Easy-to-use whole-genome sequencing workflows and standardized practices to uncover hidden genetic variation in *Synechocystis* PCC 6803 wild-type and knock-out strains

Marius Theune[1], Ruben Fritsche[¶1], Niklas Küppers[¶2] , Marko Böhm[1], Petra Kolkhof[2], Florian Paul[1], Ovidiu Popa[2,4], Ellen Oldenburg[3,4,5], Anika Wiegard[2], Ilka M. Axmann[2], Kirstin Gutekunst[1*]

¶ Contributed equally

[1]Molecular Plant Physiology, Bioenergetics in Photoautotrophs, Institute for Biology, University of Kassel, D-34132 Kassel, Germany

[2]Institute for Synthetic Microbiology, Biology Department, Heinrich Heine University Düsseldorf, D-40225, Düsseldorf, Germany.

[3]Alfred Wegener Institute Helmholtz Center for Polar and Marine Research, Bremerhaven, Germany

[4]Institute of Quantitative and Theoretical Biology, Biology Department, Heinrich Heine University Düsseldorf, D-40225, Düsseldorf, Germany.

[5]Cluster of Excellence on Plant Science, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

* Corresponding author
E-Mail: kirstin.gutekunst@uni-kassel.de (KT)

Abstract

Knock-out mutants are often used to study gene function by disrupting a specific gene and comparing the mutant to a wild-type strain. Reliable interpretation, however, requires that the two strains differ only by the intended mutation and that the observed phenotype is caused specifically by the deleted gene. In the highly polyploid cyanobacterium *Synechocystis* sp. PCC 6803, this is particularly challenging because incomplete segregation can mask genetic heterogeneity or secondary suppressor mutations. The genetic variation among laboratory wild-type lines can further confound phenotypic analyses. We show that these challenges can be addressed by routine strain validation via whole-genome sequencing (WGS). To this end, we developed and tested user friendly workflows for short-read (Illumina), long-read (Oxford Nanopore Technologies; ONT), and hybrid data, providing standardized quality control, variant calling, and structural variant detection. We benchmarked their performance in detecting single-nucleotide polymorphisms (SNPs), small indels, and structural variants using simulated datasets across different coverages and mixed populations. Applying the workflows to three *Synechocystis* sp. PCC 6803 wild-type lines revealed multiple sequence and structural differences relative to the reference genome, including previously undescribed genetic variants, underscoring the importance of documenting the strain background and the value of long-read sequencing. Characterization of two independent 6-phosphogluconate dehydrogenase (*gnd*) knock-out mutants and their complemented strains highlighted how a failed rescue can reveal a phenotype unrelated to the intended knock-out. An automated literature analysis revealed that only a minority of the investigated *Synechocystis* studies that used knock-out mutants included complementation as a control (39%), whereas this practice is more common in studies involving *Escherichia coli* (63%) and *Saccharomyces cerevisiae* (55%). Based on these results, we propose a practical guide for standardizing knock-out phenotyping in *Synechocystis* PCC 6803. Combined with accessible workflows for routine whole-genome validation, this framework aims to support more robust and reproducible knock-out studies in the future.

Introduction

The unicellular cyanobacterium *Synechocystis* sp. PCC 6803 (hereafter *Synechocystis*) was one of the first organisms to be fully sequenced (1) and is widely used as a model organism. Its efficient integration of foreign DNA into the genome via homologous recombination and suitability for genetic engineering have made it a preferred model organism for basic research and green biotechnology applications (2–4). To investigate a gene's function, knock-out mutants are often employed, in which the gene of interest is disrupted or replaced, and the resulting strain is compared with a wild-type strain. A deviant phenotype is then used to elucidate a gene's function. While creating knock-out mutants is straightforward, *Synechocystis* presents a significant challenge for such genetic studies: its polyploid genome. (5). Variable copy numbers across cells and growth conditions necessitate segregation periods after genetic manipulation (6,7). While polyploidy can be beneficial when studying non-segregated knock-out strains of essential genes, the slow segregation increases the risk of secondary mutations, particularly suppressor mutations, arising under selection pressure from the intended genomic change (8,9). Controlling for secondary mutations or identifying strain-specific ones is particularly important, given the well-documented genetic diversity even among *Synechocystis* wild-type strains (10–12). These variations are usually detected using next-generation sequencing techniques, such as Illumina sequencing, and by mapping the generated reads to a reference sequence. In this technique, genomic DNA is fragmented and bridge-amplified to create paired local-amplified sequences on a chip, which are then sequenced in parallel using a cyclic reversible-termination approach. It thereby sequences entire bacterial genomes with several hundred-fold coverage at a very low per-base-pair error rate, but is limited by the short read length of around 150 bp (13). Having only short reads poses major challenges for detecting larger structural variants or sequencing repetitive or non-unique genomic regions. Several techniques have been developed to overcome this limitation; a prominent one is Oxford Nanopore (ONT) sequencing (14). This technique fundamentally differs from polymerase-based approaches in that it sequences a long DNA strand by translocating it through a pore while measuring sequence-specific current shifts. This technique generates ultra-long sequencing reads, yielding high coverage; however, the per-base-pair error rate remains significantly higher (15). Since

89 both techniques are commercially available, bacterial WGS has become accessible
90 and affordable, even for non-specialized laboratories (16–18). However, filtering and
91 analyzing the generated data is labor-intensive, potentially preventing laboratories
92 from routinely checking their mutant or wild-type strains (19). An alternative approach
93 to determining whether a phenotype is directly linked to a gene knock-out rather than
94 to a secondary mutation is to generate a complemented strain by reintroducing the
95 previously knocked-out gene. These controls are performed by either introducing the
96 gene back to its original locus (complementation in a narrow sense) or to another
97 genomic location or to an extrachromosomal plasmid, after which it is investigated
98 whether the knock-out phenotype is rescued (20). A suitable strain for comparing
99 different complementation strategies is the previously published 6-phosphogluconate
100 dehydrogenase (gnd) knock-out mutant (21). As Gnd is a key enzyme of the oxidative
101 pentose phosphate pathway (OPP), it is essential for light-activated heterotrophic
102 growth (hereafter, heterotrophic growth) (22). It was recently reported that this strain
103 contains a secondary frameshift mutation at position 399 of the glucose-6-P.
104 dehydrogenase (zwf) gene, which was detected by targeted Sanger sequencing (23).
105 This frameshift introduces an early stop codon, thereby truncating the protein and
106 abolishing Zwf enzyme activity. Consequently, this strain was renamed Δgnd_zwf^{399+A}.
107 Zwf is another key enzyme of the OPP pathway, upstream of Gnd, and its knock-out
108 mutant is also unable to grow under heterotrophic conditions (24). We created an
109 independent *gnd* knock-out strain *(Δgnd_3)* in the same wild-type and complemented
110 it using different techniques. We compare the different complementation strategies and
111 show how a failed rescue in a complemented strain can reveal a phenotype unrelated
112 to the intended knock-out. To facilitate the use of WGS, we created and validated
113 publicly available, easy-to-use pipelines for genomic variant calling from paired
114 Illumina and ultra-long-read ONT sequencing data, and used them to discover novel
115 variants across different wild-type strains. Finally, we developed a guide that outlines
116 many key steps to make *Synechocystis* knock-out studies as robust and reproducible
117 as possible.
118

Results

Design of user-friendly, Galaxy-based pipelines for whole-genome sequencing (WGS) analysis of *Synechocystis* strains

To make WGS more accessible to non-specialized users, we created data analysis pipelines on a publicly available bioinformatics server (usegalaxy.eu, The Galaxy Community, 2024). Galaxy is free to use, provides sufficient computational power, includes a large set of preinstalled tools, and offers a graphical user interface that enables easy data management and processing. Using only tools available on the server, we created two pipelines that process either paired-end short next-generation sequencing reads (e.g., Illumina), long ONT sequencing reads, or a combination of both to detect mutations by mapping reads to the provided reference genome (S3 File). The pipelines handle all necessary steps, such as quality control, filtering, adapter trimming, mapping, and, finally, calling mutations and annotating them (Fig 1; more detailed overview in S1 Fig; detailed manual in S4 Text). Important settings for filtering sequencing reads and distinguishing true variants from false-positive calls are available in the Galaxy server's graphical user interface, where they can be easily adjusted, with recommended default values provided. To increase sensitivity, two different callers were used in parallel when possible, and their results were combined. The pipelines return the called mutations and structural variations separately, in both a ".vcf" (variant call format) file for further analysis and a ".csv" (comma-separated values) file. These .csv files can be imported into Microsoft Excel using the provided macro (S5 Text), which creates a pivot table showing the mutated position, its effect and the alternate fraction, representing how many reads contain the mutation (Tab 1). The effects of each mutation are automatically reported in the effect column (EFF) alongside the called variants in our pipeline and include the effect itself, its severity on the affected gene, the changes in the genome, the resulting protein and finally the affected gene. This analysis allows easy comparison of different strains to identify specific mutations, while preserving all quality in a separate table. The filtered and mapped reads are provided for manual inspection of called mutations or for controlling the segregation of intended changes, along with a detailed quality report that can be opened in a browser. Combining multiple sequencing datasets into collections on the Galaxy server allows effective parallel data analysis. The average runtime, excluding

wait time for serving capacity, was 55 minutes for Illumina data, 18 minutes for ONT data, and 48 minutes for combined data analysis, making our pipelines a valid tool for routine whole-genome analysis of knock-out and wild-type strains.

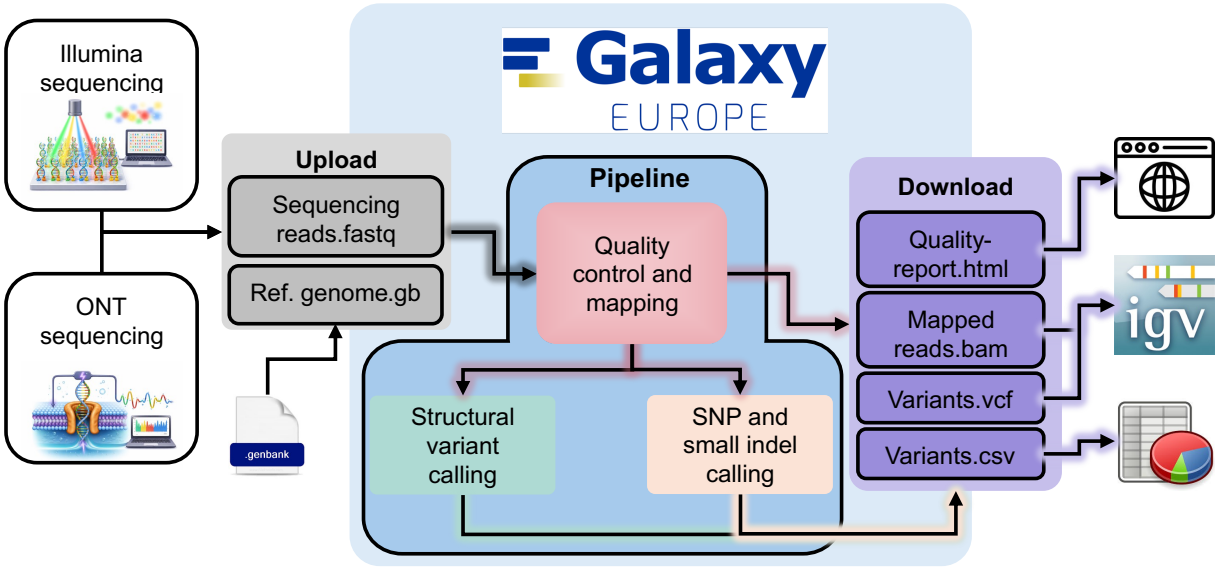

Fig 1: Schematic overview of an easy-to-use galaxy-based pipeline

Schematic overview of the workflows designed and validated here. Each pipeline can be used by simply uploading sequencing data in fastq or fastq.gz format, along with a reference genome, to the European Galaxy server. Starting the pipeline will automatically process the data by performing quality control and mapping, before using different callers for SNP and small indel or structural variant detection. Called variants are filtered and reported as .vcf and .csv files, which can be downloaded along with a detailed quality report and the mapped reads.

164 **Tab 1: Fixed genetic differences found by Illumina sequencing between wild-type**
165 **and Δgnd_zwf^{399+A}**

166 Mutations were called using Illumina sequencing and the corresponding pipeline. The resulting
167 .csv files were restructured using the provided Excel macro. Only fixed mutations (alternative
168 fraction >0.8) that differ between the two strains are shown. The presented numbers represent
169 the fraction of reads containing the called mutation. Effect collum (EFF) shows the effect on
170 genetic elements on the reference genome (based on SnpEff). Predicted effects are: Effect (
171 effect impact | functional class | codon change | amino acid change| gene name | transcript
172 BioType | gene coding | transcript ID | exon)

			SampleID		
CHROM	**POS**	**EFF**	**WT**	**Δgnd_zwf^{399+A}**	***Δgnd_3***
AP004311	8232	NON_SYNONYMOUS_CODING(MODERATE|MISSENSE|gaC/gaG|D620E|slr7010|protein_coding|CODING|BAD01912.1|1)		1	
BA000022	939030	NON_SYNONYMOUS_CODING(MODERATE|MISSENSE|cTg/cCg|L404P|ndhF3|protein_coding|CODING|sll1732|1)	1		1
BA000022	961599	FRAME_SHIFT(HIGH||tat/Atat|Y134I?|zwf|protein_coding|CODING|slr1843|1)		1	
BA000022	1415561	NON_SYNONYMOUS_CODING(MODERATE|MISSENSE|aCc/aTc|T479I|topA|protein_coding|CODING|slr2058|1)		1	
BA000022	1661788	SYNONYMOUS_CODING(LOW|SILENT|ccC/ccA|P166|NA|protein_coding|CODING|sll1232|1)		1	
BA000022	1727581	FRAME_SHIFT(HIGH||caa/caGAGCAACAa|Q103QSN?|NA|protein_coding|CODING|slr1339|1|INFO_REALIGN_3_PRIME)		0.803	
BA000022	2730516	INTERGENIC(MODIFIER||||||||)		1	
BA000022	3517416	NON_SYNONYMOUS_CODING(MODERATE|MISSENSE|tAt/tCt|Y283S|glgC|protein_coding|CODING|slr1176|1)	1		1
AP004311	8232	NON_SYNONYMOUS_CODING(MODERATE|MISSENSE|gaC/gaG|D620E|slr7010|protein_coding|CODING|BAD01912.1|1)		1	

173

174

Validation of short and ultra-long read sequencing for variant detection in simulated reads and different *Synechocystis* wild-type strains

We sequenced two non-motile (GT-T) and one motile *Synechocystis* wild-type (PCC-M) substrains previously obtained from labs in Stockholm, Sweden; Uppsala, Sweden; and Freiburg, Germany, using Illumina and ONT sequencing. We then created synthetic genomes containing different types of mutations: Single Nucleotide Polymorphism (SNPs), small insertions and deletions up to 10 bp in size (indels), larger insertions and deletions (up to 2000 bp in size) and complicated structural variants like relocations and duplications (for details see Methods section). Based on these mutated genomes and error profiles derived from the wild-type sequencing experiments, we simulated Illumina and ONT reads with varying coverages. Additionally, we simulated non-segregated variations by mixing simulated reads from the modified genomes with those from the reference genome. Because we used simulated reads, we knew all inserted mutations and used them to calculate the fraction of true positives among all called mutations (precision) and the fraction of true positives that were retrieved (recall or sensitivity) (S11 Figure). We combined by calculating the harmonic mean of the two, the F1-score, as a quality score for each technique (Fig 2). This analysis revealed that in simulated ONT sequencing data, our pipeline struggled with small mutations such as SNPs and indels, especially when the coverage was low or the mutation was not fully segregated. Simple and complex structural variants, however, were detected very reliably even at low coverage or in mixed populations, provided they accounted for more than 15% of the reads (Fig 2A, C). With simulated Illumina data, the pipeline detected SNPs, short indels, and simple structural variations even at low coverage and in non-segregated populations, but struggled with complex structural variations (Fig 2B, D). Based on these results, filters in each pipeline were adjusted, and an additional workflow was created that combines the previous two. In this hybrid version, ONT data are used to detect structural variations, while Illumina data are used to detect smaller changes, such as SNPs and indels (S10 Figure).

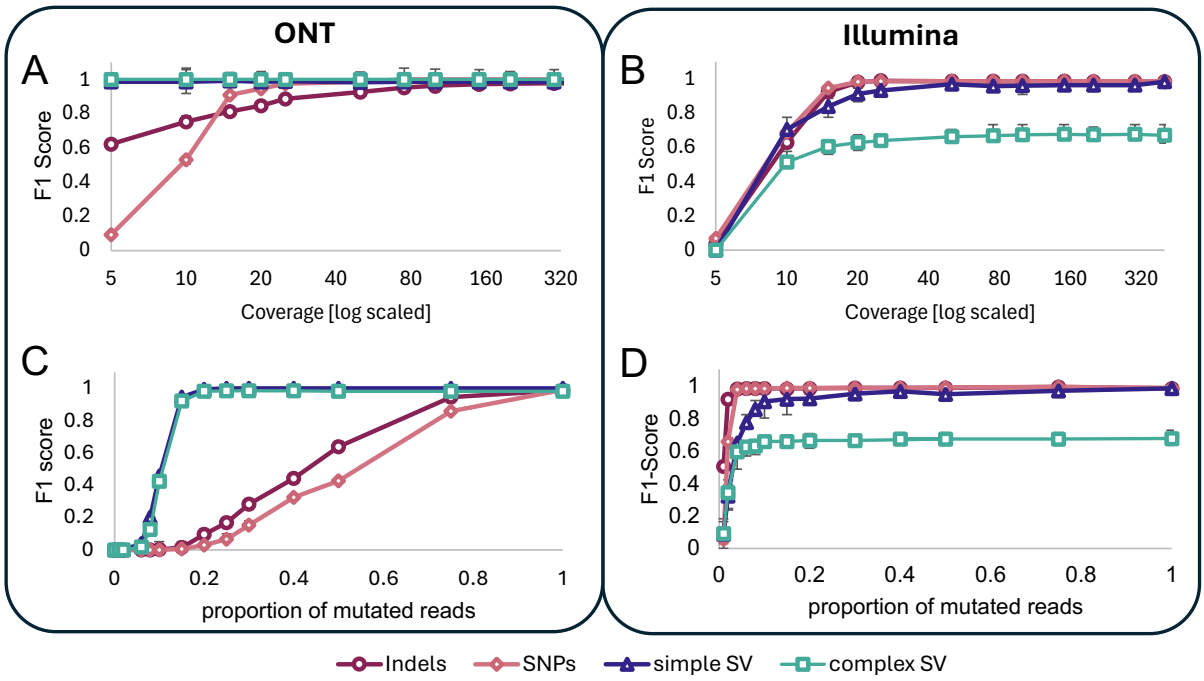

Fig 1: Validation of variant call pipeline for short and long read sequencing data

Variant call pipelines were tested using simulated reads from intentionally altered genomes (see the methods section for details). A) F1-score of variants called from simulated long ONT (Oxford Nanopore Technology) reads with different coverage. B) F1-score of variants called from simulated Illumina reads with different coverage. C) F1-score of variants called from simulated ONT sequencing reads with changing ratios of alternative reads representing non-segregated mutants (combined coverage 225x). (D) F1-score of variants called from simulated Illumina reads with changing ratios of alternative reads representing non-segregated mutants (combined coverage 500x).

Finally, we analyzed the wild-type sequencing data using our pipeline (Fig 3E). Overall, 26 distinct structural variants were identified across all strains; 10 of these (44%) were detected by both techniques, while 12 (52%) were found only in ONT data (Fig 3A). Additionally, many small mutations adjacent to structural variants on extrachromosomal plasmids were called, indicating severe differences from the reference genome at these loci (S6 Table). Apart from these regions, additional small mutations were identified and considered fixed when more than 80% of the reads showed a change at that position. Most of these 55 were identified by both techniques (93%). Additional changes that were not fixed across the population were detected more reliably by Illumina sequencing. Two of the substrains sequenced here (Freiburg

227 and Uppsala) were re-sequenced in other studies before (11,26). Most reported
228 changes were also detected here in both Illumina and ONT sequencing data (Fig 3B,
229 C). However, in both strains, additional fixed mutations were detected here, which have
230 not been reported before. Comparing the three sequenced strains based on structural
231 variants, fixed and non-fixed mutations revealed high similarity between the two non-
232 motile strains, Stockholm and Uppsala. No fixed mutation was detected that was
233 present only in one of these strains, but several non-fixed candidates could be
234 identified. In contrast, several (9) structural variants and many fixed, smaller mutations
235 (21) were found only in the motile strain from Freiburg (S6 Table).
236

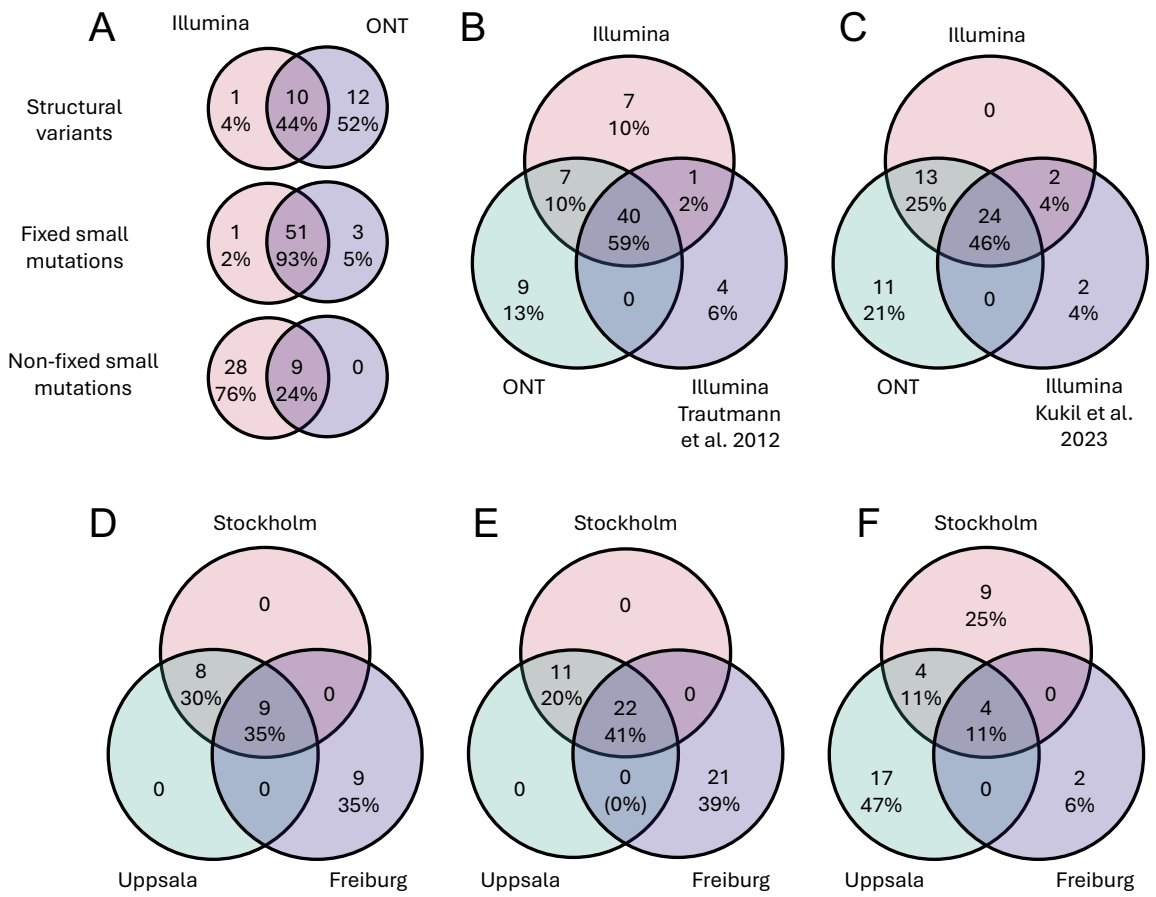

Fig 3: Genetic variations found in different *Synechocystis* wild-type strains using Illumina and Oxford nanopore technology sequencing

Three different wild-type strains from Labs in Stockholm (Sweden), Uppsala (Sweden), and Freiburg (Germany) were sequenced using the Illumina and Oxford Nanopore Technologies (ONT) platforms. Genetic variants were called using different pipelines (see the methods section for details), and the results were compared. A) Venn Diagram of unique structural variants, fixed small mutations (alternative fraction >0.8, length <10bp), and non-fixed small mutations (alternative fraction <0.8, length <10bp) found using Illumina or ONT. B) Comparison of fixed variations (structural variants and small mutations) in the Freiburg strain identified using Illumina sequencing, ONT sequencing, with those published in Trautmann et al. 2012 for this wildtype strain. C) Comparison of fixed variations (structural variants and small mutations) in the Uppsala strain identified using Illumina sequencing, ONT sequencing, with those published in Kukil et al. 2023 for this wildtype strain. D-F) Comparison of the variations found in different wild-type strains for D) structural variants, E) fixed small mutations, and F) non-fixed small mutations.

Case study of 6-phosphogluconate dehydrogenase mutants reveals the strengths and limitations of different complementation strategies and the impact of secondary mutations

To demonstrate how WGS and different complementation strategies can validate a knock-out phenotype, we compared two strains. One was the previously published *gnd* deletion strain (21), which has recently been renamed to Δgnd_zwf^{399+A} due to a secondary mutation in another key enzyme of the OPP pathway (23). The other set of mutants was generated by re-creating the original plasmid and transforming it into the same wild-type strain (Fig 4A). Six independent colonies were picked and restreaked several times to ensure segregation, which was controlled via PCR (S9 File). After that, three segregated strains *(Δgnd_3_1*, *Δgnd_3_2* and *Δgnd_3_3*, hereafter referred to as *Δgnd_3*) were selected, and their growth under heterotrophic conditions was compared with that of the previous mutant and the wild-type. *Δgnd_zwf^{399+A}* grew worse than the wild-type while *Δgnd_3* showed complete growth inhibition (24) (Fig 4B). We complemented both mutants by reintroducing *gnd* at its native locus, which restored heterotrophic growth only in *Δgnd_3*::*gnd,* as expected (Fig 4B, C). We performed Illumina sequencing of both knock-out strains and the corresponding wild-type. Analysis of the data using our pipeline identified six fixed genetic differences unique to gnd_zwf^{399+A} and two mutations only found in the wild-type (Tab 1). The changes found in gnd_zwf^{399+A} were one intergenic mutation, three non-synonymous amino acid changes and 2 frame shifts, including the known mutation in *zwf* (23). The two missense mutations identified in the current wild-type strain were also found in *Δgnd_3*, but no additional changes were identified (S2 Table). Enzyme activity measurements confirmed that only strains containing the zwf mutation, *Δgnd_zwf^{399+A}* and *Δgnd::gnd_zwf^{399+A}*, had no detectable Zwf activity (4D). To assess whether the increased copy number of plasmid-based expression influences knock-out phenotype rescue, we also reintroduced *gnd* into *Δgnd_3* using plasmid-based expression under the native promoter. This *Δgnd_3* P_{gnd}:*gnd* strain also regained heterotrophic growth; however, it still grew more slowly compared to the wild-type and the strain complemented on a genomic basis Δ*gnd_3*::*gnd* (Fig 4C). Consequently, we tested Gnd activity in all strains to estimate *gnd* expression levels (Fig 4E). Gnd activity differed markedly between the two strategies: genomic reintegration produced

287 wild-type-like Gnd activity, whereas plasmid-based expression more than doubled it,
288 indicating increased expression levels.
289

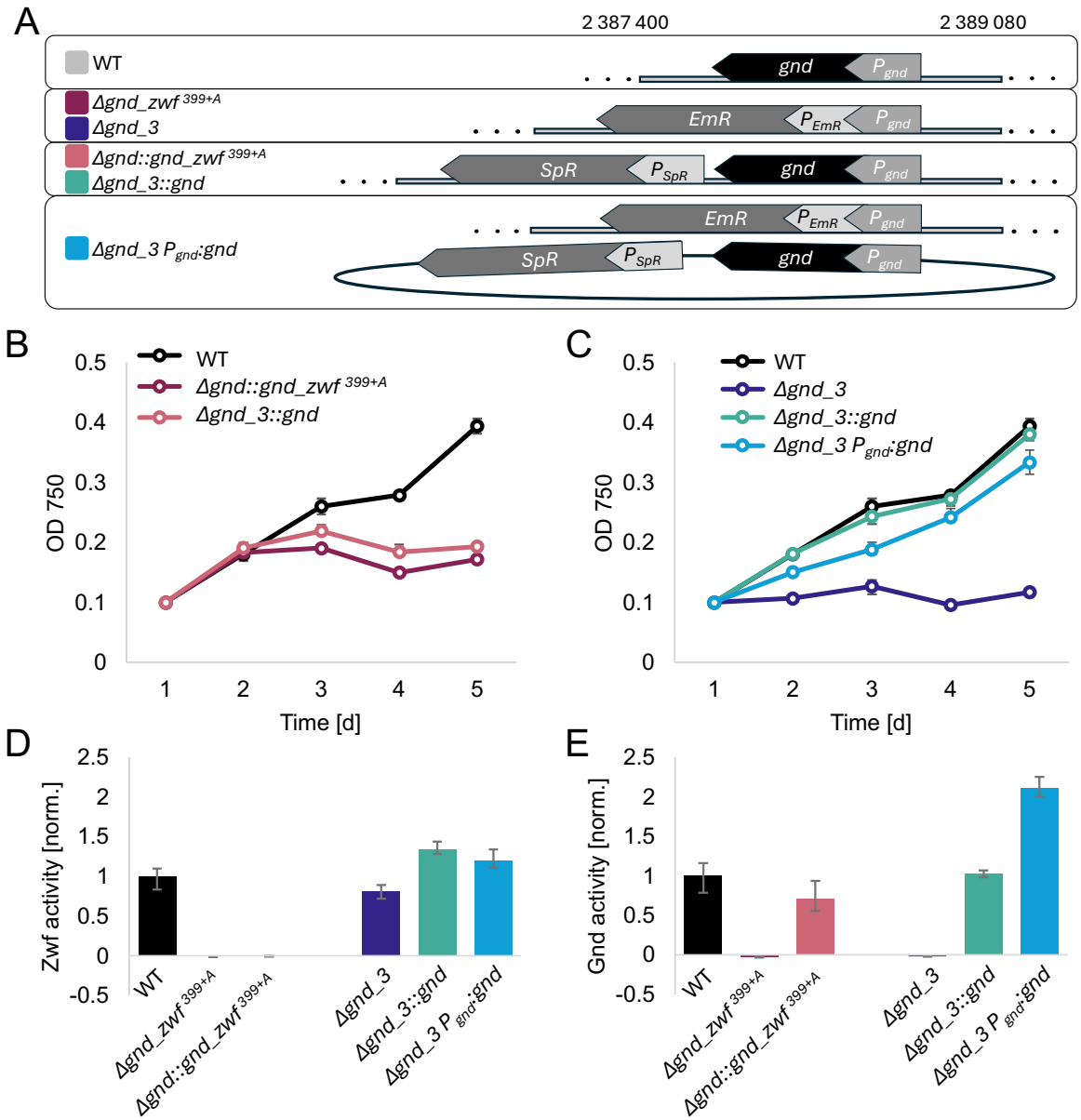

Fig 4: Heterotrophic growth and enzyme activities in different 6-phosphoguconate dehydrogenase (*gnd*) knock-out mutants and complemented strains

A) Overview of the strains used. Three dots indicate a continuing genome. **B)** Heterotrophic growth of wild-type Δ*gnd_zwf399+A* and Δ*gnd_3*) and their corresponding complemented strains were the *gnd* gene was reintroduced in its original locus. **C**) Heterotrophic growth of Δ*gnd_3* strain and different complemented strains: Δ*gnd_3*::*gnd* reintroduction of the Gnd gene at its original locus, Δ*gnd_3* P_{gnd}:*gnd* reintroduction on a stable plasmid containing the Gnd gene under its original promoter. **D**) Zwf enzyme activity in crude cell extract of the strains used in A and B, normalized to the average WT activity. **E**) Gnd activity in crude cell extract of the strains used in A and B, normalized to the average WT activity.

The majority of *Synechocystis* studies do not use complemented strains to control a knock-out phenotype

To investigate how common it is to work with knock-out mutants and to control them using their respective complemented mutants in the current *Synechocystis* literature, we conducted a large-scale, AI-based literature analysis. We screened all full-text, open-access research articles from the Europe PMC database that contained *Synechocystis* in the title or abstract using a zero-shot, large-language-model-based classification workflow (N=902; for details and prompts used, see Methods and S8 File). For each paper, we classified: (i) whether the study focused on *Synechocystis*, (ii) whether a new phenotype from a knock-out strain was described, and (iii) whether this phenotype was controlled using a complementation strain (Fig 4E). Validation of the automated approach using a manually classified random sample of 50 papers revealed precision greater than 0.93 and recall greater than 0.89 (S1 Table). This analysis revealed that after removing 178 publications that were not focused on *Synechocystis*, 337 of 740 (46%) studies reported a new phenotype using knock-out mutants, but only 131 of these 337 (39%) used complemented strains as controls (Fig 4). To determine whether this pattern is specific to studies on *Synechocystis* or common in microbiology, we applied the same method and scoring criteria to studies analyzing other model organisms, *Escherichia coli* and *Saccharomyces cerevisiae* (Fig 4B). In these studies, complemented strains were reported in approximately 63% of *E. coli* related publications (3619 of 5715) and in 55% of *S. cerevisiae* studies (2668 of 4876), showing a marked difference in the respective research communities.

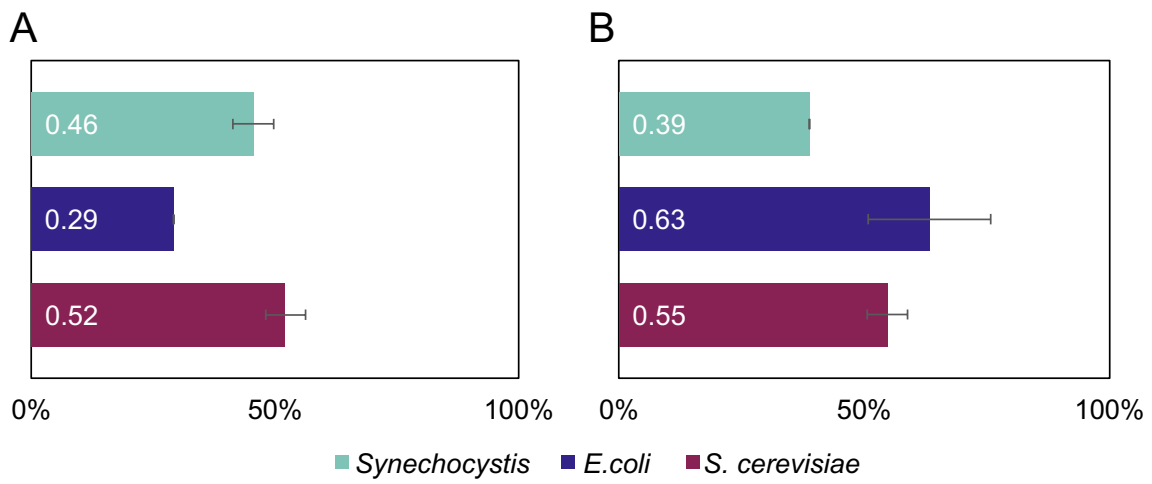

Fig 5: Large language model-based literature analysis about the use of knock-outs and complementation in different model organisms

LLM-based (Mistral-7B-Instruct-v0.3) classification of all full-text, open-access research articles about different model organisms. All publicly available original research articles that mention the model organism in the title or abstract were downloaded from the Europe PMC database (N=901 for “Synechocystis”, 14081 for “cerevisiae”, and 61528 for “coli”). Classified was: whether the study was focused on the model organism (S1_Table), (A) whether a new phenotype from a knock-out was described, and (B) whether this phenotype was controlled using a complementation strain. Error bars represent the average discrepancy relative to 50 random manually classified samples.

A practical guide to working reliably with *Synechocystis* sp. PCC 6803 knock-out mutants

Finally, we present a step-by-step guide based on our findings, suggesting a workflow for robust knock-out experiments (Fig 5). Detailed considerations for each step are provided (S7 Text), covering key aspects from planning and cloning to mutation generation and experimental design. Focusing on the common strategy of gene replacement via homologous recombination, the guide highlights potential pitfalls and offers practical approaches to minimize them using the complementation and sequencing methods described here. Designing individual studies with these recommendations in mind will improve reproducibility, reduce the risk of misleading phenotypes, and strengthen confidence in the link between observed effects and the intended gene deletion.

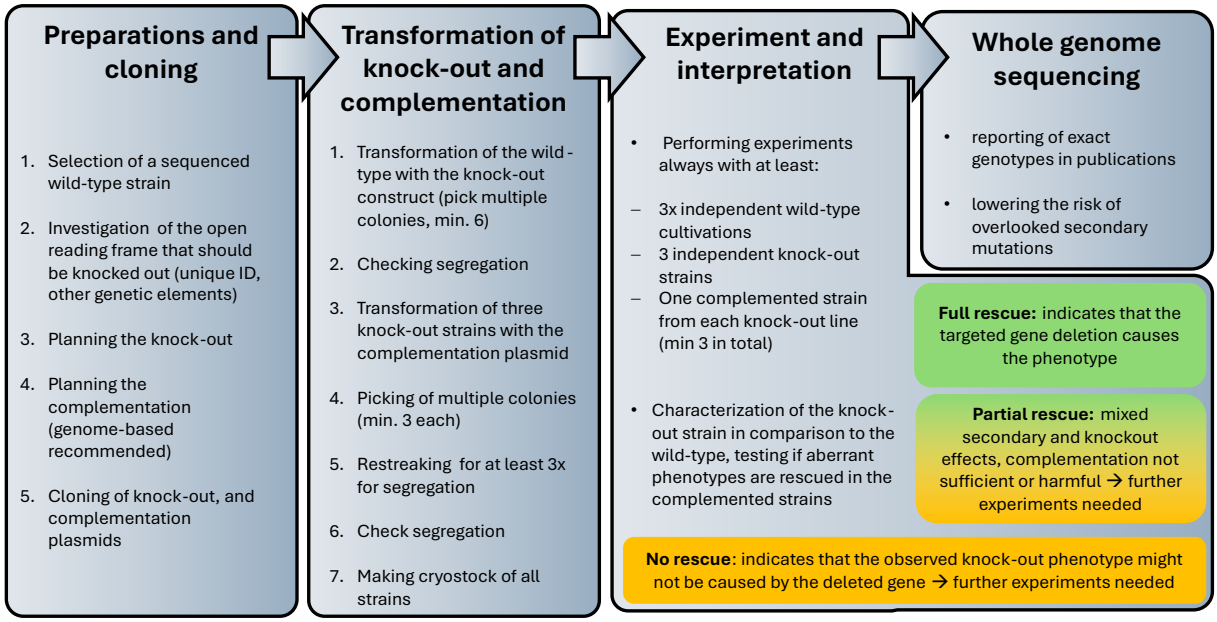

Fig 1: Proposed practical guide to work with knock-out mutants in *Synechocystis* sp. PCC 6803 facilitating robust interpretations

A short, practical guide outlining the main steps for reliable knock-out experiments in *Synechocystis*. Detailed considerations and additional information for many steps highlighted here are available in the supplementary information.

Discussion

Controlling knock-out mutants using complementation or WGS and reporting the exact wild-type strain used in a study are powerful strategies to make scientific results and publications more reliable and reproducible. Here, we investigated various strategies, highlighted their strengths and weaknesses, and provided tools for streamlining genome sequencing data analysis. By comparing short-read Illumina sequencing to ultra-long read ONT sequencing, we illustrated the known strengths and weaknesses of each technique (15,27). Both techniques and pipelines were employed to detect mutations in three different *Synechocystis* wild-type strains. The distribution of structural variants and fixed and non-fixed mutations identified by Illumina, ONT or both replicated the performance expected from simulated datasets. Comparing these results with previous investigations of the same wild-type strains showed that most known mutations were also found in this study. In addition, several new variants were identified that may be specific to the strains used in this study. However, some especially structural variants were found only by ONT sequencing. While this technique has been used to sequence *Synechocystis* before, no detailed reports on changes compared to the reference genome were made there (Battistuzzi et al., 2024, Synechocystis ATCC 27184)). Given the limitations of prior Illumina sequencing, it is likely that these variants are present in other strains as well, but have not been detected to date (11,26). Comparing the three strains revealed high similarity between the Stockholm and Uppsala strains (GT-T), while clear differences were observed between them and the Freiburg strain (PCC-M). This was expected since Stockholm and Uppsala both belong to the non-motile strain, while the Freiburg strain is motile, and many genetic differences between these lines are known (10,11). While no fixed genetic change was found between the two non-motile strains, the non-segregated mutations may reveal hidden variations either within the population or across different copies of the genome in each cell, which could become strain-specific mutations in the future. While most genetic deviations can be reliably detected using one of the techniques shown here, it cannot be ruled out that rare mutations or those in hard-to-sequence regions are still overlooked. In addition, mutations identified through WGS should always be examined by reviewing the mapped data and, if important to the hypothesis, verified using an independent method.

390 Technical advances have significantly reduced the cost of sequencing, making it
391 accessible to a wider community. The pipelines provided here, together with the cost-
392 efficient and reliable sequencing techniques available, may facilitate the sequencing
393 of wild-type and mutant strains to obtain additional information on secondary
394 mutations, genetic background, segregation, and hidden variance in the genome.
395 Publishing this information alongside the experimental data generated using the strains
396 will make experiments more reproducible across the scientific community.

397 We used a case study of two independent *Δgnd* strains to demonstrate how
398 complementation can distinguish true knock-out effects from phenotypes caused by
399 secondary mutations or polar effects. The failed rescue of *Δgnd_zwf*$^{399+A}$ indicated that
400 its phenotype relative to that of the wild-type was not only due to the *gnd* deletion but
401 also to secondary mutations. While further experiments are needed to determine which
402 mutation prevented the rescue, the reported effect can indicate the severity of the
403 genetic change (Tab 1). The effect of missense mutations or ones in intergenic regions
404 can usually only be revealed by detailed investigation or additional experiments. The
405 effect of frameshifts, however, is more severe, especially if the mutation does not occur
406 at the end of the open reading frame. The two frame shifts affected *zwf* and slr1339, a
407 gene coding for a small protein of unknown function. Because Zwf is the only enzyme
408 in *Synechocystis* that produces the substrate for *Gnd*, this mutation most likely
409 prevented complementation-mediated rescue (23). The growth phenotype of Δ*gnd_3*
410 under heterotrophic conditions was more pronounced than that of Δ*gnd_zwf*$^{399+A}$,
411 indicating a distinct phenotype in the *gnd* deletion mutant with an intact *zwf*. This
412 difference highlights the importance of complemented strains as controls, even if the
413 observed knock-out phenotype was expected based on the literature or previous
414 experiments. The slight growth defect observed in *Δgnd_3* P_{gnd}:*gnd* could indicate that
415 increased Gnd activity from plasmid-based expression affects the cells, revealing a
416 fundamental disadvantage of this strategy. The copy number of RSF1010-based
417 plasmids, such as the one used here, is reported to be 1-3 per chromosome, which is
418 well in line with the measured increase in Gnd activity by a factor of two (29,30). The
419 absence of additional mutations in the independently generated *Δgnd_3* strain relative
420 to its wild-type underscores the importance of working with multiple independent lines,
421 while mutations found in both *Δgnd_3* and the corresponding wild-type indicate lab-
422 specific variations (S2 Table). Even if independent strains would have been enough to

423 prevent misinterpretation in this case, multiple studies reported strains that
424 accumulated mutations in similar regions under high selection pressure, making the
425 use of independent lines an insufficient control and the use of complemented strains
426 essential (8,9).

427 The AI-based literature analysis revealed that such controls are not currently used in
428 the majority of knock-out studies. While our manual validation of a 50-paper subset
429 indicated high precision and recall, this sample is relatively small (approximately 5.5%
430 of the Synechocystis dataset). Furthermore, large language models can occasionally
431 misinterpret unconventional or highly complex methodological descriptions,
432 particularly if complementation strategies are not explicitly termed as such.
433 Consequently, the exact percentages reported here should be interpreted as a robust
434 statistical trend rather than an absolute metric. Nevertheless, the large sample size of
435 901 articles provides strong evidence that a significant portion of the community waives
436 these controls. A possible explanation is not a lack of awareness but rather a
437 combination of practical barriers. Due to its polyploidy and slow growth rate compared
438 to other model microorganisms, it usually takes several weeks to generate a single
439 knock-out mutant, which is doubled for complementation (Chauvat et al., 1989).
440 However, skipping these controls can have severe consequences, as shown by
441 investigations regarding the proposed glycolytic Entner-Doudoroff (ED) pathway in
442 *Synechocystis* (21,23). The overlooked secondary mutation in gnd_zwf^{399+A} led to the
443 conclusion that this pathway exists and carries at least a minor flux, despite conflicting
444 evidence (24,31). Recent investigation revealed that both plants and cyanobacteria
445 possess a promiscuous ED aldolase but no ED pathway (23,32,33). Implementing
446 these controls could have prevented this misinterpretation, redirected research efforts
447 earlier, and saved substantial time and resources across multiple laboratories over
448 several years.

449 To reduce the effort required to implement these controls, we provide tools and
450 guidelines for conducting knock-out experiments. This guide is intended as an easy
451 starting point for the workflow that allows customization for the designated purpose;
452 detailed considerations on how to adjust for each experiment are presented (S7 Fig).
453 Overall, the topic addressed in this paper has apparently been somewhat neglected in
454 parts of the *Synechocystis* community, including our labs. However, with the

emergence of novel techniques and their changing availability as costs and efforts decrease, established workflows must be constantly reevaluated and adapted. We strive to improve the robustness and reproducibility of future projects in our lab and others by outlining the necessary steps and providing the tools developed in this study.

460 ## Materials and methods

461 **Cultivation conditions for enzyme activity assays and growth curves**

462 Strains were maintained on BG11 plates at 28 °C, containing antibiotics if needed, and
463 under constant illumination (50–100 µmol m^{-2} s^{-1}).
464 Pre-cultures for growth experiments were grown autotrophically, first in baffled shake
465 flasks for at least 3 d (in 50 ml BG11). For mutants, antibiotics were used in the pre-
466 cultures, but not in growth experiments, unless noted otherwise. For growth
467 experiments, *Synechocystis* sp. PCC 6803 WT (GT strain obtained initially from the
468 McIntosh lab, Michigan State University USA) and mutants were grown in 200 ml BG11
469 medium (pH 8) at 28 °C in glass tubes and equally aerated with filter-sterilized air, as
470 described previously (34). Cultures were inoculated at an OD_{750} of around 0.1 to 0.15
471 and illuminated from two sides with constant light of 50 µmol m^{-2} s^{-1}. Heterotrophic
472 cultures were kept in darkness, except for daily illumination of around 10 min, and 10
473 mM Glucose was added at the beginning of the experiment. To monitor growth,
474 samples were taken from the tube, and OD_{750} was measured daily using a 96-well
475 plate and a plate reader (Infinite M Nano+, Tecan Group AG). In each experiment,
476 three biological replicates were used.
477 The following antibiotic concentrations were used: kanamycin 50 µg ml^{-1},
478 spectinomycin 20 µg ml^{-1}, gentamycin 10 µg ml^{-1} in agar plates, 2.5 to 10 µg ml^{-1} in
479 liquid culture.
480

481 **Cloning of the knock-out plasmids**

482 Plasmids used to create the new *Δgnd_3* strain were obtained by first creating a *gnd*
483 knock-out plasmid. Therefore, the genomic region ~230 bp upstream and downstream
484 of the *gnd* locus was amplified and cloned into a pBluescript plasmid cut with SmaI.
485 Because the *Δgnd_zwf399+A* strain was used as a template, the fragment already
486 contained the Erythromycin resistance cassette and the flanking regions needed for
487 homologous recombination.
488

489 **Cloning of the complementation plasmids**

490 To complement the knock-out strains, three different plasmids were created. First, a
491 plasmid for reintegrating the *gnd* gene into its original locus was constructed.
492 Therefore, a region ~250 bp upstream of the *gnd* gene and the *gnd* gene itself, along

493 with a 250 bp downstream region, were amplified. The fragments were combined into
494 a pBluescript plasmid opened with *Kpn*I and *Sac*I using Gibson assembly. An *EcoR*V
495 restriction site, included in the overlapping primer, enabled the insertion of a
496 spectinomycin resistance cassette downstream of the *gnd* gene and upstream of the
497 *gnd* downstream region.
498 The other two plasmids were based on the RSF1010-based plasmid pSHDY-Prha-
499 mVenus_rhaS (Addgene [a1] plasmid #137662, https://www.addgene.org/137662/,
500 RRID: Addgene_137662; (35)). From this, extraneous *Nde*I restriction sites were
501 eliminated and a spectinomycin resistance was introduced, resulting in pSSRV. From
502 pSSRV, the intermediate plasmid pSSR_C_and_N_6His was generated, in which
503 mVenus was replaced with a synthetic insert encoding both N-terminal and C-terminal
504 His-tags flanked by unique restriction sites, by removing the mVenus insert by cutting
505 with *Nde*I and *Sti*I and replacing the insert with two fragments with overlapping ends.
506 The complementation plasmid was created by cutting this plasmid with *Pst*I and *Hind*III,
507 removing the insert and the rhamnose inducible promoter and the His-tags, and
508 replacing it with a fragment containing the *gnd* gene and a 130bp upstream region
509 presumably containing the *gnd* promoter.
510
511 **DNA extraction and sequencing: Illumina**
512 Cells for Genomic DNA for Illumina sequencing was isolated from *Synechocystis* using
513 a phenol-chloroform method adapted from Makowka et al. (2020). For this, 50 mL liquid
514 culture was centrifuged (8,000 g, 10 min) and the cells were resuspended in 500 µL
515 sterile TE buffer (10 mM Tris and 1 mM EDTA, pH 7.4). Cells were broken by the
516 addition of 10 µL 10 % (w/v) SDS, 500 µL phenol-chloroform-isoamylalcohol solution
517 (25:24:1 (v/v/v)), 500 µL glass beads (0.17-0.18 mm diameter), and by vortexing for 3
518 min at 4°C. The lysate was then centrifuged (2 min, 10,000 xg, RT) and the upper
519 phase, containing the nucleic acids, was transferred to a new centrifugation cup. For
520 further purification, 500 µL phenol-chloroform-isoamyl alcohol solution (25:24:1 (v/v/v),
521 4 °C) was added, and the mixture was centrifuged (2 min, 10,000 xg, RT). The upper
522 phase was transferred to a new cup, mixed with 500 µL chloroform-isoamyl alcohol
523 (24:1 (v/v)), and centrifuged (2 min, 10,000 xg, RT). This step was repeated twice, and
524 the resulting supernatant was transferred to a new cup. A volume of 3 M sodium
525 acetate (pH 5), equivalent to around 10% of the supernatant volume, was added, and

526 the cup was mixed by inverting before adding a volume of absolute ethanol (-20 °C)
527 equivalent to around 250% of the supernatant volume. To precipitate the DNA, the
528 samples were incubated at -20 °C overnight, then centrifuged (15 min, 10,000 xg, 0
529 °C). The supernatant was discarded, and the pellet was washed by adding 1 mL of 70
530 % (v/v) ethanol (-20 °C), followed by centrifugation (5 min, 10,000 g, RT). After
531 discarding the supernatant, the pellet was dried at 50°C for around 5 min and dissolved
532 in 20 µL TE buffer. Optionally, the DNA concentration was measured with a
533 spectrophotometer (DS-11 FX, Biozym, Hessisch Oldendorf, Germany). Samples were
534 normalized to 20 ng/µl, and 500 ng were sent in for short read NovaSeq 150
535 sequencing (Mircrobe-EZ, Genewiz, Germany)
536
537 **Whole-genome sequencing data analysis**
538 All software, unless stated otherwise, was used on the EU Galaxy server(25).
539 **Synthetic genome mutations**
540 All scripts used are available in the supplement. Synthetic mutant genomes were
541 generated using custom Python scripts that insert specified mutations at random sites
542 within the reference genome, while generating Variant Call Format files (.vcf) and tab-
543 separated value files (.tsv) that contain all modified sites and were used as truth
544 states(36). For testing SNPs and small INDELS, 200 sites were inserted with a
545 minimum gap of 3500 bp between them, and 20 % of the inserted mutations were
546 elongations or shortening of a single nucleotide repeat of at least 6 bp. INDEL sizes
547 were equally distributed between -10 bp and 10 bp. Larger insertions and deletions
548 were inserted in a similar way. Sizes were 100 bp, 200 bp, 500 bp, 1000 bp, and 2000
549 bp for insertions and deletions, each repeated 3 times for a total of 30 mutations.
550 Relocations and duplications were performed with equal sizes, also 3 times; sources
551 and targets were chosen with a minimum distance of 3500 bp between them, randomly
552 as before. For relocations, the deletion at the original locus and the insertion at the
553 target were treated as a single mutation, yielding a total of 45 mutations. For each kind
554 of mutation (SNPs, small Indels, simple structural variants, and complex structural
555 variants), five modified genomes were created. Based on the mutated synthetic
556 genomes, synthetic reads were simulated for the Illumina data using ART Illumina (see
557 the supplement for specific tool settings) (37). Different ratios of reads from
558 synthetically mutated genomes and reference ones with a combined coverage of 500x

559 were used to simulate non-fixed mutations. The ONT reads were generated using
560 Nanosim on a local machine with a custom error profile from reads (38). As Nanosim
561 can only operate in' circular' mode when given a single sequence, the template
562 genome was split into its constituent contigs, and the reads were merged at the end of
563 read generation to represent circular multi-chromosome sequencing data. Different
564 coverage was created by subsampling simulated reads. Different ratios of reads from
565 synthetically mutated genomes and reference ones with a combined coverage of 225x
566 were used to simulate non-fixed mutations.
567
568 **Illumina variant discovery workflow**
569 Short reads are first filtered and trimmed using fastP (39) before being mapped using
570 BWA-MEM2 (2.3) (see supplement for specific tool settings) (40). For structural variant
571 calling, the BAM is passed on to two callers, DELLY (0.9.1) and Manta (1.6) (21,41).
572 These caller create variant f iles (.vcf) independently from each other and were filtered
573 based on their internal quality filters, read depth and quality before being merged using
574 JasmineSV (1.0.11) and being passed onto SNPeff for effect annotation (42–44).
575 Simultaneously, the BAM file created by BWA-MEM2 (2.3) was passed onto two SNP
576 and INDEL callers, FreeBayes (1.3.10) and LoFreq (2.1.5) (45,46). These callers
577 independently create variant files (.vcf), which are filtered for quality, read depth,
578 mapping quality, and strand bias before being merged and passed to SNPeff for effect
579 annotation (42)
580
581 **ONT variant discovery workflow**
582 Long reads are first filtered for quality and length using Fastplong before being mapped
583 using Minimap2 (2.28) (see the supplement for specific tool settings) (39,47). For
584 structural variant calling, the BAM file is passed to two callers, namely Sniffles (2.5.2)
585 and cuteSV (2.1.3) (48,49). These caller create variant files (.vcf) independently from
586 each other and were filtered based on quality, internal filters, variant length and reads
587 supporting it filters before having their insertion site and size was refined by IRIS (1.0.5)
588 and subsequently being merged using JasmineSV (1.0.11), to account for the
589 possibility of an caller calling no variants, a pseudo variant was appended to each file
590 and filtered again after merging (44). After merging, the VCF is sorted again and
591 passed on to SNPeff (5.2) for variant effect annotation (43). At the same time, the BAM

592 file created by Minimap2 (2.28) was passed to Clair3 (1.0.10) to call indels and SNPs
593 (50). Its output was again filtered, sorted and additionally normalised and left aligned
594 before the effects were again annotated by SNPeff (5.2) (36).
595
596 **Joint Illumina and ONT variant discovery workflow**
597 The joint variant discovery workflow utilized the strengths of both sequencing methods.
598 Large structural variants called by Sniffles (2.5.2) and cuteSV (2.1.3) using ONT reads,
599 SNPs and Indels being called by FreeBayes (1.3.10) and LoFreq (2.1.5) using Illumina
600 reads (45,46,48,49).
601
602 **Workflow benchmarking**
603 The outputs of the variant discovery workflows were compared with the true table of
604 synthetic genome mutations at the 5' site. True positives were counted if the called
605 positions were within a 10 bp window of the true one for SNPs and small INDELS, and
606 a 100 bp window for structural variants. False positives were counted if the called
607 positions were not in the start or end set of the true table, using the same buffer as
608 before.
609
610 **DNA extraction and sequencing: ONT**
611 ONT sequencing was performed for three *Synechocystis* WT substrains. The
612 Stockholm strain, generously provided by Paul Hudson, is a non-motile glucose-
613 tolerant substrain (GT-T). The Uppsala strain, a glucose-tolerant, non-motile
614 *Synechocystis* substrain, was kindly gifted by Pia Lindberg (GT-T). The motile PCC-
615 M, strain from Freiburg, which was kindly sent years ago by Annegret Wilde.
616 Cryopreserved *Synechocystis* WT strains were streaked on 1% agar plates with BG11
617 and incubated in a Multitron HT Incubator (Infors, Switzerland) (30 °C, 150 rpm
618 shaking, 1 % CO2, 75 % humidity, 32 µmol(photons)·m-2 ·s-1. 20 ml BG11 in 100 ml
619 Erlenmeyer flasks were inoculated and cultures were incubated under the same
620 conditions and shaking at 150 rpm until an OD750 nm of ~7-8. Cells from 2 ml culture
621 were harvested by centrifugation for 15 min at 10,000 rcf, resuspended in 1ml 0.1M
622 EDTA (pH 8), and incubated for 3 ½ h at 300 rpm and 24°C to remove EPS. After
623 centrifugation at 10,000 rcf for 10 min, the pellets were either washed once with 1 ml
624 MiliQ (Freiburg strain) or directly stored at -20°C (Uppsala and Stockholm strains) for

625 at least one night. Cells were resuspended in 400 µl XS buffer (1% potassium
626 ethylxanthogenate, 800 mM NH_4OAc, 100 mM Tris-HCl pH 7.4, 20mM EDTA, pH8.0,
627 1% SDS) and incubated for ~2h at 65°C, with three freeze-thaw cycles in liquid
628 nitrogen. For DNA extraction, 1 volume of phenol:chloroform: isoamyl alcohol (25:24:1)
629 was added and mixed gently by inversion. After centrifugation at 10,000 rcf and 20°C
630 for 10 min, the aqueous phase was retrieved, treated with 5µl RNAseA (4.5U in 10
631 mM Tris pH8.5, 15mM NaCl) and subsequently incubated at 37°C and RT overnight.
632 Phenol:chloroform:isoamylalcohol extraction were repeated four times before adding
633 1 volume of chloroform or chloroform:isoamylalcohol. Samples were gently inverted
634 and centrifuged at 10,000 rcf and 20°C for 10 min, and the aqueous phase was re-
635 isolated twice with chloroform:isoamylalcohol. DNA in the aquesous phase was
636 precipitated with 0.6 ml ice-cold EtOH and 0.2M NaCl (final concentration) at -20°C for
637 at least 1 night. DNA was pelleted by centrifugation at 10,000 rcf for 30 min, washed
638 twice with 0.4 ml 70% EtOH, air-dried, and resuspended in 20 µl EB buffer (Tris-HCl)
639 by subsequent incubation at 50°C and RT overnight.
640 DNA samples were further processed by the Genomics and Transcriptomics
641 Laboratory (GTL) at HHU. The DNA concentration was measured using the Qubit
642 dsDNA HS assay. Sample quality was assessed photometrically using a
643 NanoDropOne. The 260/280 and 260/230 ratios were both ≥1.92. DNA integrity
644 (fragment length) was controlled via capillary gel electrophoresis using the DNF-464
645 kit and a Fragment Analyzer 5300. The mean fragment size ranged between 24000-
646 35000 bp and all samples had an integrity QC value ≥9.9. The ONT sequencing was
647 performed by GTL on a PromethION using an R10.4.1 nanopore and Kit 14 adapter
648 barcoding, and basecalled with the Highaccuracy model v4.3.0 (400 bps).
649
650 **Enzyme activity assays**
651 Glucose-6-P dehydrogenase and 6-P-gluconate dehydrogenase activity was
652 measured by first harvesting 50 ml of cells from the exponential growth phase
653 cultivated in Kniese tubes (as described above) at an OD750 ~1. Cells were
654 resuspended in 500 µl Tris-HCl 0.1 M pH 7.4 and lysed by vortexing (6 cycles, 10 s
655 max speed) with 200 µl 0.18 mm glass beads. Samples were centrifuged at 12000 rcf
656 for 1 min to remove unbroken cells and glass beads. The supernatant was transferred
657 to a new cup and centrifuged for 20 min at 12000 rcf and 4°C to sediment the

658 membrane fraction. The soluble fraction was transferred to a new cup. The absorption
659 at 650 nm (phycobilisomes) was used to normalize the samples. 50 µl of the
660 normalized cell extract was mixed with 915 µl of Tris-HCl (0.1 M, pH 7.4) and 10 µl 0.1
661 M $NADP^+$. 195 µl of this mixture was added to a 96-well plate three times, and the
662 absorbance at 340 nm was measured for 3 min at 30°C in a plate reader (Tecan,
663 Switzerland). 5 µl glucose-6-P (for glucose-6-P dehydrogenase assays) or 6-P-
664 gluconate (for 6-P-gluconate dehydrogenase assays) was added using the injection
665 system of the plate reader, and the absorption at 340 nm was measured continuously
666 for 10 minutes. The slope after substrate injection was calculated and compared to
667 estimate enzyme activity.
668

669 **Plasmids, primers and strains used**

670 Detailed information about the plasmids, primers, and strains used in this
671 study is available in the supplementary information.

672 **Large language model-based literature analysis**

673 First, all open-access, original research articles containing "*Synechocystis*" were
674 downloaded from Europe PMC as XML files using its API and converted to a CSV file
675 using a Python script (51). Then the large language model Mistral-7B-Instruct-v0.3 (52)
676 was downloaded and installed on the University of Kassel's high-performance Linux
677 (Rocky Linux 9). Using a custom Python script title, abstract, results, and methods
678 sections of each paper were combined with a prompt containing several definitions,
679 detailed commands on how to format the answer and three questions: i) Does the study
680 present original experimental data obtained using or about *Synechocystis*? ii) Does the
681 study present original experimental data obtained using a *Synechocystis* knock-out
682 mutant? iii) Does the study include experimental data from a complementation/rescue
683 strain to test whether the phenotype is specifically due to the gene knock-out? The
684 answers were phrased to elicit Boolean responses, reasoning, and evidence for each
685 question, and the results were combined with the input data into a single CSV file using
686 Python. Fifty papers were randomly selected and categorized using the same
687 questions to enable precision and recall calculations. All scripts and prompts are
688 provided (S8 File).

Data availability

All sequencing data are available at BioProject number: PRJNA1450533 https://dataview.ncbi.nlm.nih.gov/object/PRJNA1450533?reviewer=easj36tflkd8u1bk3mt1pkgavu

Galaxy Pipelines:
Illumina variant calling:
https://usegalaxy.eu/u/niklaskueppers/w/illumina-variant-discovery
ONT variant calling:
https://usegalaxy.eu/u/niklaskueppers/w/ont-variant-discovery
Hybrid variant calling:
https://usegalaxy.eu/u/niklaskueppers/w/joint-illumina-ont-variant-discovery

Acknowledgments

ONT sequencing was performed at the Genomics and Transcriptomics Laboratory at Heinrich-Heine University Düsseldorf. We thank Tobias Lautwein and Karl Köhrer for their support. Computational infrastructure and support were provided by the Center for Information and Media Technology at Heinrich Heine University Düsseldorf. The authors acknowledge the support of the Freiburg Galaxy Team: Bioinformatics, University of Freiburg (Germany). We thank Marc von der Heiden, Aleksandar Dobrota and Andrej Dobrota for initial assistance in method testing and pipeline exploring. Many thanks to Annegret Wilde, Pia Lindberg and Paul Hudson for kindly providing the *Synechocystis* WT strains. We thank Michelle Gehringer for the DNA extraction protocol and her valuable advice.

Funding

This study was supported by grants from the German Science Foundation (DFG Gu1522/5-1, GRK2749/1, Ax 84/10-1 - Project ID: 573342561, SFB1535 – Project ID 45809066), the Dietmar-Hopp-Stiftung, Bundesministerium für Bildung und Forschung (BMBF) in the framework of the project CyFun (03SF0652A).

Conflict of interest

The authors declare that they have no conflicts of interest.

Author contributions

MT, RF, MB, FP, IMA and KG conceptualized the study. MT, RF, NK, MB, PK and FP investigated, MT visualized, MT, NK, PK, OP and FP development of methology. OP, AW, EO, IMA and KG supervised. IMA, KG acquired funding. MT wrote the original draft, all authors reviewed and edited the manuscript.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author(s) used ChatGPT 5.2 (OpenAI, 2025) to correct Python code and ChatGPT 5.2 (OpenAI, 2025), Grammarly and Paperpal to optimize language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the published article.

Supporting Information

S1 Table	AI based classification data
S2 Table	Table of mutations found in *Δgnd* and corresponding wild-type
S3 File	*Synechocystis* PCC 6803 reference genome
S4 Text	Manual for pipeline usage
S5 Text	Macro for Excel to import CSV results and create a pivot table for comparison
S6 Table	Mutations found in Synechocystis wild-types
S7 Text	Detailed considerations for creating knock-outs in *Synechocystis*
S8 File	Python scripts and prompts
S9 File	Primers, plasmids and strains used, validation of segregation via PCR
S10 Figure	A more detailed overview of the different pipelines
S11 Figure	Recall and precision of ONT and Illumina pipelines

References

1. Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, et al. 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. 1996 Jan 1;3(3):109–36. doi:10.1093/dnares/3.3.109

2. Kufryk GI, Sachet M, Schmetterer G, Vermaas WFJ. Transformation of the cyanobacterium *Synechocystis* sp. PCC 6803 as a tool for genetic mapping: optimization of efficiency. FEMS Microbiol Lett. 2002 Jan;206(2):2. doi:10.1111/j.1574-6968.2002.tb11012.x

3. Lindberg P, Park S, Melis A. Engineering a platform for photosynthetic isoprene production in cyanobacteria, using Synechocystis as the model organism. Metab Eng. 2010 Jan;12(1):70–9. doi:10.1016/j.ymben.2009.10.001 PubMed PMID: 19833224.

4. Klaus O, Hilgers F, Nakielski A, Hasenklever D, Jaeger KE, Axmann IM, et al. Engineering phototrophic bacteria for the production of terpenoids. Curr Opin Biotechnol. 2022 Oct;77:102764. doi:10.1016/j.copbio.2022.102764 PubMed PMID: 35932511.

5. Chauvat F, Rouet P, Bottin H, Boussac A. Mutagenesis by random cloning of an Escherichia coli kanamycin resistance gene into the genome of the cyanobacterium Synechocystis PCC 6803: selection of mutants defective in photosynthesis. Mol Gen Genet MGG. 1989 Mar;216(1):51–9. doi:10.1007/BF00332230 PubMed PMID: 2499763.

6. Zerulla K, Ludt K, Soppa J. The ploidy level of Synechocystis sp. PCC 6803 is highly variable and is influenced by growth phase and by chemical and physical external parameters. Microbiology. 2016 May 1;162(5):730–9. doi:10.1099/mic.0.000264

7. Pope MA, Hodge JA, Nixon PJ. An Improved Natural Transformation Protocol for the Cyanobacterium Synechocystis sp. PCC 6803. Front Plant Sci. 2020 Apr 15;11. doi:10.3389/fpls.2020.00372

8. Kobayashi M, Okada K, Ikeuchi M. A Suppressor Mutation in the α-Phycocyanin Gene in the Light/Glucose-sensitive Phenotype of the psbK-disruptant of the Cyanobacterium Synechocystis sp. PCC 6803. Plant Cell Physiol. 2005 Sep 1;46(9):1561–7. doi:10.1093/pcp/pci169

9. Nishijima Y, Kanesaki Y, Yoshikawa H, Ogawa T, Sonoike K, Nishiyama Y, et al. Analysis of spontaneous suppressor mutants from the photomixotrophically grown pmgA-disrupted mutant in the cyanobacterium Synechocystis sp. PCC 6803. Photosynth Res. 2015 Dec;126(2–3):465–75. doi:10.1007/s11120-015-0143-8 PubMed PMID: 25869635.

10. Kanesaki Y, Shiwa Y, Tajima N, Suzuki M, Watanabe S, Sato N, et al. Identification of Substrain-Specific Mutations by Massively Parallel Whole-Genome

Resequencing of Synechocystis sp. PCC 6803. DNA Res. 2012 Feb 1;19(1):67–79. doi:10.1093/dnares/dsr042

11. Trautmann D, Trautmann D, Voß B, Voß B, Wilde A, Wilde A, et al. Microevolution in Cyanobacteria: Re-sequencing a Motile Substrain of Synechocystis sp. PCC 6803. DNA Res. 2012. doi:10.1093/dnares/dss024

12. Koskinen S, Kurkela J, Linhartová M, Tyystjärvi T. The genome sequence of Synechocystis sp. PCC 6803 substrain GT-T and its implications for the evolution of PCC 6803 substrains. FEBS Open Bio. 2023;13(4):701–12. doi:10.1002/2211-5463.13576

13. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016 May 17;17(6):333–51. doi:10.1038/nrg.2016.49 PubMed PMID: 27184599; PubMed Central PMCID: PMC10373632.

14. Mikheyev AS, Tin MMY. A first look at the Oxford Nanopore MinION sequencer. Mol Ecol Resour. 2014 Nov;14(6):1097–102. doi:10.1111/1755-0998.12324 PubMed PMID: 25187008.

15. Purushothaman S, Roloff T, Egli A, Seth-Smith HM. Benchmarking Illumina and Oxford Nanopore Technologies (ONT) sequencing platforms for whole genome sequencing of bacterial genomes and use in clinical microbiology. BMC Med Genomics. 2026 Jan 15. doi:10.1186/s12920-025-02305-2

16. Bertelli C, Greub G. Rapid bacterial genome sequencing: methods and applications in clinical microbiology. Clin Microbiol Infect. 2013 Sep 1;19(9):803–13. doi:10.1111/1469-0691.12217 PubMed PMID: 23601179.

17. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, et al. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics. 2015;15(2):141–61. doi:10.1007/s10142-015-0433-4 PubMed PMID: 25722247; PubMed Central PMCID: PMC4361730.

18. Segerman B. The Most Frequently Used Sequencing Technologies and Assembly Methods in Different Time Segments of the Bacterial Surveillance and RefSeq Genome Databases. Front Cell Infect Microbiol. 2020 Oct 19;10. doi:10.3389/fcimb.2020.527102

19. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends Genet TIG. 2008 Mar;24(3):142–9. doi:10.1016/j.tig.2007.12.007 PubMed PMID: 18262676; PubMed Central PMCID: PMC2680276.

20. Watson JD. Molecular Biology of the Gene. Pearson; 2014. 872 p.

21. Chen X, Schreiber K, Appel J, Makowka A, Fähnrich B, Roettger M, et al. The Entner–Doudoroff pathway is an overlooked glycolytic route in cyanobacteria and plants. Proc Natl Acad Sci. 2016 May 10;113(19):19. doi:10.1073/pnas.1521916113

22. Jansén T, Kurian D, Raksajit W, York S, Summers ML, Mäenpää P. Characterization of trophic changes and a functional oxidative pentose phosphate pathway in Synechocystis sp. PCC 6803. Acta Physiol Plant. 2010 May 1;32(3):511–8. doi:10.1007/s11738-009-0428-7

23. Ojha RS, Theune M, Fritsche R, Makowka A, Boehm M, Peraglie C, et al. Story about honest mistakes: The cyanobacterium Synechocystis has a promiscuous Entner-Doudoroff (ED) aldolase but no functional ED pathway. [Internet]. bioRxiv; 2026 [cited 2026 Apr 3]. p. 2026.04.01.715859. Available from: https://www.biorxiv.org/content/10.64898/2026.04.01.715859v1 doi:10.64898/2026.04.01.715859

24. Makowka A, Nichelmann L, Schulze D, Spengler K, Wittmann C, Forchhammer K, et al. Glycolytic Shunts Replenish the Calvin–Benson–Bassham Cycle as Anaplerotic Reactions in Cyanobacteria. Mol Plant. 2020 Mar;13(3):3. doi:10.1016/j.molp.2020.02.002

25. The Galaxy Community. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. Nucleic Acids Res. 2024 Jul 5;52(W1):W83–94. doi:10.1093/nar/gkae410

26. Kukil K, Englund E, Crang N, Hudson EP, Lindberg P. Laboratory evolution of *Synechocystis* sp. PCC 6803 for phenylpropanoid production. Metab Eng. 2023 Sep 1;79:27–37. doi:10.1016/j.ymben.2023.06.014

27. Linde J, Brangsch H, Hölzer M, Thomas C, Elschner MC, Melzer F, et al. Comparison of Illumina and Oxford Nanopore Technology for genome analysis of Francisella tularensis, Bacillus anthracis, and Brucella suis. BMC Genomics. 2023 May 12;24(1):258. doi:10.1186/s12864-023-09343-z

28. Battistuzzi M, Morlino MS, Cocola L, Trainotti L, Treu L, Campanaro S, et al. Transcriptomic and photosynthetic analyses of Synechocystis sp. PCC6803 and Chlorogloeopsis fritschii sp. PCC6912 exposed to an M-dwarf spectrum under an anoxic atmosphere. Front Plant Sci. 2024 Jan 18;14. doi:10.3389/fpls.2023.1322052

29. Ng WO, Zentella R, Wang Y, Taylor JSA, Pakrasi HB. phrA, the major photoreactivating factor in the cyanobacterium Synechocystis sp. strain PCC 6803 codes for a cyclobutane-pyrimidine-dimer-specific DNA photolyase. Arch Microbiol. 2000 Jun 1;173(5):412–7. doi:10.1007/s002030000164

30. Huang HH, Camsund D, Lindblad P, Heidorn T. Design and characterization of molecular tools for a Synthetic Biology approach towards developing cyanobacterial biotechnology. Nucleic Acids Res. 2010 May;38(8):2577–93. doi:10.1093/nar/gkq164 PubMed PMID: 20236988; PubMed Central PMCID: PMC2860132.

31. Schulze D, Kohlstedt M, Becker J, Cahoreau E, Peyriga L, Makowka A, et al. GC/MS-based 13C metabolic flux analysis resolves the parallel and cyclic photomixotrophic metabolism of Synechocystis sp. PCC 6803 and selected deletion mutants including the Entner-Doudoroff and phosphoketolase pathways.

868 Microb Cell Factories. 2022 Apr 22;21(1):69. doi:10.1186/s12934-022-01790-9
869 PubMed PMID: 35459213; PubMed Central PMCID: PMC9034593.

870 32. Evans SE, Franks AE, Bergman ME, Sethna NS, Currie MA, Phillips MA. Plastid
871 ancestors lacked a complete Entner-Doudoroff pathway, limiting plants to
872 glycolysis and the pentose phosphate pathway. Nat Commun. 2024 Feb 6;15(1):1.
873 doi:10.1038/s41467-024-45384-y

874 33. Xie N, Sharma C, Rusche K, Wang X. Phosphoketolase and KDPG aldolase
875 metabolisms modulate photosynthetic carbon yield in cyanobacteria. Plant Cell.
876 2024 Dec 23;37(1):koae291. doi:10.1093/plcell/koae291

877 34. Theune ML, Hildebrandt S, Steffen-Heins A, Bilger W, Gutekunst K, Appel J. In-
878 vivo quantification of electron flow through photosystem I - Cyclic electron transport
879 makes up about 35% in a cyanobacterium. Biochim Biophys Acta Bioenerg. 2021
880 Mar 1;1862(3):3. doi:10.1016/j.bbabio.2020.148353 PubMed PMID: 33346012.

881 35. Behle A, Saake P, Germann AT, Dienst D, Axmann IM. Comparative Dose–
882 Response Analysis of Inducible Promoters in Cyanobacteria. ACS Synth Biol. 2020
883 Apr 17;9(4):4. doi:10.1021/acssynbio.9b00505

884 36. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The
885 variant call format and VCFtools. Bioinformatics. 2011 Aug;27(15):2156–8.
886 doi:10.1093/bioinformatics/btr330

887 37. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read
888 simulator. Bioinformatics. 2012 Feb 15;28(4):593–4.
889 doi:10.1093/bioinformatics/btr708

890 38. Yang C, Chu J, Warren RL, Birol I. NanoSim: nanopore sequence read simulator
891 based on statistical characterization. GigaScience. 2017 Apr 1;6(4):gix010.
892 doi:10.1093/gigascience/gix010

893 39. Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and
894 deduplication using fastp. iMeta. 2023 May;2(2):e107. doi:10.1002/imt2.107

895 40. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-
896 MEM [Internet]. arXiv; 2013 [cited 2026 Jan 13]. Available from:
897 http://arxiv.org/abs/1303.3997 doi:10.48550/arXiv.1303.3997

898 41. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural
899 variant discovery by integrated paired-end and split-read analysis. Bioinformatics.
900 2012 Sep;28(18):i333–9. doi:10.1093/bioinformatics/bts378

901 42. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
902 Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug;25(16):2078–9.
903 doi:10.1093/bioinformatics/btp352

904 43. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for
905 annotating and predicting the effects of single nucleotide polymorphisms, SnpEff:
906 SNPs in the genome of Drosophila melanogaster strain w^{1118}; iso-2; iso-3. Fly

(Austin). 2012 Apr;6(2):80–92. doi:10.4161/fly.19695

44. Kirsche M, Prabhu G, Sherman R, Ni B, Battle A, Aganezov S, et al. Jasmine and Iris: population-scale structural variant comparison and analysis. Nat Methods. 2023 Mar;20(3):408–17. doi:10.1038/s41592-022-01753-3

45. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing [Internet]. arXiv; 2012 [cited 2026 Jan 14]. Available from: https://arxiv.org/abs/1207.3907 doi:10.48550/ARXIV.1207.3907

46. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 2012 Dec;40(22):11189–201. doi:10.1093/nar/gks918

47. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor. Bioinformatics. 2018 Sep;34(18):3094–100. doi:10.1093/bioinformatics/bty191

48. Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, et al. Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. 2020 Dec;21(1):189. doi:10.1186/s13059-020-02107-y

49. Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol. 2024 Oct;42(10):1571–80. doi:10.1038/s41587-023-02024-y

50. Zheng Z, Li S, Su J, Leung AWS, Lam TW, Luo R. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling [Internet]. Bioinformatics; 2021 [cited 2026 Jan 14]. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.12.29.474431 doi:10.1101/2021.12.29.474431

51. Rosonovski S, Levchenko M, Bhatnagar R, Chandrasekaran U, Faulk L, Hassan I, et al. Europe PMC in 2023. Nucleic Acids Res. 2023 Nov 22;52(D1):D1668–76. doi:10.1093/nar/gkad1085 PubMed PMID: 37994696; PubMed Central PMCID: PMC10767826.

52. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas D de las, et al. Mistral 7B [Internet]. arXiv; 2023 [cited 2026 Mar 17]. Available from: http://arxiv.org/abs/2310.06825 doi:10.48550/arXiv.2310.06825