Two Cobalt Chelatase Subunits Can Be Generated from a Single *chlD* Gene via Programed Frameshifting

Ivan V. Antonov*[,1,2]

[1]Institute of Bioengineering, Federal Research Centre Fundamentals of Biotechnology, Moscow, Russia

[2]Department of Biological and Medical Physics, Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia

***Corresponding author:** E-mail: ivan.antonov@gatech.edu.

Associate editor: Rebekah Rogers

Article

Abstract

Magnesium chelatase chlIDH and cobalt chelatase cobNST enzymes are required for biosynthesis of (bacterio)chlorophyll and cobalamin (vitamin B12), respectively. Each enzyme consists of large, medium, and small subunits. Structural and primary sequence similarities indicate common evolutionary origin of the corresponding subunits. It has been reported earlier that some of vitamin B12 synthesizing organisms utilized unusual cobalt chelatase enzyme consisting of a large cobalt chelatase subunit (*cobN*) along with a medium (*chlD*) and a small (*chlI*) subunits of magnesium chelatase. In attempt to understand the nature of this phenomenon, we analyzed >1,200 diverse genomes of cobalamin and/or chlorophyll producing prokaryotes. We found that, surprisingly, genomes of many cobalamin producers contained *cobN* and *chlD* genes only; a small subunit gene was absent. Further on, we have discovered a diverse group of *chlD* genes with functional programed ribosomal frameshifting signals. Given a high similarity between the small subunit and the N-terminal part of the medium subunit, we proposed that programed translational frameshifting may allow *chlD* mRNA to produce both subunits. Indeed, in genomes where genes for small subunits were absent, we observed statistically significant enrichment of programed frameshifting signals in *chlD* genes. Interestingly, the details of the frameshifting mechanisms producing small and medium subunits from a single *chlD* gene could be prokaryotic taxa specific. All over, this programed frameshifting phenomenon was observed to be highly conserved and present in both bacteria and archaea.

Key words: **programed ribosomal frameshifting, programed transcriptional realignment, magnesium chelatase chlIDH, cobalt chelatase cobNST, cobalamin (vitamin B12), chlorophyll.**

Introduction

A number of enzymes from all domains of life require a non-protein compound known as "cofactors" for their functioning. A special group of cofactors, the tetrapyrrole cofactors, are used to perform very different functions in living cells—from photosynthesis in plants to oxygen transfer in the human body. Still, the functional core of these molecules is formed by the tetrapyrrole structure with a metal ion coordinated in the middle (Dailey 2013). The type of the metal ion largely defines the function of the corresponding cofactor. For example, chlorophyll, cobalamin (vitamin B12), heme, and coenzyme F430 contain the magnesium, cobalt, iron, and nickel ions, respectively (Guilard et al. 2003). The insertion of the ion into tetrapyrrole is performed by specialized enzymes called chelatases. Each chelatase is usually specific to one metal ion and tetrapyrrole derivative (Willows and Hansson 2003).

In total, there are seven bacterial phyla which contain chlorophototrophs: *Acidobacteria*, *Chlorobi*, *Chloroflexi* (these three groups are frequently called green bacteria), *Proteobacteria* (commonly called purple bacteria), *Gemmatimonadetes*, *Cyanobacteria*, and *Firmicutes* (Thiel et al. 2018). There are several different types of bacteriochlorophylls, including the Zinc-containing bacteriochlorophylls, that can be found in various phototrophic bacteria (Wakao et al. 1996; Thweatt et al. 2019). On the other hand, chlorophyll is synthesized by *Cyanobacteria* as well as chloroplasts of algae and plants that are believed to have common evolutionary origin (Keeling 2004). The biosynthesis of Mg^{2+}-containing chlorophylls and bacteriochlorophylls requires the magnesium chelatase that performs the insertion of the magnesium ion (Mg^{2+}) into protoporphyrin IX (Walker and Willows 1997). This enzyme consists of small (I), medium (D), and large (H) subunits (Bollivar et al. 1994; Gibson et al. 1995). The corresponding genes are called *chlI*, *chlD*, and *chlH* in the genomes of chlorophyll-producing organisms and *bchI*, *bchD*, and *bchH* in bacteriochlorophyll producers. To reduce the number of different gene names in this work, we will use the *chlI*, *chlD*, and *chlH* terminology for both groups of genes.

Another chelatase, the aerobic cobalt chelatase cobNST, is highly similar to the magnesium chelatase and these two enzymes are believed to have common evolutionary origin (Fodje et al. 2001; Lundqvist et al. 2009). Cobalt chelatase is involved in de novo cobalamin (vitamin B12) biosynthesis and performs insertion of cobalt ion (Co^{2+}) into hydrogenobyrinic acid a,c-diamide (Raux et al. 2000; Martens et al. 2002).

Like the Mg-chelatase, it also consists of small (S), medium (T), and large (N) subunits that are encoded by the *cobS*, *cobT*, and *cobN* genes, respectively (Debussche et al. 1992). Moreover, some prokaryotic genomes encode the small and the medium subunits of the Co-chelatase using the corresponding Mg-chelatase genes, so their chimeric Co-chelatases are encoded by the *chlI*, *chlD*, and *cobN* genes (Rodionov et al. 2003).

Translational or programed ribosomal frameshifting (PRF) is a special type of recoding utilized for expression of some genes and can be found in all domains of life (Baranov et al. 2002a). On PRF, a ribosome changes the initial reading frame at a specific location in mRNA known as frameshifting signal. It usually consists of a "slippery site" (i.e., the point where the translating ribosomes actually change the reading frame) and the stimulatory element(s) around it that increase the frameshifting efficiency (Atkins et al. 2016). It should be noted that such signals are most frequently present in viral genes and their derivatives (such as mobile elements), but very rarely observed in the cellular genes (Sharma et al. 2011). In fact, there are only a handful of native prokaryotic genes (including *prfB* [Craigen and Caskey 1986], *dnaX* [Blinkova et al. 1997] and *copA* [Meydan et al. 2017]) that utilize this mechanism for their expression. In all these cases, there are specific biological functions associated with these signals. For example, the frameshifting signal in some genes (e.g., *dnaX*, *copA*, *gag-pol*) allows the synthesis of two different proteins (a short one and a long one) from the same mRNA. In such genes, the frameshifting efficiency (i.e., the fraction of the ribosomes that change the reading frame at the PRF signal) defines the ratio between the two products. In other cases (e.g., the *prfB* gene in *Escherichia coli* or the *OAZ1* gene [Ivanov and Atkins 2007] in the human genome), the PRF signal can function as a biosensor and provide the autoregulation of the gene expression at the translational level. These important biological functions impose evolutionary constraints on the components of the frameshifting signal (i.e., the slippery site and the stimulatory elements). Particularly, the conservation of the slippery site where the actual frameshifting occurs can be observed for all the known PRF signals. Thus, the ab initio frameshift prediction algorithm (Antonov and Borodovsky 2010) combined with the comparative genomics approaches have allowed us to identify several new gene families with functional PRF signals (Antonov, Baranov, et al. 2013).

Genes from one of these families were annotated as "magnesium chelatase." They contained putative PRF signals located near the predicted frameshift positions. The signals from one bacterial (*Delftia acidovorans SPH-1*) and two archaeal (*Methanocaldococcus fervens AG86* and *Methanocaldococcus* sp. *FS406-22*) genes have been experimentally validated in the *E. coli* model system and have shown the ability to efficiently (up to 60%) divert translation into the −1 alternative reading frame (Antonov, Coakley, et al. 2013). This suggested that the functioning of the frameshifting signals may allow the correction of the frameshift mutations and the generation of the full-length products from the corresponding genes. Thus, the goal of the present study was to determine the possible biological function of the conserved and highly efficient PRF signal in the "magnesium chelatase" genes.

Here, we showed that the identified genes with the validated PRF signals were more similar to the *chlD* genes that encoded the medium Mg-chelatase subunit. Importantly, it has been shown that the small chelatase subunit has similarity to the N-terminal part of the medium subunit. Consequently, it has been suggested that the *chlD* and *chlI* are paralogous genes and *chlI* originated via partial duplication of the *chlD* (Xiong et al. 1998). The frameshifting signals in the *chlD* genes were located at the end of the *chlD* region with putative homology to the *chlI* gene. Given the similarity between these two genes, we hypothesized that the translational frameshifting may allow the *chlD* mRNA to produce two functional proteins—the small and the medium chelatase subunits. The results obtained in the present study supported this theory. Various genotypes and molecular mechanisms that can be used by different prokaryotes to encode the Mg-/Co-chelatases were also considered in this work.

Results

The *chlD* Genes with the Validated Frameshfiting Signals May Encode Subunits of the Aerobic Cobalt Chelatase

Our first goal was to determine which Mg-chelatase subunit was encoded by the three frameshifted genes (from *D. acidovorans SPH-1*, *M. fervens AG86*, and *Methanocaldococcus* sp. *FS406-22*) containing the experimentally validated PRF signals (supplementary table 1, Supplementary Material online). The lengths of the full-length proteins (610 aa, 681 aa, and 679 aa, respectively) corresponded to the medium chelatase subunits (supplementary fig. 1, Supplementary Material online). To determine their relationship to the *chlD*, *bchD*, and *cobT* genes, we prepared a reference phylogenetic tree where the annotated genes formed three distinct branches (supplementary fig. 2*A*, Supplementary Material online). The products of the three genes with the validated frameshifting signals were more similar to the cyanobacterial ChlD proteins (supplementary fig. 2*B*, Supplementary Material online). Thus, in this work, we will refer to these frameshifted genes as "fs-*chlD*."

The product of the *chlD* gene is usually involved in biosynthesis of (bacterio)chlorophyll as a part of the Mg-chelatase. However, the living conditions of the three organisms containing the fs-*chlD* genes with validated frameshifting signals suggested that they were unable to obtain energy from photosynthesis. Indeed, *D. acidovorans SPH-1* is a soil bacteria (and a rare human pathogen; Bilgin et al. 2015) and *Methanocaldococcus* are deep-sea archaea living in hydrothermal vents (Mehta and Baross 2006).

To investigate whether the *D. acidovorans SPH-1*, *M. fervens AG86*, and *Methanocaldococcus* sp. *FS406-22* were able to synthesize (bacterio)chlorophyll, we compared the sets of the genes present in their genomes with the gene sets that can be found in the genomes of known bacterial phototrophs (proteobacteria *Rhodobacter sphaeroides* and cyanobacteria *Nostoc punctiforme*). This analysis revealed that the three

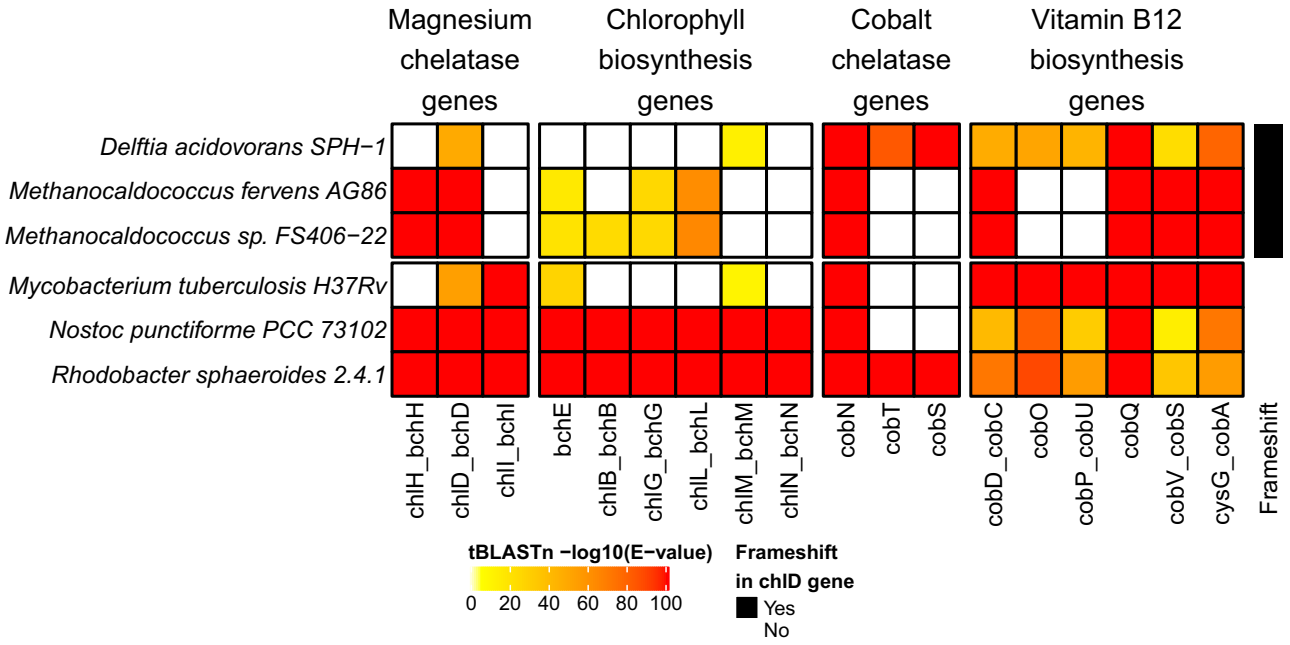

Fig. 1. Identification of the genes encoding magnesium and cobalt chelatase subunits as well as the other genes from the chlorophyll and cobalamin (vitamin B12) biosynthesis pathways. The color of the cells corresponds to the tBLASTn −log10(*E*-value) produced by a set reference proteins that were used as queries (see supplementary data, Supplementary Material online). White colors indicate the gene absence. The top three genomes contain frameshifted *chlD* genes. The bottom three organisms possess cobalamin biosynthesis pathway genes. Additionally, *Rhodobacter sphaeroides 2.4.1* and *Nostoc punctiforme PCC 73102* are able to synthesize (bacterio)chlorophyll as well.

genomes containing the frameshifted genes lacked the genes encoding the small Mg-chelatase subunit (*chlI* or *bchI*) as well as the other key genes required for de novo (bacterio)chlorophyll biosynthesis (fig. 1, left). Additionally, the *D. acidovorans SPH-1* genome did not have the *chlH* gene encoding the large subunit of the Mg-chelatase. This indicated that these three species were not able to synthesize (bacterio)chlorophyll. Thus, we concluded that the function of the frameshifted "magnesium chelatase" genes may currently be misannotated, because this enzyme may not even be produced by these prokaryotes.

It has been suggested earlier that the ChlD and ChlI proteins can function as subunits of the cobalt chelatase from the aerobic cobalamin (vitamin B12) biosynthesis pathway (Rodionov et al. 2003). It should be noted that in addition to chlorophyll, *R. sphaeroides* and *N. punctiforme* are able to synthesize vitamin B12 as well (Warren and Deery 2009). Additionally, it has been reported that *Mycobacterium tuberculosis* has the ability for de novo cobalamin biosynthesis (Gopinath et al. 2013). Thus, we compared the genes from *Methanocaldococcus* and *D. acidovorans SPH-1* with the genes present in these three cobalamin producers. Indeed, many genes from the vitamin B12 biosynthesis pathway were present in the three genomes including the *cobN* gene encoding the large Co-chelatase subunit (fig. 1, right). Moreover, in all three genomes, the frameshifted *chlD* genes were colocalized with the *cobN* genes that additionally supported their possible functional connection (supplementary fig. 3, Supplementary Material online). Additionally, the *cobS* and *cobT* genes encoding the small and the medium Co-chelatase subunits were absent in the two archaeal genomes (fig. 1). These observations indicated that the products of the frameshifted *chlD* genes were likely to function in the vitamin B12 biosynthesis pathway as subunits of cobalt chelatase.

The Hypothesis: The Small and the Medium Chelatase Subunits May Be Generated from the Same *chlD* Gene via Frameshifting

It has been known for a long time that the small (ChlI or CobS) chelatese subunit is highly similar to the N-terminal part of the medium (ChlD or CobT) chelatase subunit (Walker and Willows 1997). At the same time, all the three validated frameshifting signals in the *chlD* genes were located at the end of the region with similarity to the small subunit and in between the gene parts encoding protein domains. Thus, given the stochastic way of PRF functioning, the two different proteins can be produced from the *chlD* gene with a frameshifting signal (fig. 2 and supplementary table 2, Supplementary Material online). On one hand, the conventional translation (i.e., without frameshifting) generates the short protein (due to the premature stop codon) that is similar to the small Mg-chelatase subunit ChlI. On the other hand, translation with −1 frameshifting results in the long product with similarity to the medium Mg-chelatase subunit ChlD. Therefore, we hypothesized that the function of the PRF signal in the *chlD* gene is to provide synthesis of the two chelatase subunits from the same gene. Consequently, one can expect that a genome with a frameshifted *chlD* gene would lack a separate gene encoding small chelatase subunit—that is, something similar to what was observed in the two *Methanocaldococcus* genomes (fig. 1). To investigate this hypothesis further, our next goals were to identify all the

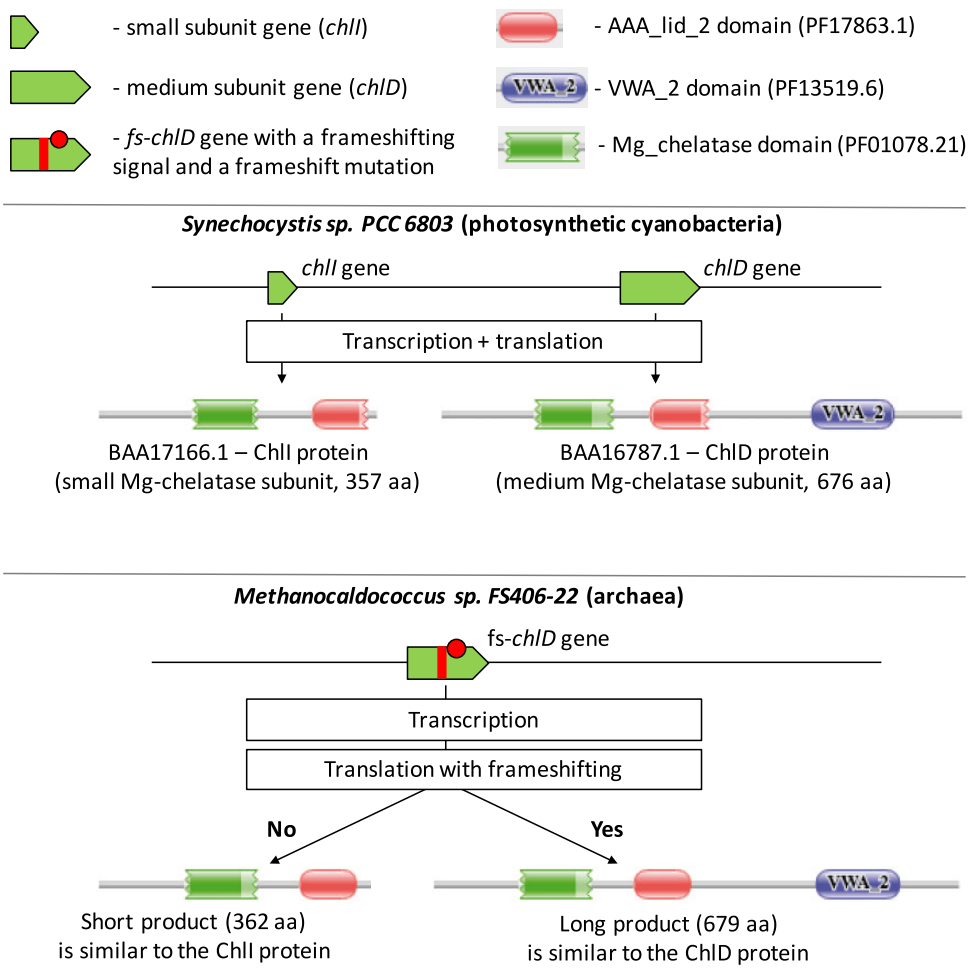

Fig. 2. The domain structures of the annotated ChlD and ChlI proteins (produced from two different genes) and the two putative products of the fs-*chlD* gene with validated frameshifting signal.

prokaryotic genomes with the *chlD* genes (frameshifted or not) and to analyze the presence/absence of the other genes encoding the Mg-/Co-chelatase subunits.

The *chlD* Genes Are Present in Diverse Prokaryotic Genomes

In order to identify all organisms containing the *chlD* gene, the TBlastN tool was applied to the prokaryotic genomes from the Reference and Representative NCBI sets. Pairs of TBlastN searches using the three *chlD* genes with validated frameshifting signals as queries allowed us to find the frameshifted as well as the normal *chlD* genes (see Materials and Methods). In total, among the 5,616 analyzed complete prokaryotic genomes, there were 1,207 genomes (22%) with the medium chelatase subunit genes (supplementary table 3 and fig. 4, Supplementary Material online). Among them, there were 167 genomes containing 169 frameshifted genes. All the identified frameshifted genes were more similar to the medium subunit of the Mg-chelatase (*chlD* and *bchD* genes) rather than to the *cobT* gene of the Co-chelatase (supplementary fig. 2C, Supplementary Material online). Most frequently, they were present in the genomes of proteobacteria, actinobacteria, chloroflexi, and euryarchaeota.

To determine whether the identified fs-*chlD* genes function as subunits of magnesium or cobalt chelatase, we predicted possible phenotypes of the corresponding species, that is, whether they were able to synthesize (bacterio)chlorophyll or cobalamin (or both). Search for the representative genes from these pathways revealed that only ~18% of the chlD-containing genomes may be able to synthesize (bacterio)chlorophyll, whereas >62% of organisms contained many genes from the cobalamin biosynthesis pathway (fig. 3). Thus, the majority of the fs-*chlD* genes may encode cobalt chelatase subunits.

Overview of Different Chelatase Genotypes

The results obtained in this and other studies suggested that the Mg- and Co-chelatases can be synthesized from several different gene sets ("chelatase genotypes"). Namely, the three component magnesium chelatase can be encoded by a full set of three separate genes (the "*chlH*, *chlD*, *chlI*" genotype—fig. 4A). Similarly, the cobalt chelatase cobNST can be produced from the "*cobN*, *cobT*, *cobS*" genotype (fig. 4B). Additionally, the small and the medium Mg-chelatase subunits can also function as parts of the chimeric Co-chelatase (Rodionov et al. 2003) (fig. 4C and D). Finally, the translational

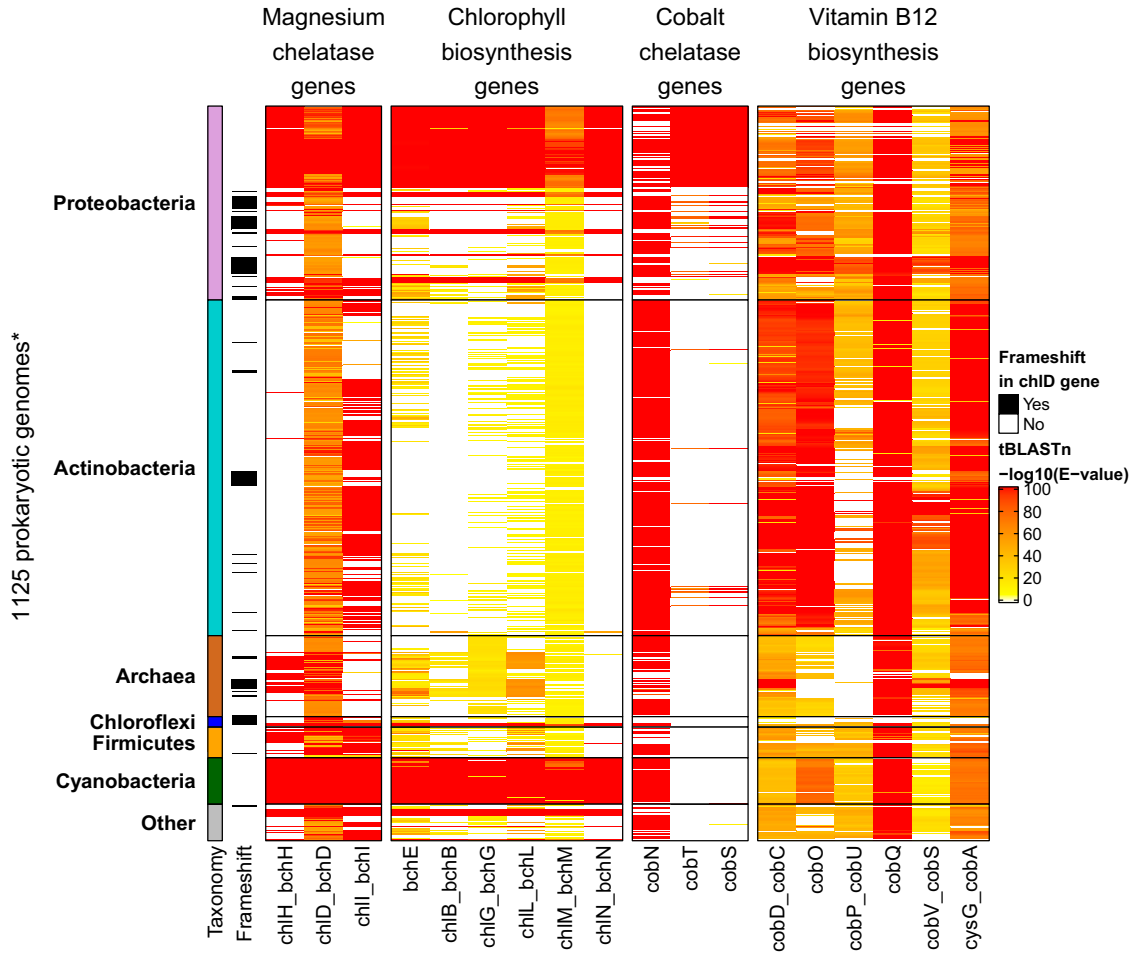

Fig. 3. Identification of the possible phenotypes of the organisms containing *chlD* genes. The heatmap was built in the same way as the heatmap in figure 1. The species from each taxonomic group were ordered according to the 16S rRNA phylogenetic tree. *Only genomes with annotated 16S rRNAs were included.

frameshifting may allow chelatase synthesis from a reduced two-gene genotype (fig. 4*E*).

To find out the frequencies of different genotypes among the 1,207 species, we analyzed the presence/absence of the genes encoding chelatase subunits. A number of genomes had the complete three-gene genotypes as well as the reduced two-gene genotypes (supplementary fig. 5, Supplementary Material online). Notably, there were 78 genomes with the "*cobN*, fs-*chlD*" genotype, that is, they contained one *cobN* (the large subunit of the cobalt chelatase) and one fs-*chlD* gene, whereas the gene(s) encoding the small chelatase subunit (*chlI*, *bchI*, or *cobS*) was absent. Overall, the small chelatase subunit gene was absent in 113 out of the 139 (81%) genomes that had the fs-*chlD* gene and at least one large chelatase subunit gene (supplementary fig. 6, Supplementary Material online). This observation supported the proposed hypothesis.

Unexpectedly, the "*cobN*, *chlD*" genotype was one of the most frequent genotypes. Taking into account that the *chlD* gene did not have a frameshift, it was unclear how the small chelatase subunit can be generated from these reduced gene sets. Moreover, this genotype demonstrated evolutionary conservation as it was common for actinobacteria, proteobacteria, and euarchaeota. Based on the analysis of the frameshifted *chlD* genes, we additionally hypothesized that some of the normal (i.e., not frameshifted) *chlD* genes may also contain embedded signals that may provide the synthesis of the small chelatase subunit (fig. 4*F*). Thus, for each of the phylogenetic group, we also searched for the possible "in-frame" signals inside the *chlD* genes. Possible molecular mechanisms that can be used to generate all chelatase subunits from different gene sets are considered below.

Proteobacteria: −1 Translational Frameshifting

More than half of all the identified fs-*chlD* genes (87 out of the 169) were found in proteobacteria and all of them were of the −1 type. This included the frameshifted gene from *D. acidovorans SPH-1* containing the −1 frameshifting signal that has been experimentally validated in our previous work (Antonov, Coakley, et al. 2013). Importantly, we observed a statistically significant enrichment of the fs-*chlD* genes in the genomes with the reduced genotypes (Fisher's exact test *P* value = $1 \times 10 - 6$—see supplementary fig. 7, Supplementary Material online). This indicated that the presence of a

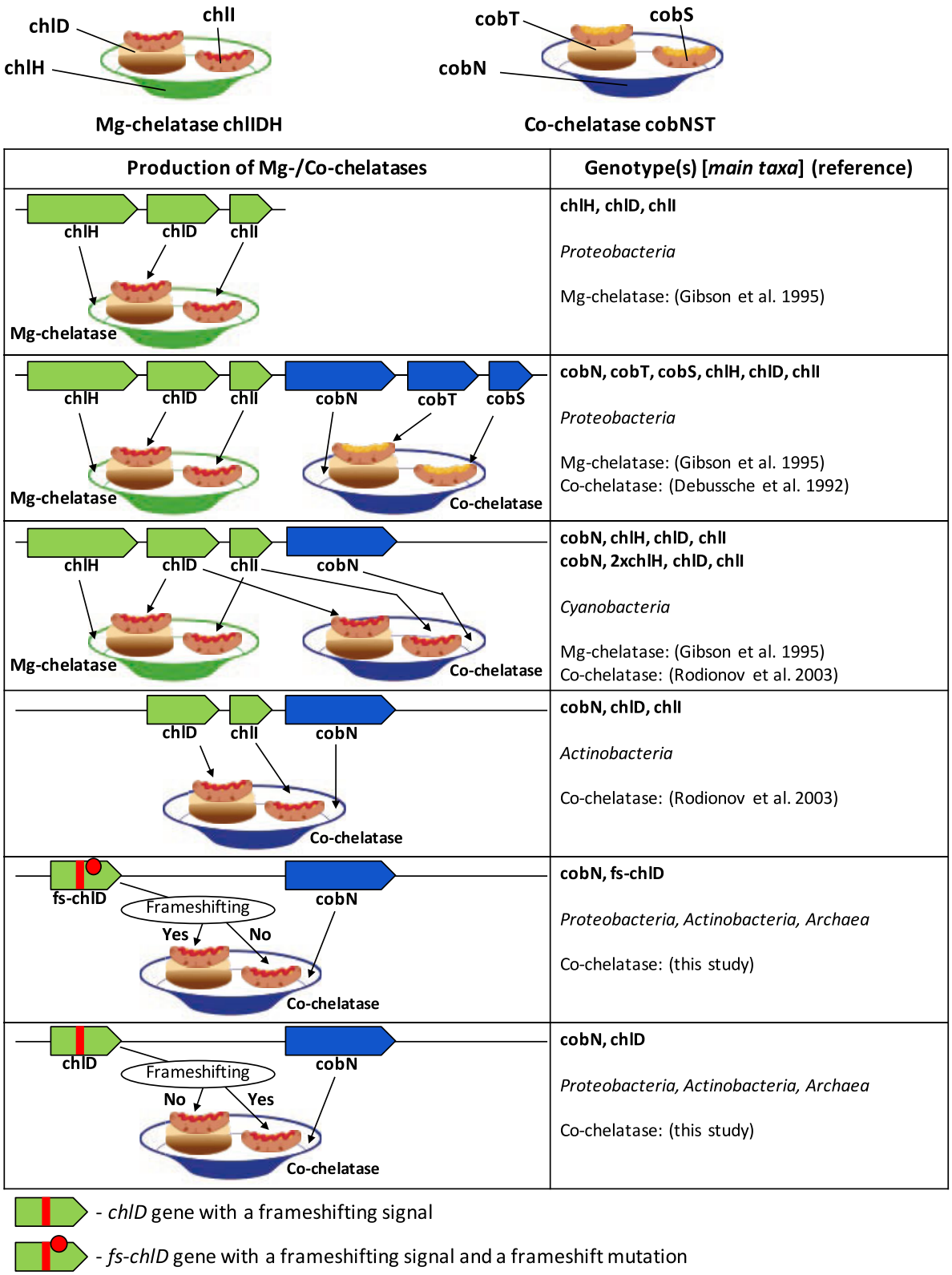

FIG. 4. The proposed mechanisms of the Mg-/Co-chelatase synthesis from different genotypes. For visualization purposes, a chelatase is depicted as a sausage (small subunit) and a hot dog (medium subunit) on a plate (large subunit). This representation highlights that facts that each of the chelatases consists of three different subunits and that the small subunit is similar to a part of the medium subunit (in other words, "a sausage is a part of a hot dog").

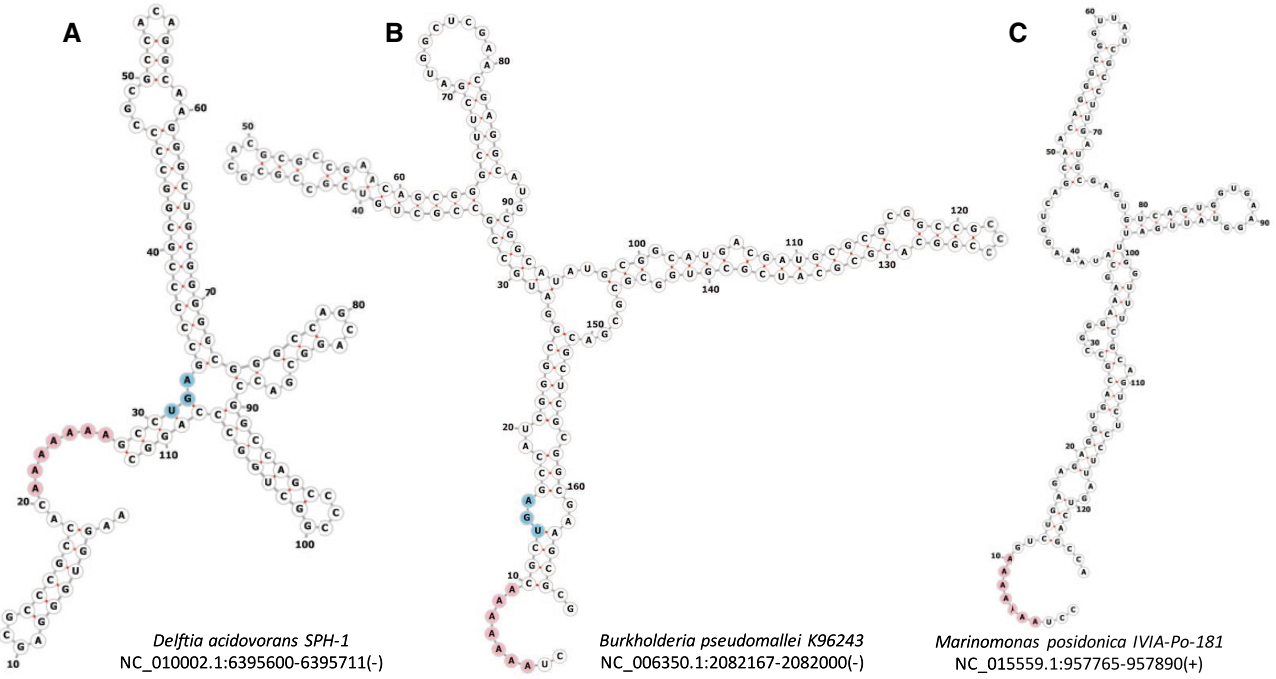

FIG. 5. Predicted −1 frameshifting signals in the *chlD* genes from (*A*) *Delftia acidovorans SPH-1*, (*B*) *Burkholderia pseudomallei K96243*, and (*C*) *Marinomonas posidonica IVIA-Po-181*. The first nucleotide in each sequence corresponds to the first codon position in the original reading frame. The putative poly-A slippery sites and in-frame stop codons are marked by colors. Importantly, the *D. acidovorans SPH-1* and *Pseudomonas aeruginosa PAO1* sequences contain a frameshift, so that the normal translation (without frameshifting) encounters a premature stop codon and produces a truncated protein (a putative small chelatase subunit). In contrast, the *Marinomonas posidonica IVIA-Po-181* gene does not have a frameshift at the DNA level and its normal translation produces a full-length medium chelatase subunit. A −1 frameshifting at the slippery site would lead to the premature translation termination due to the out-of-frame stop codon and the production of the putative small chelatase subunit.

frameshift mutation in the *chlD* gene can compensate for the absence of a separate gene encoding the small subunit. Indeed, the translational frameshifting can provide the production of two chelatase subunits from fs-*chlD* genes.

The validated frameshifting signal from *D. acidovorans SPH-1 chlD* gene contained the well-known −1 slippery site A_AAA_AAA immediately upstream of the premature stop codon (fig. 5*A* and supplementary table 1, Supplementary Material online). In total, this slippery site was present in 84 out of the 87 proteobacterial fs-*chlD* genes (supplementary fig. 8, Supplementary Material online). Additionally, many genes had strong secondary structures downstream of the predicted slippery sequences. This included the *chlD* genes from the important human pathogens *Pseudomonas aeruginosa PAO1* and *Burkholderia pseudomallei K96243* that were also missing the small chelatase subunit genes (fig. 5*B*). Together, these observations suggested strong conservation of the −1 frameshifting signal in the proteobacterial fs-*chlD* genes.

It should be noted that there were 47 proteobacterial genomes with the "*cobN*, *chlD*" genotype and many of these species were cobalamin producers. It was unclear how the small subunit of the aerobic cobalt chelatase can be generated from this reduced genotype. We hypothesized that its functioning can be similar to the "*cobN*, fs-*chlD*" genotype. The main difference between them is that upon frameshifting the fs-*chlD* gene produces the medium subunit, whereas the frameshifting in the normal *chlD* gene leads to premature termination of translation and generation of the small subunit. Such a mechanism would require the presence of the frameshifting signal in the *chlD* genes from the "*cobN*, *chlD*" genomes. Our search for such "in-frame" signals revealed putative slippery sites in 9 (19%) out of the 47 *chlD* genes (supplementary table 4, Supplementary Material online). Moreover, the strong secondary structure and the stop codon in the −1 alternative frame were also present downstream of the putative slippery site (see fig. 5*C* for a representative case). Thus, these regions were likely to function as −1 frameshifting signals as well.

The frameshifting signal can allow a *chlD* gene (frameshifted or not) to encode both the small and the medium chelatase subunits. However, such "dual coding" was not needed if a small chelatase subunit gene was present in the genome. Consequently, we expected that the *chlD* genes from the genomes with the full set of the chelatase genes (such as "*cobN*, *chlD*, *chlI*," "*chlH*, *chlD*, *chlI*," "*cobN*, *cobT*, *cobS*, *chlH*, *chlD*, *chlI*") did not have such signals. Indeed, only 2 out of the 112 *chlD* genes from the genomes with the complete genotypes contained the putative slippery sites. Thus, we observed a statistically significant enrichment of the putative slippery sites in the *chlD* genes from the reduced genotypes (Fisher's exact test P value $= 1.1 \times 10^{-22}$—see supplementary fig. 9, Supplementary Material online). This supported the hypothesis that two chelatase subunits can be generated

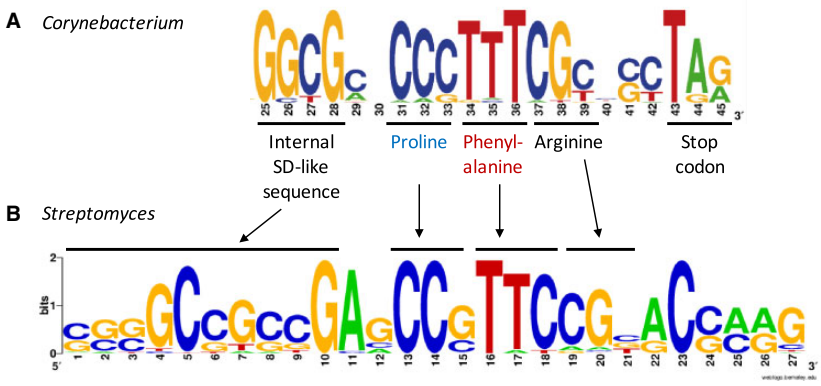

FIG. 6. The logo of the predicted +1 frameshifting signal in the *chlD* genes from the (*A*) *Corynebacterium* and (*B*) *Streptomyces* genera. The logos were generated from the gap-free codon alignments.

from the *chlD* genes (with or without a frameshift mutation) due to the presence of the frameshifting signal.

Actinobacteria: +1 Translational Frameshifting

Among the 38 fs-*chlD* genes from Actinobacteria (see supplementary table 3, Supplementary Material online), 27 were present in the genomes with the "*cobN*, fs-*chlD*" genotype (supplementary fig. 5, Supplementary Material online). Interestingly, in contrast to proteobacteria, the majority of the actinobacterial fs-*chlD* genes had the +1 frameshifts, that is, the ribosome needed to skip a single nucleotide (or move two nucleotides backward) in order to bypass the premature stop codon and continue reading the long ORF. Namely, there were 29 and 9 fs-*chlD* genes with +1 and −1 frameshifts, respectively. The majority of the fs-*chlD* genes with +1 frameshift were predicted in *Corynebacterium* genomes (including the human pathogen *Corynebacterium diphtheriae*). Our analysis of these genes revealed a putative +1 frameshifting signal located upstream of the premature stop codon. The signal consisted of two consecutive "slow" codons (encoding proline and phenylalnine) as well as an internal Shine–Dalgarno like sequence (fig. 6*A*). It should be noted that proline codons have been reported to induce +1 programed frameshifting (O'Connor 2002; Maehigashi et al. 2014) and the internal Shine–Dalgarno like site can stimulate the frameshifting efficiency (similarly to the well-studied +1 signal in the *prfB* gene encoding the release factor 2; Baranov et al. 2002b). Thus, it is likely that actinobacteria from the *Corynebacterium* genus use +1 translational frameshifting to produce two chelatase subunits from the single *chlD* mRNA.

Although the majority of the Actinobacterial genomes had the complete "*cobN*, *chlD*, chlI" genotype, there were 139 "*cobN*, *chlD*" genomes (supplementary fig. 5, Supplementary Material online). More than half of the actinobacterial genomes with this genotype were from the *Streptomyces* genus including the well-known cobalamin producer *Streptomyces olivaceus* (Hall et al. 1953). We identified a similar conserved +1 frameshifting signal in these frameshift-free *chlD* genes as well (fig. 6*B*). We also analyzed the nine actinobacterial *chlD* genes with −1 frameshifts and one of them contained a putative frameshifting signal that resembled the signals from some of the proteobacterial fs-*chlD* genes (supplementary fig. 10, Supplementary Material online). Therefore, a number of actinobacterial *chlD* genes contained putative frameshifting signals and, similarly to proteobacteria, may be able to generate two chelatase subunits from the single *chlD* gene.

Cyanobacteria

The majority of the cyanobacterial genomes contained many genes from both the chlorophyll and the cobalamin biosynthesis pathways (fig. 3). In 75 out of the 76 analyzed genomes, the subunits of the magnesium chelatase were encoded by the complete set of genes (the "*cobN*, 2x*chlH*, *chlD*, chlI" and "*cobN*, *chlH*, *chlD*, chlI" genotypes—see supplementary fig. 5, Supplementary Material online). Consequently, we did not observe any fs-*chlD* genes in these genomes.

In addition to the Mg-chelatase genes, most of the cyanobacterial genomes also contained the *cobN* gene (encoding the large subunit of the Co-chelatase) as well as the other genes from the cobalamin biosynthesis pathway. Thus, a number of proteobacteria and cyanobacteria may be able to synthesize both the cobalamin and the (bacterio)chlorophyll. However, the required Co- and Mg-chelatases were encoded by different sets of genes in these two groups of bacteria. A number of proteobacteria encoded both chelatases with the two complete sets of genes (i.e., the "*cobN*, *cobT*, *cobS*, *chlH*, *chlD*, *chlI*" genotype). On the other hand, the majority of Cyanobacteria had a complete set of the magnesium chelatase genes, whereas the cobalt chelatase was represented by a single *cobN* gene only (the "*cobN*, *chlH*, *chlD*, chlI" and "*cobN*, 2x*chlH*, *chlD*, chlI" genotypes). Thus, the aerobic cobalt chelatase in cyanobacteria is likely to be assembled from the products of the *cobN*, *chlD*, and *chlI* genes (Rodionov et al. 2003).

Archaea: Possible Transcriptional Slippage

All three frameshifting signals that have been tested in our previous work (Antonov, Coakley, et al. 2013) have shown high efficiency (up to 63%) to induce −1 frameshifting (supplementary table 1, Supplementary Material online). However, in that study, we have not investigated the details of the molecular mechanisms. The presence of the strong

secondary structure downstream of the poly-A slippery site suggested that the bacterial signal induced the translational slippage. On the other hand, the two validated archaeal signals lacked strong RNA basepairing and had longer poly-A slippery sites (supplementary fig. 11, Supplementary Material online). This indicated that in archaea the stochastic correction of the frameshift mutation in the *chlD* may happen during transcription in the process known as transcriptional slippage or programed transcriptional realignment.

In total, we identified 25 archaeal genomes with 27 frameshifted *chlD* genes (including the two validated cases). Stretches of nine or more consecutive T's or A's immediately upstream of the premature stop codon were observed in 21 out of the 27 archaeal fs-*chlD* genes indicating the conservation of the putative slippery sites (supplementary fig. 12, Supplementary Material online). A characteristic property of the archaeal genotypes was the presence of several genes encoding the large chelatase subunits. Still, the frameshifted *chlD* genes were mainly present in the genomes, where the genes encoding the small chelatase subunit were absent. Thus, the production of two subunits from the single *chlD* gene is likely to happen in archaea as well.

Discussion

In this work, we performed a comparative genomics analysis of the magnesium and cobalt chelatase genes from >1,200 prokaryotic genomes. This analysis was motivated by our previous work where we identified and experimentally validated programed frameshifting signals from three frameshifted *chlD* genes (Antonov, Coakley, et al. 2013). The *chlD* gene encodes the medium subunit of the magnesium chelatase, whereas the large and the small subunits of this enzyme are encoded by the *chlH* and *chlI* genes, respectively. To investigate the role of the frameshifting signal, we searched for normal as well as frameshifted *chlD* genes in a representative set of complete prokaryotic genomes. Surprisingly, the majority of the identified species were cobalamin producers and their genomes contained aerobic cobalt chelatase cobNST that was similar to the Mg-chelatase chlIDH. As the two chelatases are highly similar to each other, some of their genes can be interchangeable. Thus, in each chlD-containing genome, we identified all the genes encoding subunits of the chlIDH and cobNST enzymes. Next, we analyzed different gene sets that can be used to encode the magnesium and/or cobalt chelatases.

The photosynthetic bacteria encoded the Mg-chelatase using the full set of three genes (fig. 4A–C). Additionally, many photosynthetic proteobacteria were also cobalamin producers and they encoded the two chelatases with two full sets of genes (the "*cobN*, *cobT*, *cobS*, *chlH*, *chlD*, *chlI*" genotype—fig. 4B). In contrast, in cyanobacteria and some actinobacteria (Rodionov et al. 2003), the products of the *chlD* and *chlI* genes were likely to function as subunits of the magnesium as well as the cobalt chelatases (fig. 4C and D).

We noticed that the fs-*chlD* genes were mainly present in the genomes of the cobalamin, rather than chlorophyll synthesizing organisms. Additionally, the small chelatase subunit genes were missing from these genomes, that is, many of them had the "*cobN*, fs-*chlD*" genotype. Notably, this included such human pathogens as *P. aeruginosa PAO1*, *B. pseudomallei K96243*, *C. diphtheriae*, and *Nocardia brasiliensis ATCC 700358*. Given the high similarity between the ChlI protein and the N-terminal part of the ChlD protein, we hypothesized that translational frameshifting may allow the *chlD* mRNA to produce both the small and the medium subunits of the aerobic cobalt chelatase (fig. 4E). The obtained results from different prokaryotic phyla supported this theory.

Surprisingly, yet another genotype with missing small chelatase subunit gene ("*cobN*, *chlD*") was frequently observed in our analysis. It was common for three different prokaryotic groups (actinobacteria, proteobacteria, and archaea) indicating its conservation and functionality. As many of the corresponding species were able to synthesize cobalamin aerobically, we assumed that the corresponding *chlD* genes also contained the frameshifting signals that would allow them to produce the small subunit as well (fig. 4F). Indeed, we were able to identify such putative signals in some of the *chlD* genes from these genomes.

It should be noted that many fs-*chlD* genes are currently annotated as two separate adjacent genes or as frameshifted pseudogenes. Based on the obtained results, we argued that many of them actually constitute functional genes with frameshifting signals. This feature may allow the production of both the small and the medium cobalt chelatase subunits. The correction of the current annotations may help to avoid possible confusions regarding the missing chelatase genes in some genomes.

Further studies of the *chlD* genes and the embedded recoding signals may also be interesting from the evolutionary point of view. In fact, the *chlD* gene may have appeared very early in the evolution—it has recently been short listed among the genes that can be traced back to the last universal common ancestor (Weiss et al. 2016). Thus, it seems possible that the modern *chlD*, *bchD* and, probably, *cobT* genes may have originated from that proto-gene. Consequently, it was surprising to observe various recoding mechanisms in the *chlD* genes from different prokaryotic groups. Namely, putative transcriptional slippage, −1 and +1 translational frameshifting may be utilized by archaea, proteobacteria, and actinobacteria, respectively. Such a diversity of molecular mechanisms may be a result of a convergent evolution. Apparently, the ability to encode two proteins in the same *chlD* gene provided evolutionary advantages such as the possibility for cotranslational assembly of the chelatase protein complex. The synthesis of two proteins from the same mRNA ensures their spacial proximity for immediate interaction. Indeed, it has been shown that the small and the medium subunits form a separate protein complex that binds to the large subunit to make the functional chelatase (Debussche et al. 1992). At the same time, this mechanism was not observed in the Mg-chelatase genes possibly indicating that some required mutations in the *chlD* and/or *chlI* genes made it impossible to encode these two subunits in the

same gene. Partial duplication of the *chlD* gene early in the evolution of photosynthesis might have resolved this genetic conflict (Xiong et al. 1998).

It is important to note that there is an anaerobic biochemical pathway of cobalamin biosynthesis. The two alternative pathways are used by anaerobic and aerobic prokaryotes. Although a number of enzymes function in both pathways, the corresponding cobalt chelatases are different and do not share the same evolutionary origin. Taken into account that aerobic cobalt chelatase cobNST is likely to be homologous to the magnesium chelatase chlIDH, it seems reasonable to assume that aerobic vitamin B12 pathway appeared in the evolution after the appearance of the oxygenic photosynthesis. On the other hand, the anaerobic pathway may have been utilized by the ancient prokaryotes for cobalamin biosynthesis before the accumulation of the oxygen in the atmosphere. Such a logic may provide useful guidelines for the order in which different biochemical pathways appeared in the evolutionary history of life. Thus, the future studies of the frameshifting signal in the *chlD* genes may shed a new light on the early evolution of chlorophyll and/or cobalamin biosynthesis pathways.

Materials and Methods

The TBlastN (Camacho et al. 2009) tool was used to identify the genomes containing *chlD* gene(s). The translations of the three *chlD* genes with the validated frameshifting signals were used as queries (see supplementary table 5, Supplementary Material online). To allow the identification of the other frameshifted *chlD* genes, two separate tBASTn searches (*E*-value threshold $= 10^{-6}$) were performed using the N-terminal (i.e., before the frameshift) or the C-terminal (after the frameshift) part of each query protein. The genomic regions with adjacent hits (in the correct order) with up to one frameshift were classified as *chlD* genes.

In order to identify the small and the large chelatase subunit genes as well as the other genes from the chlorophyll and cobalamin biosynthesis pathways, TBlastN search was performed for a set of reference query proteins. Namely, the query list included four large (*chlH*, *bchH*, and two CobN proteins) and four small (*chlI*, *bchI*, and two *cobS* proteins) chelatase subunits from diverse phylogenetic groups (see supplementary data, Supplementary Material online). Additionally, we considered the 461 aa long "magnesium chelatase" protein (WP_085243324.1) from *Mycobacterium europaeum* as a putative small chelatase subunit because it had similarity to the annotated ChlI proteins (TBlastN *E*-value $< 10^{-15}$), contained the "Mg-chelatase subunit ChlI" domain (COG1239) and was frequently present in the actinobacterial genomes, where the reference small chelatase subunits did not produce any hits.

Given the similarities between different chelatase subunits, we imposed additional constraints on their lengths. Namely, we required that the small chelatase subunits were between 250 aa and 500 aa, medium—between 500 aa and 800 aa and large—between 1,000 aa and 2,000 aa (these values corresponded to the lengths of the annotated chelatase subunits—see supplementary fig. 1, Supplementary Material online). The identified genes were automatically annotated (i.e., assigned a gene name like *chlD*, *bchD*, or *cobT*) by the best reciprocal hit approach.

The phylogenetic trees were constructed using RAxML version 8.2.12 (Stamatakis 2014). Subsequent phylogenetic analyses and visualizations were performed by the "ape" R package (Paradis and Schliep 2019). All the heatmaps were generated using the ComplexHeatmap R package (Gu et al. 2016).

The RNA secondary structures were predicted using the RNAfold web server (Lorenz et al. 2011) available at http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi (last accessed November 25, 2019). The sequence logos were generated using the WebLogo server (Crooks et al. 2004) available at https://weblogo.berkeley.edu/logo.cgi (last accessed October 15, 2019).

Data Availability

The code that was used to generate the images and the corresponding raw data files are available at https://github.com/vanya-antonov/article-chelatase.

Supplementary Material

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

Acknowledgments

We would like to thank Pavel Baranov (University College Cork, Ireland) and Mark Borodovsky (Georgia Institute of Technology, USA) for the important contributions to the original discovery of the PRF in the chelatase gene and for the helpful suggestions regarding this study, Maria Zamkova (Russian N.N. Blokhin Cancer Research Center, Russia) for the useful discussions. This work was supported by the Russian Foundation for Basic Research (Grant No. 18-34-00589 to I.A.).

References

Antonov I, Baranov P, Borodovsky M. 2013. Genetack database: genes with frameshifts in prokaryotic genomes and eukaryotic mRNA sequences. *Nucleic Acids Res*. 41(D1):D152–D156.

Antonov I, Borodovsky M. 2010. Genetack: frameshift identification in protein-coding sequences by the Viterbi algorithm. *J Bioinform Comput Biol*. 08(03):535–551.

Antonov I, Coakley A, Atkins JF, Baranov PV, Borodovsky M. 2013. Identification of the nature of reading frame transitions observed in prokaryotic genomes. *Nucleic Acids Res*. 41(13):6514–6530.

Atkins JF, Loughran G, Bhatt PR, Firth AE, Baranov PV. 2016. Ribosomal frameshifting and transcriptional slippage: from genetic steganography and cryptography to adventitious use. *Nucleic Acids Res*. 44(15):7007–7078.

Baranov PV, Gesteland RF, Atkins JF. 2002a. Recoding: translational bifurcations in gene expression. *Gene* 286(2):187–201.

Baranov PV, Gesteland RF, Atkins JF. 2002b. Release factor 2 frameshifting sites in different bacteria. *EMBO Rep*. 3(4):373–377.

Bilgin H, Sarmis A, Tigen E, Soyletir G, Mulazimoglu L. 2015. *Delftia acidovorans*: a rare pathogen in immunocompetent and immunocompromised patients. *Can J Infect Dis Med Microbiol*. 26(5):277–279.

Blinkova A, Burkart MF, Owens TD, Walker JR. 1997. Conservation of the *Escherichia coli* DNAx programmed ribosomal frameshift signal in *Salmonella typhimurium*. *J Bacteriol*. 179(13):4438–4442.
Bollivar DW, Suzuki JY, Beatty JT, Dobrowolski JM, Bauer CE. 1994. Directed mutational analysis of bacteriochlorophyll a biosynthesis in *Rhodobacter capsulatus*. *J Mol Biol*. 237(5):622–640.
Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. Blast+: architecture and applications. *BMC Bioinformatics* 10(1):421.
Craigen WJ, Caskey CT. 1986. Expression of peptide chain release factor 2 requires high-efficiency frameshift. *Nature* 322(6076):273–275.
Crooks G, Hon G, Chandonia J, Brenner S. 2004. Weblogo: a sequence logo generator. *Genome Res*. 14(6):1188–1190.
Dailey HA. 2013. Illuminating the black box of B12 biosynthesis. *Proc Natl Acad Sci U S A*. 110(37):14823–14824.
Debussche L, Couder M, Thibaut D, Cameron B, Crouzet J, Blanche F. 1992. Assay, purification, and characterization of cobaltochelatase, a unique complex enzyme catalyzing cobalt insertion in hydrogenobyrinic acid a, c-diamide during coenzyme B12 biosynthesis in *Pseudomonas denitrificans*. *J Bacteriol*. 174(22):7445–7451.
Fodje M, Hansson A, Hansson M, Olsen J, Gough S, Willows R, Al-Karadaghi S. 2001. Interplay between an AAA module and an integrin I domain may regulate the function of magnesium chelatase. *J Mol Biol*. 311(1):111–122.
Gibson L, Willows RD, Kannangara CG, von Wettstein D, Hunter CN. 1995. Magnesium-protoporphyrin chelatase of *Rhodobacter sphaeroides*: reconstitution of activity by combining the products of the bchH, -I, and -D genes expressed in *Escherichia coli*. *Proc Natl Acad Sci U S A*. 92(6):1941–1944.
Gopinath K, Moosa A, Mizrahi V, Warner DF. 2013. Vitamin B12 metabolism in *Mycobacterium tuberculosis*. *Future Microbiol*. 8(11):1405–1418.
Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 32(18):2847–2849.
Guilard R, Kadish KM, Smith KM, Guilard R. 2003. The porphyrin handbook. Vol. 18. New York: Academic Press.
Hall HH, Benedict RG, Wiesen CF, Smith CE, Jackson RW. 1953. Studies on vitamin B12 production with *Streptomyces olivaceus*. *Appl Microbiol*. 1(3):124–129.
Ivanov IP, Atkins JF. 2007. Ribosomal frameshifting in decoding antizyme mRNAs from yeast and protists to humans: close to 300 cases reveal remarkable diversity despite underlying conservation. *Nucleic Acids Res*. 35(6):1842–1858.
Keeling PJ. 2004. Diversity and evolutionary history of plastids and their hosts. *Am J Bot*. 91(10):1481–1493.
Lorenz R, Bernhart SH, Zu Siederdissen CH, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. Viennarna package 2.0. *Algorithms Mol Biol*. 6(1):26.
Lundqvist J, Elmlund D, Heldt D, Deery E, Söderberg CA, Hansson M, Warren M, Al-Karadaghi S. 2009. The AAA+ motor complex of subunits cobs and cobt of cobaltochelatase visualized by single particle electron microscopy. *J Struct Biol*. 167(3):227–234.
Maehigashi T, Dunkle JA, Miles SJ, Dunham CM. 2014. Structural insights into+ 1 frameshifting promoted by expanded or modification-deficient anticodon stem loops. *Proc Natl Acad Sci U S A*. 111(35):12740–12745.
Martens JH, Barg H, Warren MJ, Jahn D. 2002. Microbial production of vitamin B12. *Appl Microbiol Biotechnol*. 58(3):275–285.
Mehta MP, Baross JA. 2006. Nitrogen fixation at 92 c by a hydrothermal vent archaeon. *Science* 314(5806):1783–1786.
Meydan S, Klepacki D, Karthikeyan S, Margus T, Thomas P, Jones JE, Khan Y, Briggs J, Dinman JD, Vázquez-Laslop N, et al. 2017. Programmed ribosomal frameshifting generates a copper transporter and a copper chaperone from the same gene. *Mol Cell*. 65(2):207–219.
O'Connor M. 2002. Imbalance of tRNApro isoacceptors induces+ 1 frameshifting at near-cognate codons. *Nucleic Acids Res*. 30(3):759–765.
Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35(3):526–528.
Raux E, Schubert H, Warren M. 2000. Biosynthesis of cobalamin (vitamin B12): a bacterial conundrum. *Cell Mol Life Sci*. 57(13–14):1880–1893.
Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS. 2003. Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes. *J Biol Chem*. 278(42):41148–41159.
Sharma V, Firth A, Antonov I, Fayet O, Atkins J, Borodovsky M, Baranov P. 2011. A pilot study of bacterial genes with disrupted ORFs reveals a surprising profusion of protein sequence recoding mediated by ribosomal frameshifting and transcriptional realignment. *Mol Biol Evol*. 28(11):3195–3211.
Stamatakis A. 2014. RaxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
Thiel V, Tank M, Bryant DA. 2018. Diversity of chlorophototrophic bacteria revealed in the omics era. *Annu Rev Plant Biol*. 69(1):21–49.
Thweatt JL, Canniffe DP, Bryant DA. 2019. Biosynthesis of chlorophylls and bacteriochlorophylls in green bacteria. In: Grimm B, editor. *Metabolism, structure and function of plant Tetrapyrroles: introduction, microbial and eukaryotic chlorophyll synthesis and catabolism*.Vol. 90. Amsterdam: Elsevier. p. 35–89.
Wakao N, Yokoi N, Isoyama N, Hiraishi A, Shimada K, Kobayashi M, Kise H, Iwaki M, Itoh S, Takaichi S. 1996. Discovery of natural photosynthesis using Zn-containing bacteriochlorophyll in an aerobic bacterium *Acidiphilium rubrum*. *Plant Cell Physiol*. 37(6):889–893.
Walker CJ, Willows RD. 1997. Mechanism and regulation of Mg-chelatase. *Biochem J*. 327 (2):321–333.
Warren MJ, Deery E. 2009. Vitamin B12 (cobalamin) biosynthesis in the purple bacteria. In: The purple phototrophic bacteria. Springer. p. 81–95.
Weiss MC, Sousa FL, Mrnjavac N, Neukirchen S, Roettger M, Nelson-Sathi S, Martin WF. 2016. The physiology and habitat of the last universal common ancestor. *Nat Microbiol*. 1(9):16116.
Willows R, Hansson M. 2003. Mechanism, structure, and regulation of magnesium chelatase. San Diego: Academic Press.
Xiong J, Inoue K, Bauer CE. 1998. Tracking molecular evolution of photosynthesis by characterization of a major photosynthesis gene cluster from *Heliobacillus mobilis*. *Proc Natl Acad Sci U S A*. 95(25):14851–14856.