.* (2018) The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res.*, **46**, D477–D485.
53. Altenhoff,A.M., Levy,J., Zarowiecki,M., Tomiczek,B., Warwick Vesztrocy,A., Dalquen,D.A., Muller,S., Telford,M.J., Glover,N.M., Dylus,D. *et al.* (2019) OMA standalone: orthology inference among public and custom genomes and transcriptomes. *Genome Res.*, **29**, 1152–1163.
54. Kriventseva,E.V., Kuznetsov,D., Tegenfeldt,F., Manni,M., Dias,R., Simao,F.A. and Zdobnov,E.M. (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.*, **47**, D807–D811.
55. Uchiyama,I., Mihara,M., Nishide,H., Chiba,H. and Kato,M. (2019) MBGD update 2018: microbial genome database based on hierarchical orthology relations covering closely related and distantly related comparisons. *Nucleic Acids Res.*, **47**, D382–D389.
56. Pillonel,T., Tagini,F., Bertelli,C. and Greub,G. (2020) ChlamDB: a comparative genomics database of the phylum *Chlamydiae* and other members of the *Planctomycetes-Verrucomicrobiae-Chlamydiae* superphylum. *Nucleic Acids Res.*, **48**, D526–D534.
57. Reyes-Prieto,M., Vargas-Chavez,C., Llabres,M., Palmer,P., Latorre,A. and Moya,A. (2020) An update on the Symbiotic Genomes Database (SymGenDB): a collection of metadata, genomic, genetic and protein sequences, orthologs and metabolic networks of symbiotic organisms. *Database (Oxford)*, **2020**, baz160.
58. Natale,D.A., Galperin,M.Y., Tatusov,R.L. and Koonin,E.V. (2000) Using the COG database to improve gene recognition in complete genomes. *Genetica*, **108**, 9–17.
59. Yutin,N., Puigbo,P., Koonin,E.V. and Wolf,Y.I. (2012) Phylogenomics of prokaryotic ribosomal proteins. *PLoS One*, **7**, e36972.
60. Nikolaeva,D.D., Gelfand,M.S. and Garushyants,S.K. (2020) Simplification of ribosomes in bacteria with tiny genomes. *Mol. Biol. Evol.*, doi:10.1093/molbev/msaa1184.
61. Ostergaard,E., Rodenburg,R.J., van den Brand,M., Thomsen,L.L., Duno,M., Batbayli,M., Wibrand,F. and Nijtmans,L. (2011) Respiratory chain complex I deficiency due to NDUFA12 mutations as a new cause of Leigh syndrome. *J. Med. Genet.*, **48**, 737–740.
62. Rak,M. and Rustin,P. (2014) Supernumerary subunits NDUFA3, NDUFA5 and NDUFA12 are required for the formation of the extramembrane arm of human mitochondrial complex I. *FEBS Lett.*, **588**, 1832–1838.
63. Novakovsky,G.E., Dibrova,D.V. and Mulkidjanian,A.Y. (2016) Phylogenomic analysis of type 1 NADH:quinone oxidoreductase. *Biochemistry (Mosc)*, **81**, 770–784.
64. Commichau,F.M., Heidemann,J.L., Ficner,R. and Stülke,J. (2019) Making and breaking of an essential poison: the cyclases and phosphodiesterases that produce and degrade the essential second messenger cyclic di-AMP in bacteria. *J. Bacteriol.*, **201**, e00462-18.
65. Gundlach,J., Mehne,F.M., Herzberg,C., Kampf,J., Valerius,O., Kaever,V. and Stülke,J. (2015) An essential poison: synthesis and degradation of cyclic di-AMP in *Bacillus subtilis*. *J. Bacteriol.*, **197**, 3265–3274.
66. Latoscha,A., Drexler,D.J., Al-Bassam,M.M., Bandera,A.M., Kaever,V., Findlay,K.C., Witte,G. and Tschowri,N. (2020) c-di-AMP hydrolysis by the phosphodiesterase AtaC promotes differentiation of multicellular bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 7392–7400.
67. Yin,W., Cai,X., Ma,H., Zhu,L., Zhang,Y., Chou,S.H., Galperin,M.Y. and He,J. (2020) A decade of research on the second messenger c-di-AMP. *FEMS Microbiol. Rev.*, doi:10.1093/femsre/fuaa1019.
68. Dibrov,P., Dibrov,E., Maddaford,T.G., Kenneth,M., Nelson,J., Resch,C. and Pierce,G.N. (2017) Development of a novel rationally designed antibiotic to inhibit a nontraditional bacterial target. *Can. J. Physiol. Pharmacol.*, **95**, 595–603.
69. Dibrov,P., Dibrov,E. and Pierce,G.N. (2017) Na^+-NQR (Na^+-translocating NADH:ubiquinone oxidoreductase) as a novel target for antibiotics. *FEMS Microbiol. Rev.*, **41**, 653–671.