DR. ASUNCION CONTRERAS (Orcid ID : 0000-0001-6591-9294)

Received Date : 22-Dec-2019

Revised Date : 14-Feb-2020

Accepted Date : 12-Mar-2020

Article type : Research Articles

TITLE

The default Cyanobacterial Linked Genome (dCLG): an interactive platform based on cyanobacterial linkage networks to assist functional genomics

AUTHORS

Jose, I. Labella, Antonio Llop, Asuncion Contreras

AFFILIATIONS

Departamento de Fisiología, Genética y Microbiología, Universidad de Alicante, Alicante, Spain.

Contact: contrera@ua.es

ABSTRACT

A database of Cyanobacterial Linked Genomes that can be accessed through an interactive platform (https://dfgm.ua.es/genetica/investigacion/cyanobacterial_genetics/dCLG/) was generated on the bases of conservation of gene neighborhood across 124 cyanobacterial species. It allows flexible generation of gene networks at different threshold values. The default cyanobacterial linked genome (dCLG), whose global properties are analyzed here, connects most of the cyanobacterial core genes. The potential of the web tool is discussed in relation to other bioinformatics approaches based on guilty-by-association principles, with selected examples of networks illustrating its usefulness for genes found exclusively in cyanobacteria or in cyanobacteria and chloroplasts. We believe that this tool will provide useful predictions that are readily testable in *Synechococcus elongatus* PCC7942 and other model organisms performing oxygenic photosynthesis.

KEYWORDS

Cyanobacteria; synteny; gene linkage; functional genomics; function prediction; COG; co-occurrence; DUF

INTRODUCTION

Cyanobacteria, phototrophic organisms performing oxygenic photosynthesis, constitute an ecologically and biotechnologically important phylum. The cyanobacterium *Synechococcus elongatus* PCC 7942 (hereafter *S. elongatus*), is a model system to address fundamental questions concerning the photosynthetic lifestyle. *S. elongatus* is so far the only photosynthetic organism for which the contribution of each gene to fitness has been evaluated [1]. Despite important breakthroughs in the genetic analysis of cyanobacteria, there is still a remarkable proportion of genes of unknown function in this phylum, many of which are presumably involved in functions relevant to the biology of cyanobacteria.

Bioinformatic approaches based on guilty-by-association principles such co-expression, synteny, co-occurrence, text-mining, protein binding or genetic interactions provide relevant clues or hypothesis to investigate gene function. STRING [2] and theSeed [3] are two popular web tools that take advantage of these principles to functionally connect proteins. However, the contribution of experimental data from cyanobacteria is very scarce and thus development of analytic tools that are not biased towards better studied phylogenetic groups is required to advance our understanding of cyanobacterial gene functions. In line with this, bioinformatics approaches relying on guilty-by-association strategies [4-6] have already been applied in the context of cyanobacterial genomes for particular purposes with specific outcomes. For instances, the Gecko3 tool [7] is specifically aimed at predicting gene or operon function and is particularly informative for, but limited to, robust and highly conserved genomic clusters that are present in the photoheterotrophic cyanobacterium *Synechocystis sp.* PCC 6803 (hereafter *Synechocystis*).

With the aim of contributing to cyanobacterial functional genomics, we took into account two indicators of evolutive pressure, conserved gene neighborhood and gene co-occurrence, in a phylogeny aware context to generate specific gene networks using as reference the gene set of *S. elongatus,* the model system for cyanobacterial genetics. An interactive graphical representation platform has been generated with the option of altering network size according to the synteny and connectivity parameters of the corresponding edges. The potential of our web tool to provide working hypotheses to speed functional genomics in cyanobacteria is discussed in relation to other bioinformatics approaches based on guilty-by-association principles.

RESULTS AND DISCUSSION

Generation of a Linkage Score reflecting gene order and co-occurrence in cyanobacteria

The workflow for construction of a cyanobacterial synteny network is summarized in Fig 1. *S. elongatus* translated Coding DNA Sequences (CDSs) were used as query in a BLASTp search against 124 cyanobacterial CDS sets (S1 Table). Specific blast hits were obtained establishing 2 filters: at least 50% alignment and 25% identity [8]. Whole CDS order was obtained from the BLASTp outputs for each cyanobacterial genome and used to calculate distances between all possible pairs of CDSs as the number of intergenic regions separating them.

Next, we generated an initial “Raw Linkage Score” for each CDS pair, reflecting both neighborhood and co-occurrence, which are indicators of functional connections, each one with pros and cons. In particular, gene neighborhood analysis requires a significant conservation of the genes involved, while co-occurrence analysis is not informative for highly conserved genes with no impact in the co-occurrence pattern. With this in mind, the score we present here takes into account the calculated number of genomes where synteny is observed (distance lower than 4) as a first descriptor of linkage, and corrects it by a co-occurrence ratio (the number of genomes where both CDSs are present divided by the number of genomes where at least one of the two CDSs of the pair is present)(eq. (1)). Because it relies on the frequency of neighborhood and occurrence events, the score increases with the number of closely related genomes and it is therefore very sensitive to phylogenetic bias. To minimize this bias, we applied a correction based on the average nucleotide identity (ANI), a good estimator of species boundaries [9] which, in contrasts with a phylogenetic tree from concatenated sequences, is also affected by genomic rearrangements, thus providing additional information on gene arrangement. As expected, groups of high genomic identity, formed by very close strains or substrains were detected using ANI scores (red squares in ANI heatmap in Fig 1). A total of 150500 scores were corrected (eq. (2)) and scaled to the interval [0-1]. Since the corrected Linkage Score (hereafter LS) is proportional to the number of genomes where ortholog CDS are separated less than 4 intergenic distances apart, the ortholog search requires that hits are independently located close-by in a significant number of genomes, providing robustness to the ortholog retrieval method.

A snapshot from the Cyanobacterial Linked Genome (CLG)

Once we obtained a LS for each pair of CDS (hereafter gene or protein), we studied how the generated relationships responded in terms of numbers of genes and individual networks included as a function of the LS value threshold. As expected, the total number of genes located in networks diminished as LS values increased, while the number of networks showed a more complex distribution (Fig 2a). Since the maximum number of individual networks (262) was reached at a LS threshold value of 0.3464, coinciding with the change in the slope reflecting genes decreasing in networks, this value was chosen to set the limits of the default Cyanobacterial Linked Genome (dCLG hereafter), 916 nodes connected by a maximum of 7 weighted edges or links (Fig 2b). For simplicity, networks were primarily numbered according to size.

Next, we compared the dCLG output with results obtained by other cyanobacterial works where co-occurrence [4] or synteny [7] was the main strategy. The co-occurrence analysis, which is not informative for highly conserved genes, recovered only 318 *S. elongatus* genes, of which 112 were within the 916 gene set of the dCLG. On the other hand, the synteny approach used for Gecko3 identifies only 167 *S. elongatus* genes, 138 within the dCLG, and here a relatively low number was anticipated on the basis of the restrictions imposed by the specific location of the corresponding genes in the *Synechocystis* genome, used as reference.

Since Gecko3 does not measure the direct relationship between pair of genes but the significance of their clusters, to make additional comparisons between dCLG networks and Gecko3 clusters, we used descriptors to order the 262 CLG networks according to their confidence: the average link score (ALS, eq. (3)) and links per node (LPN, eq. (4)) (S2 Table). Since the ALS reflects the general degree of linkage of the network but it is rather sensitive to the dispersion of the values, while the number of links per node LPN reflect the connectivity of the network but is affected by network size, a combined index (ALS x LPN) was considered more appropriated to reflect the robustness of dCLG networks. Accordingly, networks occupying the top positions always have their Gecko3 counterpart, and in some instances networks contain 2-3 Gecko3 clusters (see also Fig 2c), while the opposite did not occur. 233 out of the 262 dCLG networks did not have correspondence in Gecko3, in agreement with the higher connectivity of individual dCLG networks. It follows that by taking into account linkage in a large number of genomes and correcting the bias introduced by the occasional and apparently random dispersion of functionally related genes in individual cyanobacterial genomes, the CLG results informative for any cyanobacterium of interest.

A Web interface to provide open access and facilitate network analysis

To facilitate access to the information generated by the CLG, we built a web application (https://dfgm.ua.es/genetica/investigacion/cyanobacterial_genetics/dCLG/) that allows dynamic generation of CLG networks at the desired threshold value. From a query gene (COGXXXX ID or Synpcc7942_XXXX format) and for the user set threshold, the application generates a stable graphical representation for the corresponding network. Gene IDs and LS can be displayed in a table, with links to the corresponding entries on KEGG database. The threshold value of CLG networks can be modified at will, producing stabilized network displays based on the strength of the connections, thus providing a simple way of testing the strength of particular links, allowing elimination or incorporation of low confidence nodes. It should be noted that, in contrast to previous gene neighborhood analysis biased towards the cyanobacterium of reference [7], the CLG or the web tool detect networks of genes that can be scattered across the genome of *S. elongatus.*

Overlap between the dCLG and cyanobacterial core genomes

Highly conserved genes in a phylogenetic group, the so called core genes, are generally believed to be involved in essential functions for maintaining and replicating the cell, and there is a interrelationship between conservation and essentiality [5, 10, 11]. To gain insights into this interrelationship in cyanobacteria, we defined genes with conservation higher than 98% as persistent and took advantage of the experimentally determined classification of *S. elongatus* genes into essential and beneficial [1]. Persistent genes constitute 26% and 59% of the *S. elongatus* chromosome and the dCLG set, respectively (Fig 3a and S3 Figure). These proportions are, respectively, 25% and 47% for essential genes 6% and 9% for beneficial genes. In addition, 45% of the *S. elongatus* chromosome gene set and 54% of dCLG set are both persistent and essential, that is, coincidence is higher than expected at random (18.7% and 40%, respectively), supporting the interrelationship between conservation and essentiality and further indicating that conserved and essential genes have a high tendency to show synteny in cyanobacterial genomes.

To investigate possible coincidences with the cyanobacterial core genome, we compared the dCLG with four other sets of core genes previously reported [12-15]. As shown in Fig 4a, although each of the compared sets contain genes that are absent in the others, there is a significant overlap between all five gene sets and, importantly, most dCLG genes (790) were found in at least one of the other four gene sets. Although a significant number of genes (104) identified in core sets 1-4 were not in the dCLG, most of these gene nodes (83) have LS values of at least 0.3, and could thus be considered nodes of a slightly larger but less robust CLG. Importantly, repeating the analysis using just the 691 dCLG genes classified here as persistent increased the overlaps with all four core sets (Fig 4b), particularly at their intersection. The 126 dCLG genes absent from core sets 1-4 (S4 Table) correspond to relatively poorly intraconnected genes (1.03 links per node versus the dCLG average of 2.26) that are present in at least 51 of our 113 genomes. With the small exception of genes without a clear occurrence pattern, these non-core dCLG genes are mainly genes absent in *Prochlorococcus* genomes, and a fraction of them are also absent from marine *Synechococcus* species (S5 Figure). While the occurrence patterns of the non-core dCLG genes suggest their absence in many cyanobacteria is a consequence of genome size reduction and/or habitat specialization, their retrieval illustrates the potential of our bioinformatics approach to produce phylum specific networks of phylogenetic or environmental interest.

The dCLG is slightly enriched in polycistronic genes, many of which do not belong to conserved operons

As expected, our analysis favored detection of polycistronic genes, which constitute 55% of *S. elongatus* chromosomal genes [16], 66% of the dCLG and 73% of the non-core dCLG (Fig 3a). The connectivity between the different classes of dCLG genes (persistent, essential or polycistronic) was next analyzed alongside the subset of genes assigned with a COG function [17], for which we presumed no significant connectivity, since "know function" is not an intrinsic property of genes. As shown in Fig 3b, persistent and essential genes showed high intraconnectivity and low interconnectivity, mainly establishing links with other persistent and essential genes, respectively. Interestingly, a similar but far less pronounced bias was observed for polycistronic genes. While the finding that genes forming part of operons in *S. elongatus* do not show very high connectivity in cyanobacteria appears counter-intuitive, it suggests that genes involved in a given process are more likely to be separated by recombination when they participate in non-essential processes.

A significant number (approximately 33%) of the pairs of genes that form operons in *S. elongatus* are either absent (one or both genes of the pair) from the dCLG or found apart in up to 50 of the cyanobacterial genomes analyzed. Despite the reasonable assumption that co-transcription is *a priori* a more robust indicator of functional association than (cyanobacterial) synteny, the scenario revealed here challenges this view, suggesting that, for a subset of *S. elongatus* genes, co-expression could be less informative than synteny. A paradigmatic example of this is provided by *rbcX,* a gene encoding a Rubisco chaperonin [18] (Fig 5A).

In *S. elongatus*, functional connections between RbcX and Rubisco include co-localization and effects on carboxisome formation and spatial distribution [19]. In line with this, *rbcX*, a "non-core dCLG" (74% of conservation) from Group II is within a 3-node network with the Rubisco *rbcLS* genes (Fig 5, network 129). Lowering the LS value to 0.29 results in the inclusion of carboxisome and NAD(P)H dehydrogenase (NDH-1) genes (network 22), implicated in CO2 uptake [20-23]. Therefore, the extended network is consistent with a functional link between RbcX, Rubisco, carboxisomes and enzyme complexes involved in facilitation of CO2 uptake in *S. elongatus*, thus agreeing with the proven role of RbcX as a Rubisco chaperonin in organisms performing oxygenic photosynthesis. It seems that, for some of *S. elongatus* genes, CLG networks may be more informative than transcriptomic data.

Functional categories are differentially represented in the dCLG

Since the relationship between linkage and functional properties of persistent genes in bacterial genomes is rather complex [24] and tend to be associated with functional categories related to essential biological processes [25], we wondered to which extent functional categories were differently represented in the dCLG set. As shown in Fig 6a, categories J (translation, ribosomal structure and biogenesis) and H (coenzyme transport and metabolism) were the most represented in the dCLG set and the least represented in the non-core dCLG subset. The same pattern, but to a lesser extent, applies to category E (amino acid transport and metabolism). Categories C (energy production and conversion) and M (cell wall/membrane/envelop biosynthesis) were overrepresented in the dCLG and to higher extent in the non-core dCLG subset, while O (posttranslational modification, protein turnover, chaperones) was overrepresented in the dCLG, with a moderated contribution to the non-core dCLG subset. Categories T (signal transduction) and P (inorganic ion transport and metabolism) were the least represented in the dCLG.

To get additional insights into the observed bias, we calculated the intra- and inter-connectivity of nodes from the different COG categories (Fig 6b). The highest intra-connectivity was provided by two of the three categories overrepresented in the dCLG (J and C), while the other one (H) showed a relatively low intra-connectivity. These discrepancies may be explained by the different propensity of proteins in certain functional categories to engage in multiprotein complexes and the requirement of a precise stoichiometry for the translation and energy production machineries, that would not apply to coenzyme transport and metabolism.

Interestingly, the highest inter-connectivity was provided by one of two categories dramatically underrepresented in the dCLG (T), probably reflecting the plasticity inherent to regulatory processes and the diversity of interactions mediated by different signaling proteins. Exceptionally, few regulators were nevertheless highly connected. For instances, the two-component proteins with the highest LS values were the OmpR-type transcription factor RpaB (“regulator of phycobilisome association B”) [26] and the sensor histidine kinase NblS/Hik33 (“non-bleaching sensor”) [27], with LS values of 0.78 and 0.63 respectively. These two proteins, although cognate partners in a phosphorylation pathway are both part of a complex regulatory network controlling photosynthesis and other processes [28-32]. Both are essential and have been conserved across the cyanobacterial chloroplast barrier [31].

Our analysis shows that genes related to essential processes are kept together in the genome while, with remarkable exceptions worth investigating, regulatory genes and genes involved in dispensable metabolic pathways show little or no connectivity.

Function prediction for DUF proteins: comparing dCLG networks with STRING, Gecko3 and theSeed

Many of the genes or proteins classified as unknown function proteins contain DUF (domains of unknown function) domains, unique sequences and folds present in all kingdoms of life and representing one third of the current bacterial domains [33], for which there is no or very little information. When found as part of larger proteins, neighboring domains can provide clues to possible functions. However, for proteins containing exclusively DUF domains, synteny may provide the firsts clues to function prediction.

65 *S. elongatus* DUF genes (according to [34]) from the dCLG have no other recognizable domain, and we wondered whether their gene connections were coincident with those provided by STRING, Gecko3 and theSeed, for which we made several considerations. Gecko3 and theSeed rely exclusively on gene neighborhood, but they differ in output. While Gecko3 generates significance levels for gene clusters, theSeed generates individual scores for each gene pair, and thus the default option was considered appropriate for this analysis. STRING also generates scores for gene pairs based on multiple parameters. To ensure a reasonable level of confidence, the noise-introducing co-occurrence parameter of STRING was disabled, and the default score set to 0.7 (for the combined parameters of neighborhood, text-mining, databases, co-expression, experiments and gene fusion). Using STRING and theSeed we found respectively (S6 Table and Fig 7a): complete coincidence (28 and 23 cases), partial coincidence (13 and 17) and discrepancies (5 and 2 cases). In contrast, only 7 connections were retrieved by Gecko3, which is limited to the specific gene clusters present in the genome of reference (*Synechocystis)* genome.

Representative examples of relatively simple DUF-containing dCLG networks showing link strength and coincidences or discrepancies with link predictions by STRING and/or theSeed are shown in Fig 7b-e to illustrate cases in which: (i) DUF links are detected by all the methods, (ii) none of the alternative methods detect DUF connections and (iii) there is a discrepancy with STRING. The fact that some of the DUF genes have recently been "de-DUFed" is used to illustrate the pros and cons of the different prediction approaches.

Complete coincidence: DUF448 (Synpcc7942_2021), one of the 5 DUFs for which all four methods were fully coincident, is part of a highly scoring (ALSxLPN 2.5) 4-node network with *nusA*, *infB* and *rimP/DUF150* (Fig 7b) and constitutes an interesting example of conserved synteny. All four genes are present in most Bacteria and their linkage is conserved in many bacterial groups. They form a gene cluster in Gecko3 and the complete 4-node network with 6 links is also retrieved by STRING and theSeed. Their known functions involve, respectively, regulation of transcription termination/antitermination, (*nusA*) translation initiation *(infB)* and 30S subunit maturation (*rimP/DUF150)* [35], suggesting the involvement of this small basic protein in the global control of bacterial gene expression, a role recently confirmed in *B. subtillis*, where the corresponding protein (YlxR) is nucleoid associated and involved in regulation of multiple transcripts [36].

Connection only detected by the dCLG: DUF3067 (Synpcc7942_1077*)*, is detected as part of a CLG 4-node network with two cytochrome b6f subunits (CytF/PetA and PetC) and the C subunit (TatC) of the twin-arginine translocation (Tat) system, involved in the translocation of complex cofactor-containing (completely folded) proteins is required for targeting the Rieske iron-sulfur protein PetC into the thylakoid membrane in both chloroplasts [37] and cyanobacteria [38, 39]. This network (Fig 7c) is still detected at 0.43, that is, well above the dCLG threshold. While the interconnections TatC-PetA and PetC-PetA are predicted by STRING, only the CLG is able to make connections for Synpcc7942_1077. Our CLG-informed prediction is that DUF3067 proteins, whose orthologs are found in cyanobacteria and chloroplasts, and for which the first structural representative has very recently been solved by NMR [40], play a role in translocation of photosynthesis related proteins and/or regulation of cytochrome b6f activity.

Discrepancy: The 2-node network involving DUF561 (Synpcc7942_1137*)* and Synpcc7942_1138 (tRNA(Ile)-lysidine synthase (Fig 7d) is not detected by any of the alternative methods using the parameters defined above. STRING links Synpcc7942_1137 with Synpcc7942_0158 (rodanase-like) on the bases of gene arrangement, co-expression of putative homologs, and text mining derived from other bacterial phyla. However, the bacterial homologs supporting the STRING prediction do not contain a DUF561 domain, which is otherwise restricted to cyanobacteria and plants. Interestingly, the dCLG prediction connects two “plant” genes: *YCF23* (*synpc7942_1137*) and *YCF62* (*synpcc7942_1138*) with a common phylogenetic distribution (cyanobacteria and chloroplast), that are more likely to be functionally connected in cyanobacteria.

These examples support the notion that the CLG is particularly informative in the context of genes that are specifically related with the photosynthetic lifestyle, such as genes found exclusively in cyanobacteria or in cyanobacteria and chloroplasts. To obtain quantitative evidence in this context, we took advantage of the greencut2 gene set [41]. In particular, we compared for both CLG and STRING the number of links of *S. elongatus* genes belonging (*GreenCut2*) and not (*Others*) to this set of genes from the cyanobacteria-chloroplast lineage. Remarkably, the numbers of links per node for the dCLG but not for STRING is significantly greater in *GreenCut2* than in *Others* (Fig 8), in line with the anticipated potential of the dCLG in the context of genes characteristic of organisms performing oxygenic photosynthesis.

A case study: the PipX network.

PipX (DUF3539) is a unique cyanobacterial regulator identified by us in a yeast-two hybrid screen using the nitrogen signaling protein PII as bait (PII-interacting protein X) [42]. It provides a challenging example, since it is a protein involved in signal transduction (a functional category underrepresented in the dCLG, Fig 6a) that interacts physically with at least three other regulatory proteins [43] and the corresponding genes are not linked in cyanobacterial genomes. PipX forms complexes with PII and with the global transcriptional regulator NtcA according to the levels of 2-oxogularate and the ATP/ADP ratio [44-47]. PII-PipX complexes interact in yeast two-hybrid assays with the transcriptional regulator PlmA [8]. Although there is no evidence of protein binding, functional interactions were demostrated between PipX and PipY, a pyridoxal-phosphate binding protein involved in vitamin B6 homeostasis [48, 49].

Genetic and structural analyses suggest that PipX may perform additional roles [44-46, 50-52]. In this context, the N-terminal domain of PipX, whether in complex with either NtcA or PII [47] shows the same fold [51], a TLD/KOW motif found in the C-terminal domain of the NusG family which is involved in interaction with the ribosome [51, 53]. Furthermore, the C-terminal domain of PipX contains an R-rich basic patch that provides interaction determinants in non-canonical RNA binding proteins [54]. Our working hypothesis is that PipX is also involved in yet unknown regulatory processes in cyanobacteria and that some of the nodes from the dCLG network may provide clues.

PipX is part of a relatively robust (ALSxLPN 2.43) 6-node network (Fig 7e) for which Gecko3 makes no predictions, while STRING and theSeed produced coincident links for 3 out of the 12 dCLG connections. One of these coincident connections is PipX-PipY, anticipated on the bases of text-mining and neighbourhood ([55], STRING and theSeed).On the other hand, STRING retrieves the extensively characterized interactions of PII and NtcA with PipX, a case of functionally important interactions between signaling proteins for which the CLG is not informative.

As the dCLG appears to be particularly informative in the context of genes that are found exclusively in cyanobacteria or in cyanobacteria and chloroplasts, we are presently studying the connections between PipX with *engA* and *synpcc7942_2341* products, as a proof of concept. EngA is an eubacterial GTPase with a unique structure in which two G domains are tandemly repeated [56] that plays essential roles in ribosome biogenesis in both bacteria and chloroplasts [57-59]. It has been found membrane associated in *E. coli* [60] and in *Arabidopsis thaliana* thylakoids, where it has been connected with the photosystem II repair cycle [61]. Chloropast proteins (EngA1) are the closest homologs of Cyanobacterial EngA [62]. The *synpcc7942_2341* product shares homology with the transmembrane component (T unit) of ECF (Energy Coupling Factor) systems and, based on this homology, Pfam wrongly assigns *synpcc7942_2341* a CbiQ motif. Recently designated as cyano_T genes [63], *synpcc7942_2341* orthologs are not associated with recognizable *ecfA* genes in the genomes, as its happens with canonical ECF transporter genes and, importantly, they form together with their closely related plant homologs a separated branch from the *ecfT* family [63]. Recently obtained evidence supporting regulatory connections between PipX, cyano_T and EngA in *S. elongatus* confirms the potential of CLG networks to provide new hypothesis and functional predictions in cyanobacteria. Additional work is now required to elucidate the molecular details involved.

MATERIALS AND METHODS

Data acquisition and ortholog search

Cyanobacterial genome sequences and their Coding DNA Sequence (CDS) files at assembly level of "complete" or "chromosomal" were downloaded from the RefSeq repository [64] (S1 Table). CDS sequences, in nucleotides, were translated to protein using the Bacterial, Archaeal and Plant Plastid Code (translation table 11). CDS were labelled with their ordinal position in their corresponding chromosome. Ortholog retrieval was carried out using *S. elongatus* protein sequences as individual queries in BLASTp searches against the set of translated CDS of each genome. To ensure a minimum degree of specificity in the ortholog retrieval, two criteria were required: a minimum alignment length of 50% of the query protein and a minimum identity of 25% [8]. When two or more *S. elongatus* queries retrieved the same protein sequence, only the ones with the highest score were considered. For each ortholog of *S. elongatus* CDS, its ordinal position in the corresponding chromosome was retrieved and subsequently used for distance and co-occurrence calculations.

Pairwise distance and co-occurrence calculation

For each pair of CDS in each genome, the minimum distance between them was calculated as the difference between the retrieved ordinal position values in the ortholog search independently of physical distance and gene orientation. Minimum distance between two adjacent genes was always 1, even considering overlapping genes. As expected, the maximum possible number of distance values was 124, the numbers of genomes considered (only obtained for gene pairs showing full conservation). The following formula, taking into account chromosome circularity, was used:

$$\text{Distance} = nCDS + vm - vM$$

Where *nCDS* is the number of CDS in the chromosome and *vm* and *vM* are the minimum and maximum of the two position values, respectively. Additionally, for each pair of *S. elongatus* CDS the number of genomes where at least one and both of the CDS were present was retrieved.

Linkage Score and Phylogenetic correction

For each pair of CDS an initial Linkage Score, referred here as Raw Linkage Score was calculated using the number of genomes where the distance between genes was lower than 4. This relatively small window of 3 genes upstream and 3 genes downstream allows us to keep track of synteny relationships between pairs of genes even in the presence of small insertions or inversions. A correction ratio of co-occurrence / occurrence was also introduced in the following formula, to consider different degrees of conservation between pair of genes:

$$RawLinkageScore = N_o \frac{N_{co}}{N_o}$$

(1)

Where *Ns* is the number of organisms with distance between the given pair of CDS below the threshold (4), *Nco* is the number of organisms where both CDS are present, *No* is the number of organisms where at least one of the two CDS is present.

Average Nucleotide Identity (ANI) was calculated for each pair of organisms using pyANI. ANI values were scaled following the formula:

$$ANIscaled = \frac{ANI_{i,j} - ANI_{\min}}{ANI_{\max} - ANI_{\min}}$$

For each link Raw Linkage Score was corrected using the following formula, which reduces the score proportionally to the similarity of the subset of genes where the genes are neighbours compared to the similarity of the subset where the genes (at least one of them) are present:

$$LinkageScore = RawLinkageScore \frac{1 - ANI_S}{1 - ANI_O}$$

(2)

Where *ANIs* is the mean value of the scaled ANI between each pair of organisms where the distance is below the threshold and *ANIo* is the mean value of the scaled ANI between each pair of organisms containing both CDS.

Nodes and Networks descriptors

Nodes

Gene names, essentiality and descriptions were obtained from [1]. Cluster of orthologs (COG) classification was downloaded from cyanobase [65]. Operon IDs were obtained from DOOR2 database [16]. In each case, gene conservation was calculated as the percentage of genomes containing an ortholog of the given gene. To exclude highly similar genomes, this conservation was calculated in a subset of 113 cyanobacterial genomes where all ANI values between them were lower than 99.9%.

Networks

ALS was calculated using the following formula:

$$ALS = \frac{\sum_{i=0}^{N-1} \sum_{j=i+1}^{N} score_{i,j}}{L}$$

(3)

Where “N” is the number of nodes in the network, “L” is the number of links above the threshold and “score$_{i,j}$“ is the LS of the link connecting nodes “i” and “j”. When “score$_{i,j}$” is below the threshold is considered 0.

LPN was calculated with the following formula:

$$LPN = \frac{\sum_{i=0}^{N} nLinks_i}{N} \tag{4}$$

Where “N” is the number of nodes in the network and “nLinks$_i$” is the number of links above the threshold for the node “i”.

Number of **genomic neighborhoods** was calculated by assigning genes to groups where the minimum distance of each gene with at least another gene of the group was 3 or less. The resulting number of groups was considered the number of genomic neighborhoods.

Link expected distribution

For the calculus of the expected distribution of 0, 1 or 2 links, the proportion of persistent, essential, polycistronic and known function genes of the dCLG was obtained. The expected numbers of links were calculated as follows:

$$Exp0 = L * p^2$$
$$Exp1 = L * 2 * P * p$$
$$Exp2 = L * P^2$$

Where “Exp0”, “Exp1” and “Exp2” are, respectively, the expected number of links of type 0, 1 and 2 for a given characteristic (persistence, essentiality, polycistronic and known function). “p” is the proportion of genes which do not fit the characteristic and “P” is the proportion of genes that do. In the case of essentiality, “p” corresponds to non-essential and “P” to essential genes, respectively. L is the total number of links in the dCLG (1036).

CLG web interface

The CLG web application produces networks searching recursively for genes connected with links over the given threshold directly or indirectly to the query. To generate a visual display consistent with the connectivity and linkage of the nodes of the network we used the cytoscape web graph theory (network) library for visualization and analysis [66], applying the Euler layout to the generated network. Future updating of the data will include the incorporation of additional cyanobacterial genomes and the modification of the reference gene set to include model organisms.

DUF searches in STRING, theSeed and Gecko3

S. elongatus DUF genes (S6 Table) were used as query protein in STRING searches with disabled co-occurrence score and a confidence threshold of 0.700. For comparison with dCLG networks only nodes connected directly to the DUF query were considered. In the case of theSeed, dCLG predictions for DUF genes were compared against functionally coupled proteins using the *S. elongatus* DUF gene as query. Functionally coupled proteins were extracted from the feature evidence page for the DUF in *S. elongatus* genome (code 1140.7). For Gecko3, all genes inside a gene cluster were assumed to be interconnected.

For all methods compared, coincidence was considered complete when all dCLG links were found, partial when not all but at least one link was found, and no coincidence when no common links where found.

DATA AVAILABILITY

The datasets generated during and/or analysed during the current study are available from the corresponding author on request.

CONTRIBUTIONS

J.I.L. designed and performed the gene linkage analysis, analysed data and wrote the manuscript; A.L. analysed data; A.C designed the analyses, analysed data and wrote the manuscript.

COMPETING INTEREST

The author(s) declare no competing interests.

ACKNOWLEDGMENTS

We thank F. Rodriguez-Mateos and R. Dixon for constructive discussion. The authors thank the Spanish Government (MINECO) for grants BFU2015-66360-P and PEJ-2014-P-00298, the Conselleria d'Educació, Investigació, Cultura i Esport from the Generalitat Valenciana for grant PROMETEO/2017/129, and the University of Alicante for PhD fellowship (FPUUA59) to J.I.L.

REFERENCES

1. Rubin, B. E., Wetmore, K. M., Price, M. N., Diamond, S., Shultzaberger, R. K., Lowe, L. C., Curtin, G., Arkin, A. P., Deutschbauer, A. & Golden, S. S. (2015) The essential gene set of a photosynthetic organism, *Proc Natl Acad Sci U S A.* **112**, E6634-43.

2. Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J. & Mering, C. V. (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets, *Nucleic acids research.* **47**, D607-D613.

3. Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., de Crecy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E. D., Gerdes, S., Glass, E. M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A. C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G. D., Rodionov, D. A., Ruckert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O. & Vonstein, V. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes, *Nucleic acids research.* **33**, 5691-702.

4. Beck, C., Knoop, H. & Steuer, R. (2018) Modules of co-occurrence in the cyanobacterial pan-genome reveal functional associations between groups of ortholog genes, *PLoS genetics.* **14**, e1007239.

5. Tiruveedula, G. S. S. & Wangikar, P. P. (2017) Gene essentiality, conservation index and co-evolution of genes in cyanobacteria, *PloS one.* **12**, e0178565.

6. Kreula, S. M., Kaewphan, S., Ginter, F. & Jones, P. R. (2018) Finding novel relationships with integrated gene-gene association network analysis of Synechocystis sp. PCC 6803 using species-independent text-mining, *PeerJ.* **6**, e4806.

7. Winter, S., Jahn, K., Wehner, S., Kuchenbecker, L., Marz, M., Stoye, J. & Bocker, S. (2016) Finding approximate gene clusters with Gecko 3, *Nucleic acids research.* **44**, 9600-9610.

8. Labella, J. I., Obrebska, A., Espinosa, J., Salinas, P., Forcada-Nadal, A., Tremino, L., Rubio, V. & Contreras, A. (2016) Expanding the Cyanobacterial Nitrogen Regulatory Network: The GntR-Like Regulator PlmA Interacts with the PII-PipX Complex, *Front Microbiol.* **7**, 1677.

9. Richter, M. & Rossello-Mora, R. (2009) Shifting the genomic gold standard for the prokaryotic species definition, *Proceedings of the National Academy of Sciences of the United States of America.* **106**, 19126-31.

10. Dilucca, M., Cimini, G. & Giansanti, A. (2018) Essentiality, conservation, evolutionary pressure and codon bias in bacterial genomes, *Gene.* **663**, 178-188.

11. Fang, G., Rocha, E. P. & Danchin, A. (2008) Persistence drives gene clustering in bacterial genomes, *BMC genomics.* **9**, 4.

12. Delaye, L., Gonzalez-Domenech, C. M., Garcillan-Barcia, M. P., 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.

13. Simm, S., Keller, M., Selymesi, M. & Schleiff, E. (2015) The composition of the global and feature specific cyanobacterial core-genomes, *Front Microbiol.* **6**, 219.

14. Mulkidjanian, A. Y., Koonin, E. V., Makarova, K. S., Mekhedov, S. L., Sorokin, A., Wolf, Y. I., Dufresne, A., Partensky, F., Burd, H., Kaznadzey, D., Haselkorn, R. & Galperin, M. Y. (2006) The cyanobacterial genome core and the origin of photosynthesis, *Proc Natl Acad Sci U S A.* **103**, 13126-31.

15. Shi, T. & Falkowski, P. G. (2008) Genome evolution in cyanobacteria: the stable core and the variable shell, *Proceedings of the National Academy of Sciences of the United States of America.* **105**, 2510-5.

16. Mao, X., Ma, Q., Zhou, C., Chen, X., Zhang, H., Yang, J., Mao, F., Lai, W. & Xu, Y. (2014) DOOR 2.0: presenting operons and their functions through dynamic and integrated views, *Nucleic acids research.* **42**, D654-9.

17. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database, *Nucleic acids research.* **43**, D261-9.

18. Saschenbrecker, S., Bracher, A., Rao, K. V., Rao, B. V., Hartl, F. U. & Hayer-Hartl, M. (2007) Structure and function of RbcX, an assembly chaperone for hexadecameric Rubisco, *Cell.* **129**, 1189-200.

19. Huang, F., Vasieva, O., Sun, Y., Faulkner, M., Dykes, G. F., Zhao, Z. & Liu, L. N. (2019) Roles of RbcX in Carboxysome Biosynthesis in the Cyanobacterium Synechococcus elongatus PCC7942, *Plant physiology.* **179**, 184-194.

20. Shibata, M., Ohkawa, H., Kaneko, T., Fukuzawa, H., Tabata, S., Kaplan, A. & Ogawa, T. (2001) Distinct constitutive and low-CO2-induced CO2 uptake systems in cyanobacteria: genes involved and their phylogenetic relationship with homologous genes in other organisms, *Proceedings of the National Academy of Sciences of the United States of America.* **98**, 11789-94.

21. Ogawa, T. (1991) A gene homologous to the subunit-2 gene of NADH dehydrogenase is essential to inorganic carbon transport of Synechocystis PCC6803, *Proceedings of the National Academy of Sciences of the United States of America.* **88**, 4275-9.

22. Marco, E., Ohad, N., Schwarz, R., Lieman-Hurwitz, J., Gabay, C. & Kaplan, A. (1993) High CO2 concentration alleviates the block in photosynthetic electron transport in an ndhB-inactivated mutant of Synechococcus sp. PCC 7942, *Plant physiology.* **101**, 1047-53.

23. Cai, F., Sutter, M., Cameron, J. C., Stanley, D. N., Kinney, J. N. & Kerfeld, C. A. (2013) The structure of CcmP, a tandem bacterial microcompartment domain protein from the beta-carboxysome, forms a subcompartment within a microcompartment, *The Journal of biological chemistry.* **288**, 16055-63.

24. Bratlie, M. S., Johansen, J. & Drablos, F. (2010) Relationship between operon preference and functional properties of persistent genes in bacterial genomes, *BMC genomics.* **11**, 71.

25. Glover, N. M., Daron, J., Pingault, L., Vandepoele, K., Paux, E., Feuillet, C. & Choulet, F. (2015) Small-scale gene duplications played a major role in the recent evolution of wheat chromosome 3B, *Genome biology.* **16**, 188.

26. Ashby, M. K. & Mullineaux, C. W. (1999) Cyanobacterial ycf27 gene products regulate energy transfer from phycobilisomes to photosystems I and II, *FEMS microbiology letters.* **181**, 253-60.

27. Forchhammer, K. & Schwarz, R. (2019) Nitrogen chlorosis in unicellular cyanobacteria - a developmental program for surviving nitrogen deprivation, *Environ Microbiol.* **21**, 1173-1184.

28. Lopez-Redondo, M. L., Moronta, F., Salinas, P., Espinosa, J., Cantos, R., Dixon, R., Marina, A. & Contreras, A. (2010) Environmental control of phosphorylation pathways in a branched two-component system, *Mol Microbiol.* **78**, 475-89.

29. Moronta-Barrios, F., Espinosa, J. & Contreras, A. (2013) Negative control of cell size in the cyanobacterium *Synechococcus elongatus* PCC 7942 by the essential response regulator RpaB, *FEBS Lett.* **587**, 504-9.

30. Espinosa, J., Boyd, J. S., Cantos, R., Salinas, P., Golden, S. S. & Contreras, A. (2015) Cross-talk and regulatory interactions between the essential response regulator RpaB and cyanobacterial circadian clock output, *Proc Natl Acad Sci U S A.* **112**, 2198-203.

31. Riediger, M., Hihara, Y. & Hess, W. R. (2018) From cyanobacteria and algae to land plants: The RpaB/Ycf27 regulatory network in transition, *Perspectives in Phycology.* **5**, 13-25.

32. Piechura, J. R., Amarnath, K. & O'Shea, E. K. (2017) Natural changes in light interact with circadian regulation at promoters to control gene expression in cyanobacteria, *eLife.* **6**.

33. Goodacre, N. F., Gerloff, D. L. & Uetz, P. (2013) Protein domains of unknown function are essential in bacteria, *mBio.* **5**, e00744-13.

34. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E. & Finn, R. D. (2019) The Pfam protein families database in 2019, *Nucleic acids research.* **47**, D427-D432.

35. Nord, S., Bylund, G. O., Lovgren, J. M. & Wikstrom, P. M. (2009) The RimP protein is important for maturation of the 30S ribosomal subunit, *Journal of molecular biology.* **386**, 742-53.

36. Ogura, M. & Kanesaki, Y. (2018) Newly Identified Nucleoid-Associated-Like Protein YlxR Regulates Metabolic Gene Expression in Bacillus subtilis, *mSphere.* **3**.

37. Molik, S., Karnauchov, I., Weidlich, C., Herrmann, R. G. & Klosgen, R. B. (2001) The Rieske Fe/S protein of the cytochrome b6/f complex in chloroplasts: missing link in the evolution of protein transport pathways in chloroplasts?, *The Journal of biological chemistry.* **276**, 42761-6.

38. Aldridge, C., Spence, E., Kirkilionis, M. A., Frigerio, L. & Robinson, C. (2008) Tat-dependent targeting of Rieske iron-sulphur proteins to both the plasma and thylakoid membranes in the cyanobacterium Synechocystis PCC6803, *Molecular microbiology.* **70**, 140-50.

39. Frain, K. M., Gangl, D., Jones, A., Zedler, J. A. & Robinson, C. (2016) Protein translocation and thylakoid biogenesis in cyanobacteria, *Biochimica et biophysica acta.* **1857**, 266-73.

40. Aramini, J. M., Petrey, D., Lee, D. Y., Janjua, H., Xiao, R., Acton, T. B., Everett, J. K. & Montelione, G. T. (2012) Solution NMR structure of Alr2454 from *Nostoc* sp. PCC 7120, the first structural representative of Pfam domain family PF11267, *Journal of structural and functional genomics.* **13**, 171-6.

41. Karpowicz, S. J., Prochnik, S. E., Grossman, A. R. & Merchant, S. S. (2011) The GreenCut2 resource, a phylogenomically derived inventory of proteins specific to the plant lineage, *The Journal of biological chemistry.* **286**, 21427-39.

42. Burillo, S., Luque, I., Fuentes, I. & Contreras, A. (2004) Interactions between the nitrogen signal transduction protein PII and N-acetyl glutamate kinase in organisms that perform oxygenic photosynthesis, *J Bacteriol.* **186**, 3346-54.

43. Forcada-Nadal, A., Llacer, J. L., Contreras, A., Marco-Marin, C. & Rubio, V. (2018) The PII-NAGK-PipX-NtcA Regulatory Axis of Cyanobacteria: A Tale of Changing Partners, Allosteric Effectors and Non-covalent Interactions, *Frontiers in molecular biosciences.* **5**, 91.

44. Espinosa, J., Forchhammer, K., Burillo, S. & Contreras, A. (2006) Interaction network in cyanobacterial nitrogen regulation: PipX, a protein that interacts in a 2-oxoglutarate dependent manner with PII and NtcA, *Mol Microbiol.* **61**, 457-69.

45. Espinosa, J., Forchhammer, K. & Contreras, A. (2007) Role of the *Synechococcus* PCC 7942 nitrogen regulator protein PipX in NtcA-controlled processes, *Microbiology.* **153**, 711-8.

46. Espinosa, J., Castells, M. A., Laichoubi, K. B. & Contreras, A. (2009) Mutations at *pipX* suppress lethality of PII-deficient mutants of *Synechococcus elongatus* PCC 7942, *J Bacteriol.* **191**, 4863-9.

47. Llacer, J. L., Espinosa, J., Castells, M. A., Contreras, A., Forchhammer, K. & Rubio, V. (2010) Structural basis for the regulation of NtcA-dependent transcription by proteins PipX and PII, *Proc Natl Acad Sci U S A.* **107**, 15397-15402.

48. Labella, J. I., Cantos, R., Espinosa, J., Forcada-Nadal, A., Rubio, V. & Contreras, A. (2017) PipY, a Member of the Conserved COG0325 Family of PLP-Binding Proteins, Expands the Cyanobacterial Nitrogen Regulatory Network, *Frontiers in microbiology.* **8**, 1244.

49. Tremiño, L., Forcada-Nadal, A., Contreras, A. & Rubio, V. (2017) Studies on cyanobacterial protein PipY shed light on structure, potential functions, and vitamin B6 -dependent epilepsy, *FEBS Lett.* **591**, 3431-3442.

50. Espinosa, J., Rodriguez-Mateos, F., Salinas, P., Lanza, V. F., Dixon, R., de la Cruz, F. & Contreras, A. (2014) PipX, the coactivator of NtcA, is a global regulator in cyanobacteria, *Proceedings of the National Academy of Sciences of the United States of America.* **111**, E2423-30.

51. Cantos, R., Labella, J. I., Espinosa, J. & Contreras, A. (2018) The nitrogen regulator PipX acts in *cis* to prevent operon polarity, *Environmental microbiology reports*.

52. Laichoubi, K. B., Espinosa, J., Castells, M. A. & Contreras, A. (2012) Mutational analysis of the cyanobacterial nitrogen regulator PipX, *PLoS One.* **7**, e35845.

53. Burmann, B. M., Schweimer, K., Luo, X., Wahl, M. C., Stitt, B. L., Gottesman, M. E. & Rosch, P. (2010) A NusE:NusG complex links transcription and translation, *Science.* **328**, 501-4.

54. Jarvelin, A. I., Noerenberg, M., Davis, I. & Castello, A. (2016) The new (dis)order in RNA regulation, *Cell communication and signaling : CCS.* **14**, 9.

55. Labella, J. I., Cantos, R., Espinosa, J., Forcada-Nadal, A., Rubio, V. & Contreras, A. (2017) PipY, a member of the conserved COG0325 family of PLP-binding proteins, expands the cyanobacterial nitrogen regulatory network, *Front Microbiol.* **8**, 1244.

56. Verstraeten, N., Fauvart, M., Versees, W. & Michiels, J. (2011) The universally conserved prokaryotic GTPases, *Microbiol Mol Biol Rev.* **75**, 507-42, second and third pages of table of contents.

57. Hwang, J. & Inouye, M. (2006) The tandem GTPase, Der, is essential for the biogenesis of 50S ribosomal subunits in *Escherichia coli*, *Mol Microbiol.* **61**, 1660-72.

58. Bharat, A. & Brown, E. D. (2014) Phenotypic investigations of the depletion of EngA in *Escherichia coli* are consistent with a role in ribosome biogenesis, *FEMS microbiology letters.* **353**, 26-32.

59. Jeon, Y., Ahn, C. S., Jung, H. J., Kang, H., Park, G. T., Choi, Y., Hwang, J. & Pai, H. S. (2014) DER containing two consecutive GTP-binding domains plays an essential role in chloroplast ribosomal RNA processing and ribosome biogenesis in higher plants, *Journal of experimental botany.* **65**, 117-30.

60. Lee, R., Aung-Htut, M. T., Kwik, C. & March, P. E. (2011) Expression phenotypes suggest that Der participates in a specific, high affinity interaction with membranes, *Protein expression and purification.* **78**, 102-12.

61. Kato, Y., Hyodo, K. & Sakamoto, W. (2018) The Photosystem II Repair Cycle Requires FtsH Turnover through the EngA GTPase, *Plant physiology.* **178**, 596-611.

62. Suwastika, I. N., Denawa, M., Yomogihara, S., Im, C. H., Bang, W. Y., Ohniwa, R. L., Bahk, J. D., Takeyasu, K. & Shiina, T. (2014) Evidence for lateral gene transfer (LGT) in the evolution of eubacteria-derived small GTPases in plant organelles, *Frontiers in plant science.* **5**, 678.

63. Rempel, S., Stanek, W. K. & Slotboom, D. J. (2018) Energy-Coupling Factor-Type ATP-Binding Cassette Transporters, *Annual review of biochemistry*.

64. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A.,

Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D. & Pruitt, K. D. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation, *Nucleic acids research.* **44**, D733-45.

65. Nakao, M., Okamoto, S., Kohara, M., Fujishiro, T., Fujisawa, T., Sato, S., Tabata, S., Kaneko, T. & Nakamura, Y. (2010) CyanoBase: the cyanobacteria genome database update 2010, *Nucleic acids research.* **38**, D379-81.

66. Franz, M., Lopes, C. T., Huck, G., Dong, Y., Sumer, O. & Bader, G. D. (2016) Cytoscape.js: a graph theory library for visualisation and analysis, *Bioinformatics.* **32**, 309-11.

FIGURE LEGENDS

Figure 1. Workflow for generation of the Linkage Score. Cyanobacterial Coding Sequences were downloaded from refseq and translated to protein. *S. elongatus* sequences were used as query in a BLASTp search to determine the ordinal position of each *S. elongatus* ortholog gene in each genome (represented as a sequence of row colored squares in the “Ordinal Position Information” section). From ordinal positions the distance between each pair of genes in each genome was calculated. Occurrence information, also retrieved from the BLASTp search, was used together with the distance information to calculate a provisional score that was subsequently corrected using similarity between cyanobacterial genomes (ANI values obtained with pyANI) to generate the final linkage score.

Figure 2. The default Cyanobacterial Linked Genome (dCLG). a) Distribution of the number of networks (blue, left y-axis) and genes (red, right y-axis) resulting from different LS thresholds (x-axis). The selected threshold is indicated with an arrow pointing to a dashed line. b) Cloud representation of networks with LS threshold 0.3464 (dCLG). The intensity of the color of the nodes reflects the degree of conservation of the corresponding gene in cyanobacteria. Link width correlates with the corresponding LS value. c) Coincidences between network 1 and Gecko3 clusters 83, 146 and 282 are illustrated in pink, green and yellow, respectively.

Figure 3. Composition and connectivity of dCLG genes. a) Percentage of persistent, essential (according to [1]), polycistronic (according to [16]) and known function (according to COG categories) genes in the indicated gene sets. The percentage of beneficial genes are indicated in stacked transparent boxes on top of the bars for essential genes. b) Log_2 of the ratio of observed/expected links according to the kind of genes connected. For the given property “2”, “1” and “0” indicate links connecting 2, 1 or 0 genes matching the corresponding property.

Figure 4. The dCLG overlaps with core genomes. Venn diagrams representing the coincidence between a) the dCLG gene set or b) persistent genes (98% conservation in our dataset) and four available cyanobacterial core genomes (core 1, [12]; core 2, [13]; core 3, [14]; and core 4, [15]).

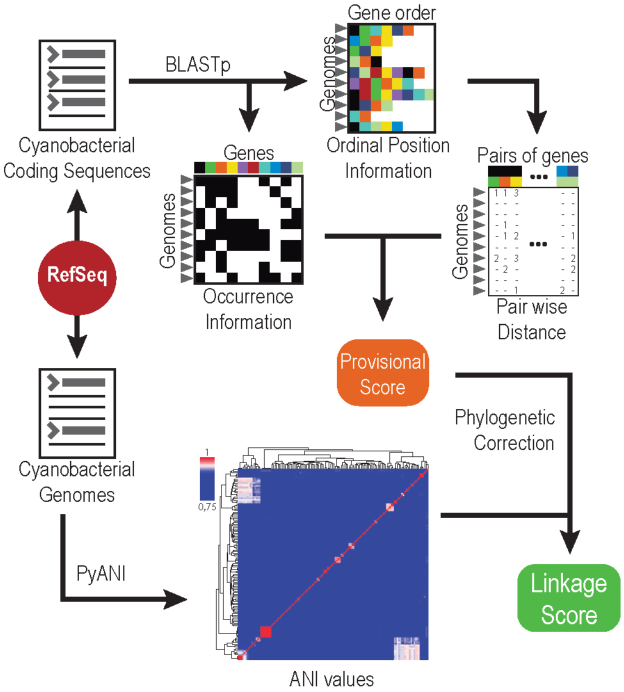
Figure 5. Network context of the *rbcX* gene. Schematic representation of the *rbcX* network at threshold value 0.29, including dCLG networks 129 (composed of genes: *rbcL*, *rbcS, rbcX*) and 22 (*ccmN*, *ccmM*, *ccmL*, *ccmK*, *ndhF4*, *ndhD4* and *chpX*) as well as genes *ccmP and ccmO* (non-connected at the threshold LS default). Black and grey links indicate scores of at least 0.3464 (default threshold) and 0.29, respectively. Link width is proportional to the LS. Rubisco, carboxisome and Ndh-1 genes are coloured in green, pink, and yellow, respectively.

Figure 6. Distribution of functions and dCLG genes. a) Log_2 of the ratio between the percentage of genes in the indicated functional categories for the indicated sets of *S. elongatus*

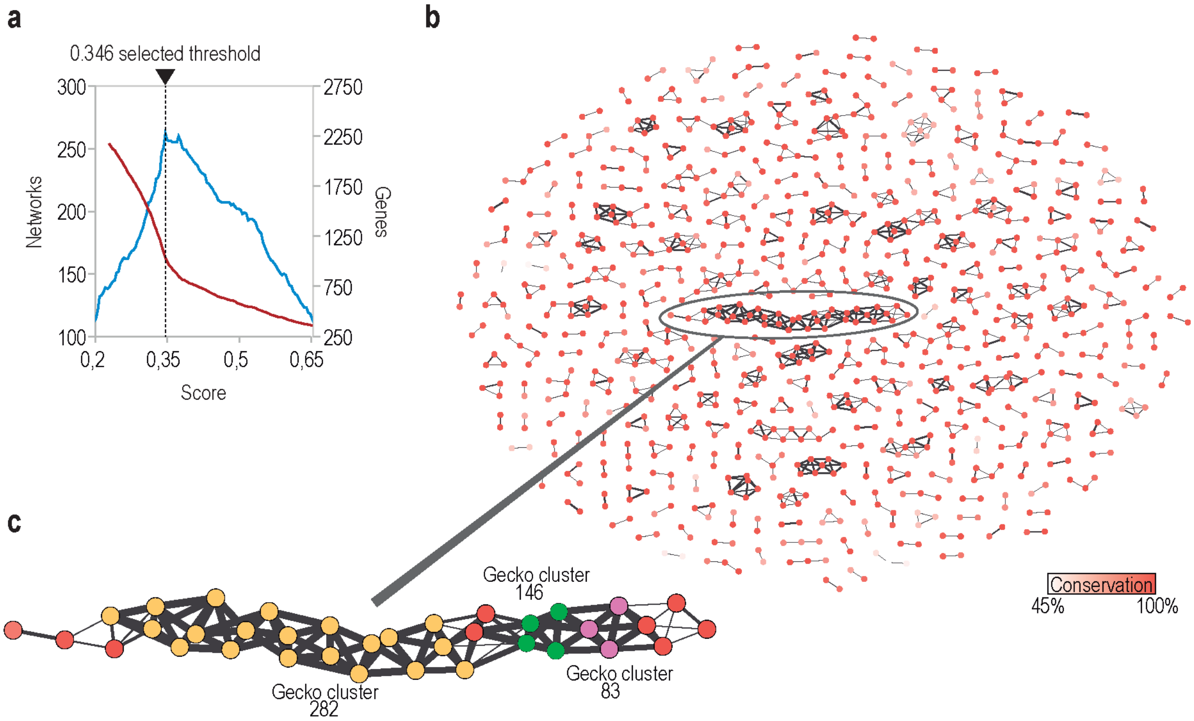
genes and the whole gene set. Only categories representing more than 2% of the *S. elongatus* gene set are plotted. b) Frequency of links connecting two (intra-category) or just one node from the same category (inter-category) normalized by the number of dCLG nodes on that category.

Figure 7. Functional predictions for DUF motifs. a) Venn diagram showing the number of dCLG DUF nodes with connections in Gecko3, theSeed and STRING (confidence level 0.7, co-occurrence disabled). b-e) Representative networks (threshold value 0.3464, except for DUF3067 network, for which 0.43 was used) containing a DUF protein (in bold). Link width is proportional to LS and link color within the networks refers to the particular method(s) where the links are found, for which the same color code that is used in a) to denote the predicting methods is used. Gene name or the shortened ID is indicated on each node. Node color indicates its phylogenetic distribution (white: bacteria, archaea and eukaryotes; orange: bacteria; purple: bacteria and plants; yellow: gram-positive bacteria and cyanobacteria; green: cyanobacteria and plants; blue: cyanobacteria). Other proteins discussed in the text are shown alongside the corresponding dCLG networks.

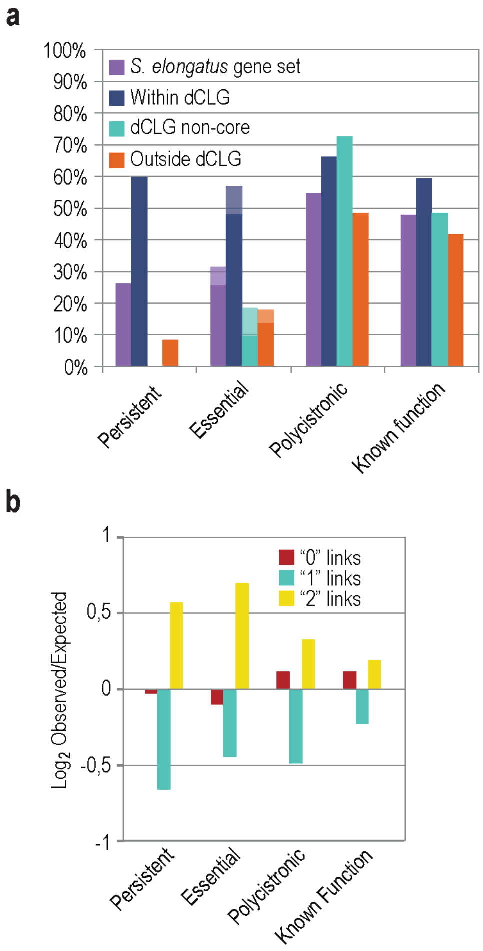
Figure 8. STRING and CLG performance on *S. elongatus GreenCut2* genes. Boxplot of the number of links per node normalized by the average number of links per node in each tool. Genes are separated in *GreenCut2* and non-*GreenCut2* (Others) according to [41]. Links are retrieved using the same criteria as in the main text.



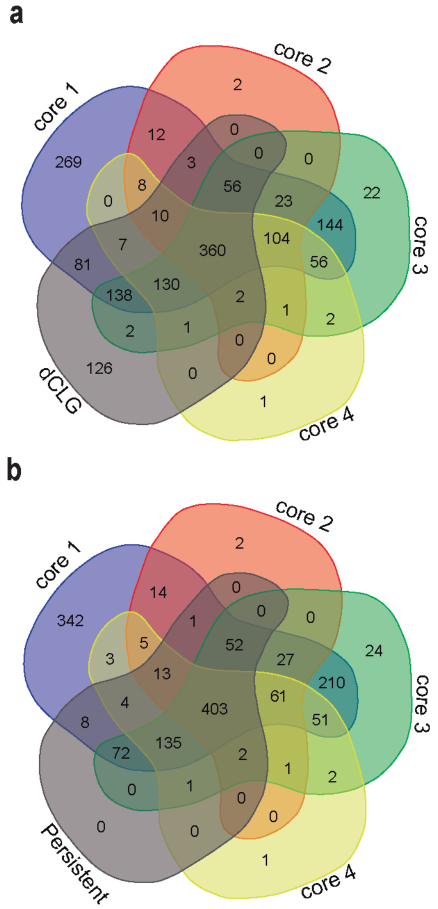
feb2_13775_f1.tif



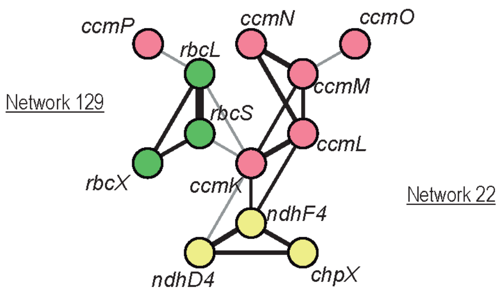
feb2_13775_f2.tif



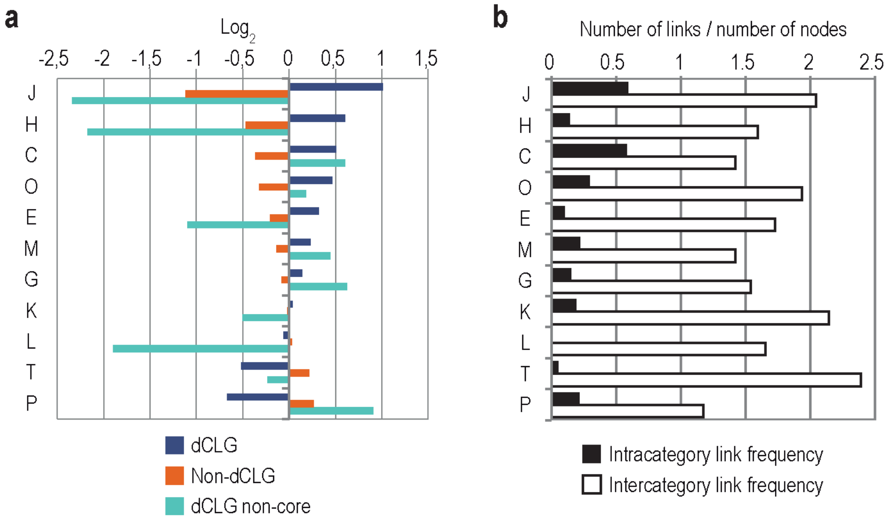
feb2_13775_f3.tif



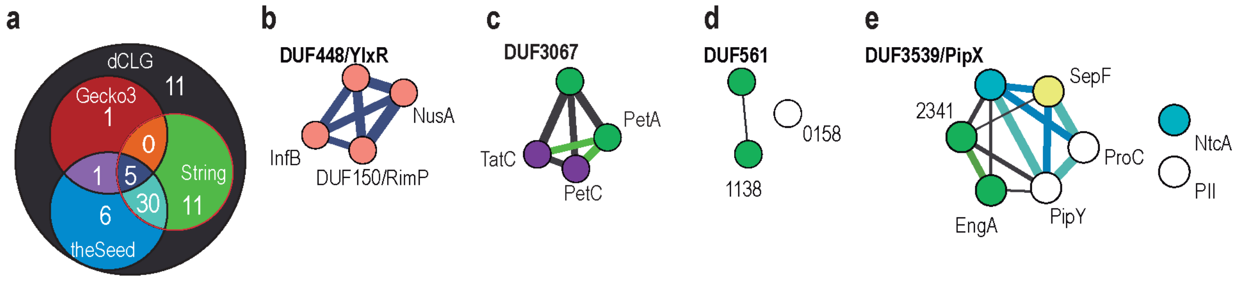
feb2_13775_f4.tif



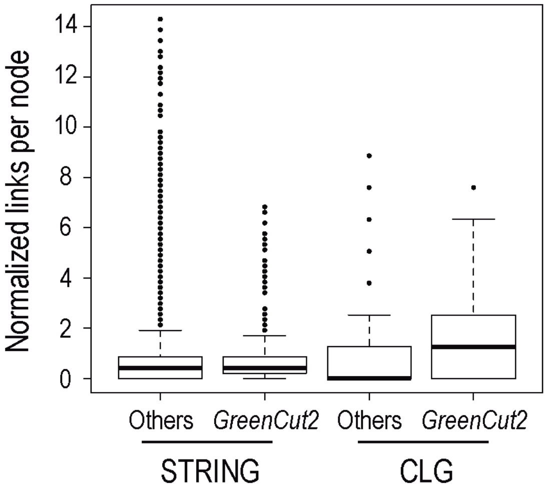
feb2_13775_f5.tif



feb2_13775_f6.tif



feb2_13775_f7.tif



feb2_13775_f8.tif