A peer-reviewed version of this preprint was published in PeerJ on 23 May 2018.

View the peer-reviewed version (peerj.com/articles/4806), which is the preferred citable publication unless you specifically need to cite this preprint.

Kreula SM, Kaewphan S, Ginter F, Jones PR. 2018. Finding novel relationships with integrated gene-gene association network analysis of *Synechocystis sp.* PCC 6803 using species-independent text-mining. PeerJ 6:e4806 https://doi.org/10.7717/peerj.4806

Finding novel relationships with integrated gene-gene association network analysis of *Synechocystis sp.* PCC 6803 using species-independent text-mining

Sanna M Kreula [1, 2], **Suwisa Kaewphan** [2, 3], **Filip Ginter** [3], **Patrik R Jones** Corresp. [4]

[1] Department of Biochemistry, University of Turku, Turku, Finland
[2] University of Turku Graduate School, University of Turku, Turku, Finland
[3] Department of Future Technologies, University of Turku, Turku, Finland
[4] Department of Life Sciences, Imperial College London, London, United Kingdom

Corresponding Author: Patrik R Jones
Email address: p.jones@imperial.ac.uk

The increasing move towards open access full-text scientific literature enhances our ability to utilize advanced text-mining methods to construct information-rich networks that no human will be able to grasp simply from 'reading the literature'. The utility of text-mining for well-studied species is obvious though the utility for less studied species, or those with no prior track-record at all, is not clear. Here we present a concept for how advanced text-mining can be used to create information-rich networks even for less well studied species and apply it to generate an open-access gene-gene association network resource for *Synechocystis sp.* PCC 6803, a representative model organism for cyanobacteria and first case-study for the methodology. By merging the text-mining network with networks generated from species-specific experimental data, network integration was used to enhance the accuracy of predicting novel interactions that are biologically relevant. A rule-based algorithm was constructed in order to automate the search for novel candidate genes with a high degree of likely association to known target genes by (1) ignoring established relationships from the existing literature, as they are already 'known', and (2) demanding multiple independent evidences for every novel and potentially relevant relationship. Using selected case studies, we demonstrate the utility of the network resource and rule-based algorithm to (*i*) discover novel candidate associations between different genes or proteins in the network, and (*ii*) rapidly evaluate the potential role of any one particular gene or protein. The full network is provided as an open source resource.

Finding novel relationships with integrated gene-gene association network analysis of *Synechocystis sp.* PCC 6803 using species-independent text-mining

Sanna M. Kreula, Suwisa Kaewphan, Filip Ginter, and Patrik R. Jones*

Molecular Plant Biology, Department of Biochemistry, University of Turku, FI-20014 Turku, Finland (S. Kreula); University of Turku Graduate School (UTUGS), University of Turku, FI-20014 Turku, Finland (S. Kreula, S. Kaewphan);Turku Center for Computer Science (TUCS), Joukahaisenkatu 3-5, 20520 Turku, Finland (S. Kaewphan); Department of Future technologies, University of Turku, FI-20014 Turku, Finland (S. Kaewphan, F. Ginter); Department of Life Sciences, Imperial College London, London, SW7 2AZ, UK (P. R. Jones)

* To whom correspondence should be addressed. E-mail: p.jones@imperial.ac.uk

ABSTRACT

The increasing move towards open access full-text scientific literature enhances our ability to utilize advanced text-mining methods to construct information-rich networks that no human will be able to grasp simply from 'reading the literature'. The utility of text-mining for well-studied species is obvious though the utility for less studied species, or those with no prior track-record at all, is not clear. Here we present a concept for how advanced text-mining can be used to create information-rich networks even for less well studied species and apply it to generate an open-access gene-gene association network resource for *Synechocystis sp.* PCC 6803, a representative model organism for cyanobacteria and first case-study for the methodology. By merging the text-mining network with networks generated from species-specific experimental data, network integration was used to enhance the accuracy of predicting novel interactions that are biologically relevant. A rule-based algorithm was constructed in order to automate the search for novel candidate genes with a high degree of likely association to known target genes by (1) ignoring established relationships from the existing literature, as they are already 'known', and (2) demanding multiple independent evidences for every novel and potentially relevant relationship. Using selected case studies, we demonstrate the utility of the network resource and rule-based algorithm to (*i*) discover novel candidate associations between different genes or proteins in the network, and (*ii*) rapidly evaluate the potential role of any one particular gene or protein. The full network is provided as an open source resource.

INTRODUCTION

Synechocystis sp. PCC 6803 (hereafter *Synechocystis* 6803) was the first photobiological organism to be sequenced in 1996 (Kaneko et al. 1996). It is a unicellular prokaryote with a compact genome (~3.5 Mbp) that is capable of non-facilitated DNA-uptake and homologous recombination. It has been extensively studied as a model for photosynthesis and cyanobacteria in general (Ikeuchi and Tabata 2001), and more recently it has been considered also as a potential host for biotechnology in which solar energy is directly converted into chemical energy and feedstock (Rosgaard et al. 2012).

Compared to other photobiological model species, such as *Arabidopsis thaliana (De Bodt et al. 2012)*, there is still a relative lack of systems biology resources for *Synechocystis* 6803 and cyanobacteria in general. The online 'Cyanobase' portal has played an important role in providing information from genome sequencing data for the cyanobacteria community (Nakao et al. 2010). However, as far as we are aware, there is only one other online database for easy access of a collection of omics data sets (CyanoEXpress (Hernandez-Prieto and Futschik 2012), microarray data repository). Transcriptome data sets included in the CyanoEXpress repository have mainly been analyzed in respective original publications by differential or simple clustering analysis; Efforts to utilize cyanobacteria systems biology data sets for graph-based network analysis are otherwise rare (Bhadauriya et al. 2007). Similarly, there is only one online graph-based network analysis platform that includes cyanobacteria species (STRING (Franceschini et al. 2013)). The STRING network, however, lacks cyanobacteria-specific data apart from its genome sequence. To complicate matters further, the majority (55.1%) of genes in Cyanobase remain "unknown" (Fujisawa et al. 2017). In part this reflects the early date of the first sequencing and persistence of historical archives of annotations in some databases, however, it also reflects the fact that very few studies have been carried out with *Synechocystis* 6803 in comparison with other model species. For example, 350,630 articles including the term '*Escherichia coli*' were found in PubMed July 2017 whilst only 3,853 included the term '*Synechocystis*' (Fig. 1).

Figure 1. There are few publications for cyanobacteria in comparison to other model species such as *Escherichia coli*. The search terms 'Synechocystis' (representing *Synechocystis sp.* PCC 6803), 'Cyanobacteria', 'Arabidopsis" (representing the model plant *Arabidopsis thaliana*) and 'Escherichia coli' were entered into PubMed (http://www.ncbi.nlm.nih.gov/pubmed) July 2017. The numbers shown in the figure were obtained from this website by selecting "Results by year".

Text-mining is a developing technology with increasing potential for scientific utilization, especially given the recent trend towards open access in the scientific literature (Gonzalez et al. 2016, Van Landeghem et al. 2011). One opportunity with text-mining is to aggregate knowledge from the massive volume of available literature and generate detailed maps of knowledge that would be difficult to obtain otherwise. Naturally, the utility of such network-based aggregation depends on the quantity and quality of the source data (Fig. 1), as well as the method of extracting the information, aggregating it and visualizing it in a meaningful manner for humans. The lack of existing literature for poorly (or not at all) studied organisms is typically addressed

by clustering homologous genes into groups (gene families) based on sequence homology (Van Landeghem et al. 2011) . Relationships between any two gene families can then be extracted from the entire accessible literature, allowing species-independent bibliome networks to be created. This has significant implications for lesser studied species as it considerably broadens the quantity of available data for network construction.

Intuitively, a text-mining network comprises interactions that are already 'known' and thus not 'novel' in the strict sense. Novel interactions can be hypothesized, through indirect connections that involve two or more known connections. Furthermore, when the species-independent network is expanded, the network depth increases and the likelihood of uncovering correct novel relationships (both direct and indirect) decreases even further due to a reduction in the overall accuracy (i.e. by increasing the chance for false positives). In order to identify novel connections that are more likely to be true, we integrated the bibliome network with complementary networks created using available large-scale experimental data sets (transcriptome, protein-protein interaction). The criteria for a genuinely interesting novel relationship was then set to require at least two independent pieces of 'evidence'. Hence, in order to facilitate the search for potential novel gene-gene associations in large networks, we developed a rule-based algorithm to identify only those interactions that are (1) not directly linked by text-mining events yet (2) supported by links from multiple data sources. This then allows a search for both novel genes in sub-systems of interest and identification of a context (and thereby possible biological role) for orphan genes aided by gene ontology analysis. This study illustrates that text-mining not only helps identify novel genes with particular physiological, regulatory or metabolic roles but also allows network clusters and patterns with likely coordinated functions to be identified.

We are interested in the metabolism of cyanobacteria, as a potential host for sustainable biotechnology. As a proof of concept, we therefore first applied this methodology to create a network resource for the cyanobacterium *Synechocystis sp.* PCC 6803 and provide case study examples with a focus on metabolic processes of interest, including the metabolism of NADPH, nitrogen, Fe-S and alkanes.

METHODS

Construction of the networks

Molecular interaction networks were retrieved and constructed from publicly available databases and from the literature, as follows:

Networks constructed using microarray and yeast-2-hybrid data

To create a S*ynechocystis* 6803 co-expression network, 68 data sets from a large scale transcriptomics study (Singh et al. 2010) were used. The transcriptome data was collected and stored as fold change (log2 (treatment/control)) of gene expression values in tab-delimited text files (Hui et al. 2008). The data was thereafter subjected to further analyzing after importation

into the analyzing and visualizing platforms Cytoscape 2.8.2, 3.0.1 and 3.3.1 (Smoot et al. 2011, Shannon et al. 2003), depending on available plugins. The ExpressionCorrelation plugin (Hui et al. 2008) was employed to generate a co-expression network using the expression values. A similarity matrix was calculated using the Pearson correlation coefficient with a strength threshold of ±0.7. The obtained co-expression based gene network (1886 nodes and 10187 edges) is referred to as CoEx. A second yeast two-hybrid (abbreviated Y2H) protein-protein interaction network was constructed by importing into Cytoscape a list of identified protein-protein interactions from an available data set (Fields and Song 1989, Sato et al. 2007).

Text-mining data

The network from the EVEX database is composed of two data sets following the different releases of EVEX namely, EVEX-2011 and EVEX-2013. EVEX-2011 is the first public release of the EVEX text-mining database which covers the literature up until June 2011 (http://www.evexdb.org/) (Van Landeghem et al. 2011, Van Landeghem et al. 2013) . EVEX-2013 was released with the extended coverage of articles from June 2011 up to June 2012 and an updated gene family assignment. Both of the EVEX data sets (EVEX-2011 and EVEX-2013) were combined and used in the present study.

EVEX data was generated using natural language processing tools primarily based on machine learning (ML) to automatically extract cellular processes and interactions among genes and their products such as RNAs and proteins (genes for short). The tools perform three significant steps namely “name entity recognition”, “event extraction” and “name entity normalization”, which will be discussed here briefly. Firstly, the tools perform name entity recognition by identifying the gene mentions in the documents. The systems then extract the biological events for each gene mention by identifying words or phrases discussing cellular process such as *regulation* and *phosphorylation* and link them to corresponding genes. Finally, to be able to link the genes to information in other databases, genes are normalized by mapping to the Entrez Gene database and respective family identifiers. In case of organism ambiguity, i.e when the organism is not explicitly stated for a particular mention thus preventing it from being normalized to a single unique identifier, the mention is only mapped to a gene family. Full details of the EVEX text-mining pipeline generating has been described previously (Van Landeghem et al. 2013).

In this work, the text-mining network was constructed by retrieving genes from *Synechocystis* 6803 extended also by genes from other organisms that belong to the same gene families in the Ensembl resource (Kersey et al. 2012). We restricted their relationships to binding and regulatory events. The nodes of the networks and gene families were labeled with *Synechocystis* 6803 identifiers. The edges define each association (binding, regulation or indirect regulation) between genes in the families extracted from the literature. The definition of ‘binding’ and ‘regulation’ was adopted from Gene Ontology (GO) by GENIA corpus (Kim, Ohta, and Tsujii 2008). The event annotations in GENIA corpus were used for training the text-mining system. For example, GO defines regulation of phosphorylation as 'Any process that modulates the frequency, rate or extent of addition of phosphate groups into a molecule'

(http://amigo.geneontology.org/amigo/term/GO:0042325). Indirect regulation is a pairwise abstraction EVEX uses for representing regulation, co-regulation and common binding partners which are not a part of the GENIA annotation. Co-regulation and common binding partners describe the associations between two genes that regulate and bind the same target gene, respectively (Van Landeghem et al. 2012). As shown in this example (http://evexdb.org/events/45700333/), EVEX describes the association between folP and MiaB as 'indirect regulation'. The binding forms non-directed edges, while the regulation and indirect regulation form directed edges. All self-interactions were removed from the network as the focus of utility was placed on identifying new partners, and in order to minimise the number of false positives.

The networks are further supported with extra information obtained from the EVEX database. The edge attributes include organisms where the relationships between genes were studied, arbitrarily calculated taxonomic distance between the studied organisms and *Synechocystis* 6803, fine-grained details of relationship such as types of the regulation (positive, negative and unspecified), speculation, negation and text-mining prediction confidence score. The node attributes also include *Synechocystis* 6803 gene descriptions, symbols, synonyms, Entrez Gene identifiers and “gene family descriptions”.

In the NCBI Entrez Gene record, the functions of a well-characterized gene are described by human annotators based on experimental evidence. However, oftentimes the description gives no extra benefit, e.g. for genes annotated as “hypothetical”. Also, new sequences with no supporting evidence naturally lack this annotation altogether. In the latter two cases, we obtained meaningful functional annotations by assigning the single most prevalent function among the genes belonging to the same gene family.

The gene family descriptions were taken from the Entrez Gene descriptions of gene members in each family. For a small gene family (i.e. <5 genes), the diverse descriptions can be manually combined and selected to represent the common functions of genes in a given family. However, this process is not suitable for a large family with thousands of genes. To solve this problem, we used the method called “canonicalization” described in (Van Landeghem et al. 2011) to select the representative description of the family. First, we collected the descriptions of all genes in a family from NCBI Entrez Gene records. We then reduced the orthographic differences by lowering the case and removing all non-alphanumeric characters such as empty space, parentheses and apostrophes. The description of the gene family is the most common canonical form of descriptions shared by most genes in the family.

The three networks, CoEX, Y2H and EVEX, were integrated using the Cytoscape tool “Advanced network merge”. The merge was carried out based on the Entrez Gene identifiers. For those data sets that did not contain such node identifiers, these were obtained by mapping through Cyanobase gene identifiers. The resulting merged network is provided as supplementary file S1 and the attribute annotations are listed in supplementary file S2.

Annotations for genes defined as "unknown" and "hypothetical"

In this study, we were interested in the information gained for non-annotated proteins when integrating multiple types of data. We primarily used annotation data from CyanoBase downloaded on 22nd of June 2012. Genes which were not annotated or annotated as 'unknown' or 'hypothetical' in CyanoBase were instead annotated with their gene family description from Entrez Gene as described above.

Automated rule- and pattern-based sub-network detection using a script

Guilt-By-Association networks were created by extending a set of nodes in a network to include also their direct neighbors, an automated process in Cytoscape termed "First neighbors of selected nodes (undirected)". The automated rule- and pattern-based script was developed to find triangular patterns (three nodes connected by three edges, also called a triad motif (Milo et al. 2002)) from the integrated network, in order to identify relationships between selected key genes (i.e. known or relevant genes for the interested study) and candidate genes (potentially related to key gene) that are most likely to be of interest. The rules were defined as follows, except where indicated: (*i*) The triangular pattern needs to have at least two different data-types and (*ii*) no direct EVEX edge originating from *Synechocystis* 6803 is allowed between a key-gene and a candidate gene, as it is therefore already known. The ranking of the entire pattern was given according to the following order: 1) EVEX (link coming from article based on *Synechocystis* 6803), 2) EVEX (link coming from article based on any other organism than *Synechocystis* 6803), 3) CoEx, 4) Y2H. Additional ranking rules were constructed to classify the most relevant candidate genes; (*j*) does the putative candidate have additional interactions with other key genes, (*jj*) do genes with direct interactions have additional indirect links and (*jjj*) do additional direct or indirect interaction exist in the extracted pattern. These rules prioritize candidates that are well connected within the network and more related to the metabolism involving key genes.

The script for pattern candidate ranking was written in Python to query the integrated network via a Cytoscape plugin, CytoscapeRPC (Bot and Reinders 2011). CytoscapeRPC recognizes the script as client and allows the script to query or modify the networks. The developed script was adapted for the integrated network of *Synechocystis* 6803 based on EVEX, CoEx and Y2H data. The main usage of the script was not only to identify candidate genes (CG script) related to known key genes in metabolism of interest, but also to allow functional prediction of "hypothetical protein" (HP script), i.e. by identifying the function of unknown proteins from a group of functional proteins they are associated with. The ranking is identical to the key-gene script where we only substituted the role of "key genes" and "candidate genes" (Supplementary file S3) with "functional protein" (i.e. proteins with verified function) and "hypothetical protein" (Supplementary file S4) respectively.

Computational requirements and potential applications on other organisms

The *Synechocystis* 6803 network is relatively small compared to other organism networks such as humans which have in general both larger numbers of nodes and edges (e.g. 13,418 nodes and

265,738 edges) (Hakala et al. 2013). The time required to generate networks is thus only a matter of seconds on a general desktop machine. However, the integration of the network requires identifier compatibility, a general problem in integrating data from different database sources, e.g. NCBI Entrez Genes and Taxonomy databases. In this study, this task took us a few hours to manually ensure the compatibility and accuracy of the data.

Text mining Performance

Due to the variance and ambiguity inherent to human language, extracting biological knowledge from text is fundamentally a demanding task requiring a complex system composed of multiple components. While most individual components of the systems are typically evaluated in isolation by their respective developers, evaluating the integrated system is difficult due to the relative lack of gold-standard data sets. In our previous work, we estimated the performance of TEES, the text-mining system used in creating the EVEX database, by manually evaluating the text-mining network of *E. coli* NADPH metabolism. The result showed that the system can perform well on event extraction and gene family assignment, achieving 53% and 72% accuracy, respectively (Kaewphan et al. 2012). The two estimates roughly correspond to, and further verify the evaluation results of TEES on human metabolism (Björne et al. 2010). Therefore, we can expect the accuracy of the system in the extraction of the *Synechocystis* 6803 network to be similar as well.

RESULTS AND DISCUSSION

A major challenge in the evaluation of complex biological networks that have not been manually curated is to know if any of its relationship links (i.e. network edges) are (1) novel and (2) correct. By integrating networks built from experimental data and text-mining it should be possible to rapidly tell whether relationships suggested from experimental data are already known *a priori* from the literature or, the reverse. If the underlying analytical data is independent and complementary to the text-mining data, it should also be possible to boost our ability to evaluate the relative likelihood that a relationship in the integrated network is true or not (through cognitive or rule-based interpretation). This assumes that multiple pieces of evidence from genuinely independent experimental data, all implying a similar conclusion, will increase the likelihood that a suggested relationship is true. In the present study, these two concepts were applied to create a meta-network based on two network-types: (1) experimental ((*i*) transcriptome and (*ii*) protein-protein interactome) and (2) literature. The methodology was applied to the metabolism of *Synechocystis* 6803 as a specific case study.

Network construction

A species-independent text-mining network (here abbreviated EVEX) was created by first assigning all genes in the *Synechocystis* 6803 genome to gene families using Ensembl Genomes (Kersey et al. 2012). All events extracted using the TEES software (Van Landeghem et al. 2013) for these selected gene families were thereafter compiled and imported into Cytoscape (Cline et al. 2007). The thus created text-mining network was therefore composed of all machine-readable

interactions (defined *a priori*, i.e. 'examples of event triggers') between any two gene families that contain at least one homolog in *Synechocystis* 6803, accessing all literature for all species in PubMed abstracts and PubMed Central Open Access full-texts up to June 2012. In this network, the nodes represent *Synechocystis* 6803 gene symbols and edges linking the nodes represent relationships (grouped into categories of binding, regulation or indirect regulation) between gene families. As a comparison, the text-mining network created using publications studying only *Synechocystis* 6803 (79 nodes, 74 edges) was significantly smaller than that using the species-independent approach (806 nodes, 3023 edges) (Fig. 2).

Figure 2. Species-independent text-mining generates a larger network compared to a species-specific network. Text-mining network extracted from EVEX using events extracted from (A) all accessible articles or (B) only those articles including the organism name *Synechocystis* 6803. The same layout was used in both cases. In the case of (B), only those edges, and their connecting nodes, originating from literature using the species '*Synechocysis* 6803' were retained.

For the transcriptome-based network (here abbreviated CoEx), a co-expression network was constructed using a collection of published microarray data that until now only had been collectively studied with a data-degrading normalization using discrete values (Singh et al. 2010). We created a co-expression network (1886 nodes, 10187 edges) with the Cytoscape plugin ExpressionCorrelation (Hui et al. 2008). For the protein-protein interaction network (here abbreviated Y2H), we used an available qualitative protein-protein interaction data set (1920 nodes, 3236 edges) generated in a high-throughput screening with the yeast-two hybrid method (Fields and Song 1989, Sato et al. 2007).

The integration of all three networks in Cytoscape using the advanced network merge plugin resulted in a combined network (IntNet) of 2,842 nodes and 16,446 edges (Supplementary file S1), representing 76% of the genome and all of its native plasmids (Kaneko et al. 1996) (Fig. 3). An overview of the nodes that are common in the three constructed networks is presented in Figure 4. In order to ensure that all the three integrated networks were independent, two edges in the EVEX network (slll0041-sll0269, sll0041-slr1636), which originated from the paper first reporting the data used for the Y2H network, were removed from IntNet.

Figure 3. Overview of the approach – Integration of networks created using three distinct data-types. A) The selected data sets Y2H, microarray and text-mining were retrieved and pre-processed. B) Networks were constructed in Cytoscape and C) merged (IntNet) with the "advanced network merge"- plugin. D) As an example, the NADP(H)-metabolism key gene slr1843 was extracted by guilt-by-association (GBA). Automated rule-based prediction was used to extract patterns with possible novel candidate genes. A spring embedded layout was used to construct the Cytoscape view. Data-types are visualized with different colours (Y2H, red; CoEx, green; EVEX blue) to easily distinguish between them.

Figure 4. The distribution of nodes across the three (Y2H, CoEx and EVEX) networks.

Global properties

Overall, IntNet displayed surprisingly little overlap between different data-types. While 52% of the nodes (1468) are represented in at least two networks, only 11% are represented in all three (Fig. 4). The distribution of source organisms used in the species-independent text-mining network is summarized based on domains and supergroups in Figure 5. Most relationships in the EVEX network originate from studies with bacteria, the same domain of life as *Synechocystis* 6803 (Fig. 5). Within the Bacteria domain *Escherichia coli* dominates, reflecting the number of publications in PubMed (Fig. 1). The second most represented group of organisms that contributed to the *Synechocystis* 6803 text-mining network belongs to the Metazoa, with human, rat and mouse being the most common contributors.

Figure 5. The phylogenetic origin of the text-mining events used to construct the species-independent network. *Escherichia coli* K-12 is the most studied organism as demonstrated by the biggest red (number of events) and blue (number of articles) circles. Only the species (all prokaryotes) that contributed most to the species-independent network are shown.

An additional benefit with the integration of different data-types was the enhancement in the number of meaningful annotations afforded by combining annotations in CyanoBase (Nakao et al. 2010) with those provided by the gene family assignments. In the microarray data set 1913 genes (46.5% of genome) were annotated (from CyanoBase) as 'hypothetical' or 'unknown'. The integration with the species-independent text-mining network increased the number of meaningful annotations in the complete network (IntNet) by 401 additions (from 53.5% to 67.6% of the genome) through the addition of gene family annotations (listed in Supplementary file S5).

Automated rule-based selection of candidates with a high likelihood of real relationship

Smaller first neighbour (Guilt-By-Association, GBA) sub-networks were first constructed for each of the case study key gene (KG) sets. Our impression was that although GBA networks were very useful, the associated cognitive interpretation (here defined as 'manual') was dominated/biased by already existing knowledge and/or relationships only supported by a single data type. In addition, it is possible that potentially interesting relationships may not be perceived owing to the daunting complexity of larger GBA networks. We therefore developed an automated rule-based script to identify smaller motifs (also called clusters) that would enhance the search for potentially novel and relevant relationships between selected key genes (KG, known or relevant genes for the study of interest) and candidate genes (CG, having potential relationship to KG). The rule of the script was set to demand at least two different data-types between a KG and CG, of which one is direct and the second is indirect (i.e. via a third node of any type). In order to enhance the chance to identify potentially novel relationships, direct EVEX edges between KGs and CGs were allowed only if they did not originate from a study using *Synechocystis* 6803. Patterns were further divided according to the source organism of the EVEX edges,

distinguishing between edges originating from *Synechocystis* 6803 and all other species. This information is used in ranking the candidates, as described below.

The patterns were ranked in descending order of importance as follows: 1) EVEX (indirect link originating from an article based on *Synechocystis* 6803), 2) EVEX (direct or indirect link coming from article based on any other organism than *Synechocystis* 6803), 3) CoEx, 4) Y2H. The output from the automated clustering script is both different and complementary to a conventional GBA analysis since (1) patterns are ranked according to their chance of being relevant and correct, (2) relationships based only on existing knowledge (i.e. direct EVEX edges) with KGs are discarded, and (3) only patterns with multiple supportive evidence (i.e. more than one edge-type) are accepted. Despite these efforts, an unknown proportion of the edges in IntNet, and motifs extracted therefrom using the rule-based selection, are still likely to be false positives.

Utilization of the integrated network to obtain novel biological insight

What can we use IntNet and its filtered derivative networks for? The diverse utility of interaction networks has been described previously (Franceschini et al. 2013). Apart from general properties and patterns on a genome-scale level (as described above) we considered two utilities of particular value for biological studies using lesser studied species: (1) To identify novel CGs with potential relationships between a known KG or a set of KGs representing an important biological process, and (2) to probe the possible role of an otherwise unknown gene or gene set that has been identified by other means. Utility 1 would be particularly valuable with poorly studied organisms for the collation of members of pathways or other similar systems that do not display co-existence in the form of operons. Utility 2, on the other hand, would be important as a follow-up to other studies that have identified genes or proteins by experimental means (e.g. affinity chromatography, yeast-2-hybrid). To evaluate these utilities, we employed KG sets from selected case studies (Table 1) to (*i*) extract first neighbor GBA-clusters and (*ii*) sub-clusters generated from all CGs (and associated triangular patterns) derived using the automated script. The KG sets were decided prior to the study based on the research interests of the group. The clusters and networks generated by both methods were thereafter evaluated manually in order to verify potentially interesting and novel CGs and to benchmark the overall approach.

Case Study 1 - Novel candidates with a potential relationship to SigE

SigE (sll1689) is a sigma factor that has been demonstrated to influence central carbon metabolism with broad impact, as evidenced by a shift in the distribution of central carbon metabolites in response to the deletion of *sigE* or over-expression of SigE (Sundaram et al. 1998; Kloft, Rasch, and Forchhammer 2005). The first neighbour GBA and script-based clusters are shown in Figure 6A, including several interesting candidates. Firstly, we noted a link between SigE and slr1055 (ChlH), a light- and Mg^{2+}-dependent anti-sigma factor shown previously to have specificity for SigE (Osanai et al. 2009). However, this link was not based on the article that demonstrated this relationship in the first place (Osanai et al. 2009). Instead, SigE connects with ChlH through edges of all three network types, a direct Y2H edge, the lead to identifying the role of slr1055 in the first place (Osanai et al. 2009), and indirect edges via sll0306 (SigB, EVEX) and

sll1886 (hypothetical protein, CoEx). The experimentally confirmed relationship between ChlH and SigE therefore verifies the conclusion of the relationship that can be drawn from the present network even in the absence of the direct text-mining link.

Figure 6. Cluster analysis with SigE (sll1689). (A) The first neighbor GBA network using only SigE as KG. (B) The combined network of motifs extracted with the rule-based script. (C) Network generated by STRING database August 23, 2014, using standard settings and sll1689 as input. Solid EVEX edges originate from any organism other than *Synechocystis* 6803. Dotted EVEX edges originate from *Synechocystis* 6803. Black edges originate from STRING database. The KG is indicated by a white node.

Several known proteins with an established role in nitrogen-metabolism (e.g. NtcA, PII (Kloft, Rasch, and Forchhammer 2005)), or the circadian clock (KaiB (Hitomi et al. 2005)) were also found to be connected to SigE, in addition to others without any meaningful annotation. The automated script (Fig. 6B) suggested a central role for sll1886 (annotated as hypothetical protein) with a close connection to SigE. Sll1886 harbors a putative zinc binding domain and shows weak homology to di-haem cytochrome C (Vandenberghe et al. 1998), suggesting the possible involvement of electron-transfer. Interestingly, a manganese transport component (MntB, sll1600) was also part of the script-based cluster which is relevant given that ChlH is Mg^{2+}-dependent.

In comparison, we also searched for CGs to SigE using STRING-db (Franceschini et al. 2013) with sll1689 as input (Fig. 6C). This produced a network of 11 nodes at the default setting. When the script- and STRING-db based networks were compared, the intersection between the two networks contained only three genes; sll1689, sll1423 (ntcA) and sll0687 (sigI). Interestingly, whilst the STRING network contained an association with glnA, the script-based network contained an association with glnB - both genes have important roles in nitrogen metabolism (Herrero, Muro-Pastor, and Flores 2001). Overall, many of the nodes in the STRING network (Fig. 6C) are related to gene transcription (RNA polymerase related gene products), whilst the script-network (Fig. 6B) is dominated by genes with a known role in nitrogen metabolism, as has also been confirmed experimentally (Muro-Pastor, Herrero, and Flores 2001). The former network has no 'unknown' members, whilst at least one completely unknown, yet intricately connected, member (sll1886) is present in the latter network. Notably, sll1886 is co-located on the genome to a "two-component sensor histidine kinase" (sll1888) which also is a member of the same CoEx network as sll1886 and ntcA (sll1423) (Fig. 6A, 6B). This strengthens the argument that sll1886 may play an important role in nitrogen metabolism.

Case study 2 –NADP(H)-metabolism

The role of the pentose phosphate pathway (PPP) in cyanobacteria under daylight conditions is not entirely clear given that $NADP^+$ is a major electron acceptor of electrons generated by water-splitting photosynthesis. A part of the metabolic flux through the carbon fixing CBB cycle has been measured to pass through the oxidative branch of PPP (oxPPP) under daylight conditions

(Young et al. 2011) though the optimal solution for biomass flux in stoichiometric models did not incorporate any oxPPP flux (Knoop et al. 2013). We were curious about the metabolic role that key-enzymes responsible for $NADP^+$-reduction in fermentative microorganisms may have in an autotrophic system and how they are regulated. The objective in the following analysis was therefore to use the network analysis in order to identify novel CGs.

A first neighbour GBA of IntNet with all pre-defined six NADPH KGs generated a complex network of 72 nodes and 194 edges (Fig. 7A) (Supplementary file S6), including OpcA, the unique cyanobacterial Zwf activator (Hagen and Meeks 2001). In contrast, only two of the 6 KGs listed for NADP(H)-metabolism were retained by the script (Fig. 7B, 18 nodes and 50 edges): Zwf (slr1834, catalyzing the first committed step of metabolic flux into PPP) and Icd (slr1289), catalyzing the only $NADP^+$-reducing step of the TCA-"cycle".

Figure 7. Cluster analysis with NADPH-related genes. (A) The first neighbor GBA network using all NADPH-related KGs (Table 1). (B) The combined network of motifs extracted with the rule-based script. (C) Predicted pattern extracted from the script result B. (D) First neighbor GBA using PntA (slr1239) or PntB (slr1434) as input. (E) Red dotted box indicates members of the Pap operon. Solid EVEX edges originate from any organism other than *Synechocystis* 6803. Dotted EVEX edges originate from *Synechocystis* 6803.

Looking closer at the script-based network, Zwf forms a motif with slr0952 (annotated as fructose-1,6-bisphosphatase (FBPase)) and sll0508 (annotated 'unknown protein') via three different data-types (Fig. 7C). Sll0508 has low similarity to other proteins and there are no hits from a search with the SIB Motif Scan (incl. Pfam, PROSITE, HAMAP etc.). This slr0952-containing motif is interesting as it suggests a link between oxPPP and gluconeogenesis. In other cyanobacteria, multiple FBPases have been identified and some of the encoding genes are co-located with *zwf* (Summers et al. 1995).

Another interesting CG, found only in the CoEx network, is Slr1194. This node is annotated as a '1 protein' that exhibits a high percentage similarity to a 'Mo-dependent nitrogenase family' protein in *Cyanothece sp.* PCC 7424, and links to Zwf via slr1793 (talB) and slr1734. The latter gene is a homolog of OpcA, an allosteric regulator and activation factor of Zwf in other cyanobacteria (Hagen and Meeks 2001).

Zwf also forms several motifs with *rpaB* (slr0947) that also include the PPP genes *gnd* (sll0329) and *talB* (slr1793) (Fig. 7B). RpaB is a regulator involved in controlling energy transfer between phycobilisomes and PSII or PSI. The relationship between RpaB and genes encoding enzymes in PPP suggests the possibility that also PPP flux may be controlled at least in part by RpaB in response to light quality and/or quantity, or another signal reflecting the internal redox-status.

Case Study 3 - Probing the role of an incompletely known gene or gene set - PntAB

Synechocystis 6803 harbors two genes (slr1239 (*pntA*) and slr1434 (*pntB*)) encoding a putative dimeric NADPH:NADH-transhydrogenase. PntAB has been shown to catalyze the proton gradient dependent transfer of electrons from NADH to NADP(H) in *E. coli* (Sauer et al. 2004). In *Synechocystis* 6803, we would expect under optimal photosynthetic conditions that $NADP^+$ is efficiently reduced by PetH, the Ferredoxin:NADP-oxidoreductase linked to PSI. PntAB may therefore only be important for the supply of NADP(H) under conditions of limiting light (e.g. during the night) and/or in order to re-oxidize NADH formed by NAD(H)-dependent reactions (Kämäräinen et al. 2017). Hence, although PntAB is well-known in fermentative microorganisms it remains unclear what role it may have in cyanobacteria, thereby falling into the category of incompletely known genes.

No motifs satisfying the criteria of the script-based filter were found including either PntA (slr1239) or PntB (slr1434). Nevertheless, a GBA-cluster was extracted using both genes as KGs (Fig. 7D). Both slr1239 and slr1434 form a co-expression based cluster with an operon (slr0144-slr0152) called Pap (Photosystem II assembly proteins) (Wegener et al. 2008) and the essential ferredoxin PetF (slr0150; Fig. 7E). The connection is quite convincing as PntA shows CoEx edges with slr0144 whilst both PntB and PetF share CoEX edges with several of the other genes in the operon, though not slr0144. The presence or absence of the Pap operon does not influence growth under so far tested conditions, although deletion mutants display a reduced capacity to evolve di-oxygen (Wegener et al. 2008). Why would there be a connection between the Pap operon and PntAB? PntAB has the role in fermentative microorganisms of catalyzing electron-transfer between one major electron acceptor-donor and another, though not ferredoxin. Several genes of the Pap operon are predicted to contain Fe-S clusters, co-factors that typically also are involved in electron transfer, the only common theme so far; this connection deserves further experimental attention to resolve.

Case study 4 - Iron sulphur cluster metabolism

As mentioned above, iron-sulphur (Fe-S) clusters are inorganic protein co-factors that are typically involved in electron transfer. They are assembled in cyanobacteria using the SUF system, even though genes with homology to members of the ISC system (the dominant system in *E. coli*) also are present in the *Synechocystis* 6803 genome (Balasubramanian et al. 2006). It has been established that SufR (sll0088) is an Fe-S containing negative transcriptional regulator of the remaining *SUF* members (*sufA*, *sufB*, *sufC*, *sufD*, *sufS*) (Wang et al. 2004). Interestingly, a first neighbour GBA with all of the above KGs (Fig. 8A, Supplementary file S6) resulted in a single cluster with two divided parts, an upper part containing all the catalytic SUF members, and a second lower part containing SufR. Even though SufR is clearly the transcriptional regulator of the other SUF members, there is surprisingly no direct connection between SufR and the other SUF members. Instead, SufR forms an intense CoEx cluster with a series of genes annotated mainly as 'hypothetical'. Three of these are iron-related proteins: PerR (slr1738), sll1202 (homolog to iron transporters) and BfrA (sll1341; bacterioferritin homolog). In contrast, the upper SUF operon cluster contains four genes encoding predicted Fe-S containing proteins: The PSI subunit *psaA* (slr1834), *bioB* (slr1364), *sll0031* (hypothetical) and *spoT* (slr1325). A possible

reason for the lack of a direct association between SufR and the remaining SUF operon may be that SufR is not the only regulatory factor controlling SUF expression, or that its control is conditional.

Figure 8. Cluster analysis with Iron Sulfur cluster related KGs. (A) The first neighbor GBA using all members of the SUF operon as KGs (Table 1). Red asterisks indicated genes encoding proteins with a predicted Fe-S cluster binding motif. (B) Two motifs generated by the rule-based filtering script using the same KGs. Solid EVEX edges originate from any organism other than *Synechocystis* 6803. Dotted EVEX edges originate from *Synechocystis* 6803.

The rule-based script of IntNet using all Fe-S KGs generated two smaller clusters (Fig. 8B). Whilst no obvious insight was obtained from the SufR-containing motif, the second cluster contained three SUF operon members connected both by EVEX and CoEx. Interestingly, all text-mining edges originated from a diverse collection of bacteria that did not include any cyanobacteria.

Case study 5 - Alkane biosynthesis

The two genes encoding the catalytic enzymes of the alkane biosynthesis pathway (Schirmer et al. 2010), and which is uniquely present in most but not all cyanobacteria, forms an extended apparent operon in most species where it is found (Klähn et al. 2014). Since the alkane biosynthesis reaction so far does not work as efficiently as needed for economically sustainable fuel production (Eser et al. 2011, Kallio et al. 2014), we were curious whether missing elements required for effective catalysis could be represented in this apparent operon. In *Synechocystis* 6803, however, only three of the apparent operon members are co-located on the genome, sll0207-sll0209. For the assembly of KGs, we therefore included homologs in *Synechocystis* 6803 to the most commonly observed members of the alkane biosynthesis operon in cyanobacteria in general (Table 1), even if they are not co-located on the genome in *Synechocystis* 6803. In this analysis (Fig. 9, Supplementary file S6), however, most of the operon members did not form a joint cluster with the exception of slr0426. A possible contributing reason for this outcome is that the biosynthetic system is unique to cyanobacteria (Schirmer et al. 2010) and that it has not yet been studied much. Consequently, it is not well-represented in the EVEX network.

Figure 9. Cluster analysis with members of the apparent alkane operon. The first neighbor GBA of IntNet using two genes encoding catalytic enzymes in alkane biosynthesis pathways and its four most commonly observed co-locating genes in all cyanobacteria. Solid EVEX edges originate from any organism other than Synechocystis 6803. Dotted EVEX edges originate from Synechocystis 6803. The KGs are indicated by white nodes.

Case study 6 – Screening for the role of genes annotated as 'hypothetical' or 'unknown'

We considered the possibility to utilize the script in order to obtain an insight into the possible role of all genes that are annotated as 'hypothetical' or 'unknown'. The rationale was that the

local context of genes without an annotation may provide insight into its possible role and that the script would allow the most important local context to be identified. All genes without an annotation were therefore employed, one at a time, as an entry gene for the automated script. The criteria of this script demanded as previously that more than one relationship type was present, plus the additional new demand that at least one of the members of the local context had an existing annotation. Over 5% of hypothetical/ unknown genes (112/1913) satisfied these criteria. The combined network with rule-based pattern motifs was composed of 331 nodes. Figure 10A is illustrating 112 motifs/candidate genes. (Supplementary file S7). Around 60% of these patterns were derived from a combination of CoEx/Y2H and around 40% from EVEX/(CoEX/Y2H). These 112 putative genes represent a list of potentially interesting genes to be studied further (Supplementary file S8). Many of the entry genes with highest ranking have a local context with a clear single focus. For example, sll0543 forms a cluster with genes encoding three key members of PSI (*psaC*, *psaB*, *psaD*) (Fig. 10B). In contrast, a similar analysis with STRING places sll0543 in a cluster of 8 genes annotated as 'hypothetical protein' and one as 'indole-3-glycerol-phosphate synthase'. In another example, the slr0144-48 Pap operon (see case study 3) is once again identified (Fig. 10C). Interestingly, in this search, the Pap operon genes form a cluster together with two PSI subunits (psaB and psaD): the only earlier study linked the Pap operon to PSII, not PSI (Wegener et al. 2008). Other selected findings include unknown genes slr0723 and sll1774 forming an intricate cluster with two genes encoding proteins with a role in pili biogenesis (slr0161, slr0163) and another gene linked to chemotaxis (slr1043). The 'unknown protein' slr1187 forms a cluster with three NADH dehydrogenase subunits (slr1279-81) (Fig. 10D), and the 'hypothetical protein' slr2003 forms a cluster with two nitrate/nitrite transport system components (slr1450-51) (Fig. 10E).

Figure 10. Cluster analysis for the role of genes annotated as 'hypothetical' or 'unknown'
(A) The combined network of motifs extracted with the rule-based script. (B) sll0543, as an example pattern with highest ranking, forms a cluster with genes encoding three key members of PSI (*psaC*, *psaB*, *psaD*). (C) slr0144-48 as another example (see Fig. 6D). (D) The 'unknown protein' slr1187 forms a cluster with three NADH dehydrogenase subunits (slr1279-81) (E) 'hypothetical protein' slr2003 forms a cluster with two nitrate/nitrite transport system components (slr1450-51).

CONCLUSIONS

This study incorporates species-independent text-mining for the creation and evaluation of biological networks. Although it is evaluated first with an established model organism, this approach is likely to have even greater utility with "new" species that until now have not been studied, particularly if it can be complemented by omics analysis at a sufficient depth to enable supporting networks to be constructed and integrated with the text-mining network.

Although the analysis of the *Synechocystis* 6803 network was constrained in scope, it uncovered many leads and insight into its metabolism and potentially also cyanobacteria in general. For example, the strong apparent connection between the Pap operon and both PSI and PntAB, in

addition to PSII as earlier reported. The lack of a clear connection between the alkane biosynthesis genes and other members of its apparent operon in other cyanobacteria was also surprising, though negative. Other leads included sll1886, SigE and nitrogen metabolism, sll0508/slr0952 and NADPH-metabolism, RpaB/slr0947 and PPP, sll0543 and *psaBCD*, slr1187 and *ndhCJK*, and slr2003 and *nrtAB*. Thus, a large number of candidate genes with potential involvement in important biological processes in cyanobacteria were identified in only the small selection of case studies presented here, the entire network certainly contains many more.

The automated script allows the potentially most important candidates to be selected given that it relies exclusively on connections that are supported by multiple and independent evidence. It must be pointed out, however, that these automated procedures cannot replace the need for further in-depth cognitive analysis of existing literature, though it may have an important guiding role, and final experimental verification. The script is expected to speed up the identification of the most interesting candidates and allow researchers to place a focus for further cognitive and experimental work, and in so doing contribute to reducing the proportion of 'unknown' or 'hypothetical' proteins.

The analysis of *Synechocystis* 6803 is likely to be further enhanced by future high-quality omics data sets, ideally from the same condition(s). In general, an extension of the EVEX event capture to include also metabolites would enable metabolic stoichiometric networks to also be included. Greater access to full-text articles is also likely to enhance the network richness and accumulation of multiple independent lines of evidence.

Acknowledgements

Computational resources were provided by CSC— IT Center for Science Ltd, Espoo, Finland. We would like to thank Sofie Van Landeghem (Ghent University) and Tero Aitokallio (FIMM) for their insight and valuable comments on text-mining, systems and network biology.

Supplementary Files

S1 File. Cytoscape file containing independent and merged networks. Opens with Cytoscape 3.1

S2 File. Text-file describing the annotations in the Cytoscape files
S3 File. Python file containing candidate gene script
S4 File. Python file containing hypothetical gene script
S5 File. Text-file containing annotations from EVEX/Cyanobase
S6 File. Cytoscape file containing the first neighbour GBA and script-based clusters used in the case studies. Opens with Cytoscape 3.1

S7 File. Cytoscape file containing all genes in the genome of *Synechocystis* 6803 without an annotation that forms a motif with at least two other nodes via at least two different data-types (i.e. edges), of which one is direct and the second is indirect, and at least one of the members of the motif has an existing annotation. Opens with Cytoscape 3.1

S8 File. Text-file containing list of possible candidates of hypotheticals

References

Balasubramanian, R., G. Shen, D. A. Bryant, and J. H. Golbeck. 2006. "Regulatory roles for IscA and SufA in iron homeostasis and redox stress responses in the cyanobacterium Synechococcus sp. strain PCC 7002." *J Bacteriol* 188 (9):3182-91. doi: 10.1128/JB.188.9.3182-3191.2006.

Bhadauriya, P., R. Gupta, S. Singh, and P. S. Bisen. 2007. "Physiological and biochemical alterations in a diazotrophic cyanobacterium Anabaena cylindrica under NaCl stress." *Curr Microbiol* 55 (4):334-8. doi: 10.1007/s00284-007-0191-1.

Björne, Jari, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. 2010. "Scaling up biomedical event extraction to the entire PubMed" In Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, pp. 28-36. Association for Computational Linguistics.

Bot, J. J., and M. J. Reinders. 2011. "CytoscapeRPC: a plugin to create, modify and query Cytoscape networks from scripting languages." *Bioinformatics* 27 (17):2451-2. doi: 10.1093/bioinformatics/btr388.

Cline, M. S., M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. R. Pico, A. Vailaya, P. L. Wang, A. Adler, B. R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker, and G. D. Bader. 2007. "Integration of biological networks and gene expression data using Cytoscape." *Nat Protoc* 2 (10):2366-82. doi: 10.1038/nprot.2007.324.

De Bodt, S., J. Hollunder, H. Nelissen, N. Meulemeester, and D. Inzé. 2012. "CORNET 2.0: integrating plant coexpression, protein-protein interactions, regulatory interactions, gene associations and functional annotations." *New Phytol* 195 (3):707-20. doi: 10.1111/j.1469-8137.2012.04184.x.

Eser, B. E., D. Das, J. Han, P. R. Jones, and E. N. Marsh. 2011. "Oxygen-independent alkane formation by non-heme iron-dependent cyanobacterial aldehyde decarbonylase: investigation of kinetics and requirement for an external electron donor." *Biochemistry* 50 (49):10743-50. doi: 10.1021/bi2012417.

Fields, S., and O. Song. 1989. "A novel genetic system to detect protein-protein interactions." *Nature* 340 (6230):245-6. doi: 10.1038/340245a0.

Franceschini, A., D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering, and L. J. Jensen. 2013. "STRING v9.1: protein-protein

interaction networks, with increased coverage and integration." *Nucleic Acids Res* 41 (Database issue):D808-15. doi: 10.1093/nar/gks1094.

Fujisawa, T., Narikawa, R., Maeda, S.-I., Watanabe, S., Kanesaki, Y., Kobayashi, K., Nomata, J., Hanaoka, M., Watanabe, M., Suzuki, S. E. E., Awai, K., Nakamura, Y. 2017. "CyanoBase: a large-scale update on its 20th anniversary". *Nucleic Acids Res* 45 (D1):D551-D554. doi: 10.1093/nar/gkw1131.

Gonzalez, G. H., T. Tahsin, B. C. Goodale, A. C. Greene, and C. S. Greene. 2016. "Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery." *Brief Bioinform* 17 (1):33-42. doi: 10.1093/bib/bbv087.

Hagen, K. D., and J. C. Meeks. 2001. "The unique cyanobacterial protein OpcA is an allosteric effector of glucose-6-phosphate dehydrogenase in Nostoc punctiforme ATCC 29133." *J Biol Chem* 276 (15):11477-86. doi: 10.1074/jbc.M010472200.

Hakala Kai, and Mehryary Farrokh, Kaewphan Suwisa, and Ginter Filip. 2013. "Hypothesis Generation in Large-Scale Event Networks."

Hernandez-Prieto, M. A., and M. E. Futschik. 2012. "CyanoEXpress: A web database for exploration and visualisation of the integrated transcriptome of cyanobacterium Synechocystis sp. PCC6803." *Bioinformation* 8 (13):634-8. doi: 10.6026/97320630008634.

Herrero, Antonia, Alicia M. Muro-Pastor, and Enrique Flores. 2001. "Nitrogen Control in Cyanobacteria." *Journal of Bacteriology* 183 (2):411-425. doi: 10.1128/JB.183.2.411-425.2001.

Hitomi, K., T. Oyama, S. Han, A. S. Arvai, and E. D. Getzoff. 2005. "Tetrameric architecture of the circadian clock protein KaiB. A novel interface for intermolecular interactions and its impact on the circadian rhythm." *J Biol Chem* 280 (19):19127-35. doi: 10.1074/jbc.M411284200.

Cytoscape ExpressionCorrelation plugin 1.01. Cytoscape, Cytoscape App Store.

Ikeuchi, M, and S Tabata. 2001. "Synechocystis sp. PCC 6803 - a useful tool in the study of the genetics of cyanobacteria." *Photosynth Res* 70 (1):73-83.

Kaewphan Suwisa, Kreula Sanna, Van Landeghem Sofie, Van de Peer Yves, Jones Patrik R., and Ginter Filip. 2012. "Integrating Large-Scale Text Mining and Co-Expression Networks: Targeting NADP(H) Metabolism in E. coli with Event Extraction."

Kallio, P., A. Pásztor, K. Thiel, M. K. Akhtar, and P. R. Jones. 2014. "An engineered pathway for the biosynthesis of renewable propane." *Nat Commun* 5:4731. doi: 10.1038/ncomms5731.

Kaneko, T, S Sato, H Kotani, A Tanaka, E Asamizu, Y Nakamura, N Miyajima, M Hirosawa, M Sugiura, S Sasamoto, T Kimura, T Hosouchi, A Matsuno, A Muraki, N Nakazaki, K Naruo, S Okumura, S Shimpo, C Takeuchi, T Wada, A Watanabe, M Yamada, M Yasuda, and S Tabata. 1996. "Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions." *DNA Res* 3 (3):109-36.

Kersey, P. J., D. M. Staines, D. Lawson, E. Kulesha, P. Derwent, J. C. Humphrey, D. S. Hughes, S. Keenan, A. Kerhornou, G. Koscielny, N. Langridge, M. D. McDowall, K. Megy, U.

674 Maheswari, M. Nuhn, M. Paulini, H. Pedro, I. Toneva, D. Wilson, A. Yates, and E. Birney.
675 2012. "Ensembl Genomes: an integrative resource for genome-scale data from non-
676 vertebrate species." *Nucleic Acids Res* 40 (Database issue):D91-7. doi:
677 10.1093/nar/gkr895.
678 Kim, J. D., T. Ohta, and J. Tsujii. 2008. "Corpus annotation for mining biomedical events from
679 literature." *BMC Bioinformatics* 9:10. doi: 10.1186/1471-2105-9-10.
680 Kloft, N., G. Rasch, and K. Forchhammer. 2005. "Protein phosphatase PphA from Synechocystis
681 sp. PCC 6803: the physiological framework of PII-P dephosphorylation." *Microbiology*
682 151 (Pt 4):1275-83. doi: 10.1099/mic.0.27771-0.
683 Klähn, S., D. Baumgartner, U. Pfreundt, K. Voigt, V. Schön, C. Steglich, and W. R. Hess. 2014.
684 "Alkane Biosynthesis Genes in Cyanobacteria and Their Transcriptional Organization."
685 *Front Bioeng Biotechnol* 2:24. doi: 10.3389/fbioe.2014.00024.
686 Knoop, H., M. Gründel, Y. Zilliges, R. Lehmann, S. Hoffmann, W. Lockau, and R. Steuer. 2013.
687 "Flux Balance Analysis of Cyanobacterial Metabolism: The Metabolic Network of
688 Synechocystis sp. PCC 6803." *PLoS Comput Biol* 9 (6):e1003081. doi:
689 10.1371/journal.pcbi.1003081.
690 Kämäräinen, J., T. Huokko, S. Kreula, P. R. Jones, E. M. Aro, and P. Kallio. 2017. "Pyridine
691 nucleotide transhydrogenase PntAB is essential for optimal growth and photosynthetic
692 integrity under low-light mixotrophic conditions in Synechocystis sp. PCC 6803." *New*
693 *Phytol* 214 (1):194-204. doi: 10.1111/nph.14353.
694 Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. "Network
695 motifs: simple building blocks of complex networks." *Science* 298 (5594):824-7. doi:
696 10.1126/science.298.5594.824.
697 Muro-Pastor, A. M., A. Herrero, and E. Flores. 2001. "Nitrogen-regulated group 2 sigma factor
698 from Synechocystis sp. strain PCC 6803 involved in survival under nitrogen stress." *J*
699 *Bacteriol* 183 (3):1090-5. doi: 10.1128/JB.183.3.1090-1095.2001.
700 Nakao, M., S. Okamoto, M. Kohara, T. Fujishiro, T. Fujisawa, S. Sato, S. Tabata, T. Kaneko, and
701 Y. Nakamura. 2010. "CyanoBase: the cyanobacteria genome database update 2010."
702 *Nucleic Acids Res* 38 (Database issue):D379-81. doi: 10.1093/nar/gkp915.
703 Osanai, T., M. Imashimizu, A. Seki, S. Sato, S. Tabata, S. Imamura, M. Asayama, M. Ikeuchi, and
704 K. Tanaka. 2009. "ChlH, the H subunit of the Mg-chelatase, is an anti-sigma factor for
705 SigE in Synechocystis sp. PCC 6803." *Proc Natl Acad Sci U S A* 106 (16):6860-5. doi:
706 10.1073/pnas.0810040106.
707 Rosgaard, L., A. J. de Porcellinis, J. H. Jacobsen, N. U. Frigaard, and Y. Sakuragi. 2012.
708 "Bioengineering of carbon fixation, biofuels, and biochemicals in cyanobacteria and
709 plants." *J Biotechnol*. doi: S0168-1656(12)00278-7 [pii] 10.1016/j.jbiotec.2012.05.006.
710 Sato, S., Y. Shimoda, A. Muraki, M. Kohara, Y. Nakamura, and S. Tabata. 2007. "A large-scale
711 protein protein interaction analysis in Synechocystis sp. PCC6803." *DNA Res* 14 (5):207-
712 16. doi: 10.1093/dnares/dsm021.
713 Sauer, U, F Canonaco, S Heri, A Perrenoud, and E Fischer. 2004. "The soluble and membrane-
714 bound transhydrogenases UdhA and PntAB have divergent functions in NADPH
715 metabolism of Escherichia coli." *J Biol Chem* 279 (8):6613-9.

716 Schirmer, A., M. A. Rude, X. Li, E. Popova, and S. B. del Cardayre. 2010. "Microbial
717 biosynthesis of alkanes." *Science* 329 (5991):559-62. doi: 10.1126/science.1187936.
718 Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B.
719 Schwikowski, and T. Ideker. 2003. "Cytoscape: a software environment for integrated
720 models of biomolecular interaction networks." *Genome Res* 13 (11):2498-504. doi:
721 10.1101/gr.1239303.
722 Singh, A. K., T. Elvitigala, J. C. Cameron, B. K. Ghosh, M. Bhattacharyya-Pakrasi, and H. B.
723 Pakrasi. 2010. "Integrative analysis of large scale expression profiles reveals core
724 transcriptional response and coordination between multiple cellular processes in a
725 cyanobacterium." *BMC Syst Biol* 4:105. doi: 10.1186/1752-0509-4-105.
726 Smoot, M. E., K. Ono, J. Ruscheinski, P. L. Wang, and T. Ideker. 2011. "Cytoscape 2.8: new
727 features for data integration and network visualization." *Bioinformatics* 27 (3):431-2. doi:
728 10.1093/bioinformatics/btq675.
729 Summers, M. L., Wallis, J. G., Campbell, E. L., Meeks, J. C. 1995. "Genetic evidence of a major
730 role for glucose-6-phosphate dehydrogenase in nitrogen fixation and dark growth of the
731 cyanobacterium Nostoc sp. strain ATCC 29133." *J Bacteriol* 177 (21):6184-94. doi:
732 10.1128/jb.177.21.6184-6194.1995
733 Sundaram, S., H. Karakaya, D. J. Scanlan, and N. H. Mann. 1998. "Multiple oligomeric forms of
734 glucose-6-phosphate dehydrogenase in cyanobacteria and the role of OpcA in the
735 assembly process." *Microbiology* 144 (Pt 6):1549-56.
736 Van Landeghem, S., J. Björne, C. H. Wei, K. Hakala, S. Pyysalo, S. Ananiadou, H. Y. Kao, Z. Lu,
737 T. Salakoski, Y. Van de Peer, and F. Ginter. 2013. "Large-scale event extraction from
738 literature with multi-level gene normalization." *PLoS One* 8 (4):e55814. doi:
739 10.1371/journal.pone.0055814.
740 Van Landeghem, Sofie, Kai Hakala, Samuel Rönnqvist, Tapio Salakoski, Yves Van de Peer, and
741 Filip Ginter. 2012. "Exploring Biomolecular Literature with EVEX: Connecting Genes
742 through Events, Homology, and Indirect Associations." *Advances in Bioinformatics*
743 2012:12. doi: 10.1155/2012/582765.
744 Van Landeghem, Sofie Van, Filip Ginter, Yves Van de Peer, and Tapio Salakoski. 2011. "EVEX: a
745 pubmed-scale resource for homology-based generalization of text mining predictions."
746 Proceedings of BioNLP 2011 Workshop, Portland, Oregon.
747 Vandenberghe, I., D. Leys, H. Demol, G. Van Driessche, T. E. Meyer, M. A. Cusanovich, and J.
748 Van Beeumen. 1998. "The primary structures of the low-redox potential diheme
749 cytochromes c from the phototrophic bacteria Rhodobacter sphaeroides and Rhodobacter
750 adriaticus reveal a new structural family of c-type cytochromes." *Biochemistry* 37
751 (38):13075-81. doi: 10.1021/bi981076z.
752 Wang, T., G. Shen, R. Balasubramanian, L. McIntosh, D. A. Bryant, and J. H. Golbeck. 2004.
753 "The sufR gene (sll0088 in Synechocystis sp. strain PCC 6803) functions as a repressor of
754 the sufBCDS operon in iron-sulfur cluster biogenesis in cyanobacteria." *J Bacteriol* 186
755 (4):956-67.
756 Wegener, Kimberly M., Eric A. Welsh, Leeann E. Thornton, Nir Keren, Jon M. Jacobs, Kim K.
757 Hixson, Matthew E. Monroe, David G. Camp, Richard D. Smith, and Himadri B. Pakrasi.

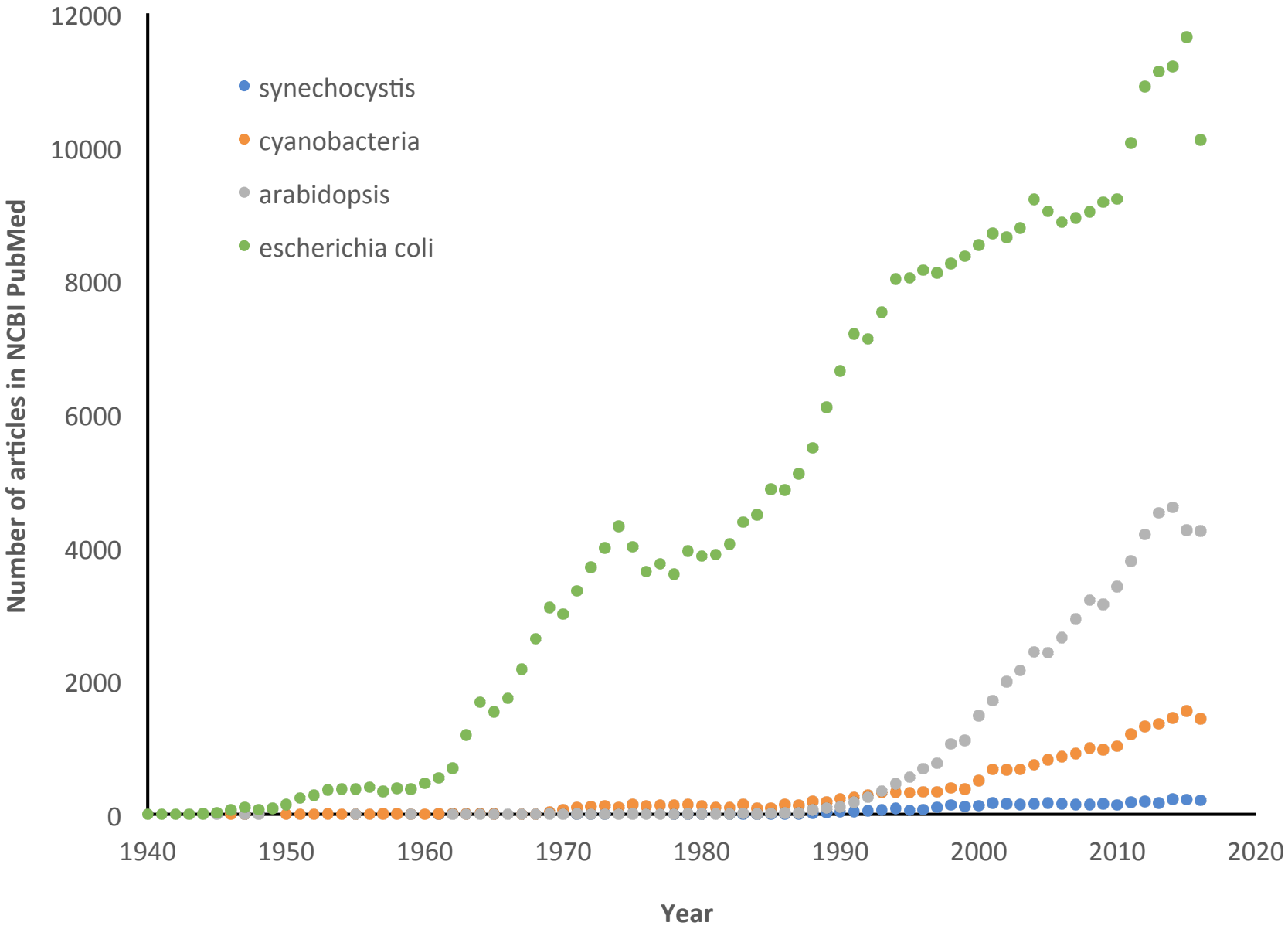
2008. "High Sensitivity Proteomics Assisted Discovery of a Novel Operon Involved in the Assembly of Photosystem II, a Membrane Protein Complex." J*ournal of Biological Chemistry* 283 (41):27829-27837.

Young, J. D., A. A. Shastri, G. Stephanopoulos, and J. A. Morgan. 2011. "Mapping photoautotrophic metabolism with isotopically nonstationary (13)C flux analysis." M*etab Eng.* doi: S1096-7176(11)00088-7 [pii] 10.1016/j.ymben.2011.08.002.

Figure 1 (on next page)

There are few publications for cyanobacteria in comparison to other model species such as *Escherichia coli*.

The search terms ‘Synechocystis’ (representing *Synechocystis sp.* PCC 6803), ‘Cyanobacteria’, ‘Arabidopsis” (representing the model plant *Arabidopsis thaliana*) and ‘Escherichia coli’ were entered into PubMed (http://www.ncbi.nlm.nih.gov/pubmed) July 2017. The numbers shown in the figure were obtained from this website by selecting “Results by year”.

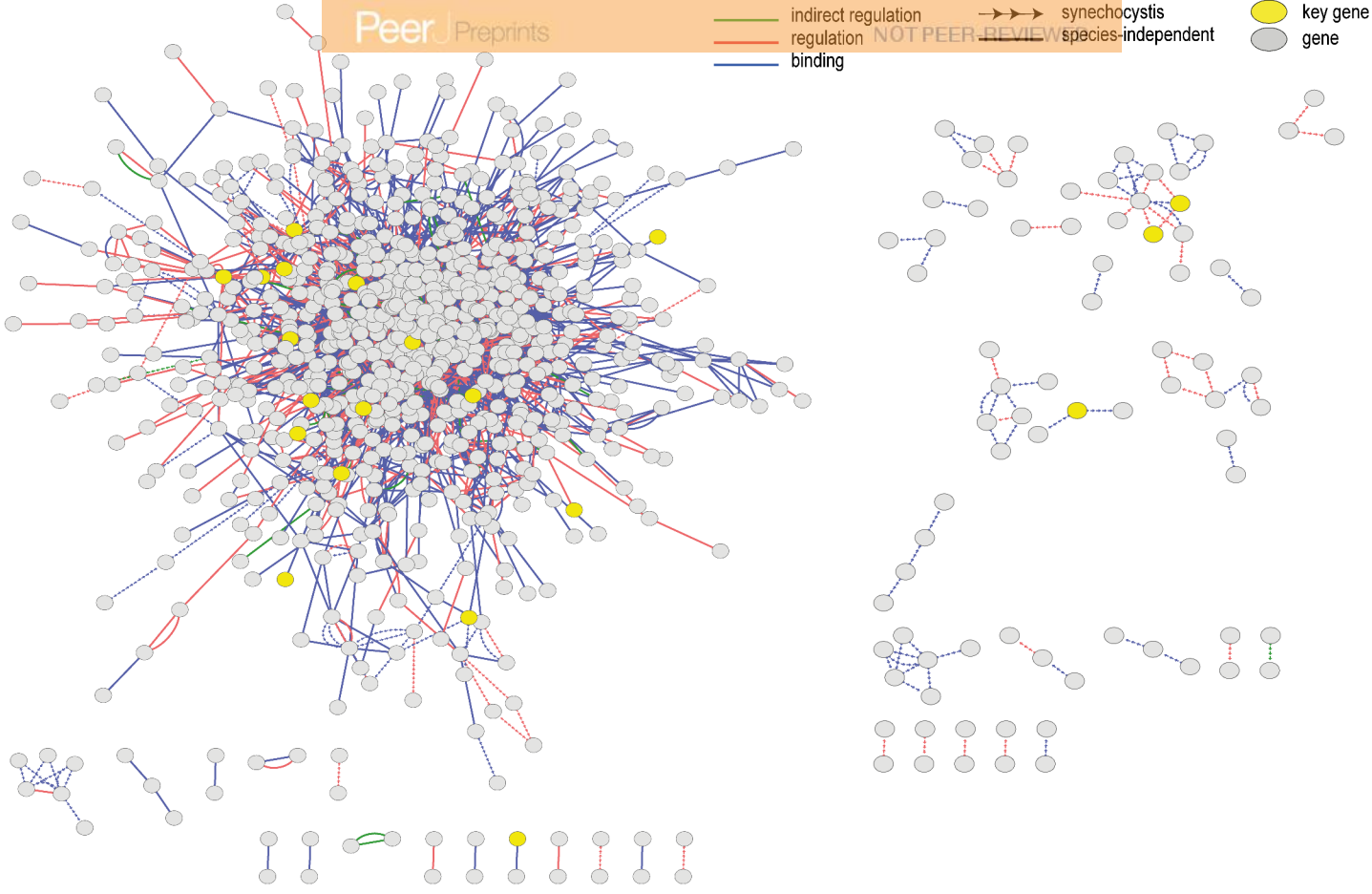


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3472v1 | CC BY 4.0 Open Access | rec: 18 Dec 2017, publ: 18 Dec 2017

Figure 2(on next page)

Species-independent text-mining generates a larger network compared to a species-specific network

Text-mining network extracted from EVEX using events extracted from (A) all accessible articles or (B) only those articles including the organism name *Synechocystis* 6803. The same layout was used in both cases. In the case of (B), only those edges, and their connecting nodes, originating from literature using the species '*Synechocysis* 6803' were retained.



Entire EVEX network (806 nodes 3023 edges)
Key genes highlighted (yellow color) in the network (17/24)

Synechocystis sp. PCC 6803 specific
EVEX network (79 nodes 74 edges)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3472v1 | CC BY 4.0 Open Access | rec: 18 Dec 2017, publ: 18 Dec 2017

Figure 3(on next page)

Overview of the approach – Integration of networks created using three distinct data-types

A) The selected data sets Y2H, microarray and text-mining were retrieved and pre-processed. B) Networks were constructed in Cytoscape and C) merged (IntNet) with the “advanced network merge”- plugin. D) As an example, the NADP(H)-metabolism key gene slr1843 was extracted by guilt-by-association (GBA). Automated rule-based prediction was used to extract patterns with possible novel candidate genes. A spring embedded layout was used to construct the Cytoscape view. Data-types are visualized with different colours (Y2H, red; CoEx, green; EVEX blue) to easily distinguish between them.

A

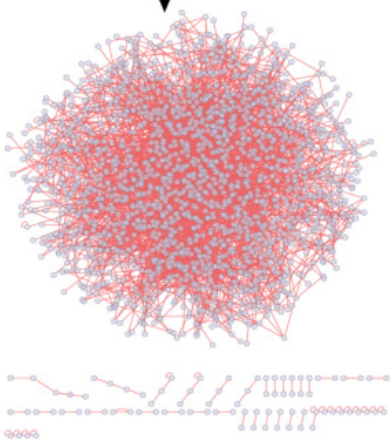
Experimental data

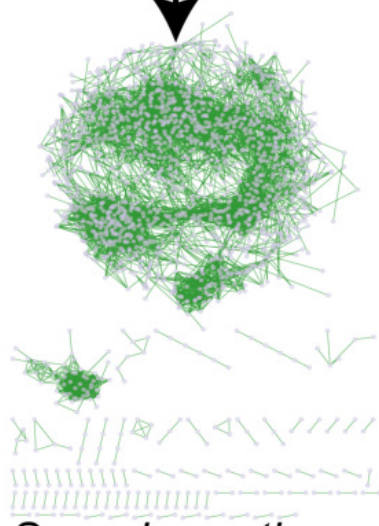
Y2H

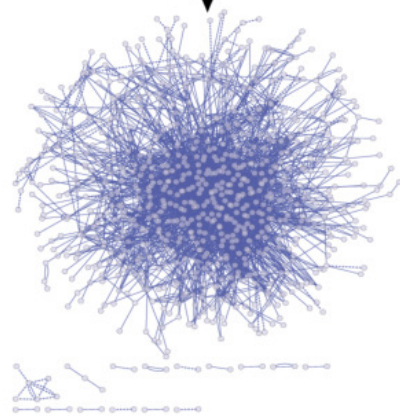
Microarray

Text mining
Event Extraction

B





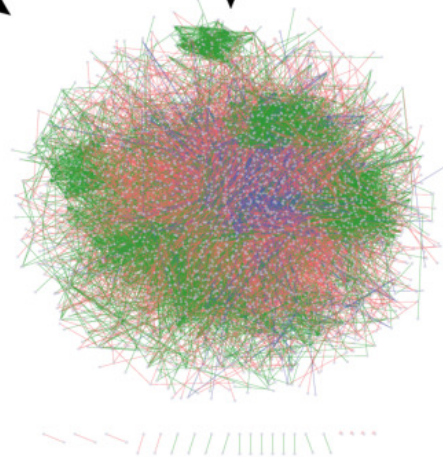


Synechocystis sp. PCC 6803 interactome

Synechocystis sp. PCC 6803 interactome

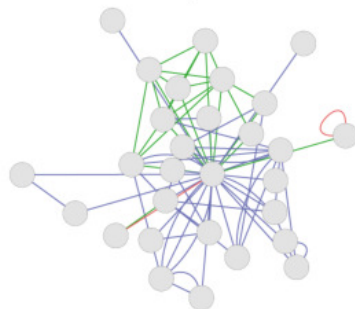
Synechocystis sp. PCC 6803 with species-independent interactome

C

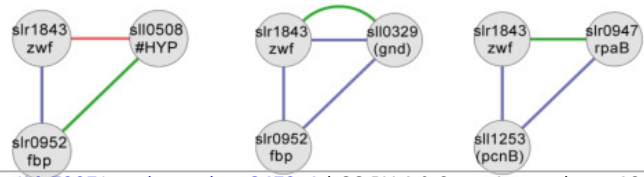


Integrated network
2842 nodes
16446 edges

D



First neighbours of NADPH-metabolism key gene slr1843 (29 nodes, 73 edges)

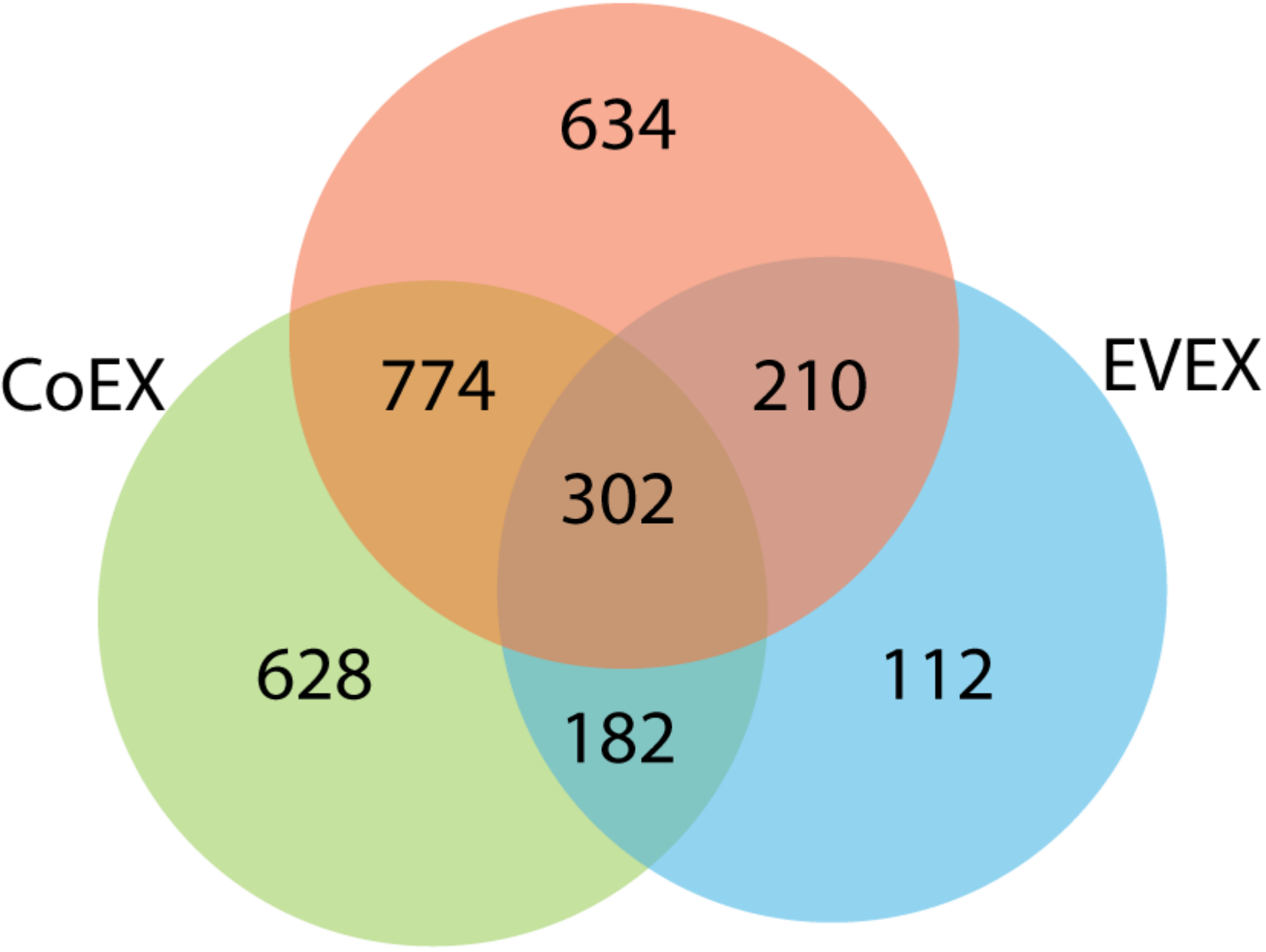


Automated rule-based prediction of NADPH-metabolism

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3472v1 | CC BY 4.0 Open Access | rec: 18 Dec 2017, publ: 18 Dec 2017

Figure 4(on next page)

The distribution of nodes across the three (Y2H, CoEx and EVEX) networks



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3472v1 | CC BY 4.0 Open Access | rec: 18 Dec 2017, publ: 18 Dec 2017

Figure 5

The phylogenetic origin of the text-mining events used to construct the species-independent network

Escherichia coli K-12 is the most studied organism as demonstrated by the biggest red (number of events) and blue (number of articles) circles. Only the species (all prokaryotes) that contributed most to the species-independent network are shown

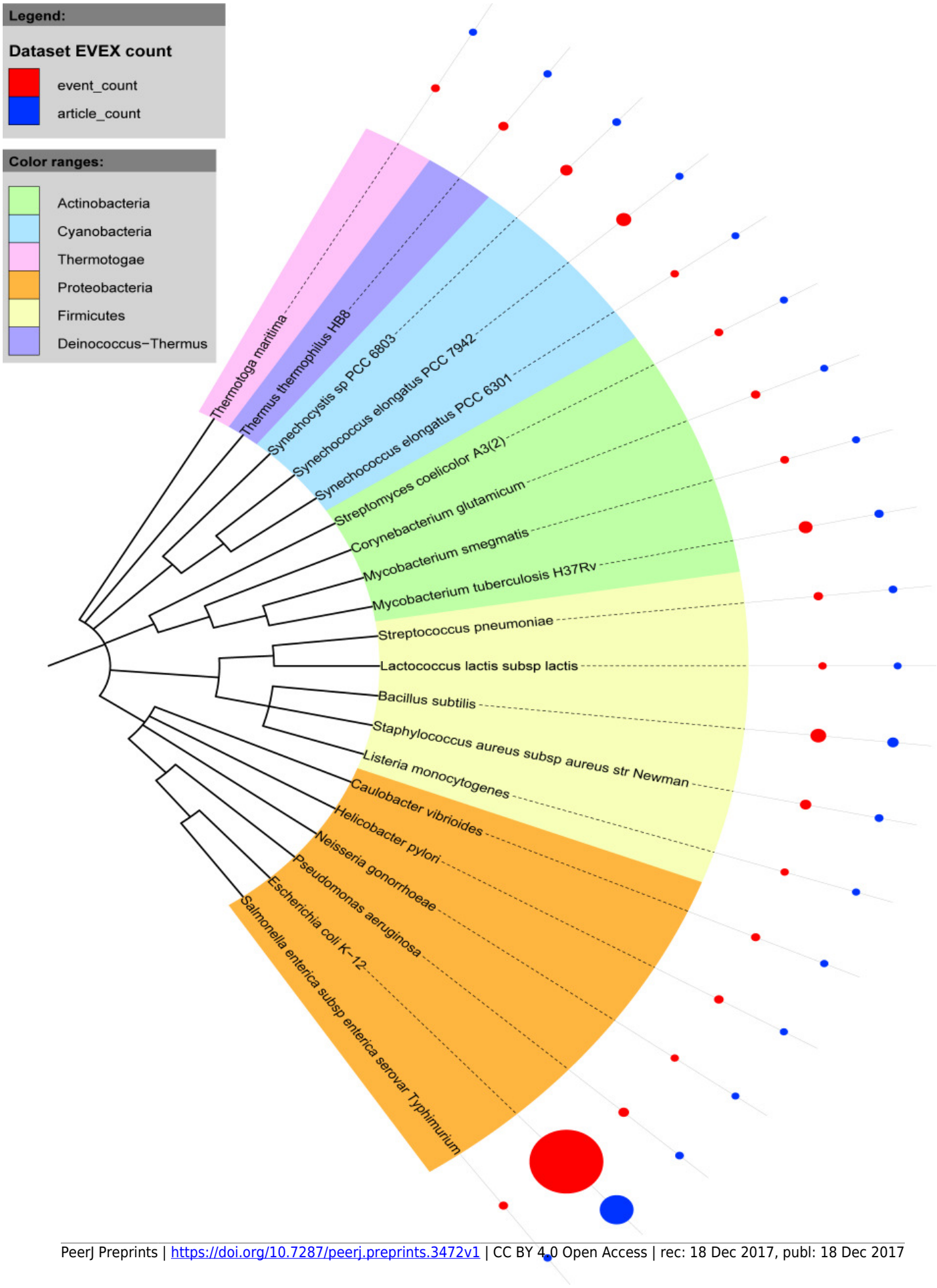


Figure 6(on next page)

Cluster analysis with SigE (sll1689)

(A) The first neighbor GBA network using only SigE as KG. (B) The combined network of motifs extracted with the rule-based script. (C) Network generated by STRING database August 23, 2014, using standard settings and sll1689 as input. Solid EVEX edges originate from any organism other than *Synechocystis* 6803. Dotted EVEX edges originate from *Synechocystis* 6803. Black edges originate from STRING database. The KG is indicated by a white node.

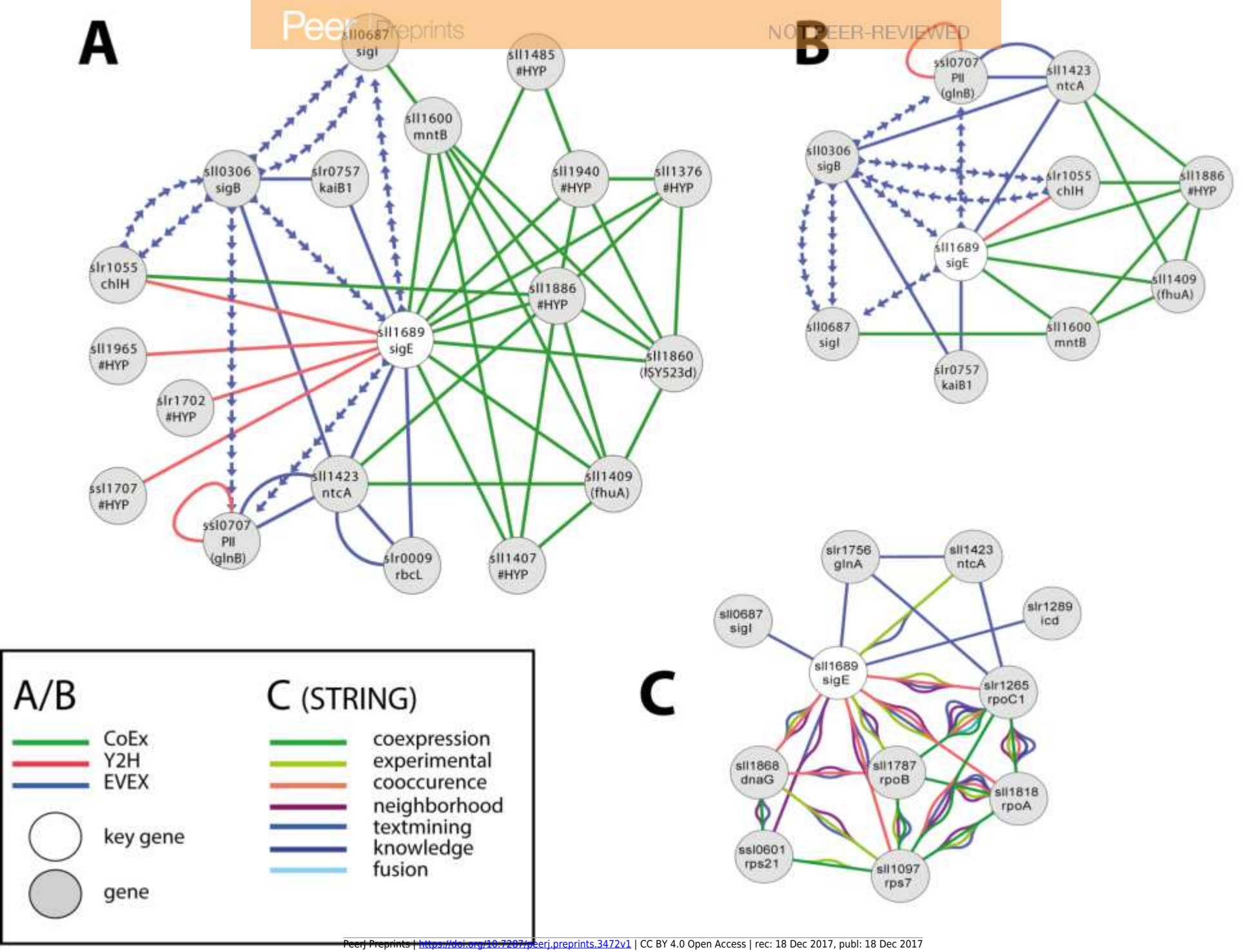


Figure 7

Cluster analysis with NADPH-related genes

(A) The first neighbor GBA network using all NADPH-related KGs (Table 1). (B) The combined network of motifs extracted with the rule-based script. (C) Predicted pattern extracted from the script result B. (D) First neighbor GBA using PntA (slr1239) or PntB (slr1434) as input. (E) Red dotted box indicates members of the Pap operon. Solid EVEX edges originate from any organism other than *Synechocystis* 6803. Dotted EVEX edges originate from *Synechocystis* 6803.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3472v1 | CC BY 4.0 Open Access | rec: 18 Dec 2017, publ: 18 Dec 2017

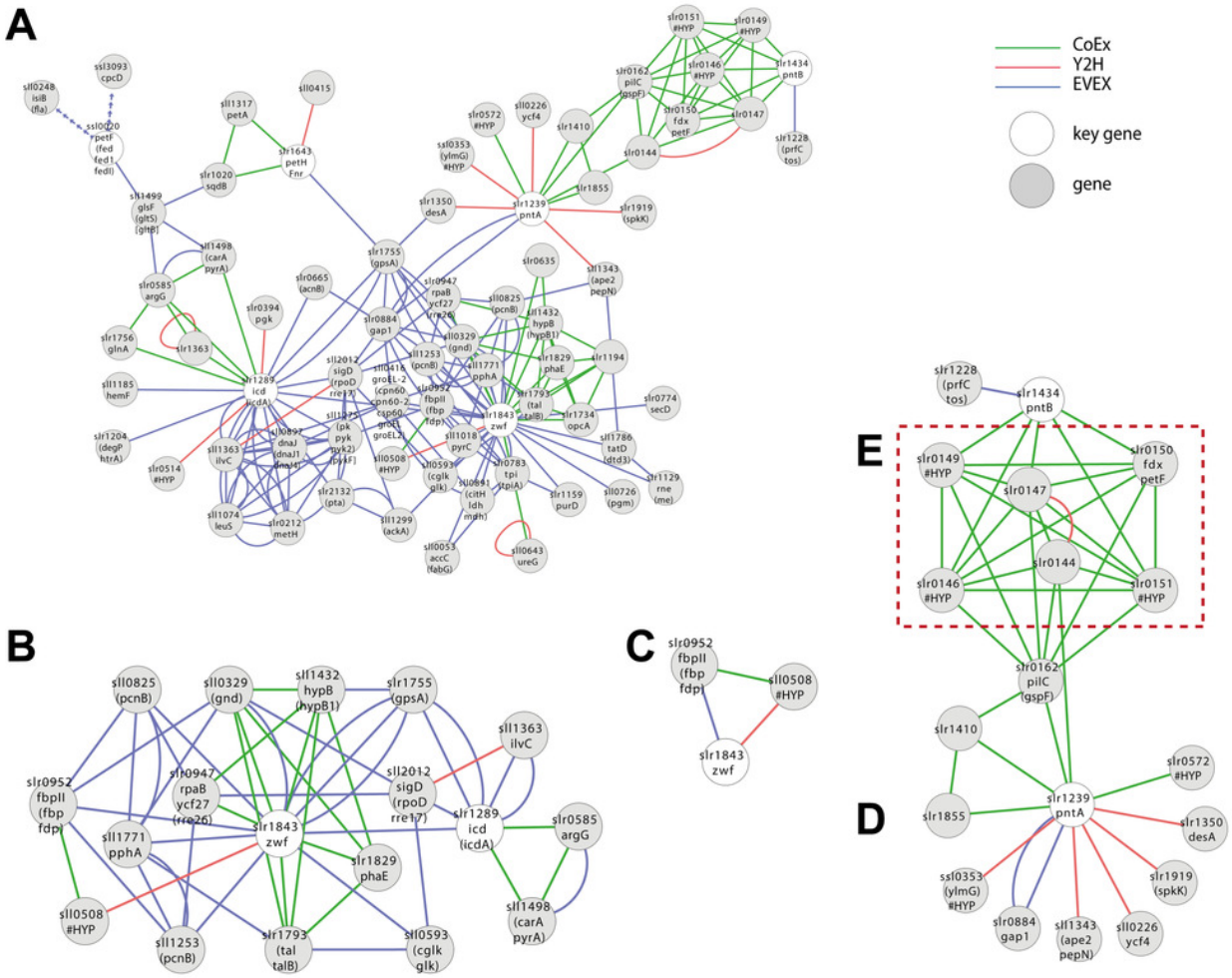


Figure 8

Cluster analysis with Iron Sulfur cluster related KGs

(A) The first neighbor GBA using all members of the SUF operon as KGs (Table 1). Red asterisks indicated genes encoding proteins with a predicted Fe-S cluster binding motif. (B) Two motifs generated by the rule-based filtering script using the same KGs. Solid EVEX edges originate from any organism other than *Synechocystis* 6803. Dotted EVEX edges originate from *Synechocystis* 6803.

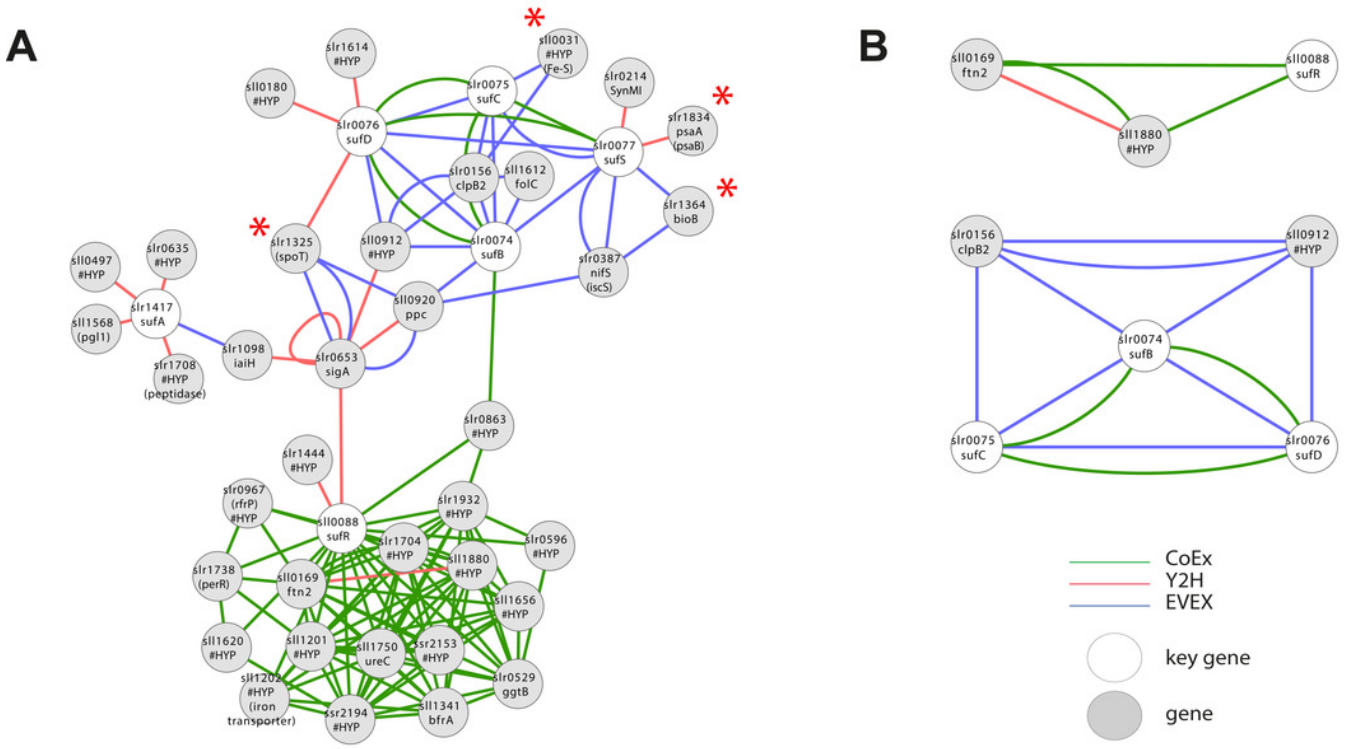


Figure 9

Cluster analysis with members of the apparent alkane operon

The first neighbor GBA of IntNet using two genes encoding catalytic enzymes in alkane biosynthesis pathways and its four most commonly observed co-locating genes in all cyanobacteria. Solid EVEX edges originate from any organism other than Synechocystis 6803. Dotted EVEX edges originate from Synechocystis 6803. The KGs are indicated by white nodes.

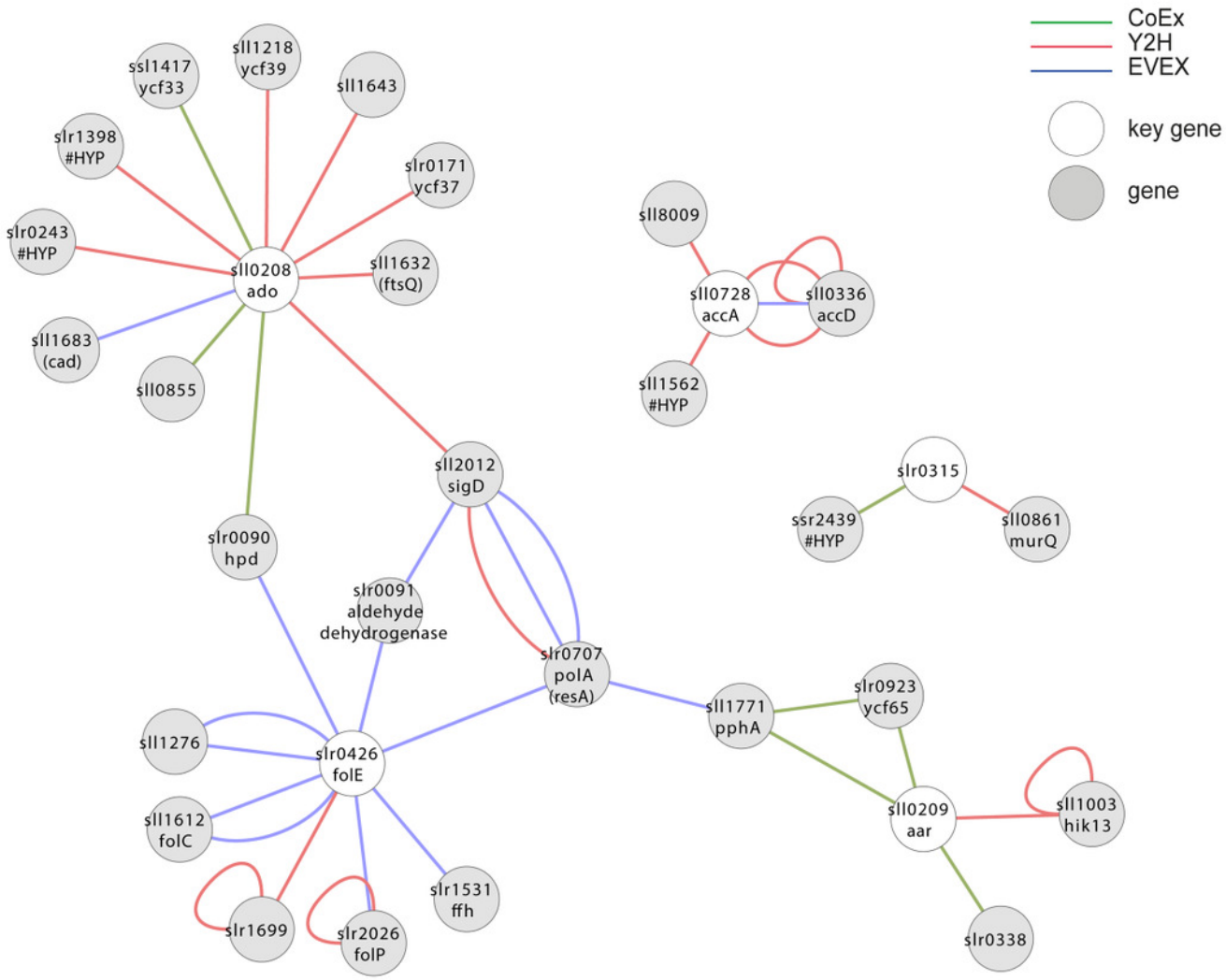
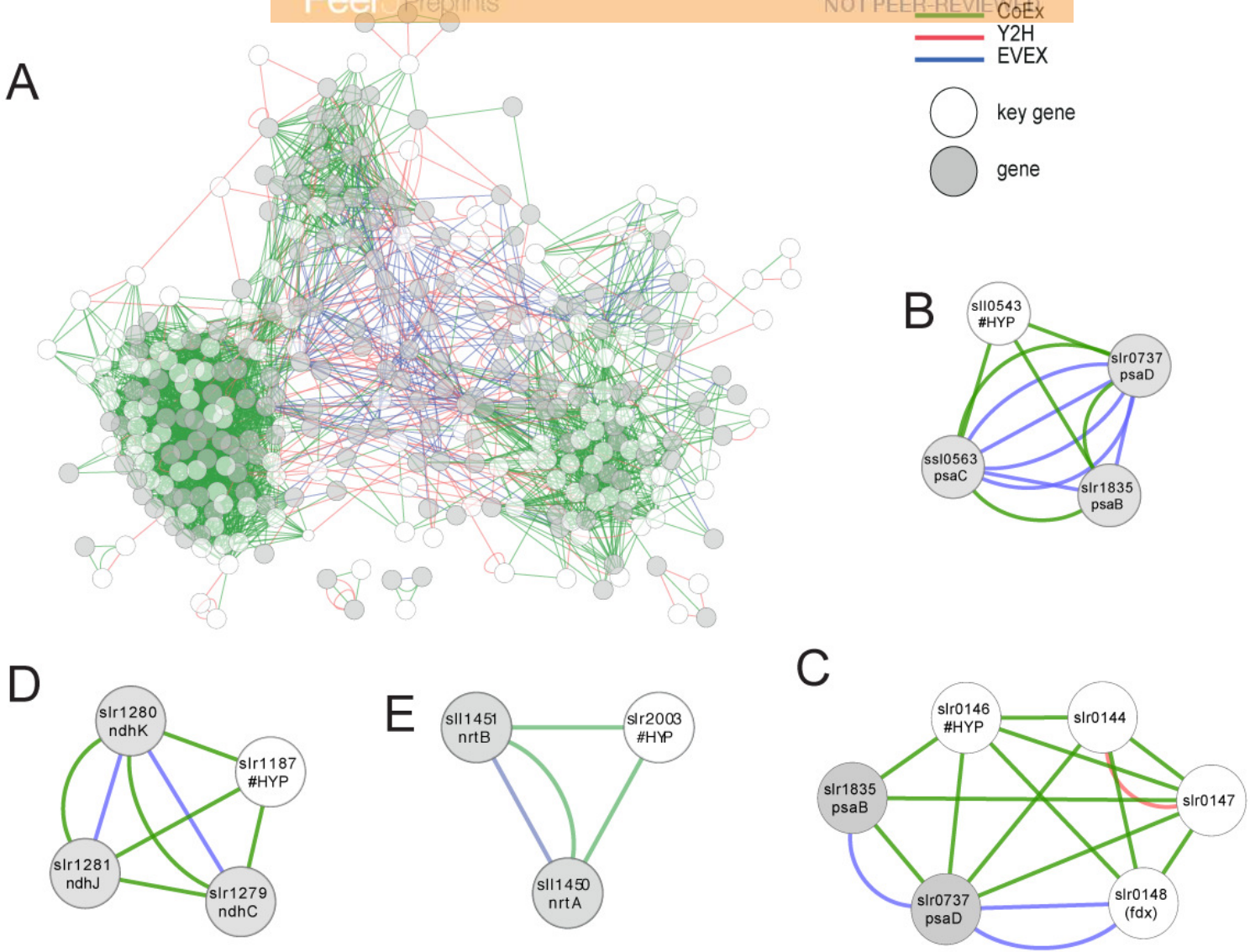


Figure 10 (on next page)

Cluster analysis for the role of genes annotated as ‘hypothetical’ or ‘unknown’

(A) The combined network of motifs extracted with the rule-based script. (B) sll0543, as an example pattern with highest ranking, forms a cluster with genes encoding three key members of PSI (*psaC*, *psaB*, *psaD*). (C) slr0144-48 as another example (see Fig. 6D). (D) The ‘unknown protein’ slr1187 forms a cluster with three NADH dehydrogenase subunits (slr1279-81) (E) ‘hypothetical protein’ slr2003 forms a cluster with two nitrate/nitrite transport system components (slr1450-51).



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3472v1 | CC BY 4.0 Open Access | rec: 18 Dec 2017, publ: 18 Dec 2017

Table 1(on next page)

List of key genes (KGs) used in the case studies

KGs identified for alkane biosynthesis were based on the consensus operon structure in cyanobacteria (Klähn et al. 2014).

NADPH metabolism	**Gene name**	**Annotation (Cyanobase)**
slr1239	*pntA*	pyridine nucleotide transhydrogenase alpha subunit
slr1434	*pntB*	pyridine nucleotide transhydrogenase beta subunit
slr1843	*zwf*	glucose 6-phosphate dehydrogenase
slr1289	*icdh*	isocitrate dehydrogenase
slr1643	*fnr* (PetH)	ferredoxin-NADP oxidoreductase
ssl0020	*petF*	ferredoxin I
Iron sulfur cluster metabolism		
sll0088	*sufR*	hypothetical protein (transcriptional regulator, *suf*)
slr0074	*sufB*	ABC transporter subunit
slr0075	*sufC*	ABC transporter ATP-binding protein
slr0076	*sufD*	hypothetical protein (FeS assembly protein)
slr0077	*sufS/nifS*	cysteine desulfurase
slr1417	*sufA*	hypothetical protein YCF57 (FeS assembly protein)
Alkane biosynthesis		
sll0209	*aar*	acyl-ACP reductase
sll0208	*ado*	aldehyde deformylating oxygenase
sll0207	*rfbA*	glucose-1-phosphate thymidylyltransferase
sll0728	*accA*	Acetyl-CoA carboxylase alpha subunit
slr0315		probable oxidoreductase
slr0426	*folE*	GTP cyclohydrolase I
Sigma factor		
Sll1689	*sigE*	group2 RNA polymerase sigma factor SigE