OXFORD UNIVERSITY PRESS Genome Biology and Evolution

Xenologs show low expression levels in the cyanobacterium Synechococcus elongatus

Journal:	*Genome Biology and Evolution*
Manuscript ID:	Draft
Manuscript Type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Delaye, Luis; Cinvestav Irapuato, Departamento de Ingenieria Genetica González-Domenech, Carmen; Universidad de Granada, Facultad de Farmacia De la Cruz, Fernando; Universidad de Cantabria-CSIC-SODERCAN, Departamento de Biología Molecular e Instituto de Biomedicina y Biotecnología de Cantabria (IBBTEC) Moya, Andres; Universitat de València, Instituto Cavanilles de Biodiversidad y Biología Evolutiva
Keywords:	Correspondence analysis, codon usage, horizontal gene transfer, translation efficiency

Xenologs show low expression levels in the cyanobacterium *Synechococcus elongatus*

Luis Delaye[1*], Carmen M González-Domenech[2], Fernando de la Cruz[3] and Andrés Moya[4]

[1*]Author for Correspondence: Luis Delaye, Departamento de Ingeniería Genética CINVESTAV-Irapuato, Km. 9.6 Libramiento Norte, Carretera Irapuato-León, 36821 Irapuato, Guanajuato, México. Tel: +52 (462) 6239669. Fax +52 (462) 6245846. LD: ldelaye@ira.cinvestav.mx.

[2]Facultad de Farmacia, Universidad de Granada, Campus de Cartuja s/n. 18071 Granada, Spain

[3]Departamento de Biología Molecular e Instituto de Biomedicina y Biotecnología de Cantabria (IBBTEC), Universidad de Cantabria-CSIC-SODERCAN, Santander, Spain.

[4]Instituto Cavanilles de Biodiversidad y Biología Evolutiva, Universitat de València, Spain.

Email addresses:

LD: ldelaye@ira.cinvestav.mx

CG: cmgodo@ugr.es

FC: fernando.cruz@unican.es

AM: andres.moya@uv.es

Abstract

Horizontal gene transfer is a central process in prokaryotic evolution. However, little is known about the factors which determine successful gene transfers. In particular, few studies have tackled the importance of gene expression levels with respect to their role in the success of a horizontally acquired gene. To answer this question, we studied the expression level of genes that clearly arrived by horizontal gene transfer into the genome of the cyanobacterium *Synechococcus elongatus* PCC 7942. We have found that xenologs show lower expression levels than the rest of the genes in the genome. We have further found that there is selection for translation efficiency in *S. elongatus*, and that the overall expression levels of xenologs, as suggested by CAI values, are low. We interpret this pattern as an indication that most successful horizontal gene transfers are those that least affect the fitness of the recipient cell. Finally, we discuss the case of the *ccoN* gene coding for a cytochrome oxidase as an example of a successful xenolog co-opted for (Fe) starvation in the context of selection of synonymous codon usage.

Keywords

Correspondence analysis, codon usage, horizontal gene transfer, translation efficiency

Background

Horizontal gene transfer (HGT) is recognized as one of the major processes in prokaryotic evolution (Zhaxybayeva et al. 2011). Starting on the micro-evolutionary scale, recent comparisons of 61 sequenced *Escherichia coli* genomes show that the core genome, i.e., the set of genes shared by all compared genomes, is comprised of only 993 gene families, while the pangenome, i.e., the set of genes shared by 61 or more genomes, is composed of 15,741 gene families (Lukjancenko, et al. 2010). This suggests that if the ancestral *E. coli* genome had approximately 4,000 genes, extant strains would have acquired about two-thirds of their genomes by HGT.

The role of HGT in the evolution of bacterial adaptation is clearly exemplified by higher levels of resistance to antibiotics among pathogenic bacteria (Maclean et al. 2010). Other 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 the sulfoquinovose synthesis in *Bacillus coahuilensis* (Alcaraz et al. 2008). Furthermore, the evolution by HGT of new metabolic pathways which degrade xenobiotics exemplifies the rapidity and efficacy of this adaptive process (Springael et al. 2004). Xenobiotics are defined as anthropogenic compounds which are believed to have been alien to extant microbial communities. For example, the *pcpB* gene, which initiates the degradation of pentachlorophenol (considered a 'true' xenobiotic), was acquired by the pentachlorophenol-degrading bacterium *Sphingomonas chlorophenolicum* (Tiirola et al. 2002) via HGT. More dramatic examples include the evolution of a pathway to degrade chlorobenzene in the *Ralstonia* spp. strain JS705. In this case, the two critical genes for chlorobenzene degradation originated from different bacterial species, one from

Ralstonia spp. strain JS745 and the other from *Pseudomonas* sp. strain B13 (Müller et al. 2003).

Finally, on the macro-evolutionary scale, the structure of the universal tree of life, as revealed by the 16/18S rRNA molecule, is being challenged by the impact of HGT in prokaryotic evolution (Doolittle, 1999; Doolittle, 2007). How to represent prokaryotic evolution in light of HGT is, as of today, a work in progress (Lazcano et al. 2011).

Despite the importance of HGT in prokaryotic evolution, the factors determining the success of acquired genes are not totally understood. Once a protein coding gene is acquired by a genome via HGT, its protein product interacts with the rest of the molecules present in the intracellular environment. The fate of the new xenolog depends upon what effects these molecular interactions have on the fitness of the recipient cell. For example, an *in silico* analysis of the metabolic network of *E. coli* has shown that the chance of acquiring a gene by HGT is up to six times higher if an enzyme that catalyses a coupled metabolite flux is already encoded in the genome (Pál et al. 2005). Another factor that seems to determine the success of a transferred gene is codon usage. In species where there has been selection on codon usage for translation efficiency, there will be a bias in the codon composition of highly expressed genes when compared to the codon usage of the rest of genes in the genome (Henry et al. 2007 and references therein). For instance, the codon adaptation index (CAI) uses this bias to identify highly expressed genes (Sharp et al. 1987). When a gene is acquired by a genome via HGT, it is very likely that its codon usage differs from the average codon usage (or the optimal codon usage) of the genes in the recipient genome. It has been proposed that with the passing of time, xenologs acquire the codon composition of the recipient genome in a process known as amelioriation (Lawrence et al. 1997). This is because xenologs suffer the same mutation pressure as the rest of the genes in the genome. However, a different

interpretation has been given to the observation that xenologs (at least some of them) resemble the codon usage of recipient genes. Accordingly, codon usage compatibility between foreign genes and recipient genomes increases the fixation probability of HGT events (Medrano-Soto et al. 2004).

Recently, it has been shown that expression level hampers HGT in *Escherichia coli* (Park et al. 2012). It was proposed that xenologs deteriorate the fitness of the recipient cells due to: a) energy expenditure in transcription and translation; b) cytotoxic protein misfolding; c) reduction in cellular transcriptional efficiency; d) detrimental protein interaction; and e) disturbance of the optimal protein concentration or cell physiology (Park et al. 2012). Here we show that the xenologs that have been transferred into *Synechococcus elongatus* PCC 7942 from non-cyanobacterial species also show low expression levels, thus corroborating and expanding previous results for *E. coli* (Park et al. 2012).

Results

We found that xenologs in *Synechococcus elongatus* PCC 7942 (hereafter *S. elongatus*) have, on average, lower expression levels than the rest of the genes in the genome (Figure 1). The difference is statistically significant at the $P < 0.001$ level (one-sided Wilcoxon test). This difference is statistically significant even when genes belonging to a large phage island are removed from the dataset. A summary of xenologs here identified is shown in Figure 2, and a more complete description of them is found in Supplementary Table 1.

It has been suggested that the Codon Adaptation Index (CAI) is a better indication of the overall expression level of a gene than the direct measurement of its expression on a

single laboratory condition (Fraser et al. 2004). To determine if xenologs also show low CAI values, we first studied to what extent bias in codon usage among *S. elongatus* genes is determined by G+C mutational bias. This is because CAI values are only meaningful in species where there has been selection of codon usage for translation efficiency (Henry et al. 2007; Puigbò et al. 2008).

One way to evaluate the dependence of bias in codon usage to G+C content is by using an effective number of codons (ENc) plot (Wright, 1990). Genes whose codon choice is constrained only by a G+C mutational bias will lie on or just below their predicted ENc values (see methods). The ENc plot for *S. elongatus* is shown in Figure 3. The correlation between the observed and "expected" ENc values is moderate (correlation coefficient r ~ 0.51). This result shows that variation in codon usage in *S. elongatus* is not entirely due to G+C content.

We have further studied the degree of determination of G+C content on the bias in codon usage among *S. elongatus* genes by performing a Correspondence Analysis (COA). In essence, COA creates a series of orthogonal axes in order to identify trends in data variation, with each subsequent axis explaining a decreasing amount of the variation (Peden, 1999 and references therein). When applied to the study of the variation in Relative Synonymous Codon Usage (RSCU) among *S. elongatus* genes, COA shows that the first axis strongly correlates with G+C content at third codon positions (GC_{3s}) (correlation coefficient r ~ -0.86, $p < 0.001$ Spearman rank correlation) and explains ~12% of all variation (Figure S1).

Although this result shows that G+C mutational bias impacts codon usage, several other aspects of the first axis are better explained by selection of codon usage for translation efficiency. These are: a) that there is an enrichment of ribosomal protein coding genes at

one extreme of the values of the first axis, and of hypothetical genes at the other extreme; b) codons predicted to be optimal by COA in *S. elongatus* are also optimal in other species known to be under selection for translation efficiency; c) the principal axis also correlates with observed ENc values (correlation coefficient r ~ 0.43, p < 0.001 Spearman rank correlation), which is an independent measure of codon bias; and d) although slightly, axis 1 also correlates with gene expression (correlation coefficient r ~ -0.06, p < 0.001 Spearman rank correlation). (See supplementary information for a complete description of COA results, Figures S1 to S4 and Table S2.)

Previous results clearly indicate that, although bias in G+C mutation does affect bias in codon usage among *S. elongatus* genes, selection for translation efficiency has been strong enough to leave a pattern in RSCU variation among genes, rendering CAI values meaningful, i.e., there has been selection for a sub-set of codons in highly expressed genes. This is supported by the correlation between CAI values and expression level (Figure 4) (correlation coefficient r ~ 0.1, p < 0.001 Spearman rank correlation).

With the exception of ORPhans, CAI values of xenologs are on average lower than the CAI values of the rest of the genes in the genome (Figure 5). The difference is statistically significant at the P < 0.001 level (one-sided Wilcoxon test).

Discussion

Here we show that xenologs in *S. elongatus* have, on average, lower expression levels than the rest of the genes in the genome. This is true for direct measurements of gene expression levels by RNA sequencing (Vijayan et al. 2011) as well as for CAI values, which are an indirect measure of gene expression. Differing from the study by (Park et al. 2012), we have further validated the use of CAI by showing that selection for

translation efficiency has been strong enough to leave a mark in codon bias in *S. elongatus*. Our results confirm and expand previous observations showing that expression level is the principal factor determining the success of an HGT (Park et al. 2012).

It has been argued that most successful xenologs are neutral or nearly neutral (Park et al. 2012; Gogarten and Townsend 2005). This is surprising given the efficacy of natural selection on prokaryotes due to their large population sizes. However, the fact that xenologs tend to show low expression levels supports previous claims about their relative neutrality, i.e., genes which least affect fitness are the ones that are the most persistent once they are acquired by HGT. If this is the case, then it is possible to picture an image where the xenome (the set of genes frequently and successfully transferred between species) is running in the background in terms of fitness, waiting for natural selection to act and pick a lucky genetic variant when there is a favorable change in environmental conditions. In those cases where the presence of a xenolog is adaptive, natural selection can further modulate its expression level to better fit its product to the inner workings of the cell.

It is likely that the above scenario explains the presence of some of the xenologs in *S. elongatus*. For instance, *ccoN* (Synpcc7942_0202) and *catD* (Synpcc7942_2603) are two genes for cytochrome oxidases coded in *S. elongatus*. The first one, *ccoN*, was acquired by HGT, while the other, *catD*, shows a phylogeny consistent with a history of vertical inheritance (Supplementary material). Interestingly, it has been shown that the expression level of the gene *ccoN* significantly increases under (Fe) starvation (Nodop et al. 2008). This is not the case for *catD*. And, unexpectedly for a xenolog, the expression level of *ccoN* as measured by RNA sequencing (Vijayan et al. 2011) and its CAI value are larger than those of the indigenous cytochrome oxidase *ctaD*. One could

argue that the expression level of *ccoN* has been adapted by modifying its codon usage as an adaptation to (Fe) starvation. Consistent with this hypothesis is the fact that in phylogeny, the branch leading to *ccoN*, the number of d_S substitutions has sensibly increased with respect to the number of d_N changes (Supplementary material).

Other xenologs that may have been co-opted for cellular functions are Synpcc7942_0458 and Synpcc7942_0459, which are predicted to code for an S-formylglutathione hydrolase and a S-(hydroxymethyl)glutathione dehydrogenase. These genes are predicted to be in a single transcription unit according to bioinformatic predictions (Caspi et al. 2012) as well as experimental determinations (Vijayan et al. 2012). And, according to the KEGG database (Kanehisa et al. 2012), these enzymes catalyze two successive steps to synthesize formate from S-hydroxy-methyl-glutatione. However, they could also participate in the synthesis of cofactors. In another example, there is experimental evidence that the proteins coded by the xenologs *moaA* and *moaC* (Synpcc7942_1282 and Synpcc7942_1285 respectively) together with other genes in the *narA* locus, participate in molybdenum cofactor biosynthesis for nitrate reduction (Rubio et al. 1998). More studies are needed to find out if these enzymes were initially neutral and subsequently retained by natural selection.

Xenologs here identified are certainly a subset of all genes acquired by HGT in *S. elongatus*. This is because we only identified genes acquired from non-cyanobacterial origins. It remains to be seen whether this pattern also applies to xenologs acquired from other cyanobacteria.

Material and methods

Xenologs identification. The protein sequences from *S. elongatus* PCC 7942 were downloaded from Genbank database (Benson et al. 2011). The genome sequence data for this organism was produced by the US Department of Energy Joint Genome Institute (www.jgi.doe.gov). BLAST (Camacho et al. 2009) searches were conducted locally for all *S. elongatus* PCC 7942 protein sequences in the non-redundant (nr) database from NCBI (ftp.ncbi.nlm.nih.gov/blast/db/FASTA/). Only homologs showing an e-value smaller than 0.001 and a minimum of 70% coverage of the query relative to the sequence of the nr database were retrieved. For each *S. elongatus* protein, up to 50 homologs were considered. The above criteria were used to retrieve only the closest homologs to *S. elongatus* PCC 7942 protein sequences.

We then reconstructed a phylogeny for each set of homologs as follows. Identified homologs were aligned with MUSCLE v3.8 (Edgar, 2004), and for each multiple alignment, the best-fit model of evolution was identified with PROTTEST v2.4 according to the Akaike Information Criterion (Abascal et al. 2005). Phylogenies for each set of homologs were reconstructed with PHYML v3.0 following the best model of evolution detected with PROTTEST (search parameter BEST for NNI and SPR, and six rate categories) (Guindon et al. 2010). For statistical support of branches, we used aLRT. And finally, by using Perl scripts and the Python package E.T.E. v2.1 (Huerta-Cepas et al. 2010), we classified as *xenologs* all *S. elongatus* PCC 7942 proteins having homologs only from species other than cyanobacteria.

Effective Number of Codons (ENc) *plot*. We studied the relationship between codon and mutational bias with a formula that predicts the expected codon usage solely as a function of G+C in the third codon position (Wright, 1990).

$$ENc = 2 + S + \left(29/(S^2 + (1 - S)^2)\right)$$

Where S is the frequency of G+C in the third codon position. Genes whose codon choice is constrained only by a G+C mutational bias will lie on or just below the curve predicted by the above equation.

Correspondence Analysis (COA). The software CodonW was used for COA (Peden, 1999). Genes with fewer than 50 codons were initially excluded from the analysis to reduce signal noise. Then, by using the Relative Synonymous Codon Usage (RSCU), the vectors of the COA were generated from those genes with more than 50 codons. After vectors were generated, those genes with less than 50 codons were added to the COA. Finally, to identify optimal codons, the RSCU of the 5% of genes in the most extreme values of Axis 1 were contrasted.

Codon Adaptation Index (CAI). The CAI values were calculated by using CodonW software (Peden, 1999).

Rates of synonymous (d_S) and non synonymous (d_N) substitutions. To infer the number of synonymous and non-synonymous substitutions in the branch leading to the *ccoN* gene, we used CODEML from the PAML v4.4 package (Yang, 2007). We contrasted two models by using a Likelihood Ratio Test (LRT). The first model had one parameter ($\omega = d_N/d_S$) for all branches; and the second model had two parameters, one for the branch leading to *ccoN* ($\omega_1 = d_N/d_S$) and the other for the rest of the branches ($\omega_0 = d_N/d_S$).

Statistical analysis. All statistical analyses were conducted in R (www.r-project.org/).

List of abbreviations

COA, RSCU, LRT, CAI, ENc, HGT.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

LD identified homologs, carried the statistical analysis and drafted the manuscript. CG classified genes and drafted the manuscript. FC conceived the study and made important contributions to draft the manuscript. AM conceived the study and made important contributions to draft the manuscript. All authors read and approved the final manuscript.

Description of additional data files

Supplementary Fig. 1. G+C content at third codon position versus Axis 1 of COA.

Supplementary Fig. 2. Occurrence of genes coding for ribosomal proteins (red), for proteins participating in the photosystem (green) or for hypothetical proteins (gray) along Axis 1 of COA. Each bar denotes the number of occurrences along 50 genes.

Supplementary Fig. 3. Observed ENc values versus Axis 1 of COA.

Supplementary Fig. 4. Expression level versus Axis 1. Expression level is measured as Mean RNA Sequencing over ORF (MRSO).

Supplementary Fig. 5. The number of silent substitutions (d_S) has sensibly increased in the branch leading to *ccoN*. a) Maximum-Likelihood tree of *ccoN*. The branch leading to *ccoN* is indicated in green and with a hash tag (#). *ccoN* is indicated by the gi number

of the coded protein (81299013). The number of d_N and d_S substitutions is shown below the branch. Only bootstrap values ≥ 50 are shown; b) the distribution of d_N and d_S substitutions. The number of d_S substitutions in the branch leading to *ccoN* is indicated with an arrow. The tree with two parameters (omegas), one for the branch leading to *ccoN* and other for the rest of the branches, explains better the variation in d_N and d_S than the tree with only one parameter ($P = 0.0045$).

Supplementary table 1.

Xenologs in *S. elongatus* acquired from non-cyanobacterial sources. Contiguous genes in the genome are indicated in blue. MRSO: expression level in "Mean Sequences over ORF" according to (Vijayan et al. 2011). In violet we show genes belonging to a large phage related island.

Supplementary table 2. Position of each codon by Axis 1 in COA. Those codons that occur significantly more often ($p < 0.01$) are indicated in yellow, and those codons with $p < 0.05$ in green. Typically optimal and rare codons are from *E. coli*, *B. subtilis*, *S. cerevisiae*, *S. pombe*, and *D. melanogaster* (Peden, 1999 and references therein).

Supplementary file of xenologs. The phylogenies of all xenologs in newick format.

Authors' information

LD: Full professor of Genetics; CG: postdoctoral specialist in bioinformatics; FC: Full professor of Genetics; AM: Full professor of Genetics.

Acknowledgements

This work was supported by Consejo Nacional de Ciencia y Tecnología CONACYT CB-2010-01 [grant number 157220]. This work was further funded by the grants SAF2009-13032-C02-01, and SAF2012-31187 from the Spanish Ministry of Economy and Competitiveness, Prometeo/2009/092 from Generalitat Valenciana (Spain), and ST-FLOW (EU).

References

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics. 21:2104-2105.

Alcaraz LD, 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 U S A. 105:5803-5808.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. 2011. Nucleic Acids Res. 39(Database issue):D32-7.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10:421.

Caspi R, et al. 2012. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 40(Database issue):D742-53.

Doolittle WF. 1999. Phylogenetic classification and the universal tree. Science. 284:2124-2129.

Doolittle WF, Bapteste E. 2007. Pattern pluralism and the Tree of Life hypothesis. Proc Natl Acad Sci U S A. 104:2043-2049.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792-1797.

Fraser HB, Hirsh AE, Wall DP, Eisen MB. 2004. Coevolution of gene expression among interacting proteins. Proc Natl Acad Sci U S A. 101:9033-9038.

Gogarten JP, Townsend JP. 2005. Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol. 3:679-687.

Guindon S, et al. 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.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 40(Database issue):D109-14.

Lawrence JG, Ochman H. 1997. Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol. 44:383-397.

Lazcano A. 2011. Natural history, microbes and sequences: shouldn't we look back again to organisms? PLoS One. 6(8):e21334.

Lukjancenko O, Wassenaar TM, Ussery DW. 2010. Comparison of 61 sequenced *Escherichia coli* genomes. Microb Ecol. 60:708-720.

Maclean 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.

Medrano-Soto A, Moreno-Hagelsieb G, Vinuesa P, Christen JA, Collado-Vides J. 2004. Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes. Mol Biol Evol. 21:1884-1894.

Mongodin EF, et al. 2005. The genome of *Salinibacter ruber*: convergence and gene exchange among hyperhalophilic bacteria and archaea. Proc Natl Acad Sci U S A. 102: 18147-18152.

Müller TA, Werlen C, Spain J, Van Der Meer JR. 2003. Evolution of a chlorobenzene degradative pathway among bacteria in a contaminated groundwater mediated by a genomic island in *Ralstonia*. Environ Microbiol. 5:163-173.

Nodop A, et al. 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. 2005. Horizontal gene transfer depends on gene content of the host. Bioinformatics. 21 Suppl 2:ii222-3.

Park Ch, Zhang J. 2012. High expression hampers horizontal gene transfer. Genome Biol. Evol. 4: 523-532.

Peden J. 1999. (http://codonw.sourceforge.net).

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. vol. 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.

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.

Springael D, Top EM. 2004. Horizontal gene transfer and microbial adaptation to xenobiotics: new types of mobile genetic elements and lessons from ecological studies. Trends Microbiol. 12:53-58.

Tiirola MA, Wang H, Paulin L, Kulomaa MS. 2002. Evidence for natural horizontal transfer of the *pcpB* gene in the evolution of polychlorophenol-degrading sphingomonads. Appl Environ Microbiol. 68:4495-4501.

Vijayan V, Jain IH, O'Shea EK. 2012. A high resolution map of a cyanobacterial transcriptome. Genome Biol. 12:R47.

Wright F. 1990. The effective number of codons in a gene. Gene. 87:23-29.

Yang Z. 2007. PAML 4: a program package for phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24: 1586-1591

Zhaxybayeva O, Doolittle WF. 2011. Lateral gene transfer. Curr Biol. 21:R242-6.

Figure legends

Fig. 1. Xenologs have low expression levels. The expression level of xenologs is on average lower than the expression level of the rest of the genes of the genome (*p* value < 0.001, one sided Wilcoxon-Mann-Withney test). Each box denotes the median, upper and lower quartiles, the inter quartile range (denoted by the whiskers), and the outliers (denoted by dots), of the gene expression level distribution for each gene category. The number of genes in each category is as follow: (o) ORPhans, 136; (r) rest of the genes, 2273; (ps) genes from the photosystem, 50; (rp) ribosomal protein coding genes, 52; (x) xenologs, 101.

Fig. 2. Xenologs classified according to function. All genes related to phage shown here are found in a single locus in *S. elongatus*.

Fig. 3. ENc plot showing the relation between predicted and observed bias in codon usage according to G+C content. Red dots indicate the expected codon usage according to G+C content at third codon positions. Gray dots show observed codon usage for each gene.

Fig. 4. Expression level versus CAI. The number of genes in each category is as follow: (o) ORPhans in blue, 136; (r) rest of the genes in gray, 2273; (ps) genes from the photosystem in green, 50; (rp) ribosomal protein coding genes in red, 52; and (x) xenologs in orange, 101.

Fig. 5. Xenologs have low CAI values. The CAI values of xenologs is on average lower than the expression level of the rest of the genes of the genome (*p* value < 0.001, one sided Wilcoxon-Mann-Withney test). Each box denotes the median, upper and lower quartiles, the inter quartile range (denoted by the whiskers), and the outliers (denoted by dots), of the gene expression level distribution for each gene category. The

number of genes in each category is as follow: (o) ORPhans, 136; (r) rest of the genes, 2273; (ps) genes from the photosystem, 50; (rp) ribosomal protein coding genes, 52; (x) xenologs, 101.

Tables