

# Identification of Xenologs and Their Characteristic Low Expression Levels in the Cyanobacterium *Synechococcus elongatus*

Gilberto Álvarez-Canales<sup>1</sup> · Guadalupe Arellano-Álvarez<sup>1</sup> · Carmen M. González-Domenech<sup>2</sup> · Fernando de la Cruz<sup>3</sup> · Andrés Moya<sup>4</sup> · Luis Delaye<sup>1</sup>

Received: 3 June 2014 / Accepted: 28 May 2015 / Published online: 4 June 2015  
© Springer Science+Business Media New York 2015

**Abstract** Horizontal gene transfer (HGT) is a central process in prokaryotic evolution. Once a gene is introduced into a genome by HGT, its contribution to the fitness of the recipient cell depends in part on its expression level. Here we show that in *Synechococcus elongatus* PCC 7942, xenologs derived from non-cyanobacterial sources exhibited lower expression levels than native genes in the genome. In accord with our observation, xenolog codon adaptation indexes also displayed relatively low expression values. These results are in agreement with previous reports that suggested the relative neutrality of most xenologs. However, we also demonstrated that some of the xenologs detected participated in cellular functions, including iron starvation acclimation and nitrate reduction, which corroborate the role of HGT in bacterial adaptation. For example, the expression levels of some of the xenologs detected are known to increase under iron-limiting conditions. We interpreted the overall pattern as an indication

that there is a selection pressure against high expression levels of xenologs. However, when a xenolog protein product confers a selective advantage, natural selection can further modulate its expression level to meet the requirements of the recipient cell. In addition, we show that ORFans did not exhibit significantly lower expression levels than native genes in the genome, which suggested an origin other than xenology.

**Keywords** Cyanobacteria · Codon usage · Horizontal gene transfer · Translation efficiency

## Background

Horizontal gene transfer (HGT) is recognized as a major process in prokaryotic evolution (Zhaxybayeva and Doolittle 2011). Furthermore, HGT has consequences that reach the planetary scale and that challenge our interpretation of the universal tree of life (Fournier and Gogarten 2008; Lazcano 2011). For example, the role of HGT in the evolution of bacterial adaptation is clearly illustrated by the increasing incidence of resistance to antibiotics among pathogenic bacteria (Maclean et al. 2010).

Despite the importance of HGT in prokaryotic evolution, the internal factors that determine the success of the acquired genes remain unclear. Once a protein coding gene is acquired by a genome via HGT, its protein product interacts with the molecules present in the intracellular environment. Furthermore, the fate of the new xenolog depends in part upon the effects that these molecular interactions have on the fitness of the recipient cell. For example, an in silico analysis of the metabolic network of *Escherichia coli* (*E. coli*) revealed that the chance of acquiring a gene by HGT and the long-term survival of the

**Electronic supplementary material** The online version of this article (doi:10.1007/s00239-015-9684-x) contains supplementary material, which is available to authorized users.

✉ Luis Delaye  
ldelaye@ira.cinvestav.mx

<sup>1</sup> Departamento de Ingeniería Genética, CINVESTAV-Irapuato, Km. 9.6 Libramiento Norte, Carretera Irapuato-León, 36821 Irapuato, Guanajuato, Mexico

<sup>2</sup> Facultad de Farmacia, Universidad de Granada, Campus de Cartuja s/n, 18071 Granada, Spain

<sup>3</sup> Departamento de Biología Molecular e Instituto de Biomedicina y Biotecnología de Cantabria (IBBTEC), Universidad de Cantabria-CSIC-SODERCAN, Santander, Spain

<sup>4</sup> Institut Cavanilles de Biodiversitat i Biologia Evolutiva, Universitat de València, Valencia, Spain

acquired gene in the recipient genome are up to six times higher if an enzyme that catalyzes a coupled metabolite flux is encoded in the genome (Pál et al. 2005a).

Recently, Park and Zhang (2012) demonstrated that high xenolog expression levels hamper HGT in *E. coli* and other bacteria. The authors proposed that xenologs diminish the fitness of recipient cells due to (a) energy expenditure during transcription and translation; (b) cytotoxic protein misfolding; (c) reduction in cellular transcriptional efficiency; (d) detrimental protein interaction; and/or (e) disturbance of the optimal protein concentration or cell physiology.

Here we show that xenologs from non-cyanobacterial sources transferred into *Synechococcus elongatus* PCC 7942 (*S. elongatus*) exhibit low expression levels, which corroborates and broadens previous results (Park and Zhang 2012). We interpret this finding as evidence of selection against xenologs that exhibit high expression levels. However, we also demonstrate that some of the xenologs detected participated in important cellular functions, thus supporting the important role of HGT in bacterial adaptation and evolution.

Additionally, we show that ORFans and genes with atypical nucleotide composition exhibit expression levels that do not differ significantly from the expression of native genes in the *S. elongatus* genome, which suggests a different evolutionary origin than the xenologs detected.

*Synechococcus elongatus* is a genetically transformable, unicellular, fresh-water cyanobacterium initially described by von Nägeli (1849); it is a model bacterium for studies of prokaryotic photosynthesis, bioenergy production, systems biology, and circadian rhythms (Koksharova and Wolk 2002; Kondo et al. 1993; Dong and Golden 2008; Rust et al. 2011). Two *S. elongatus* strains (PCC 6301 and PCC 7492) have been sequenced (Sugita et al. 2007). More recently, gene expression levels in *S. elongatus* were measured in unprecedented detail by RNA sequencing (Vijayan et al. 2011). Hence, the available information represents an outstanding opportunity to study the expression levels of xenologs in a model photosynthetic organism.

## Results

### Xenologs in *S. elongatus* Exhibit Low Expression Levels

Using three complementary approaches, we identified genes that were introduced into the *S. elongatus* genome by HGT. First, we conducted BLASTp searches of all protein sequences encoded by *S. elongatus* against a non-redundant (nr) GenBank protein database and identified those proteins

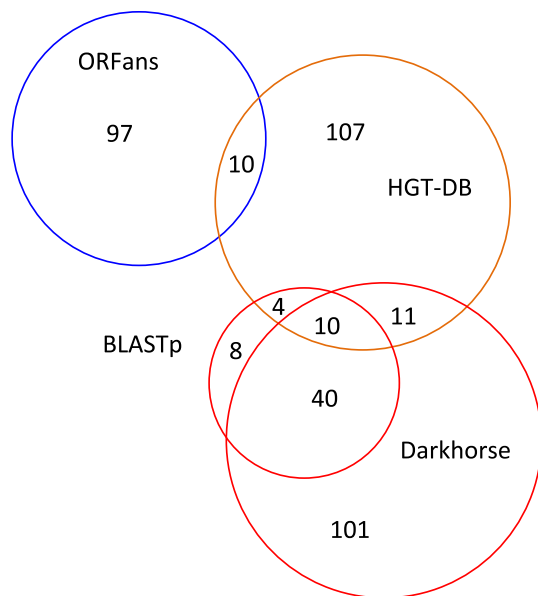
in *S. elongatus* that displayed only non-cyanobacterial homologs ( $e$  value  $< 0.001$ ) among the first 500 BLAST hits. Using this method, we were able to identify 62 chromosome-coded proteins that were probable xenologs in *S. elongatus*. This analysis also allowed us to identify 107 orphan genes (ORFans) without known homologs in *S. elongatus*. See methods for our operational definition of ORFans.

Next, we used the DarkHorse algorithm to identify potential xenologs in *S. elongatus* (Podell et al. 2008). DarkHorse works by first performing BLASTp searches of query proteins against an nr database (the same nr database used previously) and then ranking the BLASTp hits according to their bits scores and taxonomic affiliation. These rankings are then used to calculate a lineage probability index (LPI). The LPI ranges in value from 0 to 1 and is used to identify potential xenologs. After appropriate adjustment of the DarkHorse parameters, we designated all *S. elongatus* proteins with LPI scores below 0.6 as putative xenologs. Our approach was consistent with a calibration generated for 955 microbial genomes that indicated that LPI values below 0.6 are potential candidates for inter-class or inter-phylum HGT (<http://darkhorse.ucsd.edu/index.html>). Using this method, we were able to identify 162 putative xenologs.

Last, we retrieved the list of potential xenologs in *S. elongatus* provided by the Horizontal Gene Transfer Database (HGT-DB; Garcia-Vallvé et al. 2003). HGT-DB is a database centered on compositional properties of annotated genes in prokaryotic genomes that is dedicated to the identification of potential xenologs. Accordingly, HGT-DB enables the identification of all genes that display atypical G + C content, codon usage, or amino acid composition of coded proteins. These atypical genes are then classified as potential xenologs. Using HGT-DB, we retrieved 142 putative xenologs.

The final set of xenologs is presented in Supplementary Table (ST1), and the intersection of the three approaches is detailed in Fig. 1. As shown, most xenologs identified by our BLASTp search were also identified using the DarkHorse algorithm (50 out of 62). The nucleotide composition approach used to identify HGT-DB xenologs appeared to generally complement the BLASTp and DarkHorse results. Finally, only a small proportion of ORFan genes were identified as xenologs using HGT-DB. A summary of the annotated functions of these xenologs is provided in Fig. 2.

Of the xenologs detected, 124 were also detected in a previous analysis (Delaye et al. 2011). A detailed comparison of the xenologs detected in the current and prior analysis is shown in Supplementary Table 2 (ST2). In accord with prior results (Delaye et al. 2011), most transfers to *S. elongatus* that were acquired from non-cyanobacterial sources originated



**Fig. 1** Xenologs and ORFans identified in *S. elongatus*. Xenologs identified by BLASTp and the DarkHorse algorithm are encircled in red; xenologs identified by HGT-DB compositional methods are encircled in orange; ORFans are encircled in blue (Color figure online)

from proteobacteria (Fig. 3). This finding was in accord with experimental data demonstrating that conjugation from *E. coli* to *S. elongatus* readily occurs by various proteobacterial conjugative plasmids (Encinas et al. 2014).

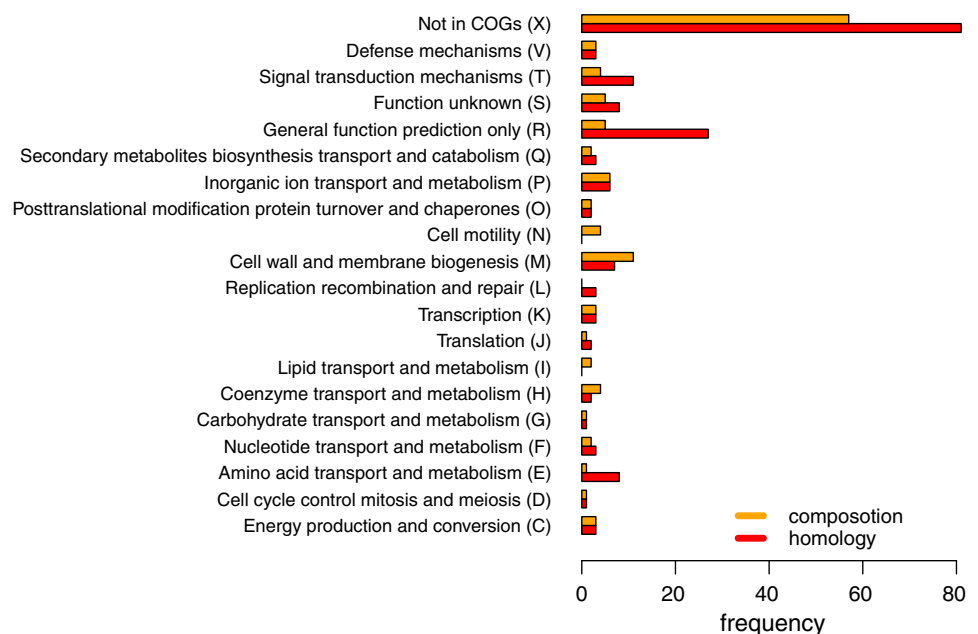
Surprisingly, the best hits for most xenologs detected by compositional methods (those retrieved from HGT-DB) were among cyanobacteria (Fig. 3). This pattern was unexpected because there is no rational explanation why

compositional methods would preferentially detect xenologs from the same phylum (in this case cyanobacterium) as the recipient organism (in this case *S. elongatus*). In principle, xenologs from phyla other than cyanobacteria should be more divergent in terms of composition due to their greater phylogenetic distance from the recipient organism.

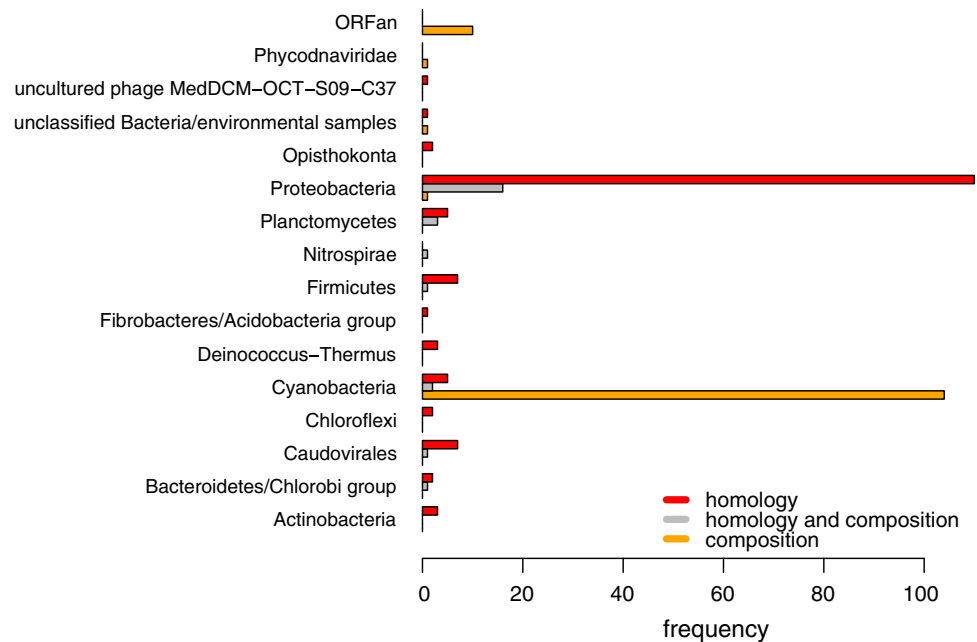
In addition, with only a few exceptions, xenologs detected by compositional methods tended to cluster in the genome separately from xenologs detected by BLASTp and DarkHorse, which also tended to cluster (Fig. 4). In fact, if we disregard the xenologs identified by compositional and homology-based methods (the 25 genes in the intersection in Fig. 1), there were 47 cases in which a xenolog detected by a homology-based method was next to another xenolog detected by the same method, 57 cases in which a xenolog detected by a compositional method was next to a xenolog detected by a compositional method, and only 6 cases where a xenolog detected by a homology method was next to another xenolog detected by the compositional method. Our findings reinforce the fact that these gene sets have different phylogenetic origin.

We then proceeded to compare the expression levels of the identified xenologs with the expression levels of the remainder of the genes in the genome based on the expression levels of *S. elongatus* genes reported by Vijayan et al. 2011. For the sake of comparison, we classified all *S. elongatus* genes into the following six categories: (i) ORFans (genes without known homologs); (ii) genes participating in the two photosynthetic reaction centers; (iii) ribosomal protein coding genes; (iv) xenologs identified by homology-based methods (BLASTp and

**Fig. 2** Functional classification of xenologs. Functional classification followed that of COG in Cyanobase (<http://genome.microbedb.jp/cyanobase>)



**Fig. 3** Origin of *S. elongatus* xenologs. The origin of xenologs is suggested by best BLASTp hit criteria



DarkHorse); (v) xenologs identified by compositional methods (HGT-DB); and (vi) a category comprising all other genes in the genome. Because the xenologs identified were most likely introduced into the genome after the divergence of *S. elongatus* from the other sequenced cyanobacteria, we refer to genes other than xenologs as “native genes” in contrast to the genes acquired more recently by lateral transfer.

Generally, we found that xenologs detected in *S. elongatus* by homology-based methods had lower expression levels than native genes in the genome (Fig. 5; Table 1). Furthermore, the differences in gene expression levels were statistically significant ( $p$  value =  $8.976e^{-16}$ ), and the differences were significant even when genes belonging to a large phage island (from position 718666 to 759037 in the genome) were removed from the dataset ( $p$  value =  $1.164e^{-14}$ ). To improve our understanding of the statistical significance of the results, we randomly selected 174 *S. elongatus* genes and performed a one-sided Wilcoxon test to evaluate whether the expression levels of these selected genes differed significantly from the remainder of the genes in the genome. We then repeated this procedure a thousand times, the results of which are provided in Supplementary Fig. 1. In addition, an empirical cumulative distribution clearly shows that xenologs detected by compositional methods had lower expression levels (Fig. 6).

The mean expression level of xenologs detected by compositional methods was also lower than the mean expression level of native genes (Table 1). However, the difference was not statistically significant ( $p$  value = 0.09676). Similarly, the differences in the mean expression

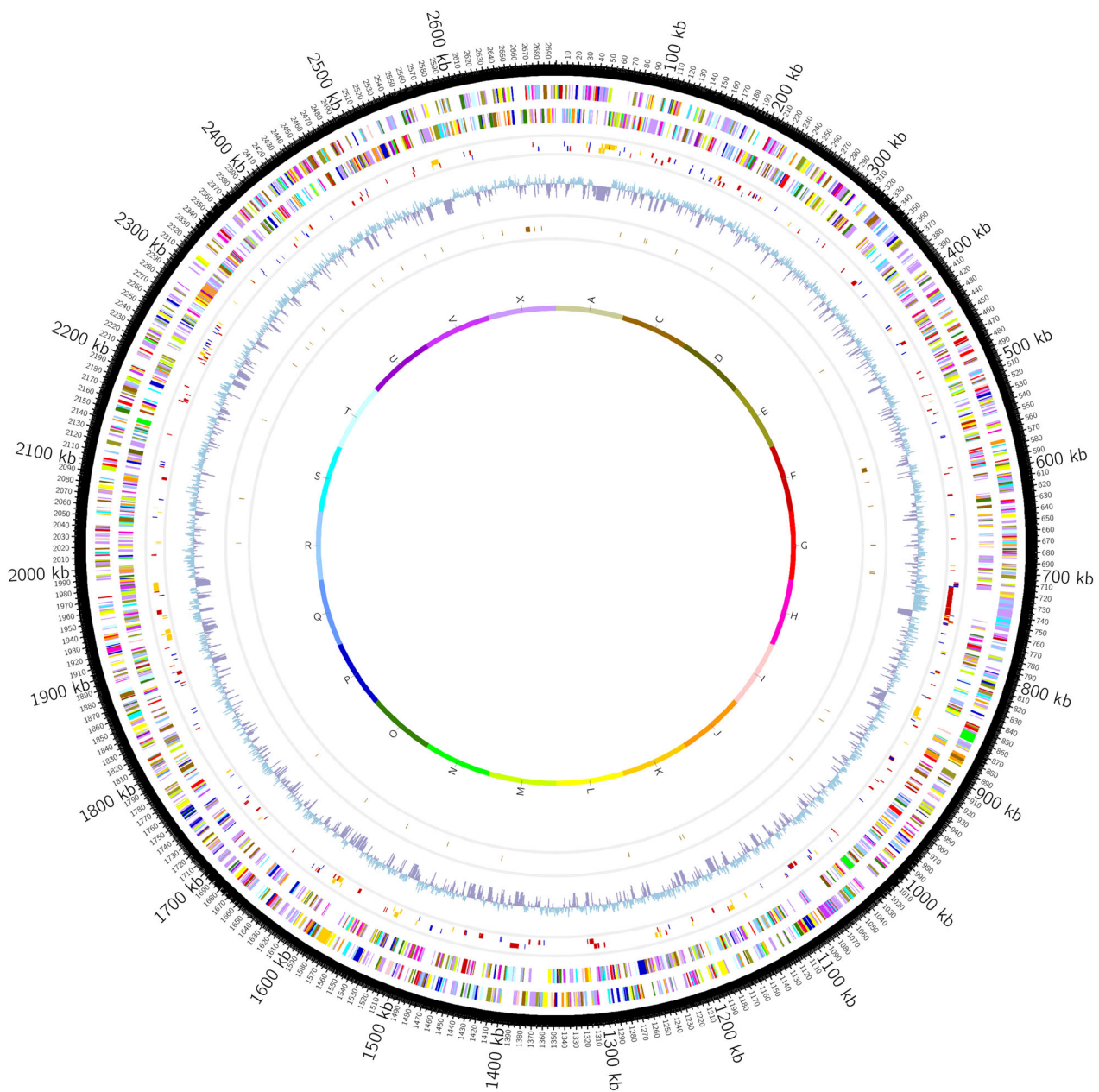
levels of ORFans and native *S. elongatus* genes were not statistically different ( $p$  value = 0.1916).

The difference in gene expression between xenologs detected by homology methods and native genes was not due to bias in RNA sequencing introduced by differences in genes sizes between both groups of genes. A statistical test revealed that the size of xenologs detected by homology methods was not statistically different from the size of native genes ( $p$  value = 0.1652). However, this finding did not hold for xenologs detected by compositional methods and ORFans, because these two groups of genes were significantly smaller than native genes (Table 1; Supplementary Fig. S2).

### Xenologs in *S. elongatus* Exhibit Low Codon Adaptation Index Values

We investigated whether the codon adaptation index (CAI; Sharp and Li 1987) of xenologs was also significantly lower than the CAI of native genes in *S. elongatus* because it has been argued that CAI values are a better indication of the overall expression level of a gene than the direct measurement of gene expression in a single laboratory condition (Fraser et al. 2004). However, CAI values are only meaningful in species that have undergone selection of codon usage for translation efficiency (Henry and Sharp 2007; Puigbò et al. 2008). Therefore, we first sought to determine the extent to which bias in codon usage among *S. elongatus* genes is determined by selection for translational efficiency.



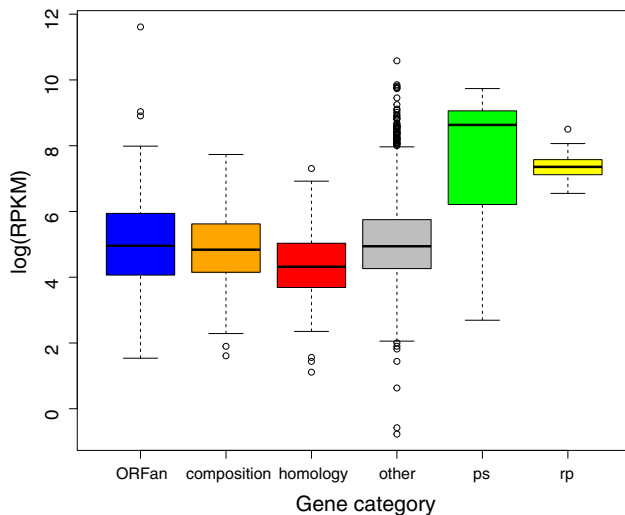


**Fig. 4** Localization of xenologs along the genome of *S. elongatus*. Starting with the outermost circle: (i) forward and reverse protein coding genes colored according to COG categories; (ii) xenologs identified by homology methods are in red, those identified by compositional methods in orange, and ORFans are shown in blue; (iii) deviation of G + C content from the mean; (iv) rRNA and tRNA genes; (v) color code for COG categories. (C) Energy production and conversion; (D) Cell cycle control, mitosis, and meiosis; (E) Amino acid transport and metabolism; (F) Nucleotide transport and metabolism; (G) Carbohydrate transport and metabolism;

(H) Coenzyme transport and metabolism; (I) Lipid transport and metabolism; (J) Translation; (K) Transcription; (L) Replication, recombination, and repair; (M) Cell wall/membrane biogenesis; (N) Cell motility; (O) Posttranslational modification, protein turnover, and chaperones; (P) Inorganic ion transport and metabolism; (Q) Secondary metabolites biosynthesis, transport, and catabolism; (R) General function prediction only; (S) Function unknown; (T) Signal transduction mechanisms; (V) Defense mechanisms; (X) Not in COGs (Color figure online)

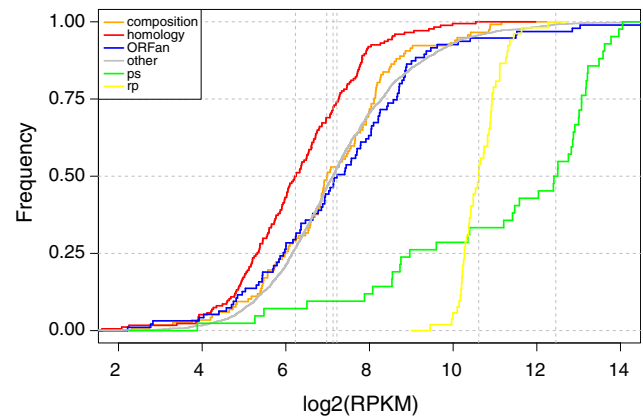
The calculation of a statistic known as  $S$  is one way to evaluate if there has been selection for translation efficiency among *S. elongatus* genes (dos Reis et al. 2004;

Sharp et al. 2005). However, different authors have calculated this statistic in different ways. In one case, dos Reis et al. (2004) constructed their statistic ( $S_i$ ) to measure the



**Fig. 5** Xenologs exhibit low expression levels. The expression level of xenologs identified by homology methods (BLASTp and Dark-Horse) was on average lower than the expression level of native genes ( $p$  value < 0.001). Each *box* denotes the median, *upper* and *lower* quartiles, the inter-quartile range (denoted by the *whiskers*), and the outliers (denoted by *dots*). The number of genes in each category is as follow: ORFans, 97; xenologs detected by compositional methods, 117; xenologs detected by homology methods, 174; other genes, 2126; genes from the reaction centers of the photosystem (ps), 50; ribosomal protein (rp) coding genes, 52

degree to which deviation from random codon usage among genes is explained by the diversity and abundance of coded tRNA.  $S_t$  is a correlation and can range in value from  $-1$  to  $1$ , where positive values of  $S_t$  (and values significantly distinct from  $0$ ) indicate that bias in codon usage among genes in the genome is due to selection for translation efficiency. Conversely, Sharp et al. (2005) designed the  $S$  statistic ( $S_r$ ) to measure the strength of the selection for translation efficiency in genes coding for ribosomal proteins.  $S_r$  can be a negative or positive value, and positive  $S_r$  values indicate a bias in codon usage due to selection for translation efficiency in ribosomal protein coding genes. A previous study suggested that both statistics are correlated (Sharp et al. 2005). However,  $S_r$  is more sensitive because it focuses only on a sub-set of highly expressed genes, whereas  $S_t$  is more stringent



**Fig. 6** Empirical cumulative distribution curves. Xenologs identified by homology methods exhibited lower expression levels than genes in other categories. *Vertical lines* denote the median of each category

because it requires that the bias in codon usage extends to most coding genes in the genome (Sharp et al. 2005).

The  $S_t$  statistic indicated that the bias in codon usage among genes in the *S. elongatus* genome was not correlated to the diversity and abundance of coded tRNA ( $S_t = 0.068$ ; Supplementary Fig. S3). Specifically, the bias in codon usage along the genome was not explained by selection for translation efficiency. However,  $S_r$  exhibited a positive value ( $S_r = 0.744$ ), which suggested that ribosomal protein coding genes display codon usage bias when compared to the rest of genes in the genome. In comparison, *E. coli* K12 DH10B, which is known to be under a strong selective pressure for translation efficiency, exhibits an  $S_t$  value =  $0.717$  and  $S_r = 1.481$ , whereas *Buchnera aphidicola* Aps, a mutualistic endosymbiotic bacteria in aphids that is known to evolve by neutral process due to recurrent population bottle necks and lack of recombination, exhibits and  $S_t$  value =  $0.01$  and  $S_r = -0.063$ .

To understand the discrepancy between  $S_t$  and  $S_r$  statistics in *S. elongatus*, we performed a correspondence analysis (COA) on codon usage (Peden 1999). Essentially, COA creates a series of orthogonal axes used to identify trends in data variation, with each subsequent axis explaining a decreasing amount of variation. When applied to the study of the variation in relative synonymous codon usage (RSCU) among *S. elongatus* genes, COA revealed

**Table 1** Gene expression levels and CAI values of xenologs detected by homology or compositional methods

| Category            | $N$  | Median gene size | Median expression level (RPKM) | Median codon adaptation index (CAI) |
|---------------------|------|------------------|--------------------------------|-------------------------------------|
| Homology            | 174  | 717              | 75.04*                         | 0.566*                              |
| Compositional       | 117  | 564*             | 126.07                         | 0.440*                              |
| ORFan               | 96   | 322*             | 141.62                         | 0.554*                              |
| Long resident genes | 2126 | 834              | 146.52                         | 0.594                               |

An asterisk denotes significance at  $p < 0.001$  when compared with long resident genes (one-sided Wilcoxon test)

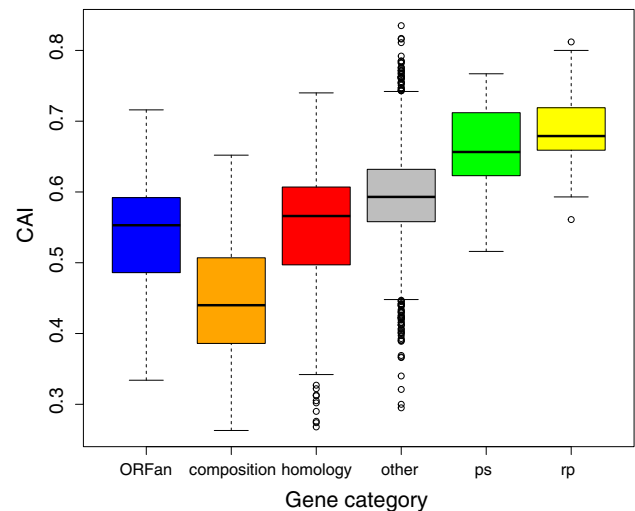
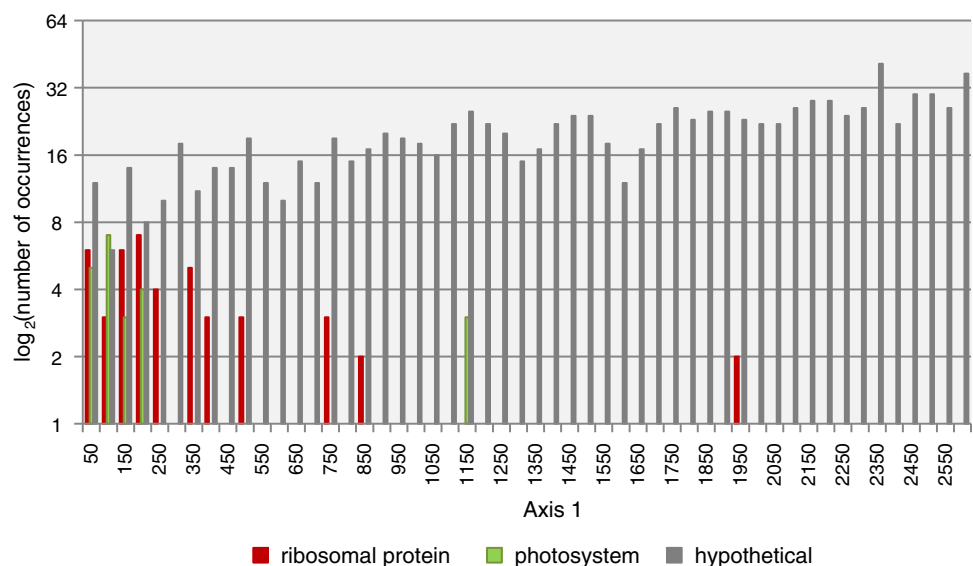
that the first axis, which explained 12 % of all variation, strongly correlated with G + C content at the third codon positions ( $GC_{3s}$ ) [correlation coefficient ( $r$ )  $\approx -0.86$ ,  $p < 0.001$ ; Supplementary Fig. S4], which demonstrated that G + C content has a large influence on codon usage bias in *S. elongatus*.

Nevertheless, two other aspects of the first axis are better explained by selection of codon usage for translation efficiency. First, codons predicted as optimal by COA in *S. elongatus* are also optimal in other species known to be under selection for translation efficiency (Supplementary Fig. S5; Peden 1999). Second and most important, there is an enrichment of ribosomal protein coding genes at one extreme of the first axis of COA values and of hypothetical genes at the other extreme (Fig. 7).

Previous results clearly indicate that although G + C content largely affects bias in codon usage among *S. elongatus* genes (in accordance with  $S_t$  close to 0), selection for translation efficiency has been strong enough to cause a pattern in RSCU variation among ribosomal protein coding genes (as shown by the positive value of  $S_r$  and the enrichment of ribosomal protein coding genes at one extreme of the first COA axis), which renders CAI values meaningful because it indicates the selection for a sub-set of codons in highly expressed genes. Once we demonstrated that selection for translation efficiency in ribosomal protein coding genes in *S. elongatus* had occurred, we evaluated the CAI values of xenologs to determine if the values were low relative to native genes.

As expected, the xenologs detected by compositional methods exhibited the lowest CAI values (Fig. 8), followed by ORFans, and then by xenologs detected using homology methods. As shown in Table 1, median CAI values of xenologs detected by compositional as well as by homology methods were statistically different from the median

**Fig. 7** The localization of genes coding for ribosomal proteins and for photosynthetic apparatus proteins along the first axis of codon variation in the COA analysis suggests selection for translation efficiency. Occurrence of genes coding for ribosomal proteins (*red*), for proteins participating in the photosynthetic reaction centers (*green*), and for hypothetical proteins (*gray*) along axis 1 of COA. Each division of axis 1 denotes clusters of 50 genes (Color figure online)



**Fig. 8** Xenologs exhibit low CAI values. The CAI values of ORFans and xenologs were on average lower than the CAI values of the rest of the genes in the genome ( $p$  value  $< 0.001$ ). Each *box* denotes the median, upper, and lower quartiles, the inter-quartile range (denoted by the *whiskers*), and the outliers (denoted by *dots*). The number of genes in each category was as follows: ORFans, 97; xenologs detected by compositional methods, 117; xenologs detected by homology methods, 174; other genes, 2126; genes from the reaction centers of the photosystem (ps), 50; ribosomal protein (rp) coding genes, 52

CAI value of native genes ( $p$  values  $2.2e^{-16}$  and  $1.356e^{-12}$ , respectively). Additionally, the mean ORFans CAI value was also statistically different from native genes ( $p$  value =  $1.312e^{-12}$ ).

### Contribution of Xenologs to *S. elongatus* Physiology and Adaptation

The xenologs detected exhibited decreased levels of gene expression compared to other genes in the genome. In

particular, the mean expression level of xenologs detected by homology methods was even lower than ORFan expression levels, which directed us to explore whether or not the xenologs contributed to *S. elongatus* survival or were only transient genes waiting to be lost from the genome. To assess whether the xenologs contributed to *S. elongatus* fitness, we investigated if they changed their expression levels in previous experiments involving iron (Fe) starvation. Changes in *S. elongatus* gene expression levels during Fe starvation have been studied using whole genome DNA microarray technology (Nodop et al. 2008). Low Fe levels frequently occur in aquatic habitats and severely limit biomass production in photosynthetic organisms (Tortell et al. 1999). Hence, genes that increase their expression level during Fe starvation might confer an adaptive advantage for *S. elongatus* under this condition.

The Fe starvation acclimation study performed by Nodop et al. (2008) resulted in the identification of 64 genes with significantly increased transcription levels after 72 h, and 24 genes with diminished expression levels at the same time point. According to our analysis, two of the genes with increased transcription levels were ORFans, and six were xenologs (Table 2). Interestingly, the genes *isiA*, *isiB*, and *isiC*, which are located in a single operon (Supplementary Fig. S6), were among the genes with the highest increases in expression level upon iron starvation. These xenologs were detected by compositional methods and were acquired from other cyanobacteria according to best BLASTp hit criteria. Prior studies have indicated that IsiA participates in several functions, including protection against the deleterious effects of light (Ihalainen et al. 2005), whereas IsiB is a flavodoxin expressed under Fe limitation in cyanobacteria and algae (Lodeyro et al. 2012), and IsiC is homologous to the *E. coli* protein FesA, which hydrolyzes ester bonds of internalized ferrin-enterobactin siderophores (Andrews et al. 2003).

Another xenolog that significantly increased in expression level during Fe starvation was *ccoN*, which codes for cytochrome oxidase *cbh*<sub>3</sub>-type subunit I (Synpcc7942\_0202). This gene is located in the same transcription unit as cytochrome oxidase c subunit (*ccoO*), which is also a

xenolog and likewise exhibits changes in expression level under Fe starvation, but the differences are not significant. The xenolog *ccoN* has a homolog in *S. elongatus* designated *ctaD* (Synpcc7942\_2603) that codes for a cytochrome-c oxidase *aa*<sub>3</sub>-type. Phylogenetic analyses revealed that *ccoN* was most likely obtained from proteobacteria by HGT, while *ctaD* has an evolutionary history that is consistent with 16SrRNA in which *S. elongatus* appears as the sister group of marine picocyanobacteria (*Prochlorococcus* and *Synechococcus* spp.; Supplementary Figs. S7, S8). Interestingly, the expression level of *ctaD* does not change under Fe starvation, whereas changes in *ccoN* expression have been observed (Nodop et al. 2008). Clearly, different roles are played by *ccoN* and *ctaD* in the physiology of *S. elongatus* under Fe starvation conditions.

Other xenologs co-opted for cellular functions in *S. elongatus* and experimental evidence of their function exists for the genes from the *narA* locus (*moaA*, *moaE*, *moaCB*, and *moeA*). Together with *moaD*, which resides in the same locus, these genes participate in the biosynthesis of molybdopterin, a molybdenum cofactor required by nitrate reductase for nitrate reduction, and are essential for nitrate reduction (Rubio et al. 1998). The *moaCDEA* genes are arranged in an operon, whereas *moeA* is coded next to the operon (Supplementary Fig. S6). According to the best BLASTp hit criteria, the whole locus appears to have been acquired from proteobacteria (Supplementary Material). Finally, an operon containing five genes related to the aerobic biosynthesis of cobalamin (*cobG*, *cobH*, *cobI*, *cobJ*, and *cobB*) was predicted by compositional methods to have been introduced into the genome by HGT (Supplementary Fig. S6). According to the best BLASTp hit criteria, these genes were most likely derived from *Oscillatoriohyphycideae* cyanobacteria.

## Discussion

We showed that on average, xenologs in *S. elongatus* have lower expression levels than native genes in the genome, which hold true for direct measurements of gene

**Table 2** Xenologs with a significantly increase in its expression level after 72 h of iron starvation

| Locus           | Gene        | Annotation  | Best BLAST hit       |
|-----------------|-------------|---|----------------------|
| Synpcc7942_0037 | –           | Hypothetical protein                                | ORFan                |
| Synpcc7942_0202 | <i>ccoN</i> | C <sub>bb3</sub> -type cytochrome oxidase subunit 1 | Gamma proteobacteria |
| Synpcc7942_1287 | –           | Hypothetical protein                                | Alpha proteobacteria |
| Synpcc7942_1540 | <i>isiC</i> | Putative hydrolase                                  | Cyanobacteria        |
| Synpcc7942_1541 | <i>isiB</i> | Flavodoxin  | Cyanobacteria        |
| Synpcc7942_1542 | <i>isiA</i> | Iron-stress chlorophyll-binding protein             | Cyanobacteria        |
| Synpcc7942_2150 | –           | Linear amide C–N hydrolase                          | Gamma proteobacteria |
| Synpcc7942_2169 | –           | Hypothetical protein                                | ORFan                |



expression levels by RNA sequencing (Vijayan et al. 2011) as well as for CAI values, which are indirect measures of gene expression. Hence, our results confirmed and expanded previous observations demonstrating that horizontally transferred genes exhibit low expression levels (Park and Zhang 2012).

In contrast to a previous study that employed data from microarray experiments to assess the expression level of genes (Park and Zhang 2012), we used RNA sequencing data, which is reported to be a more precise measure of gene expression (Vijayan et al. 2011). Despite this difference, our results were very similar to prior analyses in terms of relative gene expression levels between groups of genes. According to Park and Zhang (2012), the difference in expression levels of native genes in *E. coli* is approximately twofold greater than expression levels of recently acquired genes, which was in accord with the results of our analyses (Table 1).

In addition, Park and Zhang (2012) did not find an association between the acquisition of a gene by HGT and its classification as informational or operational. However, unlike the results of Park and Zhang (2012), we found a statistically significant association. Accordingly, we classified genes as informational if the gene belonged to the Translation (J), Transcription (K), or Replication, recombination, and repair (L) category within Clusters of Orthologous Groups (COG), or as operational if the gene belonged in any other category (Supplementary Table ST3). We then counted the number of times a gene was or was not a xenolog, and the number of times a gene was classified as informational or operational. Of the 291 xenologs reported, 12 were classified as informational and 291 as operational. The resulting *p* value from the Fisher's exact test for the distribution was  $<0.001$ , which suggested a statistically significant association. However, we acknowledge that additional in-depth studies are required to test the hypothesis that the successful transfer of genes that participate in operational processes occurs more often than the effective transfer of genes that function in informational processes (Jain et al. 1999).

Likewise, we assessed the extent to which selection for translation efficiency shaped codon usage bias among *S. elongatus* genes before using CAI, which differed from the approach of Park and Zhang (2012). As noted above, the fact that ribosomal protein coding genes cluster together in one of the extremes in axis 1 of the COA suggests that selection for translation efficiency was strong enough to influence codon usage variation despite the large bias due to G + C content. Although the prior analysis (Park and Zhang 2012) is sound, CAI values are meaningless if selection for translation efficiency did not shape codon usage bias among genes (Henry and Sharp 2007). We believe that an improvement to the Park and Zhang study

(2012) would be to first identify the genomes that demonstrated a bias in codon usage due to selection for translation efficiency among the 133 genomes analyzed and then to measure their respective CAI values.

### Adaptive Value of Xenologs

It has been argued that most successful xenologs are neutral or nearly neutral (Gogarten and Townsend 2005; Park and Zhang 2012), which is surprising given the effectiveness of natural selection in prokaryotes due to their large population sizes. However, the fact that xenologs tend to exhibit low expression levels supports previous claims regarding their relative neutrality, including the notion that genes that least affect fitness are the ones that have the highest probability of being accepted by the recipient cell for reasons that were previously described (Park and Zhang 2012).

Due to the large population sizes of bacteria, it is likely that only a small fraction of neutral xenologs is fixed in the population by genetic drift. For instance, population genetics theories predict that even beneficial mutations are lost to some extent by random factors once they appear by mutation (Gillespie 2004). Therefore, only those xenologs that confer a selective advantage will be maintained long term in the population or species. If this is the case, we imagine a scenario regarding fitness in which the xenome (the set of genes transferred between species) is functioning in the background and waiting for natural selection to choose a genetic variant when a favorable change in environmental or genetic conditions occurs. In cases where the presence of a xenolog is adaptive, natural selection could further modulate its expression to better suit the inner workings of the cell. This seems to be the case in xenologs that are up-regulated under iron starvation conditions. Furthermore, the adaptive process has been described as slow, as indicated by *E. coli* xenologs, for which it was suggested that fine-tuning expression spans approximately 8–22 million years (Lercher and Pál 2008).

Some of the xenologs detected appeared to confer an adaptive advantage for *S. elongatus*. As shown, the expression levels of six xenologs increased significantly in *S. elongatus* under Fe starvation (Nodop et al. 2008). In particular, *isiA*, *isiB*, *isiC*, and *ccoN* seemed to play a pertinent role in the acclimation to low Fe conditions. In addition, genes in the *moaCDEA* operon and *moeA* are essential for nitrate reduction. Similar examples of adaptation by HGT include the acquisition of several genes involved in salt tolerance in the halophilic bacterium *Salinibacter ruber* (Mongodin et al. 2005), and the adaptation to low phosphorous levels by the acquisition of genes for sulfoquinovose synthesis in *Bacillus coahuilensis* (Alcaraz et al. 2008). In particular, the finding that *ccoN*,

*moaA*, *moaC*, and *srpC* were acquired by HGT (Supplementary Figs. S10, S11, S12 and S13) was consistent with previous studies that demonstrated that xenologs are the main source of biochemical diversity used by prokaryotes to adapt to specific environments (Pál et al. 2005b; Treangen and Rocha 2011).

### Are Genes Detected by Compositional Methods True Xenologs?

As demonstrated, only a small fraction of the set of xenologs identified by homology methods intersected with the set of putative xenologs identified by compositional methods. In addition, the expression levels of genes with atypical base composition were not decreased, which cast doubts on the origin of these genes by HGT. Phylogenetic analysis of genes in the *isiA*, *isiB*, *isiC*, and the *cobG*, *cobH*, *cobI*, *cobJ*, and *cobB* operons did not exhibit phylogenies that were consistent with the 16S rRNA (Supplementary material Figs. S9 and S14, S15, S16, S17, S18, S19, S20 and S21). However, these trees had low corresponding bootstrap values, which could be explained by other processes, including differential gene loss and unrecognized paralogy. Hence, further analyses are needed to identify the fractions of genes with atypical sequence composition that are true xenologs.

### Where Do ORFans Come From?

The origin of ORFans is a controversial topic. One emerging view suggests that ORFans are acquired from phages by HGT (Daubin and Ochman 2004; Yu and Stoltzfus 2012). However, the fact that ORFans exhibit nearly the same expression level as native genes does not support their origin by xenology. Consequently, an alternative view suggests that ORFans are *de novo* genes (DeLaye et al. 2008; Kim et al. 2009; Fellner et al. 2014).

### Conclusion

Xenologs in *S. elongatus* derived from non-cyanobacterial sources exhibited decreased gene expression levels compared to native genes. Most of these xenologs appeared to originate from proteobacteria. Similarly, xenologs derived from cyanobacteria displayed decreased levels of gene expression, though the differences were not statistically significant. Surprisingly, the best BLAST hit criteria indicated that most of the putative xenologs detected by compositional methods were transferred from other cyanobacteria as suggested.

Our results confirmed previous reports of the relative neutrality of most xenologs. Nevertheless, some of the

xenologs detected were known to participate in cellular functions including the acclimation to Fe starvation or nitrate reduction, which corroborated the role of HGT in bacterial adaptation. In addition, ORFans exhibited nearly the same expression levels as native genes, which cast doubts regarding their origin by HGT. Although our results build upon and confirm those of Park and Zhang (2012), to our knowledge this is the first report of differences in gene expression levels between xenologs derived from both the same phylum and a phylum that is distinct from the recipient genome.

## Materials and Methods

### Xenolog Identification

Protein sequences from *S. elongatus* PCC 7942 were downloaded from the GenBank database ([www.ncbi.nlm.nih.gov/genbank/](http://www.ncbi.nlm.nih.gov/genbank/); Benson et al. 2011). The genome sequence data for this organism were produced by the US Department of Energy Joint Genome Institute ([www.jgi.doe.gov](http://www.jgi.doe.gov)). Two BLAST-based approaches were performed to identify potential xenologs. First, BLASTp (Camacho et al. 2009) searches were conducted locally for all *S. elongatus* PCC 7942 protein sequences against the non-redundant (nr) NCBI database ([ftp.ncbi.nlm.nih.gov/blast/db/FASTA/](http://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/)). Only the first 500 BLASTp hits were considered, and only homologs with an e-value smaller than 0.001 were retrieved. In this set of homologs, we identified all *S. elongatus* proteins showing only non-cyanobacterial homologs and classified the proteins as potential xenologs. We were also able to identify genes without homologs, also known as orphan genes (ORFans). The classification of ORFan genes is dependent on the representation of sequences in databases and on the method used to identify homologs. Accordingly, our approach to identify ORFans was dependent on the sequence content of the nr database and on BLASTp.

Secondly, we used the DarkHorse algorithm on the same nr database to identify putative xenologs in *S. elongatus* (<http://darkhorse.ucsd.edu>; Podell and Gaasterland 2007). As recommended by the authors for optimal DarkHorse performance, we adjusted the filter threshold by measuring the maximum number of candidate matches found for any single query in the genome for different threshold values and by identifying the threshold values at which the maximum number of candidate matches stabilized before arriving at the top number of BLAST hits (in this case 500). As shown in Supplementary Table ST4, the curve for *S. elongatus* was similar to the previously reported *Thermoplasma acidophilum* curve (Podell and Gaasterland, 2007), which indicated that filter values well below 0.1 are

appropriate for *S. elongatus*. Accordingly, we selected a filter threshold of 0.01, which resulted in the smallest slope, and then checked the LPI score histogram for *S. elongatus*. According to the histogram, queries with LPI scores below 0.6 were potential xenologs (see Supplementary Table ST5).

In addition, we tested the DarkHorse algorithm using two different parameter values pertaining to the percentage of the query that must align to the database sequence in order to classify the BLAST hit as a xenolog. The DarkHorse authors recommended using 70 % coverage of the query sequence as a threshold minimum (default option). However, if such a large cut-off threshold is applied, the DarkHorse algorithm might miss close homologs due to domain loss or recombination. Therefore, to avoid missing potential homologs, we ran the DarkHorse algorithm twice, first with 10 % and then with 70 % query coverage (with a filter threshold of 0.01 and an LPI score below 0.6 in both cases), and only selected the proteins detected using both approaches as putative xenologs. As shown in Supplementary Table ST6, greater than 90 % of the proteins identified as xenologs were found using both approaches, which resulted in the identification of 162 putative xenologs.

Finally, to identify xenologs in *S. elongatus* by compositional methods, we consulted the HGT-DB (<http://genes.urv.cat/HGT-DB/>). Briefly, HGT-DB is dedicated to the identification of potential xenologs based on the compositional properties of annotated genes. The HGT-DB methodology entails classifying all genes that exhibit atypical G + C content, codon usage, or amino acid composition of coded proteins as potential xenologs.

### Gene Expression Levels

Gene expression levels were obtained from column F of Supplementary Table 5 in Vijayan et al. (2011) and converted to reads per kilobase per million mapped reads (RPKM). Column F (Vijayan et al. 2011) contains the following information for each *S. elongatus* gene:

$$\frac{\text{(number of nucleotides covered by sequencing reads within a gene)}}{\text{(length of a gene)}}$$

Therefore, to obtain the number of reads mapped to a gene ( $C$ ), we multiplied this quantity by the corresponding gene length and then by the average length of the sequenced fragment (approximately 28 nt; Vijayan, personal communication). Next, RPKMs were calculated using the following formula:

$$\text{RPKM} = (10^9 \times C) / (N \times L),$$

where  $N$  is the total number of sequenced reads and  $L$  represents gene length.

### Correspondence Analysis (COA)

COA was performed using CodonW software as previously described (<http://codonw.sourceforge.net>; Peden 1999). Genes with fewer than 50 codons were initially excluded from the analysis to reduce the signal-to-noise ratio. Next, using Relative Synonymous Codon Usage (RSCU), COA vectors were generated from the genes with greater than 50 codons. After the vectors were generated, the genes with fewer than 50 codons were added to the COA. Finally, the RSCU of 5 % of the genes in the most extreme values of axis 1 were contrasted to identify optimal codons.

### Codon Adaptation Index (CAI)

The CAI values were calculated using CodonW software (Peden 1999).

### Codon Usage Bias

Selection for translation efficiency ( $S$ ) was evaluated using the  $S_i$  index developed by dos Reis et al. (2004) and the  $S_r$  index developed by Sharp et al. (2005). Both indexes were calculated using Perl scripts that are available upon request.

### Phylogenetic Reconstruction

Phylogenetic analyses were performed by first aligning the identified homologs by using MUSCLE v3.8 ([www.drive5.com/muscle/](http://www.drive5.com/muscle/); Edgar 2004). Next, the best-fit model of evolution was identified using PROTTEST v2.4 according to the Akaike Information Criterion ([http://darwin.uvigo.es/software/prottest2\\_server.html](http://darwin.uvigo.es/software/prottest2_server.html); Abascal et al. 2005) for each multiple alignment. Phylogenies for the *moaA*, *moaE*, *moaC*, *moeA*, *isiA*, *isiB*, *isiC*, *cobB*, *cobG*, *cobH*, *cobI*, and

*cobJ* protein coding sequences were reconstructed using PHYML v3.0 ([www.atgc-montpellier.fr/phyml/](http://www.atgc-montpellier.fr/phyml/)) following the best model of evolution detected by PROTTEST (search parameter BEST for nearest-neighbor interchange [NNI] and subtree pruning and regrafting [SPR], and six rate categories; Guindon et al. 2010). One hundred

bootstrap replicates were used for statistical support of branches. Finally, phylogenies were analyzed and visualized using the Python package E.T.E. v2.1 (<http://ete toolkit.org>; Huerta-Cepas et al. 2010) and MEGA5 (<http://www.megasoftware.net>; Tamura et al. 2011). The *ccoN* and *ctaD* phylogenies were analyzed in MEGA5.

## Genome Figure

The genome figure was constructed using Circos software (<http://circos.ca>; Krzywinski et al. 2009).

## Statistical Analysis

All statistical analyses were conducted in R ([www.r-project.org/](http://www.r-project.org/)).

**Acknowledgments** This work was supported by the Consejo Nacional de Ciencia y Tecnología CONACYT CB-2010-01 (Grant Number 157220). This work was also funded by Grants from the Spanish Ministry of Economy and Competitiveness (SAF2009-13032-C02-01 and SAF2012-31187), from Generalitat Valenciana (Prometeo/2009/092; Spain), and from ST-FLOW (EU). Work in the FdIC laboratory was financed by the Spanish Ministry of Economy and Competitiveness (BFU2011-26608) and the European Seventh Framework Program (612146/FP7-ICT-2013-10 and 282004/FP7-HEALTH-2011-2.3.1-2). We also thank Professor Peter Gogarten for his useful comments on the manuscript, and Professor Cei Abreu for useful comments on the gene expression analysis.

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

- Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21:2104–2105
- Alcaraz LD, Olmedo G, Bonilla G et al (2008) The genome of *Bacillus coahuilensis* reveals adaptations essential for survival in the relic of an ancient marine environment. *Proc Natl Acad Sci USA* 105:5803–5808
- Andrews SC, Robinson AK, Redriguez-Quinones F (2003) Bacterial iron homeostasis. *FEMS Microbiol Rev* 27:215–237
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2011) GenBank. *Nucleic Acids Res* 41(Database issue):D36–D42. doi:10.1093/nar/gks1195
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi:10.1186/1471-2105-10-421
- Daubin V, Ochman H (2004) Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. *Genome Res* 14:1036–1042
- Delaye L, Deluna A, Lazcano A, Becerra A (2008) The origin of a novel gene through overprinting in *Escherichia coli*. *BMC Evol Biol* 8:31
- Delaye L, González-Domenech CM, García-Barcillan MP, Pereto J, de la Cruz F, Moya A (2011) Blueprint for a minimal photoautotrophic cell: conserved and variable genes in *Synechococcus elongatus* PCC 7942. *BMC Genomics* 12:25. doi:10.1186/1471-2164-12-25
- Dong G, Golden SS (2008) How a cyanobacterium tells time. *Curr Opin Microbiol* 11:541–546
- dos Reis M, Savva R, Wernisch L (2004) Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res* 32:5036–5044
- Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797
- Encinas D, Garcillán-Barcia MP, Santos-Merino M, Delaye L, Moya A, de la Cruz F (2014) Plasmid conjugation from proteobacteria as evidence for the origin of xenologous genes in cyanobacteria. *J Bacteriol* 196:1551–1559
- Fellner L, Bechtel N, Witting MA, Simon S, Schmitt-Kopplin P, Keim D, Scherer S, Neuhaus K (2014) Phenotype of *htgA* (*mbiA*), a recently evolved orphan gene of *Escherichia coli* and *Shigella*, completely overlapping in antisense to *yaaW*. *FEMS Microbiol Lett* 350:57–64. doi:10.1111/1574-6968.12288
- Fournier GP, Gogarten JP (2008) Evolution of acetoclastic methanogenesis in *Methanosarcina* via horizontal gene transfer from cellulolytic *Clostridia*. *J Bacteriol* 190:1124–1127
- Fraser HB, Hirsh AE, Wall DP, Eisen MB (2004) Coevolution of gene expression among interacting proteins. *Proc Natl Acad Sci USA* 101:9033–9038
- García-Vallvé S, Guzman E, Montero MA, Romeu A (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res* 31:187–189
- Gillespie JH (2004) Population genetics a concise guide, 2nd edn. The Johns Hopkins University Press, Baltimore
- Gogarten JP, Townsend JP (2005) Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* 3:679–687
- Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59:307–321
- Henry I, Sharp PM (2007) Predicting gene expression level from codon usage bias. *Mol Biol Evol* 24:10–12
- Huerta-Cepas J, Dopazo J, Gabaldón T (2010) ETE: a python environment for tree exploration. *BMC Bioinformatics* 11:24. doi:10.1186/1471-2105-11-24
- Ihalainen JA, D’Haene S, Yermenko N, van Roon H, Arteni AA, Boekema EJ, van Grondelle R, Matthijs HC, Dekker JP (2005) Aggregates of the chlorophyll-binding protein IsiA (CP43’) dissipate energy in cyanobacteria. *Biochemistry* 44:10846–10853
- Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci USA* 96:3801–3806
- Kim W, Silby MW, Purvine SO, Nicoll JS, Hixson KK, Monroe M, Nicora CD, Lipton MS, Levy SB (2009) Proteomic detection of non-annotated protein-coding genes in *Pseudomonas fluorescens* Pf0-1. *PLoS ONE* 4:e8455. doi:10.1371/journal.pone.0008455
- Koksharova OA, Wolk CP (2002) Genetic tools for cyanobacteria. *Appl Microbiol Biotechnol* 58:123–137
- Kondo T, Strayer CA, Kulkarni RD, Taylor W, Ishiura M, Golden SS, Johnson CH (1993) Circadian rhythms in prokaryotes: luciferase as a reporter of circadian gene expression in cyanobacteria. *Proc Natl Acad Sci USA* 90:5672–5676
- Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. *Genome Res* 19:1639–1645. doi:10.1101/gr.092759.109
- Lazcano A (2011) Natural history, microbes and sequences: shouldn’t we look back again to organisms? *PLoS ONE* 6(8):e21334. doi:10.1371/journal.pone.0021334
- Lercher MJ, Pál C (2008) Integration of horizontally transferred genes into regulatory interaction networks takes many millions of years. *Mol Biol Evol* 25:559–567



- Lodeyro AF, Ceccoli RD, Pierella Karlusich JJ, Carrillo N (2012) The importance of flavodoxin for environmental stress tolerance in photosynthetic microorganisms and transgenic plants. Mechanism, evolution and biotechnological potential. *FEBS Lett* 586:2917–2924. doi:[10.1016/j.febslet.2012.07.026](https://doi.org/10.1016/j.febslet.2012.07.026)
- Macleán RC, Hall AR, Perron GG, Buckling A (2010) The evolution of antibiotic resistance: insight into the roles of molecular mechanisms of resistance and treatment context. *Discov Med* 10:112–118
- Mongodin EF, Nelson KE, Daugherty S et al (2005) The genome of *Salinibacter ruber*: convergence and gene exchange among hyperhalophilic bacteria and archaea. *Proc Natl Acad Sci USA* 102:18147–18152
- Nägeli KW (1849) *Gattungen einzelliger Algen*. Schulthess, Zurich
- Nodop A, Pietsch D, Höcker R, Becker A, Pistorius EK, Forchhammer K, Michel KP (2008) Transcript profiling reveals new insights into the acclimation of the mesophilic fresh-water cyanobacterium *Synechococcus elongatus* PCC 7942 to iron starvation. *Plant Physiol* 147:747–763
- Pál C, Papp B, Lercher MJ (2005a) Horizontal gene transfer depends on gene content of the host. *Bioinformatics* 21(Suppl 2):ii222–ii223
- Pál C, Papp B, Lercher MJ (2005b) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* 37:1372–1375
- Park Ch, Zhang J (2012) High expression hampers horizontal gene transfer. *Genome Biol Evol* 4:523–532
- Peden, J (1999) Analysis of Codon Usage. PhD. thesis. University of Nottingham. <http://codonw.sourceforge.net>. Accessed 15 May 2013
- Podell S, Gaasterland T (2007) DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol* 8(2):R16
- Podell S, Gaasterland T, Allen EE (2008) A database of phylogenetically atypical genes in archaeal and bacterial genomes, identified using the DarkHorse algorithm. *BMC Bioinformatics* 9:419
- Puigbò P, Romeu A, García-Vallvé S (2008) HEG-DB: a database of predicted highly expressed genes in prokaryotic complete genomes under translational selection. *Nucleic Acids Res* 36(Database issue):D524–D527
- Rubio LM, Flores E, Herrero A (1998) The *narA* locus of *Synechococcus* sp. strain PCC 7942 consists of a cluster of molybdopterin biosynthesis genes. *J Bacteriol* 180:1200–1206
- Rust MJ, Golden SS, O’Shea EK (2011) Light-driven changes in energy metabolism directly entrain the cyanobacterial circadian oscillator. *Science* 331:220–223
- Sharp PM, Li WH (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15:1281–1295
- Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE (2005) Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res* 33:1141–1153
- Sugita C, Ogata K, Shikata M, Jikuya H, Takano J, Furumichi M, Kanehisa M, Omata T, Sugiura M, Sugita M (2007) Complete nucleotide sequence of the freshwater unicellular cyanobacterium *Synechococcus elongatus* PCC 6301 chromosome: gene content and organization. *Photosynth Res* 93:55–67
- Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28:2731–2739
- Tortell PD, Maldonado MT, Granger J, Price NM (1999) Marine bacteria and biogeochemical cycling of iron in the ocean. *FEMS Microbiol Ecol* 29:1–11
- Treangen TJ, Rocha EP (2011) Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet* 7(1):e1001284. doi:[10.1371/journal.pgen.1001284](https://doi.org/10.1371/journal.pgen.1001284)
- Vijayan V, Jain IH, O’Shea EK (2011) A high resolution map of a cyanobacterial transcriptome. *Genome Biol* 12(5):R47. doi:[10.1186/gb-2011-12-5-r47](https://doi.org/10.1186/gb-2011-12-5-r47)
- Yu G, Stoltzfus A (2012) Population diversity of ORFan genes in *Escherichia coli*. *Genome Biol Evol* 4:1176–1187
- Zhaxybayeva O, Doolittle WF (2011) Lateral gene transfer. *Curr Biol* 21(7):R242–R246. doi:[10.1016/j.cub.2011.01.045](https://doi.org/10.1016/j.cub.2011.01.045)