D758–D764 *Nucleic Acids Research, 2022, Vol. 50, Database issue* *Published online 6 October 2021*
https://doi.org/10.1093/nar/gkab891

CyanoOmicsDB: an integrated omics database for functional genomic analysis of cyanobacteria

Peng Zhou[1,†], Li Wang[1,†], Hai Liu[1], Chunyan Li[1], Zhimin Li[1,2], Jinxiang Wang[1] and Xiaoming Tan[1,*]

[1]State Key Laboratory of Biocatalysis and Enzyme Engineering, Environmental Microbial Technology Center of Hubei Province, and School of Life Sciences, Hubei University, Wuhan 430062, China and [2]College of Bioscience and Bioengineering, Jiangxi Agricultural University, Nanchang 330045, China

Received August 10, 2021; Revised September 17, 2021; Editorial Decision September 17, 2021; Accepted September 20, 2021

ABSTRACT

With their photosynthetic ability and established genetic modification systems, cyanobacteria are essential for fundamental and biotechnological research. Till now, hundreds of cyanobacterial genomes have been sequenced, and transcriptomic analysis has been frequently applied in the functional genomics of cyanobacteria. However, the massive omics data have not been extensively mined and integrated. Here, we describe CyanoOmicsDB (http://www.cyanoomics.cn/), a database aiming to provide comprehensive functional information for each cyanobacterial gene. CyanoOmicsDB consists of 8 335 261 entries of cyanobacterial genes from 928 genomes. It provides multiple gene identifiers, visualized genomic location, and DNA sequences for each gene entry. For protein-encoding genes, CyanoOmicsDB can provide predicted gene function, amino acid sequences, homologs, protein-domain super-families, and accession numbers for various public protein function databases. CyanoOmicsDB integrates both transcriptional and translational profiles of *Synechocystis* sp. PCC 6803 under various environmental culture coditions and genetic backgrounds. Moreover, CyanoOmicsDB includes 23 689 gene transcriptional start sites, 94 644 identified peptides, and 16 778 post-translation modification sites obtained from transcriptomes or proteomes of several model cyanobacteria. Compared with other existing cyanobacterial databases, CyanoOmicsDB comprises more datasets and more comprehensive functional information. CyanoOmicsDB will provide researchers in this field with a convenient way to retrieve functional information on cyanobacterial genes.

INTRODUCTION

Cyanobacteria are the only prokaryotes that can perform oxygen-evolving photosynthesis (1). Cyanobacteria have been emerging as popular model organisms for fundamental and biotechnological research because they are amenable to genetic engineering and possess a relatively fast growth rate and good tolerance to environmental stresses (2–4).

In 1997, *Synechocystis* sp. PCC 6803 became the first cyanobacterium whose genome was sequenced entirely (5). Since then, the number of sequenced cyanobacterial genomes increased rapidly, especially when high-throughput next-generation sequencing (NGS) became a reliable and routine technique. Although hundreds of cyanobacterial genomes were sequenced within the last two decades, only limited cyanobacterial genes were functionally characterized in a few model cyanobacteria.

Microarray or RNA-sequencing (RNA-seq) based transcriptomic analysis, a powerful tool for linking genes with their functions, was also normally used to identify the differentially expressed genes in cyanobacteria under different environmental conditions. In addition, transcriptional start sites were systematically identified in multiple cyanobacterial species using primary transcriptome analysis (6–9). These transcriptional data are useful in demonstrating the biological functions of genes.

A comprehensive online database including genomic, transcriptomic, and reference information of cyanobacteria is essential for researchers in this field to *in silico* analysis of gene functions before conducting laboratory experiments. CyanoBase (http://genome.microbedb.jp/cyanobase) was first established as a genome database for *Synechocystis* sp. PCC 6803 in 1998 and has been updated several times in the last two decades (10–13). Currently, CyanoBase comprises 86 complete and 290 draft genomes and has become the most popular cyanobacterial database in this field. However, CyanoBase does not contain any transcriptomic data. CyanoEXpress

[*]To whom correspondence should be addressed. Tel: +86 18986280641; Email: xiaoming.tan@hubu.edu.cn
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

(http://cyanoexpress.sysbiolab.eu/) currently comprises visualized expression data of 3 078 genes from *Synechocystis* sp. PCC 6803 in response to 178 environmental and genetic perturbations (14). CyanOmics (https://lag.ihb.ac.cn/index.html) comprises a few omics datasets for *Synechococcus* sp. PCC 7002 (15). In addition, CyanoClust (http://cyanoclust.c.u-tokyo.ac.jp/) contains protein homology information for cyanobacteria and plastids (16), and CyanoLyase (http://cyanolyase.genouest.org/) contains the sequences and motifs of phycobilin lyases and related proteins from cyanobacteria, red algae, and cryptophytes (17).

Herein, we describe CyanoOmicsDB, an integrated web database containing genomic data of 928 cyanobacterial strains, 56 independent transcriptomic datasets, 3 primary transcriptomic datasets, and 15 proteomic datasets, which is currently the most comprehensive omics database for cyanobacteria.

MATERIALS AND METHODS

Data retrieval

To comprehensively investigate the genomic sequences of cyanobacteria, we downloaded all available cyanobacterial genomic sequences and their annotation from the NCBI assembly database (https://www.ncbi.nlm.nih.gov/assembly/) using the NCBI-datasets tool. The amino acid sequences of each protein-coding gene were used as queries to search against the Integrative Protein Signature Database (InterPro) using InterProScan 5.45–80.0 (18). Gene Expression Omnibus (GEO) Series Matrix files containing gene expression profiles were downloaded from the GEO database (19) using the GEOquery package (20), whereas the raw reads of RNA-seq were download from the Sequence Read Archive (SRA) database (21) using the SRA-toolkit (https://github.com/ncbi/sra-tools) (Figure 1).

Genomic data aggregation and processing

The metainformation of these genomes was extracted using the NCBI data format tool and formatted as a tab-separated values (TSV) file. The basic information for each gene, including locus_tag, gene symbol, old_locus_tag, genomic location, protein id, and encoding product, was extracted from the general feature format (GFF) files. The Enzyme Commission (EC) (22), Gene Ontology (GO) (23,24), Protein Families Database (Pfam) (25), MetaCyc (26), and Kyoto Encyclopedia of Genes and Genomes (KEGG) (27,28) identifiers for each gene were extracted from the resulting InterProScan outputs. The data for each gene were aggregated as another TSV file (Figure 1 and Supplementary Dataset S1).

Local blast for searching homologs

All amino acid sequences were retrieved from the raw fasta files (FAA) and combined into a single fasta file. The name of each amino acid sequence was formatted as its locus_tag. The local Basic Local Alignment Search Tool (BLAST) database was established from the resulting formatted fasta file using the makeblastdb command (29). For searching homologs of a gene, the resultant database was searched against using the BlastP command with the amino acid sequences of the gene as a query and with '-qcov_hsp_perc 70 and -evalue 1e-5' as a set of parameters. The locus_tag of each hit was used to recover protein id, product, and species name from the metadata of all cyanobacterial genes. The results were combined with the original BlastP output and shown as a new table in the HOMOLOGS module.

Retrieval of nucleotide and amino acid sequences on request

The nucleotide sequences were obtained from the fasta files (FNA) containing genomic sequences according to chromosome accession number and the start and end positions of the recovered gene. The amino acid sequences were recovered from the above-formatted amino acid fasta file according to the locus_tag of the recovered gene.

Visualization of genomic data using JBrowse

JBrowse (30) was installed and set up according to its official documentation. The genomic sequences and annotations were formatted as reference sequences and feature tracks, respectively, using the perl scripts provided by JBrowse (Figure 1).

Transcriptomic data aggregation and processing

Raw transcriptomic data of *Synechocystis* sp. PCC 6803 was from either GEO or SRA database. Culture conditions and experimental groups were extracted from GPL files in the GEO database or description information in the SRA database. GEO microarray data series was downloaded using the GEOquery (20) package and analyzed using the Limma package (31) in R language. For RNA-seq data, raw reads were downloaded from the SRA database, aligned to reference genomes (GCA_000009725.1 and GCF_000009725.1) of *Synechoccystis* sp. PCC 6803 using the Bowtie2 (32). Raw read counts of genes were computed using the HTSeq-count program (33). The summarized count matrix was analyzed using the DESeq2 (34) package. And the output Log$_2$Foldchange and adjusted p-value of different comparisons were aggregated as a TSV file (Supplementary Dataset S2).

Proteomic data aggregation and processing

The identified peptides and the post-translational modifications (PTMs) of cyanobacterial proteins were obtained from the reported publications on cyanobacterial proteomics (35–39) and further integrated into the gene information collection. Furthermore, genomic positions of nucleotide sequences coding for these peptides were confirmed by mapping their arrangements to the corresponding reference genomes and generating JBrowse tracks. Differential expression profiles of cyanobacterial proteins were also obtained from reported publications (40–47) and integrated into the differential expression collection.

Reference data aggregation and processing

For the recent publication page, the reference information containing title, authors, journal, digital object identifiers (DOIs), and abstract was retrieved from PubMed using

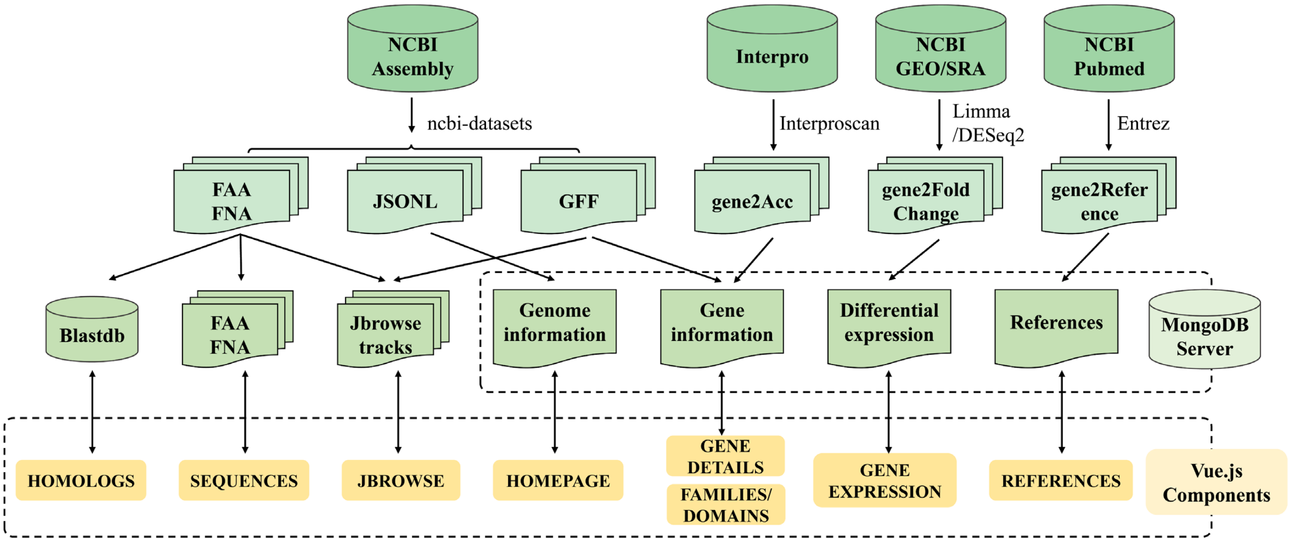

Figure 1. Overview of the process of creating CyanoOmicsDB. Raw data were recovered from public databases and then assessed to establish associations between cyanobacterial genes and their nucleotide sequences, amino acid sequences, annotations, accession numbers to various databases (gene2Acc), differential expression profiles (gene2FoldChange), and references (gene2Reference). Except for the gene-sequence association, all other associations were formatted and integrated as four collections in the back-end MongoDB database. Amino acid sequences were used to create a BLAST database (Blastdb), and the nucleotide sequences and their annotations were used in configuring JBrowse tracks on the CyanoOmisDB server. Vue.js was used as the front-end application framework of CyanoOmicsDB. Vue components (indicated by upper cases) were created to retrieve and display detailed information of the retrieved gene or genome from the back-end.

Entrez package in Python language script with 'cyanobacteria' or 'cyanobacterium' as keywords. The recent 200 publications were integrated into a TSV file. For the reference page of each gene, the information was recovered using the same script with the keywords containing both the species name and the locus_tag or gene symbol.

Web app implementation

The associations between cyanobacterial genes with their functional information were formatted and integrated as four collections in the back-end MongoDB database (Figure 1). A popular Quasar framework was used for building concise user interfaces of CyanoOmicsDB. In addition, self-constructed search engines and sortable, filterable, and paginated data tables were created for displaying information from each dataset or search result.

RESULTS

Database content

At the time of writing (7 April 2021), CyanoOmicsDB contained 8 335 261 entries of unique cyanobacterial genes from 186 complete and 742 draft cyanobacterial genomes (Supplementary Table S1). We developed a pipeline to download cyanobacterial genome datasets from the NCBI assembly database and extract, format, and import information for each gene into the CyanoOmicsDB. For each gene, CyanoOmicsDB provides basic information, gene annotations and six functional analysis modules containing JBrowse (for genome visualization), sequence, homologs, families/domains, differential gene expression, and references (Figure 1).

Basic information

For most genomes, CyanoOmicsDB provides both GenBank (GCA) (48) and RefSeq (GCF) (49) genome assemblies separately. Locus_tags, gene symbols and old_locus_tags were extracted from all annotated GFF files of GenBank or RefSeq assemblies and set as retrievable gene identifiers in our database. Further, the lengths, genomic location, and gene-coding type are also provided in the basic information module.

Gene annotations

Based on the annotated GFF files and the InterProScan output results, we extracted and collected the accession numbers for each gene from multiple widely used annotation databases, including EC,GO, Pfam, MetaCyc and KEGG. These accession numbers depicted on the CyanoOmicsDB are linked to the corresponding databases for more detailed information. Additionally, transmembrane domains of each protein-coding gene were identified using TMHMM 2.0 (50) and used as a criterion to determine whether the protein is a membrane protein or a soluble protein. The InterProScan output result for each protein-coding gene is indicated in the module tab named 'FAMILIES/DOMAINS' in which the conserved domains and homologous super-families are depicted graphically.

Identified peptides and post-translational modifications

To provide useful information for expressed proteins, we obtained identified peptides and their PTMs from some published proteome data of *Synechocystis* sp. PCC 6803 (35,36,38), *Synechococcus* sp. PCC 7002 (37,39), or *Synechococcus* sp. WH 8102 (51) (Supplementary Table S2). In total, 94 644 unique peptides and 16 778 PTM sites from 6 967 cyanobacterial proteins (Supplementary Dataset S3) were integrated into CyanoOmicsDB. For each gene, peptides identified in the same publication are indicated as separated lines in CyanoOmicsDB. Each PTM is indicated as a word combining its amino acid position, the modified amino acid residue, and the abbreviated modification type. Full names of PTM modification types were indicated in mouseover texts. By clicking on peptides or PTMs, users will be directed to the source publications in which these peptides or PTMs were experimentally identified.

Genome visualization using JBrowse

In the JBROWSE module, the retrieved gene's genomic region is displayed on the corresponding reference genome track with the recovered gene highlighted in yellow background. Users can freely view the adjacent genome areas in the same genome track by dragging or zooming in or out. The features of genes encoded on different strands are exhibited in different colors. By left-clicking on the features, users can access detailed web pages of neighboring genes. The 23 689 transcriptional start sites or transcriptional units of *Synechocystis* sp. PCC 6803, *Nostoc* sp. PCC 7120, and *Synechococcus elongatus* UTEX 2973, which were systematically identified by primary transcriptome analysis (7,9,52), are displayed as independent tracks named 'TSS' or 'Transcript_unit'. Also, 94 644 peptides and 16 778 PTM sites identified from the proteomes of *Synechocystis* sp. PCC 6803, *Synechococcus* sp. PCC 7002, and *Synechococcus* sp. WH 8102 are shown as independent tracks named as 'Peptides' and 'PTMs', respectively.

Sequence retrieval and analysis

By default, CyanoOmicsDB exhibits both nucleotide and amino acid sequences of the retrieved genes in the SEQUENCE module tab. Users can freely set the start and end positions of a gene to update nucleotide sequences. CyanoOmicsDB also provides links to further sequence analysis in the SEQUENCE module, including KEGG, InterProScan, local and online BlastP, and online BlastN and BlastX. Notably, CyanoOmicsDB can conduct local BlastP against a local amino acid database containing only amino acid sequences of cyanobacterial proteins, with a default parameter described in the Materials and Methods section. Therefore, if users want to set up custom parameters, including search database and expect threshold, they should choose online Blast interfaces.

Gene homologs

A local BlastP can be automatically performed using the default parameters when loading the HOMOLOGS module to show homologous genes of the retrieved gene. The results of BlastP will be outputted in the tab-separated format and parsed as a table containing the locus_tag, protein id, encoding products and species names of hits. Additionally, the percentage identity of identical residues between target and query sequences (identity), query coverage and *E*-value will be included in the same table. By clicking on the locus_tag of each hit, users will be directed to the detailed webpage of the hit.

References

The academic literature was recovered from PubMed using both gene identifiers and species names as keywords. Only the literature containing the recovered species name in its title/abstract and gene identifiers in the main texts was linked to specific gene entry. Basic information, including titles, authors, journals, PMIDs and DOIs of the related literature collected for each gene, will be indicated in the REFERENCES module tab, if any. The texts containing keywords can be extracted and displayed after the basic literature information. The detailed abstract and full texts can be accessed by clicking on either titles or DOIs.

Gene expression

To investigate gene function, gene expression profiles under different environmental conditions or genetic backgrounds are informative. Raw transcriptomic data of *Synechocystis* sp. PCC 6803 deposited in 40 GEO datasets and 16 SRA transcriptome studies were collected and reanalyzed using the Limma and DESeq2 packages, respectively. The differential expressions of proteins were directly collected from publications on cyanobacterial proteomes. In total, CyanoOmicsDB contains 203 pairwise transcriptome comparisons and 25 proteome comparisons among different culturing conditions and genetic backgrounds (Supplementary Table S3). Transcriptional or translational changes of each gene in various comparisons, the corresponding conditions, and GEO/SRP/PMID accession numbers were combined and displayed in the GENE EXPRESSION module. By clicking on these accession numbers, users will be directed to the GEO Accession viewer, the SRA Run Selector, or the published literature for more detailed descriptions of the comparisons.

CyanoIdMapping tool

For conversion of gene identifiers from different databases, CyanoOmicsDB provides the online CyanoIdMapping tool (http://www.cyanoomics.cn/lz/id-mapping). Users can convert RefSeq gene identifiers of cyanobacteria to the corresponding GenBank identifiers or vice versa.

Search engines

To quickly and accurately retrieve data, CyanoOmicsDB provides multiple ways for users to search data. First, CyanoOmicsDB provides a search bar on the top panel of each webpage. Using this bar, users can search any fields in either species or gene datasheets. Second, CyanoOmicsDB provides filters in any listing pages, using which users can conveniently narrow down the search results.

DISCUSSION

Comparison of CyanoOmicsDB with other similar databases

So far, there are several reported databases of cyanobacteria with different contents, for example, CyanoBase, CyanoExpress, CyanOmics, CyanoClust and CyanoLyase. A detailed comparison of CyanoOmicsDB with these databases is shown in the Supplementary Table S4. Compared with CyanoLyase and CyanoClust that focus on a limited number of genes, CyanoOmisDB collects all genes for each included genome. Compared with CyanOmics that collects several transcriptome and proteome datasets for *Synechococcus* sp. PCC 7002, CyanoOmicsDB integrates more omics datasets for sequenced cyanobacterial strains. Compared with CyanoEXpress that collects the most transcriptomic datasets of *Synechocystis* sp. PCC 6803 so far and only provides gene transcription information, CyanoOmicsDB integrates genomic, transcriptomic, and proteomic data together and provides comprehensive functional information for each gene of *Synechocystis* sp. PCC 6803, including gene annotation, homologs, gene expression, references and so on.

Undoubtedly, CyanoBase represents one of the most comprehensive databases for cyanobacteria until now and has been widely used in academia in the past two decades. Compared with CyanoBase, CyanoOmicsDB includes more cyanobacterial gene entries and provides more diverse interfaces to other gene function databases for each gene. Besides genomic data, both transcriptomic and proteomic data deposited in public databases were mined and incorporated into CyanoOmicsDB, which will give valuable clues for inferring gene functions.

Different gene identifiers from the GenBank and RefSeq genome assemblies

The GenBank and RefSeq genome assembly records for all cyanobacteria, except seven strains that were sequenced recently, are included in CyanoOmicsDB (Supplementary Table S1). Normally, RefSeq assemblies have the same nucleotide sequences as corresponding GenBank assemblies. Because of the reannotation by NCBI, gene annotation in the RefSeq geneset is well maintained and not always the same as that in the GenBank genesets. Especially, gene identifiers in the RefSeq genesets are entirely different from those in the GenBank genesets. And it is confusing that different gene identifiers were used to represent the same cyanobacterial gene in academic literature and databases. For example, CyanoBase and CyanoExpress use gene identifiers from GenBank assemblies, but CyanOmics uses those from RefSeq assemblies. Thus, associations between different gene identifiers and the id-mapping tool will be helpful for researchers in this field.

A gene can be retrieved in CyanoOmicsDB using locus_tag, gene symbol or old_locus_tag, although locus_tags are used as the primary gene identifiers in CyanoOmicsDB. Further, locus_tags in the GenBank genesets are linked to old_locus_tags in the RefSeq genesets for genes that are annotated by both genome assemblies. Alternatively, gene identifiers can be mutually converted in batches using the CyanoIdMapping tool.

Data mining of cyanobacterial transcriptomic data

The gene expression profiles can provide valuable information on the biological functions of genes. There are massive amounts of transcriptomic data accumulated in the public databases, and this amount is still increasing. However, they are obtained from different experiments and platforms. Take the microarray data analyzed in this work as an example, these data are generated from several different microarray types, which contain different numbers of probes and chip designs. Furthermore, both the culture condition and the genetic background of each sample are difficult to directly extract from the metadata and need to be manually checked. Therefore, it is time-consuming to deal with these public transcriptomic data. Until now, CyanoOmicsDB integrates 56 transcriptomic datasets of *Synechocystis* sp. PCC 6803, which encompass almost all the transcriptomic data of this species. Data mining of other cyanobacterial transcriptomic data is still ongoing, and the results will be incorporated into CyanoOmicsDB in the future.

In conclusion, CyanoOmicsDB provides a convenient and alternative means to retrieve and analyze gene functions of cyanobacteria and will be helpful to the research community.

DATA AVAILABILITY

All raw genomic data can be found at NCBI assembly database using the accession numbers listed in Supplementary Table S1. All raw reads and microarray data series, whose accession numbers are listed in Supplementary Table S3, can be downloaded from SRA and GEO databases, respectively. Processed datasets (Supplemental Dataset 1–3) supporting both this article and CyanoOmicsDB are available in the Figshare repository (https://figshare.com/s/c3729a3e623c6f003fc0). CyanoOmicsDB is freely available at http://www.cyanoomics.cn/.

All genomic and transcriptomic data were from the public NCBI assembly, GEO, and SRA databases. Accession numbers are listed in Supplementary Table S1 and S3.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Tao Zhu, Dr He Zhang, and Ms. Huili Sun for their valuable suggestions. We are also grateful to Dr Zhuo Chen for his help in collecting proteomic data.

FUNDING

National Natural Science Foundation of China [31871303 to X.T.]; Open Funding Project of the State Key Laboratory of Biocatalysis and Enzyme Engineering [SKLBEE2019015 to Z.L.]; Key Laboratory of Biofuel, Chinese Academy of Sciences [CASKLB201802 to X.T.]; State Key Laboratory of Freshwater Ecology and Biotechnology [2018FB09 to X.T.]. Funding for open access charge: Open Funding Project of the State Key Laboratory of Biocatalysis and Enzyme Engineering [SKLBEE2019015].

Conflict of interest statement. None declared.

REFERENCES

1. Kirsch,F., Klahn,S. and Hagemann,M. (2019) Salt-regulated accumulation of the compatible solutes sucrose and glucosylglycerol in cyanobacteria and its biotechnological potential. *Front. Microbiol.*, **10**, 2139.
2. Hitchcock,A., Hunter,C.N. and Canniffe,D.P. (2020) Progress and challenges in engineering cyanobacteria as chassis for light-driven biotechnology. *Microb. Biotechnol.*, **13**, 363–367.
3. Hagemann,M. and Hess,W.R. (2017) Systems and synthetic biology for the biotechnological application of cyanobacteria. *Curr. Opin. Biotechnol.*, **49**, 94–99.
4. Savakis,P. and Hellingwerf,K.J. (2015) Engineering cyanobacteria for direct biofuel production from CO_2. *Curr. Opin. Biotechnol.*, **33**, 8–14.
5. Kaneko,T. and Tabata,S. (1997) Complete genome structure of the unicellular cyanobacterium *Synechocystis* sp. PCC6803. *Plant Cell Physiol.*, **38**, 1171–1176.
6. Mitschke,J., Georg,J., Scholz,I., Sharma,C.M., Dienst,D., Bantscheff,J., Voss,B., Steglich,C., Wilde,A., Vogel,J. *et al.* (2011) An experimentally anchored map of transcriptional start sites in the model cyanobacterium *Synechocystis* sp. PCC6803. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 2124–2129.
7. Kopf,M., Klahn,S., Scholz,I., Matthiessen,J.K., Hess,W.R. and Voss,B. (2014) Comparative analysis of the primary transcriptome of *Synechocystis* sp. PCC 6803. *DNA Res.*, **21**, 527–539.
8. Pfreundt,U., Kopf,M., Belkin,N., Berman-Frank,I. and Hess,W.R. (2014) The primary transcriptome of the marine diazotroph *Trichodesmium erythraeum* IMS101. *Sci. Rep.*, **4**, 6187.
9. Tan,X., Hou,S., Song,K., Georg,J., Klähn,S., Lu,X. and Hess,W.R. (2018) The primary transcriptome of the fast-growing cyanobacterium *Synechococcus elongatus* UTEX 2973. *Biotechnol. Biofuels*, **11**, 218.
10. Nakamura,Y., Kaneko,T., Hirosawa,M., Miyajima,N. and Tabata,S. (1998) CyanoBase, a www database containing the complete nucleotide sequence of the genome of *Synechocystis* sp. strain PCC6803. *Nucleic. Acids. Res.*, **26**, 63–67.
11. Nakamura,Y., Kaneko,T. and Tabata,S. (2000) CyanoBase, the genome database for *Synechocystis* sp. strain PCC6803: status for the year 2000. *Nucleic. Acids. Res.*, **28**, 72.
12. Nakao,M., Okamoto,S., Kohara,M., Fujishiro,T., Fujisawa,T., Sato,S., Tabata,S., Kaneko,T. and Nakamura,Y. (2010) CyanoBase: the cyanobacteria genome database update 2010. *Nucleic. Acids. Res.*, **38**, D379–D381.
13. Fujisawa,T., Narikawa,R., Maeda,S.I., Watanabe,S., Kanesaki,Y., Kobayashi,K., Nomata,J., Hanaoka,M., Watanabe,M., Ehira,S. *et al.* (2017) CyanoBase: a large-scale update on its 20th anniversary. *Nucleic Acids Res.*, **45**, D551–D554.
14. Hernandez-Prieto,M.A. and Futschik,M.E. (2012) CyanoEXpress: a web database for exploration and visualisation of the integrated transcriptome of cyanobacterium *Synechocystis* sp. PCC6803. *Bioinformation*, **8**, 634–638.
15. Yang,Y., Feng,J., Li,T., Ge,F. and Zhao,J. (2015) CyanOmics: an integrated database of omics for the model cyanobacterium *Synechococcus* sp. PCC 7002. *Database*, **2015**, bau127.
16. Sasaki,N.V. and Sato,N. (2010) CyanoClust: comparative genome resources of cyanobacteria and plastids. *Database (Oxford)*, **2010**, bap025.
17. Bretaudeau,A., Coste,F., Humily,F., Garczarek,L., Le Corguille,G., Six,C., Ratin,M., Collin,O., Schluchter,W.M. and Partensky,F. (2013) CyanoLyase: a database of phycobilin lyase sequences, motifs and functions. *Nucleic Acids Res.*, **41**, D396–D401.
18. Jones,P., Binns,D., Chang,H.Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
19. Clough,E. and Barrett,T. (2016) The gene expression omnibus database. *Methods Mol. Biol.*, **1418**, 93–110.
20. Davis,S. and Meltzer,P.S. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, **23**, 1846–1847.
21. Kodama,Y., Shumway,M., Leinonen,R. and International Nucleotide Sequence Database, C. (2012) The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
22. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.
23. Gene Ontology, C. (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
24. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
25. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J. *et al.* (2021) Pfam: The protein families database in 2021. *Nucleic. Acids. Res.*, **49**, D412–D419.
26. Caspi,R., Billington,R., Keseler,I.M., Kothari,A., Krummenacker,M., Midford,P.E., Ong,W.K., Paley,S., Subhraveti,P. and Karp,P.D. (2020) The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res.*, **48**, D445–D453.
27. Kanehisa,M., Furumichi,M., Sato,Y., Ishiguro-Watanabe,M. and Tanabe,M. (2021) KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.*, **49**, D545–D551.
28. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acid. Res.*, **28**, 27–30.
29. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
30. Buels,R., Yao,E., Diesh,C.M., Hayes,R.D., Munoz-Torres,M., Helt,G., Goodstein,D.M., Elsik,C.G., Lewis,S.E., Stein,L. *et al.* (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, **17**, 66.
31. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
32. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
33. Anders,S., Pyl,P.T. and Huber,W. (2015) HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.
34. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
35. Spät,P., Klotz,A., Rexroth,S., Maček,B. and Forchhammer,K. (2018) Chlorosis as a developmental program in cyanobacteria: the proteomic fundament for survival and awakening. *Mol. Cell. Proteomics*, **17**, 1650–1669.
36. Spat,P., Macek,B. and Forchhammer,K. (2015) Phosphoproteome of the cyanobacterium *Synechocystis* sp. PCC 6803 and its dynamics during nitrogen starvation. *Front. Microbiol.*, **6**, 248.
37. Yang,M.K., Qiao,Z.X., Zhang,W.Y., Xiong,Q., Zhang,J., Li,T., Ge,F. and Zhao,J.D. (2013) Global phosphoproteomic analysis reveals diverse functions of serine/threonine/tyrosine phosphorylation in the model cyanobacterium *Synechococcus* sp. strain PCC 7002. *J. Proteome Res.*, **12**, 1909–1923.
38. Ma,Y., Yang,M., Lin,X., Liu,X., Huang,H. and Ge,F. (2017) Malonylome analysis reveals the involvement of lysine malonylation in metabolism and photosynthesis in cyanobacteria. *J. Proteome Res.*, **16**, 2030–2043.
39. Chen,Z., Zhang,G., Yang,M., Li,T., Ge,F. and Zhao,J. (2017) Lysine acetylome analysis reveals photosystem II manganese-stabilizing protein acetylation is involved in negative regulation of oxygen evolution in model cyanobacterium *Synechococcus* sp. PCC 7002. *Mol. Cell. Proteomics*, **16**, 1297–1311.
40. Borirak,O., de Koning,L.J., van der Woude,A.D., Hoefsloot,H.C., Dekker,H.L., Roseboom,W., de Koster,C.G. and Hellingwerf,K.J. (2015) Quantitative proteomics analysis of an ethanol- and a lactate-producing mutant strain of *Synechocystis* sp. PCC6803. *Biotechnol. Biofuels*, **8**, 111.
41. Xiong,Q., Feng,J., Li,S.T., Zhang,G.Y., Qiao,Z.X., Chen,Z., Wu,Y., Lin,Y., Li,T., Ge,F. *et al.* (2015) Integrated transcriptomic and proteomic analysis of the global response of *Synechococcus* to high light stress. *Mol. Cell. Proteomics*, **14**, 1038–1053.
42. Wegener,K.M., Singh,A.K., Jacobs,J.M., Elvitigala,T., Welsh,E.A., Keren,N., Gritsenko,M.A., Ghosh,B.K., Camp,D.G. 2nd, Smith,R.D. *et al.* (2010) Global proteomics reveal an atypical strategy for carbon/nitrogen assimilation by a cyanobacterium under diverse environmental perturbations. *Mol. Cell. Proteomics*, **9**, 2678–2689.

43. Huang,S., Chen,L., Te,R., Qiao,J., Wang,J. and Zhang,W. (2013) Complementary iTRAQ proteomics and RNA-seq transcriptomics reveal multiple levels of regulation in response to nitrogen starvation in *Synechocystis* sp. PCC 6803. *Mol. Biosyst.*, **9**, 2565–2574.
44. Qiao,J., Huang,S., Te,R., Wang,J., Chen,L. and Zhang,W. (2013) Integrated proteomic and transcriptomic analysis reveals novel genes and regulatory mechanisms involved in salt stress responses in *Synechocystis* sp. PCC 6803. *Appl. Microbiol. Biotechnol.*, **97**, 8253–8264.
45. Liu,J., Chen,L., Wang,J., Qiao,J. and Zhang,W. (2012) Proteomic analysis reveals resistance mechanism against biofuel hexane in *Synechocystis* sp. PCC 6803. *Biotechnol. Biofuels*, **5**, 68.
46. Tian,X., Chen,L., Wang,J., Qiao,J. and Zhang,W. (2013) Quantitative proteomics reveals dynamic responses of *Synechocystis* sp. PCC 6803 to next-generation biofuel butanol. *J. Proteomics*, **78**, 326–345.
47. Qiao,J., Wang,J., Chen,L., Tian,X., Huang,S., Ren,X. and Zhang,W. (2012) Quantitative iTRAQ LC-MS/MS proteomics reveals metabolic responses to biofuel ethanol in cyanobacterial *Synechocystis* sp. PCC 6803. *J. Proteome Res.*, **11**, 5286–5300.
48. Sayers,E.W., Cavanaugh,M., Clark,K., Ostell,J., Pruitt,K.D. and Karsch-Mizrachi,I. (2019) GenBank. *Nucleic Acids Res.*, **48**, D84–D86.
49. 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. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
50. Sonnhammer,E.L., von Heijne,G. and Krogh,A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 175–182.
51. Li,Y.Y., Chen,X.H., Xue,C., Zhang,H., Sun,G., Xie,Z.X., Lin,L. and Wang,D.Z. (2019) Proteomic response to rising temperature in the marine cyanobacterium *Synechococcus* grown in different nitrogen sources. *Front. Microbiol.*, **10**, 1976.
52. Mitschke,J., Vioque,A., Haas,F., Hess,W.R. and Muro-Pastor,A.M. (2011) Dynamics of transcriptional start site selection during nitrogen stress-induced cell differentiation in *Anabaena* sp. PCC7120. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 20130–20135.